A Microcomputer-Based Vital Records Data Base with Interactive Graphic Assessment for States and Localities

DANIEL WARTENBERG, PHD, VINCENT J. AGAMENNONE, MS, DAVID OZONOFF, MD, AND R. J. BERRY, MD

Abstract: Vital records data bases describe large populations over long periods of time, yet their organization and size often preclude or discourage their use. We constructed a microcomputer-based data base of all singleton births in Massachusetts, 1975–84. The original data were stored in 700,000 records, each 174 bytes long, occupying a total of over 120 megabytes (MB). By removing redundant information and unique identifiers, and packing the data, we store 21 fields of this information in a 16-byte record resulting in a data base of 11.1 MB, a saving of over 90 percent of disk space. By using programs written expressly for this data base, we can display a birth weight frequency plot of the entire data set in under 65 seconds on an IBM-compatible PC-AT. Comparable assessments in SAS-PC took over 105 minutes and in main frame SAS on an AS-9000 took over 37 CPU seconds. Implementing similar systems for state registries on births, deaths, cancers, and birth defects potentially offers investigators easy access to vast stores of information and would enable public health officials to produce timely reports, initiate a variety of surveillance activities, and respond rapidly to residents' inquiries about clusters and anomalous disease patterns. (*Am J Health* 1989; 79:1531–1536.)

Introduction

For each of the 3.5 million births and 2 million deaths in the United States per year, a formal record is made of each event and a certificate is filed with the local health department which includes information on a selection of socioeconomic, demographic, and occupational variables. These birth and death certificates contain a vast store of information which can be used for many types of epidemiological studies and for surveillance purposes.[1–5] Historically, these vital records data were kept in file cabinets in state government agencies. Later, much of the information was transferred to punched cards. Now, most states use computer-based systems, with records listed by date of filing (not date of event) on a main frame computer-based magnetic tape. To access most vital records data bases is still difficult, slow, and cumbersome. Further, a variety of state-specific confidentiality restrictions and lack of formal access procedures may prevent investigators from using these data.

As an example of the use of these data, public health officials frequently need descriptive statistics on reproductive outcome, cause of death, or cancer rates. Developing these data requires laborious compilation of data from registries, integration with census-based population estimates, and evaluation. If anomalous patterns are found, the process may have to be repeated to provide summary information from nearby municipalities. Completion of the entire evaluation often takes weeks or months.

To make rapid display and analysis possible, we devised a novel approach to data storage and retrieval using new but readily available technology. We designed software to take advantage of current microcomputer hardware. While the system we describe is for Massachusetts birth certificates, we envision similar systems for rapid retrieval, analysis and display of death certificate data, reportable and occupational diseases, Surveillance, Epidemiology and End Result (SEER) registry data, and various other registry and health outcome data.

In developing our system, we minimized both the computer space and operator time requirements. We also designed the system for a relatively inexpensive and commonly available type of microcomputer (an IBM-compatible 80286 AT-type computer, 8Mhz, no wait states, with an EGA graphics card). We believe that this type of computer is well within the financial resources of even small, local health departments and that using our approach will make it easier to review and analyze vital records and reportable disease information on a routine basis. While many others have conducted such studies using similar data, the specific program design features of our system (e.g., data packing, record retrieval via pointer files) make it unusually fast and easy to use even for computer novices. Below, we present an overview of the design of our system and the first program we have developed to review and analyze the data. We believe that our system has the potential to benefit public health investigations and surveillance, and also has the capacity to enable local officials to respond rapidly and efficiently to their constituents.

Methods

Creating the Data Base

The research project motivating development of this data system was a study of the statistical distribution of birth weights in Massachusetts requiring rapid display and summary of data for exploratory statistical analyses. In addition, we wanted to identify municipalities within unusual birth weight distributions that might warrant in-depth investigation or surveillance.

For our study we obtained the birth certificate file of all births in Massachusetts from 1975–84 from the Massachusetts Department of Public Health. The file contains more than 735,000 records. The format of the original data tape is shown in the Appendix. Data were culled to exclude all records of births to out-of-state residents, all non-singleton births (which we chose not to analyze), and all records without birth weight information. The resulting data base contains 696,460 records.

Each birth record, as originally provided, has 174 bytes of information. (Each byte is a single character of information such as an individual numeral or letter.) Personal identifiers (baby's name, certificate number, mother's maiden name,

Address reprint requests to Daniel Wartenberg, PhD, Assistant Professor, Department of Environmental and Community Medicine, UMDNJ—Robert Wood Johnson Medical School, Piscataway, NJ 08854. Mr. Agamennone is also affiliated at UMDNJ; Dr. Ozonoff is with the Environmental Health Section, Boston University School of Public Health; Dr. Berry is with the Division of Birth Defects and Developmental Disabilities, CDC, Atlanta. This paper, submitted to the Journal July 18, 1988, was revised and accepted for publication April 13, 1989.

etc.—63 bytes) were removed from all records before we received them to protect individual privacy. This left 111 bytes of relevant information. We dropped 26 bytes of information not related to our study (e.g., multiple birth information, baby's vital status at time of report, etc.) leaving 85 bytes.

Among the 85 bytes retained, we identified 33 bytes of redundant information. These represented fields in which information was coded directly from the birth certificate, such as birth weight in pounds and ounces or date of last menses, and then was converted to another unit or type of reference, such as birth weight in grams or gestational age. Because of the need to minimize space, we removed all such calculated values, reasoning that we could recalculate the values later, if needed. Finally, two 1-byte fields of complementary data were combined into a single byte. This reduced the records to 51 bytes of usable information. This information still required over 36 million bytes or 36 megabytes (MB) for storage.

To further reduce the mass storage requirements of our data base, we packed our data in fractional bytes rather than using the traditional integer representations for these categorical variables (see Appendix for details). While most computer programs access data on a byte-by-byte basis, we have packed data on a bit-by-bit scale (each byte has 8 bits). Given the nature of our data, this has resulted in a three-fold reduction in space utilization. This condensation of information does not result in any loss of information but rather excludes unused space not usually considered in the storage of data. The packed records contained all the information in the 51 byte records but stored it in only 16 bytes of mass storage per record, a 67 percent space savings, and a 80 percent space savings over the 85 bytes of retained information (see Table 1). Once packed, the total data base of 696,460 records occupied 11,143,360 bytes of mass storage, a savings of more than 90 percent from the original 120 MB, a savings of nearly 70 percent over the reduced 36 MB file.

Each year's data were stored in a separate disk file. This allows for easy handling and manipulation of the data and facilitates update of the data base as additional years of data become available. After the packed data files were created, we downloaded the data, year by year, from the main frame to the microcomputer.

Organizing the Data Base

For our project, there were only two indices by which we wanted to access each data record: date of birth and town of mother's residence. At times, we would want to summarize the entire data base; at other times, we would want all births only from one or a few towns over all years, all births for all towns for one year, or some combination thereof. Processing for each of these choices had to be easy and rapid for repeated review and reassessment. We therefore organized the data by date of birth and by town, and used pointer files for rapid access.

Within each year, the data were sorted by town of mother's residence, and within each town the data were further sorted by date of birth. A pointer file then was constructed listing, for each town, the record number in each file where that town's data began. Access was possible by two methods—if all towns were selected, the program would read the selected year files sequentially, without use of the pointers. If data only from specific towns were required, however, selected either by town index number or name, the pointer file for each year requested would be used to identify the required record numbers for the current run, and only these records would be accessed. This direct access method of data retrieval, as opposed to sequentially reading the entire data file, greatly speeded processing if only a few towns were selected.

Interactive Access and Display

Once the data base was downloaded to our microcomputer, was sorted and pointer files were constructed, we developed a series of programs for interactive display and analysis. We do not report the details of all these programs because their design was responsive to particular objectives of our substantive study. We will give one example.

The program we describe allows the user to specify any combination of years and towns, and then pick one of four types of plots: 1) frequency distribution of birth weights; 2) frequency distribution of gestational ages; 3) box plots of birth weight versus any one of 14 covariates; and 4) sex-ratio versus parity. Each of these has additional specifications that are described below. The first two options can process the entire data set in slightly more than one minute. The latter two take approximately two minutes. Selecting a subset of the data base results in proportionately faster operation.

First, the user selects the geographic and temporal extent of data to be retrieved. To define the temporal selection, the user may pick any combination of years, including all years. To define the geographical selection, the user has three choices: pick all towns, enter a list of town names, or use an on-screen map to pick towns one-by-one. The on-screen map displays the town name, town index number, population, and the total population selected up to that point.

Next, the user selects the type of plot desired and chooses a series of plot-specific run-time options. For frequency distributions of birth weight or gestational age, the user must specify a simple frequency distribution curve, a cumulative distribution curve, or its inverse normal transform. The latter is useful in assessing departure from normality—a normal distribution yields a straight line. Very low, low, and high birth weight categories are plotted on the screen as well as the birth weight distribution. Then, the user must specify whether the individual years of data should be pooled and plotted or if each year should be plotted individually on the same set of axes. For the birth weight frequency distribution, the birth weight data are stripped out of each record accessed (i.e., only the relevant bits are interpreted), and counters in an array of 256 integer variables for each year, each variable corresponding to a specific weight, are incremented. If requested, the data are summed over years and/or transformed, and the results are displayed on the screen. A sample output is shown in Figure 1A as it comes from the computer.

For gestational age, rather than weight categories, the array categorizes gestational age by number of days and treats this array similarly to the birth weight array. About 20 percent of all the records are missing gestational age information, so additional notation is needed to indicate what proportion of the records selected are used in the displayed plot. This is shown in Figure 1B as it comes from the computer.

For box plots, the user specifies which of 14 covariates is desired as the abscissa. These covariates, as provided on the birth certificate, are: age, education or race of either the mother or the father, mother's parity, mother's marital status, month prenatal care began, number of prenatal visits, Apgar score at one minute, Apgar score at five minutes,

gestational age, or baby's sex. After selecting a covariate, data are stripped out of each record and stored in a doubly dimensioned array, one index for birth weight (1–256) and one index for the value of the covariate selected. Distributional information (quantiles) is derived from this array, and box plots[6] illustrating this information are drawn on the screen. This provides the user with information about the average tendency of birth weight as the covariate varies as well as the variability of birth weights as the covariate varies. At the bottom of each plot is written a symbol equal to the log base 2 of the number of births at that value of the covariate (1–2 = a; 3–4 = b; 5–8 = c; etc.) to indicate how many observations on which the data summary is based. A sample plot is shown in Figure 1C as it comes from the computer.

The sex-ratio plot program requires no additional information after the year and towns are specified. The data are extracted and stored as number of males and number of females at each parity between 1 and 16. Then, using each value of parity as the abscissa, the percentage of males of the total is plotted as the oridinate, and a symbol equal to the log base 2 of the number of births in that parity category is printed at the top of the screen. The number of records used in the plot and the number with unknown parity are also listed in the upper right corner of the screen. There are virtually no records missing information on the sex of baby. A sample plot is shown in Figure 1D as it comes from the computer.

Evaluation

To evaluate the performance of our programs, we ran a benchmark versus SAS-PC, version 6.03. For the SAS run, the entire data base was reformatted as 3-byte ASCII records, each representing one birth weight in ounces and no covariate information. The data set was read into SAS, and PROC FREQ was run to tabulate the frequency distribution of the data. The time required for PROC FREQ, after the data set was already read in, was 1 hour and 48 minutes. The comparable run using our program, reading all records for all years and all towns, computing the frequency distribution of these data and plotting the data took 62 seconds. This is a savings of more than 99 percent in real time. (While one could optimize SAS performance, e.g., by specifying variable length, as suggested by one reviewer, our goal is to compare our system to methods likely to be used by naive users.) A box plot run for all years and all towns, the slowest option in our program, took 1 minute and 57 seconds. A comparable run is not available as a simple procedure call in SAS.

We also conducted a similar comparison with a mainframe AS-9000 computer using the original Massachusetts Department of Health tapes and full 174-byte records. Tabulating the frequency distribution of all birth weights (since the system is not interactive we could not get screen plots) took 37 CPU seconds. At a cost of $1,800 per CPU hour this amounts to a cost of $19.91 per run. Real time processing varied considerably depending on the load on the system and priority allocated to this job. In general, main frame runs took considerably longer in real time than PC runs.

Discussion

The development of rapid and efficient computer programs to access large health outcome data bases has been explored by a variety of other investigators. Connell and co-workers[7] reviewed the use of large data bases in health care studies, suggesting that proliferation of computers has made these data readily available to researchers. They pointed out that most large data bases are unwieldy and expensive to analyze, particularly using standard statistical software. Thus, they suggested eliminating redundant or superfluous variables, combining variables, sampling records, or aggregating records into summaries. While we agree with the elimination of redundant or superfluous variables, we have overcome the other problems by designing a system expressly for analyzing a particular data set and maintaining full access to each individual record. By retaining individual records, we can avoid some of the statistical pitfalls they warn of, such as the ecological fallacy.

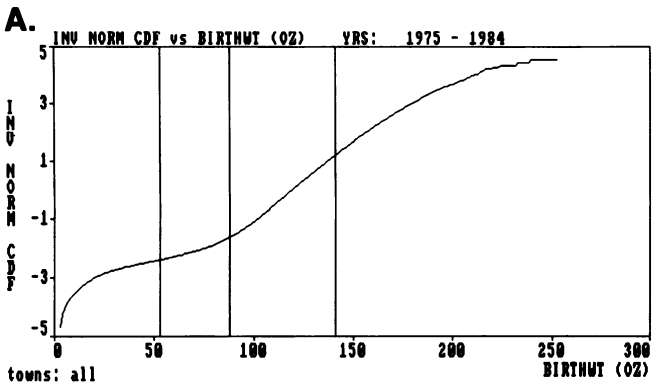
Gittelsohn[8] described an information recovery technique that he developed for analysis of US mortality data available for the National Center for Health Statistics. By limiting data fields stored, and coding the retained data in binary representations rather than character codes (similar to our data packing), his system realized a 250-fold space savings and a processing time savings of over two orders of magnitude. But, in contrast to our system, Gittelsohn's system is main frame-based and excludes data that may be pertinent for certain types of studies. He does not retain individual specific records and has no opportunity for including names or other unique identifiers.

Investigators at the University of Texas are pioneering work using WORMs and CD-ROMs on PCs for data storage.* They have created a system that allows users to specify and retrieve any set of case records from the Hispanic Health and Nutrition Examination Survey (HHANES) study. The data are output to a disk file for input to standard statistical packages. The HHANES data set is comprised of approximately 14,000 individuals with varying amounts of information for each, totaling about 40MB of data. Retrieval set ups take about 15 minutes and retrieval an additional 15–120 minutes. With this program, all analysis still is done in standard statistical packages.

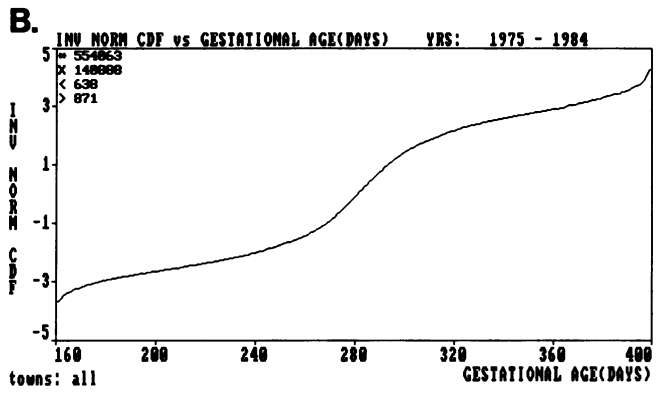
Our total package for 10 years of birth weight data takes about 12 MB of disk space and responds to inquiries in seconds. We maintain individual records for each event and have built analytic tools into our program package. We note that the design and development costs of the system we have constructed likely outweigh the costs we would have incurred with limited data analysis on a main frame system. However, for routine analysis and summarization, as is frequently needed, the system we describe is easier and faster. It has the added capability of instantaneous data review, interactive processing, and exploratory data analysis (EDA) methodology not available on most main frame computers for a data base of this size.

We believe that our system can help epidemiologists, public health officials, and vital statistics registrars meet the demands of their professions. It will enable researchers easier access to data already collected and validated by government health officials. It will enable public health officials to carry out up-to-date surveillance and to respond to public inquiries in a timely fashion, generating data summaries and reports in a matter of hours or days instead of weeks or months. And, it will ease the data collection, review, and summarization process handled routinely by government registries. It could also assist officials in identifying regions showing anomalous incidence rates or clusters. Systems like the one we describe

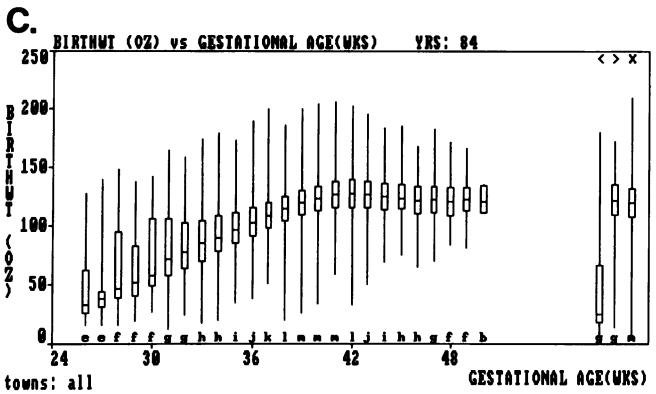
*Personal communication, Lindsay Reed and Steven Bates, University of Texas School of Public Health, Houston, February 10, 1989.



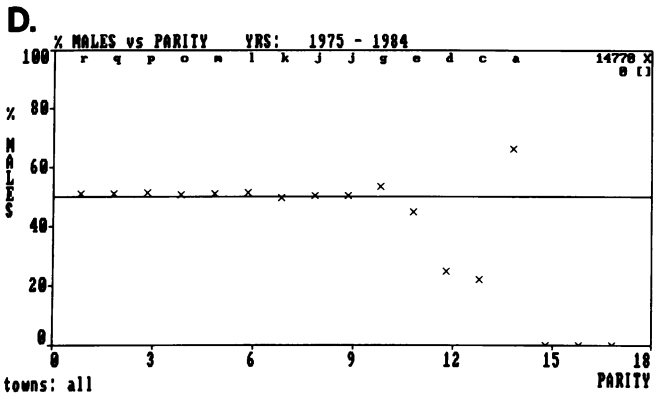
A. Birth Weight Frequency Distribution—This figure shows the inverse normal transform of the cumulative distribution function of birth weights. The three vertical lines represent weights less than 53 ounces ("very low birth weight"), less than 88 ounces ("low birth weight") and more than 156 ounces ("high birth weight"). The birth weight in ounces is on the abscissa, and the inverse normal transform of the cumulative frequency distribution is on the ordinate.



B. Gestational Age Frequency Distribution—This figure shows the inverse normal transform of the cumulative distribution function of gestational age. The gestational age in days is on the abscissa, and the inverse normal transform of the cumulative frequency distribution is on the ordinate. The value next to the asterisk is the number of records used in the plot. The value next to the X is the number of records missing gestational age. The value next to < is the number of records with gestational age less than 160 days. The value next to > is the number of records with gestational age greater than 400 days.



C. Box Plots—This figure shows the distribution of birth weight at each value of the covariate selected, in this case, mother's age. At each value of the covariate, the diagram shows the range, quartiles, and median of the distribution. The abscissa of the plot is the value of the selected covariate, and the ordinate is the birth weight. The letters below each diagram are letters corresponding to the base 2 logarithms of the number of records at that value of the covariate.



D. Sex Ratio versus Parity—This figure displays the sex ratio of births at each parity for the selected data. The horizontal line is at 50% males. The value next to the X is the number of records missing parity information. The letters above each value of the covariate are letters corresponding to the base 2 logarithms of the number of records at that value of the covariate. The abscissa is parity and the ordinate is the percentage of males of the total births.

FIGURE 1—Screen Dumps—This figure shows actual screen images produced using the packed data base and our software. All displays except 1C use the entire data base (1975–1984, all towns). Figure 1C uses all towns for 1984 only.

could be implemented for all types of data bases (e.g., death certificate files, cancer registries, birth defect registries, health survey data) on inexpensive and readily available hardware. Officials can have desktop computer data bases at their finger tips with which to study epidemiological phenomena and researchers can begin to surmount some of the limitations of costly data collection by gaining access to vast data stores of health outcome registries.

APPENDIX
Packing the Data

The most important design feature of our data system in terms of performance is efficient storage of the data. Minimizing the space required by the data base limits the resources needed to hold the data. Additionally, minimizing the size of each record maximizes the speed with which data can be accessed, data input being the rate limiting step in most data base systems. While most computer programs access data on a byte-by-byte basis, we have packed data on a bit-by-bit scale (each byte has 8 bits). Each data record as provided has 174 bytes long. We culled 51 relevant bytes of information and packed these 51 bytes into 16 bytes using a bit-packing system.

Each byte consists of 8 bits, each of which can have the value 1 or 0. Thus there are 2^8 possible combinations of 1s and 0s that each byte can assume, 256 different values. For traditional data storage, many of the 256 bit combinations are unused. For example, for baby's sex only 3 out of the 256 possible values are used: 1 for male, 2 for female and 0 for missing value. By allowing only the necessary number of bits of a single byte to present these 3 possible values we can store all the information in 2 bits. This leaves 6 bits for storage of another variable. Thus, we can pack more than one variable into a single byte with no loss of information (e.g., bits 1–2 can represent a variable having up to four difference categories, such as baby's sex and bits 3–8 can represent a variable having 64 different categories, such as mother's age). We can pack both baby's sex and mother's age in 1 byte with no loss of information. This yields a three-fold space savings. Birth weight, though theoretically a continuous variable, is recorded in pounds and ounces and stored on the original tape in 4 bytes. However, all weights between 0 lbs, 1 oz and 16 lbs, 0 oz represent only 256 different values. We have coded this variable as 1 byte, a four-fold saving of space. (More frequently, birth weight is recorded in grams, 454 grams equaling 1 lb. All weight from 0 lbs to 16 lbs, or 0 to 7264 grams, in 50 gram increments, or approximately 1.8 oz also could be stored in a single byte.) This type of space reduction was applied to each field retained and resulted in records requiring 128 bits each, or 16 bytes of mass storage rather than the 51 used on the original tape (see Appendix Table 1). The total data base of 696,460 records occupies 11,143,360 bytes of mass storage, a savings of more than 90 percent from the original 120 MB, a savings of nearly 70 percent over the reduced 36 MB file.

We chose to pack the data before downloading to our microcomputer. The packing program was written on a mainframe system (Rutgers University's AS 9000) so that only the packed data would be downloaded to the microcomputer. This main frame computer has a FORTRAN language compiler, but neither Pascal nor C language compilers. Using FORTRAN necessitated extra care in packing because bits are not easily accessible. The packing program was written so that variables would end on even-word boundaries for each of 4, 32-bit FORTRAN integers. To pack the data, the value of each variable was coded, added to a 32-bit integer, and then the integer was shifted left by the number of bits allocated to that field. The value of the next variable for that record was coded, added to the integer, and then the integer was shifted left for the next field. This was continued until this integer was fully packed. Thus, each variable was inserted at the right of the integer and pushed left by the insertion of the next variable until all the designated fields were entered (see Figure 2). All the data were packed into 16 byte (128 bit) records which were stored in yearly files.

To ensure the integrity of the packed data transmission, we downloaded the data using error-checking software (KERMIT v. 2.29, February 19, 1987). We transmitted over a 2400-baud modem and each year's data took one to three hours to send, depending on the computing load from other users on the main frame. (More rapid transfer could have been accomplished by using a hard-wired terminal with 9600-baud or faster transfer rates.) We also ran a few test programs to verify that the microcomputer data files were identical to those on the main frame. Once downloaded, these files were sorted and indexed as described in the text.

If it had been necessary to keep lengthy identifying information, such as a person's name and address, the size and access time of the data base would increase. However, rather than complicating the handling of the health outcome data, identifier data could be kept in a separate, cross-linked file and accessed only when such information is required. (We have constructed such a system for New Jersey death certificates). For most statistical inquiries, these identifiers would not be used. For confidentiality protection, only those with proper clearance could be given copies of (or access to) the personal identifier files while all interested researchers could be given copies of the data base.

APPENDIX TABLE 1—Data Packing Format

Bytes	Status	Bits	Item
6	I		Certificate Number
3	D		Town of Birth Occurrence
22	I		Name of Baby
1	K	2	Sex of Baby
1	D		Plurality
1	D		Birth Order
6	K	13	Birth Date
1	K	2	Mother's Marital Status
2	D		State of Father's Birth
2	K	6	Father's Age
12	I		Mother's Maiden Name
2	D		State of Mother's Birth
2	K	6	Mother's Age
3	K	9	Town of Mother's Residence
4	D		Boston Census Tract
1	D		Birth Attendant
4	D		Facility of Birth Occurrence
10	I		Baby's Medical Record Number
1	K	4	Father's Race
2	K	3	Father's Education
1	K	4	Mother's Race
2	K	3	Mother's Education
1	R		Previous Live Births Now Living
1	R		Previous Live Births Now Dead
1	C		Previous Terminations before 20 Weeks
1	K	4	Previous Terminations after 20 Weeks
4	K	9	Date of Last Live Birth
6	K	13	Date of Last Menses
1	K	4	Number of Prenatal Care Began
2	K	6	Number of Prenatal Care Visits
1	D		Complications or Concurrent Illness
1	D		Birth Injuries to Baby (to 1977 only)
4	K	8	Birth Weight (lbs. oz.)
4	K	9	Congenital Malformation Code
1	D		Type of Patient Care Received
1	D		Baby Alive at Time of Report
4	K	9	Date of Last Termination
3	D		City/County of Mother's Out-of-State Residence
1	K	4	APGAR Score at 1 Minute
1	K	4	APGAR Score at 5 Minutes
2	R		Gravidity
11	R		Blank Field
2	R		Health Service Area of Mother's Residence
1	R		Blank Field
2	R		County of Mother's Residence
2	R		Health Service Area of Birth Occurrence
1	R		Blank Field
2	R		County of Birth Occurrence
4	R		Birth Weight (grams)
1	R		Generated Race of Baby
7	I		Welfare Facility Vendor Number (1975, 1976 only)
1	R		Adequacy of Prenatal Care Index
1	D		Neonatal Match Indicator
6	I		Death Certificate Number if Neonatal Match
2	K	6	Parity
2	R		Gestation Period
174		**128**	**TOTAL**

DATA PACKING SUMMARY

Bytes	% of Total	Type
63	36	(I) Identifiers
26	15	(D) Dropped
33	19	(R) Redundant (or blank)
1	1	(C) Combined
51	29	(K) Kept
174	**100**	**TOTAL**

Note that 128 bits = 16 bytes = 9.2% of 174 bytes.

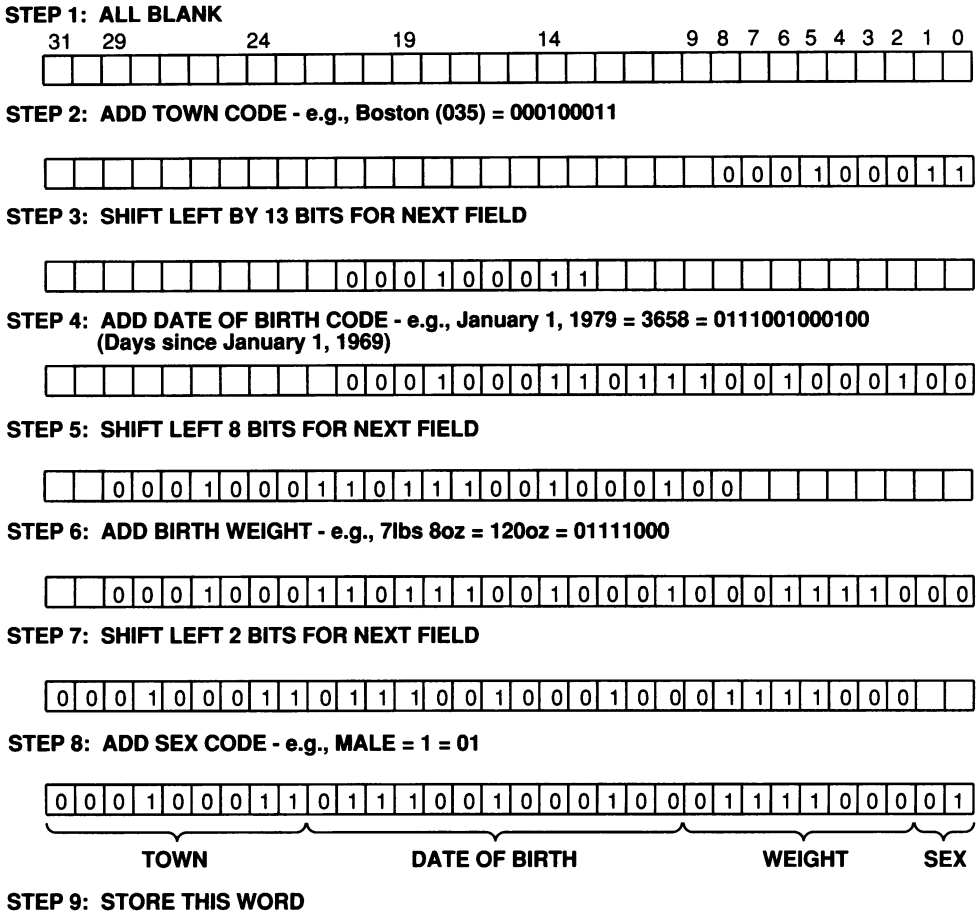


FIGURE 2—Packing the First Word—This figure shows the process by which data are coded and combined into integer words for storage in the packed data base. The example shows this process for the first word only. A similar process is used for the additional three words.

ACKNOWLEDGMENTS

We thank Elaine Trudeau, Registrar of Vital Records, Massachusetts Department of Public Health, for providing the data tapes and helping us to understand the data collection and evaluation procedures used in their management. Two reviewers and two editors provided useful comments and suggestions. Judy Bloem assisted in the manuscript preparation. This research was supported in part by funds from the Comprehensive Environmental Response, Compensation, and Liability Act trust fund through Cooperative Agreement no. U50/CCU101045-02-1 from the Centers for Disease Control and the Agency for Toxic Substances and Disease Registry, US Public Health Service.

REFERENCES

1. Dubrow R, Sestito JP, Lalich NR, Burnett CA, Saig JA: Death certificate-based occupational mortality surveillance in the United States. Am J Ind Med 1987; 11:329–342
2. Milham S: Occupational mortality in Washington State 1950–1979. Cincinnati: National Institute for Occupational Safety and Health, 1983. DHHS[NIOSH] pub. no. 76-175.
3. Archer VE: Association of nuclear fallout with leukemia in the United States. Arch Environ Health 1987; 42:263–271.
4. Thomas TL, Fontham ETH, Norman SA, Stemhagen A, Hoover RN: Occupational risk factors for brain tumors. Scan J Work Environ Health 1986; 12:121–127.
5. Milne KL, Sandler DP, Everson RB, Brown SM: Lung cancer and occupation in Alameda County: A death certificate case-control study. Am J Ind Med 1983; 4:565–575.
6. Tukey JW: Exploratory Data Analysis. Reading, MA: Addison-Wesley Publishing, 1977.
7. Connell FA, Diehr P, Hart LG: The use of large data bases in health care studies. Annu Rev Public Health 1987; 8:51–74.
8. Gittelsohn AM: On the distribution of underlying causes of death. Am J Public Health 1982; 72:133–140.