# FOOD SAFETY DATA BASES

D. Jesse Wagstaff

People express ideas about which foods are good or bad for them virtually throughout their lives. National concern in the United States was formalized in 1906 by passage of the sister acts regulating safety of 1) meats and 2) other foods and drugs.

The Food and Drugs Act of 1906 has been amended a number of times and has become the Food, Drug, and Cosmetic Act. The Food and Drug Administration (FDA) implements its provision. Prevention of food contamination by organisms or chemicals capable of causing disease was thought of to benefit public health. Although it is widely believed that our society actually has received some of the expected health benefits, it is difficult to provide strong scientific evidence that food safety regulatory activities in fact have led to improved public health. Acquisition of such evidence required 3 types of information: 1) the count and characterization of the population, 2) exposure, which in this case is food analyses and food intake, and 3) health outcomes.

If substances in food affect health, then disease rates in population groups in the U.S. should be related to food intake. In a classic application of such methods, Dr. John Snow helped control a 19th century cholera epidemic in London by comparing the geographic distributions of cholera cases with drinking water source.

The U. S. population is large, is exposed to a large number of potentially harmful agents, and is afflicted by many types of disease. Information is collected from the public for different purposes which have expanded beyond the historic objectives of governing, taxing, and raising armies as exemplified by the Domesday Book. As data collection has expanded and purposes for those data have diversified, so also have file formats become dissimilar. The population, exposure, and health outcome data needed to evaluate human food safety are collected by a variety of agencies for a variety of purposes. Those data are then tabulated in different ways and released in various massive dissimilar published tables and computer file structures. In the U.S. government alone there are an estimated 250,000 data files. No master catalog or inventory listing these files or describing their contents has ever been compiled.

The logistical problems of using large, complex national files have generally restricted evaluations of human food safety data to small or local studies or to use of summarizations of national data. Although automated data processing methods and equipment such as computers were originally developed to process population data, scientific evaluators, outside of a few organizations, have not had adequate access to automatic data processing. Also computers until recently lacked the speed and storage capacity to support any large number of studies using national files. With few exceptions the problems of massive files, file dissimilarity, and inadequate computer access have prevented optimum use of national files. The earliest year for which machine readable national files are available for all of the three areas of population, food intake, and health outcome is 1965.

The magnitude of the problem of evaluating data on human food safety at the national level dictated that the approach had to be extensive but systematic. To this end the FDA has launched an Epidemiology Information System (EIS) to enable us to acquire, store, analyze, and display the needed information. Such an effort required the cooperation of many individuals, agencies, and organizations.

Emphasis in the first phase of the development of the EIS was on bibliographic indexing. Through an interagency agreement between the FDA and the National Library of Medicine (NLM) an index to the more important reprints, books, and other documents of the Epidemiology Unit is being created by the Toxicology Information Response Center at Oak Ridge National Laboratory. There are two main issues in creating collections of documents and indexing them. The first is the number of documents to collect and the criteria of selecting those documents. The second issue is selection of a nomenclature or coding scheme for indexing the selected documents. Resolutions of these problems were not obvious at the outset but have evolved as our experience has grown. Financial and space limitations dictated that only a moderately small number of documents could be housed and indexed. Criteria for selection of documents to index and include in the collection has been a matter as much of determining things outside our scope as determining things within our scope. Animal toxicity tests generally are outside while human epidemiology is in. Drugs are out but foods and cosmetics are in. Infectious agents are out but environmental chemicals which are food contaminants are in. The subject areas selected include epidemiologic studies and reviews on human safety of foods and cosmetics. Memoranda, letters, and preliminary reports are generally not indexed. Documents are selected for indexing by system users. Meetings are held periodically to review the selection process and to discuss operation of the system.

With the possible exception of chemical compounds we have not been able to find universally accepted standard nomenclatures or codes. We chose to work with the biomedical nomenclature used by the NLM for MEDLINE. It is called Medical Subject Headings (MeSH). Using this controlled vocabulary all documents dealing with a particular subject are indexed under a single keyterm regardless of the different synonyms that may have been used in the document itself. For example, articles regarding vitamin C are indexed under ascorbic acid. System documentation aids the user in determining correct keyterms. Most documents are indexed under several keyterms. The searcher can retrieve references using almost any combination of authors and keyterms connected with Boolean connectors such as "and" and "or." For example, a person could readily find all documents in the system indexed under both ascorbic acid and colonic neoplasms.

Each document selected for indexing in the EIS is microfiched and is assigned a unique identification number which designates its location in the files. Because of extensive crossindexing there is no need to file multiple copies of a document dealing with several subjects. In a manner similar to MEDLINE the user can select references using authors and keyterms. The computer system is RECON, operated by the Department of Energy. Subscriptions to RECON are available to the public.

Another bibliographic index on food safety available from the Bureau of Foods but not in the EIS is the Natural Toxicants Database. This computer-indexed collection of documents was originally limited to mycotoxins but has since been expanded to cover other areas. The index is accessible by subscription to the Parklawn Computer Center, FDA, Rockville, Md.

The need to obtain information on research in progress in epidemiology was recognized by the Epidemiology Work Group of the Interagency Regulatory Liaison Group of which FDA is a member. Through a contract with the Smithsonian Science Information Exchange, an Epidemiology Research Projects Directory has been published and may be purchased from the National Technical Information Service, Springfield, VA. This hard copy of the annual Epidemiology Research Projects Directory supplements the creation by the Epidemiology Work Group of the RPROJ subfile of the NLM's TOXLINE, containing monthly updates of epidemiology and toxicology research. TOXLINE represents a database which has a totally uncontrolled vocabulary. Instead, the computer is directed to match words in the request with character strings in the title and abstract. The user has to be concerned about synonyms and even misspelled words.

The remainder of this presentation will emphasize files of numeric data, particularly those files related to food safety. These files come from different sources. Population data originates in the Bureau of Census, national food intake surveys are conducted by such organizations as the U.S. Department of Agriculture (USDA) and mortality data are compiled by the National Center for Health Statistics. Even a simple comparison of disease-specific death rates for two population groups of differing dietary intakes required bringing together or integrating data from a number of sources. Computer systems called Integrated Database Systems (IDBs) have been developed to enable researchers and reviewers to efficiently and easily perform operations of this type.

Because of the long development time and the high costs required, most agencies including FDA share use of existing IDBs rather than develop their own. At the national level there are four major IDBs in the areas of health and environment. These are SEEDIS, GEOECOLOGY, DIDS and UPGRADE.

SEEDIS, the Socio-Economic Environmental Demographic Information System, developed by Lawrence Berkeley Laboratory, is used for applications in human health, air pollution, employment, unemployment, energy planning, environmental impact analysis, and land use analysis. SEEDIS supports 50 different geographic levels from nations down to minor civil divisions. SEEDIS has the ability to automatically aggregate or disaggregate data from one geographic level to another. Presently, 12,000 different data elements are archived in 31 files.

Geoecology Data Base is an integrated data base of diverse environmental resource information developed by the Oak Ridge National Laboratory. Data on terrain and soils, water resources, forestry, vegetation, agriculture, land use, wildlife, air quality, climate, natural areas, endangered species, and human population are stored in Systeme Internationale units at the county level for the coterminous United States with some data available at the subcounty level in a few cases.

DIDS, the Decision Information Display System, is an interactive color mapping system developed in the Executive Office of the President to display statistical data. DIDS provides very rapid retrieval, display, and manipulation of thematic map images using avant garde but expensive color television types of computer techniques.

UPGRADE is an acronym for the User Prompted Graphics Data Evaluation system developed by the Council on Environmental Quality (CEQ) to enable noncomputer-oriented scientists to analyze environmental data by graphing, statistical analyses, data manipulations, and mapping. UPGRADE was first used for the 1976 Annual Environmental Quality report. Users can key in their own data, or can use data on the system. UPGRADE presently is the only one of the major IDBs which is available nationally and internationally through computer time sharing. Subscriptions at competitive commercial or government rates can be obtained from Sigma Data Corporation, Washington, DC. Any standard printing terminal can provide access for retrieval and analysis but graphic displays require a Tektronix type of terminal or appropriate plotter. Devices to adapt microcomputers to emulate Tektronix have been advertised but to my knowledge have not been tested with UPGRADE.

A high level of cooperation and exchange of information exists among producers and users of these IDBs, as was seen at the First Integrated County Level Data User's Workshop held last fall in Reston, VA. A second workship is being planned for this October, again to be held in the Washington, DC area. People interested in integrating data from different subject areas are invited to attend.

The IDB used by the EIS is UPGRADE. it was originally developed by the CEQ, but has since also been supported by a number of cooperating agencies including the Geologic Survey, Environmental Protection Agency, USDA, National Cancer Institute, and as a cooperative venture of the EIS of FDA. Prior to involvement of FDA, emphasis in UPGRADE was on environmental datasets; population and health datasets were available only in summarized form.

After the log-on, the entire UPGRADE session is conducted in English. The user is asked yes or no questions or presented with menus for selection of tasks to be performed. Detailed explanations can be requested at any time in the session by entering "help." Steps can be repeated by entering "back." "Exit" terminates a session. In the first section of an UPGRADE session called the setup section, the user is prompted for terminal type and line speed. Also the user may select full or abbreviated prompting by selecting verbose or terse mode. In the next section called the data selection section, the user selects data from one or more datasets. One field or a number of fields may be selected from each dataset. The selected fields from the various datasets are integrated into a user dataset which may be used immediately or saved for use in future sessions. UPGRADE does not actually search millions of records at the time of the request; rather it searches indexes to enable it to quickly retrieve data which previously had been selected, pruned, and sorted. In the FDA UPGRADE project, which is one of the functions of the EIS, the goal is to prepare food safety data sets for English language response to about 90% of the most commonly expected types of requests. Preparation for unusual requests which may

never be made would not be an efficient use of resources. In the last section, called the graphic and statistical analysis section, the user directs the processing and display of the data selected in the data selection section. The UPGRADE user is presented several choices for display of data either before or after analyses. The user can print statistical results in rows and columns, construct a table of the data, plot data points, connect the points, draw a regression line through the points, or draw a bar graph. With the advent of SASGRAPHICS, it should be possible to use enhancements of all of these display procedures and to have new ones.

Of course it is possible to make nonsensical comparisons or perform illogical or incorrect analyses. Responsibility for quality is in three parts. First, the organization which generated a dataset is responsible for faithful execution of the requests entered at the keyboard to retrieve, analyze, or display data. And third, responsibility for interpretation of results rests with the user. USDA cannot expect to pass responsibility for accuracy of its food survey data on to an IDB or user. Nor should English-language capability in an IDB be expected to make every user an expert statistician. Some users will only use simple descriptive statistical procedures or displays while others more expert in statistics may wish to use the fullest capabilities of available statistical packages.

One of the most interesting display techniques in UPGRADE is mapping. Any data that can be referenced to geographic areas can be mapped. The user selects the data, the intervals into which the range of data are to be divided, the shading and color patterns considering the capabilities of the terminal or plotter, the geographic area, and the heading and labeling. Presently UPGRADE can map by county and by state. In addition some natural bounded areas, such as river drainage areas can be mapped. Maps can be created for point data but it is difficult to relate point data to area data.

In the FDA UPGRADE project detailed health datasets are being added to UPGRADE in the three areas of population, exposure, and health outcomes. Census data or estimates for each year from 1965 to 1976 are being added for each U. S. county by age, race, and sex. Some religious groups such as Mormons and Jews follow unique dietary practices and thus are of value in food safety studies. The Bureau of Census used to collect data on church membership but now that information must be sought from private sources. Several files of food intake and related exposure data have been acquired by UPGRADE including food intake surveys conducted by USDA in 1965 and 1977-78; food intake interview data from the first Health and Nutrition Examination Survey (HANES I) a fish intake survey obtained from the National Marine Fisheries Service; food analyses done by FDA for pesticides, metals, and other contaminants; occupational data in the County Business Patterns files; and the 1974 Census of Agriculture. Detailed health outcome files recently made available to UPGRADE are U. S. detailed mortality in the years from 1965 to 1976, except for 1972 for which a complete file is not available, and medical interviews and physical examination results from HANES I.

Some of the things which have been done during the enhancement of UPGRADE with these new files are as follows. Population census or estimates have been printed for different geographic areas by age, race, and sex. Fish

intake in the Great Lakes area has been tabulated by species to aid in the human food safety evaluation of dioxin contamination of fish in the Great Lakes. Beef intake statistics from the USDA 1977-78 food intake survey have been analyzed to assist in evaluating safety of drugs and feed additives given to cattle.

Effort in the EIS has been concentrated on system development but some general comments about food intake and its relationship to health can be made. Types and amounts of foods consumed change during normal lifespans. For example, milk is the only food of the newborn and is the major beverage during childhood. But as adulthood is reached other drinks including coffee and tea are substituted for milk, the substitution being greater for women than for men. In old age food intake patterns tend to return to those of childhood. In this country we are fortunate to have an abundant food supply. In fact we eat more than is necessary to maintain normal function. On average we accumulate adipose tissue as we age.

However, we are living longer than our ancestors. Also the causes of death have changed since the last century. Then, most people died of infectious disease and often at an early age. Now most fatal disease is chronic and strikes the elderly age group. Half the present day deaths are due to cardiovascular diseases. The next largest group of fatal diseases are neoplasms. Trauma is the third category. Respiratory diseases, although not generally highly publicized, still account for about 4% of deaths, while all other diseases, including those of the digestive system, infancy and congenital anomalies and miscellaneous causes, account for less than 20%.

In a typical recent year, 1975, only 0.22% of deaths were due to nutritional problems other than alcohol. In many of these cases the people literally ate themselves to death as evidenced by a listing of obesity on the death certificate as the primary cause of death. Others of these deaths were probably due more to malabsorption or other physiologic impairment than to dietary deficiency. The total number of nutritional deaths is less than the number of deaths due to septicemia or pneumococcal pneumonia or chronic nephritis or cardiac congenital anomalies and is not far different from the number of deaths from paralysis agitans or hyaline membrane disease or amyotrophic lateral sclerosis. Six times more people drink themselves to death, nine times more people kill themselves with guns or other means, seven times more people are murdered, and fourteen times more people are killed in traffic accidents than die from nutritional problems.

Why then the great continuing national debate about diet and food safety? Probably the answer is that scientists and laymen alike feel that health and food are more closely related than the above statistics indicate. Perhaps diet is involved in the etiology of the great killers of our time, cardiovascular disease and cancer. Decisions about food safety, whether made by individuals or by regulatory agencies, should be based on the best scientific evidence available at the time. The strength of the evidence underlying the decision may be overwhelmingly strong, as in the case of measures to avoid botulinum toxin or the evidence may be controversial, as in the case of nitrites. Improvement of scientific evidence to support better food safety decisions can only be obtained through considerable effort.

In conclusion, there are large amounts of information available on human food safety but difficulties exist in organizing and accessing it. FDA is addressing the issue through development of the EIS to index research in progress and published reports and to integrate data on population, exposure, and health outcomes. Financial support is becoming critical for producers and users of IDBs. Development of these systems may slow. But as critical as the issue of financial support is, an even more important issue is demonstrating that creation of databanks on foods and related subjects is of sufficient benefit to society to justify continued use of public funds. This will require increased cooperation of individuals and organizations. The EIS expresses eagerness to cooperate in efficient use of food safety information.