LINKAGE OF FOOD INTAKE AND HUMAN HEALTH DATA TO DEMONSTRATE PUBLIC HEALTH BENEFITS OF NUTRIENT DATA.

by
D. J. Wagstaff
Epidemiology and Clinical Toxicology Unit
Bureau of Foods
Food and Drug Administration

ABSTRACT

The ultimate purpose of information in a nutrient databank is to improve human health through better nutrition and safer food. Determining whether this purpose is achieved is a very large complex task which must be approached in a systematic fashion. Pertinent data files such as mortality files must be identified and practical means for accessing and analyzing records in them should be available. Hindrances to this access and analysis include shortage of funds, large volumes of data, and nonuniformity in data recording and coding. Some approaches to solve these problems are cooperation of researchers, division of labor and bartering to replace actual transfer of funds, use and integration of small machines such as word processors and microcomputers, shared use of large computers and analysis support staffs, sorting and indexing of large files, and creation of tables of equivalent codes.

DEFINITION OF A NUTRIENT DATABANK

This is my second year attending these conferences. I have enjoyed them and found them worthwhile but admit to a little uncertainty as to what a nutrient databank is. I will use the term in a generic sense to mean sources of data on human food intake and food components rather than any particular data file or computer center.

An ordinary bank provides a medium of exchange for all sorts of goods and services. That medium, which we call money, can be used to obtain the necessities of life or other things we desire. There are many different kinds of money but they all have equivalent values.

An ordinary bank safeguards money in vaults and safety deposit boxes and of equal importance the bank maintains an accurate record of the money it holds at all times.

An ordinary bank defines to the last penny its obligations to other organizations and individuals in our society. These obligations are carefully detailed even down to the exact procedures for fulfilling each of those obligations.

An ordinary bank has buildings, offices, vaults, safety deposit boxes, tellers, and officers to enable it to achieve its purposes. It works closely with its investors, depositors, and borrowers.

A nutrient databank is obviously not an ordinary bank. The goods and services this bank deals with are information about nutrients and capabilities to make that information useful. The media of exchange are the symbols of language and science. Some of these symbols have universal meanings but other symbols or codes have meaning to only a select few in our own organizations. Other people can't decipher these codes. Much progress has been made but we still can not easily exchange detailed information on nutrient intake and its effects on humans. To a certain degree, it seems we are perpetuating the confusion of the Tower of Babel. Nutrient databanks have not yet achieved standards to provide maximum ease and efficiency in data transfer and exchange.

Nutrient databanks generally do not have vaults or safety deposit boxes to safeguard their information but they are moving towards the standards of other organizations which have very elaborate vaults to hold information which they consider valuable. There is not a master nutrient databank which has an accurate inventory of its data holdings and a daily balancing of its account books.

The obligations of nutrient databanks to other organizations and individuals are not clearly defined. Perhaps some would argue that a nutrient databank has no obligations but I think this would not be appreciated by the taxpayers, hospital patients, stockholders, or food purchasers who are paying our way.

Nutrient databanks generally have no buildings, offices, or tellers to do their business but usually have some officers.

A nutrient databank or at least a master nutrient databank seems to be more of a concept or ideal than a physical bank. Its structure, activities, and objectives are evolving.

PURPOSE OF A NUTRIENT DATABANK

The main purpose of information in a nutrient databank is to improve and maintain human health. This assumption that health is the purpose of the bank is based on 2 things. First, the definition of nutrient itself is that it is something taken into the body to sustain life and maintain health. If the purpose of foods and nutrients is to maintain health then it logically follows that the purpose of information about nutrients is also to maintain health. Second, most of the attendees at this conference seem to be from food or health organizations and presumably believe that their ability to improve food or health will be enhanced by attendance here. There may be economic or other types of benefits but here we will assume that improved health is the major benefit of a nutrient databank.

JUSTIFICATION OF A NUTRIENT DATABANK

The bank should be supported by society only if society can be convinced that the purpose is worthwhile and that the bank can achieve that purpose. There is a need to determine whether the benefits of nutrient databanks justify their costs. It is not enough to just assume that we are on the side of the angels and that everyone accepts that as fact. There are great men and women of science as well as ordinary people who have doubts about the wholesomeness and safety of our food. Ideal body weight is being challenged. The ideal diet is also being challenged. How many of you this morning used saccharin in total disregard of the warning about cancer in animals?

So-called health foods and natural foods are said to promote health while those bought in ordinary grocery stores, supermarkets, or cafes are said to promote disease. In most cases we do not know whether our efforts are beneficial. We believe that probably we are helping but it is not outside the realm of possibilty that we are injuring and hindering. We should be prepared to consider the unthinkable that some of the things we are doing are causing health to deteriorate or even to cause the loss of life.

We shall here concentrate on defining what pieces of scientific evidence are needed and to look at some of the steps which must be undertaken to acquire that evidence. Everything can not be done at once; a databank has to be built and used in a logical systematic fashion. It will take time to get to where we want to be.

I do not advocate cynicism or that we become paralyzed in decision making because we don't have ultimate proof of everything. Of course we must make decisions as they are needed based on the best information available at the time and then revise that decision if necessary when better information becomes available.

The point to this philosophical tangent is to emphasize the importance of being farsighted in building nutrient databanks, to make sure that they serve the needs of society and that society knows that it is being well served.

MEANS OF DETERMINING WHETHER PURPOSE IS BEING ACHIEVED

Health is the length and quality of life. We would all like to live full lives to the very end and then die quickly, painlessly, and gracefully. Health is freedom from pain and disability, ability to perform bodily functions and fulfill physical and mental potential. Health is action not structure. Health is physiology not morphology. It is not anatomy, it is not biochemistry, it is not clinical chemistry, and it is not nutrition. Knowledge of these things may aid us in predicting health but they are not health.

I don't care if my blood iron level is zero and my spleen is green with pink polka dots if my health is not affected. Nutrient levels in the diet and in the body are indicators and predictors of health but they are not health. We do not know that their prediction of health is correct in all cases. The predictions of Jimmy the Greek before the Superbowl are not the same thing as the score at the end of the game. Opinion polls predicted victory for Thomas Dewey but Harry Truman won the election. A nutrient or food under a given set of conditions and at given levels maintains or improves health only to the extent that health itself is maintained or improved and not to the extent that an intermediate or predictor such as a biochemical concentration in the body is altered. The ability to run and not tire is important but the blood level of iron is an interesting scientific concept knowledge of which hopefully will help maintain ones capacity to run. I have heard the term nutritional health but I do not know what it means. If equine health is the health of horses, bovine health is the health of cattle, and human health is the health of humans, then is nutritional health the health of nutritionals? Nutritional health is a poor term which tends to confuse ends and means. The end is a healthy person and an entire population of healthy people rather than a person or population which particular chemical patterns in their foods or bodies.

If we accept that health is the desired end and that we want the best evidence possible that nutrient intake causes good health, then as in getting evidence of other things by science, we would need to have study subjects and the ability to measure the responces in those exposed compared to those not exposed. We are of course not permitted to risk the health of people in experiments. So to obtain human data we must rely on natural experiments, those exposures and outcomes which normally occur. Ideally we could do prospective studies on all important questions and follow groups of people in the future who are exposed to food components at different levels. This is a good approach except that it takes more resources and time than will ever be available except for a very few cases such as cigarette smoking. Another approach in which the basic unit is the individual human is the case-control study which is done retrospectively and thus saves much time. The records of people who have a certain outcome, e.g., bladder cancer, are compared to records for people alike in all respects except they do not have bladder cancer to see if the exposure to artificial sweeteners is significantly different. However, even this approach is very expensive and can only be done in a few

cases. A recent study of the relationship of bladder cancer and artificial sweeteners cost more than a million dollars. The alternative is to do descriptive epidemiologic studies and analyze the results carefully to generate the most likely theories for more definitive tests. This often involves study of the records of groups or populations of people rather than individuals. But still health records must be linked with food intake records.

Record Linkage

Record linkage is simply bringing together different records pertinent to the same subject. Today in our laboratories the total record for an experimental rat seldom is written on a single sheet of paper. Body weight and feed consumption are often written many times on separate sheets. Samples of body fluids, tissues, and excreta are collected and may be examined by different people in different organizations over a long period of time. Yet all the records for that rat can be linked because the separate records contain an animal number or other unique identification. The records for the group from which that rat came could be linked in the same way.

Records pertaining to human beings are identified by a legion of code numbers. It is a rare person who can remember or even who has a duplicate copy of the identification numbers for all of the documents in his or her wallet. It is a challenge to link all the records of one type for a person, say all the records of social security payments. But is is presently impossible to link all of the different types of records for a person. Indeed our society seems to be saying they don't want any organization to have the ability to link all their records. This fear may offer some protection against potential blackmailers or snoopy loose-lipped clerks. But it handicaps our ability to relate cause of disease to the disease itself, be that cause in food or elsewhere, be that cause an excess or a deficiency. If the cause and effect can not be linked, the disease will go on unchecked.

However there are many food consumption and health records which are in the public domain and which can be linked. Possible causes of disease can be looked for in food consumption pattern data. In most files, individuals are not identified in such a way as to enable linking their food intake and health records but groups of people are identified. In addition to such things as age, race, and sex, the major group identifier is geographic area such as region of the country. But before matches of food intake and health records can be made, the coding schemes must be equivalent. The steps to getting the health records properly sorted and coded for a correct match with food intake are not trivial challenges.

First the available human health data must be identified. The major health records for human populations are mortality and morbidity, i.e., tallies of deaths and illnesses. As a starting point we have chosen to work with mortality records. But, first a few background words about vital records in this country. Vital records are records of the vital events of birth, marriage, and death. We have no central respository of human records in the United States. We have no national library; the Library of Congress is the closest thing. The National Archives have a mandate to collect all national records but their resources are limited. There are national libraries in certain disciplines such as the National Library of Medicine and the National Library of Agriculture but they serve different constituencies and thus there are overlaps and gaps between them.

Our system of vital records is derived from our philosophy of government. In many European countries where there were state churches, that church recorded vital events of the people. In the course of history, this function was taken over by the national government. The basic jurisdiction was first church and then national government. In this country, even in colonial times, there was no state church. Jurisdiction was and still is with local government.

In general, the basic jurisdiction for vital records in the United States is the county. There are about 3150 counties depending on your definition of county. There are some complexities. In New England the town has been a record center. Alaska does not have counties; thus we usually have to consider Alaska as a county as a whole. There are several independent cities in the United States which are not within the jurisdiction of a county including Carson City, Nevada; St. Louis, Missouri; Baltimore, Maryland; and several cities in Virginia. There are counties in which the major city and the county are handled as if they were the same such as Denver, Colorado. In some files the boroughs of New York City are listed separately and in other files they are grouped together. Most counties were formed before 1900 but even now counties are being formed or merged. The main trouble area is Virginia with the coming and going of the independent cities, but also this has occurred in other states such as Georgia. There is no uniformity in the size, shape or population of counties. Generally, counties in the East are smaller than in the more sparsely settled West. And of course, Texas has far more counties than any other state. County names reflect our rich and varied heritage. The most popular names for counties are early American patriots but counties are named for famous Indians even Indians of poetry, and for animals, birds, fishes, rivers, and religious figures. There are even county names of totally unknown origin.

The national responsibilty for compiling vital records from the local governments is vested in the National Center for Health Statistics. Vital records are published annually in bound volumes and computer tapes. Due to the considerable time and effort of gathering and compiling, publication is always a few years after the events.

For each death certificate, data is recorded on the public release tapes for cause of death given as a code from the International Classification of Diseases (ICD), county of usual residence of the decedent, county in which the death occurred, date of death, sex, race, age, and some identification codes. However, there are no person identifiers. Only the agencies with primary jurisdiction could identify the decedent. The National Death Index being produced by the NCHS may provide some help on this in the future. It is a long way from knowing that mortality and food intake tapes exist to being able to link the information in these tapes. And this is where the process of record linkage really begins.

PROBLEMS

There are several problems which will take years to resolve. Some of the problems are shortage of funds, nonuniformity in data recording and coding, and large volume of data. The shortage of funds is well known. Last year we talked of 4 large national Integrated Database systems. At the present time all of the 4 are shrinking, laying off staff and taking other painful steps. Probably some of them will disappear before this time next year. You don't need to be told of the decreases in federal grant and contract funds.

Nonuniform recording and coding of data is a problem well recognized by this group. For us comparing apples and oranges is more than just a trite saying, it is a real life problem. Populations of people are commonly identified by geographic areas and those areas carry different codes. For example, the county this meeting is being held in has one code in mortality files and another code in food intake files.

The large volume of data is illustrated by mortality files. At nearly 2 million deaths per year, the volume of mortality records grows so rapidly that in just 10 years this room could not hold all the records if they were in card files. Without proper organization it would be humanly impossible to retrieve any given set of records. Even large computers have difficulty if organization is lacking. A researcher spent $20,000 for computer time alone to retrieve the death records for just one cause of death.

APPROACHES TO SOLVE THE PROBLEMS

Some approaches to solving these problems are suggested but due to the large effort required, not all of these problems have actually been solved. This is a presentation of issues and approaches. More work must be done to produce a definitive system.

The problem of shortage of funds can be addressed in a couple of ways. The first way is improved cooperation. Researchers in government, academia, and industry can work together more closely. There are some pieces of information which are confidential and even some applications of public knowledge which can not be divulged or shared. But these confidential items can be safeguarded while at the same time sharing publically accessable data, ideas, and efforts to the increased benefit and decreased expense of all. I would welcome contact with any of you who feel we may have some interests in common. Tasks could be divided between researchers who individually could not handle the whole project. Bartering could to a large extent replace direct transfer of funds to accomplish the same research in the end.

The second suggestion to decrease costs is use of small machines such as microcomputers and word processors which are becoming common in offices, laboratories, and homes. Use of them can increase productivity and decrease costs. Through telecommunication and other means, data can be transfered almost anywhere. Data tables, figures, text, and bibliographies can all be produced and melded into finished products without cutting, pasting, photocopying, or undue redlining and hand editing.

The problem of nonuniform recording of data and coding was addressed in the previous talk. Suffice it to say here that production of standard codes should be strongly encouraged.

Solving the problem of large volume requires effort and effective organization. Organization of mortality files has reduced retrievel costs to less than 1% that of searching the original files. This organization consisted first of reducing the volume as far as possible and then arranging the residue for efficient access. The steps are strip, delete, fold, sort, index, and subset. For a one time only project, one could skip directly to the subset step but if more subsets were needed they would each cost as much as the first. Going through all the steps produces a system to greatly reduce subsetting costs in the future.

Strip

As received from NCHS, each death record is 160 characters long or the equivalent of two 80 column punch cards. Some of the data is recoded; e. g., age is grouped into 10 year age intervals in one field and into 5 year intervals in another field. By eliminating blank columns and recoded columns we reduced the record length to 44 characters.

Delete

After stripping out blank and redundant data, the organizer may decide that some of the data for each record are not needed and delete them. By this process, combined with binary coding, we reduced down to only 4 bytes per record but we paid because the resulting file was machine dependent and we couldn't use it when we lost acess to the machine it was built on. A philosophy we adopted early in our system development efforts was that the system should be capable of responding to 90% of the expected requests while the other reqestors would have to reevaluate their requests or bear the cost of processing larger records.

Fold

Sometimes there are separate records with identical data. Space can be saved by folding or summing over these records, e.g., in the mortality records we fold down to 5 year age intervals. The technique of folding is of even greater value in a subset where the structure of the file does not have to be maintained as does that in the main system file.

Sort

A decision has to to be made as to what orientation users have in retrieving records. In virtually all requests we have seen for mortality records the user is interested in certain causes of death with specific cancer types leading the list. Therefore, the mortality file for each year was sorted by the cause of death code which is ICD (International Classification of Disease).

Index

The next step is to create an index to tell which part of the file contains the records for each cause of death. Retrieving now becomes rather efficient because the tape drive can spin down to the first record of the set desired and read any number of records ignoring all the rest of the tape.

Subset

A file can be created of just those records needed for a project. The subset can then be further processed by the mainframe or downloaded for processing by a smaller computer.

After the mortality records are in a practical system we will be able to link death records with food intake records. But, because populations in most national food intake surveys are not identified down to the county level but only to the multi-state census division level, it will be necessary to aggregate death records up to the census division level for data linkage.