FOOD COMPOSITION DATA BASE FEATURES

Jean A. T. Pennington, Ph.D., R.D.
Food and Drug Administration, Washington, D.C.

Introduction

As new or prospective users of food composition data bases, you may be (or will soon be) developing your own data base or adopting or modifying a previously developed data base. The large number of food composition data bases within government, academic, food industry, and private institutions emphasizes the necessity of designing data bases with unique features to serve specific needs. Because the development of a data base is a costly and time-consuming venture, it is important to identify at an early stage those features that will serve your specific needs so that you may incorporate them into your data base as you develop it, or you may select a data base that contains these needed features.

The features of a data base are usually a direct reflection of the functions served by the base. The most common use of food composition data bases is to assess the nutrient levels of diets. These diets may have been collected in food consumption surveys or they may be the diets of patients or clients who are consulting with dietitians or nutritionists. However, even the data bases developed for the purpose of evaluating diets vary considerably depending upon the population groups evaluated, the nutrients evaluated, and the purpose of the analysis.

In addition to assessing the nutrient content of diets, food composition data bases have numerous other uses. The Food and Drug Administration (FDA) uses food composition data bases to develop definitions for imitation foods and food substitutes; to develop standards of identity for specific foods; and to develop regulatory policies concerning nutrient fortification and use of food additives. The United States Department of Agriculture (USDA) uses food composition data bases to determine the adequacy of the United States (U.S.) food supply to meet nutritional needs. Food composition data bases are used to plan menus for hospitals, military feeding facilities, school lunch programs, and other group feeding programs. The food industry uses food composition data bases to develop nutritional labeling and labeling claims for their products.

To serve these various purposes, food composition data bases have characteristics that are specific for user needs. These characteristics include the types of foods; the number of foods; the food descriptions; classification or coding schemes; the types and number of nutrients; and the expression of the nutrient values.

Development of a Data Base

The characteristics of a data base generally reflect the effort and decisions made by the data compiler(s). Food composition data for data bases are gathered primarily from USDA publications, the scientific

literature, the food industry, and unpublished laboratory reports. At the first level of data collection, the compiler gathers the original reports or publications and transfers specific information to a file which may be written or computerized. This information includes the name of the food, its descriptors, and nutrient data in the original units. The compiler should also record the bibliographic reference, analytical methodology, quality control information (e.g., duplicates, blanks, standard reference materials), number of samples, method of sampling, and any other information (e.g., maturity, season, soil type, geographical location) that may be pertinent to the ultimate use of the data base. The best method for a hand-written version is to use one page per food per source. The nutrient composition data should then be converted to standard units (usually values per 100 grams wet weight of food and/or values per typical serving or other practical portion).

An evaluation of the compiled data may reveal that there are many or several sources for the nutrient content of some foods, but only partial data for other foods. There will probably also be conflicting data from various sources. The data compiler (a person knowledgeable about food composition, analytical methodology, and statistics) must carefully scrutinize the data from the various sources and make decisions regarding: combining partial data from various sources; aggregating data from various sources; and not using data that appear erroneous. If data are to be compiled from several sources, it is essential that the food descriptions and levels of some of the more basic nutrients (e.g., water, protein, calories) agree. The aggregation process may involve calculating averages or weighted averages for nutrient values. For practical purposes, the end result should be a separate single entry for each nutrient in each unique food product. The manner in which the values are derived must be documented, and the compositional data must continuously be reevaluated as newer data become available.

Foods to Include

The foods included in a food composition data base will reflect the functions of the base. A data base developed for a food company may only need compositional data for company and perhaps competitor products. A hospital data base should include the institutional foods, special recipes, and medical foods used by that institution. Similarly, data bases used by other institutions (military feeding facilities, school lunch programs, dormitories, prisons, boarding schools) should reflect the foods commonly served by these institutions. Data bases used for assessment of patient/client diets, epidemiological studies concerning diet and health, or other nutrition studies that include the general population should reflect the current food supply to include traditional, commercial, homemade, and restaurant foods.

To select the foods for a data base to be used in food consumption studies, it must be clear what dietary information will be elicited during the interview or by the food record instructions. If the subject is not asked the appropriate questions regarding the identification of each food, the subject will probably not provide the information. The code selected

for a product such as a cheeseburger could reflect a generic product as easily as a homemade or fast food brand name product. Likewise, low-fat milk may be a single code, or it may have several codes depending upon the level of fat (1/2%, 1%, or 2%) and the presence of nonfat dry milk solids.

The concern is the level of detail needed for your work. Do you need to know the brand name of the cheeseburger or the percent of fat in low-fat milk, or will the more generic terms suffice? It is possible that for some foods you will need more detail than others, depending upon the nutrients you wish to emphasize. You must also be realistic about the level of description you can expect from the participants in a food consumption survey. If your sample is large and the amount of time with each person is short, the amount of information to be elicited will be much less than if you have a small sample, repeated subject contact, and enough time to ask detailed questions.

If your data base only has specifically described food items, and the subject is unable to remember food details, the coder must guess at appropriate codes. Many such guesses might affect the validity of your results. The greater the number of food codes, the greater is the burden upon the coders and the greater is the chance of error.

If the purpose of your data base is to evaluate diets, you need only include foods as consumed. In this case you will not want to waste space with food items like raw meat, dry oatmeal, unpopped popcorn, etc. Your system should allow for easy addition and deletion of foods, as appropriate, to keep pace with the ever changing food supply.

Some thought should be given to including water as well as vitamin/mineral supplements, other supplements (protein powder, bone meal, bran, wheat germ oil, cod liver oil), and medications that have energy or nutrient values (e.g., cough syrup or antacids containing calcium carbonate) or that may interfere with nutrient absorption. Regional or local analyses for the nutrient content of water will probably be needed to assure the accuracy of the water data. The inclusion of water, supplements, and medications in a data base is optional and depends upon the purposes for which of the data base will be used.

Number of Foods

The number of foods in a food composition data base becomes a function of the detail of your food descriptors. The greater the level of detail per food, the more foods you will need to include. For example, it may be sufficient to have raw spinach and cooked spinach in your data base, or you might need more detail about the cooked spinach such as whether it was fresh cooked; canned; frozen, leaf; frozen, chopped; or frozen, creamed. Virtually every simple food item in a data base can be expanded in this way. The expansion is even greater if you choose to include special dietary products (e.g., foods with low or reduced sodium, fat, or cholesterol) or brand names of products.

Brand names for some products are important for product identification. Subjects undergoing a dietary interview might be more likely to report brand names for some candy bars or ready-to-eat cereals than to identify them as "chocolate covered nougat and caramel" or "O-shaped oat rings." Brand names may also be important for frozen entrees and fast foods because of differences and proportions of ingredients and hence nutrient values from one brand to another. However, a data base for a food consumption study will probably not need to include different brands of foods with similar nutrient values (e.g., corn flakes, chocolate cake from mix, mashed potatoes from instant, macaroni and cheese from box mix, frozen orange juice). A data base for a regulatory agency or for a food industry may, however, need such brand name data, especially if there is concern with label compliance, nutrient fortification, and/or label claims.

It may be that nutrient data for a certain brand named product (e.g., Brand Y macaroni and cheese) is used for a group of similar products (e.g. all macaroni and cheese mixes). This fact should be documented in your data base files at the first level of data collection. The brand name should not, however, be associated with the food description record in the coder's manual as this will cause confusion for the coders.

The number of foods in currently available data bases vary considerably. The data base used by the Neonatal Intensive Care Unit at the Milwaukee County Medical Complex contains 57 foods, all of which are infant formulas; the FDA's Total Diet Study has 234 foods; and the Minilist of the University of California, Berkeley has 235 foods. In contrast, the data base of the Ohio State University contains over 8,500 foods. The data bases used in the two most recent national food consumption surveys, the USDA 1977-78 Nationwide Food Consumption Survey (NFCS) and the Second National Health and Nutrition Examination Survey (NHANES II) contained approximately 3,700 and 2,600 foods, respectively. These data bases were sufficient to evaluate 24-hour recalls and two-day diaries for 30,000 persons (NFCS) and 24-hour recalls for 20,000 persons (NHANES II). The diets were those of subjects selected to be representative of the non-institutionalized U.S. population with regard to age, sex, income, race, region, and urbanization.

The number of foods in a data base should reflect the unique nutritional differences of individual foods and the importance of these nutritional differences to the uses of the data base. For example, a study concerned with fat intake and its relationship to cardiovascular disease may have replicate entries for the same item (e.g., fried eggs, homemade chocolate cake, pancakes, french fries) depending upon the type of fat used (e.g., butter, margarine, corn oil, vegetable shortening, peanut oil, palm oil, lard, etc.). This information could be very important in evaluating the relationship between intake of fatty acids and cholesterol with the incidence of heart attacks or strokes. This same data base may, however, have broad food aggregations for foods that are low in fat (e.g., fruits and vegetables). A data base developed for an epidemiological study of diet and cancer would probably focus on foods that are major sources of carotenes, dietary fibers, vitamin C, and fat. Again, there are no right or wrong choices for data bases, only appropriate choices to suit specific needs.

FDA's Total Diet Study monitors the levels of contaminants and minerals in the diets of selected age-sex groups though yearly analysis of 234 "core" foods. Each food represents a group of similar foods. For example, the frozen commercial apple pie represents all commercial, homemade, and restaurant fruit pies, turnovers, pastries, and strudels. These "core" foods are collected, prepared for consumption, and analyzed four times per year for 11 essential elements and over 200 pesticides residues, industrial chemicals, and toxic elements. The heavy analytical burden of this program requires that the number of food samples be kept relatively low. Though the numbers of foods are low, the Total Diet Study has successfully monitored the levels of contaminants and nutrients in the U.S. food supply since 1961. Included among the Total Diet Study nutrient findings were increases in iodine from dairy and grain products and a decrease in iron intake of infants due to the decrease in the fortification level of this mineral in infant cereal.

Food Descriptions

Depending upon the uses of the data base, food descriptions may range from the general (e.g., fried chicken) to the very specific (e.g., chicken, roaster, thigh, batter dipped, fried in hydrogenated cottonseed oil). Some data bases such as those for national food consumption surveys must include the full range of descriptions from the general to the specific. The data base for the NFCS contained many "not further specified" (NFS) entries (e.g., milk, NFS; meat, NFS; sandwich, NFS) because many of the participants could not adequately describe the foods they consumed. Nutrient values for these NFS foods were estimated so that nutrient intakes could be estimated. This feature is far superior to having only specifically defined foods and forcing the coder to select what they consider to be the "best" food code.

The level of detail you want in your food descriptions is dependent upon the types of questions you wish to address. If your data base is used to estimate nutrient intake, the level of food description should parallel the information collected from study participants. This is dependent upon the food consumption methodology and the level of intelligence and patience of the subject and interviewer. Different food consumption methodologies (24-hour recalls, food diaries, food frequencies, etc.) do elicit somewhat different food descriptions and detail.

For the food descriptions of your data base, consider whether you need to know the source of the food (e.g., homemade, commercial, fast food, deli, restaurant); their preservation method (fresh, frozen, salted, dehydrated); the packaging material (metal, paperboard, glass, cellophane); the preparation method (boiled, roasted, fried, microwaved, pressure cooked, steamed, barbequed, stir fried). For mixed dishes you may want a recipe file which lists the ingredients of commercial and restaurant foods and quantities of ingredients for home prepared items.

It is often difficult to determine what is a single food (requiring no recipe) as opposed to a mixed dish (which requires a recipe). Foods to which only water, salt, herbs, and spices (including garlic and onion) have been added are usually considered single foods. Foods with added fat

or prepared with a fat, sauce, or gravy are often considered single foods as well. For some foods you need to decide if you will have entries for the mixed food or will code added ingredients separately. For example, for a baked potato with added butter, margarine, sour cream, or cheese sauce, the baked potato and topping may be coded separately or there may be separate codes for baked potato with a specific topping. The same is true for coffee with milk or cream and/or sugar.

You may wish to establish an order to the descriptive terms of foods in your data base. It is often convenient to list the raw/fresh item first, followed (if applicable) by the cooked item, followed by various processed forms of the item listed alphabetically:

peas, green, immature
    raw
    fresh, cooked
    canned
    frozen
    frozen, boil-in-bag
    frozen, in butter sauce
peas and carrots
    canned
    frozen
peas and mushrooms, frozen
peas and onions
    frozen, in butter sauce
    frozen, in cream sauce

It is extremely useful to develop a dictionary of descriptive terms. This is important if several persons are maintaining and updating the data base. The dictionary should also include "use" and "used for" terms (e.g., salted: use salt added; cowpeas: used for black-eyed peas). The dictionary is most easily handled as a computerized file.

Classification or Coding Schemes

A classification system is of particular importance to the person who must code food consumption data. To assess the nutrient content of a diet, the coder must find the appropriate food code for each food using a code manual. The food codes and quantities of foods consumed may then be entered into a computer, and a software program may perform the necessary calculations to determine nutrient intakes. The most common classification scheme used for food composition data bases is based on food groups (meat, dairy, fruit, vegetable, etc.). There may be subsequent categories within these major food groups. For example, meat may be subdivided among beef, pork, lamb, veal, and game, and beef may be further categorized by round, rump, loin, hamburger, porterhouse steak, T-bone steak, etc. The foods within a subcategory are usually listed alphabetically, and each food is assigned a code number.

The USDA NFCS classified foods into 10 major groups and various major and minor subgroups. These three types of classifications are denoted by the first three numbers of a seven digit code number. This classification and coding scheme allows data retrieval for the major food groups, major food subgroups, and minor food subgroups in addition to retrieval of data for individual foods. For example, one could use appropriate codes to estimate consumption of whole fresh, fluid milk (a specific food); all fluid milks (a minor food subgroup); all milk and milk drinks (a major food subgroup); or all milk and milk products (a major food group).

Classification of foods by food groups produces a strict hierarchical scheme. Although useful for some food consumption studies, it is not efficient for other data base uses such as identifying foods by other characteristics such as food source (soy, beef, apple) preservation method (freezing, canning); or packaging materials (plastic, paperboard, metal). Also a food group classification leads to problems in placing many foods. For example, Irish coffee could be grouped with coffee or with alcoholic beverages; cafe au lait could be grouped with coffee or milk beverages; mixed dishes could be grouped together as a "mixed dish group" or each mixed dish could be grouped according to its major ingredient. The NFCS uses the latter method. It is often difficult to determine the major ingredient in mixed dishes such as beef-vegetable stew or lasagne. Fast foods may be grouped together as a "fast food group" or placed according to individual food type (e.g., sandwiches, vegetables, beverages). Commercial baby foods may also be grouped together as a "baby food group" or placed in appropriate food categories (e.g., meats, fruits, vegetables, cereals). Soups could be placed in a "soup group" or placed according to major ingredient (vegetable, meat, pasta, bean, etc.). These are decisions to be made by the data base compiler.

The FDA has developed an internal coding sytem to allow retrieval of data on the basis of food descriptors. This system can be applied to any food composition or food consumption data files. The FDA system describes foods on the basis of 11 factors: product type; food source; part of plant or animal; physical state, shape or form; degree of preparation; treatment applied; preservation method; packing medium; container or wrapping; food contact surface; and user group. Product type refers to food group. Food source and part of plant or animal describe the origin of the food or the major ingredient if it is a mixed dish. Degree of preparation; treatment applied; and preservation method refer to industry processing of the food. Packing medium, container or wrapping, and food contact surface refer to industry packaging materials. User group refers to the major consumers of the food. There are almost 2,000 factor terms which are fully defined by the FDA thesaurus which can be used in various combinations to retrieve foods from data files. For example, one could retrieve all canned foods that contain mushrooms or all soy-based foods in plastic containers.

If you are developing a data base for international use, you will probably need to include the scientific name (genus and species) in the food

description. If you are coding foods from other countries, you will need to retain the original name (in the native language), and determine the English translation, the scientific name, the language used, and the geographical location.

Types and Numbers of Nutrients

The nutrients included in a data base (like the foods included) remain a function of the purposes for which the base will be used. A data base used for institutional meal planning may include only the proximate nutrients and the major vitamins and minerals. A data base used by a food industry may include those nutrients used for food labeling. Data bases for epidemiological studies will focus on those specific nutrients believed to be associated with the diseases/disorders of concern to the study. A multipurpose data base tends to be all-inclusive (i.e., to include all available nutrients for all available foods). This often leaves many missing values in the data base as nutrient values for some foods are quite extensive, while values for many other foods are only available for the major nutrients.

The number of nutrients available in currently developed data bases ranges from 4 to over 100. Many data bases retain only the basic 17 nutrients of the original USDA Handbook 8. Others may also include vitamin B-6, vitamin B-12, folacin, carotenes, dietary fiber, and trace minerals such as zinc, copper, manganese, magnesium, iodine, and selenium. Other data bases include individual sugars, fatty acids, and amino acids. The necessity of including crude fiber and ash in the data base may be questioned, although you may wish to retain them in the first level of data collection. Some data bases have separate entries for plant and animal sources of protein and/or iron.

Expression of Nutrient Values

It is important that the data in the data base be on a fresh (wet) weight basis. The compiler may convert dry weight values to fresh weight values, but unfortunately, most references do not give the residual moisture of a dried food. The best that can be done (if communication with the authors of the references is not feasible) is to assume zero percent moisture in the dry product and footnote the fact that your nutrient values are estimated from dry weight values. One must be careful to discern between nutrient values listed as "percent dry weight" versus "percent dry ash."

For most purposes, nutrient values per 100 grams are most useful in a food composition data base. Information on the weight of standard servings or single service portions should also be collected. This information is essential for converting food consumption data into nutrient consumption data. If you are evaluating food frequency data which is not quantitated by serving portions, you will need to retain the nutrient data in your base by standard serving portions.

Data that have been imputed or calculated from recipe ingredients should be footnoted as such. Missing values should be imputed, especially for

foods that are recognized sources of particular nutrients. Otherwise tabulations of total daily intakes should be denoted as lacking values from one or more foods. This is of particular concern when intakes appear to be marginal compared to Recommended Dietary Allowances or Estimated Safe and Adequate Daily Dietary Intakes. The imputed and calculated values should remain flagged in the data base and replaced when possible by analytical values.

The ideal expressions for nutrient levels of foods are means with standard deviations and medians. Unfortunately, most data bases can only include means, and certainly those who use data bases to assess intakes of nutrients are primarily concerned with mean values for each nutrient in each food. However, those who use data bases for other purposes (food labeling, regulatory purposes, data comparisons on single products) may need to know the extent of nutrient variability and the reliability of the data. USDA has begun a practice of using confidence codes to indicate the reliability of nutrient data.

Data compilers also need to be concerned about the units used for nutrient measures. For several nutrients, there are several acceptable units (e.g., international units and retinol equivalents for vitamin A, international units and milligrams for vitamin E). Vitamin E may be expressed as total or as alpha-tocopherol. Energy may be expressed as kilocalories or kilojoules. Consideration must be given to the units desired and the units most commonly used and/or available. In some cases, it might be desirable to include values for some nutrients in several units.

Summary

There are many currently used food composition data bases. Even though most of them incorporate food composition data from similar sources, there are unique characteristics and features of these data bases that reflect their individual functions. If you are at the point of developing, adopting, or modifying a data base, think carefully about the functions that your base will serve. If you are in academia, your base may need to serve the needs of research in several departments plus analysis of student diets; if you are in private practice your base may serve mainly to analyze the diets of individual clients and patients. Data bases for industry may need only company products and only those nutrients listed on food labels. Data bases for large diet-health epidemiological studies or national food consumption studies must consider variables associated with eating habits such as age, sex, race, region, urbanization, season, income, religion. Food consumption studies directed at specific population groups may contain a smaller number of foods.

With regard to nutrients, consider the questions you wish to address and consider all the nutrients that might be of concern. It is difficult and inaccurate to go back and add nutrients to an already developed data base. One might attempt to include all available nutrient values in a data base, but only use those that are essentially complete for evaluating food consumption data. A review of the data bases that are currently in use may help to determine those characteristics regarding foods, descriptors, nutrients, and data expressions that would be appropriate for your data base. An awareness of these features at an early stage will aid in the development or selection of a data base to serve your needs.