DESIGNING A COMPUTERIZED NUTRIENT DATABASE FOR MEDICAL RESEARCH APPLICATIONS

I. Marilyn Buzzard, Ph.D., R.D., Yvonne A. Sievert, M.S., R.D., Sally Schakel, R.D., and Mary Stevens

Nutrition Coordinating Center
University of Minnesota

Purpose and development of the NCC nutrient database

The nutrient database developed at the University of Minnesota's Nutrition Coordinating Center (NCC) was designed to facilitate nutrient calculation for a standardized system of collecting and analyzing dietary data for medical research studies. The NCC Nutrition Data System (NDS) was established in 1974 through the collaborative efforts of experts in nutrition, computer science, education and statistics (1). The development and maintenance of the NDS has been supported by the National Heart, Lung and Blood Institute (NHLBI). The original emphasis on intake of dietary lipids for cardiovascular research has been maintained as the system has been expanded to include dietary components of research interest to all major diseases.

Food specificity

The design of the NCC nutrient database differs from that of most large nutrient databases in the manner in which foods are described and in the flexibility provided for specification of fat and sodium intake. Through the use of computerized algorithms to calculate added fat, salt and flour used in food preparations and the use of numerous coding guides and coding rules to specify the coding for brand name products and various food combinations, the system can accommodate nutrient calculations for over 150,000 foods using a database of less than 2,000 entries. A schematic representation of the design of the NCC nutrient database to minimize database redundancy while maximizing specificity of food descriptions is shown in figure 1.

To permit maximum specificity for fats used in preparation, few entries for foods cooked with fat are included in the database. The computerized food preparation algorithms permit detailed specificity of the type of fat used in preparation. The amount of the designated fat, in addition to the appropriate amounts of salt and/or flour, are automatically calculated based on the amount of the food consumed. For example, there is no entry in the database for breaded and fried chicken. Coding of breaded and fried chicken requires selection of one of four entries for chicken (light or dark meat, with or without skin) and selection of one of a possible 141 cooking fats included in the current database. Thus, a total of 564 combinations of chicken and cooking fat are possible as shown in figure 2. Preparation with or without salt increases the possible number of preparations to 1,128. If all of these preparation options were separate entries in the database, the result would be a massive database that could not be efficiently maintained.

Detailed specificity for fats is also permitted for recipes containing fat as a major ingredient; any one of the current 141 fats in the database may be selected for the recipe. On-line coding procedures which facilitate rapid access to all recipes in the database allow further specificity for recipe ingredients such as fat levels of milk or ground beef. Ingredients and amounts may be altered to reflect individual documentation of recipe modifications. For example, if oyster stew was documented as having been prepared with 1% milk, without salt and with only half as much margarine as specified in the recipe, the recipe would be accessed on the coding screen, the whole milk code replaced by the 1% milk code, the salt deleted and the amount of the selected margarine reduced by one-half.

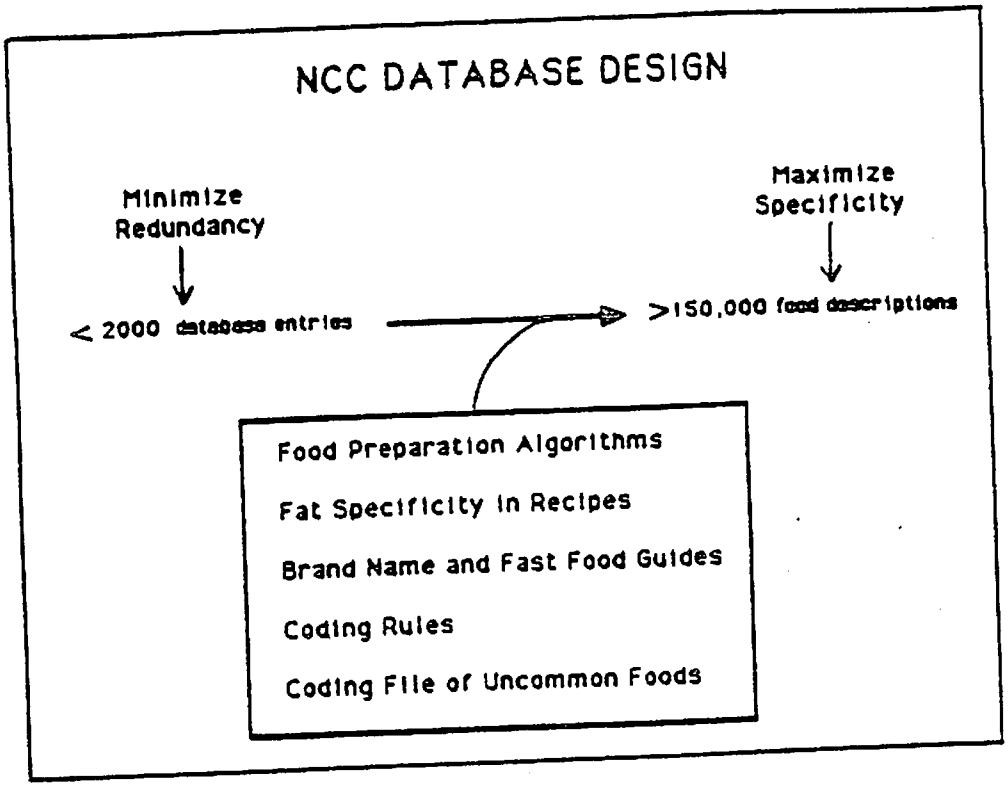

Figure 1. Schematic representation of the design of the NCC nutrient database.

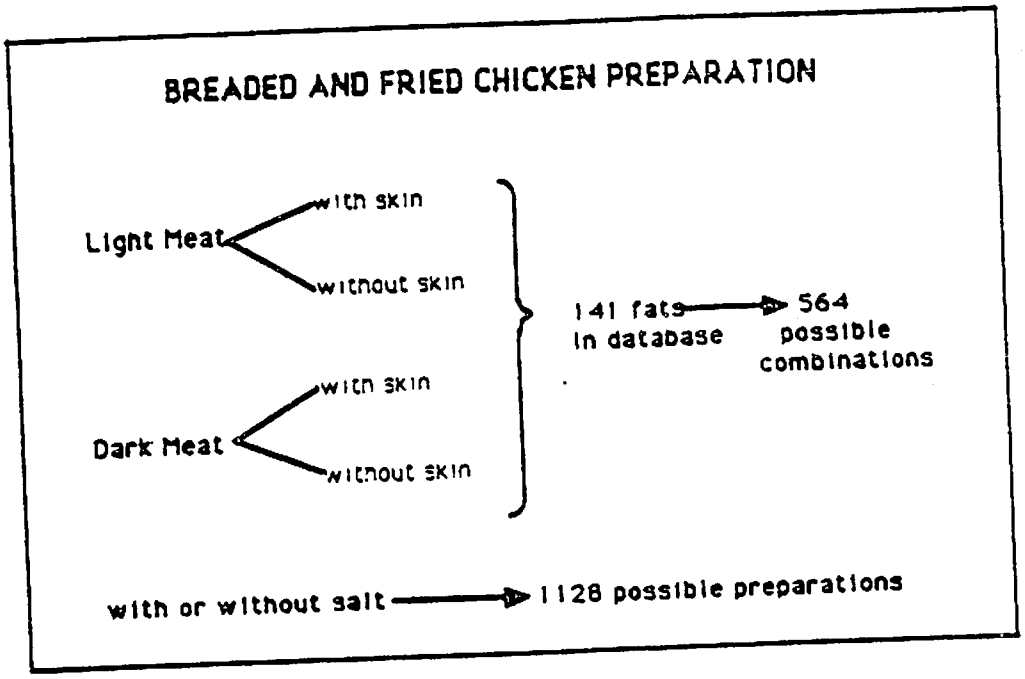

Figure 2. Diagram of number of possible breaded and fried chicken preparations using the NCC system.

All commonly consumed commercial products are listed in brand name coding guides which specify the appropriate database entry to use for each brand name item. The 39 brand name coding guides currently used for the NCC system include foods such as cookies, crackers, margarine, salad dressings, cheeses and vitamin-mineral supplements. For example, the current database includes 22 generic entries for salad dressings based on ingredient oils, total fat and sodium. The salad dressing guide lists 207 brand name salad dressings and specifies which of the 22 salad dressing entries in the database best represents each brand name product. Similarly, the margarine guide directs the coding of approximately 350 brand name margarine based on 82 entries in the database.

Other NCC procedures for minimizing database redundancy include the use of standardized coding rules which specify the coding for numerous food combinations and the maintenance of a file of coding directions for infrequently consumed foods. These procedures ensure standardized coding while further reducing the number of entries that must be maintained in the database.

Nutrients

The most recently released version of the NCC nutrient database (Version 14) includes values for 73 nutrients. These nutrients are listed in table 1. The working version of the nutrient database, known as the Reference Food Table, contains values for 107 nutrients. The additional 34 nutrient fields are in preparation for future research use but are not yet considered complete enough for use in nutrient calculations. The incomplete nutrient fields include 18 amino acids, manganese, chromium, cellulose, hemicellulose, lignin, gums, and 10 sugars and sugar alcohols.

Quantification of amounts

Food amounts may be entered in any standard unit of weight or volume since the calculation software includes conversion factors for all standard units. Conversion factors for food specific units such as a slice of bread, a large egg or a stalk of celery are also maintained in the database. Multiple densities are maintained for foods that may be consumed in more than one form. This allows all forms of the same food to be represented by a single entry. For example, cheddar cheese measured as a solid piece, a cubed volume or a grated volume are all represented by the same database entry using the appropriate density for each form. The calculation software can also be modified to accept amounts entered in terms of study specific food models.

Sources, selection and documentation of nutrient data

The NCC nutrient database is a compiled database derived primarily from USDA sources including the Nutrient Data Base for Standard Reference, the revised sections of Handbook 8, and tables of provisional data. Other data sources include the scientific literature, information from manufacturers, international tables of food composition and standard cookbook recipes. Approximately 30 scientific journals are routinely reviewed for food composition data; others are accessed on an ad hoc basis as relevant articles are identified through standard reference indices. Information on thousands of commercial products is maintained in a manufacturers' reference file. Updated ingredient and nutrient information is requested at least annually from several hundred major food manufacturers. This information is used to update database entries and brand name guides. An average of five brand name guides are updated each month. Label information is occasionally used to specify the coding for a particular product if the required information cannot be obtained from the manufacturer. For example, it is sometimes difficult to obtain manufacturers' nutrient information on products distributed under store name brands since the distributors are not the manufacturers. Label information is also used to specify product ingredients when manufacturers do not provide this information with the nutrient data.

When multiple sources of data are available, selection of nutrient values for the database is based on criteria such as the source of the data, the analytic methodology used, laboratory

quality control procedures including use of reference standards, representativeness of the sample and the sample size. If USDA data are available, they are given priority. Refereed scientific journals are prioritized over non-refereed journals, and scientific publications are preferred to non-scientific published materials. Preferred laboratory methods are based on USDA guidelines whenever possible.

Occasionally, when data cannot be obtained from any other source, the NCC generates the required data. For example, the NCC funded two studies conducted by the University of Minnesota Food Science and Nutrition Department; one study involved the determination of fatty acid values in candies, and the other involved the estimation of the uptake of sodium in cooked pasta. Results of these studies have now been published (2,3). Generation of food density data by the NCC involves activities such as the preparation of recipes, measuring the volumes of candy bars, and weighing measured volumes of dry mixes.

The NCC nutrient database contains a minimum of missing values. Since missing values are calculated as zeros in the analysis of research data, missing values are permitted only when there is no available information on whether or not a nutrient is present in a food. Table 1 includes the percent complete for each nutrient n the current database as well as the percent of imputed values. Assumed zeros, such as cholesterol in plant products, are not classified as imputed values. Most imputed values are either values substituted from a similar food (such as the micronutrient values for low sodium cheeses substituted from the regular cheeses) or calculated values. Calculated values are based on one of the following procedures: 1) calculation from another form of the same food using yield and/or retention factors such as raw to cooked conversions; 2) calculation from a related nutrient in the same food such as retinol and beta carotene from total vitamin A content; or 3) calculation from a product formula or a product ingredient list. Imputed values are replaced by analytical values as soon as they become available.

A computerized reference code system documents the source of data for every nutrient and non-nutrient value in the database. A free-text comment section is used to describe data sources in more detail such as the actual calculations used to impute values, analytical methods used to obtain certain values or recipe sources. A paper file is also maintained for every entry in the database. All changes ever made to the entry and the rationale for every change are included in this file.

Maintaining database stability and reanalysis capability

To meet the needs of long term research studies, the NCC must provide database stability for ongoing studies as well as the capability to reanalyze data collected in the past using an updated version of the Food Table. This is accomplished by limiting all updating to the Reference Food Table which is the working version of the nutrient database. Periodically (approximately every 6-12 months), all updating effort ceases on the Reference Food Table, and the database nutritionists spend several weeks verifying the accuracy and internal consistency of the database using a series of computer generated integrity reports. Prior to the release of a new version of the database, nutrients are calculated for a test set of dietary intake records, and the calculated values are compared with those from the previous database version; all differences must be verified as due to intentional changes in the database since the previous version.

At the beginning of a study, the most recent version of the nutrient database is assigned to the study. For the duration of the study, no changes are permitted in the database other than the addition of new entries for new foods that come on the market. At the end of the study, investigators may elect to rerun all study data on a newer version of the nutrient database. To allow the capability for reanalysis, no entries are ever deleted from the database, even though a product is discontinued. Database entries for discontinued products are "deactivated" which means that they are no longer used in coding. When a product reformulation affects the nutrient

profile of the item, the previous database entry is deactivated and the reformulated product is assigned a new food code. Deactivated entries are maintained in the database so that previously collected data can be reanalyzed on an updated version of the database.

Availability of NCC nutrient data*

The NCC nutrient database is available to the public on magnetic tape. Computer generated reports can be produced to meet individual needs. Other available items are the 500 page coding manual and monthly updates of the coding guides. A nutrient calculation software system designed for use on an IBM-XT or compatible microcomputer with hard disk is under development. Features of the NCC Nutrient Calculation System (NCS) include data entry using food names or food codes, prompts for added fat and salt in food preparation, screen display and running totals for five nutrients selected by the user, access to 73 nutrient fields for each food or an entire menu or food record, recipe breakdown screens for recipe modification, and vitamin/mineral supplement screens. A master index of common food descriptions will include brand names. The system is targeted for completion by the end of 1987.

* Requests for information on NCC materials may be directed to: Jeanne Nelson, Nutrition Coordinating Center, University of Minnesota, 2829 University Avenue S.E., Minneapolis, Minnesota 55414; 612-627-4871.

REFERENCES

1. Dennis, B., Ernst, N., Hjortland, M., Tillotson, J., and Grambsch, V. The NHLBI Nutrition Data System. J. Am. Diet. Assoc. 77:641, 1980.
2 Ahn, P.C., Kassim, N. and Hegarty, P.V.J. Total Fat and Fatty Acid Composition of Commonly Available Chocolate Candies. J. Am. Diet. Assoc. 79:552, 1981.
3. Albrecht, J.A., Asp, E.A., and Buzzard, I.M. Contents and retentions of sodium and other minerals in pasta cooked in unsalted or salted water. *Cereal Chemistry*, in press, 1986.