Maintenance of Food Composition Databases

Sally Schakel
Brian Westrich
University of Minnesota
Minneapolis, Minnesota

Introduction

The Nutrition Coordinating Center (NCC) at the University of Minnesota maintains two food composition databases: 1) a Nutrient Database that contains descriptions, nutrient values and reference codes for representative foods in the American diet, and 2) a Brand Name Database that contains nutrient, ingredient, density and serving information provided by food manufacturers for brand name products. The Brand Name Database is used to provide information for foods in the Nutrient Database and to aid in determining coding for brand name products by matching them to similar Nutrient Database entries.

Maintenance of the Nutrient Database

The NCC Nutrient Database contains approximately 1500 elemental (non-recipe) foods and 1000 recipes or formulations of commercial products. For more efficient maintenance of the Nutrient Database, entries are limited to foods in the form in which they are consumed (e.g. entries include only cooked, not raw potatoes), and foods with similar nutrient contents are grouped together into a single database entry. Components of each database entry include a food code, food description, nutrient values per 100g of food, and reference codes for each nutrient value.

The Nutrient Database is used to calculate dietary intakes for numerous research studies and must be maintained to meet the following needs of these studies:

1) *Provide nutrients of interest.*

The Nutrient Database was developed in 1974 to calculate dietary intakes for two major cardiovascular studies. Because these studies were especially interested in intake of total fat, fatty acids and cholesterol, the database reflected that emphasis by including these nutrients along with the other proximates and selected vitamins and minerals. Since that time, requests have come from research studies dealing with hypertension, cancer and diabetes to use the Nutrient Database. To accommodate these studies, additional nutrient fields, such as sodium, dietary fiber and sugars, have been added. Currently the database contains 93 separate nutrients, including 23 individual fatty acids and 18 amino acids.

2) *Provide specificity of foods to obtain differences in nutrients of interest.*

When a new nutrient is added to the database, existing database entries are often split into two or more entries so that differences in the new nutrient of interest can be seen in otherwise similar foods. For example, the database was expanded to include separate entries for canned and frozen vegetables when sodium became a nutrient of interest. When dietary fiber was added to the database, the number of cereal entries increased to account for different levels of dietary fiber in these products, and entries for pasta expanded to white and whole wheat.

3) *Provide an updated database.*

As new or better nutrient values become available, it is important to update the Nutrient Database with the more current data. The primary source of data is the USDA Nutrient Database for Standard Reference. Each NCC food that is based on a USDA entry contains that USDA code number in the field for "reference source." Values from a new release of the USDA database can be transferred directly into the NCC database using the reference code as a link. Differences between the old and new USDA values are flagged and verified as correct before the new value becomes a part of the NCC database. Additional data sources used are other USDA publications such as handbooks, provisional tables, and the survey database; scientific literature; foreign food tables; and food manufacturers via the Brand Name Database.

4) *Provide a complete database.*

Because missing values are calculated as zeros in dietary intakes, it is important for each database entry to have a complete nutrient profile. It is better to have a reasonable estimate of a nutrient amount for the food than leave the field blank. This often involves imputing nutrient values. Missing values can be imputed from a similar food, from a different form of the same food (e.g. raw to cooked using retention factors), from a related nutrient, or by developing a formula of the product ingredients that can be used to calculate the missing nutrients. For example, a manufacturer may provide basic nutrient values for a cocoa powder, but no values for sugars, dietary fiber or fatty acids. The ingredients of the cocoa powder are entered into a computer program in the order that they appear on the label. Using food formulary books or a reasonable guess, the amounts of

each ingredient per 100g are entered. Then the nutrients given by the manufacturer are entered so that the total nutrients of the ingredients can be compared with the nutrients given by the manufacturer. Guidelines for an acceptable difference between a calculated nutrient value and that of the manufacturer have been established. If the nutrient comparisons are not close enough, the outlying nutrient is flagged by the computer and the formula must be adjusted until all nutrient differences fall within acceptable limits. When the formula has been determined, the computer calculates all missing nutrient values by totaling the values from the ingredients in the formula. These calculated values are then entered into the Nutrient Database entry. Each nutrient value is accompanied by a source code so that imputed nutrient values can be recognized and replaced with analytic data when they become available.

5) *Provide an accurate database.*
Validation of the Nutrient Database involves several procedures used to locate errors. First, edit limits are used to identify values incorrectly entered into the database. Within the computer program are maximum allowed nutrient values for foods within a particular food group. These limit values cause the computer to flag any nutrient value that appears too large for the type of food that is being entered. The nutritionist must check all flagged nutrients and verify that the values are correct.

The most thorough validation check is a review of a new entry or changes to an existing entry by a second nutritionist. All values for nutrients, serving sizes, and densities are checked against the original source of the data. Calculations and assumptions made about the food are corroborated by this nutritionist.

A third series of checks are run on the internal consistency of the database before a version of the Nutrient Database is released. Calculations are done to compare the total weight of the proximate nutrients to 100g; to compare actual calories with calculated caloric values; to compare weight of individual fatty acids to total fat; amino acids to protein; and fiber, sugars and starch to total carbohydrate. Large differences are flagged by the computer and the nutrient values for that food are verified by a nutritionist. Other nutrients, such as vitamins and minerals, are checked by comparing values of foods within the same food group (e.g. compare riboflavin of all dairy products) and identifying outliers that need to be checked.

Maintenance of the Brand Name Database

The first step in maintaining a brand name database is to contact manufacturers to acquire brand specific nutrient data, a process detailed in an earlier presentation. NCC also refers to printed publications and contacts food retailers to obtain brand name product information. Data are obtained for over 6,000 brand name products annually.

Because more than 12,000 new brand name products are introduced into the market each year, the next step in the process is to determine which of these products need to be entered into the database. NCC has developed criteria to help make these decisions. For example, the Brand Name Database includes only foods whose nutrient content varies significantly from brand to brand. Therefore, different brands of cookies are included in the database, while different brands of canned peaches are not. In addition, only nationally distributed products are included in the database.

The Brand Name Database is designed to a) allow entry of data in the same format as it is supplied to NCC, b) preserve time-related data, and c) provide data validation at the time of data entry. Nutrient values received from manufacturers are often expressed using different units of measure (grams vs. milligrams) and methods of reporting (nutrients per serving vs. nutrients per %USRDA). The Brand Name Database allows entry of all types of values. In addition, multiple values for each nutrient can be stored in the database. For example, both analytical and label values for the same product can be stored, as well as nutrients both per 100g and per serving size. During data entry, the source of the nutrient information can be specified by selecting from over 50 different sources, including product label, calculated, or analytic data. Because the Brand Name Database can store an unlimited number of nutrients for each product, and each nutrient can have an unlimited number of values, a date is entered with each value. This allows use of the most up-to-date nutrient information available, while at the same time preserving an historical record of changes in foods over time. A "reason" code notes whether newer information is due to better data or to a change in product formulation. Currently, only limited data validation occurs at the time of data entry. This includes checking the validity of codes entered for nutrient name, source, and method of measurement. However, in the near future, a more thorough system of quality control will be implemented similar to validation procedures used for the Nutrient Database. In addition, quality control checksspecifically designed to handle the multiple nutrient values possible in the Brand Name Database will be added, as well as duplicate entry of randomly selected products.

Sometimes manufacturers do not provide sufficient data to calculate nutrient values. For example, nutrient content of a candy bar may be reported per bar, but the weight of the bar is not provided. In the case of missing data, NCC contacts the food manufacturer for this additional information or obtains it from the product label. Currently, the Brand Name Database contains over 5,000 brand name products, 127 different nutrients, and 70,000 individual nutrient values. The database size is projected to expand to 6,000 products and 90,000 nutrient values by the end of this year.

Reports can be generated from the Brand Name Database to match brand name products to the most appropriate entry in the Nutrient Database. This matching is based on nutrient composition. For each food type, "key" nutrients are selected for which the food type is a significant source. For example, calcium is a key nutrient for dairy products. Key nutrient values for each brand name product are then compared with the corresponding nutrient values for existing entries in the Nutrient Database. Differences must be within an acceptable range established for each nutrient. If there is no match between a product and a Nutrient Database entry, a new Nutrient Database entry may be created. This new entry is usually based on a product formula developed using the procedure described in section 4 above. These practices permit NCC to accommodate nutrient differences between brand name products without greatly increasing the number of Nutrient Database entries.