Methods of Imputation Used in the USDA National Nutrient Database for Standard Reference

Susan E. Gebhardt and Robin G. Thomas, USDA-ARS Beltsville Human Nutrition Research Center

USDA ARS Beltsville Human Nutrition Research Center
Serving The Nation Since 1906. Improving Health Thru Research.

Introduction

A question the Nutrient Data Laboratory (NDL) is often asked is, "Where do the nutrient values in the USDA National Nutrient Database for Standard Reference (SR) come from?" Data derivation codes were developed by NDL to give the users of SR information about how each nutrient value was obtained. The new Nutrient Data Bank System, initiated in 2001, assigns from 1 to 4 letters as codes for nutrient data compiled for food items. Analytical data are typically not available for all nutrients in all foods. A subset of SR, about 2,800 foods, must have complete nutrient data for the 65 nutrients that are used as the foundation of the Food and Nutrient Database for Dietary Surveys, the basis for USDA's dietary survey component of NHANES. NDL uses standardized procedures to estimate missing values for food items when analytical data are not available (*Haytowitz, et al. 2008, Schakel, et al. 1997*). The objective of this study is to present the predominant methods of imputing used to estimate nutrient values for foods in the current release of USDA National Nutrient Database for Standard Reference (SR20).

Methods and Materials

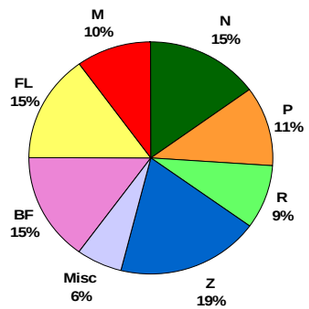
The number and types of data derivation codes that were reported in SR20 were reviewed. The Nutrient Data file was queried to determine how frequently each derivation code was used. The individual codes were grouped into categories and the percentage of each non-"A" category was calculated. "A" category identifies analytical or derived from analytical data.

Results

Of the 300,000 nutrient values in SR with data derivation codes, 200,000 are not "A" codes; number will continue to increase as data for food items are compiled through the new NDBS. Fig. 1 shows use of non-"A" categories.

- The codes most often used in SR20 are Z-19%, BF-15%, FL-15%, and N-15%.
- Codes BF, FL, P, R, and Z are further described in Tables 1-5.
- **M** indicates data submitted by a manufacturer.
- **N** indicates a nutrient value was calculated based on other nutrient values, e.g., carbohydrate by difference and energy (based on other proximates).
- Examples of miscellaneous codes are T, C, and LC:
- **T** indicates data taken from another source, such as NDL Special Interest Databases.
- **C** codes are used when analytical values for a class of foods are averaged to get a value that is used to estimate the content of that nutrient in other foods in that class; e.g., analytical values for choline in 4 species of raw finfish were averaged to give a mean value used for imputing choline in other species.
- **LC** data are based on the nutrient label claim, used for a few processed foods.

Figure 1. Percentage of non-analytical derivation codes used in SR20



BF = Based on similar food	**P = Physical composition**
FL = Formulation based on ingredients	**R = Recipe**
M = Manufacturer's data	**Z = Assumed zero**
N = Based on other nutrients	**Misc = all other non-"A" codes**

Table 1. BF Codes Used in SR20 (Based on another form or similar food)

Frequency*	Derivation Code	Further Defined (corresponds to third letter in code)
30	BFAN**	Adjusted for ash
50	BFCN	Adjusted for carbohydrate
9,800	BFFN, BFFY	Adjusted for fat
2,800	BFNN, BFNY	Adjusted for non-fat solids
3,600	BFPN, BFPY	Adjusted for protein
6,900	BFSN, BFSY	Adjusted for solids
800	BFYN	Adjusted for yield
9,600	BFZN, BFZY	No adjustment

*Rounded; **Last letter N = retention factors not used; Y = retention factors used

BF codes are used when the calculation of the nutrient value is based on another form of that food or a similar food. Specific BF codes are assigned depending on any adjustments that were made (such as based on fat or solids) and if any nutrient retention factors were used in the calculation.

For example, the value for total sugars in unsweetened frozen blueberries (NDB 09054) was imputed from an analytical value for sugar in raw blueberries (NDB 09050) as follows:

(10.0 g sugar raw * 13.4 g solids frozen) / 15.7 g solids raw = 8.5 g sugar frozen

Since the sugar was adjusted for solids and no retention factors were used, the resulting derivation code for that nutrient in NDB 09054 is BFSN. NDB 09050 is referred to as the Reference NDB number (Ref_NDB_No):

NDB_No	Long_Desc	NutrDesc	Nutr_Val	Src_Cd	Deriv_Cd	Ref_NDB_No
09054	Blueberries, frozen, unsweetened	Sugars, total	8.5	4	BFSN	09050

Table 2. FL Codes Used in SR20 (Estimated formulation based on ingredient list)

Frequency*	Derivation Code	Further Defined
17,300	FLA	Based on analytical data
6,700	FLC	Based on label claim values
8,600	FLM	Based on manufacturer's calculated data

*Rounded

FL codes are assigned when formulations are used to estimate nutrient values based on a list of ingredients along with data for a limited number of nutrients in that food (*Haytowitz, et al. 2008, Schakel, et al. 1997*).

Table 3. P Codes Used in SR20 (Physical composition)

Frequency*	Derivation Code	Further Defined
7,200	PAE	Derived from analytical data; estimated physical composition
11,300	PAK	Derived from analytical data; known physical composition
300	PIE	Derived from imputed data; estimated physical composition
4,500	PIK	Derived from imputed data; known physical composition

*Rounded

P codes are used for meat and poultry when data for lean meat and fat and possibly skin are combined in appropriate proportions to create a specific cut of meat.

Table 4. R Codes Used in SR20 (Recipe)

Frequency*	Derivation Code	Further Defined
3,900	RA	Approximate ingredient proportions (e.g. combination of several recipes)
2,200	RC	Cookbook
4,500	RKA, RKI	Known formulation; no adjustments; all analytical or combination
2,700	RP	Per package directions (e.g. cake mix)
6,000	RPA, RPI	Per package directions; no adjustments; all analytical or comb.

*Rounded

R codes are used when adding ingredients such as milk or butter to a packaged mix, when a formulation is known, and for home prepared foods.

Table 5. Examples of Derivation Code Z Used in SR20 (Assumed zero)

Nutrient	Frequency*	Primary Foods
Vitamin B-12	1400	Spices, Fruit, Vegetables, Legumes, Nuts and Seeds, Grains
Cholesterol	1500	Spices, Breakfast Cereals, Grains, Fruit, Vegetables, Legumes, Nuts, Seeds
Alcohol	1500	All foods except foods/beverages containing alcohol
Dietary Fiber	1900	Meat, Fish, Poultry, Eggs
Caffeine	2500	All foods except coffee, tea, chocolate, and foods/beverages with added caffeine
Retinol	2500	Spices, Breakfast Cereals, Beef, Fruits, Vegetables, Legumes, Grains, Beverages
Folic Acid	4400	All foods except fortified grain products
Carotenoids	8500	Dairy, Meat, Poultry, Fish, Fats and Oils

*Rounded

Z code is assigned when the food has an insignificant amount of that nutrient or when none occurs naturally in the food. For example:

- dietary fiber does not occur in meats, so an "assumed zero" Z code is assigned
- cholesterol, not normally found in plant products, is assigned an "assumed zero" Z code for fruits, vegetables, grains, and other plant foods

Discussion and Conclusions

Data derivation codes provide important information concerning the sources of data and methods of imputation used to estimate nutrient values in the USDA National Nutrient Database for Standard Reference. This is particularly useful to other database developers who may have to use imputation for their database applications.

The derivation codes, along with other variables such as Reference NDB Number, are located in the Nutrient Data (Nut_Data) file which can be downloaded along with the other SR files from the Nutrient Data Lab Web site, www.ars.usda.gov/nutrientdata.

References

Haytowitz, D.B., Lemar, L.E., Pehrsson, P.R. 2008. USDA's Nutrient Databank System---A tool for handling data from diverse sources. J. Food Comp. Anal. (Submitted)

Schakel, S.F., Buzzard, I.M., Gebhardt, S.E. 1997. Procedures for estimating nutrient values for food composition databases. J. Food Comp. Anal. 10, 102-114.