INTEGRITY CHECKS FOR NUTRIENT DATA

Suzanne P. Murphy, Ph.D., R.D.

University of California
Berkeley, California

Although the importance of using automated methods to check the integrity of nutrient databases is generally recognized, the process is not implemented by all developers. Developers who carefully check the validity of their databases need to be distinguished from those whose procedures do not include these types of checks. Research results can be seriously biased if diets are analyzed using databases with inaccurate values, yet users usually have no consistent way of determining the quality of nutrient data. Although the use of automated integrity checks certainly does not ensure accurate nutrient data, it is one step in a process that all developers should undertake. At the 11th National Nutrient Databank Conference, I gave a brief report on the procedures we routinely use at U.C., Berkeley, and Marilyn Buzzard, Yvonne Sievert, and Sally Schakel gave a comprehensive overview of their procedures at the Nutrition Coordinating Center at the University of Minnesota.[1,2] In response to many requests from my colleagues, I have undertaken to formalize these procedures, and extend them to additional nutrient databases. This paper describes the preliminary results of the project.

PURPOSE

The primary goal in undertaking this project was to develop automated methods of checking the integrity of nutrient databases. Specifically, a set of edit limits would be developed which could be used by developers in checking nutrient values on their databases. Several desirable attributes of the edit limits were considered:

- They should be flexible so they could be responsive to changes in foods and nutrient values over time;
- They should be generated from the data, rather than being pre-set by subjective criteria;
- They should identify a manageable subset of nutrient values that should be examined for possible re-validation.

A secondary goal was to develop a reporting format that developers and users would find useful in describing the integrity of nutrient databases.

BACKGROUND AND METHODS

USDA's Nutrient Data Base for Individual Food Intake Surveys (SDB), Release 2.1 (1986), was used to generate edit limits. This database was chosen because it is a large, well-validated nutrient database, developed by the federal agency with the responsibility for maintaining national food composition data (HNIS, USDA), and in wide use for nationwide dietary assessment.

Food groups

Food groups were defined using the coding scheme from the SDB. The SDB foods were divided into 12 groups, based primarily on the high-order digit of the food code (1 through 9). Group 9 (sugars, non-alcoholic beverages, and alcoholic beverages) was divided into three groups (groups 9, 10, 11). Mixed dishes from four groups were combined into a twelfth group, based on the high-order 2 digits of the food code.

Nutrient variables

Edit limits per 100 grams of food were developed for the nutrients on the SDB: water, energy, protein, fat, saturated fat, monounsaturated fat, polyunsaturated fat, cholesterol carbohydrate, dietary fiber, alcohol,

vitamins A, E, C, B6 and B12, carotene, thiamin, riboflavin, niacin, folacin, calcium, phosphorus, magnesium, iron, zinc, copper, sodium, and potassium.

Calculated variables

Five calculated variables were developed as follows:

SUM PROX (proximate) = sum of water, protein, fat, carbohydrate, and alcohol. Ash values are not carried on the SDB, so ash was excluded from this calculation.

KCAL DIFF = difference between the energy (kcal) calculated: (4 x protein) + (4 x carbohydrate) + (9 x fat) + (7 x alcohol); and the energy value recorded on the database.

KCAL % DIFF = difference between calculated and recorded energy as a percent of recorded energy: (KCAL DIFF/KCAL)*100.

FAT DIFF = difference between the fat (g) calculated: saturated fat + monounsaturated fat + polyunsaturated fat; and the fat value recorded on the database.

FAT % DIFF = difference between the calculated and recorded fat as a percent of recorded fat: (FAT DIFF/FAT)*100.

Generation of edit limits

The 5245 unique food items on USDA's SDB were used to generate edit limits for nutrient values. For SDB food items with multiple fat codes, the default entry was used; if there were salted and unsalted options, the default entry was used. Univariate statistics were examined for all foods on the database, and for foods divided into 12 food groups. The 1st and 99th percentile for each food group was chosen as a logical edit limit -- approximately 50 foods on this large database should fall below the 1st and above the 99th percentile for each nutrient. Using the minimum and maximum values as edit limits gave very large ranges, which seemed less useful. For smaller data bases of 1000 to 2000 food items, a developer would need to examine only 10 to 20 high and low values for each nutrient. To partially avoid the need to examine foods which were only slightly outside the limit, the upper limit was rounded up and the lower limit was rounded down to the closest integer.

Edit checks

An example of a reporting format for food items in the meat-fish-poultry group is shown in the attached table, using food items from the University of California, Berkeley, Minilist. Similar tables (not shown) were generated for two other nutrient databases (Home and Garden Bulletin No. 72 Data Set and Nutritionist III). All responded well to the integrity checks applied.

SUMMARY

It is hoped that these procedures will be useful to nutrient database developers. The edit limits were successfully used to provide pertinent information about three quite different (in number of foods and number of nutrients) nutrient databases, which indicates that the edit limits are reasonable, and may be used with some confidence by other developers.

When examining nutrient values that fall outside the edit limits, it is important to consider the possible reasons for out-of-range values.

- Approximately 1% of values are expected to fall outside the ranges, since the limits are based on the 1st and 99th percentiles.

- Differences in the specific foods chosen for databases (versus those on the SDB) may result in values that are correctly outside the limits.

- For databases which carry nutrients per serving size, rounding errors for foods with small serving sizes may be magnified into large errors when nutrients per 100 grams are calculated.

- Errors may exist in the SDB, and thus may generate incorrect edit limits.

- Errors may exist in the database being checked.

Note that only one of the five possibilities is that the values on the database being examined are actually in error. Examination of out-of-range values requires the assistance of a person with adequate knowledge of food composition data to distinguish among the possibilities.

This methodology will find nutrient values that are clearly out of the normal range, but does not identify errors of smaller magnitude. Thus, these checks do not ensure integrity, and certainly do not replace careful validation of nutrient values. However, checking for edit limits can be a useful final step when developing nutrient databases.

The development of these edit limits is still in the preliminary stage, and much remains to be done to refine and extend the methodology. Comments and suggestions are welcome.

I would like to acknowledge, and express my appreciation of the help I received from Marilyn Buzzard and Sally Schakel of the Nutrition Coordinating Center at the University of Minnesota, Betty Perloff of the Human Nutrition Information Service, USDA, and Laurie North of N-Squared Computing.

For more information, or to obtain a copy of the edit limits, contact:

Suzanne P. Murphy, Ph.D.
Department of Nutritional Sciences
119 Morgan Hall
University of California
Berkeley, CA 94720

Phone: 415-642-5572
Bitnet: MURPHY8 @ UCBCMSA

REFERENCES

1. Murphy, S.P. "Data integrity -- methods for data validation." In: *Proceedings, 11th National Nutrient Databank Conference*. Athens, GA: Georgia Center for Continuing Education, The University of Georgia, 1986.

2. Buzzard, I.M., Sievert, Y.A., Schakel, S. "Database validation procedures." In: *Proceedings, 11th National Nutrient Databank Conference.* Athens, GA: Georgia Center for Continuing Education, The University of Georgia, 1986.

RANGE CHECKS FOR MEAT/FISH/POULTRY GROUP
UCB MINILIST, 1985 VERSION
(N=34)

Limit	Low Limit	High	%Low	%High	%Missing
Proximate sum (w/o ash)	88	101	0	0	0
Calc - reported kcal	-10	3	0	0	0
Percent kcal diff	-7	2	0	0	0
Calc - rep't fat (g)	-3	0	-1	-	-
Percent fat diff	-28	-3	-	-	-
Water (%)	23	81	6	3	0
Energy (kcal)	93	535	3	6	0
Protein (g)	8	34	3	3	0
Fat (g)	0	46	0	6	0
Sat fat (g)	0	17	0	3	0
Mono fat (g)	0	22	-	-	-
Poly fat (g)	0	7	0	6	0
Cholesterol (mg)	37	628	3	0	0
Carbohydrate (g)	0	16	0	6	0
Dietary fiber (g)	0	1	0	0	0
Alcohol (g)	0	0	-	-	-
Vitamin A (IU)	0	37709	-	-	-
Vitamin A (RE)	0	11313	0	3	0
Carotene (RE)	0	13	-	-	-
Vitamin E (mg ATE)	0	6	0	0	0
Vitamin C (mg)	0	35	0	0	0
Thiamin (mg)	0	1	0	0	0
Riboflavin (mg)	0	4	0	3	0
Niacin (mg)	1	16	3	3	0
Vitamin B6 (mg)	0	2	0	0	0
Folacin (mcg)	1	471	0	0	0
Vitamin B12 (mcg)	0	72	0	3	0
Calcium (mg)	5	183	6	3	0
Phosphorus (mg)	50	481	0	3	0
Magnesium (mg)	8	72	3	3	0
Iron (mg)	0	11	0	0	0
Zinc (mg)	0	28	0	6	0
Copper (mg)	0	8	0	0	0
Sodium (mg)	49	2684	6	6	0
Potassium (mg)	54	575	0	3	0

[1] A dash means nutrient values are not available. Fat difference cannot be calculated since values for monounsaturated fatty acids are not available.