

THE PROBLEMS OF USING SMALL NUTRIENT DATA BASES

Loretta W. Hoover
Professor
Department of Human Nutrition, Foods,
and Food Systems Management
College of Home Economics
University of Missouri-Columbia

Determining what size of nutrient data base to use is a decision which many face when planning to use computers. Reduced nutrient data bases are of two basic types: (1) grouped data consisting of weighted averages for nutrients used to represent all foods in the group and (2) direct substitution where the nutrients for one food are substituted for the nutrients in a somewhat similar food. Pennington (1) has described a procedure for establishing a "mini-list" by using correlation to determine the co-occurrence of seven index nutrients in foods. Investigators have demonstrated that both of these types of small nutrient data bases were functional for their intended purposes.

Major studies such as the USDA Nationwide Food Consumption Survey (NFCS) (2) and the HHS National Health and Nutrition Examination Survey (NHANES) (3) have identified foods commonly eaten by the U.S. population. These foods have served as the basis for smaller data bases.

What distortions in nutrient values can one expect when fewer foods are included in a nutrient data base? In response to that question, the findings from a study involving the use of two reduced nutrient data bases are summarized in the first segment of this paper. The final portion of the paper is focused on my personal opinions and concerns relative to some potential problems associated with the use of small nutrient data bases.

COMPARISON OF LARGE AND SMALL NUTRIENT DATA BASES

With funding from USDA*, a project was conducted at the University of Missouri-Columbia to examine the effect of food item specificity on mean nutrient intake values reported for different sex/age groups using data from the first quarter of the 1977-78 Nationwide Food Consumption Survey of Individuals. The project consisted of four phases: (1) loading the USDA for use on the UMC computer, (2) analysis of food consumption frequencies, (3) developing a cross-reference file for substitutions, and (4) reducing the nutrient data base and testing the results for two substitution levels. The nutrient data base contained values for energy and fourteen nutrients for 100 gram portions of 4,404 food items. The individual intake data processed in the study were three-day dietary records for 7,914 individuals grouped into 22 sex/age groups.

* Funding for this research was provided by the U.S. Department of Agriculture under Research Agreement Number 53-32U4-0-201.

Analysis of Frequency of Consumption

During the second phase of the project, the frequencies of consumption were determined for each food item. Several types of output were generated to list the frequencies in several formats: (1) descending order of frequency, (2) in order as maintained in the nutrient data base and (3) in descending order in each major subgroup. Items eaten several times a day had higher frequencies. The consumption frequencies were displayed according to sex/age group in the third format. The total frequencies for the 4,404 foods were tabulated into three categories: 1371 foods had zero frequency, 2574 foods were consumed 1 to 99 times, and 459 foods had frequencies greater than 100. Thus, over 25% of the food items were not consumed during the first quarter of the survey. The relationship between rank order and frequency was established with a rank of 1 assigned to whole fluid cow's milk with a consumption frequency of 14,142. A reciprocal relationship existed since the consumption was ranked in descending order; a food with a consumption frequency of 88 was assigned a rank of 500.

Nutrient Data Base Reduction

Prior to reduction of the nutrient data base, the food items were analyzed for similarity in nutrient profile. A total of 1044 food items had nutrient values identical to another item in the nutrient data base and differed only in food item description. To reduce the nutrient data base, frequently consumed foods which were unique in nutrient composition and could be representative of a group of similar foods were designated as Retained Items. Food items were designated as Inactive by grouping with a Retained Item assigned as a substitute in a cross-reference file which was created to maintain the substitution assignments. The nutrient data base was not altered. The clusters of Retained and Inactive items were analyzed for similarity in nutrient profile. In Substitution Level 2, the goodness-of-fit between nutrient profiles for the Retained and Inactive items were analyzed using regression analysis. Greatest emphasis was placed on achieving a better fit, for the most frequently consumed foods with a rank greater than 500. R^2 values for 80% of those 499 foods ranged between .9 and 1.0 and the Beta values ranged between .7 and 1.3 for 83% of those foods indicating satisfactory substitution assignments for most of the frequently consumed foods. Although a similar analysis was not made for Substitution Level 1, the fit was assumed to be better since more foods were retained.

Substantial reductions were made in the size of the nutrient data base. In Substitution Level 1, 396 food were retained resulting in a 91% reduction. In Substitution Level 2, a 95.4% reduction was achieved by reducing the nutrient data base to 200 foods. The effective reduction was really somewhat less since some foods were not consumed and some foods had nutrient values identical to another food maintained as a Retained Item. The number of foods retained in the two substitutions levels for each major subgroup are shown in Table 1.

Consequences of Nutrient Data Base Reduction

The nutrient values per sex/age group were computed for the total sample using the original nutrient data base to serve as the baseline for comparison with the values generated using the reduced nutrient data bases. This analysis was also performed to assure that the computer program was accurate. The nutrient values were computed for both substitution levels and were compared with the baseline nutrient values.

Percent differences from the baseline values were computed for each day and the three-day average for each nutrient for each of the 22 sex/age groups. The minimum and maximum absolute percent differences were determined for each nutrient for both substitution levels. The maximum values were larger in Substitution Level 2 and many of the larger values were attributed to the Under 1 age children since very few baby foods were retained. For Substitution Level 1, the maximum values were under 10% for all nutrients except magnesium and Vitamin A. In Substitution Level 2, Vitamin A and Vitamin B12 were associated with percent differences greater than 10% after eliminating the differences attributed to the Under 1 age group.

A two factor analysis of variance model was processed for each nutrient for both substitution levels to statistically analyze the effect of nutrient data base reduction. In Substitution Level 1, F-ratios were significant at the .05 level for fat, carbohydrate, calcium, iron, magnesium, Vitamin A, thiamin, and Vitamin B6. In Substitution Level 2, only Vitamin B6 was not associated with a significant F-ratio for either data base type or interaction between data base type and sex/age group. Thus, the results for Substitution Level 2 were more complex and difficult to interpret. A statistical analysis of the power of the F-test was performed to determine minimum detectable differences in mean nutrient values.

The consequences of the reductions were also analyzed by performing t-tests for each nutrient for each of the 22 sex/age groups. A total of 330 t-tests were performed for each substitution level; 199 were significant at the .05 level in Substitution Level 1 and 207 were significant in Substitution Level 2.

Since nutritional adequacy is often expressed in terms of a standard such as the Recommended Dietary Allowances (RDA), that standard as adapted by USDA staff (2) was used for comparison with the baseline nutrient values and those from Substitution Levels 1 and 2. The comparisons were made by computing percent differences using the three-day average values for 13 nutrients. Except for six instances, the differences all carried the same sign. Using the reduced nutrient data base, the conclusions about nutritional adequacy were essentially the same as those made using the original nutrient data base.

Conclusions and Recommended Research

After considering the details of these analyses, several conclusions seemed appropriate. If a nutrient data base is tailored to consumption practices, the nutrient intake of a large group can be approximated with a smaller data base. However, translating those results into nutritional guidance may be difficult since the specific food selections are not known. Small nutrient data bases are probably inadequate for analyzing nutrient intake for specific subsets of the population and may not be suitable for intake data from other quarters of the NFCS. Small nutrient data bases are not suitable for individual dietary records or data from small groups if numerous substitutions must be made. One of the chief reasons for questioning the use of a small nutrient data base is that specific food consumption practices of a diverse population cannot be monitored effectively. Thus, a sound basis for nutritional guidance may not be feasible using small nutrient data bases which require many substitutions.

More research is needed to determine the effects of reduced nutrient data bases on other quarters of data from the NFCS and to determine the effect on nutrient values for subsets of the population. Also, the suitability of a reduced nutrient data base larger than those tested in this project should be evaluated to determine the value of a mid-size nutrient data base.

LIMITATIONS OF SMALL NUTRIENT DATA BASES

Some individuals use small nutrient data bases for a number of reasons. Small nutrient data bases have fewer options to consider when coding, fewer items to keep up-to-date, require less computer data storage and memory, are not cluttered by many rarely eaten foods, and may be tailored to meet the needs of a specific project. Some regard the types of variations identified in the research project described above as within the range of normal variation for nutrients in foods and are not concerned about the distortion introduced by data analysis. Other developers prefer smaller nutrient data bases since the probability of missing values may be reduced. However, the potential problems with small nutrient data bases should be considered.

Potential Problems of Small Nutrient Data Bases

Potential problems may be associated with using small nutrient data bases. Whether or not these problems exist may depend on the size, contents, and intended use of a small nutrient data base. Some of the problems are itemized below:

1. May lack food item specificity.
2. May not represent foods eaten in various seasons.
3. May not contain foods eaten by sex/age groups, geographic areas, or ethnic populations.
4. May not contain a variety of mixed dish items.
5. May not reflect various food preparation methods.
6. May not reflect intake of some nutrients due to lack of specificity resulting in incorrect estimation.
7. May discourage specificity in data collection if fewer food items are available for coding.
8. May require substitutions for numerous foods.
9. May result in more inconsistency in coding food items and poorer intercoder reliability.
10. May require more judgment in coding food item substitutions and result in frustration for coders.
11. May provide inaccurate diagnostic information if numerous substitutions are required resulting in an imprecise screening tool.

12. May bias results through decisions made while coding.
13. May not have a sound basis for dietary guidance.
14. May not satisfactorily reflect nutrients of concern in future studies.
15. May be limited to use in certain studies since not suitable as a general purpose tool.
16. May not be appropriate for analysis of individual dietary records.
17. May not include forms of food needed for calculation of nutrients for recipes.
18. May need custom software to select and update records when new nutrient data are released.
19. May multiply coding effort by coding several items to represent a single mixed dish item.

Summary

Each professional has the responsibility to select a nutrient data base which is appropriate for intended uses. With the advances being made in computer storage technology, size limitations on nutrient data bases used on microcomputers are being eliminated and larger nutrient data bases can be accommodated.

References

1. Pennington, J. A.: Dietary Nutrient Guide. Westport, Conn: AVI Publishing Company, Inc., 1976.
2. Food and Nutrient Intakes of Individuals in 1 Day of the United States, Spring, 1977. USDA Nationwide Food Consumption Survey, 1977-78. Preliminary Report No. 2. Science and Education Administration, Washington, D.C., 1980.
3. Dietary Intake Findings, United States, 1971-1974. Data from the National Health and Nutrition Examination Survey. DHEW Pub. No. (HRA)77-1647, 1977.

TABLE I

NUMBER OF RETAINED FOOD ITEMS IN NUTRIENT DATA BASES

FOOD GROUP	NUMBER OF FOODS IN ORIGINAL NDB	SUBSTITUTION LEVEL 1	SUBSTITUTION LEVEL 2
Milk and Milk Products	321	36	24
Eggs, Mixtures and Substitutes	51	8	3
Fats, Oils, and Salad Dressings	70	11	5
Sugars, Sweets, and Beverages	392	48	14
Grain Products	956	68	47
Dry Legumes, Nuts, and Seeds	157	13	10
Meat, Poultry, Fish, and Mixtures	1307	107	60
Vegetables	677	67	23
Fruits	473	38	14
Total	4404	396	200
Average Percent Reduction	—	91%	95%