

DATA INTEGRITY - METHODS FOR DATA VALIDATIONS

Suzanne P. Murphy

The UCB Minilist, developed by the Nutritional Sciences Department at the University of California, Berkeley, is a relatively small but extensively used nutrient data base. We take pride in its integrity, and go to some length to validate the new entries every time it is changed.

Updates of existing nutrient values are the most common type of change that we make to the Minilist. We generally try to batch updates to coincide with new releases of USDA's standard reference tape. For a new "version" of the Minilist, a set of change sheets is generated showing the old and new nutrient values side by side, along with the percent difference. These are checked by our staff, and then circulated to all current users of the Minilist. In addition, we calculate old and new mean values for all nutrients, to be certain there are no drastic changes. Finally, we run several "model" diets through the system, and check to be sure the changes in the total are correct.

We can also run several validation programs: one adds the proximate nutrients (protein, fat, carbohydrate, water and ash) and compares the total to 100%; another calculates energy values by multiplying carbohydrate, fat, and protein, by 4, 9, and 4, and compares the result to the value on the data base for energy (alcoholic beverages are done separately); a third program adds the amino acids and compares the sum to the protein value. Any discrepancies that turn up are carefully checked.

Additions of new food items require more extensive validation, but the scale is usually smaller (one or two foods), so the checking can be more intensive. The new values are entered from a worksheet, printed, and sight checked against the original values. In addition, the validation programs mentioned above can be run.

Expansion of the nutrient data base to include new nutrients is by far the most complex of our maintenance tasks. It normally involves changes in our programs, as well as the data base. Obviously, extensive research is involved in order to find the best analytic values, and impute missing values. After each value is entered from a work sheet, it is sight-checked against a printed listing. Then we run model diets and check against hand-calculated totals. Finally, we try to find diets that have been analyzed for the new nutrient, and compare our calculated totals for the same diet.

Overall validation of the system is performed by comparing our results to those reported by other researchers using either analyzed values or different nutrient data bases. Our most recent validation was performed using over 1000 NHANES II diets of young women. The resulting nutrient totals were compared to those reported by NCHS for nutrients in common, and to literature values for the remaining nutrients. Discrepancies were noted, and their sources investigated.

In summary, we believe data validation procedures should be an integral part of every diet analysis system. They are often time-consuming, but valid research results often depend on an accurate and unbiased system. Data base users should always question developers about their methods of ensuring integrity.