THE HIGHS AND LOWS OF BUILDING A DATA BASE

Judith E. Brown, R.D., MPH, Ph.D.
Director and Associate Professor
Program in Public Health Nutrition
School of Public Health
University of Minnesota
Minneapolis, MN 55455
and
President, Nutrition Support Headquarters, Inc.

I am here today representing one of many who have taken the road from nutrient data base ignorance to enlightenment. One who has traveled the road that leads us through the highs and lows of building, maintaining, and implementing a computerized, data base baby. As I have traveled this road and learned where the potholes are, I have taken notes. Today I will give you a book report on the notebook's contents in the hopes of saving those who will take a similar road time and trouble.

PASSAGES IN THE LIFE OF A DATA BASE BUILDER

Building a nutrient data base has its own set of passages. I was introduced to nutrient composition in my freshman year in college. Remember the exercise of recording your diet, looking the foods up, painfully converting household units to grams, and so on? I had no doubt that the results were perfect. I ate 12.3 mg iron each day. Then I became a teacher. I gave that assignment. When reports came in with footnote numbers into the calculations, I began to have my doubts about how accurate these calculations were. So, we went to computerized nutrient analysis, the type that required coding of foods and converting amounts into grams, for better results.

This computer solution failed to take us to nutrient nirvana. Coding and converting measures to grams was a terrific hassle. Once foods were coded, the code was frequently not recognized (because it wasn't included) by the computer program. How many grams are in a half cup of grapes, anyway? Why didn't the program include zinc? And why did we have to make so many substitutions, like entering lasagna because ravioli wasn't in the data base? (Students still thought the results appearing on the "hard copy" were perfect. I was pulling my hair out.)

The Dietary Analysis and Assessment System (nicknamed DAS by users of the system) we were compelled to develop, was built out of total frustration with existing systems. Something had to be done about the coding and converting hassles--these were the two main culprits that took the joy out of nutrient analysis. Why not develop a micro-computer program that would accept foods by their common names and everyday, household measuring units? Why not include those nutrients, for which nutrient composition data exists, that represent important nutrition/health relationships? That defined my dream for a nutrient analysis system.

MAKING THE DREAM COME TRUE

With massive amounts of help from a scientific programmer and hardware specialist, and a substantial financial investment, the dream did come true. During the process, it was up to me to assemble and maintain the data base, and up to the others to get it to work according to the plan.

My life has changed since that day in 1979 when I started to systematically collect food composition data. I went from being the mother of two children to being the mother of two children plus one nutrient data base. All three require extensive nurturing and care, and provide learning experiences. I have learned that there is no data base that is without multiple errors, that contains complete data for all of the nutrients you want, and that includes all of the foods that people eat.

LEARNING TO LIVE WITH ERRORS

> "Honey, if you can't live with mistakes, don't get into the nutrient data base business."
>
> (Dr. E.M. Widdowson, 1983)

That there are multiple errors in data bases comes as no surprise when you consider all of the opportunities for errors. In a data base, say of 1400 foods that includes 35 nutrients per food, there are about 196,000 digits, or 196,000 chances for making errors. Humans enter each of the 196,000 digits, and even with multiple checks, it s difficult to guarantee that each entry is correct.

OTHER SOURCES OF ERRORS

In addition to data base entry errors, bloopers creep into data bases principally in two other ways.

I. HUMAN

- A. Calculating Edible Portion
- B. Calculating/Given Amount
- C. Converting Measurement Units
- D. Recording, Entering, Cross-Checking (Eye-Crossing)
- E. Dietary Data Entry

II. METAPHYSICAL (Sic)

- A. Glitches

When researching a data base, you are sometimes presented two different values for the same food. Pick the wrong one, and you have unwittingly entered an error. Here are three sets of examples of the kinds of choices you face when data differs from reference to reference:

Lowfat Yogurt w/Nonfat Milk Solids, 1c. kcal = 143
Yogurt From Whole Milk, 1c. kcal = 141
Parsley, Raw, Chopped, 1 T = 3.5 g
Parsley, Raw, Chopped, 1 T = 10 g
Fryers, Cooked, Fried, Breast, w/o Ribs, 94 g, KCal = 160
Roasters, Roasted, Light Meat w/o Skin, 94 g, KCal = 171

"WHAT? REALLY? THAT MUST BE WRONG . . ."

Some bloopers can be identified by the outrageous nutrient values assigned. The ones I have kept notes on include:

Rice Krispies, 1c Phosphorous = 0.39
Beef Stroganoff, Frozen, 200 g, kcal = 86
Hollandaise Sauce, 1/4 c, kcal = 180
Round Steak, Bottom, Lean Only, Broiled, 114 g, Riboflavin = 376 mg
Sirloin Steak, Lean, 125 g, Thiamin = 125 mg
Beef Enchiladas, Homemade, 200 g, Vitamin A = 6000 IU

NOTING WHAT'S SAID IN FOOTNOTES

Some of the errors that find their way into data bases can be traced back to their source--the footnotes that are tucked away in small print at the bottom of the pages of nutrient composition tables. Here are three examples of getting tripped by footnotes:

[2]Source of data does not indicate whether raw or cooked, assumed values are for raw food.
[138]Based on total contents of can. If bones are discarded, value will be greatly reduced.
[175]Values range from 60mg to 1,000mg per 100g

USDA MEETS MANUFACTURER'S DATA

Another component of the nutrient data base business is the collection of manufacturer's data. Manufacturer's data are particularly useful because they represent nutrient composition information for foods and brand names that often cannot be found elsewhere. But how does manufacturer's data compare with USDA data when both sets of information are available? My assumption that they would be comparable fell short of the truth:

USDA and Manufacturer's Data
Ex: Cranberry Juice Cocktail, 6 fl. oz.

	Ocean Spray (1978)	USDA #456 (1975)	% Difference USDA/Ocean Spray
Calories	105.7	124	15
Protein	.19 g	.2 g	0
Fat	.19 g	.2 g	0
CHO	26.4 g	31.4 g	16
Calcium	6 mg	10 g	40
Phosphorus	2.3 mg	6 mg	62
Potassium	34.4 mg	19 mg	81
Sodium	3.0 mg	2 mg	50
Iron	.3 mg	.6 mg	50
Thiamin	.02 mg	.02 mg	0
Riboflavin	.03 mg	.02	50
Niacin	.09 mg	.1 mg	0
Vitamin C	60 mg	30 mg	100

Range (0-100%)
= 36% Difference USDA/Ocean Spray

MISSING VALUES ARE A SOURCE OF ERROR

The final source of error that needs to be mentioned is that of missing values. Missing values are to nutrient data bases what the boll weevil is to cotton farmers. Many data base developers boast of the large number of food items and nutrients included in their data bases. What isn't mentioned is the percent of missing values that automatically come with these high figures:

Missing Values*

	Total Foods n	Sample n	No. Food Components	% Missing
USDA #8 (1963)	2483	550	14	13
Bowes & Church's (1980)	4768	450	26	49

*Based on systematic sampling by page number

The question of the extent to which accuracy is lost when a data base is replete with missing values has not been studied. One can only suspect that significant underestimations for particular nutrients results.

Missing data is a particular nemesis when a nutrient/health risk is suspected but data on the particular nutrient is lacking. Data on beta-carotene and simple sugars composition of foods are very incomplete; making it difficult to study the relationships between the nutrients and the incidence of disease. For now, besides undertaking the needed food analyses, we have to live with the data we have. Or, as Margaret Moore so aptly described the situation with simple sugars data:

> "Well, you know, you can't make a silk purse out of a sow's ear".
>
> (personal communication, 1982)

REALITIES OF NUTRIENT DATA BASE IMPLEMENTATION

The final chapter in my notebook was a cause of consternation. Now that my sense of humor has been revived, it is a cause for chuckles. The chapter notes experiences in the real life implementation of DAS. My concerns about the accuracy and completeness of DAS were overshadowed by the realities of dietary data collection. While we wring our hands over the difference in thiamin composition of french cut vs. uncut green beans, there is a health professional out there submitting food records to us on behalf of patients.with entries such as the below:

Food Record Quotes

Food/Description	Amount
Turkey Salad	Large paper plate full
Potato Chips	1 handful
Peanut Butter Bar	1 (stole 2 or 3 off food cart)
Bologna Sandwich	1 bite
Hamburger Casserole (Hamburger, noodles, tomatoes, garlic, cheddar cheese)	1 C.
Brownie	1 - 2" x 1"
Fruit	4 oz.
Meat	2 T.
Cutter's Insect Repellent	2 sprays

Submitters of food records are not without a sense of humor. One particularly poor food record was submitted along with an attached note from the nutritionist that read:

> "I think she'd take in more nutrients if she ate the printout."
>
> Barosso, G. (personal communication, 1983)

NUTRIENT BLOOPER SNOOPERS

Two major problem areas have been highlighted: the errors and missing values in nutrient data bases and the inaccuracies in dietary data collection. The latter is a problem for methodologists and implementors of dietary data collection. The former will become less of a problem as new food analyses are performed. With 200 new food products being offered each month, and other products becoming history, maintaining data bases will forever be with us. The error creep in data bases might be alleviated, in part, if users of nutrient data bases cooperated in the identification of errors. Thus the Nutrient Blooper Snoopers are born. Rewards of, say, one dollar per identified error could be offered. T-Shirts proclaiming "I'm a Nutrient Blooper Snooper" could be presented to those who found the culprits. The point is that, because consumers of nutrient data generally hold the utmost trust in the analysis results, it is incumbent upon us to make data bases as accurate as is humanly possible.