NUTRIENT CALCULATION SYSTEMS FOR EDUCATORS AND CLINICIANS
II. EVALUATING DATA BASES FOR MICROCOMPUTER PROGRAMS

Phyllis J. Stumbo

Clinical Research Center
University of Iowa
Iowa City, Iowa

The computer has transformed nutritional data into a consumer commodity. Nutritional data graces restaurant menus, cookbooks, and newspaper and magazine articles. The computer makes detailed nutrient data immediately available and educators and clinicians are increasingly expecting more than speed from the computer -- they are demanding more convenience and easier-to-use programs.

During the early years nutritionists needed relatively few entries to describe the food supply. For example, two entries - chocolate or vanilla pudding - described all pudding. Today nutritionists need to account for differences in cooked, canned, refrigerated, frozen, and instant products, some of which can be made with whole, skim, 2% or 1% fat milk. The quest for convenience while accounting for the wide range of food products available today has led to a demand for brand-name specific data base entries.

The ever popular cookie illustrates why brand-specific data is sought. A recent addition to my grocery store was a cookie called "Graham Bites". This product is a very small graham cracker-like product. Imagine, if you will, a client who ate 44 of these tiny cookies and, considering the list of cookies from USDA's Continuing Survey of Individual Intakes data base shown in Table 1, select an item to describe 44 Graham Bites. I asked a colleague to choose something from this list that she thought approximated the amount and nutrient composition of 44 Graham Bites and she selected 5 graham crackers. Actually four graham crackers and one sugar cookie better characterize the higher fat content of the Graham Bites, but making this choice required examining weight and nutrient values of both products, information not usually at the nutritionist's fingertips when evaluating intake records.

Figure 1 shows how cookie sizes and composition vary. The Nabisco company produces many cookies including at least 22 of the sandwich-type. Composition of most sandwich cookies is similar, but a few popular ones differ considerably. The 10 gram "USDA" cookie in Figure 1 provides 50 Calories and is a standard for comparison. This standard corresponds to the nutrient composition of Nabisco Oreo cookie as well as many other sandwich cookies. However, choosing this cookie for all sandwich cookies causes considerable error. If a client eats 5 Lance cookies that are made to dispense from a vending machine he consumes 65 fewer Calories than represented by 5 standard cookies. If, on the other hand, the client actually eats 5 Burry cookies he consumes 215 MORE Calories than 5 of the standard product. Figure 1 illustrates differences in these cookies.

If this sounds like a lot of detail about a trivial product, consider that cookies were the fourth most frequently consumed food by children 6 to 11 years of age in NHANES II (1). Cookies provided over 3% of these children's overall Calorie and fat intake, and over 5% of their polyunsaturated fat intake. In a population of 8 to 10 year old Iowa children cookies provided 10% of their saturated fat, and 13.5% of their polyunsaturated fat intake. (2) It is clear that mistakes in identifying cookie intake could be a major source of error for some children's intake.

These examples illustrate the greater precision available when nutrient data bases provide brand-specific information, but my real purpose here is to alert this group to the menace lurking within many data bases that offer brand-specific information. I will use data from seven nutrient analysis programs for the microcomputer to illustrate the way information is organized on data bases. Objectives for this paper are to: 1.) Compare generalized and brand-specific nutrient profiles, 2.) Demonstrate limitations of industry-provided data, and 3.) Suggest safeguards for ensuring data integrity.

Background

The first written report I am aware of that showed differences between nutrient data bases was a comparison by Loretta Hoover that was conducted by participants of this conference in 1976 (3). Figure 2 compares evaluations of a standard menu using 11 data bases which resulted in a difference of 571 Calories between the lowest and highest values. Differences in serving size were assumed to be part of the problem (foods were listed as an "average apple" or "a small potato", for example) so gram weights were assigned to each food and a second analysis conducted which reduced the discrepancy to 456 Calories. A third analysis conducted with even more detailed information such as describing peach in gelatin salad as "canned solids with liquid in heavy syrup" reduced the difference between lowest and highest values to 357 Calories. Participants researched their data bases and found that ten were based on 1972 USDA data tapes, and one used 1963 tapes, so different data sources could account for some discrepancies. Most developers had also added nutrient values from other sources including industry generated data, so the precise causes of discrepancies were not identified. The coder who selects inappropriate foods can also introduce considerable variation, and this was not investigated.

Shanklin authored one of the few published reports that identified sources of nutrient value discrepancies between computerized data bases (4). Differences in the NHANES II data base and the Nutrient Dietary Data Analysis (NDDA) system developed at the University of Illinois using the Ohio State University nutrient data base as the primary data source are shown in Figure 3. The greatest errors reported were observed in infant products where potassium data was missing from the NHANES II data tape for Enfamil With Iron. Thiamin in Similac With Iron differed by a magnitude of 10 with the NHANES value being the most plausible. This disparity could have occurred from misplacing a decimal point. The 2-fold differences in carbohydrate Calories is consistent with choosing the concentrate rather than the ready-to-serve product on the NDDA. Disparities in the adult diet were traced to a chicken entry on the NDDA which was missing several nutrients. In this study both missing values and incorrect values contributed errors.

In offering guidelines for selecting a dietary analysis system Gail Frank wrote, "a subtle but important point to consider about the nutrient file is the extent of missing values (5)." I would argue that this is not a subtle matter. I have evaluated seven microcomputer programs for extent of missing values and find the problem a major source of error. In the remainder of this paper I will illustrate problems encountered when data bases are missing nutrient data and recommend safeguards to ensure data integrity.

Comparison of Generalized and Brand-Specific Nutrient Values

As people use convenience products and eat away from home more often, they are less able to provide needed detail about their meals since they neither purchase the ingredients nor prepare the product. For example, people generally do not know the type of fat used in recipes in restaurants, whether margarines are 50 or 80% fat, whether sauces are made with milk or cream, or the fat content of ground beef used to make chili, to list a few specifics. So-called "fast food" restaurants have a strong share of the dining-out market. Since their products are highly standardized, it simplifies interviews to document intake of foods from these restaurants. Most nutrient software packages provide some information about "fast foods". McDonald's restaurant is clearly the leader in this market, and most food programs either list some McDonald's products, or provide representative fast food products. Table 2 shows the USDA Dietary Analysis Program's (DAP) list of fast food sandwiches which are representative of most fast food restaurant sandwich menus.

A number of programs offer thousands of foods and 50 or more nutrients. I contend this is an illusion. The illusion is that very few brand specific entries have values for nutrients beyond calories, the macronutrients, and a few micronutrients required for nutritional labelling. In order to compare the ability of programs to represent convenience and fast foods I listed a variety of products and examined nutrient values on 7 nutrient calculation programs. The program names are given in Table 3. This comparison showed missing values to be a serious problem. To illustrate this problem I chose a McDonald's meal composed of a Big Mac, french fries, vanilla milkshake, and fried apple pie for comparison since these were the most universally available products on programs I examined. Since DAP uses USDA values I will show data for a comparable meal from this program for comparison. The DAP meal consisted of a double decker deluxe hamburger, regular order of french fries, a 12 ounce vanilla shake, and a fast food fried fruit pie. Only Nutripractor 4000 (NP4000) gave McDonald's values for all four meal components. Nutritionist III (N3) provided the Big Mac and I chose representative USDA food entries for the other 3 meal components.

All programs I reviewed provided Calories and grams of carbohydrate, protein, and fat for these foods. Programs agreed substantially on these nutrients. Figure 4 shows the Caloric value for the McDonald's meal as reported by each program. Serious problems were encountered with other nutrients, however. For example, 5 of the 6 programs offered copper values (DINE does not provide copper information). Four of 5 programs agree within 25%. As Figure 5 shows, one program, NP4000 offers copper, but does not provide copper data for foods in this meal.

Another problem arises when fat components in this meal are considered. Total fat is in good agreement among programs as shown in Figure 6, but values for fatty acids are incomplete, and total fatty acid values are misrepresented by two programs because zero quantities are assumed where data is missing. (See Figures 7, 8, and 9.) DINE does not offer grams of total fat, but this information can be calculated from the percent of calories provided by fat. NP4000 gives no fatty acids at all for foods in this meal, and N3 lacks fatty acid values for the Big Mac, which was the only part of the meal assigned values provided by the manufacturer - - USDA values represented the remainder of the meal for Nutritionist III's calculation. Figure 7 shows moderate agreement for saturated fat for the McDonald's Meal among four programs. Greater differences are observed for monounsaturated and polyunsaturated fats as seen in Figures 8 and 9. Nutritionist III would be more in agreement if oleic acid was selected instead of monounsaturated fat, and linoleic acid was selected rather than polyunsaturated fat, since the USDA apple pie entry contained nutrient values for these specific fatty acids, but not for the two larger classes of fatty acids. The user would have to be very well acquainted with the data base to avoid these kinds of discrepancies.

Unfortunately, the extent of missing values is often overlooked by the clinician. If this McDonald's meal were the evening meal and lunch was a ham sandwich, tossed salad, apple, and milk, and breakfast was Cheerios, milk, and orange juice, and you used the total daily values provided by each program, the P/S ratio would vary from 0.4 to 1.3 for the entire day (see Table 4). Lamentably, summations may be the only values examined by the clinician. The high P/S ratio of 1.3 resulted from having no fatty acid values contributed by the McDonald's meal on the NP4000 program when it actually contained around 70 grams of fat.

Limitation of Industry-Provided Data

I attempted to compare nutrient values for frozen meals, but was unable to obtain comparable products from the programs evaluated. Data from Computrition illustrates the problem of missing values in industry generated data for these products. Figure 10 depicts thiamin content of Lean Cuisine and Weight Watcher's spaghetti compared to the USDA spaghetti value and the Weight Watcher's pizza compared to the USDA pizza value. It is not possible to determine right or wrong especially since thiamin in both products are probably added as enrichment. However, Figure 11 compares folacin values for these two products, and we find industry has not provided folacin information. These examples illustrate limitations of data bases when nutrient data for processed food is simply not available.

Pennington demonstrated the problem posed by missing values using the USDA standard reference tape to evaluate the Total Daily Diet (TDS) Study Foods (6). Figure 12 shows that chemical analysis of TDS foods was greater than calculations for copper and manganese. However, when appropriate values were estimated for foods which lacked values for these minerals on the standard reference tape, the calculated and analyzed values agreed closely. This illustrates the importance of using a complete data base.

Proposed Safeguards for Ensuring Data Integrity

We have a dilemma, especially in reference to the use of industry-provided nutrient data. The industry data is helpful in identifying serving size and the amount of total fat, but if fatty acids and micronutrients are missing, then summations are incorrect. Programs offering numerous nutrients but delivering a handful of empty cells are giving not just less information, but misinformation. If the consumer insists on brand-specific information on nutrient data bases there are several options that should be considered. Three options are described below.

OPTION 1 - Generalized Food Profiles: A set of general food profiles can represent many specific food products. USDA uses this principle by providing food profiles for restaurant foods and mixed dishes which are representative of a wide variety of products (see Table 2). Effectively using this list requires familiarity with food nomenclature and composition. The University of Minnesota Nutrition Coordinating Center has the most elaborate procedures of any program for imputing representative values for nutrients they offer on NDS. Developers of this program estimate they can represent 16,000 foods with less than 2000 nutrient profiles (7,8). For example, a few years ago I examined their data base and found that approximately 400 different cookie brands were coded with the nine cookie codes shown in Table 5. Approximately 550 cookies are currently represented with 25 or so cookie profiles, and a new revision has just been released which will expand this list. An added bonus to this program is that representative profiles are automatically selected by the program when a specific brand name is specified so the user is not required to be familiar with all food products on the grocery shelf. I would be remiss if I did not mention the tremendous cost of providing this detailed brand-specific information in as complete a manner as presented by the NDS program. Table 3 shows the cost of programs described herein. If brand-specific information is sought, the cost is high and the buyer should be aware of the constant need for updating the data base. Eight thousand new items were added to the grocery selves in 1988 alone (8). Because the variety of available processed foods is constantly changing, brand-specific nutrient data bases require frequent updating if they are to remain useful. Using generalized food profiles to represent specific brand name products is an efficient way to have complete nutrient data bases.

OPTION 2 - Bypassing missing values: A second option is a programming routine that would prevent the use of nutrient fields that are missing data. The program could suppress sums for any nutrient if foods were selected that had data missing for that nutrient. Or nutrients could be chosen first, and the program only offer foods which had complete values for the selected nutrients. This would probably frustrate users, but it would safeguard the integrity of data generated.

OPTION 3 - Data Base Certification: The third option is a certification process for microcomputer data bases to help users select appropriate software. The certification process might be conducted by an impartial committee with approved data bases certified as complete and their values judged representative. Such a certification process might involve a fee paid by software developers desiring certification. An expert would examine the data base using the criteria of completeness and representativeness. Certified programs would carry a document instructing users to report questionable values they encounter to the certification board, and the certification would be renewed annually provided user observations had been reviewed and problems corrected.

Two speakers at the 14th National Nutrient Databank Conference predicted that we should not expect complete brand-name specific nutrient data in the foreseeable future, at least not during the 1990's (9, 10). Therefore, we must learn to use the tools we have now in a responsible manner, so our nutrition information does not become misinformation. Educators and clinicians should evaluate their data base and know it's limitations to improve their ability to correctly interpret nutrient data and translate that nutrient information accurately for their clients.

References

1. C Dresser, and M Carroll: Major contributing sources of macro and micronutrients for persons ages 1-74 years by age and race: United States, 1971-80. National Center for Health Statistics, Hyattsville, MD. Sept. 1986. Edited unpublished tables.

2. DISC study data, personal communication, Phyllis Stumbo, 1990.

3. LW Hoover. Computerized nutrient data bases: I. Comparison of nutrient analysis systems. J Am Dietet Assoc 82:501-505, 1983.

4. D Shanklin, JM Endres, and M Sawicki. A comparative study of two nutrient data bases. J Am Dietet Assoc 85:308-313, 1985.

5. GC Frank and S Pelican. Guidelines for selecting a dietary analysis system. J Am Dietet Assoc 86:72-75, 1986.

6. JAT Pennington and DB Wilson. Daily intakes of nine nutritional elements: Analyzed vs. calculated values. J Am Dietet Assoc 90:375-381, 1990.

7. FNIS, Microcomputer software collection, National Agriculture Library, Mimeo, June 1990.

8. SF Schakel, RA Warren, and IM Buzzard. Imputing nutrient values from manufacturers' data. Proceedings of the Fourteenth National Nutrient Databank Conference, 155-165, 1989.

9. SS Harris, The government's experience in data development. Proceedings of the Fourteenth National Nutrient Databank Conference, 3-9, 1989.

10. HD Hurt, Nutrient databanks: The role of the food industry, Proceedings of the Fourteenth National Nutrient Databank Conference, 11-15, 1989.

Table 1. Cookie choices available with USDA-CSFII:

	Wt, g		Wt, g
1. Assorted, commercial	8.6	10. Maccaroons	19.0
2. Butter, thin rich	5.0	11. Marshmallow	17.0
3. Chocolate chip, commer	10.0	12. Molasses	32.0
4. Chocolate chip, home	10.0	13. Oatmeal w/ rsn	13.0
5. Coconut bars	9.0	14. Peanut	21.8
6. Fig bars	14.0	15. Raisin	17.8
7. Gingersnaps	7.0	16. Van wafers	4.0
8. Granola cookie	13.0	17. Graham cracker	7.0
9. Lady Finger	28.0		

Table 2. "Fast Food Restaurant" Food Profiles
USDA Dietary Analysis Program (DAP)

Hamburger, plain, on bun	Steak and cheese submarine
Hamburger, 1/4 lb., plain, bun	Meatball submarine
Hamburger, double, plain, on bun	Pizzaburger
Cheeseburger, plain, on bun	Ham and cheese sandwich
Cheeseburger, 1/4 lb., plain, bun	Bacon, lettuce, and tomato
Cheeseburger, double, plain, bun	Club sandwich
Hamburger, double decker deluxe	Cold cut submarine
Cheeseburger, bacon, on bun	Frankfurter, plain, on bun
Beef, roast, sandwich	Chicken fillet sandwich
Beef barbecue on bun	Fish Sandwich
Gyro	

Table 3. Microcomputer programs compared:

Cost	
$7,000	NDS - Nutrient Data System, University of Minnesota
1,250	NP4000 -Nutripractor 4000, Practorcare Corp.
695	Computrition - Computrition, Inc.
495	N-3 - Nutritionist III, N-Squared Computing
245	DINE - DINE Windows, DineSystems
295	FP2 - Food Processor - 2, ESHA Corporation
60	DAP - Dietary Analysis Program, USDA

Table 4. Sample Menu and P:S Ratio

BREAKFAST:	Cheerios 2% milk Orange juice	P:S RATIO:
		FP2 - 0.2
		N-3 - 0.3
LUNCH:	Ham sandwich Fresh Apple Cola	DINE - 0.3 NCC - 0.5 DAP - 0.6 NP4000 0.8
DINNER:	McDonald's meal	

Table 5. University of Minnesota Cookie Profiles

Nine Cookie Profiles:	----Per 100 gm:---- Fat, g	Sodium, Mg
High fat, High sodium	23.1	295
High fat, Med sodium	23.1	185
High fat, Low sodium	23.1	125
Medium fat, High sodium	16.1	295
Medium fat, Med sodium	16.1	185
Medium fat, Low sodium	16.1	125
Low fat, High sodium	10.4	295
Low fat, Med sodium	10.4	185
Low fat, Low sodium	10.4	125

Figure 1.

Size and caloric value of selected sandwich cookies

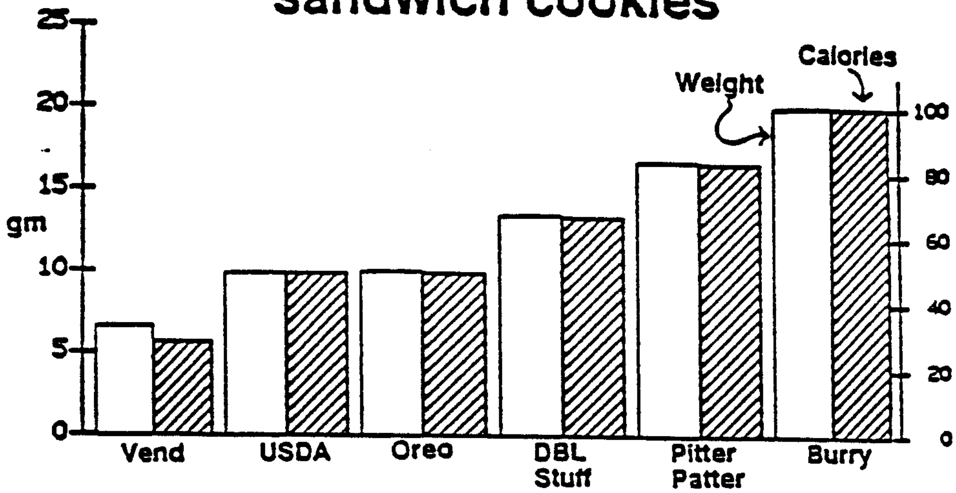
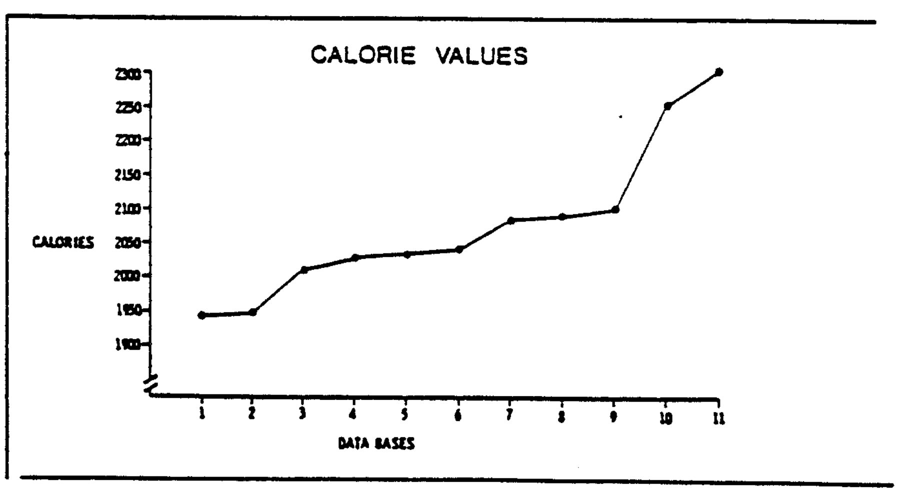


Figure 2.



Hoover, JADA, May 1983

Figure 3

NHANES II and NDDA Systems' Databases Compared

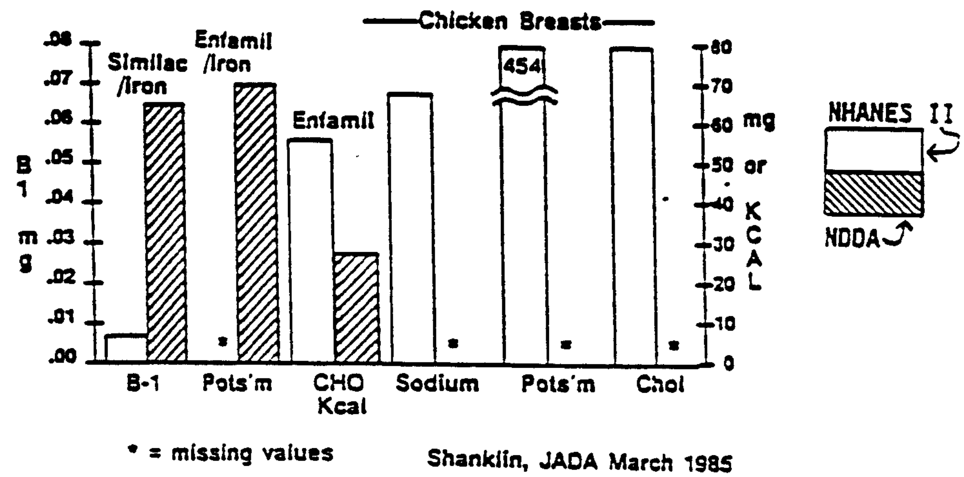


Figure 4.

Calorie content of McDonald's Meal from 6 databases.

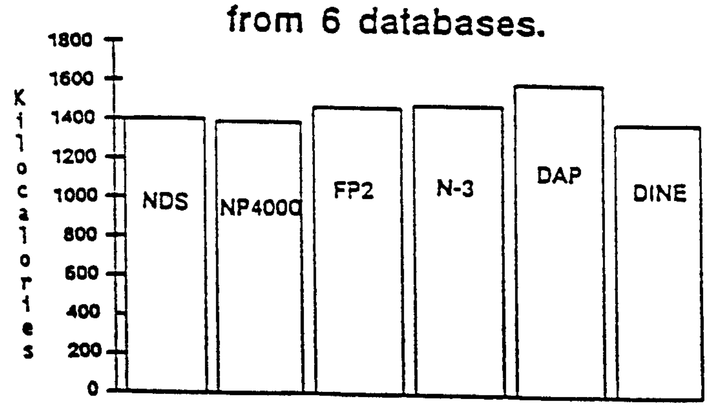
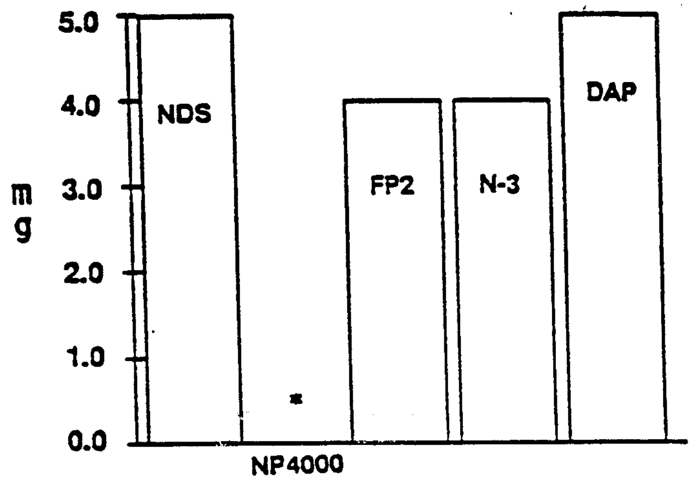


Figure 5.

Copper content of McDonald's Meal from 5 databases.



* = missing values

Figure 6.

Total fat content of McDonald's Meal from 5 databases

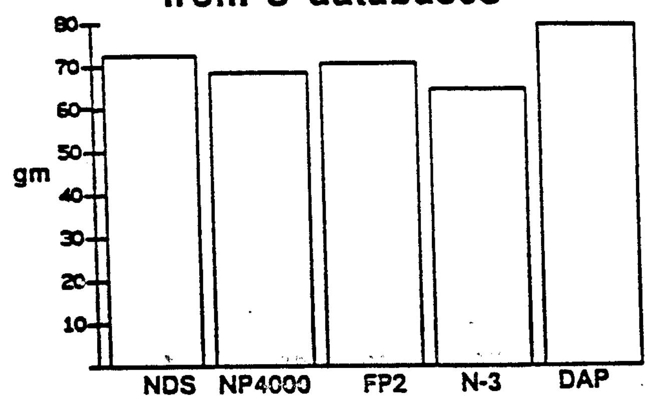
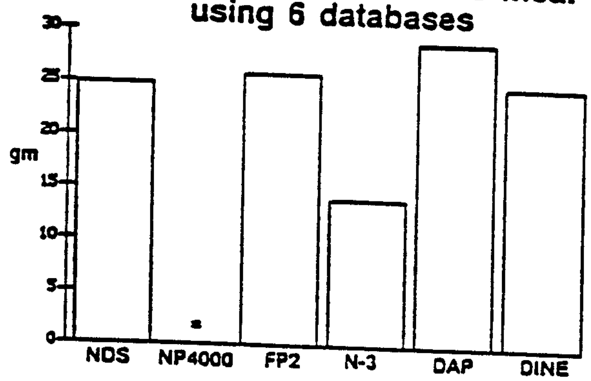


Figure 7.

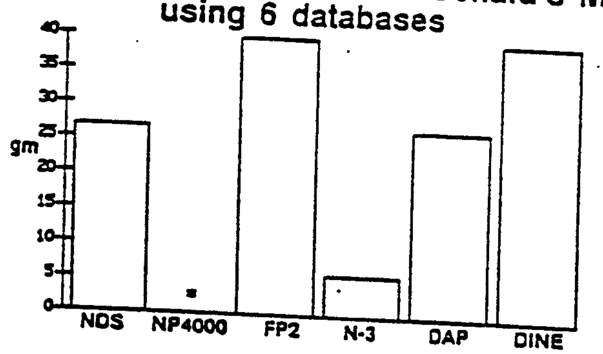
Saturated fat in McDonald's Meal using 6 databases



* = missing data

Figure 8.

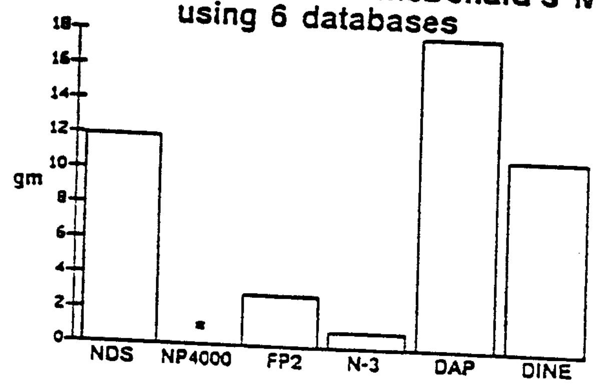
Monounsaturated fat in McDonald's Meal using 6 databases



* = missing data

Figure 9.

Polyunsaturated fat in McDonald's Mea[l] using 6 databases



* = missing data

Figure 10.

Thiamin content of selected frozen meal entries from Computrition

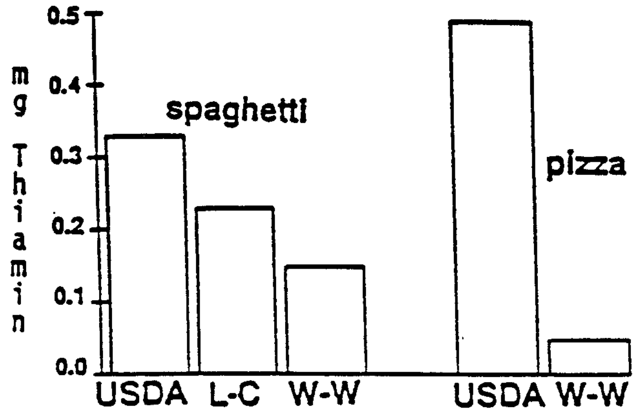
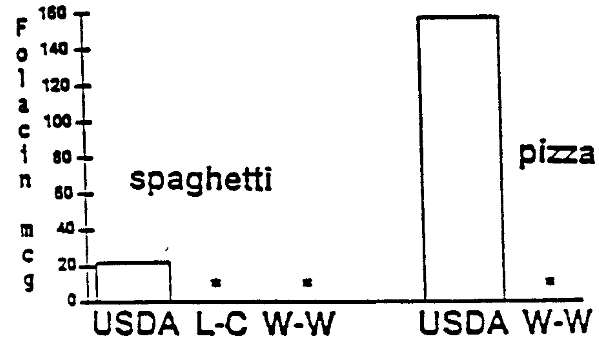


Figure 11.

Folacin content of selected frozen meal entries from Computrition

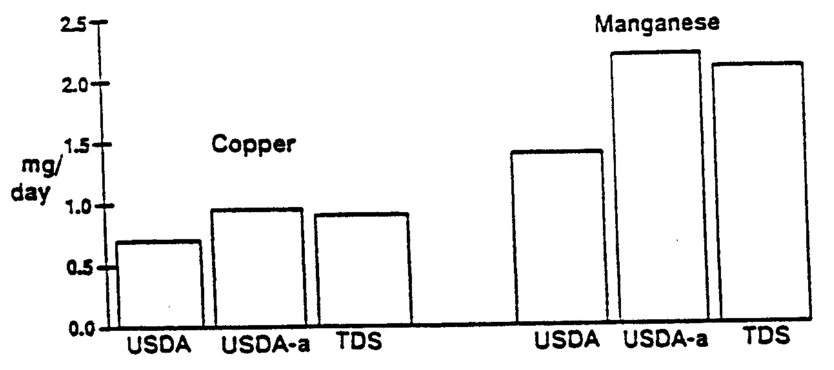


* = missing values

L-C = Lean Cuisine
W-W = Weight Watchers

Figure 12.

Missing elements for Total Diet Study Foods using USDA Standard reference - diets for women age 25-30.



USDA-a: adjusted to eliminate missing values.

Pennington, JADA March 1990