

# Developing and Maintaining a Nutrient Data Base for Food Frequency Questionnaires

by Helaine R. H. Rockett, MS RD FADA, and Lisa B. Litin, RD CNSD

## Food Frequency Questionnaire

The increased interest in nutritional etiology of disease has prompted development and evaluation of techniques to measure dietary intake (1). The food frequency questionnaire is an efficient, accurate method to assess individuals' diets in small and large studies (2,3). Food frequency questionnaires (ffq) are designed to measure average long-term diet rather than to provide a precise estimator of short-term intake such as that provided via a 24-hour dietary recall or multiple day dietary records (1). The epidemiological objective of the food frequency questionnaire is to categorize individuals by their nutrient intake (2).

Over the past 15+ years, investigators at Channing Laboratory, Harvard Medical School and Harvard School of Public Health have been developing and refining a semi-quantitative food frequency questionnaire used extensively in nutritional research studies. The earliest version of this food frequency questionnaire (ffq) consisted of a 61-item questionnaire which was shown to provide a reasonable measure of dietary intake among 200 registered nurses in the Boston area when compared with 4 one-week dietary records (1). This initial ffq has undergone constant evaluation and continues to be used in the ongoing Nurses' Health Study of 200,000 women and the Health Professionals' Follow-Up Study of 50,000 men. Most recently, it has been reformatted and validated for an adolescent population. The current version of the adult ffq consists of approximately 140 food items.

The food frequency questionnaire is a list of foods with response categories answering "On average how often have you consumed this food in a specified time period (often 1 year)?" A typical set of frequency responses includes: never, 1-3/month, 1/week, 2-4/week, 5-6/week, 1/day, 2-3/day, 4-5/day and greater than or equal to 6/day.

Initially, the choice of foods on a questionnaire is often driven by a specific research interest. For example, the original Nurses' Health Study ffq sought to categorize individuals by their intake of nutrients and food items that were hypothesized related to cancer and heart disease(2,3). Therefore some of the nutrients and food items of interest were protein, fat and fatty acids, vitamin A, carotene, and cruciferous vegetables (2).

After determining which nutrients or foods are to be the focus of the research, three characteristics should be met when choosing a food for a ffq: the food should be eaten often by a considerable number of the sample studied, the food should have the nutrient(s) being studied in a substantial amount, and the amount eaten should vary between persons (1). Using the above criteria, a long list of foods is initially generated and systematically reduced.

To ascertain more specific nutrition information, it is often beneficial (barring cost and space constraints) to incorporate open-ended questions to identify specific brands of multiple vitamins, margarine, cooking oils, cereals and foods consumed at least once a week that were not asked on the ffq itself.

To quantify intake, the frequency response is given a weight (ie response of "1" or "never" receives a weighting factor of 0.00, response of "2" or "1-3/month" receives a weighting factor of 0.07, response of "3" or "1/week" receives a weighting factor of 0.14, etc) and the nutrients for a specific food are calculated by multiplying this frequency weight by the amount of the nutrient in a standard portion. For example, if the subject reports an average intake of 1 cup of skim milk 2-3 times / day, the nutrient values for 1 cup of skim milk are multiplied by the weighting factor of 2.5. This process continues for all food items in the ffq and the values are summed across all foods and vitamin supplements in order to derive a daily nutrient intake value across all nutrients. This method is useful in ranking subjects according to food or nutrient intake (i.e. by decile) so that extremes can be identified and compared.

The ffq is a relatively inexpensive tool because it is self-administered and therefore minimizes costs related to interviewer time. Optically scannable forms can be used to facilitate data entry of the responses and minimize errors. Due to its low cost, the ffq is a very popular dietary assessment tool for large scale dietary studies (1). However, as with any nutritional methodology, some drawbacks do exist. There is a potential bias in that the correct response is dependent upon the subjects' memories(1). Individuals may answer questions as they think would be viewed as better nutrition, that is, under-reporting foods considered unhealthy and over-reporting foods considered healthy(1). The food frequency questionnaire does not retrieve unique individual details of the diet unless, specifically designed to do so (1). Therefore, the items on the "typical ffq" may be inappropriate for subjects with culturally diverse food consumption patterns unless the ffq is designed for the specific population (4).

### **The Basis of Harvard's ffq Database:**

The database used for Harvard's food frequencies questionnaires' nutrient analysis is a specifically designed program documented as "harvardsffq.date." The foundation of the database is the US Department of Agriculture Standard Reference supplemented with the Food Consumption Survey data and additional information from McCance and Widdowson's *The Composition of Foods* (4th and 5th editions), journals, and manufacturers (5-7).

US Department of Agriculture obtains data from USDA analyses and other government laboratories as well as food industry and literature. This is the number one source of nutrient values because of the standard experimental procedures and representative large sampling of American foods (5).

McCance and Widdowson's *The Composition of Foods* (4th and 5th editions) provide additional foods and nutrients not available in the Handbook 8 series. Again the majority of the analyses of the foods listed in the book were conducted by the government laboratories or other contracted facilities (6,7).

Scientific journals are used for specific nutrients that are not provided by USDA or provide the most recent analyses. This is used particularly when one specific nutrient is being studied by an epidemiologist.

Manufacturer's information is used primarily in our cereal and vitamin tables that are brand specific. Additional information from manufacturers is used for new foods that have recently been created or are not available from the usual sources.

## How a Ffq Database is Different from a Dietary Record/Recall Database:

To understand the difference between a ffq database and a dietary record database, you must consider the instrument used to gather the individuals diet information. A food frequency questionnaire, as discussed above, is different than a dietary record or recall. The database therefore must also be different. The harvardsffq database has core foods that appear on the different versions of the questionnaire. Approximately 200 core foods (for several different food frequency questionnaires) are in our database. The preparation of the food in the database is specified by the questionnaire (ie raw, fresh, cooked, etc). So, each food item on the ffq has a corresponding food item with its nutritional composition in the database. In a dietary record database, one food might be a combination of food items in the database ( ie fried fish = fish + breading + fat) thus individualizing the nutrients eaten.

Following this same format the harvardsffq database has a specific portion size that each food on the questionnaire states or implies (bread - 1 slice, milk - 1 cup) and this is how it is found or calculated in the database. Portion sizes can be natural units (1 slice of bread or 1 medium apple) or commonly used portions based on literature (8). Some food frequency questionnaires, such as Block (9), have variable portion sizes (small, medium, large) but these are also rigid sizes predetermined and the specific food frequency is multiplied by this portion. This is again different than the diet record database where portion sizes can vary from individual to individual.

The cereal, oil, margarine, and vitamin open-ended or write-in sections on the Harvard ffq are brand specific. The distinctive brands of these food items require maintenance of additional nutrient files for cereals, oils, margarines, and vitamins. The databases for these sections may be similar to a dietary record database. However, they are processed following the ffq format using weighted frequencies multiplied by specific nutrient content of that food. The frequencies of these items are based on the frequency response for that question using a standard portion size. For example, if the cereal write-in section has "Rice Krispies" as a response, the computer program first picks up the frequency of the cold cereal question provided in the ffq and then picks up nutrients specific for "Rice Krispies" in our cereal table. This occurs on every questionnaire for vitamins, oils, margarines and cereals. In addition to these write-in sections, there is also the additional foods section. If a food is consumed at least once a week and is not included in the frequency section of the questionnaire, it can be written in by the subject with any specified amount and frequency. The food is then coded with an appropriate portion and frequency, and then analyzed in a similar manner as the rest of the foods on the questionnaire.

The foods listed on the questionnaire and the nutrients stored in the database are driven by what epidemiologists are studying. This is another difference from the dietary record database. The dietary record database is based on all foods that the population eats. It does not differentiate foods in the database by those three criteria that put a food on an ffq: that a food must be eaten often by a considerable number of the population, foods must have the nutrient(s) being studied in a substantial amount, and the amount eaten must vary between persons. In addition to the core foods found on the Harvard questionnaires, specific foods are added to and deleted from the questionnaires based on what is being studied. New foods that have been added and/or deleted over time include onions, garlic, soy sauce, raw carrots, raw spinach, and coleslaw. In addition to the core nutrients in the database, other nutrients are entered based on new research that is looking at these "new" nutrients or non-nutrients. Some new nutrients and non-nutrients that we have updated our database with are: flavonoids(quercetin, kaempherol, and myricetin), fibers (Englyst,

Southgate, AOAC, insoluble, cellulose, hemicellulose, lignin, soluble), trans fatty acids, phytate, and oxalate.

With more than 130 nutrients, there are foods in our database with blanks for specific nutrients because there is no resource available. In our nutrient analysis blanks or missing values are treated as zero. Therefore, the objective in the harvardffq database is to not leave a blank unless the nutrient is assumed to be negligible. To fill in these missing values, nutrients are imputed from either similar foods or from a recipe designed for that food. If there is not a similar food or a viable recipe cannot be written to fit its nutrient breakdown it is then left as a missing value. There are very few such foods and these have been documented to be blank. This may not be true in record databases that may allow more blanks.

Finally, the harvardffq reference system may be different from record databases in that it is designed for the investigators using the database. In research today, verification of the source of each nutrient is necessary, particularly when the research not only receives public attention through the media but also is shaping policy and health recommendations. For example, folate has been found to reduce the prevalence of neural tube defects (10) as well as reducing the incidence of colorectal cancer (11). Due to these and other on-going studies involving heart disease, the Recommended Daily Allowance for folic acid is being reevaluated. The source of the folate then must be accurate and exact. Our reference system documents the reference source for each nutrient of each food. This is very specific and each new reference has a specific number that can be used all through the database. There is also a date attached to the reference code to note when it was added or changed.

## **Updating the Database**

A nutrient data base needs to be constantly updated and expanded as new foods and supplements become popular on the market or updated nutritional data becomes available. With the help of a computer programmer, an on-line editing system can be set up for the data entry of new foods and nutrients. Data entry by hand is the primary method for adding information to the data base that has been obtained from scientific journals or manufacturers. However, the majority of foods and nutrients in current databases are from data provided by USDA. In the past, this information was available on tape to purchase and thus minimized data entry time and inaccuracies. More recently, the entire USDA Handbook 8 Series as well as nutrients from the Nationwide Food Consumption Surveys are available over the Internet. Our programming is devised to electronically update or add new foods or nutrients to the system. Advanced technology has greatly enhanced the data entry process. This saves an extraordinary amount of time and minimizes errors associated with manual entry.

In order to maintain internal quality control of the database, various cross checks are performed on a regular basis. Our first quality control step is that all updates and additions that are hand-entered are cross checked for accuracy. Updates and additions that are computer driven are checked by the dietitian doing the update. The values that are imputed or recipe driven are linked throughout the data system so that if one food is updated then any other food that has it as an ingredient or is dependent on it will automatically be updated or flagged.

The second quality control step has three parts or programs. Before the temporary files, where all editing occurs, are updated to permanent files which cannot be altered, three programs are run to additionally check for errors. The first is a "nutdiff" program that compares the temporary file to the

permanent file and records the differences. These changes are verified for accuracy. The second is a "checknut" program that checks all subcomponent parts of a specific nutrient that they add up to that nutrient. For example, this program would make sure that the sum of the amino acids do not exceed the total protein, the sum of the fatty acids do not exceed the total fat, the caloric content is approximately equal to  $4 * (\text{carbohydrate} + \text{protein}) + 9 * \text{fat} + 7 * \text{alcohol}$ . The last program of our second quality control step is a standard test file that is maintained with an analysis output. The standard test file is run each time the files are updated with the new permanent data file and the old analysis is compared to the new analysis. Differences must all be verified.

The third quality control step is added when the permanent version is made; the version is dated and a comment area is attached to the file to note what revisions (changes in the foods or nutrients) have been made. This provides additional control over a large database with numerous versions. In addition when an analysis is generated, the dated nutrient files and the analysis program are referenced on the output files. This provides a record of the specific files used for the analysis and any analysis can be reproduced if necessary.

## **Future**

The future direction of food frequency databases will, of course, follow the demands of research. Some of the new areas that are being studied in nutritional epidemiology are the non nutrient components of food (phenols), the use of food frequency questionnaires with new groups (eg. older children and adolescents as well as elderly), and the use of additional biochemical markers and corresponding nutrients to validate questionnaires. Our changing food supply will also direct our efforts to maintain a database that is current with what people are eating. This includes the ever increasing cereal and vitamin markets as well as the different trends in formulated foods such as "no fat" bakery products as well as the nutrients they provide. Finally the continual changing technology of computers requires that we do things faster, more accurately and in a smaller space. Preparing the pc operating version of the harvardffq database is in progress and will be our next technological change to our database.

In conclusion, the harvardffq database has changed tremendously since its inception in 1980. We have more than tripled the nutrients and quadrupled the number of foods. With the ever increasing pace of change brought on by computers and the information highway, we are now providing more information to more studies looking at what people eat and how this affects all of us.

## References

1. Willett WC. *Nutritional Epidemiology*. New York, NY: Oxford University Press; 1990.
2. Willett WC, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am J Epidemiol*. 1985; 122:51-65.
3. Willett WC, et al. The use of a self administered questionnaire to assess diet four years in the past. *Am J Epidemiol*. 1988; 127:188-99.
4. Block G. Human Dietary Assessment: Methods and issues. *Prev Med*. 1989; 18:653-660.
5. Composition of Foods. Washington, DC: US Dept of Agriculture; 1976-1992. Agriculture handbooks no 8-1 to 8-21.
6. Paul AA, Southgate DAT. *McCance and Widdowson's The Composition of Foods*. 4th ed. London, England: Her Majesty's Stationery Office; 1976.
7. Holland B, et al. *McCance and Widdowson's The Composition of Foods*. 5th ed. Cambridge, UK: The Royal Society of Chemistry and Ministry of Agriculture, Fisheries, and Food; 1991.
8. Church CF, Church HN. *Food Values of Portions Commonly Used*. 12th ed. Philadelphia: J.B. Lippincott; 1975.
9. Block G, et al. A data-based approach to diet questionnaire design and testing. *Am J Epidemiol*; 124:453-469.
10. Melunsky A, et al. Multivitamin/folic acid supplementation in early pregnancy reduces the prevalence of neural tube defects. *JAMA*. 1989; 262:2847-2852.
11. Giovannucci E, et al. Folate, methionine, and alcohol intake and risk of colorectal adenoma. *J Natl Cancer Inst*. 1993; 85:875-884.