JOURNAL OF FOOD COMPOSITION AND ANALYSIS (2002) **15,** 339–348
doi:10.1006/jfca.2002.1082
Available online at http://www.idealibrary.com on IDEAL®

## ORIGINAL ARTICLE

# Development of a Multi-nutrient Data Quality Evaluation System

Joanne M. Holden[1], Seema A. Bhagwat, and Kristine Y. Patterson

*Nutrient Data Laboratory, Bldg. 005, Rm. 201, BARC West, 10300 Baltimore Blvd., Beltsville, MD 20705-2350, U.S.A.*

Received October 26, 2001, and in revised form April 12, 2002

The U.S. Department of Agriculture's (USDA) Nutrient Data Laboratory (NDL) has redesigned the software of the USDA Nutrient Databank to provide a system for data acquisition, compilation, and dissemination, and as part of this system has developed a module to facilitate the evaluation of analytical data quality. USDA's first data evaluation procedures were developed as a manual system to assess the quality of analytical data for iron, selenium and carotenoids in foods. These procedures have been modified and expanded for multi-nutrient data quality evaluation. The same five evaluation categories; sampling plan, number of samples, sample handling, analytical method and analytical quality control, from the original work have been maintained but the evaluation questions have been made more objective. The rating scales for each category have been changed from discrete steps to a more continuous scale. These ratings are combined to give a Quality Index (QI), and Confidence Code (CC) that is disseminated with the nutrient data to indicate the level of confidence in the value. The algorithm for combining ratings from the five categories was revised from earlier systems to avoid the possibility that the aggregation of several mediocre data points would together, merit the highest CC.
© 2002 Published by Elsevier Science Ltd.

*Key Words:* data evaluation; nutrient database; analytical data.

## INTRODUCTION

The U.S. Department of Agriculture's (USDA) Nutrient Data Laboratory (NDL) disseminates the Nutrient Database for Standard Reference (SR) which provides values for more than 80 components in more than 6000 foods. USDA's food composition databases are used widely by the scientific community to monitor food intakes, to conduct nutrition research, to develop food and nutrition policy, and to develop public and private feeding programs. In addition, the data are used as the basis of other databases worldwide. As a first step the available nutrient values are collected from various sources. USDA's food composition data may be generated by

[1]To whom correspondence and reprint requests should be addressed. Fax: +1 301 504 0632. E-mail: jholden@rbhnrc.usda.gov

0889-1575/02/040339 + 10 $35.00/0 © 2002 Published by Elsevier Science Ltd.

chemical analysis of food samples, may be the result of calculations (by recipes or formulations) or may be the result of expert estimation or "imputation".

Analytical data may be obtained from the food industry, from other government agencies, from the scientific literature or from USDA-initiated contracts. As a result, the data gleaned from diverse sources may be of uneven quality and lacking in detailed supporting documentation. Therefore, it is of paramount importance to evaluate the quality of the analytical data for its reliability and to provide an indicator of data quality to guide the scientific community in various data applications. Finally, acceptable analytical data from various sources may be combined to provide the most representative estimates for nutritional components in foods.

USDA's data evaluation procedures were first developed as a manual system to assess the quality of analytical data for iron in foods (Exler, 1983). Each datum was evaluated by the criteria set for three categories—documentation of analytical method, sample handling and appropriateness of analytical method and analytical quality control. Data were rated on a scale of 0–3 for each category. This conceptual model was refined and expanded into five categories to evaluate data on selenium (Bigwood *et al.*, 1987; Holden *et al.*, 1987; Schubert *et al.*, 1987), copper (Lurie *et al.*, 1989) and carotenoids (Mangels *et al.*, 1993; Holden *et al.*, 1999). Two new categories, sampling plan and number of samples were added; analytical method and sample handling were separated to make two distinct categories and analytical quality control remained a separate category. All criteria used to assess data quality were revised. Specific questions were defined for each category and arranged in a series of "decision trees". Each question was linked to other questions: the direction of the query process was determined by the previous answer, usually in the form of a "yes" or "no" response. This process resulted in a discrete numerical rating for each category with the rating value as an indication of the relative quality of the data. For the most part the ratings were based on the acceptability and/or the completeness of the documented details which accompany the data values. These ratings for the five categories were combined to yield a single numeric "Quality Index (QI)" for a nutritional component in a food from each source of data. The combination of acceptable data for the nutrient component in a food resulted in a "Confidence Code (CC)" determined from the combination of all the QI results from all the data sources. The CC is an indicator of the relative quality of the data and of the confidence a user can have in each value. These quality codes were included with the data dissemination. With the development of the system for evaluating the quality of selenium data, unique software modules were developed to facilitate the data evaluation process (Bigwood *et al.*, 1987; Holden *et al.*, 1987). Upon completion of the system for evaluating the quality of data for five individual carotenoids (Mangels *et al.*, 1993; Holden *et al.*, 1999) USDA investigators noted many similarities and some differences in the detailed evaluation of data for different nutritional components, including both organic and inorganic components. This information was useful for developing a more universal system.

The NDL has redesigned the software of the USDA Nutrient Databank, Architecture and Integration Management-Nutrient Data Bank System (AIM-NDBS), to provide a system for data acquisition, compilation, and dissemination. As part of this system the NDL developed a module to facilitate the evaluation of analytical data quality. Since the USDA databank releases values for more than 80 nutrient components it was necessary to develop a multi-nutrient model to evaluate the analytical data for the five categories previously mentioned. The objective of this paper is to report the results of USDA's continuing research in the area of data quality evaluation and to present an expanded concept for multi-nutrient data quality evaluation. We will discuss the complexity of developing a general system for many nutrients and the importance of detailed documentation.

## METHODS

In 1995, a small international workshop was held at USDA offices to review the data evaluation system. Subsequent informal discussions with other experts at national and international meetings and during various sessions of the International Post-graduate Course on the Production, Management and Use of Food Composition Data (FoodComp) resulted in a list of constructive comments about the previous evaluation systems. Comments indicated that the five categories were useful for evaluation purposes and the idea of assigning ratings and confidence codes was reaffirmed. However, various experts stated that it was necessary to revise ''subjective'' questions to make them as quantitative and objective as possible. The concept of the number of samples analyzed had been ambiguous to some and needed clarification. It was also suggested that the discrete nature and the range (0–3) of the existing rating scale were not sufficiently broad or refined enough to permit evaluation of the essential detailed and specific documentation. And finally, a different algorithm for combining ratings was recommended to avoid the possibility that the aggregation of several mediocre data points would, together, merit the highest Confidence Code.

### *Categories of Evaluation*

The same five categories, sampling plan, number of samples, sample handling, analytical method and analytical quality control, were maintained as indicated by the comments. To facilitate the rating of each category, several criteria were developed to include distinct and specific details. Criteria for each category take the form of questions. These questions were reviewed and revised to make them as objective as possible. Rating points were assigned to the response to each specific question. These points accumulate to yield the specific rating for each category. Discussion of the five categories follows.

*Sampling plan.* The rating for sampling plan reflects the representativeness of the food sample units procured for analysis with regard to relevant factors (i.e., the food type, brand, cultivar, geographic origin, and/or market share). Historically, many published reports indicate the use of ''convenience'' samples obtained from local markets or experimental plots. Others conducted limited national or regional sampling. The ideal sampling plan would be based on statistical theory and address a demographic framework for national, regional (within a nation) or local sampling as well as the profiles and attributes of the food or foods to be sampled. Statistical guidelines for sampling can be found in various text books including Cochran (Cochran, 1977). The theoretical assumption includes random selections of cities, stores, lots, etc. to assure the opportunity for sampling according to probability. Regions can be set up by population density and distribution, ethnic diversity, or geography. Sampling schemes should include a balance of urban and rural locations.

In addition to the demographic framework for the sampling plan, one must include full descriptions of the food sample units as well as the details of the selection of the foods. These details include the determination of which types of foods (e.g., retail, wholesale, or home grown, etc.) as well as forms (e.g., preservation state such as frozen, dried, canned, cooked, prepared or preserved), sizes, cultivars or breeds, brands or trademarks, market share or sales volume, etc. Sampling strategies based on mathematical probabilities of occurrence may be difficult to plan and conduct due to the lack of relevant market statistics for production, sales, or consumption. In the

absence of these data, informal or qualitative information about consumption patterns, commercial market share or production and "disappearance" statistics may be used.

The USDA's new system attributes higher ratings to nationally representative collections of sample units. The system gives the highest sampling plan ratings for sampling designs which select units to represent foods in type and number currently being consumed by the population of interest. Compositing of sample units is one way to assure the representativeness of the mean while keeping the costs of analysis low. It can, however, result in the compression and underestimation of the variability estimate.

Sample units collected at retail stores, at wholesale outlets or at production sites may each be representative of the national distribution. There are criteria within the databank system to evaluate sampling plans for both retail and point of production sample collection. Foods that are available only in limited locations or are ethnic foods consumed by a limited population are referred to as the "unique foods"; the system is capable of evaluating sampling plans for these as well.

For samples obtained from retail sources, sampling plan parameters evaluated include the number of regions (each encompassing several states) represented, the number of cities within each region, the number of food outlets sampled within each locality or city, and the number of separate lots obtained as well as the number of seasons represented. Within the USDA system, the United States was divided into four regions—Northeast (NE), Midwest (MW), South (S), and West (W) using information from the National Bureau of the Census. Postal codes were used to define cities and locations. Sampling plans that were probability based (e.g., related to the frequency distribution of the randomly selected towns, stores, brands, specific packages, etc.) were assigned higher ratings than those that were non-probability based (i.e., samples selected for convenience, experimental purposes, or non-random sample collection schemes planned without regard to population distribution or sales volume).

In retail sampling plans, the highest rating would be given to a statistically designed plan that samples from all four regions, from three cities in each region and two locations (stores) in each city picking up two different lots at each location with additional credit if the samples are obtained in two different seasons. This would total 48 sample units. The scoring system is designed to give the most emphasis to the number of regions represented, i.e., if resources were limited, collecting from one city in each of four regions would result in a higher rating than two cities in each of two regions. Likewise, the number of cities is given more weight than the number of locations in each city or the number of lots collected at each location. The allocation of points is based on the assumption that variability in nutrient content among sample units obtained from different regions will be greater than variability among sample units from different lots. A well-designed plan could receive a high, but not optimal rating, with fewer than 48 sample units and still be reasonably representative of the nation's food supply.

Sampling plans for unique foods distributed at retail outlets, as defined above, are rated in a similar way with the exception that "regions" are not preset within the system, but are defined with regard to the availability of the food and consumption patterns. Knowledge of detailed demographic characteristics and food properties are essential for evaluation of this category as well as for planning future studies.

Entering information into the evaluation system for sample unit collection at points of production (i.e., manufacturing plants, orchards, feed lots, etc.), whether a widely consumed food or a unique food, requires a good understanding of market share as well as production locations and distribution. The ratings reflect how well

the sample units represent the product that is available to the consumer. For the point of production sampling, the system considers the % of national (total) production represented by the sample units collected instead of the number of regions and the number of cities sampled in the retail scheme. For example, if the production locations sampled accounted for 90% of the product produced and available to the consumer, that portion of the rating would be high. Other factors considered in the rating would be the number of days/lots from which samples were collected as well as the number of seasons represented. Again, the rating would be higher if the sampling was statistically planned. While a highly rated sampling plan will include selection of sample units from different lots, days and seasons, the goal is to determine a reliable estimate of mean nutrient content for the food. This approach does not necessarily permit the quantitation of sources of variabilities such as specific agricultural conditions.

The United States is a large country with a diverse population. Therefore, the "Sampling Plan" category is important and the criteria are complex. Other countries may not need such a complex set of criteria, but should consider the underlying principles of demography and food production patterns to obtain nationally representative samples.

*Number of samples.* Assessment of the number of samples analyzed is critical to the estimation of the mean as well as to the magnitude of variability for a component in a food. The analysis of a small number of independent samples limits the ability to estimate the mean and variability. On the other hand, the analysis of adequate numbers of samples (either individual or composite samples) permits the more accurate calculation of the mean. In addition, information concerning the standard deviation of the nutrient values is desirable to estimate variability. The rating for number of samples is determined by the number of sample aliquots analyzed independently. Repeated analyses from the same homogenate (from extraction to analysis) validates the homogeneity of the sample and repeated analyses of the same extract validates instrument precision. Although it is important to assess analytical precision, they do not necessarily strengthen the estimation of the mean or provide data to assess or calculate the variability in the nutrient values for different sample units, or varieties, cultivars, etc. Therefore, when a specific sample is analyzed more than once (e.g., replicated) it is counted as one sample. Likewise, the analysis of a composite of many sample units is counted as one sample for this category (the collection of many sample units is rated under the sampling plan category). The highest rating is given for 12 or more analytical samples. This is, in part a practical decision since in most studies of food composition it is unlikely to obtain data for as many as 12 or more different samples of any food. The complexity of food composition and consumption makes it difficult to characterize a food in a definitive way when one does not have the resources to analyze hundreds of samples, with every permutation of brand, breed, or environment. While it must be recognized that a better estimate of the mean value can be obtained through the analysis of samples composited of several to many units the analysis of composite samples obscures or confounds any determination of serving to serving variability. At this time the USDA system does not consider standard deviation in the assignment of rating since this information is frequently not documented. Theoretical considerations indicate the sample size needed to obtain an error bound having a given confidence level is proportional to the between unit variance and inversely proportional to the length of the error bound. Basing future outcome projections on historical data from the NDL's database it strongly suggests that more than 25 analytic samples would be needed to improve the reliability of the estimated mean.

*Sample handling*. To assure general nutrient stability of the food matrix, nutrient content and representativeness of a sample, proper handling of sample units and composites is critical. The system evaluates sample handling including all the steps from the point the sample is acquired to the point when the analysis begins. The rating is based on information about storage temperature, processing (dissecting, cooking, weighing), homogenization, and moisture content (indication of proper storage) information. Optimal ratings are assigned to sample-handling conditions which would preserve the original nutrient content and the integrity of the food matrix.

*Analytical method*. Valid and meticulously applied analytical methodology is critical for obtaining accurate nutrient data. The system evaluates analytical methods in two parts; (1) the method itself is evaluated using optimal criteria established to judge the validity of processing, identification and quantitation steps employed, and (2) whether the laboratory generating the analytical data has a demonstrated ability to use the method to obtain accurate data. In view of the importance of this category we have organized committees of international experts to identify the critical methodological steps for each method for each nutrient (or a group of related nutrients). These critical steps are used to validate the method from processing of samples (extraction, digestion, etc.) through analysis (detection/identification, limits of detection, percent recoveries) and quantitation (limits of quantitation, calibration curves, calculation algorithms, certified/standard reference materials, etc.). Each step of the method is assigned points depending on its importance for accurate analysis. These steps are then used to prepare a framework of questions and rating points for that particular nutrient/method combination and are incorporated in the AIM-NDBS. The method evaluator selects the appropriate nutrient/method combination for the specific documented method and answers the questions presented, e.g., Hertog *et al*. (1992), flavonols by high performance liquid chromatography (HPLC). Figure 1 illustrates the method evaluation pathway for flavonols by HPLC. The system generates an ''interim rating'' for the method in question which is stored in the system to be retrieved when data generated by this method are evaluated.

For the second part of the evaluation, the performance of the analyst using this published method is validated by examining their results for precision (repeatability) and accuracy (values within the accepted range of a reference material, or percent recoveries of spikes or internal standards along with comparison with another laboratory or method). The use of a certified reference material to determine accuracy receives a higher rating than the use of reference or consensus materials having values of less certainty. The interim rating is summed with the laboratory performance rating to give the analytical method rating. A rating of less than 5 out of 20 possible points for analytical method indicates the need for further review of the data.

*Analytical quality control*. Analytical quality control refers to the accuracy and precision in the day-to-day performance of the analytical method. Accuracy and precision are judged as described above. A quality control (QC) material (certified, reference, consensus or in-house material which has been developed for particular nutrients) should be used with each batch or on each day of analysis. Analytical values for in-house materials should be confirmed by concurrent analysis of a certified material or confirmatory analyses done by a second method or by another laboratory. However, values within the accepted range of a certified reference material get higher ratings than those for other materials. In addition, the frequency of the use of a QC material is one of the criteria for evaluating quality control.

**Example of Flavonol Method Evaluation:**
**Evaluation of HPLC Methods**

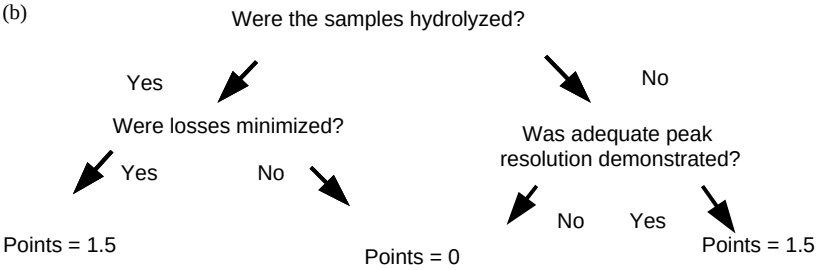
Software presents questions based on information flow illustrated below

**Sample Processing:**

(a)

| | Points for "Yes" |
|---|---|
| Were samples protected from oxidation? | 0.5 |
| Were samples protected from UV light? | 0.25 |
| Was the sample size ≥ 1 gram (dry weight)? | 0.5 |
| Was optimization of extraction reported? | 1.25 |

(b)




**Analysis & Quantification:**

(c) Flavonoid peaks identified by:

| One method Points=1 | Two or more methods Points=2 |
|---|---|
| Possible Methods: | |
| Retention time | 2nd HPLC method (diff. mobile phase) |
| Co-chromatography | Nuclear magnetic resonance |
| Thin layer chromatography | Mass spectrometry |

(d)

| | Points for "Yes" |
|---|---|
| Was the purity of flavonoid standards verified? | 0.4 |
| Were there ≥ 3 standards used for the calibration curve?* | 0.4 |
| Was linearity of the standard curve demonstrated? | 0.4 |
| Was the calibration curve correlation coefficient (r) ≥ 0.99? | 0.4 |
| Was the instrument response checked frequently? | 0.4 |

*If the answer is "No", the scores for both C and D are zero

Analytical Method Interim Rating = A + B + C + D



FIGURE 1. Example of rating criteria: flavonols by HPLC.

*Rating Scales, Quality Index and Confidence Code*

The format for rating computation was changed. Rating points were assigned to questions incrementally thus creating a more continuous rating scale. Previously, the rating scale was discrete with possible integer values in the range of 0–3. This range was expanded to 0–20 to permit more detailed examination of various aspects of the documentation.

In the databank system, nutrient data from several acceptable sources are combined to give an overall estimate of the nutrient content of that food. Ratings are recalculated for these aggregated data based on the ratings from all the sources. For sampling plan, while individual sources may not have received high ratings, the rating for the aggregation of similar items may be higher if those sources together represent more regions. The system considers the number of regions, number of seasons, percent market share or percent of production volume represented by the aggregate. A sampling plan for an aggregate with all sources having systematic probability designs receives a higher rating for the same criteria. The number of samples rating for the aggregate is computed from the total number of samples in all the sources. At aggregation, the sample handling, analytical method and analytical quality control ratings are calculated as weighted averages in the same manner as the nutrient values are weighted by the market share of the particular food.

After each nutrient value is rated for all five categories, the ratings are summed to yield a Quality Index (QI) to determine the overall acceptability of the value. Since each category has a maximum rating of 20, the highest possible score for a QI is 100. At aggregation, QI is calculated by summing the adjusted ratings for each category of the aggregate giving a maximum score of 100. This method of calculating a QI represents a revision of the method used in previous evaluation systems. It addresses the concern of reviewers that the aggregation of numerous mediocre data sets could merit a higher Confidence Code (CC). The QI is used to determine the CC which is a letter (A, B, C or D) assigned to the nutrient estimate of the food, 'A' being the highest quality and it is disseminated with the food value.

## RESULTS AND DISCUSSION

Recently, the USDA data quality evaluation system has undergone major revision to become a module in the new AIM_NDBS. This multi-nutrient expanded version builds on USDA's experience in developing various prototype systems for data quality evaluation since 1983. The evaluation of food composition data quality is a complex process which, in large part, relies on the completeness of documentation of information for five categories including sampling plan, sample handling, number of samples, analytical method, and analytical quality control. While the five categories and the application of a rating system have been reaffirmed, various changes have been made to revise specific questions and criteria in each category as described above.

The extension of the maximum rating scale for each category from 3 (earlier version) to 20 and use of a continuous scale allows better scrutiny of the categories. In the earlier system a QI was calculated by averaging the ratings of the five categories and at aggregation QI values were summed over several sources to get a CC. With this method of calculation a CC depended as much on the number of data points as on the quality of each data point. The modified system emphasizes the quality of the data rather than the total number of data points.

The software system has been designed to allow data and other information to be entered into the database either manually or electronically. The electronic entry is the most efficient, particularly for data which comes from NDL contracts where complete information is available and received in a standard electronic format. Data from some sources can present problems in making accurate assessments for a variety of reasons including (1) insufficient or unclear documentation of the origin of the food samples, (2) failure to either report or to use QC materials, (3) confusion regarding the number of samples analyzed, and (4) an overall lack of details. Often complete information that can support the validity of the data is not documented in published reports. Sometimes it is possible to contact the authors to obtain additional information, but this is time consuming. Authors reporting nutrient data should be aware that inclusion of as many relevant details as possible facilitates the evaluation of their data.

The complexity of developing a multi-nutrient evaluation system requires creation of nutrient/method specific criteria and extensive documentation. We plan to develop analytical method evaluation criteria for most of the nutrients of importance. With the assistance of many experts we have developed criteria for carotenoids, flavonoids, vitamin K, sugars, and dietary fiber and we are actively working on riboflavin, nitrogen, ash, moisture, and minerals. This information will provide guidelines for the evaluation of analytical data with particular regard to analytical methods. As analytical data are evaluated we will generate and release confidence codes with the analytical nutrient values in the future USDA Nutrient Databases for Standard Reference.

## CONCLUSIONS

To summarize the achievements of the new system: The data evaluation system allows (1) systematic and standardized evaluation of data quality, (2) identification of critical points which determine data quality, (3) archiving of detailed information used or considered for evaluation allows for retrieval of documentation for each source, and (4) recompile and reevaluation of data when changes in analytical methodology or specific processing/preparation techniques occur. Indications of data quality will assist in establishing priorities for the analysis of foods where existing data are of poor quality. The evaluation process has emphasized the need for certified reference materials for emerging nutrients. Once again, we emphasize the importance of giving complete documentation for all aspects of compositional studies to permit the proper assessment of the data quality.

## REFERENCES

Bigwood, D. W., Heller, S. R., Wolf, W. R., Schubert, A., and Holden, J. M. (1987). SELEX: an expert system for evaluating published data on selenium in foods. *Anal. Chim. Acta* **200,** 411–419.

Cochran, W.G. (1977). Sampling Techniques. John Wiley & Sons, New York.

Exler, J. (1983). Consumer Nutrition Division, Human Nutrition Information Service, USDA, HERR Number 45.

Hertog, M. G. L., Hollman, P. C. H., and Venema, D. P. (1992). Optimization of a quantitative determination of potentially anticarcinogenic flavonoids in vegetables and fruits. *J. Agric. Food Chem.* **40,** 1591–1598.

Holden, J. M., Schubert, A., Wolf, W. R. and Beecher, G. R. (1987). A system for evaluating the quality of published nutrient data: selenium, a test case. *Food Nutr. Bull.* **9**(Suppl.), 177–193.

Holden, J. M., Eldridge, A. L., Beecher, G. R., Buzzard, I. M., Bhagwat, S., Davis, C. S., Douglass, L. W., Gebhardt, S., Haytowitz, D., and Schakel, S. (1999). Carotenoid content of U.S. foods: an update of the database. *J. Food Comp. Anal.* **12,** 169–196.

Lurie, D. G., Holden, J. M., Schubert, A., Wolf, W. R., and Miller-Ilhi, N. J. (1989). The copper content of foods based on a critical evaluation of published analytical data. *J. Food Comp. Anal.* **2,** 298–316.

Mangels, A. R., Holden, J. M., Beecher, G. R., Forman, M. R., and Lanza, E. (1993). Carotenoid content of fruits and vegetables: an evaluation of analytic data. *J. Am. Diet. Assoc.* **93,** 284–296.

Schubert, A., Holden, J. M., and Wolf, W. R. (1987). Selenium content of a core group of foods based on a critical evaluation of published analytical data. *J. Am. Diet. Assoc.* **87,** 285–296, 299.