COMMENTARY

Data Quality in Sampling, Analysis, and Compilation

David A. T. Southgate

8, Penryn Close, Norwich NR4 7LY, Norfolk, U.K.

The focus of this paper is on Data Quality as it applies to Nutritional Compositional Databases. Quality implies fitness for the purposes for which the data will be used. An appreciation of the requirements of the users of the database is thus a primary determinant of the standards that should be applied during the generation of the data. Most users expect that the compositional data will be representative of the foods consumed by the population being studied and thus Sampling is the first critical element. It is also expected that the data will be analytically sound so the second critical element is Analytical Quality. Finally, the Compilation process involves making judgements about the data to be included and their presentation and in this the users expect that the compilation process will adopt to focus on the quality of the data that will meet their needs. The integration of these elements during the production of data for inclusion in the database will be advocated. Finally, there follows a discussion of how best to present data quality in a database so that the users can use the database appropriately and effectively.

Key Words: requirements of users; data quality; sampling; analysis; compilation.

INTRODUCTION

Data quality is a measure of the fitness for the purpose for which the data will be used. Therefore, any discussion of the quality of data in a nutritional compositional database must start from a consideration of the ways in which databases are used (Southgate and Greenfield, 1992). High-quality data meets the needs of the users well, whereas low-quality data may be unsuitable for the needs of most users, while it may well be adequate for the less demanding types of use.

Databases are used in a large number of ways and providing high quality for each and every type of use is a very difficult task. All users, however, have a number of requirements in common (Table 1).

First, they expect that the compositional data will be representative of the foods in the country or region for which the database was prepared. Second, they expect that the data will have been obtained using the most appropriate methods of analysis and that the analyses will have been carried out correctly. Third, they expect the data to reflect “accurately” the composition of the foods.

1To whom correspondence and reprint requests should be addressed. Tel.: +44 1603 465540.
The requirements for accuracy, in the strict sense, are in fact different for different types of user. Those involved in the educational use of databases or in the formulation of dietary guidance (including many dietetic uses) are using the database in a semi-quantitative way. Whereas at the other extreme those who use a database for calculating the composition of diets in metabolic studies require an accuracy that no database can achieve (Paul and Southgate, 1978). The requirements of those using databases for nutritional epidemiology fall between these two extremes of requirements for formal accuracy, i.e., closeness to the true composition of a food.

ENSURING THAT THE DATA ARE REPRESENTATIVE

This requirement implies that achieving the measure of data quality relates to the way in which the food sample to be analysed was sampled. The design of the sampling protocol is therefore a prime requirement in generating data for a database. The ways in which databases are used make this an extremely onerous and at times impossible task. The individual foods consumed by subjects in an epidemiological study are rarely representative of the same food for the country as a whole. Moreover, databases are often used over a long period of time whereas the sampling has been, inevitably, carried out earlier. Users act as though the composition of a given food or groups of foods is constant over both time and space.

All that the most quality conscious compilers can do is to ensure that the samples were representative of the foods of interest at the time of sampling.

All sampling is associated with a degree of sampling error and the confidence limit of sampling is defined first, by the observed variability of a food and the chosen level of confidence that one wishes to achieve has often to set intuitively in deciding the number of replicate samples to collect.

Often in database work the variability is not known and a value must be assigned intuitively. Moreover, the variability of some nutrients may be greater than others so the number of samples required to achieve the same confidence limits for all nutrients may differ (Holden and Davis, 1995; Makinson et al., 1987).

THE DESIGN OF SAMPLING PROTOCOLS

These protocols are based on knowledge of the foods being studied and, ideally its consumption by the population for which the database is being prepared. The
consumption data will show whether or not some stratification by regions within the
country or by population group, for example, is required. Knowledge of the food
and the methods used in its production, storage and marketing is also essential. This
will determine the pattern of sampling throughout the seasons, for example, or after
storage under certain conditions. These possible sources of variation may suggest
further levels of stratification in the sample protocol.

The knowledge of the food will also provide some information on which to base a
measure of the variability in the composition of the food. If it is possible within the
resources required the samples collected should be analysed separately so that one
has direct information of variability. Where resources are limited it may be possible
to get variability data for a limited number of nutrients for which this food is a major
source in the diet.

More commonly in database programmes the individual samples are combined
into a mixed composite sample for analysis. The confidence limits for the values
obtained on this mixed sample are those chosen in selecting the number of samples
forming the mixture. Most commonly these are set as 0.05 or 5% which is the sam-
pling error, in addition to this there is the uncertainty associated with all analytical
methods.

Designing sampling protocols that include all the known sources of variation leads
to a very extensive protocol that is consequently expensive to implement (Torelm,
1997). In most database programmes it is usual to develop complete sampling
protocols for those foods which are major components of the diet and to use simpler
protocols for those foods which are minor components. This of course implies that
the data generated will have a lower quality than if they were based on a more
complete protocol.

Once the basic features of design are established the protocol must be developed
into a complete account that will form the basis for instructing and training the
samplers. This account should ideally be developed in collaboration with those who
will be responsible for the analytical work. This ensures that the amounts of food
collected will be adequate for the projected analyses and that the conditions used in
the storage of the samples will not compromise the analyses. Logistical considera-
tions also need to be taken into account in the delivery of the samples to the
analytical laboratories.

Finally, the sampling protocol needs to be run through to identify possible
problems. This may have to be done at least as “a paper exercise”. The elements in
the design of sampling protocols are set out in Table 2.

ANALYTICAL ASPECTS OF DATA QUALITY

A great deal of attention has been given to ensure the quality of analytical data
because it is of prime importance for virtually all purposes for which foods are
analysed. Many of the principles developed are especially relevant to the generation
of data for databases (Greenfield and Southgate, 1994).

The first essential is to establish a Quality Assurance Scheme (QAS) that applies to
all aspects of data generation from the receipt of the food samples through to the
final provision of a record of the results (American Society for Quality Control,
Statistical Technical Committee, 1973). One key aspect of this scheme is that all
aspects of the work in the analytical laboratory are considered. Thus staff and staff
training, the working environment and the provision and maintenance of
instrumentation are covered. Documentation is a key to the implementation of a
QAS, so that samples can be tracked through the laboratory from receipt to the provision of analytical results.

The procedures for handling the samples as received and the preparation of analytical samples and their storage are set out in detail. The choice of analytical methods forms part of the QAS together with the procedures used to evaluate the performance of the methods. It is not sufficient to rely on performance criteria established in another laboratory, for example, in the laboratory of the author of the method, although these are valuable guides. For the procedures for controlling the quality of the analyses, standards to be used and the use of CRMs or in-house reference procedures, see Hollman et al. (1993). The logic of the algorithms for calculating results form part of the QAS together with any check procedures.

The essential features of a QAS are set out in Table 3 and it is evident that when these are developed into a manual they form a substantial tome especially in a laboratory which has a wide portfolio of methods. One important point to emphasize here in the training of staff is that this presents the Quality Assurance Manual as a working manual to use and develop with time. QA manuals have no real value unless they are used regularly.

---

### TABLE 2

Main elements in the design of sampling protocol

<table>
<thead>
<tr>
<th>Element</th>
<th>Factors considered</th>
<th>Contribution to the design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification of food</td>
<td>Type, mode of preparation and storage</td>
<td>Essential for proper identification of food to be sampled</td>
</tr>
<tr>
<td>Food variables</td>
<td>Seasonal&lt;br&gt;Regional&lt;br&gt;Storage&lt;br&gt;Marketing</td>
<td>Stratification of sampling design</td>
</tr>
<tr>
<td>Production</td>
<td>Seasonal&lt;br&gt;Regional&lt;br&gt;Storage&lt;br&gt;Marketing</td>
<td>Stratification of sampling design</td>
</tr>
<tr>
<td>Consumption</td>
<td>Seasonal&lt;br&gt;Regional&lt;br&gt;Specific groups within population</td>
<td></td>
</tr>
<tr>
<td>Nutrient variability</td>
<td>From pilot studies or by intuitive methods</td>
<td>Numbers of replicate samples required</td>
</tr>
<tr>
<td>Analytical factors</td>
<td>Amounts of food required&lt;br&gt;Specific storage or preservation conditions of samples&lt;br&gt;Logistics of integrating sampling with analysis</td>
<td>Practical matters to be considered in the detailed protocols</td>
</tr>
<tr>
<td>Preparation of detailed sampling protocols</td>
<td>Formal training and testing of protocol</td>
<td>Elimination of problems during sampling&lt;br&gt;Quality of sampling</td>
</tr>
</tbody>
</table>

---

**COMPILATION**

The compilers are responsible for ensuring that the quality of the database meets the needs of the users. In order to do this there needs to be consultation between the users and the compilers.
Where the database programme includes provision for generating the data directly (Southgate, 1974; Southgate and Greenfield, 1992) by a sampling and analytical programme for producing data, the compilers have direct control over the sampling and analytical procedures and therefore direct control over the quality of the data.

More usually compilers have to make use of data which has not been produced under their direct control, that is from published and unpublished sources, including other compilations of data in databases or printed food composition tables.

These data have to be evaluated using the same standards. This involves first, considering the sampling protocol, its design and execution, and second the choice of analytical methods and their execution and the quality assurance procedures in place.

This is one of the most difficult features of database compilation because few published sources describe the sampling procedures used in sufficient detail to assess the confidence limits that apply to the data. Analytical procedures are more commonly described in more detail but comments on the quality assurance procedures, for example, the standards or SRMs used, are rare. Some of this is the responsibility of editors and their journals’ restrictions on page numbers but often authors seem unaware that these details give greater confidence to their data.

The range of the information that compilers use in assessing data are given in Table 4.

**TABLE 3**

<table>
<thead>
<tr>
<th>Main elements of a Quality Assurance Scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory facilities</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Documentation</td>
</tr>
<tr>
<td>Analytical stages</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

All the above elements are brought together in a Quality Assurance Manual.

Where the database programme includes provision for generating the data directly (Southgate, 1974; Southgate and Greenfield, 1992) by a sampling and analytical programme for producing data, the compilers have direct control over the sampling and analytical procedures and therefore direct control over the quality of the data.

More usually compilers have to make use of data which has not been produced under their direct control, that is from published and unpublished sources, including other compilations of data in databases or printed food composition tables.

These data have to be evaluated using the same standards. This involves first, considering the sampling protocol, its design and execution, and second the choice of analytical methods and their execution and the quality assurance procedures in place.

This is one of the most difficult features of database compilation because few published sources describe the sampling procedures used in sufficient detail to assess the confidence limits that apply to the data. Analytical procedures are more commonly described in more detail but comments on the quality assurance procedures, for example, the standards or SRMs used, are rare. Some of this is the responsibility of editors and their journals’ restrictions on page numbers but often authors seem unaware that these details give greater confidence to their data.

The range of the information that compilers use in assessing data are given in Table 4.

**DEscribing data quality in a database**

Many users of databases are content to allow the judgement of the compilers in selecting data of the required quality. Others would like some assessment of the confidence that can be ascribed to the data. Data quality as said earlier is an integration of sampling and analysis procedures. Assessing the quality of analytical data has been addressed for a number of nutrients, iron, selenium and carotenes. The approach has been to establish the “best practice” for carrying out a method and to assign quality codes. Some of these apply to the details published in the paper from
which the data are taken. The concept of data quality codes is being extended (Holden et al., 2002) to include sampling and to other nutrients by examining the details of each method and establishing the key stages which determine the confidence one can ascribe to the data.

These confidence codes cannot be expressed statistically in a formal way. In using these codes it is important to remember that these codes are categories, not real numbers which can be manipulated arithmetically nor do they mean that data with a quality code of 3 is three times better than data with a quality code of 1.

As mentioned earlier the sampling protocols for foods that are minor components of the diet are commonly simpler than those used for important foods in the diet either quantitatively (frequency and amounts) or as sources of nutrients that are especially important to public health.

One aspect of data quality that applies to all data and which is important to convey to users is that all foods are biological materials and show natural variability in composition (Paul and Southgate, 1978). The best-sampling protocol can therefore only give a sample whose composition lies within the confidence limits assigned in sampling. Analytical uncertainties are part of the overall error. These are greater for some nutrients than others but it is commonly observed that the precision of analyses is a function of the composition of the analyte, so analytical confidence limits are wider for trace constituents than for macronutrients (Horwitz, 1976).

The variability in the composition of foods is one reason why all calculations from a database are associated with uncertainty that limits their accuracy. If very high accuracy is required, for example, in a metabolic study, then a database prepared from the analysis of the foods being consumed should be used.

REFERENCES


