Measuring and Improving Data Quality
Part II: Measuring Data Quality

Written by Mike Martin DVM, MPH, Clemson University

Part I of this series presented examples of "data gone bad."[1] Data problems are common sources of frustration, and it often seems impossible to improve the quality of the data we work with. However, well researched and proven methods for measuring and improving data quality do exist. To improve data quality, it is important to define what is meant by "quality" and establish methods of measuring that quality.

Data quality is not a single attribute. It can be measured on many dimensions and is often perceived differently by different customers.[2] For example, timeliness may be the most important factor to one data consumer while completeness may be most important to another. In addition, the relative importance of different quality dimensions varies from customer to customer, as do each user's required levels of quality on each dimension. For instance, a syndromic surveillance program might require data on herd investigations within the first 48 hours after each visit (to be able to respond in a timely manner) but would require findings on only very general categories.

Some commonly listed data quality attributes include:

- Validity and integrity: are the data correct?
- Accessibility: are the data readily available?
- Timeliness: is the data available when needed?
- Accurate clinical content: has the clinical picture been accurately described?
- Temporal reliability: do the meaning or intent of the data collected remain consistent over time?
- Completeness: do the data contain all relevant information?
- Precision: how well do the data reflect the full details of the original observation?

There is a tendency to focus on only validity and integrity when considering data quality. Of course, a data system must capture correct information, and data should not be changed during storage or processing. Correct information provided in a timely manner to those with need and authorization is the most familiar feature of data quality. However, there are finer points to data quality that must be considered:

Accurate clinical content refers to the fact that clinicians frequently use common terms in ways that carry very specific clinical meanings. Clinicians may use terms that they *believe* carry specific meanings when, in fact, there is actually little consensus. In 1980, Bryant and Norman[3] demonstrated that when physicians read clinical reports and assigned values to the verbally described probabilities there was almost no interclinician agreement on the meaning of the verbal expressions.[4] High quality data use only terms with clear clinical meanings or explicitly define the terms used.

Temporal reliability points to the problem of meanings that change over time. Without clear clinical case definitions, the meaning of each diagnosis changes, as does our understanding of

disease. During an outbreak or with program diseases, there is a degree of control over strict clinical case definitions and other factors affecting data collection and classification. This is not the case with more general surveillance, where each data element's quality must be assessed for factors that may change its meaning, measurement, or recording methods over time.

Completeness also is susceptible to variability as time passes and data sources change. Voluntary reporting systems are notoriously susceptible to various reporting biases. Although there are ways to "live" with these sources of bias, they should be reported accurately in any analysis of data quality.

Precision is usually thought of as the number of significant digits in numerical measurements, but the concept applies to any data. Poor data quality can result from the loss of precision or by expressing more precision than is actually present. Quantitative test results reported as simply "normal" or "high" lack precision.[5] However, expressing too much precision can also be problematic. A common mistake is recording zip code centroids as seven digit latitude and longitude values. In this case, the expression of the data implies more precision than is actually present in the original measurement. Coded values can include the degree of precision by using hierarchical coding systems that allow expression of either specific terms or a more general concept. A common example is use of Linnaean taxonomy to name the genus but not the species of an organism, e.g. *Streptococcus sp.* The Systematized Nomenclature of Medicine[6] is another example of hierarchical representation. Good data quality on the precision dimension, therefore, involves recording the full precision of the original measurement but also accurately reflecting the limits of that precision.

These examples show the complexities involved when defining measures of data quality. Each of the quality dimensions important to internal or external data customers must be defined and methods found to reliably measure these attributes. Measurement generally involves comparisons of a data sample in the system with some independently collected "truth." This sampling involves statistical methods familiar to epidemiologists.

Once we have a clear definition of what we mean by data quality in our own applications, we can begin to apply proven methods of data quality improvement. Those methods will be the topic of the third paper in this series. The series will conclude with specific recommendations for improving data systems within Veterinary Services.

---

[1] Martin MK, Measuring and Improving Data Quality; Part 1: The Importance of Data Quality, NAHSS Outlook, February 2005, http://www.aphis.usda.gov/vs/ceah/ncahs/nsu/outlook/issue4/data_quality1.pdf (accessed 02/18/2005).

[2] R. Y. Wang , M. P. Reddy , H. B. Kon, "Toward quality data: an attribute-based approach", *Decision Support Systems*, 13:3-4, pp.349-372, March 1995.

[3] G.D. Bryant, G.R. Norman, "Expressions of probability: words and numbers," *N Engl J Med.* 302(7), p. 411, Feb 14, 1980.

[4] D.L. Sacket, et.al., *Clinical Epidemiology; A Basic Science for Clinical Medicine 2nd Ed.*, Boston, Little Brown And Company, 1991, p 109.

[5] Sacket, p 101.

[6] Snomed International, Systematized Nomenclature of Medicine, http://www.snomed.org/index.html (accessed 02/15/2005).