Measuring and Improving Data Quality
Part 1: The Importance of Data Quality

Written by Mike Martin DVM, MPH, Clemson University

Veterinary medicine is a very data-intensive business. Clinical and regulatory decisions are made using information derived from a vast number of data sources. Some of these data are precise numerical values, such as antibody titers from the laboratory. Others are more subjective in nature and involve a significant range of uncertainty. But in all cases, we operate on a belief that information derived from our data is “right.” But what, exactly, do we mean by data being right? Several examples of “data gone wrong” might illustrate the importance of improving data quality.

In our first example, an area epidemiologist received a copy of ArcGIS mapping software and decided to generate a quick map of his State’s premises. He pulled FRONT_GATE_LAT and FRONT_GATE_LONG from the State’s GDB database and plotted them on the map. All of the State’s premises plotted somewhere in the North Atlantic near Greenland. There was little doubt that the values in the database were invalid; they obviously did not reflect the real location of these premises. His investigation found that the latitude and longitude coordinates had been switched and the signs reversed.

On December 24, 2003, a cow tested positive for BSE and stole Christmas. During the next five months an enhanced surveillance program and supporting data system were quickly designed. One very important issue was recording the age of the cattle sampled. The system designers understood that the veterinarians and technicians sampling in the field often would not have access to the recorded age of the animals and, therefore, would need to estimate age based on the animals’ teeth. To avoid making the age seem too precise, the system was designed to record age as a series of ranges. This seemed like a good solution until, several months into the sampling, the national managers requested a report with the age divided into categories different from those recorded in the database. The designers in this case had the right idea of not recording more precision than known, but valuable information was inadvertently lost by aggregating data in advance instead of as needed.

Another example illustrating data gone wrong involves an area epidemiologist who wanted to know how his State’s TB testing program was operating. He queried the GDB for caudal fold tests performed by practitioners, which produced numbers of negative and suspect tests. He calculated the rate of suspect tests for each practitioner to see that they were all getting the expected number of suspects (some results are suspect (false positive) even with no TB in the State.) Finally, he looked to see that each suspect test result was followed by a negative comparative cervical test. To his surprise, there were no comparative cervical tests recorded for the past year. Unless this State was truly mishandling their TB program, the GDB data were clearly incomplete. Further investigation showed that in the past it had seemed easier to simply add in the few comparative cervical tests at the time the annual report was written, rather than enter them into the GDB. The annual reports were correct, but any direct analysis of the database gave a very incomplete picture of the program.

Surveillance programs can be adversely affected by poor quality data. In order to be effective, surveillance must detect an event before it would have become apparent without an established surveillance system. It also must do so when the chosen intervention would be more effective than it would have been if applied later.[1]

Foreign animal disease surveillance is designed to detect an introduction of disease and confine it before it spreads to an entire region or throughout the country. It seems obvious that monthly or annual reporting would be an ineffective and insufficient tool for foreign animal disease surveillance. Past surveillance systems have failed because they were built on data that, while otherwise high quality, were not timely enough to be usable.

Data must also be available when needed. For example, a train accident released a large cloud of toxic chlorine gas. Reports of dead animal sightings along Horse Creek were received at the State emergency command post. The veterinary epidemiologist needed to build a map of these sightings in relation to the incident location and the health department's map of the "hot zone." Unfortunately, the health department's only access to their own map data was via a Web application that did not allow them to save or share data layers. As a result, the veterinary epidemiologist had access only to a printed version of the official hot-zone map. Data created and used by one individual or group may be of limited value if not available to other authorized users in appropriately standardized formats.

Disease surveillance depends on having consistent access to high quality data from a number of sources. Unfortunately, for some sources we may have very little control over data quality. In these cases, the best we can do is to note the true level of quality and consider the limitations in our analysis. For other sources, however, we in Veterinary Services do control the production of data. In these cases, we have a chance to apply scientific principles to the improvement of our data quality.

There is a growing body of scientific work on data quality.[2] This work breaks down roughly into two areas. The first category includes methods for measuring various aspects of data quality. Establishing data-quality requirements is often listed as a distinct type of research within this area.[3] A second type of study covers best practices for improving data quality either at the source or in the process of data warehousing.

The next paper in this series will address methods for measuring and recording data quality. The third installment will look at some of the proven methods for accomplishing data-quality improvement. Finally, we will put forward some specific recommendations for improving data quality within Veterinary Services.

[1] D.L. Sacket, et.al., *Clinical Epidemiology; A Basic Science for Clinical Medicine 2nd Ed.,* Boston, Little Brown And Company, 1991, p 153-156.

[2] R.Y. Wang, V.C. Storey, C.P. Firth, "A framework for analysis of data quality research, *IEEE Trans on Knowledge and Data Engineering"*, 7:4, pp. 623-640, Aug 1995.

[3] R.Y. Wang, H. B. Kon, S.E. Madnick, "Data quality requirements analysis and modeling," *Ninth Int Conference on Data Engineering*, Vienna Austria, April 1993, http://www.iqconference.org/documents/publications/TDQMpub/IEEEDEApr93.pdf (Accessed 12/30/04).