White Paper

Data Pedigree – Getting to Know Your Data

Ronald D. Snee, PhDSnee Associates, LLC

In the pharmaceutical world we talk a lot about Data Integrity. The FDA (2016) published A guidance for industry on matters involving data integrity. We hear analysts talking about Data Quality. Legal proceedings often question the “Chain of Custody of the Data.” Metadata is also of concern when assessing Data Integrity (FDA 2016). I conjecture that these issues can be effectively discussed, understood, and assessed by using a common structure called “Data Pedigree,” the subject of this article.

Hoerl and Snee (2019) define data pedigree as:

“Documentation all the origins and history of a data set including its technical meaning background of the process that produced it the original collection of samples measurement process used and the subsequent handling of the data including any modifications or deletions made through the present."
Every data set has a story, a history of the data. The Data Pedigree tells the story, the history.

The core elements of a data pedigree is summarized in Table 1
(Hoerl and Snee 2019).

Table 1. Core Elements of a Data Pedigree

- Basic explanation of the underlying subject matter knowledge of the phenomenon being measured.
- Description of the process that produced data such as manufacturing, healthcare, and finance.
- Description of how the samples were obtained and measured.
- Measurement process is used to assign numbers or attributes to the samples.
- Existence of recent analysis of the said measurement systems such as gauge repeatability and reproducibility studies and calibration studies.
- History of data documenting the chain of custody - who has had access to the original data what - if any changes or deletions were made and access to the original data that can be verified.

Knowing the data pedigree enables the assessment of:

- **"Data Integrity."** Can I trust these data? Are these data really what they are advertised to be? Data integrity is a big issue in the pharmaceutical industry.
- **Data Quality** - Are these data "Fit for Use" in this study? This terminology comes from Juran's definition of quality as "Fitness for Use" (Juran 1989). Quality data are the data that meet the needs of the problem-solving process. Understanding the data pedigree results in a deeper knowledge of the data (Snee and Hoerl 2012).
- **Appropriate Data Analysis Approaches** driven by type of variables and data (quantitative or qualitative), the number of variables, type of model considered, etc.
- **"Chain of Custody"** which is important in legal proceedings; persons and organizations that have been involved in the collection, storage, and distribution of the data.
- **"Metadata"** – Data and information about the data.

When assessing the pedigree of the data, two critical questions to keep in mind are:

- Do I really understand how the data were collected?
- Can I trace back and identify the origin of each data point?

Such an assessment is enhanced using data graphics and visual analysis tools. The creation of process diagrams (schematics) is always helpful in assessing data pedigree and understanding the problem. The data pedigree should be assessed before, during, and after the analysis:

- **BEFORE**: Understand the process, sampling procedure, data collection, analysis preparation and measurement system.
- **DURING**: Constantly check the data and results with the "does this make sense" test aided with extensive use of graphical displays.
- **AFTER**: Evaluate the results to make sure the results and conclusions make sense regarding what is known about the problem being investigated.

It is also important to check the assumptions of the data collection process. For example,

- Are the data from an observational study collected without a well-structured and defined protocol or are the data from a statistically planned experiment or survey? Observational data can be of low quality for several reasons (Hoerl and Snee 2012, 2020).
- Is the randomization process used understood? The method of randomization defines the appropriate model for the data.
- Has the possibility of within-experiment non-homogeneity been evaluated?
- Has the health of the measurement equipment been evaluated? On what frequency?
- Was there a protocol for data collection including sampling, and was it followed?
- Are any data clearly wrong (e.g., grossly atypical values, pregnant males, etc.) or show trends or results that do not make scientific or technical sense?

What should you look for?

Assess data quality. This example discusses a data quality issue. When comparing Carbon Monoxide (CO) data to the Air Quality Standard, it was found that one sampling station the CO second-highest value was 35 ppm with the maximum value of 39 ppm, well above the standard of 9 ppm. Researchers thought it prudent to study the hourly data used to compute the second-highest value (Snee and Pierrard 1977). A plot of the hourly CO values for the period in question showed ten consecutive hourly readings of 39 ppm, with four out of the next six hourly readings at 39 ppm and the remaining two readings at 36 ppm.

This small amount of variation over a 16-hour period is not typical of variation in hourly CO readings and does not represent an accurate characterization of the air quality around the sampler. Snee and Pierrard (1977) conclude that it is highly probable that these data are the result of equipment malfunction. A similar problem was found in the CO data from another location. Salsburg (2017), renowned Pfizer statistician tells us that “lack of variation is often the hallmark of faked data”

Assess the measurement process: When evaluating data quality, you should always think about the measurement process. For example, an improvement project was closed out in the measure phase because it was discovered that the measurement instrument had not been calibrated for two years. After calibration, the product problems completely disappeared (zero defects) and resulted in $157,000 of savings per year in scrap. Case closed in this example.

Understand how the process operates: The initial analysis of an experiment to evaluate a second source of raw material supply produced no significant effects, except a three-factor interaction involving shift, team and time. Knowledge that three factor interactions rarely occur in nature suggests that further analysis is needed. A careful discussion of how the process associated with the data operated discovered that four operating teams conducted a 24/7 three-shift operation. In effect, the shift variable in the model was measuring the time-of-day effect (shift-to-shift variation) and differences among the teams.

When the shift and team effects were added to the model as different variables, the results were better behaved. It was concluded the there was no difference between the two raw material sources, and team four—due to its greater experience—produced yields that were 5% higher than the other teams, which was a significant increase due to the high volume of product produced by the process. This unexpected finding provided a method to increase process yields.

Defective Product Case. This example covers the use of a product that had resulted in the death of a user. The producer did a special study to assess product stability, which was the suspected cause of the problem. The producer submitted 150 test cases to the involved government agency that were claimed to demonstrate that the product was stable.

The analysis of the test data showed that product was not stable over the test period. More importantly, the analysis of the data pedigree uncovered the fact that there were only 122 independent test cases, not 150. There were thirty-eight exact duplicates of the other 122 cases. The final resolution of the problem is proprietary and can't be discussed here. The important point is that the assessment of the data pedigree uncovered the erroneous conclusion submitted by the producer.

Don't Be Fooled by the Appearance

Understanding "Data Pedigree" is critical to success of any data-based study. The pedigree should be assessed throughout the analysis of the data. As President Reagan said in dealing with Russians "Trust but verify". Constant use of graphical displays is an invaluable tool to assess the data.

Always ask yourself, "Do I really understand how the data were collected? Can I trace back and identify the origin of each data point?" A good principle to remember is that data are guilty until proven innocent, not the other way around.

References

Food and Drug Administration (2016) "Data Integrity and Compliance with CGMP - Guidance for Industry", Rockville, MD.

Hoerl, R. W. and R. D. Snee (2019) "Show Me the Pedigree: Part of Evaluating the Quality of Data Includes Analyzing Its Origin and History", Quality Progress, January 2019, 16-23.

Hoerl, R. W and R. D. Snee (2020) Statistical Thinking - Improving Business Performance, 3rd Edition, John Wiley and Sons, Hoboken, NJ.

Juran, J. M. (1989) Juran on Leadership for Quality – An Executive Handbook, Free Press, New York, NY.

Salsburg, D. A. (2017) Errors, Blunders and Lies, CRC Press, Boca Raton, FL.

Snee, R. D. and J. M. Pierrard (1977) "The Annual Average: An Alternative to the Second Highest Value as a Measure of Air Quality", Air Pollution Control Association Journal, Vol. 27, No. 2, 131-133.

Snee, R. D. and R. W. Hoerl (2012) "Inquiry on Pedigree – Do You Know the Quality and Origin of Your Data?" Quality Progress, December 2012, 66-68.

Ronald D. Snee Bio

Ronald D. Snee, PhD is Founder of Snee Associates, LLC, a firm dedicated to the successful implementation of process and organizational improvement initiatives.He provides guidance to pharmaceutical and biotech senior executives in their pursuit of improved business performance that produces bottom line results. He worked at DuPont for 24 years prior to starting his consulting career. He has served as Adjunct Professor in the pharmaceutical programs at Temple and Rutgers Universities.

Ron received his BA from Washington and Jefferson College and MS and PhD degrees from Rutgers University. He is an academician in the International Academy for Quality and Fellow of the American Society of Quality, American Statistical Association, and American Association for the Advancement of Science.

Ron is an ASQ Honorary Member (Hall of Fame) and has been awarded ASQ's Shewhart, Grant and Distinguished Service Medals, and ASA's Deming Lecture, Dixon Consulting and Hahn Quality and Productivity Achievement Awards. He is a frequent speaker and has published 10 books and more than 350 papers in the fields of pharmaceuticals, statistics, quality, performance improvement and management.