# SHOW ME THE PEDIGREE

***Part of evaluating the quality of data includes analyzing its origin and history***

by Roger W. Hoerl and Ronald D. Snee

## Just the Facts

*Data analysis is a cornerstone of quality improvement. However, information about data quality often is overlooked. Statistical analyses based on tainted, unreliable or compromised data sets have undermined the integrity of the entire scientific system.*

*Quality improvement analytics—and scientific inquiry in general—could benefit significantly by adopting the same rigor in data quality as the legal profession has in its concept of chain of custody for evidence, and the U.S. Food and Drug Administration in its concept of data integrity.*

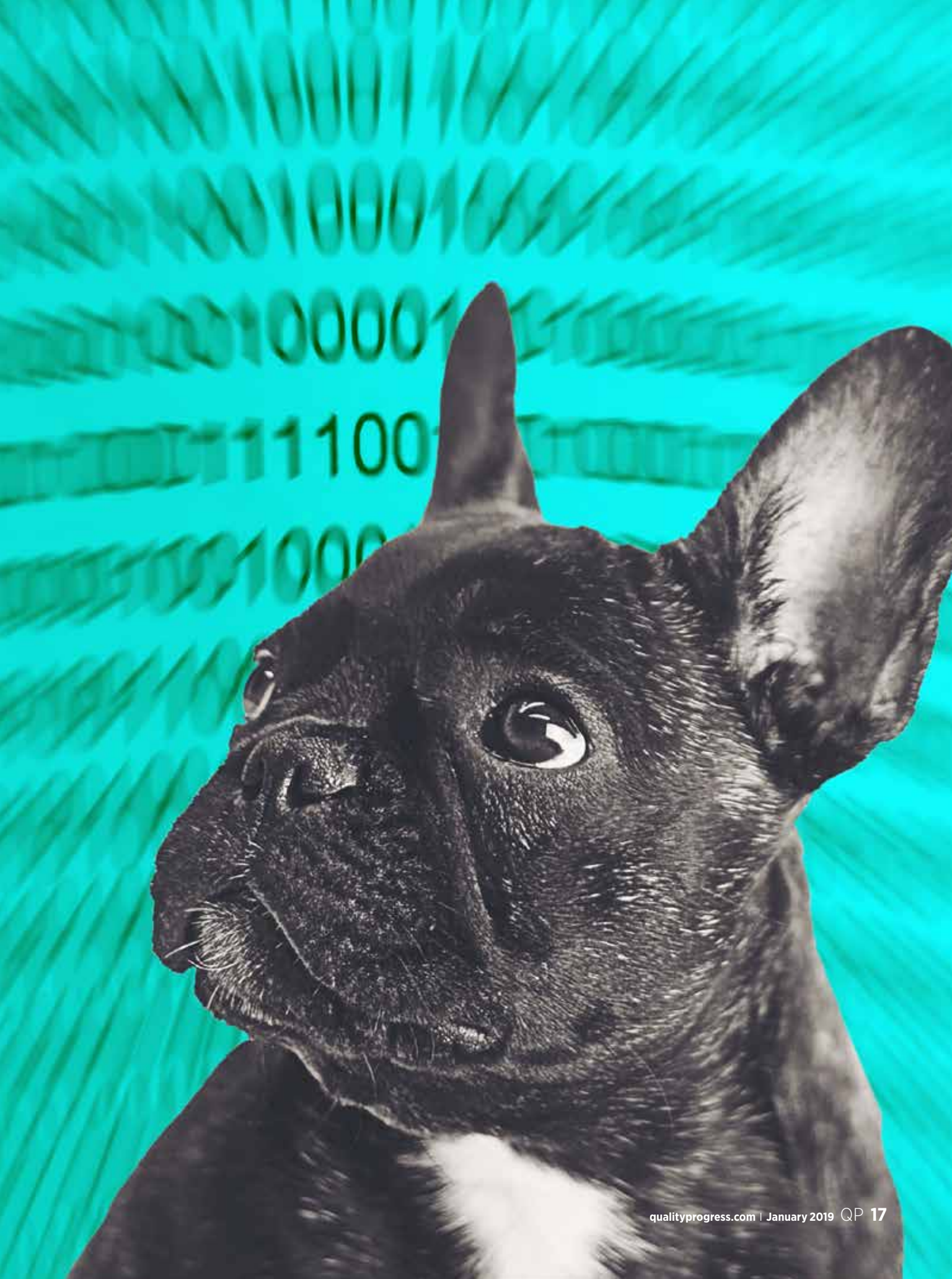
*Looking at the pedigree of the data—including how it was produced, the origin of data samples, and how it was collected and handled—must be part of any evaluation of data or information quality.*

**Data analysis is a cornerstone of quality improvement.** While there's obviously much more to quality than data and statistics, devoid of analytics, quality can devolve into a lot of hype, slogans and fanfare. This explains why quality professionals have effectively applied tools, such as statistical process control, experimental design and regression analysis, for decades.

As data sets have gotten larger and easier to download from the internet, however, information about the quality of the data often is lost, assumed or overlooked. This situation shouldn't be surprising because the clear majority of statistics textbooks present every data set as a "random sample" of impeccable quality. It is, therefore, understandable that some quality professionals would assume that data are "innocent until proven guilty." Too often, they are not even under suspicion.

## What could possibly go wrong?

Beneath the exciting sound bites appearing on Twitter and Instagram about analytics and data science, there's also growing awareness of a dark side of analytics, one in which models frequently



fail with potentially disastrous consequences. For example, many have written about the disaster at the Duke Center for Genomic and Computational Biology. There, four cancer gene signature papers were retracted because the results turned out to be invalid, in part because of discrepancies in the data the Duke researchers analyzed. Unfortunately, it is likely that women died because oncologists used the results of these papers in prescribing treatment to women battling cancer.[1]

Similarly, numerous articles in quality literature have reviewed the space shuttle Challenger disaster on Jan. 28, 1986.[2] That morning, NASA scientists held a conference call to decide whether it was safe to launch the shuttle, given the abnormally low temperatures at Cape Canaveral that day (31º Fahrenheit). The team reviewed available data on the relationship between temperature and O-ring failures. Unfortunately, a scientist had already eliminated data for which there were no O-ring failures, thinking this data was not relevant. This omission led to a decision to launch, when any reasonable analysis of the full data set would have revealed that launching at that temperature would be extremely dangerous. The entire shuttle crew of seven astronauts died.[3]

Even the flagship scientific journal *Nature* is not above publishing faulty research due to questionable data. In 2015, *Nature* retracted a 2014 paper on attitudes about same-sex marriage, in part because of "certain statistical irregularities in the responses" that were analyzed in the paper.[4]

Decisions made based on text data also have gone horribly wrong recently. The Israeli news website Haaretz reported that on Oct. 22, 2017, police arrested an unnamed Palestinian man living in the West Bank for enticing terrorism.[5] The reason for the arrest was that the man had posted "Attack them!" on his Facebook account. After further examination, however, it turned out that the man had actually posted the message "Good morning!" in Arabic. Unfortunately for him, Facebook's automated algorithms translated "Good morning!" in Arabic to "Attack them!" in Hebrew. Haaretz reported that the man subsequently canceled his Facebook account.

Another example illustrating the potential quality disasters that can result from questionable text data is the recent article in the journal *Science,* in which researchers from the Massachusetts Institute of Technology analyzed more than 4.5 million tweets on 126,000 topics. They ultimately concluded: "Falsehood diffused significantly farther, faster, deeper and more broadly than the truth in all categories of information." In fact, the spread of false information reached "critical mass" (defined as reaching 1,500 people) six times faster than true information.[6]

Are these isolated incidents unlikely to recur? Or perhaps are they growing evidence that large amounts of text or data—combined with sophisticated algorithms—do not guarantee true quality improvement? We argue that these incidents reflect one root cause of the ongoing "reproducibility crisis" in science.[7] While much of the effort to address this crisis has focused on standardizing analyses and providing original source code used by authors, a major source of problems has been overlooked: data quality.

**Data quality often has been overlooked in textbooks because it is not easy to mathematize.**

## Garbage in—garbage out

Data quality often has been overlooked in textbooks because it is not easy to mathematize. Conversely, the impact of data quantity is easy to mathematize. This is why most statistics textbooks incorporate formulas for sample sizes needed to estimate parameters with a desired degree of precision, as well as "power curves" showing how large a sample size must be to detect an effect of a given magnitude with a specific probability. But what if the original data are faulty? How useful are these formulas and power curves then? Unfortunately, some quality professionals have no formal training in data quality, hence they are likely to overlook this potential root cause.

Author Richard D. De Veaux and others provide further examples of how poor data quality limits the effectiveness of big data studies.[8] Authors Ron S. Kenett and Galit Shmueli build on this by

defining the concept of information quality (InfoQ) as the potential of a data set to achieve a specific (scientific or practical) goal using a given empirical analysis method.[9] That is, the InfoQ depends on the goals of the analysis, and the specific methods employed. Later, Kenett and Shmueli note the critical role of data and information quality in addressing the reproducibility crisis. In addition, they provide a formal, quantitative framework for evaluating information quality.[10]

Note that InfoQ depends on the specific goal of the analysis intended. Therefore, it is not an inherent attribute of the data set itself. We argue that a related but distinct concept, data pedigree,[11] which is an inherent attribute of the data set, is needed to accurately evaluate data and information quality. Before defining data pedigree, let's first review a related concept from the legal profession.

## Legal chain of custody

Interestingly, if we benchmark the legal profession, it seems to be way ahead of the quality and scientific communities when it comes to understanding the importance of data quality. In legal circles, data quality means documenting the integrity of evidence. The legal profession uses the term "chain of custody" for evidence, as opposed to data quality, but of course evidence is the "data" analyzed in a courtroom.

Basically, the chain of custody refers to documentation of how the evidence was originally obtained (for example, legal vs. illegal search), and its movement and location from that point on until it's presented in court. The Legal Dictionary states the following concerning the importance of documenting the chain of custody for legal evidence:

"Proving chain of custody is necessary to 'lay a foundation' for evidence in question by showing the absence of alteration, substitution or change of condition."[12]

Documenting the origin of data to ensure freedom from alteration, substitution, or change in condition is equally required to lay a foundation for any statistical analysis. In the Duke Genomics case, for example, *New York Times* reporter Gina Kolata noted that when statisticians began evaluating the published analysis, they "found errors almost immediately. Some seemed careless—moving a row or column over by one in a giant spreadsheet—while others seemed inexplicable. The Duke team shrugged them off as 'clerical errors.'"[13]

Similarly, in the Challenger disaster, the fundamental issue was not with the original data collected, but rather that someone decided to eliminate data without O-ring failures—that is, there was an alteration to the data set that ended up costing the lives of seven people.

In our experience, many organizations keep important data sets in Excel spreadsheets that are subsequently emailed around the organization, with no or minimal documentation of

TABLE 1

### Core elements of a data pedigree

- **What the data represent—that is, a basic explanation of the underlying subject matter knowledge of the phenomenon being measured, including units of measurement.**
- **Description of the process that produced the data, such as a financial process, healthcare process or manufacturing process.**
- **Description of how the samples were obtained from this process that were subsequently measured.**
- **The specific measurement process used to assign numbers or attributes to the samples.**
- **The existence (or lack) of recent analyses of the said measurement system, such as gage repeatability and reproducibility studies and calibration studies.**
- **The history of the data, documenting the chain of custody—who has had access to the original data, what if any changes or deletions have been made—and access to the "gold standard"—that is, access to a copy of the original data that can be verified.**

who has made changes to the spreadsheet, what these entailed or when they were made. Rarely is a "gold standard" of the original data available.

The Legal Dictionary continues: "Court-rendered judgments and jury verdicts that are based on tainted, unreliable or compromised evidence would undermine the integrity of the entire legal system ..."[14] Statistical analyses based on "tainted, unreliable or compromised" data sets have, in fact, undermined the integrity of the entire scientific system. In science, we refer to this undermined integrity as "the reproducibility crisis."

The U.S. Food and Drug Administration (FDA) uses a similar concept—data integrity— which it defines as the completeness, consistency and accuracy of data. The FDA states: "Complete, consistent and accurate data should be attributable, legible, contemporaneously recorded, original or a true copy, and accurate (ALCOA)."[15] These concepts bare obvious similarities to the concepts of chain of custody, information quality and existence of a gold standard.

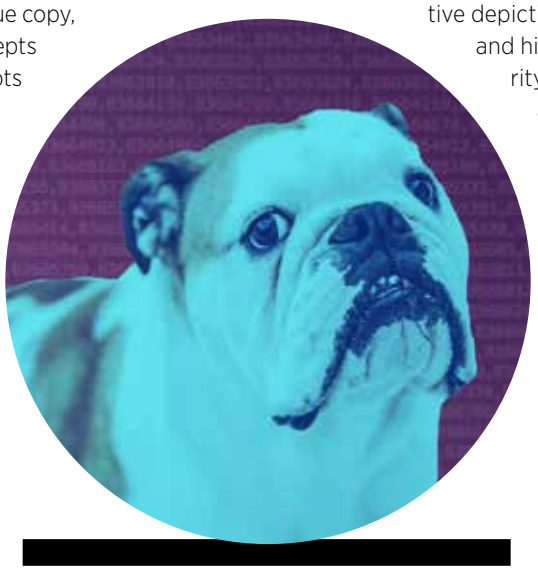
Quality improvement analytics—and scientific inquiry in general—could benefit significantly by adopting the same rigor in data quality as the legal profession has in its concept of chain of custody for evidence, and the FDA in its concept of data integrity. We refer to such an approach as documenting the data pedigree,[16] borrowing the term "pedigree" from animal husbandry—that is, show dogs and race horses, for example. Obviously, the value of a yearling racehorse depends significantly on the quality of its pedigree.

## Data pedigree

We define data pedigree as: "documentation of the origins and history of a data set, including its technical meaning, background on the process that produced it, the original collection of samples, measurement processes used, and the subsequent handling of the data, including any modifications or deletions made, through the present." A proper data pedigree should include each of the elements listed in Table 1, p. 19.

We understand that some may consider such documentation as unnecessary bureaucracy that will stifle quality improvement. We feel, however, that relatives of those who died from the Challenger disaster or from misguided clinical trials based on papers from the Duke Genomics Center would disagree. Had a complete data pedigree been available in these cases, lives would certainly have been saved.

We acknowledge, however, that common sense must be applied to the data pedigree. If someone is gathering data to improve his or her golf game, obviously the degree of rigor required in documenting the data pedigree is minimal. In medical research or when public safety is involved, however, the pedigree should be rigorous to enable evaluation of whether it meets the high standards of the FDA definition of data integrity, for example. Note that the data pedigree captures the metadata, or "data about data"[17] concerning a particular data set, no matter how good or bad it is. That is, data pedigree is not a standard, but rather an objective depiction of the meaning, condition and history of the data. Data integrity, on the other hand, defines a standard for acceptance by the FDA.



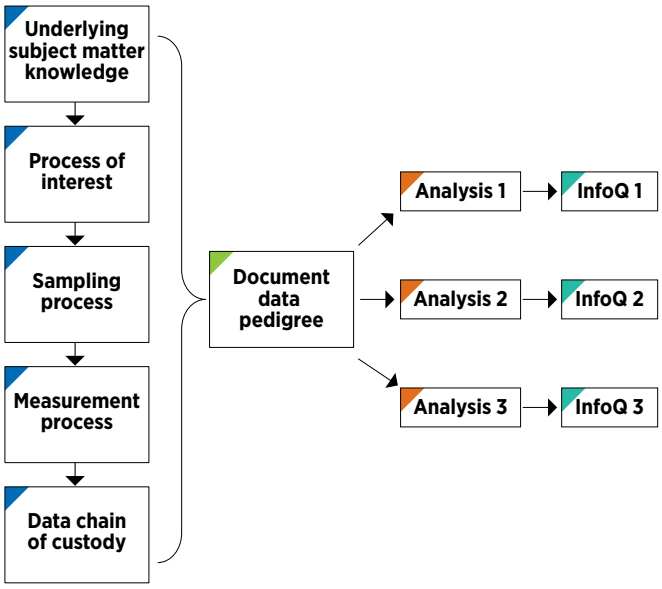
**Quality improvement analytics—and scientific inquiry in general—could benefit significantly by adopting the same rigor in data quality as the legal profession has in its concept of chain of custody for evidence, and the FDA in its concept of data integrity.**

The first element of the data pedigree is an explanation of what the data represent—that is, the underlying subject matter knowledge involved, including units of measurement. If the data involve measurements of color using the CIELAB color space,[18] for example, it is naïve to think that someone could analyze this data and properly interpret the results without understanding what L*, a* and b* represent in this color space.

Even such mundane things such as units of measurement are important. Sadly, in 1999, NASA lost a $125 million Mars orbiter when two teams working together on the project used

FIGURE 1

## Role of data pedigree in evaluation of InfoQ




***InfoQ*** = *information quality*

different units of measurement: one metric and the other English.[19]

The next element is a description of the process that produced the data. There is an old saying in statistics: "You can't understand the data unless you first understand the process that produced it." This not only refers to the measurement process, but also understanding the process that produced the samples subsequently measured. Process understanding, perhaps obtained through a supplier, input, process, output, customer diagram,[20] provides the context that enables you to draw actionable conclusions from data analyses.

For example, suppose you are handed data on the viscosity of an industrial chemical. You certainly could begin analyzing the numbers without knowing anything about the chemical process involved. If the specific chemical in question ages rapidly, however, it would be difficult to draw actionable conclusions without knowing how old the samples were. Of course, without knowing anything about the chemical process, you would not even know that this was a relevant question to ask. We argue that the old saying about understanding the process is correct. Before crunching numbers, we shiould first develop at least a fundamental understanding of the process that produced the numbers.

Next, it is important to document how the samples that were measured were originally selected. In textbooks, it is easy to state that a sample was randomly selected. But in practice, random sampling is an ideal that is rarely accomplished. In evaluating an election poll, for example, were registered votes sampled, likely voters, or some other group? If they were likely voters, how was a "likely voter" defined? What timeframe was used? How were voters who refused to answer questions handled? We argue that the devil is indeed in the details. Without understanding the sampling approach, it is impossible to know how broadly the results of the analysis might be inferred.

Measurement system evaluation is a traditional strength of the quality profession, and is critical in evaluating data pedigree. We have, unfortunately, seen numerous data sets presented without explanation of how the measurements actually were made. There is a reason that documentation of the measurement system is a key element of ISO 9000 standards.

In addition to documenting the measurement system, it is important to know if, how and when this measurement system has been formally evaluated or calibrated. Remember that one of the phases of a lean Six Sigma project is the measure step, which typically involves formal evaluation of the measurement system—for good reason. Similarly, you could argue that the core purpose of the accounting profession is to ensure accuracy and consistency of financial statements (measurements). Tainted, unreliable or

compromised measurement systems are, unfortunately, all too common.

Lastly, we feel it is important to document the chain of custody (history) of the data set, in terms of who has had access to the data and could have made modifications, including eliminating data points. Ideally, a gold standard of the original data set should be maintained—that is, what the FDA calls an original or true copy.

To understand why, consider the case of economists Carmen Reinhart and Kenneth Rogoff. In 2010, they published research on a large data set including 44 countries and spanning more than 200 years of history.[21] Their analysis demonstrated a negative growth rate for countries with a high debt to gross domestic product (GDP) ratio, which had obvious implications for economic policy. However, another set of economists sought to replicate these results, but could not.[22]

After further investigation, author Thomas Herndon and his fellow authors determined: "We ... find that coding errors, selective exclusion of available data, and unconventional weighting ... led to serious errors ... Our finding is that when properly calculated, the average real GDP growth rate for countries carrying a public-debt-to-GDP ratio of 90% is actually 2.2%, not -0.1%, as published in Reinhart and Rogoff."[23] In other words, when the data issues were addressed, the conclusions were exactly opposite of those originally published—that is, that high debt leads to increased growth, not decreased growth.

## Data pedigree and data/information quality

We already have noted that data pedigree and the FDA concept of data integrity are related: One is a standard (data integrity), while the other is an objective depiction of the data and its history (pedigree).

We also see a synergistic relationship between data pedigree and InfoQ. As explained by authors Kenett and Shmueli, InfoQ is relative—that is, it is specific to a given purpose and analysis.[24] The same data set could be considered high quality for one analysis, but low quality for another.

**Know that the data pedigree provides additional important information: insight regarding how to analyze the data set.**

Conversely, the data pedigree is absolute; it documents the origins, history and current status of a given data set, regardless of how you plan to use it.

Of course, both concepts are important and relevant. No data set is perfect; each has its own limitations. We suggest that data pedigree should be the basis of any evaluation of data quality or InfoQ. That is, we have no solid foundation on which to evaluate data or information quality if we have not first documented the data pedigree. After we have done so, this documentation can be applied to any formal evaluation of InfoQ for a given analysis that might be performed in the future. This relationship is illustrated in Figure 1 (p. 21).

## Enabling effective data analysis

Know that the data pedigree provides additional important information: insight regarding how to analyze the data set. The core elements of a data pedigree, summarized in Table 1, help identify the sources of variation in a data set. The sources of variation define the appropriate models that could be used to analyze the data.

Only after you know the potential sources of variation in a data set can you effectively create an appropriate model. Statistical tools—such as analysis of variance, regression and multivariate analysis—are all based on knowing the potential sources of variation. Some sources of variation may not be of interest, and therefore are nuisance variables. However, ignoring them in modeling is likely to produce bad results.

Knowledge of the data pedigree also helps you determine whether the data set is adequate for answering the question being asked in the first place. In some cases, unfortunately, it won't be, and time can be saved by not developing a model that will ultimately be inadequate for the questions of interest.

## Heading off failures

The quality profession has long understood the importance of measurement system evaluation. However, this is only one aspect of a data pedigree. Similarly, we are witnessing a growing body of failures of analytics in quality and scientific literature. Some of these blunders are humorous, but others are deadly. Data quality and InfoQ have been identified in the literature as one cause of these problems, and rightly so. The FDA, among many other federal agencies, has recognized the critical importance of this issue.

To properly evaluate data quality or InfoQ for a specific analysis, or determine the data integrity, we first must document the pedigree of the data. Those teaching quality improvement or statistical methods should emphasize the importance of documentation of the data pedigree before conducting formal data analyses. Similarly, if professional journals and agencies offering research grants were insistent on formal documentation of data pedigree, quality improvement studies would be much more reproducible. As a result, quality blunders would be significantly reduced. QP

**Ronald D. Snee** is president of Snee Associates LLC in Newark, DE. He has a doctorate in applied and mathematical statistics from Rutgers University in New Brunswick, NJ. Snee is an honorary member of ASQ and has received ASQ's Shewhart, Grant and Distinguished Service Medals. He is an ASQ fellow and an academician in the International Academy for Quality.

**Roger W. Hoerl** is a Brate-Peschel associate professor of statistics at Union College in Schenectady, NY. He has a doctorate in applied statistics from the University of Delaware in Newark. Hoerl is an ASQ fellow, a recipient of the ASQ's Shewhart Medal and Brumbaugh Award, and an academician in the International Academy for Quality.

**REFERENCES**

1. Gina Kolata, "How Bright Promise in Cancer Testing Fell Apart," *New York Times,* July 7, 2011, www.nytimes.com/2011/07/08/health/research/08genes.html.
2. Siddhartha R. Dalal, Edward B. Fowlkes and Bruce Hoadley, "Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure," *Journal of the American Statistical Association,* Vol. 84, No. 40, pp. 945-957.
3. Necip Doganaksoy, Gerald J. Hahn and William Q. Meeker, "Assuring Product Reliability and Safety," *Statistics, A Guide to the Unknown,* fourth edition, Duxbury Press, 2006.
4. John Bohannon, "Science Retracts Gay Marriage Paper Without Agreement of Lead Author LaCour," *Science*, May 28, 2015, www.sciencemag.org/news/2015/05/science-retracts-gay-marriage-paper-without-agreement-lead-author-lacour.
5. Yotam Berger, "Israel Arrests Palestinian Because Facebook Translated 'Good Morning' to 'Attack Them,'" *Haaretz*, Oct. 22, 2017, www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427.
6. Soroush Vosoughi, Deb Roy and Sinan Aral, "The Spread of True and False News Online," *Science*, Vol. 359, No. 6,380, March 9, 2018 pp. 1,146-1,151.
7. Darrel Ince, "The Problem of Reproducibility," *Chance*, Vol. 25, 2012, pp. 4-7.
8. Richard D. De Veaux, Roger W. Hoerl and Ronald D. Snee, "Big Data and the Missing Links," *Statistical Analysis and Data Mining*, Vol. 9, No. 6, Aug. 17, 2016, pp. 411-416.
9. Ron S. Kenett and Galit Shmueli, "On Information Quality," *Journal of the Royal Statistical Society,* Series A, Vol. 177, 2014, pp. 3-38.
10. Ron S. Kenett and Galit Shmueli, *Information Quality: The Potential of Data and Analytics to Generate Knowledge,* John Wiley and Sons, 2017.
11. Roger W. Hoerl and Ronald D. Snee, *Statistical Thinking: Improving Business Performance*, second edition, John Wiley and Sons, 2012.
12. The Free Dictionary, "Chain of custody," https://legal-dictionary.thefreedictionary.com/chain+of+custody.
13. Kolata, "How Bright Promise in Cancer Testing Fell Apart," see reference 1.
14. The Free Dictionary, "Chain of custody," see reference 12.
15. U.S. Food and Drug Administration (FDA), "Data Integrity and Compliance With CGMP: Guidance for Industry," April 2016, www.fda.gov/downloads/drugs/guidances/ucm495891.pdf.
16. Ronald D. Snee and Roger W. Hoerl, "Statistics Roundtable: Inquiry on Pedigree," *Quality Progress,* December 2012, pp. 66-68.
17. FDA, "Data Integrity and Compliance With CGMP: Guidance for Industry," see reference 15.
18. Wikipedia, "CIELAB color space," https://en.wikipedia.org/wiki/CIELAB_color_space.
19. Robin Lloyd, "Metric Mishap Caused Loss of NASA Orbiter," CNN, Sept. 30, 1999, http://edition.cnn.com/TECH/space/9909/30/mars.metric.02.
20. Hoerl and Snee, *Statistical Thinking: Improving Business Performance*, see reference 11.
21. Carmen M. Reinhart and Kenneth S. Rogoff, "Growth in Time of Debt," *American Economic Review*, Vol. 100, No. 2, 2010, pp. 573–578.
22. Thomas Herndon, Michael Ash and Robert Pollin, "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff," *Working Paper Series 322*, Political Commentary Research Institute, 2013.
23. Ibid.
24. Kenett and Shmueli, "On Information Quality," see reference 9.