Data Quality—Why Hygiene Matters

WHEN IT COMES TO ACTUARIAL WORK (or analytics work in general), few parts of the process are as important as making sure that the data used is of good quality. While the data processing part of a project can often take the most time, few parts of the process are seen as less glamorous or enjoyable than ensuring the quality of the data. (Admit it, though—there is something very satisfying about seeing totals balance...)

There is considerable attention being paid to data these days, including important discussions surrounding data privacy and data ownership (see sidebar, “Data Privacy,” page 32). But what about the quality of data being protected? Who is responsible, are the appropriate amount and types of resources being devoted, and what are the right processes to ensure that data is high quality?

Work focused on data quality is sometimes referred to as “data hygiene.” Data quality is often the forgotten part of the Big Data equation. It may seem obvious, but major issues might result if data being used for increasingly sophisticated algorithms and important purposes is not of good quality.

Data is the foundation on which all analytics is built. The very best analyst may apply the most advanced techniques that technology can support, but that analyst could never overcome the issue of bad source data. Bad data might not only lead to less-than-optimal analytics results, but also it can lead to other—and possibly more severe—consequences, such as poor servicing of customers or reputational damage.

The importance of data is illustrated in Figure 1, which we’ll refer to as the Modeling Lifecycle. The process flow can apply more broadly to general actuarial or other analytical work.

We start with the business question: “What am I trying to accomplish with my analysis?” Only then does the lifecycle move into data. Data is the raw material that fuels the analysis. The entire process stops without data. The end goal, of course, is to transform that raw material into something useful—to turn data into actionable insights.

No data is perfect, so with any analytics project, considerable effort is made to ensure that data is as useful and accurate as reasonably possible. Ideally, companies should take steps—before any analytics work is initiated—to provide for high-quality data. We’ll explore these aspects as we dig deeper into data stores, metadata, and data sets themselves.

Data Quality: Not Just a Good Idea

Based on reports from Forbes and other sources, revenues from Big Data analytics are growing by double digits annually and nearing $200 billion worldwide. The emphasis on and importance of data quality is clear. Companies strive to attain the proper value from their Big Data investment. From an actuarial standpoint, there is another solid reason to focus on quality of data. Actuarial Standard of Practice No. 23 (ASOP No. 23), Data Quality, provides guidance on what an actuary should consider during a project.

In short, ASOP No. 23 helps define what types of reviews the actuary should perform on data. In addition, ASOP No. 23 lays out what should be done in cases of potentially significant limitations of data required for analysis. Importantly, it also addresses steps that the actuary is not required to perform in the course of analysis.

Section 3.1 of ASOP No. 23 includes guidance for the actuary on how to use available data to successfully perform the analysis being undertaken. This guidance requires that the actuary clearly understand the business question; think back to the first step of the Modeling Lifecycle. If the actuary is aware of significant limitations of the data, he or she has a duty to disclose those limitations.

Section 3.2 addresses several actuarial considerations when reviewing the data for a project. Among them:

■ Whether the data is appropriate and sufficiently current for the work;
■ Whether the data is reasonable and consistent both with internal and external information the actuary is aware of;
■ Whether the data is sufficient for the project; and
■ Whether there is any alternative data that can be reasonably obtained and that would provide sufficient benefit to the work.

ASOP No. 23, Section 3.3 acknowledges that a review of the data may not uncover issues but stipulates the review should be performed unless professional judgment suggests otherwise. If a review of the data is deemed necessary and is undertaken, some of the review steps should include:

■ Working to identify the definition of each data element (see commentary on metadata)
■ Working to identify questionable or inconsistent data values, and possibly
taking steps to enhance the quality of the data—or at a minimum, documenting the significant impacts the lack of data quality would have on the analysis.

A review can also include comparative data from prior analyses for consistency, if possible.

Importantly, ASOP No. 23, section 3.8 discusses steps that an actuary is not required to take. These include:

■ Determining whether the data is falsified or intentionally misleading;
■ Acquiring other data just to search for issues in the data; and
■ Performing an audit on the data.

Certainly, actuarial standards have a lot to say about the importance of data quality and the steps actuaries need to consider in their work. But it is important to remember those data review steps not required by ASOP No. 23. Several of the steps in Section 3.8 are good to consider, both from a business practice standpoint as well as inclusion in any strong data management/data quality program.

**Data Quality Best Practices**

Actuaries can consider a number of best practices to ensure good data quality. As noted, there are several reasons why investing in data quality is important from organizational and actuarial perspectives.

Often, companies have multiple data stores built over many years (or decades) housed in multiple technologies—on premises and/or in the cloud. Unfortunately, this makes it easy to have conflicting data that results in bad analysis, lost time, and an inability to serve customers in an optimal manner. It may also result in wasted resources if a company cannot connect cross-functional data. Big Data only works if you can use it.

Consider the following when establishing a strong data management program:

■ Get support at top levels of the company—and especially from the business. There are several important tactical steps that need to occur to ensure better data quality. One of the best ways is to make certain that a data management program has buy-in and support from key stakeholders. It is important for all stakeholders to understand the value for them and for the organization (and alternatively, what horrors happen when data quality is bad!). It is also helpful to set expectations. Implementing steps for data management is not quick, easy, or cheap, so if there is not an appropriate level of initial support from top company leadership, the data quality exercise is most likely doomed for failure before it starts.

■ Share the word. After getting key top-level support, share the reasons for and the importance of the new (or enhancement of existing) data management program—broadly. This includes sharing information across multiple levels and functions of an organization, including those who will be heavily engaged in the work.

■ Develop a solid data governance framework. Successful execution of the “how” of data management includes clear definition of roles. Ideally, both IT and the business should be represented in the data governance structure to provide broader buy-in as well as
allow for diverse and knowledgeable perspectives. However, it must be clear, structurally, who is the owner (or better, “steward”) of certain aspects of data. For example, who decides how a data element is defined, where it is stored, or who has access to it? While it seems like a simple step, if roles are not clear and decision rights not defined, the data management program will flounder.

- **Assess the current state.** Before you know what needs to change within the data organization, you need to assess the quality of the current data environment. To do that, a data manager needs to determine thresholds of quality and whether those thresholds should vary by data element. For example, will the organization hold financial or statistical data to a higher standard than data that is used for more ad hoc research? Data will never be perfect, and it can be cost-prohibitive and cause a sort of organizational paralysis if there is a belief that it can be. Establish a clear plan of what data will be assessed, in what order, and with which metrics.

  - **Be willing to invest in metadata.** Defined as “data about the data,” metadata is an often overlooked yet critical component of data management. Good metadata can provide significant value and return on investment. Depending on the starting point, however, it may be a substantial undertaking. Often, metadata is stored within code, scattered across process guides and manuals … or residing in the heads of longtime employees. With a concerted effort to concentrate, consolidate, and develop good metadata, both power users and casual users of data will be able to better understand:
    - How fields are defined
    - Where fields are located

  - **How data connects**

    The good news is that many tools exist today to help with the task of developing and maintaining metadata. Once the initial, foundational work of creating a data management program is complete, there are several ways to ensure continued data quality for data stores:

    - Clearly define allowed values for data fields and program the system to only accept values that comply.
    - Perform periodic audits of data, including thorough documentation of and steps taken to perform root-cause analysis of uncovered issues.
    - Strive to fix data issues as close to the entry point (source) as possible to reduce data error proliferation and/or error corrections needed in other, multiple places.
    - Automate data entry as much as possible. For quality and consistency,

---

**DATA PRIVACY**

Data privacy, ownership, and questions over how different organizations collect, maintain, and utilize an individual’s data is a hot topic among consumers, lawmakers, and companies. It seems every few days the media announces a new data breach, often involving large, well-known companies and brands.

In addition, questions about data, from what is being stored to how it is being used by corporations, continue to draw attention. The increasing volume of data produced by vehicles, for example, creates major questions about who owns vehicle, driver, and related driving data—the automaker, the insurer or other third-party recipients, or the consumer?

In 2016, legislators in the European Union passed the General Data Protection Regulation (GDPR), effective in 2018, which has several provisions regarding handling of personal data. One such provision is the “right to be forgotten,” which allows individuals to have search engines remove results linked to their names in certain circumstances. In a September 2019 ruling, Google was required to comply with the rule only in EU countries, not worldwide. The case underscored the complexity of compliance in an increasingly intertwined, global economy.

In 2018, California passed the California Consumer Privacy Act (CCPA), which goes into effect in January 2020. Among other protections, the CCPA allows for California consumers in certain situations to be informed about the personal information that companies collect about them. The CCPA also gives consumers the right to ask that their information be deleted if certain criteria are not met, prevent companies from selling or disclosing their personal information, and pursue a limited right of action against firms that have a breach of their personal data. Companies have been working to comply with the law while debate and discussion about CCPA continues. The group partially behind the passage of...
consider using third-party vs. agent- or customer-entered data. There are an increasing number of third-party data sources across multiple domains in existence today.

- Perform balancing and reconciliation of data, like balancing and reconciliation to financial statements.
- Create a data quality “dashboard” with key metrics like accuracy, completeness, and timeliness. Share results of the dashboard and actions taken to improve results broadly to continue to build support for the data management work and program.
- Take steps to periodically refresh data. One of the most overlooked ways that data goes bad is that it gets stale.
- Require strict controls on who has “change” access to databases or code and include a rigorous change documentation process.
- Ensure that a company has highly available and regularly monitored systems so that when issues occur, they are quickly rectified.

Finally, the following are some ways to ensure data quality for reviews on datasets being used for a given analysis.

- Assess how the data compares year-to-year.
- Consider how data compares to other, related sources of data (such as industry aggregations of data).
- Review how it compares to data used in prior analyses.
- Look for anomalies in the data (e.g., duplicated data or portions of data deleted unexpectedly).
- Look for outliers in the data (such as unusually large or small values).
- Examine for values of the data that are not expected (for example, having character values in a data field that is expected to be numeric).
- Keep track of totals (records, counts, etc.) during the steps of the analysis to help ensure that data does not get dropped or duplicated during processing work.

Importantly, actuaries can also use their deep subject matter expertise to look for reasonableness of the data.

Data quality is just one step in the Modeling Lifecycle, and it may lack glamor and glory. But data quality serves a critical role in the success or failure of analytics (Big Data or otherwise). The integrity of the analysis, the importance of the results, the quality of the product, and ultimately, the ability for a company to successfully compete—now and into the future—relies on it. Companies and actuaries cannot afford the cost of overlooking the necessary investment and return in solid data management.

GREG FRANKOWIAK, MAAA, FCAS, is a senior consulting actuary at Pinnacle Actuarial Resources in Bloomington, Illinois.

the CCPA has been working on legislation that would provide additional consumer rights and privacy obligations on affected companies.

In 2019, the U.S. Congress also worked on comprehensive data privacy and protection in addition to other bills that addressed consumer data privacy and protection in part. As of this writing, it does not appear that any federal comprehensive data privacy and protection legislation will pass in 2019. The absence means that companies doing business across state lines will still have to understand and comply with varying data privacy legislation across jurisdictions. Many states, including Illinois, New York, Pennsylvania, Texas, and Washington are looking at their own potential comprehensive privacy laws or are expanding existing consumer data privacy laws already on the books. Maine and Nevada passed such laws earlier in 2019. Legislative focus on this important topic does not appear to be abating anytime soon.