Supply Chain Resource Cooperative

How important is data cleansing in spend management?

Robert Handfield, Ph.D.

I recently entered into an interesting debate with a software provider who was developing software for the healthcare industry. (See the previous post and white paper on spend management in healthcare for a complete report). This individual made the following statement:

We don't feel data cleansing is as valuable an exercise as data normalization, which is what we do. With data normalization, the system does not have to cleanse the data to make effective use of the resulting analysis.

Every time a new item is brought into the ERP, our product index layer recognizes it and enriches it with the correct manufacturer name and item number, UNSPSC code and a standardized product description. Having the correct manufacturer name and item number ensures a high match rate as the system synchronizes items between the contract file and item file. This ultimately makes for **accurate price updates** when contract managers push new contract pricing back into the ERP system. This process also helps users take advantage of new contract pricing immediately when it takes effect.

I took issue with this point of view, as I believe it is a subjective statement. Our white paper is an attempt to clarify exactly why data cleansing is so important. If data normalization was acceptable, healthcare would not be adopting GS1 standards whereby manufacturers must publish data with a "warranty" of accuracy. Clearly, healthcare values accurate and clean data for an effective supply chain, including downstream analytics. The old saying "Garbage in, Garbage Out" still applies.

Data Cleansing

Access to the right data is essential, as accurate and properly coded data provides the foundation for category management strategies, including leveraging, pricing agreements, quantity discounts, value analysis, supply base optimization and other important cost management activities. Data Cleansing is actually a process that involves four stages.

Data Acquisition

First, the user is contacted and the "raw" data is collected from different sources. Common sources of data can be the customers' MMIS, GPO and local suppliers. It is important at this stage that *all relevant spend data, including indirect spend*, is included in the analysis. Note that many providers restrict their data acquisition to only electronic EDI data, or inventory data that is readily available, thereby missing a significant "chunk" of the total spend. The net impact of this oversight is that it provides an inaccurate representation of what the healthcare system is truly spending on third party goods and services

Data Cleansing

Busch notes that from a technical perspective, first generation-spend analysis approaches were limited by the underlying architecture, development, analytical and visualization capabilities available to providers at the time. This is still a major problem for healthcare providers. The limits of relational database technology based on disk storage and traditional data warehousing approaches to storing, querying and accessing information and reports are one example of the constraints that are often encountered, due to old technology platforms. A few healthcare systems are now beginning to invest in the usage of in-memory databases that rely on main memory (or RAM, as it's better known) storage approaches that can materially increase query speeds as well as workarounds to traditional storage and query models that greatly increase both the speed with which we can search and access information as well as the ability to search information sets in the context of each other.

One of the most important challenges in healthcare is that the data coming from manufacturers/suppliers of healthcare supplies is flawed even before it reaches the hospital's analytics team! A recent study by the Department of Defense[1] conducted significant analyses of item data collected from various DoD suppliers, and found significant data disconnects between Healthcare industry trading partners. This poor connectivity included poor data accuracy between manufacturers, distributors and DoD's own internal pricing/contract management applications. Further, the study found that the process of requesting "one-off" data feeds from partners was a significant resource burden on all parties involved. Up to 20% of manufacturer data has errors that are transmitted to distributors and other third parties, with further data errors occurring in other parts of the channel as well. What this means is that much of the data that is already assumed to be "clean" that is imported into databases for spend analysis is already rife with error!!

Data Classification/ Preparation

One of the most important foundational shifts in spend analysis technology in the past 18 months has been an interest in greater flexibility and visibility into the classification process. Increasingly, more advanced organizations are starting to look for the ability to classify spend to one or more taxonomies at the same time (e.g., customized UNSPSC and ERP materials code) as well as having the ability to reclassify spend to analyze differ views and cuts of the data based on functional roles and objectives. Moreover, some organizations are looking to exert greater control over the spend visibility process; these individuals are often becoming distrustful of "black box" approaches to gathering and analyzing spend data. Coding of data is essential when conducting category analyses and clinical effectiveness studies. For example, a hospital wanting to gain strategic advantage through public reporting on clinical excellence will require an understanding of the impact of products on reducing hospital-acquired infections and contributing to the total episode of care and a preference for "smart products".[2] An example is pumps that provide feedback on accuracy of dosage delivery. Category analysis using data classification codes can also identify areas where "system internal co-sourcing" is taking place. This refers to situations where decisions regarding commodity items as well as physician preference items and actual determination of vendors where there continues to be a duplication of purchasing efforts at both hospital and system levels. This is an expensive proposition, which includes duplication of effort for identifying products and suppliers, developing and managing requests for proposal and information, optimizing proposals and obtaining offers, finalizing awards, and implementing and monitoring contracts. As systems migrate from being holding companies to operating companies, reduction of internal co-sourcing is an important strategic opportunity, but will rely on effective data cleansing and coding as the basis for analysis and action.

It is important to note here that some providers we spoke with believed that data cleansing is not as valuable as data normalization. The point was made that "normalization does not have to cleans the data to make effective use of the resulting analysis. We disagree with this point for several reasons. First, if data normalization was acceptable without cleansing, healthcare would not be adopting GS1 standards, to address the issue of manufacturers publishing data with a "warranty" of accuracy. Accurate and clean data is critical for any type of analytics or normalization effort. In this case, if the "garbage" goes in, than the resulting output is more likely to be "garbage" as well!

Database Population

Finally, the coded dataset is uploaded into the requisite application. Once uploaded, the real power of the data can be leveraged through merging with other data forms for benchmarking and cross-reference analyses. Application and data integration paradigms have already shifted in a number of non-healthcare applications from one of batch uploads from multiple source systems to real-time data queries that can search hundreds (or more) disparate sources while normalizing, classifying and cleansing information at the point of query. In the coming years, Oracle, IBM and D&B, all of which have purchased customer data integration (CDI) vendors, could begin to apply these techniques to spend and supplier information as well. Incidentally, Oracle is the first to market in the procurement space with a supply-focused product that leverages CDI technology (gained from its Siebel acquisition), although at the time of publication, this technology is not currently available from a procurement use case perspective. CDI technology actually improves the integrity of the data from individual sources, allowing users to match and link disparate information sources with varying levels of accuracy. CDI tools can correct for data-entry mistakes, such as misspellings, across different data sources to provide an accurate picture. Here again, the ability to accurately match UNSPSC codes to items is dependent on the accuracy and transparency of the original dataset!

One of the important questions to note in this four stage process is “who owns the cleansed and coded data?” ***It is important to note that not all providers will share the results of a cleansing activity with the customer.*** In some cases, they may elect to clean it only as input into their particular application (e.g. contracting, GPO services, etc.). Without direct access and ownership of the cleansed dataset, performing in-depth category analysis is not possible, which is the equivalent of restricting internal access to one’s own books! Trusting that a GPO or third party will conduct their due diligence and perform spend analyses on your behalf is a naïve assumption that merits further consideration.

Some of the providers we reviewed had systems that sould recognize a product and enrich it with the correct manufacturer name and item number, UNSPSC code, and descriptions, before uploading it into an ERP. However, these providers acknowledged that not every product code was matched, leaving an unknown number of items with no match that was not uploaded into the contract database. Here again, the importance of cleansed data is critical.

An automated process augmented with a manual process is the current standard in the healthcare industry that increases efficiency and accuracy. Customer service is important in this stage, as involving personnel with the expertise, such as clinicians to manage data is one of the key check-points for customers while choosing the vendors. Further, proper coding of the data will require engaging clinical experts, as well as other functional groups such as facilities, logistics, IT, legal, marketing, and finance to truly make sense of the data is critical to arrive at strategic sourcing decisions that will be effective. In this regard, a third party should be willing to provide the level of consulting and coordination that is consistent with the level of effort required to perform a thorough spend management project.

[1] “Creating a Source of Truth in Healthcare: Testing the GDSN as a Platform for the Healthcare Product Data Utility Results from DoD Healthcare GDSN Pilot Phase IIA”, DoD/VA Data Synchronization Study, September 2007.

[2] Schneller, Eugene, “A guide to successful Strategic Sourcing”, *Materials Management in Healthcare*, June 2010, pp. 22-25