While this column typically uses data visualizations you’ve probably seen before, I want to introduce one that perhaps you have not. This is in the realm of text analysis. When looking at FDA data, there are numerous places where the most interesting information is not in a data field that can be easily quantified, but rather in narrative text. Take, for example, Medical Device Reports of adverse events, or “MDRs.” While we can do statistical analysis of MDRs showing, for example, which product categories have the most, the really interesting information is in the
Why do we care? Those focused on product quality want to learn from history, and how better than from everyone’s history, not just your own. We can use the MDR data to find out product experiences with broad or narrow product categories. In the year 2020, for example, there were over 1.5 million such reports.

In this month’s column, I’m going to focus on cardiovascular devices. I could just as easily focus on software, or all implants or orthopedic devices or any other term that might cover a broad range of products. Because this stuff isn’t common, I’m going to include the methodology first, and then the visualization.

**Methodology**

I’m using what is known in data science as topic modeling to extract information from the event descriptions in MDRs. Topic modeling is an approach that allows us to see the big picture from a long list of documents. Essentially, topic modeling seeks to identify common topics discussed in a corpus of records.

While the output sort of looks like English, it also sort of does not. The algorithm is looking for words that are used together frequently. The algorithm strings those words together based on statistical importance, not based on how an English major would write them. Further, to help the algorithm recognize that the word “implant,” “implanted,” “implants,” and “implantation” are all pretty similar, we reduce each word to its stem. In the chart, you will see all variations of the word valve as “valv”. It takes some getting used to.

In addition, because context is usually important to understand how words are used, we look for phrases, in this case two words that are typically strung together. We could use however as many words in a phrase as we want, but more words means a lot more computer resources. In this case, two-word phrases seems to work adequately.

This exercise is a blend of art and science, and one of the judgments to make is the number of topics to consider. We typically make that decision based on what we refer to as coherence, which is a statistical measure of what is a logical principle: how well the words go together. We want to find topics that are meaningful to humans. Another area for using judgment is getting rid of words we don’t care about because they are too common and uninformative. I removed words such as “complaint” and “patient,” because they were in many of the event descriptions and didn’t add any particularly useful information.

From a technical standpoint, for those interested, I’m using a specific technique called Latent Dirichlet Allocation, a form of unsupervised learning, implemented through the Python library genism.

For cardiovascular devices, I thought it might be interesting to compare the topics that MDRs included in 2010 with those from 2020. I wanted to see how much change there might be over time. In the year 2010, there were over 45,000 MDRs for cardiovascular devices, and in 2020, there were almost 85,000.
**Charts**

A word about reading these charts. These are what are referred to as heat maps. The colors correspond with the intensity or value of a particular word, in this case to the topic. The darker the color, the more important the word in characterizing the meaning of the topic.

**Interpretation**

An expert in cardiovascular devices will undoubtedly be able to offer much greater insights into what these data mean, but it’s actually interesting that over the course of 10 years, the software found what turned out to be many very similar topics. The order changed. But many of the topics seem to be pretty similar. For those in the industry, that may be a bit depressing as it suggests a lack of progress in fixing
common problems.

Balloons rupturing still seems to be an issue. Guidewire tips still seem to be an issue. Battery problems still seem to be an issue. But we also have some new issues, such as problems with a surface cooling device (Arctic Sun®) for therapeutic hypothermia following cardiac arrest.

This technique of topic modeling can be applied as broadly or as narrowly as we wish. It can be very helpful when doing trend analysis where the valuable underlying data is in large volumes of text rather than structured data. In future columns, I will dig into other sources of regulatory text to see what information we can mine.

©2021 Epstein Becker & Green, P.C. All rights reserved.

Source URL: https://www.natlawreview.com/article/unpacking-averages-analyzing-descriptions-cardiovascular-device-adverse-events