Big Data Can Yield Big Insights On Promotional Practices

by Paul E. Greenberg and Tamar Sisitsky; Analysis Group, Inc.

Law360, New York (August 14, 2014, 11:40 AM ET)

Paul E. Greenberg

Tamar Sisitsky

In recent years, the use of big data and analytics have become more prevalent than ever before. Not surprisingly, this general trend has also affected the way government investigations and private litigation have unfolded with respect to allegations of improper promotional practices by manufacturers in the health care sector. Plaintiffs and defendants in this context have increasingly relied on big data to make their case, due in part to greater accessibility of both larger quantities and wider varieties of health care data, together with expansive increases in ever-cheaper computing power.

For pharmaceutical, biotechnology and medical device manufacturers, the ubiquity of big data at the negotiating table and in the courtroom has not only increased the analytical complexity required to address familiar litigation questions, it has also raised the bar on the types of questions being asked. Appreciating the implications of this new era of big data for health care litigation requires an understanding of several characteristics of big data: (1) increased quantity of data; (2) proliferation of new types of data; and (3) potential benefits from combining seemingly unrelated data sets in new ways.

Increased Quantity of Data

The volume of available health care data has expanded exponentially in recent years. Currently, data files in this industry are increasingly measured in terms of terabytes or petabytes of information, a scale that would not have been possible to work with efficiently just a few years ago. By way of background, a binary digit is 1 bit and there are 8 bits to a byte. In terms of disk storage, starting from 1 byte, each subsequent term on the following list is 1,000 times larger than the preceding one: byte, kilobyte, megabyte, gigabyte, terabyte, petabyte, exabyte, zettabyte, yottabyte, brontobyte, geopbyte. This progression implies that one petabyte is equivalent to 1 trillion kilobytes; in plain English, that's a lot of data.

Medicaid MAX administrative claims data, for example, is approximately six terabytes in size and contain billions of historical records of every payment made on behalf of millions of fee-for-service Medicaid beneficiaries. This includes inpatient and outpatient hospital stays/visits, emergency room visits, medical procedures, laboratory tests and pharmaceutical prescriptions fills. This immense data set covers all 50 states and the District of Columbia and has been analyzed in many contexts, including by the U.S. Department of Justice in its investigations of alleged off-label promotion by manufacturers.

Longitudinal, deidentified patient-specific data of this type can shed light on patient characteristics — including comorbidities and historical patterns of treatment — and how they affect physicians' prescribing decisions. For example, a patient with a history of drug switching within a specific therapeutic class might suggest unmet need as a key factor underlying a particular treatment choice. Such a pattern can be documented as an example of what likely would have occurred anyway even in the absence of the conduct at issue. Big data of this type can also be very helpful in identifying larger trends in treatment patterns and offer a more robust body of evidence than much smaller, cross-sectional physician surveys, for example. It can also facilitate analyses of rare diseases, specific treatments or particular patient subgroups, where it can be difficult to generate sample sizes sufficient to draw statistically meaningful inferences without access to a large overall patient population.

These vast troves of health care information often require a larger scale of raw computing power, such as parallel processing, and new computing approaches that could include machine learning procedures and predictive analytics. In addition, an increasing quantity of available data could change the nature of the analysis required to draw meaningful conclusions. As the size of the data set being analyzed increases, a threshold might eventually be crossed at which point it can become qualitatively quite different from smaller data sets, increasing the importance of human expertise in generating useful results. Moreover, since evaluations based on extremely large numbers of observations tend to be statistically significant in most circumstances, formal statistical signals may be appropriately superseded by expert judgment concerning clinical or economic importance.

New Types of Data

As the quantity of available data has increased, information has also grown more diverse, resulting in the proliferation of new types of data, such as those available from electronic health records or social media, as well as real-time patient information from next generation smart-device biometrics. Because some of this data is so new, appreciating the applicability of these information types may require a high degree of industry experience and technical proficiency.

Today, EHR use is widespread among providers, including compilation of both structured data (e.g., laboratory test results) and unstructured data (e.g., clinical notes). Such records can provide even more detailed patient information than insurance claims databases such as Medicaid MAX, albeit for much smaller sample sizes. However, given the unstructured nature of key aspects of the data, along with significant variation in the range of data captured at each venue of care, the ability to derive reliable insights from EHR often hinges not only on the integrity of the data itself but also on the expertise of the analyst.

Social media also offer new opportunities to leverage health information dissemination. These new data could help to identify alternatives to promotion that could explain uptake of a particular product (e.g., based on patient reactions to news events as recorded in discussion groups and chat forums). Although many aspects of social media data collection and analysis remain in their early stages of development, this source is likely to play a larger role over time. It would not be surprising to see such analyses transition from straightforward volume-based assessment to eventually include reliance on more sophisticated methodologies that account for cluster effects, herd effects and other complex patterns of patient behavior that can, at times, underlie the choice of one treatment over another.

Combining Standalone Datasets

The increasing quantity of available data and the promulgation of new data types have also created opportunities to merge multiple, distinct data sets to produce novel outcomes. This additional category of big data could help to clarify potential drivers of prescribing dynamics in the context of various types of disputes.

For example, in cases involving alleged kickbacks, valuable insights can be obtained by linking together some specific datasets, none of which were set up for this purpose. Overlaying physician prescribing data on speaker honoraria payment histories and event attendance lists can enrich one's understanding of the scope and impact of such manufacturer-sponsored activities. In the event prescribing and payment data are not available from the manufacturer, the combination of other publicly available sources can be instructive. This includes third-party vendor data from IMS or Wolters-Kluwer, newly released 2012 Medicare Part B administrative claims data, data that are forthcoming as a result of the Sunshine Act and data reported on the ProPublica website with extracts based on corporate integrity agreements and voluntary reporting.

Whereas it was once unthinkable to imagine combining data from disparate sources to obtain a comprehensive perspective on the medical circumstances giving rise to a particular treatment, it has now become a more realistic possibility. Indeed, the ability to combine insurance claims data, provider clinical records, patient and provider surveys and data generated from new biometric recording technologies, for example, has every potential to fundamentally affect a litigant's case.

Conclusion

The increased availability of large and complex health industry data, combined with advanced quantitative methodologies, provides a basis for detailed analysis of many different types of allegedly improper conduct by manufacturers in the health care sector. Today, many pharmaceutical, biotechnology and medical device companies, as well as other health care entities, maintain rich data sets concerning the types of marketing activities and provider relationships that are often at the heart of improper promotion and kickback allegations. This data can effectively be combined with third-party and other publicly available data to perform rigorous analyses of the conduct at issue and any corresponding impact.

Paul Greenberg is a managing principal and director in Analysis Group's Boston office, where he directs the firm's health care practice group. Tamar Sisitsky is a vice president in the Analysis Group's Boston office.