Big Data Enabling Better Pharmacovigilance
INTRODUCTION

Biopharmaceutical companies are seeing a surge in the amount of data generated and made available to identify better targets, better design clinical trials, enable translational science, and better position products in the market. Big Data is being generated from two different sources – science-based data from laboratories including genomics and proteomics data, and real world data from Electronic Health Records (EHRs), claims, wearables, surveys conducted by market research, and social media. Biopharmaceuticals can derive significant value and competitive advantage by using data in all parts of their value chain.

This white paper highlights the advantages of leveraging big data in pharmacovigilance. The paper begins with examples of big data sources and the benefits of using such data across the value chain. Then, using pharmacovigilance as an example, the paper delves into approaches of using big data and building a synergy between traditional and big data analytics. The value propositions are clear: faster and better insights.

Sources and Benefits of Using Big Data in Biopharmaceuticals

Big data for biopharmaceuticals is generated both from internal and external sources. This data comes from internal experiments in laboratories, and from patients and stakeholders.

- **Internal scientific data:** An exponential increase in genomics and proteomics data has provided significant opportunities for identifying and validating targets, understanding disease pathways from genes through disease manifestation, and optimizing leads at the patient level for efficacy and safety.

- **Patient-related data:** The data from central laboratories, prescriptions, claims, EHRs, and now Health Information Exchanges (HIEs) is providing an immense opportunity to analyze and gain insights across the entire value chain, such as:
  - **Drug Discovery:** Analyzing and spotting additional indications for a drug, disease pathways, and biomarkers.
  - **Clinical Trials:** Optimizing clinical trials through better selection of investigators and sites, and defining better inclusion and exclusion criteria. Wearable technologies can generate significant amount of data to monitor patients, such as tracking key parameters and therapy compliance.
  - **Commercial:** Increasing revenue, as illustrated by Sanofi’s use of comparative market surveillance studies to display Lantus’ superior efficacy. The use of Lantus delayed the future need for higher-priced therapy. The research forced the German Payer G-BA (Federal Joint Committee) to reverse its decision to not provide coverage for Lantus.

- **Social Media:** Patients are using Twitter, Facebook, and blogs to communicate their outcomes and impressions of the drug.
The surge in data availability, in terms of volume and variety, provides several alternatives to gain insights.

- Integrating data from multiple sources: There is significant value in looking at data integrated from several sources. A sequential view/analysis of data does not provide the synergy that comes from integrated data.

- New techniques in analyzing data: There are several new techniques available to analyze. For example, research publications can be analyzed using text mining methods to summarize the findings. Therapy or drug-based ontologies (taxonomy combined with interrelationships among entities) can be developed to refine searches. The ontologies can, in turn, be developed through machine learning.

Clearly, there is excitement for big data in biopharmaceuticals from the explosive growth in reliable information, and new technologies to search, compare, analyze, and summarize it. This enables discovering better drugs more quickly, and at lower cost. The value propositions are clear.

BIG DATA IN PHARMACOVIGILANCE

Big Data can be a foundation for pharmacovigilance for integrating and analyzing the vast variety of data. One can say that it has a great need for, and provides compelling opportunities within the pharma value chain.

- Safety concerns can have serious implications on patient health and on a company’s financial health and reputation. These concerns can derail a drug on the track to becoming a blockbuster.

- Different types of data needs to be integrated:
  a. Adverse Event (AE) cases reported directly to companies and its partners
  b. Health agency databases such as FDA Adverse Event Reporting System (FAERS) and Uppsala Monitoring Center (VigiBase)
  c. Real world data from cohort studies can be provided as both structured and unstructured data
  d. Published literature on safety and efficacy
  e. Data from social media (Twitter, Facebook, blogs), while this data is not as reliable as others, it is becoming the earliest warning signal and is creating a definite and special need to seek, store, and analyze to spot safety concerns

The FDA recently launched Sentinel System, a national electronic system to track the safety of products in the market\(^1\). It allows the FDA to actively query diverse automated health care data holders—including electronic medical records (EMRs), insurance claims, and registries—to evaluate product safety issues.
Clearly, there is excitement for big data in biopharmaceuticals from the explosive growth in reliable information... The value propositions are clear.
TRADITIONAL APPROACH: SIGNIFICANT DELAY FROM AE TO INSIGHT

AE case processing is the traditional workhorse for spotting safety in real world cases. Much of the insight comes from analysis at the end of a case’s lifecycle. A case is processed (data entry, medical coding and review), analyzed and reported either through expedited or periodic reporting requirements. Signaling is a relatively recent data mining approach applied to AEs.

Figure 1 displays the sequence of events and activities that need to be completed before data mining, analysis, and arriving at insights. The lead time to derive meaningful insight from information ingestion and processing through signal detection can be over a month even when considering serious adverse events.

Contrast this with the actual life cycle of an adverse event. The first trace of an adverse event can begin in the social media—it may be as simple as a non-serious AE. These events may not even be reported to a physician. Additionally, non-serious AEs can worsen into a serious AE. Even then, physicians are not mandated to report AEs – it is voluntary. De-identified data from hospitals can be analyzed but that requires significant ingestion and curation time, adding one more step and resulting in significant delay. Detailed analyses of causality and higher incidences (signal) occur much later (see Figure 2).

Figure 1: AE Case Processing Lead Time

Figure 2: The Lead Time before The Delay to Insight
THE NEW PARADIGM: LEVERAGING BIG DATA FOR A QUICKER INSIGHT

The data for pharmacovigilance has an additional dimension compared to data for other parts of the biopharma value chain. It not only has volumes of reliable scientific and patient-related data to analyze, but also volumes from social media that forms the other “V”: veracity. Figure 3, a conceptual view, shows the diversity of data for pharmacovigilance.

Two-Tiered Approach to Analysis

There is a need for a tiered approach to analyze the data from low and high reliability, to mine text versus data, and to understand the speed versus the reliability of insights. See figure 4. The data from social media is available faster than data from published journals, but may be from a patient who may not be a qualified physician, while the other is from a qualified physician who has spent hours thinking, collecting evidence, analyzing, and putting the concept in a structured format. The real world data and physician-reported adverse events are available faster than those from published journals, but the level of research is less reliable than the information from journals.
With the proliferation of social media and the release of data from large study endeavors, data sources are increasing, and will only continue to increase at a faster rate.
Big data tools and technologies are ideally suited for this tiered analysis approach. The Big Data layer is enabled by cost-effective technologies, newer analytics and visualization, and quicker implementation timelines.

1. A first tier (big data) that emphasizes the following:
   a. Speed to Insight through faster data ingestion
   b. A consolidated view of the data
   c. A company and therapeutic specific ontology automated through machine learning
   d. Business friendly analysis through:
      i. Data exploration
      ii. Text mining
      iii. Ontology-based searches
      iv. Quick data visualization
      v. Testing ad hoc hypothesis

2. A second tier (traditional) can leverage the information and insight gained from the output of first-tier and continue with the traditional analysis of case processing and signaling:
   a. Focused analysis of insights from Tier 1
   b. An expanded scope to text-based insights from detailed case analysis and summary reports
   c. Harmonized findings from the first tier and the traditional approach
The Big Data Reference Architecture

The Big Data layer is an essential tool in today’s pharmacovigilance operations. The reference architecture in Figure 5 shows what must be designed to deliver speed and flexibility in a cost-effective manner.

**DATA SOURCE**
- FAERS, VigilBase, Real World Data, HIE
- Publications from Journals, Narratives from Case Intake
- Drug & Product Registration
- Social Media, Registries
- Preclinical, PK/PD Data

**Value Proposition**
- Ability to include multiple large volumes of diverse data sources

**BUILDING FOUNDATION**
- Quick Ingestion, Curation, Data Lineage, etc.
- Leveraged for faster ingestion to traditional databases

**Ontology**
- Organ classes with synonyms, known adverse events for specific class of drug – drug – indication combination
- Company-specific, therapeutic specific, and indications specific ontology
- Developed continually using machine learning capabilities

**ANALYSIS**

**Business Question**
- Generate ad-hoc hypotheses leading to business questions

**Text Mining**
- Search
- Document summarization
- Theme extraction
- Interface with ontology to develop concepts
- Word cloud visualization
- New insights

**Data Exploration**
- Search, data visualization, summarization, and modeling

**CONCLUSION: WHERE ARE YOU IN YOUR JOURNEY?**

With the proliferation of social media and the release of data from large study endeavors such as the Human Genome Project, data sources are only increasing today and will only continue to increase at a faster rate. Patient registries may be able to get near real-time data from wearable technologies, and Health Information Exchanges will be an increasingly important source of information.

The increasing need to assimilate a wide variety of data in large quantities from numerous disparate sources quickly and efficiently has become important from both competitive and safety perspectives. Prospective competitors may be able to position their products by highlighting specific safety aspects as a means to drive pricing.

While the benefits and value propositions are clear, it is important to realize that Big Data is a journey. Companies typically adapt the following steps – however, each step has best and worst practices.
1. Assessment of current objectives and environment for readiness and scope: Several companies have already started or completed this assessment. This step helps in solidifying initial scope and objectives while educating all stakeholders on the big data ecosystem.

2. Roadmap and Proof of Concept: There are pros and cons to sequencing these two steps. A good proof of concept is better when there are influential stakeholders who have expressed skepticism that this is an “old wine in a new bottle” hype. A roadmap, on the other hand, precedes a proof of concept when there are enlightened influential stakeholders or a champion for the cause. In either case both steps are required.

3. Implementing the First Phase: Choosing scope for the first phase is critical to the success of the program. The scope must be of the right size, sources of data should showcase the benefits, tools should be identified, and implementation methodology should be concrete to ensure success for continued adoption of the roadmap.

For more information on how Knowledgent can assist with big data enablement of pharmacovigilance, visit www.knowledgent.com.

References:


2. FDA’s Sentinel Initiative, [http://www.fda.gov/Safety/FDASentinelInitiative/ucm2007250.htm](http://www.fda.gov/Safety/FDASentinelInitiative/ucm2007250.htm)