SciBite
an ELSEVIER company

Whitepaper

# Semantic Analytics:
# An integrated approach for pharmacovigilance teams to achieve total awareness

SciBite
an ELSEVIER company

# Semantic Analytics: An integrated approach for pharmacovigilance teams to **achieve total awareness**

**The World Health Organization (WHO) Programme for International Drug Monitoring (PIDM) was established in 1968 to facilitate the detection, assessment, understanding and prevention of adverse effects and other drug related problems, such as lack of efficacy. Despite increasingly stringent regulations, pre-marketing clinical trials designed to ensure safety do not involve a sufficiently large or diverse population and are not long enough to detect all potential safety issues, such as adverse events that have a low incidence rate[1] or those resulting from drug-drug interactions, which are thought to be responsible for up to 30% of adverse events[2].**

In most countries, adverse event reporting is now a regulatory requirement for pharmaceutical companies and has resulted in more than 100 drugs being removed from the market due to safety issues in the past 40 years. However, many of these, such as Vioxx (Rofecoxib) and Meridia (sibutramine), had already been prescribed to millions of patients worldwide and adverse events remain one of the leading cause of death in the US[3].

Clearly more comprehensive, systematic monitoring is required in order to detect, validate and act upon new adverse events as early as possible. WHO defines a safety signal as: "Reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously"[4]. Pharmaceutical companies are now obliged to scan a range of credible data sources for possible such relationships – so called 'signal detection' – rapidly assess validity and notify regulatory bodies within a reasonable timeframe, based on severity.

This places significant demands on Pharmacovigilance teams, who are challenged to maintain safety and compliance with the same, or fewer, resources amid increasingly stringent, globally diverse regulations. The exponentially growing amount of data and increasingly diverse range of sources make it almost impossible to maintain a comprehensive and up-to-date understanding. The result is that the legacy approach to signal detection, involving manually scanning biomedical sources, is no longer practical.

This whitepaper describes how Semantic Analytics enables an integrated approach to Pharmacovigilance, unlocks the potential of biomedical content and will enable pharmaceutical companies to:

- Explore, analyse, query and manage underutilized sources of safety information
- Achieve a comprehensive, up-to-date awareness
- Expedite the process of identifying, validating, reporting and acting upon adverse events...

and ultimately to protect patient safety, mitigate risk and ensure compliance in a resource-effective manner.

1 Alemenoff JS, Pattishall EN, Gibbs TG, et al. Novel statistical tools for monitoring the safety of marketed drugs. Clin. Pharmacol. Ther. 2007; **82**:157-66

2 Pirmohamed D, Orrne ML. Drug interactions of clinical importance. In: Davies D, Ferner RE, de Glanville H(eds). Davies's Textbook of Adverse Drug Reactions. London: Chapman & Hall Medical, 1998, 888–912.

3 Lazarou J, Pomeranz B, Corey P. Incidence of adverse drug reactions in hospitalized patients. JAMA: J. Amer. Med. Assoc.1998; **279**(15): 1200-5

4 World Health Organization. The importance of pharmacovigilance: safety monitoring of medicinal products. Geneva: WHO http://apps.who.int/medicinedocs/pdf/s4893e/s4893e.pdf 2002

SciBite
an ELSEVIER company

## Enabling an Effective Pharmacovigilance Strategy

To be effective, a Pharmacovigilance strategy must be built on several critical foundations. Namely the ability to:

- Identify safety signals in a growing number of sources
- Contextualise unstructured text and enable it to be treated as data
- Make connections across disparate data sources
- Ensure information is usable and accessible without having to rely solely on the limited availability of experts

Each of these is described in detail below.

## Widening the 'Search Space'

Until recently, Pharmacovigilance teams have had to rely on a limited number of structured sources to search for adverse events. Spontaneous reporting systems such FDA Adverse Event Reporting System (FAERS), EudraVigilance and Vigibase provide searchable repositories of safety-related information reported by health professionals. Whilst undoubtedly valuable, these systems are considered to have numerous drawbacks including latency, under-reporting and reporting bias[5]. Similarly clinical observational data, such as electronic health records (EHR), can also be a valuable source of information[6] but are not widely accessible, even when de-identified[7]. EHRs can also contain incorrect or incomplete clinical information and introduce potential 'noise' as they may mention the conditions of relatives or those that were considered but ruled out during diagnosis.

Regardless of their merits and limitations, by relying solely on spontaneous reporting systems and EHRs, adverse events with a low incidence rate (e.g. 1:10,000 or less) or those that manifest themselves as birth defects may not be picked up. Hence, regulatory bodies have mandated that pharmaceutical companies must maintain a comprehensive, global awareness by proactively and thoroughly scanning a broad spectrum of sources for safety events associated with their products[8,9].

A wealth of safety-related material is available within the biomedical literature, such as PubMed, and other credible resources[10], such as new drug applications (NDA) to the FDA and EMA and conference abstracts and proceedings. Yet compared with spontaneous reporting systems, the biomedical literature has thus far identified a relatively small number of safety events. As such, the effort to monitor an increasing volume of literature may at first appear to yield rapidly diminishing returns. However, several major drug recalls were initiated based on an analysis of published adverse events[11,12], thus having an impact that is disproportionately high compared to their volume. Recognising the explosion of available information, the EMA now recommends screening the internet or digital media for reports of possible adverse events[13,14].

A potentially valuable, up-to-date resource is the information generated by patients themselves, as they share unsolicited information about their experiences with medicinal products on health forums such as PatientsLikeMe as well as blogs, wikis and even Twitter. However, since there are little or no controls over what people say, finding an actionable signal amongst the unsubstantiated 'noise' of misinformation and random chatter is a challenge. Nevertheless, it is important to recognise that more and more people are turning to social media for medical information – one recent study found more potential adverse events reported on Twitter than were reported to FAERS during the same six month period[15].


5 Alvarez-Requejo A, Carvajal A, Begaud B, Moride Y, Vega T, Arias LM. Under-reporting of adverse drug reactions estimate based on a spontaneous reporting scheme and a sentinel system. Europ. J. Clin. Pharmacol. 1998; **16**(3):328-37

6 Colona PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR project. Pharmacoepidemiol. Drug Saf. 2011; **201**(1):1-11

7 Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. Journal of Biomedical Informatics 2013. **46**(5):765-73

8 World Health Organization. The importance of pharmacovigilance: safety monitoring of medicinal products. Geneva: WHO, http://apps.who.int/medicinedocs/pdf/s4893e/s4893e.pdf 2002

9 Kelman CW, Pearson SA, Day RO, Holman CD, Kliewer EV, Henry DA. Evaluating medicines: let's use all the evidence. Med. J. Aust. 2007; **186**:249-52

10 Koutkias VG, Jaulent MC. Computational approaches for Pharmacovigilance signal detection: toward integrated and semantically-enriched frameworks. Drug Saf. 2015; **38**:219-32

11 Edwards IR. What are the real lessons from Vioxx? Drug Saf. 2005; **28**:651-8

12 Levesque LE, Brophy JM, Zhang B. The risk for myocardial infarction with cyclooxygenase-2 inhibitors: a population study of elderly adults. Ann. Intern. Med. 2005; **142**: 481-9

13 HMA: Guideline on good pharmacovigilance practices (GVP) Module VI – Management and reporting of adverse reactions to medicinal products (Rev 1). http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/09/WC500172402.pdf

14 European Commission. eHealth for safety: impact of ICT on patient safety and risk management. 2007. Available at http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=3150 Accessed 20th March 2017

15 Freifeld CC1, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, Dasgupta N.. Digital DrugSurveillance: Monitoring Pharmaceutical Products in Twitter. Drug Saf. 2014 May;**37**(5):343-50.

Hence, there are now numerous initiatives such as EU-ADR[16] whose goals are to effectively mine adverse events from social media.

There is also growing interest in mining the implicit information in search engine logs – the search terms patients use as they leverage search engines to learn about the medications they are using can give valuable insights into potential adverse events[17], akin to the Google Flu Trends project[18] which has been used to track influenza and give 'early warning' signals.

Ultimately, there isn't a single definitive place to find adverse event information, and the trend is for information to be unstructured and increasingly decentralized. The need to monitor an ever growing number and variety of sources is clearly a challenge – an effective Pharmacovigilance strategy needs to be able to accommodate a growing number of disparate data sources, and remain adaptable enough to incorporate new sources that will undoubtedly arise in the future.

## Unlocking the Potential of Unstructured Data

Limiting surveillance to a limited number of structured sources is no longer sufficient, yet the vast majority of sources available today are unstructured and not designed with searching for adverse events in mind. Text mining offers the possibility to utilise the biomedical literature more effectively by enabling text to be treated as data. However, many computational approaches struggle to deal with the complexity and variability of unstructured scientific language. Multiple terms can be used to describe the same thing, so searching for 'heart attack' would miss references to 'myocardial infarction' or 'serious heart event'.

Structured domain knowledge, in the form of high quality biomedical vocabularies and ontologies such as MeSH, Snomed, RXNorm, MedDRA, DINTO and the Anatomical Therapeutic Chemical (ATC) Classification System, provides a common language that is a critical foundation of any text analysis methodology. Ranging from simple lists containing all of the known terms for the same real world "thing" to hierarchical groupings and classifications of scientifically-related concepts (such as 'all anti-inflammatories' or 'all DNA replication proteins'), they can be used to accurately detect relevant information within unstructured biomedical text.

But no single vocabulary or ontology is comprehensive in either depth or breadth. For example, no one vocabulary includes all possible terms or phrases used for every indication and drug name vocabularies do not always include the multiple trade names by which the same drug is marketed. Since adverse event surveillance is a global responsibility, relying solely on using a single ontology is akin to relying on a single 'lens' with which to review the global biomedical literature and will result in potentially important safety signals being overlooked. Similarly, to effectively mine patient-generated sources, a specific ontology would be needed to align consumer vernacular with MedDRA. Many pharmaceutical companies also have their own lists of project codes and compound IDs that are useful when mining internal documentation.

However, biomedical vocabularies and ontologies are often developed in isolation, leading to numerous standards with varying and overlapping content. A robust, holistic approach to Pharmacovigilance requires the integration of a range of complementary ontologies, enabling a rich 'net' to be cast to identify all possible uses and variants of a term of interest.

## Contextualizing and Linking Data

While adverse events can be reported as direct statements on causality, they are often only referred to by association or inference. Likewise, it may be difficult to directly associate an adverse event with a specific drug if multiple drugs mentioned in same article, which is increasingly the case due to the rise in popularity of multi-drug regimens.

The standardization that is made possible by vocabularies and ontologies is the foundation of Semantic Analytics, which applies an explicit, unique meaning and description to a term. This enables unstructured text to be contextualised so that it can be understood and used as high quality, actionable data, irrespective of its source. Without such contextualization, the natural ambiguity present in unstructured text would reduce its utility. For example, the term 'concentration' could be referring to a measurement or an activity – Semantic Analytics helps define which one is correct. Despite significant advances in the technology, sophisticated tools such as natural language processing and machine learning are unable to disambiguate terms. Semantic enrichment is a necessary prerequisite for cognitive computing and provides the logic necessary for effective artificial intelligence.

16 https://www.ncbi.nlm.nih.gov/pubmed/19745234

17 White RW, Harpaz R, Shah NH DuMouchel W, Horvitz E. Toward enhanced pharmacovigilance using patient-generated data on the Internet. Clin. Pharmacol. Ther. 2014; 96:239-46

18 Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS and Brilliant L. Detecting influenza epidemics using search engine query data. Nature 457: 1012-1014 (19 February 2009)

The ability to uniformly contextualize data by semantic enrichment provides the basis for linking data from heterogeneous biomedical sources. It facilitates the generation of a network, or semantic model, of interconnected facts. This enables Pharmacovigilance teams to build up a richer picture, find supportive information to validate potential safety signals (e.g. by determining biological plausibility) and infer additional insights that are only possible through semantically enriched, linked data. For example, by connecting DailyMed label information, it is possible to determine if a safety signal is associated with a known adverse event. Furthermore, by incorporating data sources that do not explicitly contain safety information, such as ChEMBL drug targets, it is possible to identify adverse events associated with all drugs of the same class, or that have the same target. Similarly, the ability to search and navigate linkages across data sources simplifies the identification of potential drug-drug interactions[19], for example based on gene–drug relationships[20].

## Ensuring Accessibility

Effective pharmacovigilance requires the contribution of a range of people from different disciplines, each with different needs. Data scientists will need to perform sophisticated queries to find everything that is known about a drug or class. They will benefit from a way to visualise large volumes of data, such as a network/graph view, which will enable them data to find new connections and perspectives. For example, using a potential safety signal as a starting point, network visualizations enable users to traverse the linkages in the semantic model and identify if the event is known or rare and whether it expected based on the drug class – i.e. to help determine if it is it plausible.

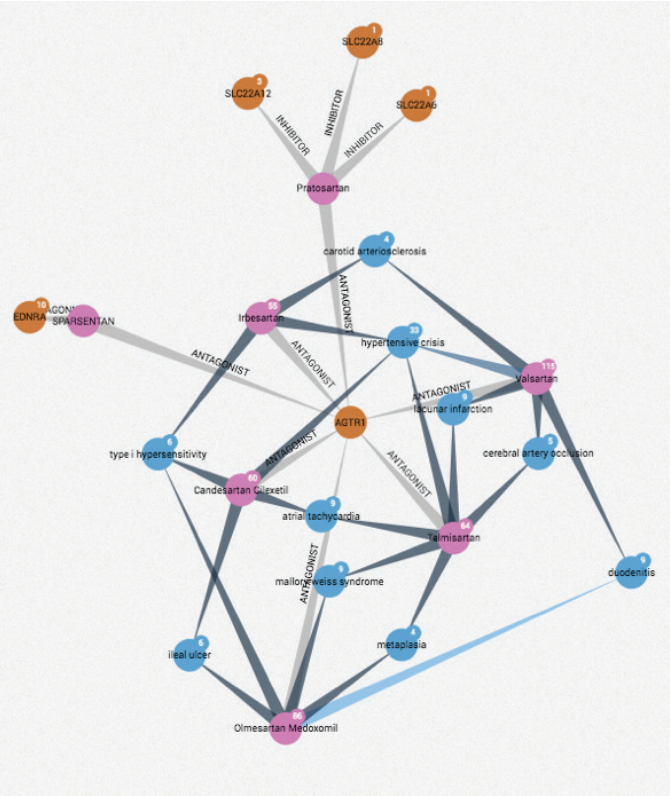


*Figure 1: A network illustrating drugs from ChEMBL (pink nodes) that have an antagonistic effect on target, angiotensin II receptor, type 1 (AGTR1: orange node). Additionally, these drugs are connected to FDA-reported adverse events (blue nodes) from FAERS (https://open.fda.gov/data/faers/). The relationships between the nodes have a type (e.g. inhibitor, antagonist) and the Drug-Adverse Event relationships are scored based on reporting frequency.*

In contrast, other users will need to be able flexibly query the linkages present in a semantic model without actually seeing the graph itself. For example, regulatory affairs will benefit from an intuitive interface that highlights important information in a manner that is easy to understand and digest.

19 Vilar S, Friedman C, Hripcsak G. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. Briefings in Bioinformatics, 2017: 1–15

20 Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. Pac. Symp. Biocomput. 2012: 410–21.

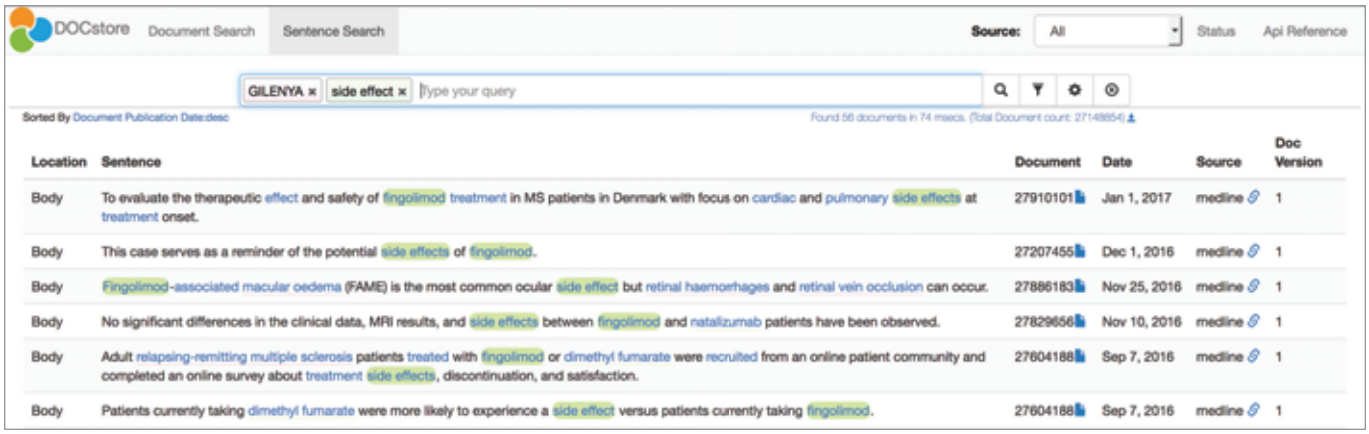


***Figure 2:*** *An intuitive interface enables users to quickly identify same-sentence co-occurrence of drug and side effect terms. The example shows text from MEDLINE relating to fingolimod, which was automatically recognised as synonymous to Gilenya*

For it to be effective, a Pharmacovigilance strategy needs to be based on the principle of democratizing knowledge. It needs to appreciate that Pharmacovigilance is not a single activity and that it is conducted by multiple people with different roles and that information needs to be surfaced to different consumers in different ways, to meet the specific needs of their workflow.

An effective Pharmacovigilance strategy must be able to:

- Integrate and connect an increasingly diverse range of data sources
- Contextualise the wealth of unstructured data found in the biomedical literature
- Cast a rich 'net' to identify all possible uses and variants of a term of interest
- Make information accessible to different consumers in different ways, to meet different workflows

## Applications of Semantic Analytics for Pharmacovigilance Teams

This section outlines a range of applications of Semantic Analytics for Pharmacovigilance teams. All applications are based on the key principles described above, with each application building on the benefits of the previous one.

## Streamlining Literature Review

Until recently, extracting useful information from the biomedical literature necessitated a manual process involving complex but imperfect searches, limited to abstracts or indexing terms and that resulted in tens or hundreds of articles to be read by experts. For example, a search of PubMed for articles with Infliximab in the title or abstract, published in 2016 results in over 900 hits.

Extending this search to include the class of drug (i.e. TNF inhibiting immunosuppressants) or the indications it is intended to treat (such as ankylosing spondylitis, psoriatic arthritis, psoriasis and rheumatoid arthritis) results in thousands of articles. Pharmacovigilance teams simply don't have the time to read all of the available content relevant to their drug. Yet they need to be confident that their strategy minimises or eliminates the risk of missing a serious adverse event or reporting deadline.

Since pharmaceutical companies are expected to maintain awareness not only of their own compounds but also compounds of same class and target, such a manual approach is prohibitively time consuming and has a high risk of missing safety signals. Furthermore, even when the work is outsourced, the laborious manual review of all the articles published in the previous week typically takes several days. When it comes to identifying severe adverse events, any delay, even a matter of days, can cost lives.

A simple but effective application of Semantic Analytics is to rapidly process a huge volume of biomedical content from a diverse range of sources. Semantic Analytics enables articles to be ranked by their relevance to a particular search and can accurately mark-up all relevant terms and concepts present within each article, without being limited by the indexing terms used by the data source.

While this does not limit the volume of content to be processed, it does enable reviewers to focus on the articles of highest relevance, rapidly identify the topics covered in the article, easily interpret the text and prioritise accordingly thus expediting the entire process. Essentially this approach uses technology to assist human review and eliminates the trade-off between the volume of content and the time available to review it.

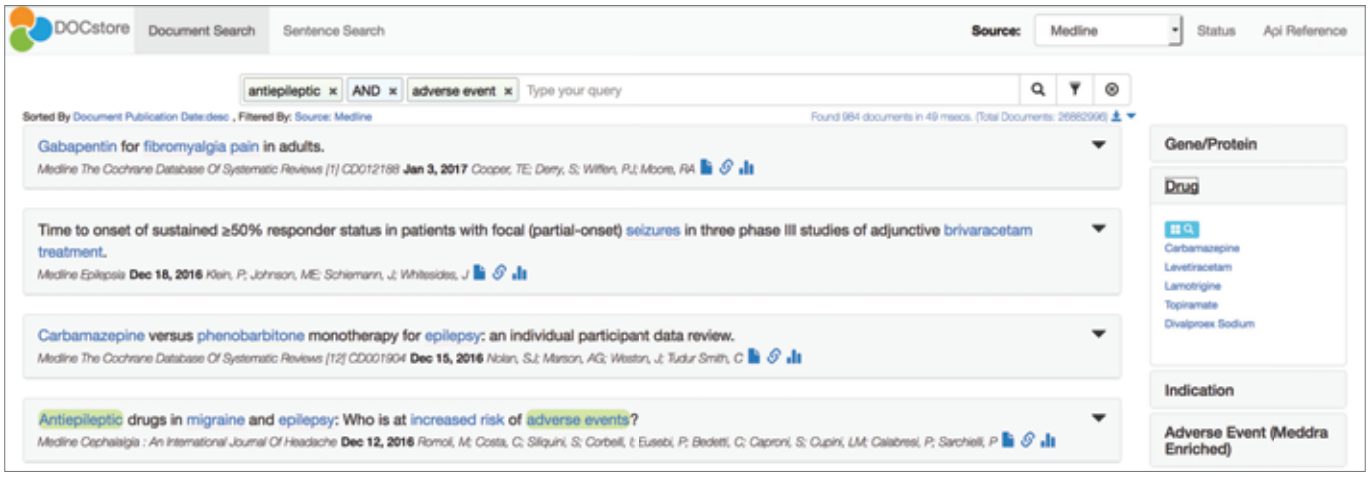


*Figure 3: Ranking query results by relevance and summarising all terms identified within the results (right hand side of the screen) to enable users to easily navigate to the most interesting articles.*

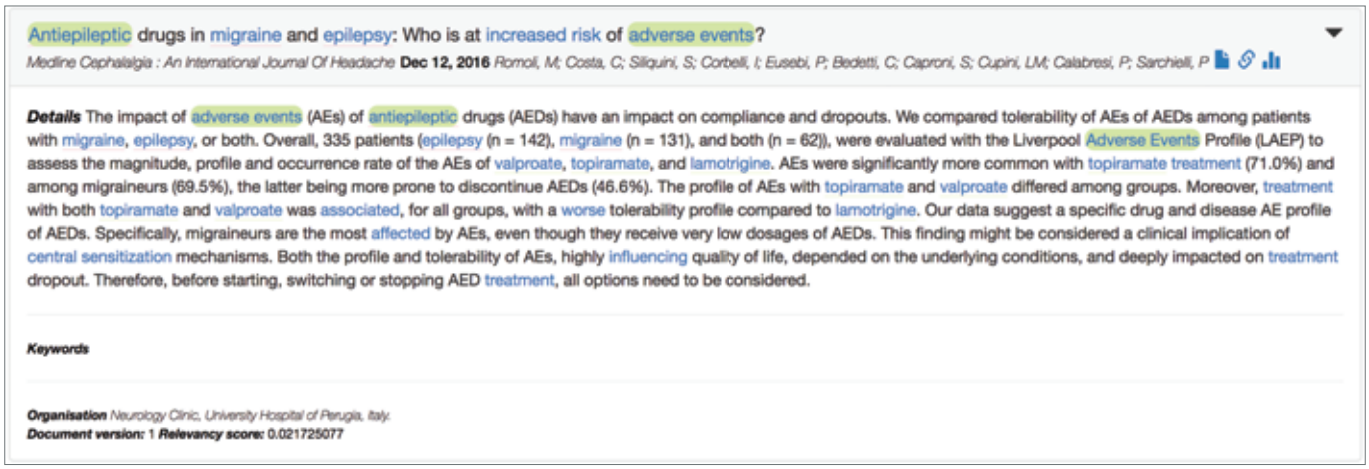


*Figure 4: Marking-up relevant terms within an article facilitates interpretation*

## Automated Surveillance

Many pharmaceutical companies apply a triage process to review the biomedical literature. The first filter involves performing literature searches for a pre-defined set of terms. The second filter, which is frequently outsourced, involves reading the full text of all search hits to check the whether the presence of the terms of interest are relevant. Any articles deemed to contain potential safety signals are sent to experts for assessment, resulting in the confirmation or rejection of the identified signal.

This approach is flawed in several ways. Firstly, the initial search process can't possibly encompass all possible synonyms for a term of interest, so signals can be missed from the outset. Secondly, relying on different people reading through search hits is a highly subjective process. The serious implications of missing a critical signal result in an incredibly risk averse process, generating many false positives. The consequence is that experts still end up reading through reams of information – the valuable time of a very limited resource is ultimately spent excluding a large proportion of the information that comes their way. The entire process is time consuming, not scalable, and is at odds with the goal of reaching timely and accurate conclusions.

Semantic Analytics can simplify the process of flagging safety signals. By combining this with business rules to determine how documents containing potential signals should be prioritized and routed, the non-value added activity of people reviewing documents in order to determine whom to send them to can be eliminated. Essentially, Semantic Analytics turns manual surveillance on its head by continually scanning all available sources and only issuing alerts when something noteworthy is found. This 'pre-screening' can occur in near real-time.

This changes the role of expert resources to one of oversight, with the need to only get involved in the detail of more ambiguous cases. Their time will be used more intelligently – automated pre-screening ensures the right people will be alerted to important safety signals in near real-time without the need to wait for the earlier stages of the triage process to be finished. By optimizing and automating the routing process, outsourcing costs can be reduced. Perhaps more 14 importantly, the costs that would arise from delays in identifying an important safety signals can be eliminated.

This application of Semantic Analytics is by no means limited to assessing safety signals – it can be applied to any process whereby prioritization and/or routing of documents can only be determined by first reviewing the content of the document.

## Predictive Analytics

With the vast and growing amount of data in the public domain, even knowing which drugs belong in a particular class can be a challenge. Semantic Analytics can ensure all existing information is considered when looking for likely side effects.

As illustrated below, there is a direct relationship between the drugs Valsartan and Olmesartan medoxomil and inflammation of the lining of the digestive system (duodenitis). However, Semantic Analytics reveals extended connections, enabling the identification of a potential relationship between Angiotensin II Receptor Type 1 (AGTR1) antagonists and duodenitis, something that would be missed by traditional keyword-based search strategies.

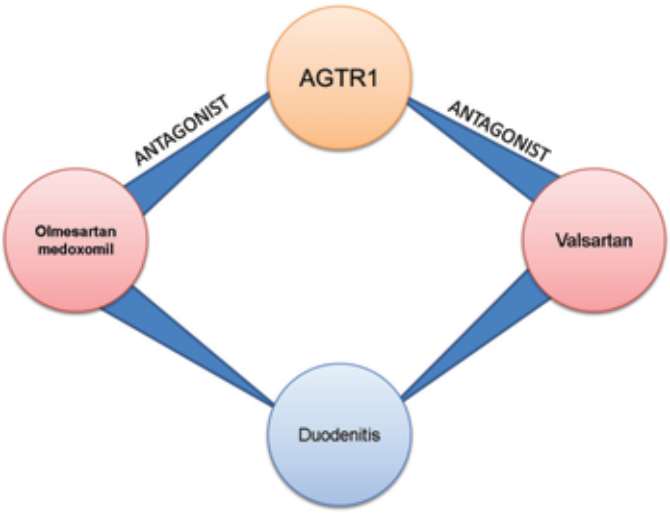


***Figure 5:*** *Identifying potential relationships from connected data sources*

Such insight is valuable both before and after a drug is marketed. Even prior to first in human trials, Semantic Analytics can be used to generate new hypotheses and help predict which adverse events may occur, however infrequently, including potential drug-drug interactions. Hence, clinicians and study teams can be armed with the knowledge of what to look out for when treating a specific condition.

Similarly, once a drug is in clinical use, if an unexpected safety issue is identified, the rich picture revealed by Semantic Analytics enables Pharmacovigilance teams to rapidly determine how biologically plausible it is based on a comprehensive, interconnected knowledge base. Hence, the confirmation or rejection of potential adverse events becomes more accurate, timely and evidence-based.

As with automated surveillance for safety signals, Pharmacovigilance teams can be automatically alerted to interesting new connections as and when they are found. Since any new knowledge can be used to refine the surveillance process, 16 a feedback loop can be established to continuously improve and expand organizational knowledge.

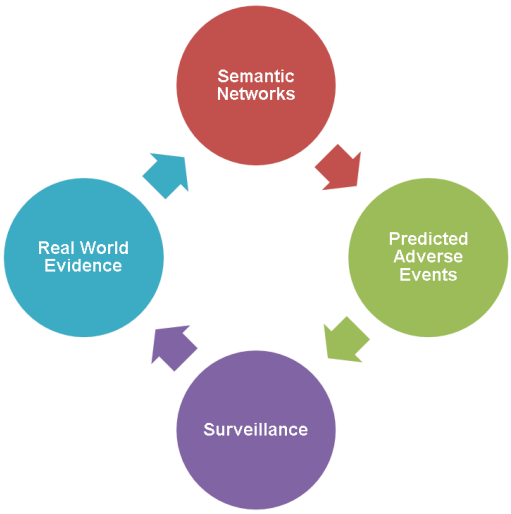


*Figure 6: Establishing a knowledge feedback loop*

## Beyond Adverse Event Detection – Additional Applications

The application of the context-rich knowledge that results from a Semantic Analytics approach is by no means limited to the detection and validation of adverse events. For example, a by-product of scanning the literature for adverse events related to all drugs in a class is that the results provide an up-to-date source of valuable competitive intelligence. It can also help focus future R&D efforts and ensure time is not wasted developing molecules that are from a class that has associated safety issues or would ultimately be entering an unattractive, prohibitively competitive marketplace.

Despite ongoing efforts towards harmonization, safety regulations still differ significantly between countries. Hence, the inclusion of regulatory documents and articles within the pool of literature to be scanned will enable regulatory teams to be alerted to relevant changes and ensure they comply with national laws and local regulations. Finally, due to rising development costs, a growing number of biotechs and pharmaceutical companies are focussing on Drug Repurposing, Repositioning and Rescue (DRPx) as a means to maximise value from their R&D investments. By identifying drugs with similar adverse event profiles, one can form the hypothesis that they have a related mechanism of action. Hence, the information managed by an integrated Pharmacovigilance strategy can be used to inform the strategy for a DRPx initiative and enables organizations to quickly identify where to focus further investigation.

## Summary

Regulatory bodies expect pharmaceutical companies to maintain an up-to-date awareness of the safety implications of not only their own drugs but also those from the same drug class and with the same target that are marketed by competitors. To ensure compliance, Pharmacovigilance teams must move on from the prevailing manual approach involving the time consuming, piecemeal review of a small range of structured sources.

Semantic Analytics provides a resource-effective solution to these challenges. It enables Pharmacovigilance teams to efficiently and comprehensively monitor a wide range of heterogeneous and cross-disciplinary sources and be fully aware of all safety signals directly and indirectly associated with one or more drugs of interest. The application of Semantic Analytics will expedite the identification and validation of adverse events, enabling pharmaceutical companies to reach timely, well-informed decisions, resulting in safer treatments for patients.

SciBite
an ELSEVIER company

SciBite's data-first, semantic analytics software is for those who want to innovate and get more from their data. At SciBite we believe data fuels discovery and we are leading the way with our pioneering infrastructure that combines the latest in machine learning with an ontology-led approach to unlock the value of scientific content. Supporting the world's leading scientific organisations with use-cases from discovery through to development, SciBite's suite of fast, flexible, deployable API technologies empower our customers, making it a critical component in scientific, data-led strategies. Contact us to find out how we can help you get more from your data.

To learn how SciBite can unlock the value of your data, speak to one of our experts today or email us at contact@scibite.com

**Head Office:**
SciBite Limited
BioData Innovation Centre
Wellcome Genome Campus
Hinxton, Cambridge CB10 1DR
United Kingdom

www.scibite.com
contact@scibite.com
LinkedIn: SciBite
Twitter: @SciBite
+44 (0)1223 786 129