Text Mining for Drug Development: Gathering Insights to Support Decision Making

Sherri Matis-Mitchell
Consultant, DataStar Insights
Oxford PA , USA
Sherrimatismitchell@gmail.com

Abstract— Drug discovery in Pharma R&D is an information driven process requiring many disparate bits of data from many different sources, both structured and unstructured. Text mining is the key methodology used to extract entities and relationships from unstructured text in the quest for the knowledge needed to bring a safe and effective drug to market and beyond. Much of the insight needed in early drug research to identify drug target to disease relationships and progress a potential drug target, comes from published literature and internal reports. Later stage drug development requires many additional sources of information including case reports, clinical trials, competitive intelligence and other diverse sources. In this publication, I will present 4 different use cases on how text mining is used to drive decision making in drug discovery and development and also how it can be used to identify patient insights from sources such as social media

Keywords—Text Mining; Drug Discovery; Pharma R&D; Social media; drug safety; patient journey;

I. INTRO TO DRUG DISCOVERY

Drug discovery began with the use of medical plants to treat illness. Later drug discovery began by extracting the pharmacologically active compound from nature products. Accompanying the genomic revolution, modern drug discovery methods shifted to identification of disease associated genes as potential drug targets, followed by discovery of a compound or biologic that would interact favorably with the target to treat disease. Because of the variability of the human population, a drug can vary in efficacy and safety so extensive preclinical and clinical testing is require by the regulatory agencies before a drug is launched on to market. The modern drug discovery process takes place over an average of 10.7 years and at a cost of about 2.6 billion dollars[1,2]. To ensure drug safety, constant post launch, monitoring of a drug or biologic is required. Finally, drugs can and do fail at any stage along the process costing millions or billion or in lost revenue and in some cases, causing harm or even death.

Drug discovery requires a lot of information to succeed and asking the right question can go very far in ensuring a quicker and surer outcome. By providing quicker understanding of disease and find better disease targets, and finding the right compound, and, identifying potential risk earlier in the process to help make it more efficient and shorten the timeline to a safe medicine. We need to select the right dose to maximize efficacy and minimize risk and develop meaningful trials in the right patient populations to ensure success Finally, even after the drug is launched, companies need to monitor for reports of adverse events that arise following drug treatment but also need to understand what patients are saying to minimize risk, understand competing therapies, and alleviate any new issues arising. Because much of this information is present in written reports, published literature or study reports, text mining can help wade thru the pages and help to uncover facts and relationships in unstructured text[3]

II. USE CASES FOR TEXT MINING

A. Text Mining in Early Drug Discovery

Many diseases like diabetes or cancer arise from a complex series of events involving multiple genes and pathways but others, including many rare diseases are associated with a single gene or even a single mutation in that gene. With the help of semantically enriched taxonomies of genes, diseases and drugs and biologics, text mining can identify both old and new relationships between genes and diseases, genes and drugs and drugs and disease[4]. New drug to disease relationships can represent potential repositioning opportunities. Most drug projects are designed to cure diseases that affect large populations but the rare disease patient community is empowered and are pushing for answers, and initiatives like the Orphan drug act offer incentives to address these rare diseases. Understanding these less complicated but rare diseases that often arise from a single gene mutation can shed light on more complex diseases and text mining has great utility in this use case[5.]

B. Mining for Preclinical and Clinical Drug Safety

The ultimate goal of preclinical testing is to accurately model the drug's safety in animals to predict what will happen in humans. The risk for adverse events can vary across different therapeutic areas and some drug classes inherently have

liabilities for certain adverse events[6]. The tolerance of side effect can also vary across therapeutic areas. Because only a fraction of drugs succeeds, there is a large amount of data from failed compound in unstructured internal reports and study documents as well as published literature.

In a small number of cases, unsafe medicines have been progressed into human trials and beyond due to lack of a preclinical safety "signal" in and much is being done to prevent this. In the 1990's, a number of drugs were found to cause a life threatening cardiac arrhythmia caused by QT prolongation and were withdrawn from the market.[7] This has led to testing of all drugs for this liability. One example of how text mining can benefit is in the building of a reference compound set for evaluation of QT prolongation. In 2015, The HESI Pro-Arrhythmia Working Group published on using text mining to identify both human and non-rodent animal studies that assessed QT signal concordance between species and identified drugs that prolonged the QT interval.[8] In this work, text mining was essential to identifying *compound to biological effect to species* relationships in the published literature for expert review.

C. The Role of Text Mining in Drug Submission

The submission of a new drug to the FDA or requires proof that the medicine is safe and effective as demonstrated by non-clinical testing and clinical trials. The submission package can contain thousands of pages of written material. Text mining can support this process in a number of ways and can positively impact the process by saving the time of project teams and potential reviewers.

A real world example of how text mining can impact the submission process will be discussed and while the example has been stripped of proprietary details, it should still demonstrate a tangible value. In this case, the team was filing for a waiver for additional safety studies and based on their knowledge of the drug's pharmacology and that of other drugs in the same class, the team felt this was warranted but still needed to convince the regulatory agency. A keyword based literature search found 3000 full text documents that needed further review to identify the smaller set of documents relevant to the specific question. The completed review and summary report was due in 7 months and an outside vendor quoted a figure of 9 months and 180,000$ to complete the review. Text mining of the full text documents and subsequent review was then completed in 2 months saving 5 months' time and 180,000$. While the monetary impact of text mining and other informatics processes R&D processes can be hard to quantitate, this example demonstrates a clear value of text mining.

D. Post-launch, Mining Social Media for Patient Insights.

Social media is a largely untapped source of information and insights for pharma, on therapy efficacy and safety, patient journey, unmet need, and customer reputation. When a patient receives a life altering disease diagnosis and subsequent treatment, many will turn to social media for support and additional information. As an industry, pharma has been using social media to communicate with patients via channels like Twitter, but this has largely been driven by the commercial to inform the public. While pharmaceutical companies are using social media to provide product information and promotional materials, they should also be using it to better understand patients' needs and experiences, and to provide additional education, particularly to those with chronic illnesses. One emerging trend is to use text mining to analyze sentiments, identify adverse events and glean insights from social media. Social media conversations also can inform R&D and pharmacovigilance efforts.[10] Social media is here to stay and pharma should be responsibly engaging in it to get in touch with patients. When companies engage, and have the right tools and analytics capabilities technologies like text mining they can gain valuable insight into what patients are saying and use those insights to make better treatments that improve the quality of lives.

REFERENCES

[1] J. A DiMasi,, H. G Grabowski, Tufts CSDD briefing on R&D cost study, 2014 http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study

[2] Z. Bian, S. Chen, C. Cheng, J. Wang, H. Xiao, H. Qin, Developing new drugs from annals of Chinese medicine, Acta Pharmaceutica Sinica B 2012;2(1):1–7

[3] R. McEntire, D. Szalkowski, J. Butler, MS Kuo, M Chang, D Freeman, S McQuay, J Patel, M McGlashen, WD Cornell, JJ Xu, Application of an automated natural language processing (NLP) workflow to enable federated search of external biomedical content in drug discovery and development. Drug Discovery Today 2016, 21 (5) :826–835

[4] D. Rebholz-Schuhmann, R. Oellrich A, Hoehndorf Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet. 2012 Dec;13(12):829-39

[5] D Sardana , C Zhu, M Zhang, RC Gudivada, L Yang, AG Jegga. Drug repositioning for orphan diseases. Brief Bioinform (2011) 12 (4):346-356.

[6] Cronin MT1, Jaworska JS, Walker JD, Comber MH, Watts CD, Worth AP." Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ Health Perspect. 2003 Aug; 111(10): 1391–1401.

[7] CE Pollard, N Abi Gerges, MH Bridgland-Taylor, A Easter, TG Hammond, and J-P Valentin "An introduction to QT interval prolongation and non-clinical approaches to assessing and reducing risk". Br J Pharmacol. 2010 Jan; 159(1): 12–21.

[8] HM Vargas, AS Bass, J Koerner, S Matis-Mitchell, MK Pugsley, M Skinner, M Burnham, M Bridgland-Taylor, S Pettit, JP Valentin "Evaluation of drug-induced QT interval prolongation in animal and human studies: a literature review of concordance". Br J Pharmacol. 2015 Aug;172(16):4002-11.

[9] M. Larkin, "Social media for Pharma- an Experts View" 2014 https://www.elsevier.com/connect/social-media-for-pharma-an-experts-view

[10] A Sarker, R Ginn, A Nikfarjam, K O'Connor, K Smith, S Jayaraman, T Upadhaya, G Gonzalez. "Utilizing social media data for pharmacovigilance: A review." J Biomed Inform. 2015 Apr;54:202-12.