*Advances in Social Science, Education and Humanities Research, volume 638*
**Proceedings of the 2021 International Conference on Public Art and Human Development (ICPAHD 2021)**

# The Application of Big Data in Pharmacovigilance: A Systematic Review

Mingshuai Han

*Faculty of Land and Food System, University of British Columbia, Vancouver, British Columbia; V6T 1Z4, Canada*
*Email: hmskevin@outlook.com*

## ABSTRACT

Since the outbreak of coronavirus, public health has become a prevalent topic revolving around people's daily life ever since. When the official announcement that the first mRNA COVID-19 vaccine candidates had successfully completed the clinical trials and was ready to be launched, the public remained deeply skeptical about the unknown contraindications that might cause severe immune compromise after administering the vaccine. Even though the mRNA vaccine was FDA-authorized, the person getting vaccinated could still be affected by adverse events after receiving the first dose. Mild symptoms could manifest quickly, and could be fatal under extreme circumstances after getting the first dose. Adverse drug events (ADEs) are commonly monitored throughout the clinical trials, to minimize any unintended consequences of specific medication use. Epidemiologists have reported a global rise in human diseases over the past few decades. The augmented demand for innovative drugs has naturally boosted drug manufacturing to become one of the most lucrative fields of the healthcare industry. Drug safety surveillance needs to be highlighted given that even a slight overdose could do harm to one's body. As such, post-market drug safety must be monitored strictly and routinely. This paper provides an overview of introducing the existing big data applications in the global pharmacovigilance market, employing algorithmic models to predict ADEs of marketed drugs to effectively improve the therapeutic effect, and challenges presented in applying big data to post-market drug safety monitoring. The specific advantages of employing algorithmic models are similar to the six Vs of big data (volume, velocity, variety, veracity, validity and value) and are pivotal to the policy-making process. This study examines the popular commercially available pharmacovigilance tools that can be found online by investigating their models and services. Then, we delved into several algorithmic models to see the detailed procedure of how big data process complex information.

***Keywords:*** *Big data application, Pharmacovigilance (PV), Adverse Drug Events (ADEs), Post-market drugs, Algorithmic Models*

## 1. INTRODUCTION

The core idea of drug safety monitoring is to collect data from patients in clinical trials and report back to the system to detect any ADEs associated with a specific medication. In fact, a significant number of rare ADEs remaining undetected due to the limited amount of test subjects in clinical trials. The sample size is not large enough to monitor and summarize all the potential ADEs at once. One of the possible solutions is launching a post-market drug safety surveillance program to keep track of ADEs. In terms of pharmacovigilance, it is a relatively new discipline, but will be accelerated with the integration of biomedical informatics, big data analytics, artificial intelligence and machine learning. When the concept of pharmacovigilance was first brought up in the healthcare space, the database was more like a manually organized spreadsheet. Over the past few years of rapid growth, it evolved into an electronic, commercial, highly functional spreadsheet that is capable of sorting and storing a massive amount of adverse event reports from different drug manufacturers [1]. Predictive analysis can be employed to tackle the main issues associated with pharmacovigilance, which are case management, signal management, and benefit-risk management. This paper summarizes the current applications of big data analytics in the field of drug safety monitoring, to provide biomedical professionals with a holistic view of all the tools they are able to utilize. Thus, improving the safe use of postmarket drugs and having a more accurate evaluation of the benefits and risks of a specific medication. The study researches the current commercially available digital tools that provide solutions to post-market drug surveillance and evaluating the pros and cons associated with their service. In addition, this study also looks into the algorithmic models which fundamentally support these tools to be able to perform certain tasks and summarizes each model's main working process.

## 2. THE INCREASING NEEDS FOR PHARMACOVIGILANCE IN POST-MARKET DRUGS USING PREDICTIVE ALGORITHMIC MODELS

Given the ever-changing nature of the current society, data analytics plays a pivotal role in multiple fields by cleaning and analyzing a complex set of data, and eventually systemically extracting valuable information from a hidden data set. The rapidly growing population across the globe generates a huge amount of data in a few seconds. The demand of collecting and storing such a huge volume of data simultaneously has challenged the traditional databases which accelerate the emergence of novel data analyzing tools. Big data analytics is a concept that enables statisticians to deal with an overwhelmingly large set of data and carry out multiple analyses in a short period of time. The adoption of big data analytics systems aims in making data-driven decisions that effectively improve the quality of the outcome. Examining big data is capable of uncovering hidden correlations, and thus, making informed decisions [2]. As big data is mainly composed of machine learning, predictive analytics, data mining and language processing, the multifunctional characteristic of big data has further expanded the scope of its applications and can be used across different disciplines.

The discovery and development of a new drug is a time-consuming process that requires the completion of several phases before the drug gets patented and released to the market. Testing the efficacy of the drug is the aim of doing clinical research and carrying out a series of clinical trials. While the efficacy aspect of the medication might be well-defined after the clinical trials phase, drug safety could take months or even years to be validated. Adverse reactions sometimes occur soon after the drug consumption due to contraindications but in most cases, ADEs could evolve over a span of several months or years due to the long latency. Premarket studies do not guarantee a drug is 100% safe, as the safety profile cannot be completely determined until a more comprehensive postmarket analysis has been done.

The data processing cycle starts with pharmacoepidemiological data collection followed by completing data entry, managing drug safety information, summarizing clinical safety data, signal detection and eventually carrying out benefit-risk analysis. Collecting ADEs reports from hospitals, private clinics and individual patients is the primary step of completing the safety profile of postmarket drugs, and undoubtedly, it is also the most pivotal procedure in the realm of pharmacovigilance. A massive amount of data consisting of ADEs reports will be accumulated due to the time-consuming nature of drug safety monitoring. In order to achieve a more effective and efficient surveillance system, the Food and Drug Administration (FDA), pharmaceutical companies and healthcare organizations can utilize the help of big data analytics, more specifically employing algorithmic models to perform predictive analysis based on data extracted from existing databases. Implementing algorithmic models can help auto-code the drugs, sorting out the adverse event information originating from clinical trials and postmarket drug safety monitoring programs, improve the efficiency of maintaining data, and lastly predict any suspected risks associated with postmarket drugs. Big data analytics provides a way to detect any previously unsuspected safety signals based on time disturbances among datasets.

## 3. EXISTING PHARMACOVIGILANCE TOOLS

### *3.1 The FDA Adverse Event Reporting System (FAERS)*

The FDA Adverse Event Reporting System (FAERS) is one of the publicly available databases that collect ADE reports directly from patients and healthcare providers. Once the reports are delivered to FAERS, drug safety evaluators and healthcare professionals in the Center for Drug Evaluation and Research's (CDER) Office of Surveillance and Epidemiology will closely monitor the safety profile of any suspicious post-market drugs that could potentially cause severe adverse events. If a safety issue is spotted, then FDA will re-evaluate the labeling information and take regulatory actions to protect public health. The FAERS dashboard provides open access to the public and more importantly, therapeutic biological product manufacturers are able to utilize the data obtained in FAERS to enhance the safety of their products. Nonetheless, FAERS has not been fully developed yet which means there are still limitations in achieving an efficient reporting system. First, duplicative reports have been found in certain instances. Extra data files clutter available spaces in the database and lead to decreased data processing rates and reduced capability to pinpoint critical information. Second, it's inevitable to have incomplete ADE reports due to unforeseen reasons, which interrupts the process of signal detecting. Third, information included in the reports has not been medically verified and cannot be viewed as evidence to estimate the occurrence rates of the ADEs and establish causation [3]. Hence, FAERS needs to be improved to meet the ever-changing nature of post-market drug safety assessment.

### 3.2 MedWatch Reporting System

Similar to FAERS, MedWatch is also a program that mainly serves as a tool to report adverse drug events to medical professionals and inform the public about clinical information on medical products. The MedWatch reporting system allows everyone to report any medical injuries or deaths potentially caused by the usage of a certain drug. Patients can voluntarily report any adverse drug reactions to MedWatch by mail using the postage-paid form, by fax, or by phone. It is a government-approved source designed to educate the public about the importance of detecting ADEs and reporting it to get feedback. Despite its advantageous role, the MedWatch reporting system has several flaws that need to be addressed. Underreporting is one of the main issues that leads to a low quantity of ADEs reports being filed. Lack of public awareness of the benefits associated with this spontaneous reporting system hinders MedWatch's ability to achieve its full capacity. In addition, MedWatch is often subject to reporting biases due to the nature of spontaneous reporting. Reporters are easily affected by the promotional claims and medical literature on social media [4].

### 3.3 Arriello

Arriello is another commercial platform specialized in providing global and domestic pharmacovigilance (PV) services. It covers a wide range of functions within the realm of pharmacovigilance. The primary step is to keep track of all the reported ADEs by establishing the Individual Case Safety Reports (ICSR's) which are fully compliant with regulatory guidelines. After data entry, the subsequent step is carrying out MedDRA coding, case evaluation, and finally reporting electronically through EudraVigilance. Arriello is favoured as it provides more personalized solutions to address post-market drug safety issues. For example, its risk management system can detect any new safety concerns that show up under their surveillance with respect to a patient's medication and the medical professionals would then work collaboratively to provide the patient with targeted solutions [5]. Despite the significant progression Arriello has made over the last few years, there are still issues related to their pharmacovigilance platform as a whole. Arriello relies heavily on the expertise of their medical professionals, to detect and manually categorize the ADEs found in the reports, which is labour-intensive and time-consuming. If more big data analytics can be applied and practiced, Arriello has the potential to carry out predictive analysis and stay in the lead of the pharmacovigilance space.

### 3.4 APCER Life Sciences

APCER Life Sciences is a biomedical company that offers pharmacovigilance services and consultancy, which includes aggregate reporting, signal detection, literature search, and risk management. They are specialized in the risk management section known as "pragmatic approaches to risk minimization measures (RMM)". The computative model is mainly composed of risk identification from safety data evaluation, signal detection, and if a risk signal is detected, routine risk minimization measures will be initiated to direct the risk to a series of pharmacovigilance activities [6]. APCER has a matured analytical model, but still has constraints. The program can be more automated in terms of identifying ADEs.

## 4. PREVIOUS AND NOVEL METHODOLOGY BASED ON ALGORITHMIC MODELS

Commercially available tools that serve the purpose of post-market drug safety surveillance are largely based on computational and statistical methods to boost the development of post-market drug pharmacovigilance. This section briefly introduces the previous and newly discovered algorithmic models that have been applied or have the potential to be applied to further strengthen the functionality of pharmacovigilance tools.

### 4.1 The Observational Health Data Sciences and Informatics Common Data Model (OHDSI CDM)

The Observational Health Data Sciences and Informatics Common Data Model (OHDSI CDM) is a new concept that aims to integrate the Spontaneous Reporting System (SRS) and Electronic Health Record (EHR) into one single database that is able to process a larger sample size and cross-validate results. However, the data structures and terminologies used in the two systems are completely different. In order to tackle the problem of inconsistency, this model called ADEpedia-on-OHDSI is developed to convert the data collected in SRS to the OHDSI CDM format. First, all clinical data is extracted from FAERS database and the details of these data are then mapped to a specific table in the OHDSI CDM. Prior to loading the data into the OHDSI CDM format, data conversion and data imputation need to be conducted. This study is meaningful as it effectively incorporates both the spontaneous reporting data and EHR data. Yet, one drawback associated with OHDSI CDM is that during the data conversion process, there is random loss of a portion of medical information which negatively affects the accuracy of the final data loaded into OHDSI CDM [7].

### 4.2 Target Adverse-event (TAE) Profile Model

The machine learning model derived from the previous pharmacological target adverse-event (TAE) profile model is able to extract data from FAERS and use the current data on FDA-approved drugs with a postmarket exposure of three years to predict postmarket adverse events related to a larger set of drugs. The model first determines the pharmacological target of the interested drug and then matches it with the comparator drugs regarding their FAERS reports, literature reports and FDA product labels to generate TAE profiles. Predictions are based on a list of possible adverse events which is called designated medical events (DMEs). Then, TAEs will be analyzed using log-likelihood ratio, associated case count, and proportional reporting ratio. Lastly, based on the lit score for each DME, medical officers in FDA will have a consultation to select the main features and classify potential ADEs. In order to validate and evaluate the overall performance of this model, bigger sample size needs to be taken into consideration. Another possible problem associated with TAE is a false-postive prediction which must be carefully addressed before the model can be a practical tool [8].

### 4.3 Sequence symmetry analysis (SSA)

Sequence symmetry analysis (SSA) is a novel method that uses computerized claims data to detect adverse drug reactions. It was applied initially as a data mining tool for studying side effects of specific medication use and more recently, SSA was developed to generate signals which also supports spontaneous reporting in the field of pharmacovigilance. Generally speaking, SSA examines the sequence of events occurring in relation to a patient's overall use of medication. SSA primarily utilizes a sequence ratio (SR) which is essentially the incidence rate ratio in statistics. The SR is calculated by taking the outcome medication and dividing by the index medication, which gives an approximation of pharmacological outcome over the exposure of drug usage. If there's no association between the medication and the potential ADEs, one would observe a symmetrical distribution of the outcome medication both prior to and after the index medication is initiated. In order to make the estimation more accurate, a hypothetical waiting distribution method is employed to discriminate and only capture incident drug users. It has been proven that with the help of SSA, the rates of ADEs detection have been significantly increased by about 21%. SSA is also computationally efficient as it requires fewer variables to initiate the analysis and is able to perform the calculation given limited data information. Nonetheless, there could be certain factors that might affect the accuracy of SSA. If the use of the outcome medication is increasing, it might lead to an excess of index medications and eventually cause an overestimate of the true incidence rate ratio. One of the solutions to address the problem of prescribing trends is to establish a null-effect SR which demands an overwhelming amount of computational work for an entire population. In order to be fully functional, SSA must provide solutions to accurately identify new users who switched between medications, decide the appropriate exposure time window, prevent protopathic bias and accurately define a signal [9].

## 5. DISCUSSION

### 5.1 Reporting Bias

Undoubtedly, medical professionals will envision a bright future for a comprehensive and mature pharmacovigilance system employing a big data analytics framework. There are still major issues that need to be addressed in order to accelerate the process. Unlike the data collection process from randomized clinical trials or any pre-approval experimental trials which directly obtain required data from test subjects, the cornerstone of post-market drug safety surveillance relies heavily on patients to voluntarily report possible ADEs. Reporting bias is a common issue results in poor signal detection. For instance, patients might report severe or less frequent ADEs than non-serious or common ADEs largely due to their perception of the spontaneous reporting system as a whole. It's difficult to distinguish between unbiased and biased ADEs being reported. When reporting bias happens, the algorithmic model will pick up the false information and carry on its workflow. The current stage of big data analytics cannot successfully discriminate the nature of the reported adverse events.

### 5.2 Complex Calculations

One of the most crucial steps in the cycle of post-market pharmacovigilance procedure is carrying out the benefit-risk analysis. It is essentially a tool to make all the necessary data (i.e. preclinical, clinical and post-market phase) into account and conclude whether a drug's benefits outweigh its associated risks under specific circumstances. Given the immense workload, the algorithmic model has to be structured and functional storage of datasets is demanded. The benefit-risk analysis must be highly precise, as it assists the pharmaceutical professionals in the decision-making processes, such as deciding whether or not to recall a medical product or not. Most data analytic tools are incapable of meeting both the goals of massive storage and selectively retrieving required information from multiple datasets at the same time.

## 6. CONCLUSION

Big data offers a robust scientific framework to assist the development of active drug safety surveillance, allowing more real-time pharmacovigilance practices to be carried out. Most of the current official or commercial applications are based primarily on constructing algorithmic models and are capable of performing case management, signal management, and benefit-risk management. Data analytics helps to maintain the advantages of previous databases while expanding their functionality and pushing the edge of post-market drug safety reporting. The majority of novel methodologies delve into building predictive models to predict the likelihood of a certain adverse event might occurring based on patterns from previous ADEs reports. Given that statisticians are working closely with medical professionals to accelerate the course of digital pharmacovigilance, it's a good sign to have more interdisciplinary practices as new sources of medical information are constantly being considered. Concurrently, there has not yet been agreement about which of the commercially available pharmacovigilance software is most likely to meet the demands of post-marketing drug safety surveillance. This paper only looks at a few representative software and algorithmic models that can be used in post-market drug safety surveillance; there are still a large number of commercially available pharmacovigilance software that needs to be investigated in terms of their role in predicting ADEs. Future studies can shed light on improving the accuracy of drug-ADEs prediction when detecting risk signals. Further enhancements need to be studied in order to continuously strengthen the overall performance.

## REFERENCES

[1] Beninger P. (2018). Pharmacovigilance: An Overview. Clinical therapeutics, 40(12), 1991–2004. https://doi.org/10.1016/j.clinthera.2018.07.012

[2] Chai, W. (2021, February). big data analytics. TechTarget. Retrieved September 13, 2021, from https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics.

[3] Center for Drug Evaluation and Research. (2020, April 2). Postmarketing surveillance programs. U.S. Food and Drug Administration. Retrieved October 3, 2021, from https://www.fda.gov/drugs/surveillance/postmarketing-surveillance-programs.

[4] Berniker, J. S. (n.d.). Spontaneous Reporting Systems: Achieving Less Spontaneity and More Reporting. Why MedWatch is problematic. Retrieved October 4, 2021, from https://dash.harvard.edu/bitstream/handle/1/8846816/Berniker.html?sequence=2&amp;isAllowed=y.

[5] Global Pharmacovigilance and Clinical Safety Services Overview - clinical, regulatory and Pharmacovigilance Services. Arriello. (2021, September 16). Retrieved October 9, 2021, from https://www.arriello.com/services/pharmacovigilance/.

[6] Pragmatic approaches to risk minimization measures. APCER Life Sciences. (2021, March 26). Retrieved October 9, 2021, from https://www.apcerls.com/pragmatic-approaches-to-risk-minimization-measures/.

[7] Yu, Y., Ruddy, K. J., Hong, N., Tsuji, S., Wen, A., Shah, N. D., & Jiang, G. (2019). ADEpedia-on-OHDSI: A next generation pharmacovigilance signal detection platform using the OHDSI common data model. Journal of biomedical informatics, 91, 103119. https://doi.org/10.1016/j.jbi.2019.103119

[8] Schotland, P., Racz, R., Jackson, D. B., Soldatos, T. G., Levin, R., Strauss, D. G., & Burkhart, K. (2021). Target Adverse Event Profiles for Predictive Safety in the Postmarket Setting. Clinical pharmacology and therapeutics, 109(5), 1232–1243. https://doi.org/10.1002/cpt.2074

[9] Lai, E. C., Pratt, N., Hsieh, C. Y., Lin, S. J., Pottegård, A., Roughead, E. E., Kao Yang, Y. H., & Hallas, J. (2017). Sequence symmetry analysis in pharmacovigilance and pharmacoepidemiologic studies. European journal of epidemiology, 32(7), 567–582. https://doi.org/10.1007/s10654-017-0281-8.