

# Big Data Analytics and Pharmacovigilance—An Ethical and Legal Consideration

<sup>1</sup>Rahat Kumar, <sup>2</sup>Pratyush Sharma, <sup>3</sup>MaheshInder Singha, <sup>4</sup>Jaswinder Singh

## ABSTRACT

Drug side effects become once the drugs enter the market. Such type of side effects also called adverse drug reactions (ADRs). ADRs are a major cause of morbidity and mortality among people and present a major health problem worldwide. Thus, what we need is effective post-marketing safety surveillance (or called as pharmacovigilance). In pharmacovigilance, adverse events are obtained from various sources such as drug manufacturers, healthcare professionals, and patients/people consuming these drugs but this presents a problem of lesser reporting as reporting is voluntary amongst these groups. The solution to this problem is the use of big data analytics. Big data” has become a more and more cited term in health care, for the potential use of the huge amount of data collected from digital medical records or administrative data (e.g., drug prescriptions, hospital dismissal forms, healthcare services, etc. and also as a support for regulatory decisions, pharmacovigilance included. Big data also references artificial intelligence, infrastructure, and services, as well as automated processing operations that facilitate the collection, storage, and analysis of data gathered and being produced in ever-increasing quantities. Big data obtained via social media is analyzed using statistics might have considerable errors and biases for individuals not conforming to group characteristics. Utilization of big data in Pharmacovigilance will bring in the potential to complement traditional spontaneous reporting systems, by allowing an epidemiological approach to determine the incidence of adverse events in the population.

**Keywords:** Adverse drug reactions (ADRs), Big data analytics, Ethical considerations, Pharmacovigilance.

**How to cite this article:** Kumar R, Sharma P, Singha M, Singh J. Big Data Analytics and Pharmacovigilance—An ethical and legal Consideration. *Curr Trends Diagn Treat* 2018;2(1):58-65.

**Source of support:** Nil

**Conflict of interest:** None

## INTRODUCTION

Many unintended side effects of drugs become evident once the drugs enter phase IV, i.e. reach the market and is used by the population. Such type of side effects called ADRs poses a significant cause of morbidity and mortality amongst the users. This requires post-marketing surveillance (or called as pharmacovigilance). Pharmacovigilance means the collection and analysis of spontaneous reports of adverse drug events obtained from a variety of sources such as drug manufacturers, healthcare professionals, and patients/people consuming these drugs. With the exclusion of manufacturers, adverse events reporting is voluntary in other groups, and thus the reporting is lesser than expected.

Moreover, the data generated needs to be compiled and analyzed for interpretation, which is a cumbersome job. So a major boost to fuel future pharmacovigilance electronic medical records (EMR), which include any data generated during routine clinical care is recorded on the cloud, and thus the big data generated is analyzed. In the era of easy access to online social networking, many applications have become popular, allowing users to communicate, interact, and share information worldwide.<sup>1</sup> The social media bridges the gap between consumers limited by geographical and social limitations by sharing their experiences in relation to drug efficacy/safety which otherwise becomes difficult in the real world. Patients using social media have made blogs/communities for sharing share their personal experiences with drug use, like efficacy, safety or any untoward experiences while taking treatment. The social media includes social networks, blogs, microblogs and discussion boards, emails, and chats, which are potential sources to explore.<sup>2</sup> Pharmacovigilance programs in the next 10 years has a potential to alter the prescribing pattern of physicians due to additional information on safety and efficacy. However, pharmacovigilance is facing challenges as:<sup>3</sup>

- *Web-based Sales and Information:* The free sale of medicines is available via the internet nowadays especially in the category of nutraceuticals, and this prescription based sale is used by many companies to sell drugs. Such deals cover the wide variety of information on drugs like prescription drugs, unregistered medicines, herbal medicines, Unani

<sup>1</sup>Technical Coordinator cum Professor, <sup>2</sup>Student Big Data Analytics, <sup>3</sup>Professor, <sup>4</sup>Senior Resident

<sup>1,3</sup>Department of Pharmacology, Shri Guru Ram Das Institute of Medical Sciences and Research, Amritsar, Punjab, India

<sup>2</sup>Department of Computer Science Engineering, Chandigarh University, Punjab, India

<sup>4</sup>Department of Radiology, Shri Guru Ram Das Institute of Medical Sciences and Research, Amritsar, Punjab, India

**Corresponding Author:** Rahat Kumar, Technical Coordinator cum Professor, Department of Pharmacology, Shri Guru Ram Das Institute of Medical Sciences and Research, Amritsar, Punjab, India, e-mail: rksharma66@gmail.com

drugs, Siddha system of medicines but there is with questionable safety, efficacy, and quality of these drugs. Unfortunately, feedback is not provided to regulatory agencies involved in the monitoring of ADRs with the use of such drugs. This creates a barrier in ADR reporting and drug usage and hence affects the pharmacovigilance programme.

- *Broader Safety Concerns:* These are due to problems arising from irrational drug use, polypharmacy and overdosage, interactions, drug abuse, self-medication, and use of substandard medicines. Current systems need to be evolved to address this broad scope adequately as to provide feedback about the use of such drugs.
- *Public Health versus Pharmaceutical Industry Economic Growth:* There are conflicting interests within the pharmaceutical industry while dealing with public health as companies' main focus is on profits not on safety. The industry needs to overcome these weaknesses of safety monitoring and need to conduct clinical trials and post-marketing surveillance.
- *Monitoring of Established Products:* The generic drugs do not ensure the good efficacy and safety of its products. The generic sector is the largest supplier of essential drugs, but there is no data available regarding their efficacy and safety versus branded drugs.
- *Attitudes and Perceptions of Drugs to Benefit and Harm:* The perception of benefit and harm and minimum acceptable risk for medicines is not considered in a meaningful way. Thus the harm caused by medicines especially experimentation with drugs has been shown to be significant amongst patients. Unfortunately, this is not reported further and hence affecting pharmacovigilance program.
- *Outcomes and Impact:* No doubt, public awareness over the safety of medicines is increased, and there is an increased awareness of accountability on the part of companies and prescribers. This needs more research into the effectiveness of pharmacovigilance with a major focus to empower health practitioners and patients with useful information that can improve individual therapy. Early diagnosis and management of medicine-induced disease, helps to reduce iatrogenic diseases.<sup>4</sup>

Upcoming new pharmacovigilance data sources:

The various sources categorized are as under:

- (a) *Fully Established:* The various sources are as patients, healthcare professionals, pharmacists, electronic medical records, claims information, spontaneous-reporting system. The problems with these methods are under-reporting, finding difficulty in reporting because of lack/poor connectivity between data evaluators, difficult to find time to send reports due to

the busy schedule of clinicians, lack of understanding about the importance of ADR reporting/lack of culture to report ADRs.

- (b) *Under Development:* Various sources under this category are as web-based, internet search (e.g., Google, Bing), social media (e.g., Facebook, Twitter), patient forums like blogs, groups, etc.

### New Source of Information

No doubt, web-based information is the best source for collecting information on ADRs, but this source is associated with various fallacies. The main problem encountered is over-reporting. This happened with Google for over-reporting of flu infection in public forum. In February 2013, Google flu trends (GFT) made headlines but not for the reason that Google executives or the creators of the flu tracking system observed the same as is shown in (Fig. 1). Nature reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness than the centers for disease control and prevention (CDC) reporting the disease. This was based on the estimates on surveillance reports from laboratories across the United States.<sup>5</sup> A drastically overestimated peak flu level in 2012 was reported and with widespread media coverage. Many flu-related searches by people extrapolated the problem and said the problem in people who were not even ill or were not having flu.<sup>6</sup> Thus blaming drug usage for even minor issues may or may not be associated with drug administration like getting viral influenza/viral infection. Thus to overcome this problem a constant adaptation and recalibration of information is needed.

### Variety of sources and volume of information

Apart from Google, there are various sources of information as is depicted in (Fig. 2). In fact, in July 2015, FDA has agreed to share information with Google and other web-based portals about using data mining to identify unknown drug side effects been reported at these sites.



Fig. 1: Over reporting of flu by Google



Fig. 2: Various sources of information for data mining.

### Traditional Approach of Adverse Event Reporting

Adverse events (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 is reported to the Vigiflow online database information site. Signaling is another data mining approach applied to AEs, showing isolated information about ADRs that can be compiled later on. The sequence of events that are completed before data mining, analysis and reporting is presented in (Figs 3 and 4).

The lead-time derives 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 and the case reported can be simple as a non-serious or life threatening AE. Additionally, non-serious AEs can worsen into a serious AE. The reporting amongst physicians is not mandatory, as it is voluntary among them.

### Current Pharmacovigilance Reporting System

It is presently based on reporting of ADRs into national and international databases like VigiFlow. VigiFlow is an individual case safety report (ICSR) management system developed and hosted by Uppsala Monitoring Centre (UMC). It is compatible with the ICH-E2B standard for electronic transmission of ICSRs. The most national pharmacovigilance reporting system was established in the 1990s. FDA Adverse Event Reporting System (FAERS) contains over 7 million records and reflects data back to 1969 when it was started. About 95% of all cases reported into FAERS received are from manufacturers rather than directly from health-care professionals.<sup>7</sup> The FAERS public dashboard is a highly interactive

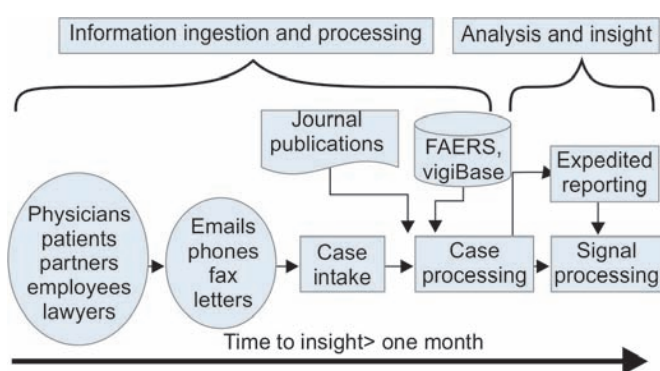


Fig. 3: Displaying the sequence of events and activities for ADRs reporting.

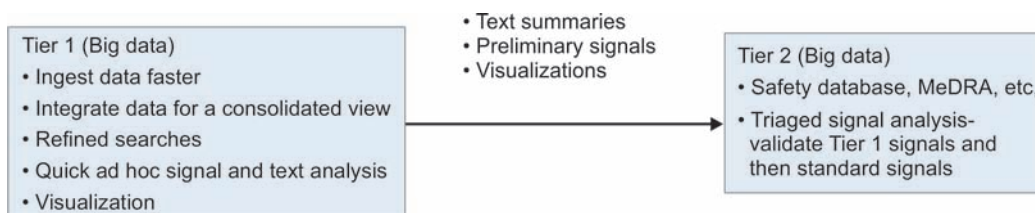


Fig. 4: Two-tiered approach for faster data retrieval/analysis

web-based tool that allows for the querying of FAERS data in a user-friendly fashion. This tool intends to expand access to FAERS data to the public to search for information related to human adverse events reported to the FDA by the pharmaceutical industry, healthcare providers and consumers.<sup>8</sup>

### Legality of Big Data Analytics in ADRs Reporting

Biopharmaceutical companies are seeing a surge for data generated and hence need to identify better targets, better design clinical trials, enable translational science, and better positioning of products in the market. When dealing with data approaching a size large enough to warrant an investigation of alternate methods, quality becomes an important factor. Quality ensures that the newly proposed methods perform just as well as the current “tried-and-true” methods. Big data provides opportunities within the pharma value chain in terms of reporting. Different types of data need to be integrated with this system are mentioned as under:

- (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 both structured and unstructured data
- (d) Published literature on safety and efficacy
- (e) Data from social media (Twitter, Facebook, blogs). However this data is not as reliable as others, but it can help in providing the earliest warning signals.

The FDA recently launched Sentinel System, a national electronic system to track the safety of products in the market. It allows the FDA to inquire about automated healthcare data holders like electronic medical records (EMRs), insurance claims and registries and safety issues. Big Data is a broad term that refers to large-volume, complex, growing data sets with multiple, autonomous sources.<sup>9</sup> According to guidelines for small and medium businesses, qualitative signal detection methods include ‘case-by-case manual review’ without comparison with cumulative data. While ‘quantitative methods’ utilize statistical analysis to identify drug-event pairs that occur with disproportionately high frequency, typically using the proportional reporting ratio (PRR) or the empirical Bayesian geometric mean (EBGM).<sup>10</sup> This large generation of data requires the pooling of the data to analyze for further interpretation. Big data analytics can be very helpful in analyzing and interpretation of such data using various tools. Four vs. as Volume, Velocity, Variety, identify big data and Veracity in context to medical data generated.

- *Volume*: Data available in radiology departments is high in volume as we can note by image size of computed tomography (CT), angiography, radiography, and mammography images. Data sets are particularly large for cardiac CT studies and angiographic studies.
- *Velocity*: This means that data are generated at what speed? Radiology data gets churned at high speed as CT and magnetic resonance imaging (MRI) scanners are already pushing freshly acquired data in Picture archiving and communication system (PACS); subsequently, images are stored in short and long-term storage using
- *Vendor Neutral Archives (VNA)*: This data gets generated in real time.
- *Variety*: Radiology data is rich in variety, as the images come from various sources; these can be from digital radiography, computed radiography, conventional radiography, interventional radiology, MRI, ultrasound, color doppler, and positron emission tomography-computed tomography (PET-CT) studies.
- *Veracity*: This relates to the authenticity and credibility of the dataset. This must be the aim of an ideal big data project; however, there is also a need to do a proper and systematic analysis of data to get accurate results. Therefore, suboptimal scan scans with motion artifacts, and incomplete studies can be deleted from the study group by quality checks. This will help in maintaining the uniformity of data sets. To study such vast data and review these for clinically relevant and useful information using standard algorithms is unthinkable considering the time required to search for useful information. However, by applying different algorithmic tools and converting raw data to transformed data from such large datasets, there is a possibility of understanding and using such data for gaining new knowledge and insights.

“Big data” has become a more and more cited term in health care, for the potential use of the huge amount of data collected from digital medical records or administrative data (e.g., drug prescriptions, hospital dismissal forms, healthcare services, etc.) and also as a support for regulatory decisions, pharmacovigilance included.

### Ethical Consideration

In the last decade, there has been a growing acknowledgment internationally that government bodies responsible for ensuring the safety and effectiveness of medicines have faced serious challenges when protecting the public from harm once people in the uncontrolled, real-world context<sup>11</sup> use these products. Thus, a more detailed ethical analysis

needs to be considered while conducting post-market drug monitoring and regulatory activities of the drug in the market. The various considerations are as under:

- Protection of the patients from the harmful effects of medicines
- Generating quality data of the correlation of the ADRs with drug treatment
- Anticipate in decision-making for the action plan in case of ADRs
- Transparency in maintaining and reporting of ADRs,
- Establish a risk-benefit profile and clearly explain to the patients,
- Minimization of risk and maximization of the benefit to patients,
- Identify the vulnerable populations with high risk of ADRs.

Thus, it is established that rational deliberation<sup>12</sup> must occur to ensure that the ethical tensions are acknowledged and addressed, and the remedial measures for considerations need not be exhaustive or decisive.

### Sources and Benefits of Using Big Data in Biopharmaceuticals

Big data for biopharmaceuticals is generated both from internal and external sources. The aim of using big data is to convert unstructured data to wisdom after extracting knowledge and information from given data.

(a) *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. This data comes from internal experiments in laboratories, and from patients and stakeholders.

(b) *External Scientific Data*: This includes patient-related data: The data from central laboratories, prescriptions, claims, electronic health records (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 a 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.

- *Social Media*: Patients are using Twitter, Facebook, and blogs to communicate their outcomes and impressions of the drug.

Big data analytics a helping hand in the process of analysis of complex data: This consists of networks between the various CT, MRI, Ultrasound, and Radiography equipment's and patient order entry systems, which are the main source of patient image data.

- *Cloud*: It means storing radiological data on offshore (remote) servers hosted by proprietary internet networks and connecting them to the local hospital computers for fast access, processing and distribution of large data.
- *Cyber*: It describes the computer processing power and memory that will be required to process the query to obtain specific answers. The complex query requires large mainframe computers and higher processing power to get desired results in real time.
- *Content/context*: This refers to the digital imaging and communications in medicine (DICOM) datasets, which can be searched, and information derived from getting relevant, meaningful results and correlation, which can help to alter patient management promptly.
- *Community*: This implies to sharing data of a particular and sensitive nature with similar available data in other health care institutes to determine consistency and to get the required information. This collaboration and sharing of radiology data can be of special interest in infectious diseases such as chest radiographic findings in H<sub>1</sub>N<sub>1</sub> influenza. Big data can be of potential help in early recognition of such imaging findings.
- *Customization*: Radiology structured, non-structured data query should be customized, and algorithms should be available on demand to answer specific queries to make this technology of clinical relevance and drive home the future of personalized healthcare. In this manner, big data can add value to radiology.
- *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.
- 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
- 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 as followings:
  - (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

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. The data from social media is available faster than data from published journals, but maybe 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 reported faster than those from published journals, but the level of research is less reliable than the source journals.

### Intricacies of Big Data

The global technological prowess and per-capita capacity to save digital information has roughly doubled every 40 months since the 1980s.<sup>13</sup> Since 2012, every single day 2.5 exabytes ( $2.5 \times 10$ ) of new data has been created, and as of 2014, each day 2.3 zettabytes ( $2.3 \times 10$ ) of data has been generated.<sup>14</sup> Current research estimates approximately 2.5 quintillion bytes of data being created each day.<sup>15</sup> Processing large data using the current relational database management systems, standard desktop statistics and visualization dashboards is a near impossible task. Instead, the need of the hour is using “massively parallel software deployed on a large number of servers.”

### Data Privacy in Big Data Analytics

Big data is obtained through social media data and is analyzed using statistics that identify the patterns for groups.<sup>16</sup> These insights might have significant errors and biases for individuals not conforming to group characteristics. Furthermore, when inferences are made about particular individuals poor quality data algorithms may not be well understood by decision-makers and hence resulting in complex ethical issues result. This could, for example, be problematic if incorrect data or an algorithm led to an unethically discriminatory decision.

On the contrary, reducing big data to a limited visual presentation that lacks information about sources and quality of data can also lead to incorrect unethical decisions.<sup>17</sup> Frequently, when big data informs decision-making, the underlying justifications of the decision cannot be easily understood. Questions related to the responsibility of human decision making arise. Data privacy is an important issue whenever information is posted allowing easy access and use of this data for research purposes. The users should be responsible for preserving the anonymity of the information pertaining to their identity. There are still ethical issues in using social media for the extraction of new medical knowledge, the most important being that each patient posting on a forum should be guaranteed to keep his anonymity.<sup>18</sup> While participation in society is becoming more and more dependent on using apps, social networks and sensors, the social evolution and awareness behind big data analytics is very slow. Current principles and guidelines for protecting an individual's privacy rights lag behind technological developments. Similarly, no guidelines are indicating if, or how, unethical consequences such as discriminations against individuals can be avoided. For example, health population data, matched with genome data, can be used to infer the relationship between gene sequences and disease. Societies need to decide how to balance these benefits with the potential loss of privacy or autonomy.<sup>19</sup> Considering data extraction for pharmacovigilance; anonymization was the main solution to guarantee data privacy such as implemented by Benton et al.<sup>20</sup> However, despite the anonymization process, patients may be identified as a posteriori as per Zimmer;<sup>21</sup> his analysis of a public release of anonymized Facebook data shows that a special mechanism should be implemented to limit access only to authorized personnel. This is the reason why we designed procedures to guarantee that only data going through a step of data minimization should be accessible for analysis by registered end users, but we kept raw data accessible in very specific circumstances to allow contacting the patient if drug withdrawal was required for safety reasons. Concerning privacy and confidentiality, much work is needed in terms of formulating guidelines to help drug developers understand the current thinking about the extent and nature of evidence from big data that would be deemed admissible in the drug approval process.<sup>22,23</sup> With respect to the quality of the data, there is a vibrant debate in the scientific community regarding whether real-world data is of sufficient quality for evidence-based medicine. If it is not, then many believe that the issue of GIGO (e.g., garbage in–garbage out) applies. With proper analytical, computational and data management

tools, big data analytics is useful and can support health policy decision-making. The maintenance of EHRs is increasingly becoming standard practice in health care. The primary purpose of EHRs is to store patient information that is used in the clinical care of the patient. However, EHR data from a large number of patients can be pooled together being linked to other databases and is queried with important scientific questions. To date, however, EHR data are under-utilized in health research and public health practice relative to the benefits they are capable of generating.

### Applications of Big Data in Shaping Healthcare Segment

Big data creates a tremendous opportunity for the world economy not only in the field of national security but also in areas ranging from marketing and credit risk analysis to medical research and urban planning. At the same time, the extraordinary benefits of big data are tempered by concerns over privacy and data protection. Though these are majorly potential big data application; boundaries are blurred as it is simultaneously working in tandem with other applications such as cloud computing, data mining, deep learning, and computer-aided diagnosis (CAD). A bigger variety of data would also result in better prediction of efficacious and safe therapeutic interventions, potentially resulting in the speedy elimination of unsafe drugs from the market based on statistically valid evidence, limitation of certain treatments to specific populations, which can benefit from their use, and expedited approval of new indications commonly used off-label. The expectation from big data is that it may ultimately lead to better and more informed decisions. For instance, big data may bring better insights in scientific and medical research, increased self-knowledge for individuals, services and medical treatments that are more personalized and thus better suited to the individual.<sup>24</sup>

### Indian Pharma Industries and Big Data Analytics for Pharmacovigilance

Indian pharma industry is now increasingly looking to bring Big Data Analytics into the pharmacovigilance space. This is to simplify the large volume of structured and unstructured information, which is difficult to process by putting in place simple database management. Big data also references artificial intelligence, infrastructure, and services, as well as automated processing operations that facilitate the collection, storage and analysis of data gathered and is being produced in ever-increasing quantities. Spontaneous reporting systems (SRSs) are inalienable components of pharmacovigilance, but on its

own can never offer a complete picture of patient safety information. The data collected often has limited patient information including medical histories, concomitant treatments, and pre-existing illness conditions. There is under-reporting of adverse reactions with both voluntary and mandatory surveillance systems. The reporting rates may vary widely for drugs. Utilization of big data in pharmacovigilance will bring in the potential to complement traditional spontaneous reporting systems, by allowing an epidemiological approach to determine the incidence of adverse events in the population. Enhanced ability to identify and investigate adverse drug reactions, which occur over a longer time, may not be identified. However, there is more significant potential for the investigation of signals across different subpopulations.

### REFERENCES

1. Harpaz R, DuMochel W, Shah N. Big Data and Adverse Drug Reaction Detection. *Clinical Pharmacology & Therapeutics*. 2016;99(3):268-269.
2. Cedric B, Badisse D, Sylvie G L, Stefan J D, Carole F, Charles H, Sandrine K, Vincent L, Suzanne P, Christophe R, Stéphane S. The Adverse Drug Reactions from Patient Reports in Social Media Project: Five Major Challenges to Overcome to Operationalize Analysis and Efficiently Support Pharmacovigilance Process. *JMIR Res Protoc* 2017;6(9):e179.
3. Jeetu G, Anusha G. Pharmacovigilance: A Worldwide Master Key for Drug Safety Monitoring. *J Young Pharm*. 2010 Jul-Sep;2(3):315-320.
4. Waller PC, Wood SM, Langman MJ, Breckenridge AM, Rawlins MD. Review of company postmarketing surveillance studies. *BMJ*. 1992 Jun 6;304(6840):1470-1472.
5. Butler D. Google Flu Trends. *Nature* 2013; 494:155.
6. Butler D. When Google got flu wrong. US outbreak foxes a leading webbased method for tracking seasonal flu. Source: <https://www.nature.com/news/whengoogle-got-flu-wrong-1.12413>. (Last accessed on 27-02-2018.)
7. U.S Food and Drug Administration. Transcript: FDA's Adverse Event Reporting System (FAERS). Source: <https://www.fda.gov/Drugs/NewsEvents/ucm577118.htm>. (Last accessed on 27-02-2018.)
8. The Sentinel Initiative National Strategy for Monitoring Medical Product Safety. May 2008. U.S. Food and Drug Administration Office of Critical Path Programs. Source: <https://www.fda.gov/downloads/safety/fdassentinelinitiative/ucm124701.pdf> (last accessed on 22-03-2018.)
9. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering [Internet]*. Institute of Electrical and Electronics Engineers (IEEE); 2014 Jan;26(1):97-107.
10. Shibata, Atsuko, and Manfred Hauben. "Pharmacovigilance, Signal Detection and Signal Intelligence Overview." Chicago, Illinois, USA, 2011.
11. Arlett P, Portier G, de Lisa R, Blake K, Wathion N, Dogne J-M, et al. Proactively managing the risk of marketed drugs: experience with the EMA Pharmacovigilance Risk Assessment Committee. *Nature Reviews Drug Discovery [Internet]*. Springer Nature; 2014 Apr 22;13(5):395-397.

12. Upshur RE. Principles for the justification of public health intervention. *Can J Public Health*. 2002 Mar-Apr;93(2):101-103.
13. Hilbert M, López P. The world's technological capacity to store, communicate, and compute information. *Science*. 2011 Apr 1;332(6025):60-65.
14. IBM What is big data? – Bringing big data to the enterprise. [Last accessed on 28 May 2018]. Available from: [www.ibm.com](http://www.ibm.com).
15. Vcloudnews.com. (2017). Every Day Big Data Statistics – 2.5 Quintillion Bytes of Data Created Daily. [online] [Last accessed on 27 Mar 2017]. Available at: <http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillionbytes-of-data-created-daily/>
16. Wigan MR, Clarke R. Big data's big unintended consequences. *Computer*. 2013 Jun;46(6):46-53.
17. Ekbia H, Mattioli M, Kouper I, Arave G, Ghazinejad A, Bowman T, et al. Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology* [Internet]. Wiley; 2014 Dec 31;66(8):1523-1545.
18. McKee R. Ethical issues in using social media for health and health care research. *Health Policy* [Internet]. Elsevier BV; 2013 May;110(2-3):298-301.
19. Newell S, Marabelli M. Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems*. 2015 Mar 1;24(1):3-14.
20. Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, Leonard CE, Holmes JH. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. *J Biomed Inform*. 2011 Dec;44(6):989-996.
21. Zimmer M. But the data is already public? On the ethics of research in Facebook. *Ethics Inf Technol*. 2010;12:25.
22. Julie Brill. Federal Trade Commission. Protecting Consumer Privacy in an Era of Rapid Change: A Discussion of the FTC's Privacy Report. <https://www.ftc.gov/public-statements/2012/03/protecting-consumer-privacy-era-rapid-change-discussion-ftcs-privacy> (last accessed on 27-08-2018).
23. European Parliament. Council of the European Union E. U. Directive, 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Off. J. EC*. 1995;23(6):348/74-348/99.
24. Opinion 7/2015. Meeting the challenges of big data. European Data Protection Supervisor. Source: [https://edps.europa.eu/sites/edp/files/publication/15-11-19\\_big\\_data\\_en.pdf](https://edps.europa.eu/sites/edp/files/publication/15-11-19_big_data_en.pdf) accessed on 26-06-2018.