Big Data and Pharmacovigilance: Data Mining for Adverse Drug Events and Interactions

C. Lee Ventola, MS

INTRODUCTION

Adverse drug events (ADEs), including drug interactions, have a tremendous impact on patient health and generate substantial health care costs.[1–9] A "big data" approach to pharmacovigilance involves the identification of drug–ADE associations by data mining various electronic sources, including: adverse event reports, the medical literature, electronic health records (EHRs), and social media.[1–4,10–22] This approach has been useful in assisting the Food and Drug Administration (FDA) and other regulatory agencies in monitoring and decision-making regarding drug safety.[1–4,6,10,23–26] Data mining can also assist pharmaceutical companies in drug safety surveillance efforts, adhering to risk management plans, and gathering real-world evidence to supplement clinical trial data.[2,3,10,27–35] The use of data mining for pharmacovigilance purposes provides many unique benefits; however, it also presents many challenges.[1–4,10,11,36–39] Various steps can be taken to improve the use of data mining for pharmacovigilance purposes in the future.[1–4,10,11]

NEED FOR PHARMACOVIGILANCE

The primary goal of drug safety regulators and researchers is to identify and observe ADEs that can cause public harm.[3] Many ADEs are identified only *after* a drug has been marketed when it is used by a larger and more diverse population than during clinical trials.[1,4] ADEs discovered after a drug is in broad use can be a significant cause of morbidity and mortality, so effective post-marketing drug safety surveillance is critical to the protection of public health.[1,3,4]

A new drug is granted regulatory approval only after its efficacy and safety have been demonstrated in a series of clinical trials.[4] Randomized, controlled, phase 3 studies are considered to be the most rigorous means for studying a drug's efficacy and safety.[4] However, these trials often enroll a relatively small number of patients according to specific inclusion and exclusion criteria that do not always represent all potential users of the drug.[3,4] Clinical trials also take place over a relatively short period, making ADEs with a long latency difficult to detect.[4] Furthermore, after regulatory approval, drug labeling and/or prescribing practices may evolve to include new indications or patient populations, off-label uses, or concomitant use with other drugs.[4] Each of these new variables may contribute to the development of ADEs that had not been observed previously during clinical trials.[4] Even over-the-counter medications, such as nonsteroidal anti-inflammatory drugs and phenylpropanolamine, have been associated with confirmed adverse drug reactions (ADRs) after regulatory approval, causing withdrawal from the market or changes in labeling.[6,23,24]

Data mining drug safety report databases, the medical literature, and other digital resources could play an important role in augmenting the information about ADEs that is obtained during short-term clinical trials.[3] Data mining for pharmacovigilance purposes may also provide an "early warning system" that could detect drug safety issues more promptly than traditional methods. For these reasons, data mining these sources for ADEs is of great interest to the FDA, the pharmaceutical industry, and drug safety researchers.[3]

The author is a consultant medical writer living in New Jersey.

BIG DATA AND PHARMACOVIGILANCE

What Is Big Data?

The term "big data" refers to a large volume of diverse, dynamic, distributed structured or unstructured data that provides both opportunities and challenges with respect to its interpretation due to its complexity, content, and size.[1,11] Traditional methods are often inadequate for processing big data because the volume of data is so large and complex.[11] Besides vast volume and variety, other features of big data include its rapid speed of accumulation and transmission.[11] A glossary of terms pertaining to big data, data mining, and pharmacovigilance is provided on the following page.

The digital revolution introduced advanced computing capabilities, spurring the interest of regulatory agencies, pharmaceutical companies, and researchers in using big data to monitor and study drug safety.[11] Significant improvements in computing power and speed have allowed the automation of drug safety surveillance signal detection in large complex databases.[4] Previously unavailable, novel sources of real-world evidence and experimental data in digital form have also become available for pharmacovigilance purposes.[1] The confluence of these events has spurred the development of automated, quantitative big data methods to analyze ADE reports to supplement and complement traditional qualitative pharmacovigilance methods.[4]

The use of big data for pharmacovigilance involves novel electronic methods that are applied to analyze the large and growing volume of information about ADEs in spontaneous reporting system (SRS) databases and other digital sources.[11] SRS databases are repositories for spontaneous reports of ADEs made to regulatory agencies by health care professionals (HCPs), consumers, medical product companies, and other sources.[11] These methods focus on the rapid electronic identification of possibly related discrete data points that would be nearly impossible to detect through a conventional manual search.[11] Big data methods are used to analyze information patterns within datasets to identify new associations among drugs, ADEs, and risk factors.[11] Because this is the purpose of pharmacovigilance, the appeal of using big data for drug safety surveillance is evident.[11]

Signal Detection

What Is a "Signal"?

The Council for International Organizations of Medical Sciences (CIOMS) has issued a definition of a drug safety sur-

GLOSSARY[40–44]

Adverse drug event—Any untoward medical occurrence in a patient or clinical investigation subject who has been administered a pharmaceutical product, which does not necessarily have a causal relationship with the treatment. Any unfavorable and unintended sign, symptom, or disease temporally associated with the use of a medicinal product, whether or not related to that product.

Adverse drug reaction—A response to a drug that is noxious and unintended and that occurs at doses normally used in humans for prophylaxis, diagnosis, or therapy of disease or for modification of physiological function.

Algorithm—A mathematical formula typically made up of a series of calculations that is placed in software to perform an analysis on a set of data with the goal of solving a specific problem.

Artificial intelligence—The development of machines and software that can perceive a situation and assess the appropriate action to take when required, and can even learn from that action.

Bias—A systematic error due to any of a variety of factors (e.g., incorrect premise, small sample size, etc.) that can falsify or distort study results.

Big data—A massive volume of structured and unstructured data that is too large, complex, and/or varied for analysis by traditional processing methods, but may have potential to be data mined for valuable information.

Co-occurrence-based method—A method based on the fact that two or more events or circumstances occurred or existed simultaneously.

Confounding variable—An extraneous factor, in addition to the factors being studied, which may have influenced study outcome. Not accounting for confounding variables decreases the validity of a study.

Data mining—An analytical process where large datasets are analyzed or "mined" in search of meaningful patterns, relationships, or insights.

Machine learning—Algorithms that analyze inputted data for the purpose of learning to make decisions based on new, not yet seen data, patterns, or events.

Natural language processing—The ability of a computer program to understand human speech as it is naturally spoken. It is based on computer science, artificial intelligence, and computational linguistics, usually for the specific goal of programming computers to effectively analyze a large volume of natural language content.

Signal validation—The process of evaluating the data or documentation that supports a detected signal. This is done to verify that there is sufficient evidence to demonstrate that a potential causal association exists, justifying further analysis of the signal.

Spontaneous report—An unsolicited communication, usually to a regulatory authority or a drug company, that describes an adverse drug event in a patient who has taken one or more medicinal products who is not involved in a clinical study or other type of organized scheme.

Structured data—Any data that has been organized into structured fields, such as a database or spreadsheet, so that it can be easily processed or analyzed.

Text mining—The application of statistical, linguistic, and machine-learning methods on text-based sources to derive meaning or insight.

Unstructured data—Information that has not been organized in a predefined manner so that it may be stored in structured data fields in a database. Examples include text, images, audio, and video.

Validated signal—A signal for which the signal validation process has demonstrated sufficient evidence for a causal association, and therefore further analysis of the signal is justified.

veillance signal.[4] CIOMS was established by the World Health Organization (WHO) and the United Nations Educational Scientific and Cultural Organization to establish guidelines for the international biomedical community concerning ethics, product development, and pharmacovigilance.[4,12] The CIOMS definition of a pharmacovigilance signal is:

> (i) It is based on information from one or more sources (including observations and experiments), suggesting an association (either adverse or beneficial) between a drug or intervention and an event or set of related events (e.g., a syndrome); (ii) it represents an association that is new and important, or a new aspect of a known association, and has not been previously investigated and refuted; and (iii) it demands investigation, being judged to be of sufficient likelihood to justify verificatory and, when necessary, remedial actions.[12]

As this definition states, once a pharmacovigilance signal is generated, it must be verified.[4] For this reason, while data mining can be used to detect potential signals and may imply hypotheses related to those signals, it cannot by itself prove a direct causal relationship between a drug and an ADE.[2] Other sources of safety data (i.e., clinical trial data, the medical literature, and others) must be analyzed to confirm the clinical significance of a pharmacovigilance signal generated by data mining.[2] If the signal is verified and evidence of causality between a drug and an ADE is established, the FDA may issue a recall, change a drug's labeling, or withdraw a medication from the market.[2] The need to evaluate other types of data to verify data-mining signals was reinforced when the FDA detected a drug–ADE association between statins and amyotrophic lateral sclerosis (ALS).[13] Although data mining had identified 91 reports of this potential drug–ADE association, data obtained in 41 clinical trials representing 200,000 patient-years of exposure did not validate this observation.[13] The clinical trial data reported nine cases of ALS for statin-treated patients and 10 for placebo-treated patients; therefore, it did not confirm an increased incidence of ALS in patients taking statins.[13]

Methods Used to Detect Pharmacovigilance Signals

Many statistical methods have been developed for data mining drug safety signals in SRS databases and other sources.[3,4] The approaches that are most often applied are "disproportionality methods" and "text mining."[3,4] A brief description follows of these and other methods that may be used for drug safety surveillance data mining. Table 1 provides examples of methods used to data mine various data sources for drug safety signals.

Disproportionality

The primary method used for data mining SRS reports employs algorithms that perform "disproportionality analyses."[3,4] These are based on statistical calculations that detect drug–ADE associations that occur at higher-than-expected frequencies.[2,4] These methods compare the actual count for an association between a drug and an ADE (or drug *combination* and ADE for drug–drug interactions [DDIs]) with the

Table 1 Examples of Data Sources and Methods Used in Data Mining for Pharmacovigilance[2–4,14,17]

Type of Data	Stage	Method	Tool	Purpose
Observational				
SRS reports	Routine	MGPS	Empirica Signal	Identify drug–ADE signals
	Developmental	Text mining	Vaccine Adverse Event Text Mining system	Identify vaccine–ADE signals; plot temporal association
	Developmental	MGPS	MASE	Molecular analysis of drug–ADE association
	Developmental	GIS	Various investigational	Manage drug safety data temporally and geographically
Electronic health records	Developmental	Text mining	Various investigational	Identify drug–ADE signals
Medical Literature				
MEDLINE	Developmental	Text mining	MeSH	Identify drug–ADE signals
Medical literature	Developmental	NLP	Linguamatics I2E	Predict drug–ADE associations based on chemical structure
	Developmental	Text mining	G-VISR	Identify vaccine–ADE signals; examine molecular information
Internet				
Social media	Developmental	Text mining	MedWatcher Social	Identify drug–ADE signals
Other Internet sources (websites, search logs, etc.)	Developmental	Text mining	Various investigational	Identify drug–ADE signals

ADE = adverse drug event; GIS = geographical information systems; G-VISR = Georgetown Vaccine Information and Safety Resources; MeSH = medical subject headings; MGPS = multi-item gamma-Poisson shrinker; NLP = natural language processing; SRS = spontaneous reporting system.

background count for the ADE for all other drugs or drug combinations in the database; this produces a proportional reporting ratio (PRR).[2,4] The PRR indicates the degree of disproportionality occurring for a drug–ADE association, compared with all other products in the database.[2,4] If a high number of drug–ADE associations are detected compared with the background count, a signal for a potential cause–effect relationship between a drug and ADE has been detected.[2]

Although this approach is effective, disproportionality methods don't adjust for small numbers of drug–ADE associations that may be present for specific drugs or in the entire database.[2] In these situations, more advanced statistical methods are used, such as the multi-item gamma-Poisson shrinker (MGPS) or the Bayesian confidence propagation neural network (BCPNN); these methods are used by the FDA and the WHO, respectively, to analyze SRS databases.[2,4]

Several data-mining software programs for pharmacovigilance purposes that can generate PRR and/or MGPS scores are commercially available, such as Empirica Signal (Oracle Corp., Redwood Shores, California), PV-Analyzer (Ennov USA, San Jose, California), and SAS (SAS Institute, Inc., Cary, North Carolina).[2] The FDA is also evaluating and advising on the development of a proprietary software tool called Molecular Analysis of Side Effects or MASE (Molecular Health, Inc., Boston, Massachusetts).[2] This tool integrates ADE report data with various chemical and biological data sources in a drug-specific manner to create a "molecular fingerprint" that can be used to evaluate the biological plausibility of drug–ADE association signals.[2,4] This software also identifies transporters, enzymes, and targets that are disproportionately associated with drug–ADE associations and proposes molecular target and enzymatic pathway hypotheses for further investigation.[2]

Text Mining

A large volume of "unstructured" or "narrative" data is often present in the text submitted in ADE reports, which requires analysis using "text mining."[2] Examples of unstructured data that can be mined for pharmacovigilance purposes include event descriptions or narratives in EHRs, the medical literature, social media, or the Internet.[1]

The FDA has developed a tool called the Vaccine Adverse Event Text Mining (VaeTM) system to detect drug–ADE associations present in the text in the SRS database that it maintains for vaccines, the Vaccine Adverse Event Reporting System (VAERS).[2] This software uses "rules" to extract diagnostic (e.g., laboratory test results), treatment, and temporal information from VAERS.[2] The extracted information is then plotted on a time axis, providing a temporal view of the development of ADEs following the administration of vaccines.[2] This feature may also be used with drug SRS databases.[2] The FDA also developed a search and retrieval framework (SARF) that can use text mining to screen for lists of items within a number of informational repositories.[2] SARF incorporates general-purpose dictionaries and state-of-the-art ontologies maintained by the FDA and National Library of Medicine (NLM).[2]

Natural Language Processing

Techniques involving natural language processing (NLP) are commonly used to mine the published medical literature for pharmacovigilance signals.[3] The main systems that employ NLP extract potential relationships from text through the use of machine-learning, rule-based, and co-occurrence-based approaches.[3] Text-mining methods that involve NLP are cost-effective and can be used for both the prediction and detection of drug–ADE relationships.[3]

Change-Point Analysis

Change-point analysis (CPA) is a statistical method used for examining large databases to determine whether a change in variability or slope has taken place in a time series or sequence.[2] The FDA can use CPA as a public health surveillance tool that can detect the longitudinal effects of ADEs and drug recalls.[2]

Geographical Information Systems Technology

Geographical information systems technology enables analysts to capture, analyze, and manage drug safety data temporally and geographically, allowing real-time identification and intervention.[2] This technology can be used to identify at-risk populations, product contamination patterns, and areas where public health education or assistance may be needed.[2]

Visualization Tools

Graphical tools, such as heat or sector maps, may be used to create a visual representation of large, complex bodies of data.[2] These tools can then be applied to visually display subgroups of related products and outcomes to assist drug safety researchers in making comparisons.[2]

Sources for Data Mining for ADEs

SRS Databases

To meet its safety monitoring responsibilities, the FDA collects and maintains SRS databases for the products it regulates.[2] The agency receives approximately 1.5 million ADE, product complaint, and user error reports each year from HCPs, consumers, companies, and other sources, concerning drugs, vaccines, and medical devices for human use.[1–3] Such reports are entered into SRS databases for each product type (drugs, vaccines, or medical devices) so that they may be analyzed to identify possible safety issues.[1,2] The number of safety reports made to the FDA annually is continuously expanding due to increases in the type and number of products the agency regulates, awareness of the importance of these reports, ease of submitting reports (i.e., digitally), and a larger population.[2]

Spontaneous reports that the FDA receives regarding drug ADEs are entered into the FDA's Adverse Event Reporting System (FAERS).[3] These reports are comprised of real-world data about suspected safety issues regarding drugs.[4] SRS reports generally include information concerning the patient, drug, event, and concomitant drugs that may have caused or contributed to the ADE.[3] Therefore, the analysis of individual or a case series of SRS reports may be an important source of information to identify potential safety concerns.[4] With the exception of pharmaceutical companies, submitting reports of suspected ADEs to FDA's MedWatch program is voluntary; this is why the process of collecting SRS reports is often described as "passive."[1]

The FAERs and VAERs databases, the European Medicines Agency's EudraVigilance, and WHO's VigiBase are among the largest SRS databases worldwide (see Table 2).[3,4] VigiBase is the largest, with more than 15 million reports from more than 100 countries.[3] More than 70 countries have established national SRS databases similar to FAERs.[4] Although each

Table 2 Major Spontaneous Reporting System Databases[2–4,45,46,47]

Name	Region	Catchment Period	Total Number of Reports	Report Sources	Report Content
FDA Adverse Event Reporting System (FAERS)	Mostly United States	1969–present	> 9,000,000	HCPs, pharmaceutical companies, patients, consumers	• Mandatory post-marketing ADE reports submitted by pharmaceutical companies • Voluntary ADE reports from HCPs and the public (via MedWatch)
FDA Vaccine Adverse Event Reporting System (VAERS)	United States	1990–present	> 640,000	HCPs, pharmaceutical companies, patients, consumers	• Reports of ADEs concerning vaccines
European Medicines Agency Eudravigilance	European Union	2001–present	> 1,240,000	Marketing authorization holders	• Reports of suspected ADEs concerning medical products
World Health Organization VigiBase	Worldwide*	1968–present	> 15,000,000	National pharmacovigilance centers	• Reports of suspected ADEs concerning medical products • Reports of ADEs from studies or other social monitoring situations

* Although more than 100 member countries participate, the majority of reports come from the U.S. and European Union.

ADE = adverse drug event; FDA = Food and Drug Administration; HCP = health care professional.

database is dedicated to a different geographical area, there may be some duplication or overlap in the reports that they contain.[4] This is particularly true for serious or severe ADEs, which are often reported to national or regional authorities, such as FAERs or EudraVigilance, respectively, as well as to the global repository, VigiBase.[4]

Post-marketing drug safety surveillance has traditionally been conducted by the systematic review of ADEs that have been reported to SRS databases.[4] This practice has been effective in identifying many types of drug–ADE associations.[1] However, the approval of numerous new drugs and the continually increasing number of reports submitted to these databases has made this approach difficult and impractical.[4] Other drawbacks inherent in this method include the significant delay that may occur before a drug–ADE association is detected, and the fact that many ADEs are underreported in SRS databases.[1] These limitations have inspired the development of other supplemental and complementary approaches to drug safety surveillance using data mining and other data sources.[1]

Electronic Health Records

EHRs contain data that allow real-time, real-world surveillance of drug safety.[1,3,10] These records include information gathered during routine clinical care, including patients' symptoms, medications, health outcomes, results of diagnostic tests, physical exam findings, and hospitalizations.[4] Because EHR databases contain real-time, real-world clinical data, they can potentially provide a more proactive approach to pharmacovigilance.[1,4]

EHR databases include extensive data concerning large populations of patients, so they can be used to complement existing pharmacovigilance approaches, which are mostly based on SRS ADE reports.[1,4] The time-stamped, population-based health care data in EHRs can be used to validate signals identified through mining SRS databases.[4] Because EHR databases contain records concerning a large number of patients, they also provide the opportunity to identify the magnitude of a drug safety problem.[1,4] Using EHR data, a cohort of patients taking the same drug for the same disease can be identified and analyzed to determine the extent of a suspected association between a drug and a particular ADE or other health outcome.[11] Analysis of EHR databases is also efficient with respect to the manpower, financial costs, and time required to complete a study.[4] EHRs are also considered to be a high-quality data source because they are usually created and maintained by HCPs.[11]

Because EHR data is routinely collected and thus longitudinal, ADEs that have a long delay between exposure and clinical effect (e.g., cancer or cardiac valvulopathy) can be identified, especially in databases that have a low patient turnover and long follow-up.[4] Because the data in EHR reports is real-time and real-world, it evolves with new ways that medications are being used.[4] Therefore, mining EHR data may detect new risks due to off-label uses, changes in indication, or the way older drugs are used.[4] EHR databases are also useful for detecting ADEs that have high background incidence rates (e.g., acute myocardial infarction) or unpredictable drug ADEs that are underreported in SRS databases because they are not suspected to be drug induced.[4] Data in EHRs about patient demographics, medications, and health services use may also permit the development of a risk–benefit profile for drugs, providing a broader perspective for regulatory evaluation and decision-making.[4]

There are studies showing that the unstructured narrative data in EHRs can yield better results for some pharmacovigilance procedures than the mostly coded, structured data in SRS reports.[3,11] One study examined unstructured data to identify DDI signals from more than 50 million clinical notes in a database of EHRs.[14] Disproportionality ratios were used to identify DDIs associated with 1,165 drugs and 14 ADEs.[14] The results were validated by a complementary study that analyzed the SRS reports in FAERS, which identified DDIs using an MGPS algorithm.[14] This study first mined EHRs for DDIs and then validated the findings with results from FAERS—this was unusual because normally the opposite is done.[14] The results of this study demonstrated that data mining a combination of structured and unstructured data for DDIs may provide signals with higher statistical confidence levels.[14]

Despite the advantages that EHR databases provide, data mining this source for drug safety surveillance signals is currently done nearly exclusively to complement, not replace, the analysis of SRS databases.[3,4] In fact, EHR databases are now most often used for validation of drug–ADE signals that have been initially detected in SRS databases.[4]

Medical Literature

The medical literature is another major source of data that is expected to improve the detection of drug–ADE associations.[1,3] Information published in clinical studies, observational studies, case reports, and other articles can be analyzed to identify these signals.[3] Both regulatory agencies and product manufacturers routinely consult and track the medical literature to identify undetected drug–ADE associations.[1] Because the content of medical literature is peer reviewed, it is considered to be a highly reliable source of ADE information.[1] Data mining published medical literature can also provide evidence for mechanisms of action for possible DDIs.[3]

Despite its promise, the medical literature is not being used to its full potential, so new computational approaches are being developed to analyze these data more effectively.[1] The FDA has partnered with the NLM to develop literature-based data-mining approaches for drug–ADE detection, based on the identification of disproportionate reporting of drug–ADE associations in MEDLINE.[2] MEDLINE includes more than 20 million abstracts and citations from medical journal articles.[2] Linguistics and cognitive science experts at the NLM have created medical subject headings (MeSH) used for indexing citations in MEDLINE, based on ADE terms in the *Medical Dictionary for Regulatory Activities*.[2] MeSH headings have also been created for drug names based on the Anatomical Therapeutic Chemical Classification System and RxNorm.[2] Additional dictionaries of terms can be selected and added by drug safety researchers for the purpose of indexing MEDLINE.[2]

The FDA has also applied text mining to study drug safety based on chemical structure information in the medical literature.[2] The ability to predict the clinical safety of a drug based on chemical structure is becoming increasingly important, especially when safety data are insufficient or inconclusive.[11] The software used to do this, Linguamatics I2E (Marlborough, Massachusetts), uses NLP to interpret unstructured text when

performing custom searches of information in the medical literature.[2] With respect to mining the medical literature for vaccine–ADE associations, the FDA partnered with Georgetown University to develop the Georgetown Vaccine Information and Safety Resources tool.[2] This tool examines the medical literature for vaccine–ADE associations and molecular information concerning individual vaccines.[2]

Social Media and Other Internet Resources

Social media has enabled the previously unprecedented public sharing of health information, including heath problems and outcomes, and patients' experiences concerning medications.[1] Patients and caregivers consult the Internet and participate in online interactions to obtain medical or drug information to supplement the guidance provided by their physicians or pharmacists.[11] As a result, social media, including social networks, chat rooms, health blogs, and patient community websites, provide a more patient-centered model of ADE reporting than SRS or EHR databases.[1,11] In addition, every post on Facebook, Twitter, Snapchat, Instagram, or YouTube accumulates with great diversity, volume, and speed, providing a large volume of data from millions of users that may reveal previously undetected ADEs when analyzed by data mining.[3,11]

Data mining the unstructured data in these platforms for combinations of terms commonly used to describe an ADE could reveal discussions that are relevant to drug safety; this may be confirmed by reviewing the online conversations that were detected.[11] Like EHRs, social media posts are often in real time, providing the opportunity for earlier detection of drug safety issues than may occur by analyzing SRS databases.[1,11] Sharing one's experience with a medication often elicits discussion of similar experience by other users within a community, triggering a spike in reports that may serve as "an early warning system" for ADEs.[11] Social media also provides an opportunity to monitor ADEs that don't need to be treated immediately or aren't required to be reported to an SRS.[1] It can also assist in investigating ADEs for drugs that treat conditions that have a social stigma, which are underreported in other sources.[3] Social media is also useful in characterizing poorly explored DDIs, such as those between drugs and dietary supplements.[3]

Social media platforms are viewed as offering great potential for monitoring drug effects on public health, so their utility for pharmacovigilance purposes is being investigated.[3] Studies have successfully detected mentions of drugs or drug combinations on Twitter or Instagram while exploring the potential use of social media for identifying information about ADEs and DDIs.[15–17] One study analyzed 5,329,720 posts on Instagram that had been left by 6,927 users between 2010 and 2015, which focused on symptoms associated with antidepressants.[17] Four dictionaries, including drug, pharmacology, and ADE terminologies, were used for text mining the data.[17] Co-mentions of drugs and potential ADEs were identified for daily, weekly, and monthly periods.[17] Associations between the terminologies and the probability that they were mentioned during the same period as the ADE were determined using proximity graphs.[17] This study demonstrated the potential of using data mining to identify associations between drug mentions and terms relevant to ADEs on Instagram.[17]

Other sources of Internet data, such as patient community websites and search engine logs, are also being investigated to determine their value when mined for pharmacovigilance purposes.[1,11] Studies have demonstrated that health care websites provide data that can be used to reliably detect drug–ADE associations and DDIs.[18–21] Websites that host large patient communities, such as MedHelp, DailyStrength, or PatientsLikeMe, may be helpful when data mining for drug–ADE signals and other health care outcomes.[3,11] Millions of people discuss medical problems on these websites, providing laboratory data and other health information.[11] PatientsLikeMe also provides an option for patients to evaluate their medications, including an opportunity to comment on ADEs.[11]

Search engine logs can also be analyzed to detect potential drug safety signals by performing text mining and a frequency analysis of search queries that concern drug–ADE associations.[1,3] For example, one study that analyzed the search logs of 80 million Internet users was able to detect a disproportionate number of searches regarding a specific drug in comparison with a previous period.[22] This study demonstrated that data-mining search logs may detect increased public activity in searching for information about a drug that may presumably occur in response to a potential ADE.[22]

Although studies that have data mined social media platforms, patient communities, and search logs for drug–ADE associations have been conducted, use of these sources for drug safety surveillance is limited.[3] More studies are needed to understand the potential role of social media and other Internet data sources in pharmacovigilance.[3] To this end, the FDA has collaborated on the development of an exploratory data-mining tool, called MedWatcher Social, which monitors several social media platforms for potential drug–ADE associations.[3] The FDA has also stated its interest in technologies that search the Web to identify drug–ADE signals because these alternative sources could provide additional information.[3,11]

APPLICATIONS OF BIG DATA IN PHARMACOVIGILANCE

FDA Drug Safety Surveillance

Regulatory agencies, the pharmaceutical industry, and drug safety researchers have studied the use of big data for pharmacovigilance purposes since the 1990s.[10] To date, a big data approach to drug safety surveillance has proven to be cost-effective, fast, and capable of identifying unsuspected statistical relationships regarding drug–ADE associations.[1,10]

The FDA has been expanding its use of data mining to analyze the increasing number of reports the agency receives, speed the identification and prioritization of potential safety issues, and free personnel to perform tasks that can't be automated.[10] The FDA has stated many advantages regarding the use of data mining for drug safety surveillance.[2] Historically, analyses have been conducted manually, causing concern that the selection, quality, and interpretation of data was subjective.[2] Because data mining is automated, data selection and interpretation is more standardized and objective.[2] Data mining also saves time by permitting the simultaneous analysis of drug–ADE associations across an entire database at once.[2] Statistical scores and prioritization of the drug–ADE signals identified are computed within minutes—much faster than manually

requesting these computer calculations.[2] Studying DDIs is also facilitated through the automatic detection of unusual reporting patterns for patients using multiple products.[2]

In addition, data-mining methods can help provide an understanding of the biological basis for signals by incorporating reference databases regarding drug chemistry and physiology.[2] Curated knowledge databases can link ADEs with the chemical properties of drugs and their effects on physiological pathways and organ systems, helping to identify the mechanisms by which ADEs develop.[1] Such systematic approaches can be used to investigate the biochemistry and pharmacogenetics of drugs and how drug–receptor interactions lead to ADEs and DDIs.[1] This may provide a more advanced approach and potentially deeper understanding than traditional drug safety surveillance systems that are based on a drug's basic targets and chemical structure.[1] In addition, such systematic approaches to drug safety surveillance can be predictive, offering the potential to identify potential ADEs before they are observed.[1]

Industry Drug Safety Surveillance

Because the FDA has had some success using data mining for drug safety surveillance, it has recommended the use of this strategy to the pharmaceutical industry.[2] Mining post-marketing safety data in observational sources (SRS reports, EHRs, and insurance claims) may assist the pharmaceutical industry in detecting drug safety signals earlier, conducting risk assessments, and gathering real-world data for the interpretation of clinical results.[30] By providing a clearer view of a drug's safety profile based on real-world evidence, data mining also supports more informed marketing decisions, as well as increased time- and cost-saving efficiencies.[30]

Due to an increasing number of drug safety issues and withdrawals, the pharmaceutical industry is facing increased demands for greater accountability from the FDA and European Medicines Agency.[30] Regulators in the U.S. and Europe now require pharmaceutical companies to conduct proactive drug risk management programs, based upon comprehensive risk evaluation and mitigation strategies in the U.S. or risk management plans in Europe; these replace voluntary drug safety management efforts by the pharmaceutical industry.[30,32,33] This mandate is designed to improve the anticipation and minimization of known or potential drug risks, as well as to augment information about patient populations that were not studied during clinical trials.[33] The drug risk management plans are legally binding commitments, containing specific requirements, timelines, and penalties for noncompliance.[32]

Because pharmaceutical companies must include strategies for safety signal detection and analysis throughout a drug's lifecycle in their risk management plans, data mining may help these firms meet these regulatory obligations.[28,30] Randomized clinical trials are the gold standard for determining drug efficacy; however, they are not fully competent to predict drug safety post-marketing, when a drug is used in the real world.[34] The use of data mining to proactively collect real-world safety data from SRS reports, EHRs, and other electronic sources may significantly contribute to the execution of a pharmaceutical company's risk management plan.[34]

Data mining also provides advantages that may support a new "innovation-promoting" model for drug regulation. The 21st Century Cures Act (similar to the "mandated pathways approach" in Europe) is an expansive piece of health legislation that was passed on December 7, 2016, by the 114th U.S. Congress.[31] The act places great emphasis on reducing the time and money needed for drug development.[31] This new regulatory paradigm allows smaller, shorter, fewer, and more flexible randomized clinical trials to be conducted prior to the approval of a new drug or indication.[10,31] Additional clinical studies, risk management plans, and more intensive drug safety surveillance using real-world data will then be required after drug approval.[10] Such measures may be fulfilled through mining observational data, statistical modeling instead of post-approval clinical studies, and/or surrogate endpoints.[31] The act also allows companies to submit "data summaries" rather than full clinical trial results when applying for new indications for already-approved drugs; the data summaries may include evidence derived from observational data, insurance claims, and/or patient experience data.[22,39] Although this legislation creates new concerns, it may partially alleviate pharmaceutical industry complaints regarding clinical trials, such as patient recruitment challenges, burdensome and obtrusive data collection requirements, and the fact that clinical study results don't necessarily apply to the overall population.[29]

The use of EHRs to conduct post-approval clinical research is increasingly being discussed as a measure to counterbalance these drawbacks of clinical trials.[29] It is envisioned that EHRs may be used as the primary data source for post-marketing observational and comparative effectiveness studies.[29] Data mining can examine EHR data from tens of millions of patients to identify a potential drug–ADE association, whereas a clinical trial typically enrolls only several hundred people.[10] Data-mining studies also are less costly and produce results within months, whereas clinical trials are expensive and results may take years.[10,27] The successful use of EHRs for post-marketing clinical research may therefore allow the pharmaceutical industry to use fewer resources to conduct more studies in less time, potentially yielding more consistent results based on real-world data.[27]

Despite enthusiasm regarding the new regulatory model in the 21st Century Cures Act, research published by numerous investigators has identified many obstacles that may prevent obtaining results that are sufficiently robust, reliable, and reproducible if observational data were a primary source for regulatory decisions.[10] To fulfill the vision of EHRs being used to justify fewer and shorter randomized clinical trials, the quality and validity standards for such data-mining studies will need to be very high.[4,10] Mining observational data may be useful for identifying interesting new perspectives regarding drug safety risks or to verify risk assessments based on other data sources.[10] However, it may be overreach to have as much confidence in mining observational data for drug–ADE associations as in the results obtained in well-conducted, randomized clinical trials.[10]

Relying on mining EHRs or other observational data to detect DDIs for drugs in development or soon after approval may also be challenging.[3] It is not feasible to clinically evaluate all possible drug combinations during the development of a drug.[3] Therefore, because many potential DDIs are unknown, using data mining to detect potential DDIs for drugs in development

may be useful.[3] However, to accomplish this goal, adequate information about a new or experimental drug must already be present in the medical literature and other data sources.[3] The use of data mining to assist in regulatory decision-making for drugs in development or early post-marketing is therefore limited when studying the safety of new agents for which little data is available.[3]

Before data-mining results are routinely involved in regulatory decision-making, the investment of substantial financial resources will be needed to study and improve the reliability of EHR data for drug risk assessments.[10] Although a single medium-sized clinical trial can cost tens of millions of dollars, it may be naïve to think that similar high-quality clinical safety data may be obtained from millions of EHRs or other observational data sources at minimal cost.[10] Regulatory decision-making is also complex and might not permit delaying decisions until information from ancillary sources is available, especially if a definitive clinical trial is the best source for the needed data.[4] Although it has potential, big data may not currently be capable of providing the necessary assurances to counterbalance new drug approval policies that reduce the number, size, duration, and rigor of randomized clinical trials.[10]

Regulatory Decision-Making

There are many examples of the important role that data mining has played in identifying drug–ADE signals that have led to regulatory action by the FDA and other regulatory agencies.[2] Examples of signals that have been identified or strengthened through mining SRS reports include the association of temafloxacin with hemolytic anemia; terfenadine and cisapride with ventricular arrhythmias; and fenfluramine with cardiac valvulopathy.[4] Data mining has even identified new ADE signals for older drugs, such as the association between propylthiouriacil and hepatotoxicity.[2] Data mining has also been useful in providing information that is helpful in further characterizing ADRs.[4] For example, the predominant cholestatic pattern and delayed time to onset and recovery found in flucloxacillin-induced hepatitis were detected by data mining SRS reports.[4]

With regard to vaccines, the first safety signal detected by the FDA involved the use of an MGPS mining method to analyze Fluzone (Sanofi Pasteur) vaccination data for the 2010–2011 flu season.[2] This analysis detected an association between febrile seizures and the administration of Fluzone to young children.[2] In another example, higher than expected reports of intussusception following administration of the RotaShield rotavirus vaccine were detected in 1999.[4] Following verification of this increased risk in epidemiologic studies, the manufacturer voluntarily removed this vaccine from the market.[4] A vaccine–ADE association between a meningococcal conjugate vaccine and risk of the development of Guillain–Barré syndrome was also first observed by data mining VAERS.[4]

Once an ADR is confirmed and the dangers it poses are verified and understood, the FDA will consider actions that may include adding a warning to the drug label or withdrawing the medication from the market.[3] Consideration of the risk–benefit balance is important in determining what action is taken, particularly with regard to withdrawal of a drug from the market.[4] How this will affect patients and whether an alternative treatment is available are factors involved in this decision.[4]

CHALLENGES IN APPLYING BIG DATA TO PHARMACOVIGILANCE

Data mining provides a unique opportunity for drug safety surveillance; however, it also presents challenges, including specific problems inherent in each data source.[1] Analytical challenges involved in data mining may include a lack of established standards and validation methods, confounding variables, false signals, data inconsistencies, bias, or too much or too little data.[1] Many of these analytical challenges arise because the data used have been repurposed from their original function.[1,4] Other challenges involving technical or cost issues may also be a factor, as well as issues regarding the ethics, privacy, and security of patient data.[1] An in-depth understanding of these challenges and limitations will allow regulators, the pharmaceutical industry, and researchers to be better equipped to make informed decisions.[2]

Problems with standardization include the absence of established statistical methods, variable or limited terminology, and few validation studies.[10] Development of standardized vocabularies is critical for the application of data mining to drug safety surveillance.[3] Lack of standardization is a particular problem for the FDA because the agency primarily uses disproportionality and CPA data-mining tools.[2] These tools work best when analyzing databases with standard terms for products, events, and demographic information.[2]

Inconsistency in the quality of all health data sources, including SRS reports, the medical literature, EHRs, and social media, presents another challenge.[2] The quality of a data source is largely dependent on the way the data have been structured and the expertise of the person entering or providing the data.[2] For example, the MeSH indexes in MEDLINE are presumably of higher quality than social media because they are structured and created by linguistic and cognitive science experts.[2] Currently, most data entry for all sources is done manually, which contributes to inconsistent quality; however, the use of text-mining tools to standardize the content of coded fields is being investigated.[2]

Challenges of Specific Data Sources

SRS Databases

SRS report databases, such as FAERS, provide a valuable source of data for drug safety surveillance efforts.[3] However, there are challenges inherent in SRS databases that limit their utility, such as sample variance and the underreporting of ADEs.[3] Underreporting may be due to lack of recognition of a potential drug–ADE association, not being informed regarding the reporting requirements or process, or fear of litigation.[2] In addition, except for drug companies, SRS reporting is voluntary, so FAERs may not include all ADEs.[3] Studies have shown that as many as 90% of serious ADEs go unreported.[38] Underreporting of ADEs can cause a delay between the introduction of a drug to the market, discovery of an ADE, and subsequent corrective regulatory action.[3,4] Low reporting also compromises the ability to detect ADEs and artificially increases risk estimates for drugs, causing false positives.[3] The reports submitted to SRS databases are also subjective because they represent the reporter's concern that there may be a drug–ADE relationship, not necessarily an actual ADE.[2] The quality of the data in SRS databanks may also vary by country or region and

the knowledge and skill level of the individual entering or submitting the report.[4]

Other challenging phenomena that may occur with SRS databanks include missing, incorrect, or vague information, and duplicate reporting.[2,3,11] Detailed information about each ADE makes it easier to interpret SRS reports, especially when diverse or severe symptoms are present due to underlying disease.[11] However, SRS case reports may not be complete or may contain inconsistent information with respect to medical history or comorbidities.[4] Because the baseline number regarding exposure to the drug in the overall population is usually not available or included in SRS databases, the true incidence of an ADE cannot be determined.[4] SRS reports where the patient was on multiple drugs also increase the chance that spurious associations between a particular drug and an ADE may be generated.[3] Data mining for DDI signals or multiple ADEs can also be computationally expensive because it involves the analysis of a large amount of data for multiple drug combinations.[3]

Challenges are also involved in interpreting the information in SRS databanks.[2] Database-specific knowledge is often required to understand the data-mining methodologies used, as well as the signals generated when these databases are analyzed.[2] Knowledge required may include coding dictionaries, data entry and coding processes, and reporting requirements; database structure architecture; identification of spam or malicious reports; and an understanding that signals generated are specific to the database.[2]

EHRs and Other Observational Data

EHR databases may provide a wealth of information regarding medication use; however, there are caveats with respect to the interpretation of the signals generated when mining this data source.[4] Because both are "observational data," EHRs share many of the limitations of SRS reports.[1,3] For example, all observational data are associated with confounding control and bias issues.[1] Like SRS reports, EHRs pose shortcomings regarding reliability, reproducibility, and statistical standards.[10] This means that further validity studies, an understanding of the ADEs that are likely to be captured poorly, and a means to improve the consistency and accuracy of this data source are needed.[10]

As with SRS reports, the analysis of EHRs is also often complicated by the presence of unstructured narrative information and/or complex, missing, or inaccurate data.[3] Concomitant medications or comorbidities, as well as a lack of standard terminology, may interfere with or confound data-mining signals when analyzing EHRs.[3] The standardization of vocabularies concerning diseases, drugs, and processes entered into EHRs could assist in improving signal detection and validation.[3] A high degree of variability in the data for individual clinical parameters in EHRs also leaves results subject to substantial bias and difficult to replicate, due to method selection or the data itself.[3,10]

When analyzing EHRs, it is unclear what methodology is most appropriate for application to a particular database to answer a specific question.[1] It is also difficult to determine how to interpret the inconsistent findings that may arise from the use of different methodologies and databases.[1] There may be too few cases in an EHR database to analyze a particular drug–ADE association or DDI because, in some cases, a large number of patients is needed to generate robust signals.[3] Too small a sample size may also be a problem when data mining for ADEs associated with orphan drugs.[11] It is possible that the limitations caused by small sample size could be overcome by networking databases of EHRs together to create a larger data source.[11] However, the capability for signal detection may still be low for drugs that are infrequently used and for very rare outcomes, even with large multicountry databases.[4]

Another challenge when data mining EHRs is the lack of access to patient medical records by drug safety surveillance experts due to patient privacy measures.[3] As a result, data concerning only small populations of patients may cause false-positive or false-negative signals due to small sample size.[3,10] For this reason, caution should be used in interpreting studies that fail to detect a drug effect because important drug harms may be masked due to a too-small population.[10] Standardizing and anonymizing patient information could expand access to larger volumes of EHRs, which will broaden the methods that can be applied—improving the accuracy, quality, and reproducibility of mining this data source.[3]

Health insurance claims are another source of observational data that may be used in data mining for drug safety surveillance.[10] However, the coding and terminology used in this source may be ill-suited for this purpose.[10] Health insurance claims may provide an insurance claim auditor's interpretation of health care data because the coding of health outcomes in these records may be biased by reimbursement policies.[4] The information in Social Security or health maintenance organization databases could also be influenced by lack of administrator or payer incentive to record accurate data.[4] Illustratively, one study showed differences in the diagnosis-related group coding and classification assigned by the physician compared with that recorded by hospital administration staff.[4] The data provided in real-world EHR or health insurance claims may also be influenced by evolving clinical practices, such as changes in disease management guidelines or shifts in preferential prescribing.[4]

Medical Literature

The vast and continuously growing amount of data in the medical literature also presents challenges for drug safety surveillance data mining.[3] Similar to other sources, the unstructured nature of the data in the medical literature makes the detection of drug–ADE associations challenging.[3] Because of its large volume, data-mining studies using this data may also be extremely time consuming.[3]

Data-mining methods used to extract DDIs from the medical literature have performed well in different extraction challenges.[3] However, although co-occurrence-based methods have been shown to yield good recall, many false positives still occur, causing low precision.[3] Accuracy may be improved by applying rule-based methods, but complex sentences present limitations.[3] Machine learning methods generally produce the best performance; however, training these algorithms with big data sets and well-annotated content is important to their success.[3]

Computational methods for the extraction and analysis of information in the medical literature are being improved

continuously.[3] NLP algorithms and other tools are being developed that automatically extract information from this source.[3] Use of these tools is expected to save time and resources and improve accuracy, making them essential to compiling a more complete compendium of drug knowledge concerning ADEs and DDIs.[3] The NLM is also continuously improving search capabilities for MEDLINE to make it more user friendly and search results more accurate.[3]

Social Media and Other Internet Sources

Despite its potential, data mining social media and other Internet sources for drug–ADE signals is most controversial. Along with issues concerning the feasibility of mining the enormous volume of information in these sources, there are concerns regarding the quality and reliability of the data and ethical issues.[1] The statistical challenges that have been identified regarding mining social media for drug–ADE signals include lack of specificity, verification difficulties, low validity, and bias.[11]

Regarding lack of specificity, a statement made by someone who had taken a medication that he or she "is not feeling well" could refer to a potential ADE or to an underlying medical condition.[11] Following up on signals from social media posts, to verify that every potential symptom identified was, in fact, an ADE, may be considered intrusive or too difficult because of privacy issues, lack of contact information, or the time required.[11] Moreover, it may be impossible to determine if a social media report is from a geographic region that is relevant to a particular regulatory agency's pharmacovigilance efforts.[11] With respect to validity, patients and caregivers are generally poorly qualified to diagnose a medical condition; therefore, they are more likely to make posts on social media that are inaccurate, misinformed, or incomplete.[11] A suspected medical diagnosis by someone who is not an HCP does not provide proof of an ADE, but may be the product of hearsay or media influence.[1,11]

Additional technical challenges regarding the use of social media to identify drug–ADE signals concern difficulties in automatically extracting ADE information from unstructured text.[1] Data mining for terms used in social media text searches may infrequently (less than 2%) identify drug–ADE associations, occurring at a rate five to 10 times lower than that detected by searching a structured data source.[37] Information from social media that likely does not use medical terminology and has not been screened by an HCP also hinders signal detection by adding background noise.[11] Data mining social media may also identify a duplicate patient-reported record of a suspected ADE that has already been reported through conventional pharmacovigilance channels.[11]

In addition, social media introduces an inherent bias toward patients younger in age, as well as cultural and socioeconomic groups who have access to electronic devices and the technical ability to use them.[11] Mining social media data may thus exclude older or sicker patients, who experience a greater risk for ADEs associated with their medications.[11] Even if all the biases and obstacles regarding data mining social media are overcome, the low-quality data contained within this source are only likely to provide value for signal detection for drugs that are broadly used.[11] Despite these challenges, attempts are ongoing to data mine social media to identify suspected ADEs within patient descriptions of drugs and events.[11]

A major challenge in mining Internet search logs for drug safety signals is the assumption that a search for drug information was made because the patient experienced that event.[11] The assumption that all searches involving terms for a drug and ADE serve as a proxy for drug exposure will confound data-mining results.[11] The inability to ascribe motivation to searches for medical information was notoriously evident when the Web service "Google Flu Trends" signaled a flu outbreak based on a surge of flu-related Internet searches.[39] This alert was false because most of the searches were made by people who were researching information reported in the media, rather than those who were ill.[11] This incident provides a real-world example of how Internet searches may be more accurately considered to be a medical information query rather than a report of an actual event.[11]

FUTURE DIRECTIONS

Despite the numerous challenges inherent in the current methods and data sources available, the promise and opportunity for big data to contribute to improvements in pharmacovigilance cannot be overlooked.[1,10] Big data can potentially identify correlations from information patterns about patients, drugs, ADEs, and risk factors within diverse datasets.[11] This is the purpose of pharmacovigilance, so the appeal of utilizing big data for drug safety surveillance is evident.[11] A description of steps being pursued to overcome the current limitations inherent in data mining for pharmacovigilance purposes follows.

Improve Consistency Among Data Sources

The results of drug safety data-mining studies need to be consistent, and results from all databases need to be reproducible.[10] Analytic challenges are expected to multiply with the availability of new data sources and new methods for submitting SRS reports, such as mobile and Web-based applications.[2] Therefore, it will be helpful for the FDA to design a standardized information technology system that enables data to be submitted, retrieved, processed, and evaluated in a consistent manner.[2] Improvements in the consistency of medical terminology, the annotation of content, and a unified criteria for a gold standard regarding methods and verification are also crucial steps for improvement in the detection of ADEs and DDIs.[3] Greater collaboration in data sharing between disciplines is also needed to encourage consistent results.[1]

Validate Utility of Data Sources

The FDA needs to further validate relatively new sources of electronic health care data, such as EHRs, insurance claims, social media, and other Internet sources.[1,2] Several large studies have been conducted to validate the use of EHRs, but the other new data sources are not yet in systematic use; therefore, whether they will ever become routine validated sources in drug safety surveillance data-mining studies is uncertain.[1] Caution is also warranted when using such diverse data sets that have different limitations and biases, even though doing so may yield more ADE signals.[3] Determining which data sources should be used to generate hypotheses and which meet a higher standard, allowing their use in confirming drug–ADE

signals, is also necessary.[1] More transparent audit trails would also be helpful in validating each analyst's selection criteria, results, and interpretations.[2]

Establish Standards for Signal Detection

Standards for properly evaluating drug–ADE associations are critical and should be implicit when undertaking any signal detection activity.[4] Guidelines need to be established to determine when a signal is substantial enough to require follow-up and verification.[4] Adjustment of signal thresholds to reflect ADE severity and the condition for which the product was prescribed may also be necessary.[2] Preferred statistical methods also must be further analyzed to refine and broaden what is known about them.[10] Standards should also be established for the interpretation of studies that don't detect any risks to prevent the potential harm that may occur from false-negative data-mining results for drug–ADE associations.[10]

Apply an Integrative Approach to Signal Detection

Drug safety surveillance efforts should integrate evidence from all possible sources in order to protect public health interests.[4] Exploring multiple sources in parallel may provide more benefit than exploring single sources sequentially.[1] Drug safety surveillance data-mining efforts may be improved by establishing a complete reference database that includes event, product, toxicology, physiology characteristics, and visual analytics, coupled to context information across multiple data resources.[2]

Integrative approaches that combine data from different resources, such as the scientific and medical literature, SRS reports, social media, and even medical images, could assist in developing robust models with improved accuracy in predicting ADEs.[1,3] All sources for data mining drug–ADE associations present some challenges, but combined, they may enable the study of drug effects from a variety of perspectives.[3]

Improve Data Mining Software and Tools

Data-mining tools need to be improved so that they: are quicker and/or require less data; process and retrieve results in real time; are scalable to accommodate growing databases; and use more advanced NLP and text-mining algorithms to more accurately extract signals from unstructured data.[2] Development of objective benchmarks to compare and quantify the performance of drug–ADE detection tools across data systems is also imperative.[1] Technological advances are also needed to overcome the biases inherent in data-mining algorithms that have been designed and coded by humans.[11] There is also a need to develop a technological platform and retrieval system to extract information from unexplored sources, such as medical images.[3]

Apply Data Mining to Other Product Safety and Regulatory Issues

The FDA has been fairly satisfied with the success of using data-mining methods to analyze SRS report data, so agency experts are developing even more sophisticated methods for use in analyzing additional internal and external databases.[2] In addition to its use for drug safety surveillance, data mining may also be applied to assist the FDA in reviewing dose–response relationships, chemosensory effects, efficacy evaluations, marketing strategies, and advertising perceptions, as well as drug initiation, cessation, and switching rates.[2] The FDA is also using data mining in field work to explore trends in safety, inspection, and recall data, so that agency resources and personnel can be allocated accordingly.[2]

CONCLUSION

Data mining has been successful in identifying new drug–ADE associations for drug safety surveillance purposes.[11] Although numerous challenges remain, the promise and opportunity for big data to make further contributions to pharmacovigilance efforts are evident.[1,10] Improved methods, tools, and data sources used in drug safety surveillance are still in the early stages of development and are likely to further advance the use of big data for pharmacovigilance in the future.[1–4,10,11] Ultimately, how well big data improve the detection of drug safety issues will be the true measure of its value.

REFERENCES

1. Harpaz R, Dumochel W, Shah NH. Big data and adverse drug reaction detection. *Clin Pharmacol Ther* 2016;99(3):268–270.
2. Duggirala HJ, Tonning JM, Smith E, et al. Use of data mining at the Food and Drug Administration. *J Am Med Inform Assoc* 2016;23(2):428–434.
3. Vilar S, Friedman C, Hripcsak G, et al. Detection of drug–drug interactions through data mining studies using clinical sources, scientific literature, and social media [published online February 17, 2017]. *Brief Bioinform* doi: 10.1093/bib/bbx010.
4. Coloma PM, Trifiro G, Patadia V, Sturkenboom M. Postmarketing safety surveillance: Where does signal detection using electronic health care records fit into the big picture? *Drug Saf* 2013;36:183–197.
5. Department of Health and Human Services. National Action Plan for Adverse Drug Event Prevention. 2014. Available at: https://health.gov/hcq/pdfs/ade-action-plan-508c.pdf. Accessed September 24, 2017.
6. McGettigan P, Henry D. Cardiovascular risk with nonsteroidal anti-inflammatory drugs: systematic review of population-based controlled observational studies. *PLoS Med* 2011;8(9):e1001098.
7. Yang H, Yang CC. Drug–drug interactions detection from online heterogeneous health care networks. Presentation at 2014 IEEE International Conference on Healthcare Informatics, Verona, Italy, September 15–17, 2014. Abstract available at: https://ieeexplore.ieee.org/document/7052464. Accessed September 24, 2017.
8. Agency for Healthcare Research and Quality. Medication-related adverse outcomes in U.S. hospitals and emergency departments, 2008. Available at: www.hcup-us.ahrq.gov/reports/statbriefs/sb109.pdf. Accessed September 24, 2017.
9. Boston University Slone Epidemiology Center. Patterns of medication use in the United States: a report from the Slone Survey 2006. Available at: www.bu.edu/slone/files/2012/11/SloneSurveyReport2006.pdf. Accessed September 24, 2017.
10. Moore TJ, Furberg CD. Electronic health data for post-market surveillance: a vision not realized. *Drug Saf* 2015;38(7):601–610.
11. Price J. What can big data offer the pharmacovigilance of orphan drugs? *Clin Ther* 2016;38(12):2433–2445.
12. Council for International Organizations of Medical Sciences. About. 2017. Available at: https://cioms.ch/about. Accessed September 22, 2017.
13. Colman E, Szarfman A, Wyeth J, et al. An evaluation of a data mining signal for amyotrophic lateral sclerosis and statins detected in FDA's spontaneous adverse event reporting system. *Pharmacoepidemiol Drug Saf* 2008;17(11):1068–1076.
14. Iyer SV, Harpaz R, LePendu P, et al. Mining clinical text for signals of adverse drug–drug interactions. *J Am Med Inform Assoc* 2014;21:353–362.
15. Carbonell P, Mayer MA, Bravo A. Exploring brand-name drug mentions on Twitter for pharmacovigilance. *Stud Health Technol Inform* 2015;210:55–59.

16. Hamed AA, Wu X, Erickson R, et al. Twitter K–H networks in action: advancing biomedical literature for drug search. *J Biomed Inform* 2015;56:157–168.
17. Correia RB, Li L, Rocha LM. Monitoring potential drug interactions and reactions via network analysis of Instagram user timeliness. *Pac Symp Biocomput* 2016;21:492–503.
18. Yang CC, Jiang L, Yang H, et al. Detecting signals of adverse drug reactions from health consumer contributed content in social media. Presentation at 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Beijing, China, August 12–16, 2012. Available at: https://tinyurl.com/y7zl7yjq. Accessed September 22, 2017.
19. Yang CC, Yang H, Jiang L, et al. Social media mining for drug safety signal detection. Presentation at 2012 International Workshop on Smart Health and Wellbeing, Maui, Hawaii, October 29, 2012. Abstract available at: https://dl.acm.org/citation.cfm?doid=2389707.2389714. Accessed September 22, 2017.
20. Yang H, Yang CC. Harnessing social media for drug–drug interactions detection. Presentation at 2013 IEEE International Conference on Healthcare Informatics, Philadelphia, Pennsylvania, September 9–11, 2013. Abstract available at: https://ieeexplore.ieee.org/abstract/document/6680457. Accessed September 24, 2017.
21. Yang H, Yang CC. Mining a weighted heterogeneous network extracted from health care-specific social media or identifying interactions between drugs. 2015 IEEE International Conference on Data Mining Workshops, Atlantic City, New Jersey, November 14–17, 2015. Abstract available at: https://ieeexplore.ieee.org/document/7395671. Accessed September 24, 2017.
22. White RW, Harpaz R, Shah NH, et al. Toward enhanced pharmacovigilance using patient-generated data on the Internet. *Clin Pharmacol Ther* 2014;96:239–246.
23. Food and Drug Administration. FDA issues public health warning on phenylpropanolamine. Available at: www.fda.gov/Drugs/DrugSafety/InformationbyDrugClass/ucm150763.htm. Accessed September 22, 2017.
24. Cantu C, Arauz A, Murillo-Bonilla LM, et al. Stroke associated with sympathomimetics contained in over-the-counter cough and cold drugs. *Stroke* 2003;34(7):1667–1672.
25. Sentinel Coordinating Center. Sentinel Initiative background. Available at: www.sentinelinitiative.org/background. Accessed September 25, 2017.
26. Sentinel Coordinating Center. Mini-Sentinel. Available at: www.sentinelinitiative.org. Accessed September 25, 2017.
27. Observational Health Data Sciences and Informatics. Who we are. 2018. Available at: www.ohdsi.org/who-we-are. Accessed April 13, 2018.
28. Global Health Network. Global Pharmacovigilance. Glossary of drug safety terms. Available at: https://globalpharmacovigilance.tghn.org/resources/glossary. Accessed September 26, 2017.
29. Cowle MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clin Res Cardio* 2017;106:1–9.
30. Lu Z. Information technology in pharmacovigilance: benefits, challenges, and future directions from industry perspectives. *Drug Healthc Patient Saf* 2009;1:35–45.
31. Korda M. 21st Century Cures Act: back to the future? *Springer Nature*. January 24, 2017. Available at: https://nbmecommunity.nature.com/users/21923- michelle-korda/posts/14629-21st-century-cures-act-back-to-the-future. Accessed September 28, 2017.
32. Balian JD, Wherry JC, Malhotra R, Perentesis V. Roadmap to risk evaluation and mitigation strategies. *Ther Adv Drug Saf* 2010;1(1):21–38.
33. Lamarque V, Pietan Y. The pharmaceutical industry and the adverse effects of drugs. *Ann Pharm Fr* 2007;65(5):308–314.
34. Andrews E, Dombeck M. The role of scientific evidence of risks and benefits in determining risk management policies for medications. *Pharmacoepidemiol Drug Saf* 2004;13(9):599–608.
35. Regulatory Affairs Professional Society. 21st Century Cures redux and what it will mean for the FDA. Available at: http://raps.org/Regulatory- Focus/News/2016/11/28/26242/Regulatory-Explainer-21st-Century-Cures-Redux-and-What-it-Will-Mean-for-FDA. Accessed September 28, 2017.
36. Banda JM, Callahan A, Winnenburg R, et al. Feasibility of prioritizing drug–drug event associations found in electronic health records. *Drug Saf* 2016;39:45–57.
37. IMS Health, Inc. Monitoring adverse events in pharma's patient support programs. Available at: www.imsbrogancapabilities.com/pdf/nexxus-social-adverse-effect-tracker.pdf. Accessed September 22, 2017.
38. Hazell L, Shakir SA. Underreporting of adverse drug reactions: a systematic review. *Drug Saf* 2006;29(5):385–396.
39. Butler D. When Google got flu wrong. U.S. outbreak foxes a leading Web-based method for tracking seasonal flu. *Nature* 2013;494:155–156.
40. Food and Drug Administration. Guideline for industry—clinical safety data management: definitions and standards for expedited reporting. March 1995. Available at: www.fda.gov/downloads/drugs/guidances/ucm073087.pdf. Accessed September 24, 2017.
41. Sloane R, Osanlou O, Lewis D, et al. Social media and pharmacovigilance: a review of the opportunities and challenges. *Br J Clin Pharmacol* 2015;80(4):910–920.
42. Simon Fraser University. Big data glossary. Available at: https://tinyurl.com/y6vhvgff. Accessed September 26, 2017.
43. Big Data Made Simple. Big data A to Z: a glossary of big data terminology. July 4, 2014. Available at: http://bigdata-madesimple.com/big-data-a-to-zz-a-glossary-of-big-data-terminology. Accessed September 26, 2017.
44. MedicalBiostatistics.com. Glossary of methodological terms. Available at: www.medicalbiostatistics.com/Glossary.pdf. Accessed September 26, 2017.
45. Uppsala Monitoring Centre. What is VigiBase? Available at: www.who-umc.org/vigibase/vigibase. Accessed September 26, 2017.
46. Food and Drug Administration. FDA Adverse Event Reporting System (FAERS) public dashboard. February 5, 2018. Available at: www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070093.htm. Accessed April 13, 2018.
47. European Medicines Agency. 2016 Annual Report on EudraVigilance for the European Parliament, the Council, and the Commission. March 16, 2017. Available at: www.ema.europa.eu/docs/en_GB/document_library/Report/2017/03/WC500224056.pdf. Accessed January 2, 2018. ■