Artificial Intelligence in the Life Sciences 1 (2021) 100005

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Signal Detection in Pharmacovigilance: A Review of Informatics-driven Approaches for the Discovery of Drug-Drug Interaction Signals in Different Data Sources

Heba Ibrahim [a,b,*], A. Abdo [a], Ahmed M. El Kerdawy [c,d], A. Sharaf Eldin [a,e]

[a] Department of Information Systems, Faculty of Computers and Artificial Intelligence, Helwan University, Egypt
[b] Egyptian Pharmaceutical Vigilance Center, Egyptian Drug Authority (EDA), Egypt
[c] Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Cairo University, Egypt
[d] Department of Pharmaceutical Chemistry, Faculty of Pharmacy, New Giza University, Egypt
[e] Faculty of Information Technology and Computer Science, Sinai University, Egypt

ARTICLE INFO

Keywords:
Machine Learning
Data Mining
Adverse Drug–Drug Interaction
Signal Detection
Pharmacovigilance

ABSTRACT

The objective of this article is to review the application of informatics-driven approaches in the pharmacovigilance field with focus on drug-drug interaction (DDI) safety signal discovery using various data sources. Signal can be a new safety information or new aspect to already known adverse drug reaction which is possibly causally related to a medication/medications that warrants further investigation to accept or refute. Signals can be detected from different data sources such as spontaneous reporting system, scientific literature, biomedical databases and electronic health records. This review is substantiated based on the fact that DDIs are contributing to a public health problem represented in 6-30% adverse drug event occurrences. In this article, we review informatics-driven approaches applied by authors focusing on DDI signal detection using different data sources. The aim of this article is not to laboriously survey all PV literature. As an alternative, we discussed informatics-driven methods used to discover DDI signals and various data sources reinforced with instances of studies from PV literature. The adoption of informatics-driven approaches can complement and optimize the practice of safety signal detection. However, further researches should be carried out to evaluate the efficiency of those approaches and to address the limitations of external validation, implementation and adoption in real clinical environments and by the regulatory bodies.

1. Introduction

1.1. Pharmacovigilance background

Drug safety monitoring identified as Pharmacovigilance (PV) is formally defined by the "World Health Organization (WHO)" as "the science and activities relating to the detection, assessment, understanding and prevention of drug-related problems" [1]. PV may be classified into two categories: (1) preauthorization PV concerning the information adverse drug events accumulated from clinical trial settings (phase 0/I through phase III) [2]; and (2) postmarketing/post-authorization PV concerning the safety information collected during post-authorization life cycle. Even though pre-authorization clinical studies (i.e. RCTs) are deliberated as the lineament of explaining a drug efficacy, they don't usually detect whole safety concerns of a certain drug before its use in the real world due to well-known limitations. Those limitations can be summarized in the restricted sample of trial participants involved in those studies compared to the total targeted population who may experience the drug whenever marketed, the limited interval of drug exposure per each study participant especially if the drug projected for chronic use, lack of possibly risky patient subpopulations who are usually omitted from clinical trials (e.g., patients with impaired organs, elderly patients, children, and childbearing woman who may be pregnant or lactating mother, patients on chemotherapy) [3]. Furthermore, RCTs can't detect rare/very rare adverse drug reactions (ADRs) (with occurrences of 1/10000 and 1/100000, respectively) or latent ADRs (> 6 months) [4]. That said, the drug efficacy information from premarketing RCTs is generally more comprehensive and reliable, whereas complete safety profiles can't be established [4]. These boundaries impose obligations of both marketing authorization holders and regularity authorities to continuously monitor, collect, assess, and draw conclusions and appropriate actions to keep positive benefit-risk ratio as long as the drug exists in market.

* Corresponding author.
E-mail address: pharma_heba@hotmail.com (H. Ibrahim).

https://doi.org/10.1016/j.ailsci.2021.100005
Received 9 March 2021; Received in revised form 26 June 2021; Accepted 27 June 2021
Available online 8 July 2021

1.2. Drug-drug interaction detection

One of the leading causes of global morbidities and treatment failure is interaction between drug agents. About 30% of unexpected adverse drug events may be accounted for Drug-drug interactions (DDIs) [5]. In acknowledgment of adverse drug effects is a chief reason of product attrition in late stages of the drug development cycle, the earlier detection of the interaction profiles of novel drug candidates is still intractable. In conventional drug development process, many of candidates of novel therapeutics are investigated afore triaging a slight quantity of eligible molecules, ordinarily depends on prior scientists' experience in the therapeutic domain and are nominated for subsequent laboratory and preclinical testing. Regular drug discovery process includes the delivery of new candidates from basic medicinal scientists into testing in preclinical phases followed by a conduction of different clinical trials. Unpredictable drug interactions may result in serious health outcomes and consequently undesirable influences on the whole drug discovery progression. Accordingly, an increasing demand for developing informatics-based frameworks to discover safety comprehension in terms of clinical practice and originate DDIs of novel therapeutic nominees earlier in drug development through multidisciplinary linking between clinical and preclinical information. [6].

Additionally, premarketing clinical studies don't usually explore DDIs but rather emphasize on investigating the efficacy and safety of individual medications [7]. Since patients on polypharmacy are not usually included in clinical trials. Furthermore, whenever DDIs are doubted, sampling biases and cohort sizes bound the capability for identifying rare adverse reactions [8]. Interactions between drugs may take place when sharing the same target proteins, pharmacological and metabolic pathways leading to affecting the safety and/or efficacy profile of pharmaceutical products. In other means, the co-administration may considerably affect the efficacy and/or safety profiles of a drug agent. Unpredictable DDIs are recognized only through signal detection practices and postmarketing life cycle [9]. The magnitude of a DDI seriousness warrants different regulatory actions grading from label information changes into market withdrawals [10–12]. Due to the limitations and difficulties in monitoring interaction profiles of drugs, multiple informatics-based researches have been emerged in recent years for DDI discovery via adopting machine learning (ML) and data mining methods from heterogeneous data sources.

DDIs can be classified by various criteria. With regard to severity, interactions are frequently characterized by three categories according to severity into (minor, moderate and major) [13]. Those interactions having minimal clinical consequences and minimal risk are classified as minor DDIs. Moderate DDI is with moderate clinical significance, usually avoid these combinations and they may be used only under special circumstances with closer monitoring and may require dosage changes. Major DDI is with high clinical significance probably leading to serious clinical outcomes should be averted due to negative benefit-risk ratios.

In terms of mechanism, DDIs are classified as either pharmacodynamic (PD) or pharmacokinetic (PK) [14]. It is noteworthy that PD – based DDIs make up a smaller class than PK-based DDIs. A Pharmacodynamic (PD) interaction occurs when the pharmacological effect of a drug is affected by another as a result of: 1) direct effect at target site, (2) interference with biological or physiological signaling pathways, resulting in additive, or antagonistic effects, or synergistic/ indirect pharmacologic effect [15]. Pharmacokinetic (PK) interaction occurs when a drug influences the disposition of another drug in the body by interfering its absorption, distribution, metabolism, or elimination (ADME) properties, causing an altered plasma concentration of the first drug that may lead to detrimental consequences (treatment failure or toxicity). Within the pharmacokinetic processes, the metabolism part covers the largest [16]. A recent study [17] has estimated that the PK-based DDIs occur at metabolic level as follows: (i) 64% of PK-related DDIs involve induction or inhibition of the hepatic cytochromes that are accountable for medications metabolism, (ii) while 20% of PK DDIs result from transporters and 2% from carriers.

1.3. The concept of Safety Signal Detection

According to the WHO, safety signal is defined as "reported information on a possible causal relationship between an adverse event (AE) and a drug, of which the relationship is unknown or incompletely documented previously" [1]. The practice of "signal detection" is typically meant for marketed drugs where a safety concern may be new in nature not recognized in premarketing phases or a novel pattern of already known adverse drug reaction in terms of (severity, high risk sub-population, frequency, etc.). Signal detection can be "qualitative" built on case-by-case assessment of individual case safety reports) or "quantitative" via adopting machine learning or mining approaches using databases of real world data from clinical trials, electronic health records, biomedical literature, pharmacological databases, etc. Quantitatively/qualitatively detected signals necessitates further validation and confirmation by clinical judgment [18].

The aims of SD practices are to earlier discover potential safety concerns (completely novel or new pattern to previously known ADR); to distinguish signal of higher-order associations; to confirm signals that have been first identified from qualitative avenues; and to probably detect class effect safety issues.

1.4. Scope

In this paper, a systematic review was mainly conducted to formulate an answer to the question "How were informatics-driven approaches in terms of data mining and machine learning applied by researchers to identify DDI safety signals? The different aspects informatics-driven models existing in the underlying article are shown below in Section 7-Fig. 5.

In order to address the main objective of this review, a search strategy was conducted using the most prestigious biomedical literature databases (PubMed, ScienceDirect, Scopus, and Google Scholar). No time constraints were applied to be able to retrieve all potential relevant articles. The following eligibility criteria were used to find relevant studies:

Inclusion criteria

- Research/regular papers and review articles provided a ML framework for DDI prediction;
- Research/regular papers provided a data mining method for drug safety signal detection;
- Included information about evaluation for an informatics framework;
- Written in English

Exclusion criteria

- Studies not related to the detection or prediction of DDIs;
- Studies not applying any machine learning or data mining techniques;
- Studies not containing any quantitative results;
- Studies with only pharmacological scope;
- Not written in English.

As a summary, the remainder of this review is organized into different sections. We first discuss common data sources used in pharmacovigilance. We then explore informatics methods based on data mining techniques typically used by regulatory bodies for the quantitative PV signal detection and categories of ML algorithms deployed in DDI prediction. Finally, we draw conclusions along with outlining limitations and recommendations for future researches.

2. Widespread data sources in support of Pharmacovigilance

2.1. Postmarketing Pharmacovigilance Data: Spontaneous reporting systems

Postmarketing PV data are mainly available in several unique data sources through Spontaneous reporting systems (SRSs). This type of spontaneous data is the main source for postmarketing monitoring until now. SRSs are large databases for collecting individual case safety reports (ICSRs) of suspected adverse events passively reported to regulatory authorities by patients, healthcare professionals, and industry. The VigiBase (the WHO-UMC global database of ICSRs) [19], US Food and Drug Administration (FDA) adverse event reporting system (FAERS) [20, 21], the European EudraVigilance [22], and the "vaccine adverse event reporting systems (VAERS)" [23] are the most prominent ICSRs management systems.

Such kind of SRSs are developed to enable continuous monitoring of pharmaceutical and biological products within pharmacovigilance systems [24]. Generally, the structure of SRSs adheres to the international safety reporting guidance issued by the International Conference on Harmonisation (ICH E2B) that sets standards for reporting individual case safety reports (ICSR) [25]. Other type of SRSs is the proprietary company safety databases that contain ICSRs of a company pharmaceutical products. They can support early safety signal detection. The merits of SRSs are represented in providing information necessary for establishing causality assessment from real data and/or developing mining/ML models that might not be available in other data sources. For example but not limited to, apparent temporal time relationship, concomitant medications, indications, de-/re-challenge information, some demographic patient data, and action taken & outcome Information. Despite multibenefits of SRSs which playing a crucial role to support wise therapeutic decision-making in regulatory bodies, they have well-known boundaries. Those limitations can be summarized in over-/under- reporting rates, limited past medical history of patients, false signaling due to misattribution to hidden confounders/risk factors, and inability to indicate real incidence rates.

2.2. Electronic Healthcare Data

Electronic healthcare data consists of multiple sources like electronic health records (EHRs), administrative and insurance claims databases. EHRs are a kind of longitudinal observational systems where patients' medical records are created from multiple systems in healthcare organization(s). Many of the EHR fields are composed of unstructured data (e.g. discharge summaries, lab test findings, nurse notifications, etc.) and non-specific adverse events. Natural language processing (NLP) studies show how using EHRs data via adopting text mining approaches can be beneficial for extracting, encoding, and detecting safety signals [26–29]. The main benefit of EHRs from the aspect of PV is their capability to conduct active surveillance from real-work data [30–32]. Merits distinguish EHR systems from SRSs can represented in: not containing duplicates, largely unaffected by under- or over-reporting as they typically derived automatically, providing consistent information on most of the subject's drug exposure periods, clinically relevant events (independently from the exposure status) and valuable information on exposed subjects without events, and much more complete assessment of drug exposure and comorbidity status. For that, EHR systems may generate safety signals earlier than SRSs. Nevertheless, the abovementioned limitations consider the key boundaries for applying in regular PV. Other limitations can be represented in: difficulty to access due to privacy issues, and very challenging policy and technical issues to integrate EHRs from multiple sources. In EHRs, elevated rates of a medical diagnosis may indicate a safety issue (if subsequent to prescription), so some studies have hypothesized that metrics based on exploring longitudinal observational dataset for temporal associations between events over time rather than person-counts could produce more accurate associations. Mining of the EHRs data provides patterns of relationships among various entities offering enhanced medical intelligence unobtainable by other means [33,34].

2.3. Biomedical Literature

In the context of PV, biomedical literature may provide important medical data, from preclinical researches, clinical trials, observational studies, and case reports, which can be missed in regulatory PV activities due to focusing on coded and structured SRSs data. Additionally, drug interactions are frequently reported in pharmaceutical journals and technical reports, making medical literature the most effective source for the detection of DDIs. Therefore, developing NLP methods aiming at mining drug safety information for either single drug-adverse-event combinations or higher-order associations have been grown in recent years [35–37]. Despite its merits, this type of data sources has a number of deficiencies that can be summarized as following: (1) Information comprehensiveness is limited due to restricted financial access to many databases, ignorance of their existence by many people, non-standardization of their interface formats, in addition non-user friendly search engines; (2) Information quality is limited because no uniform guidelines exist for contents of the major text fields in database records (abstracts, titles, keywords, descriptors); (3) Compatibility among the contents of all record fields leading to different perspectives of the same technical topic; (4) Information retrieval is limited because time, cost, technical expertise, and substantial detailed technical analyses are required to retrieve the full scope of related records in a comprehensive process; (5) Unfiltered nature of biomedical literature is too statistically noisy to allow accurate signal detection. These deficiencies can result in two serious limitations: a substantial amount of relevant literature is not retrieved, and a substantial amount of non-relevant literature is retrieved [36]. In summary, it is very challenging to detect drug safety signals due to high statistical noise of the biomedical literature, so high quality text mining algorithms and/or machine learning needed to be developed [37].

2.4. Biomedical Knowledge Databases

There are a number of freely available biomedical knowledge databases, for example, DrugBank, STITCH, and PharmGKB [38–40], support authors to develop representation frameworks targeting the prediction of safety signals. This kind of resources provides information about drugs' chemical structures, biological elements, targets of pharmacological actions, functional/metabolic pathways, etc. Making use of this type of data sources offers some privileges, for example, allowing the opportunity to predict early-stage attrition due to new therapeutic candidates' toxicities during drug-development process [41]. This type of data sources has limitations represented in: (i) difficulty of modeling and simulating multifaceted metabolic pathways for thousands of drug candidates as it is crucial to preprocess molecules extracted from those databases with a set of cleaning rules to be representative dataset; (ii) no available databases completely describe the exact mechanisms by which drugs and genes may interact; (iii) complex informatics methods are required to explore those chemical or genetic data.

3. Data mining methods for Pharmacovigilance signal detection

3.1. Disproportionality analyses

Data mining is the process of extracting the patterns, associations or relationships among raw data using different analytical techniques involving the creation of a model to derive useful knowledge [42,43]. In PV, the data mining techniques are dedicated for hypothesis generation of new possible adverse drug events". Quantitative signal detection practices in large databases (e.g. VigiBase) are founded on data mining algorithms using disproportionality analyses (DPAs) [44,45]. DPAs are

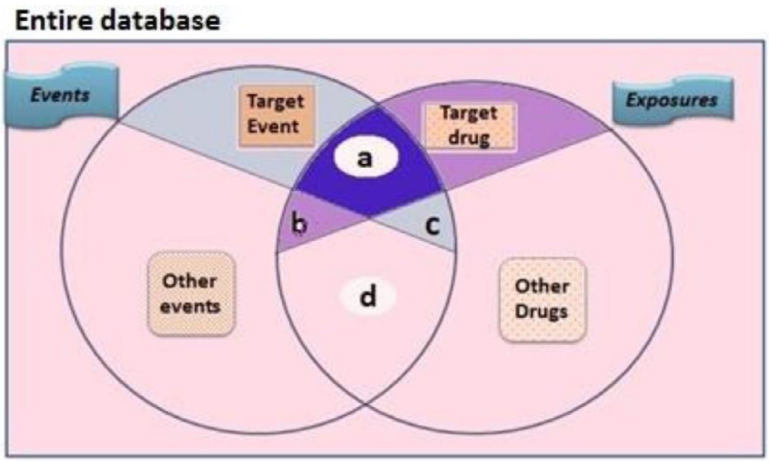

Fig. 1. Venn Diagram illustrating the entries {a, b, c, d} of contingency table. a, number of reports containing both the suspect drug and the suspect adverse event; b, number of reports containing the suspect drug with other adverse events (except the event of interest); c, number of reports containing the target adverse event with other medications (except the drug of interest); d, number of reports containing other medications and other adverse events.

categorized into two general classes (frequentist and Bayesian). Both categories utilize the entries of confusion matrix (as shown Fig. 1) to compute the magnitude of statistical association between drug-adverse-event combinations (DECs) in a PV database [46]. Their principles depend on estimating observed-to-expected reporting ratios for drug-reaction pairs using the total number of cases in background [47–52]. They quantify the unexpectedness of adverse event being reported to a drug or drug pair. All DPAs share the same basic of disproportionate reporting between observed to expected ratios, despite they differ in computation per each. Whenever a DEC's DPA computation exceeds the predefined minimum thresholds, it will be flagged as a statistical safety signal that requires further clinical assessment. Different thresholds are used for these algorithms to identify significant safety signals for DECs.

3.2. Association Rule Mining

Besides DPAs, there is an algorithm called *association rule mining* (abbreviated as ARM) [53]. ARM is considered a well-known mining algorithm for disclosing interesting patterns concealed within large databases. This algorithm was first advanced since more than a decade to be applied to the field of computer science, then expanded to various sciences [54–56]. Some studies have adopted ARM-based algorithms to detect adverse DDI patterns using SRSs [57]. Apriori algorithm is a kind of association rules mining which offers an appropriate representation of sparse data for complex computations [53]. Any association rule can be expressed in the form of $x \rightarrow y$ where x is an element from X itemset, while y is an element from Y itemset, noting that x & y are different items. To estimate directionality, the general Apriori algorithm works in two steps. The key of primary step is built on counting a frequent item set when an observed association between items exceeds the support threshold. While the second step depends on generating confident association rules fulfilling predefined minimum thresholds. From the scope of PV, X symbolizes an itemset of medications and y symbolizes an itemset of adverse-reactions. A significant association rule indicates that a specific DEC exceeds minimum thresholds of minimum confidence and support. The support of a DEC $S(X)$ is the observed numerals of reports having X. While support of an association rule $S(X \rightarrow Y)$ can be symbolized as $S(X \cup Y)$. The confidence of an association rule $C(X \rightarrow Y)$ indicates $(S(X \cup Y)) / (S(X))$ **(1)**. Confidence defines how frequently objects in Y represented in reports containing X. Confidence enables the estimation of conditional likelihood $(Pr(Y)/(X))$ **(2)** of Y in presence of X. An interesting association is flagged when a rule fulfills the minimum thresholds of both support & confidence, however some studies have suggested other measures to filter significant DDI associations [53].

3.3. Regression-based Approaches

Confounders are hidden covariates that may be hidden factors leading to either flagging spurious safety signals or delaying the detection of significant ones [58]. It is worth mentioning that confounders may infer a risk factor predisposing the adverse reaction or a key to identify higher risky patient subpopulations. A simpler type of confounders can be seen in variables (for example, age, gender, and year). They may be handled effectively by the stratification for each stratum using Mantel–Haenszel adjustments [59,60]. Nevertheless, adjusting huge numerals of possible confounders may result in missing signal detection in a timely manner [61–63]. Another limitation is represented in stratification by gender, age-group, etc. where the number of case reports are low and then infeasible to conduct subgroup analyses. Additionally, there are other forms of confounders called "innocent bystander" responsible for the occurrence of adverse events such as interacting drugs or indications of reported comedications. Unfortunately, using Mantel–Haenszel approaches for adjusting such kind of confounders is ineffective [60]. For large numbers of covariates, adopting logistic-regression (LR) approach is more efficient [61].

LR extends linear regression function by a sigmoid function to a value interval from 0 to 1 [62]. LR computes ROR by categorizing database's records as case-control records where a case is counted whenever having adverse-event of interest, whereas controls are counted whenever records having other adverse-events. In the context of PV, two studies have reported adopting LR modeling using SRSs aiming at DDI signal detection [63,64]. The comprehension of LR can be expressed according to the following formula:

$$\hat{y} = \hat{p}(x_1, x_2) = \log \frac{p}{1-p} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_n x_n \quad (3)$$

Where $\hat{y}$ denotes predicted outcome; $\hat{\beta}_0$ = coefficient of slope; x_0 = Slope; βn = coefficient of covariates n, X $_n$ = number of predictors or features covariates.

4. Paradigms of data mining and text mining studies using spontaneous reporting systems for detecting DDI signals

Nevertheless the limitations of any "spontaneous reporting system (SRS)" in regular PV practices, SRSs provide precious tool in detecting previously unknown safety concerns [65]. DPAs represent the most commonly applied algorithms in the area of quantitative signal detection within health authorities and large pharmaceutical industry [66,67].

The Shrinkage measure Ω, which is the extension of IC measure for the triplets {drug-drug-AE}, has been innovated by Norén *et al.* at Uppsala monitoring Centre (UMC) to screen potential DDI signals in the

VigiBase [9]. Authors have shown the privileges of Ω shrinkage measure compared to logistic regression in detecting interesting patterns of DDI signals. It can be calculated by estimating the proportion between observed and expected frequencies as seen in the following equation;

$$\Omega = log_2 \frac{n111 + \alpha}{E111 + \alpha} \quad (4)$$

Where E111 stands for the "expected" incidence value of drug-drug-event combination within a PV database; a is a constant equals to 0.5; Ω_{025} is the lower boundary of 95% confidence interval which flag a statistical DDI signal when its value ≥ 0 [9]. In 2013, Choi *et al.* have investigated the feasibility of using the national Korean health insurance system named "HIRA" in identifying DDI signals [68]. They adopted the shrinkage measure Omega Ω on the ICD-10 codes of diagnoses in association with a standard DDI dataset of NSAID and diuretics.

A Reporting ratios (RR) was used, as a conditional disproportionality analysis, to assess the increased of users searching for hyperglycemia-related terms given that they searched for both pravastatin and paroxetine. The results provided evidence to the potentiality of prediction of drug safety signals from search logs.

Almenoff *et al.* have examined the Bayesian disproportionality measure MGPS (90% confidence interval) in screening interaction profiles between antihypertensive drug "verapamil" and other classes of cardiovascular medications [69]. Results have shown the usefulness of MGPS disproportionality algorithm in mining DDI interesting patterns in polypharmacy conditions. Schuemie *et al.* [70] proposed an approach known as longitudinal GPS (LGPS) which is a modification of the original MGPS approach. LGPS computed the expected number of medical events during drug prescriptions based on an aggregate unexposed patient-time in exposed and unexposed patients. For protecting against confounding from dominating unexposed patients so long as the majority of the patients in a population don't receive specific medicine, authors suggested a filter named Observational Profiles of Adverse events Related to Drugs (LEOPARD) to eliminate spurious associations caused by protopathic bias. LGPS has been shown to outperform related methods, including MGPS.

In 2010, Harpaz *et al.* [71] have proposed an enhanced Apriori algorithm using FAERS data for the purpose of detecting interesting DDI patterns. Because of the unsuitability of classic Apriori metrics for drug safety field, RRR measure was utilized instead, with supplementary filter that each drug pair in association triplet should have RRR score higher than any of each drug individually. The second filter criterion was aimed to eliminate spurious DDI signals. Results exhibited nearly 35% of the drug-drug-reaction associations relating to well-known DDI, thus indicating the probable benefit of custom ARM in flagging potential DDI signals that warrant further clinical assessment. Another study, Ibrahim, H *et al.* [72] have presented an augmented data mining algorithm "hybrid Apriori", customized to PV applications, where disproportionality PRR measure was adopted as interestingness measure to discover association patterns of adverse DDIs from FAERS data. On the other side, Van Puijenbroek *et al.* [73] have considered an adjusted disproportionality measure ROR through implementing a model of logistic regression on the Netherlands's PV database "LAREB" to evaluate the adverse effects for the coadministration of "non-steroidal anti-inflammatory drugs (NSAID)" and "diuretics" on worsening the symptoms of congestive heart failure (CHF) disease. Results implied that using NSAIDs themselves has no role in increasing the risk of CHF. While upon coadministration with diuretics, NSAIDs have contributory role in aggravating CHF. The authors have concluded that adopting LR can offer advantages over bivariate disproportionality measures via guarding against false signals as a result of confounding by concomitant medications.

Thakrar *et al.* have illustrated the principal of multiplicative and additive modeling in discovering DDI signals using SRS [74]. The design of multiplicative/additive framework was built on estimating the interaction coefficient (δ) where the associated risk with drug pair coadministration is larger than those observed with each drug individually. Increased risk of a particular safety concern linked to specific drug(s) can be represented in higher incidence rates, higher observed-to-expected ratios, observed disproportionate reporting, etc. The statistical formula for a multiplicative model of "interaction coefficient δ'' associated with a drug pair can be calculated by a logarithmic linear regression equation as follows:

$$Log\ (risk\ of\ event) = \alpha + \beta\ (drug\ A) + \gamma\ (drug\ B) + \delta\ (drug\ A\ and\ B) + other\ covariates \quad (5)$$

While the statistical formula of the additive model for an "interaction coefficient δ'' associated with a drug pair can be calculated as shown in the below equation.

$$Risk\ of\ event = \alpha + \beta\ (drug\ A) + \gamma\ (drug\ B) + \delta\ (drug\ A\ and\ B) + other\ covariates \quad (6)$$

The principal of multiplicative model is based on assuming that a safety concern accompanied with a medication is multiplied to its background estimate, whereas the principal of additive model is founded on adding the safety concern linked to a drug into its background estimate. Regarding the multiplicative model, supposing the null hypothesis is real (i.e., non-DDI), the proportional risk of drug-drug-adverse-event combination is equal the product of the proportions of each drug individually ($PRRAB = PRRA \times PRRB$) (**7**). Likewise the additive model, if the proportional risk of an adverse event (AE) associated with a particular drug-pair is equal to the sum of proportions for the same AE of interest associated with each drug individually ($RRAB = PRRA + PRRB$) (**8**). DDI signal generated when multiplicative model $PRRAB/PRRA \times PRRB > 1$ or additive model $PRRAB - PRRA - PRRB > 0$ along with significant $p\text{-}value < 0.05$. The additive model in the context of DDI signal detection has shown higher sensitivity results compared to the multiplicative one. Hence, multiplicative models can be complementary evidence for DDI signals detected by the additive models.

There are different types of unstructured data sources which can be represented in narrative portions of EHRs, biomedical databases, etc. In recent years, researchers have proposed methods for mining such sources to effectively predict adverse DDI signals making up for the limitations of spontaneous reporting systems. Algorithms related to text mining and natural language processing (NLP) are commonly utilized to evolve these tasks.

Iyer *et al.* [75] have proposed an approach of natural language processing for disclosing DDI signals directly from the embedded verbatim parts of two big corpora of EHRs. They adopted adjusted disproportionality measures to detect significant associations of DDIs composed of 1165 distinct drugs and 14 distinct AEs. The authors' method has shown comparable performance between the DDIs signals discriminated from EHRs and those identified from SRSs.

Jon D. Duke *et al.* [76] have developed a new approach to mine mechanistic features from PV literature abstracts to predict potential DDI signals. Authors have specifically investigated interesting DDI patterns associated with the increased risk of myopathy and related musculoskeletal conditions. First, the authors applied two-step Information Retrieval (IR) approach to recognize the expression patterns relevant to DDI from PubMed abstracts. Then, they performed two logistic regression analyses to test each DDI effect on myopathy. They validated clinically significant DDI signals via using a database of electronic health records. Five novel DDI signals of increased risk of myopathy along with related cytochrome metabolizing enzymes were predicted.

LePendu *et al.* [77] proposed an approach for mining textual clinical notes corresponding to 1,044,979 patients derived from the Stanford Translational Research Integrated Database Environment (STRIDE) using odds ratio (OR) as frequency-based association measure. In purpose of generating hypotheses of DDIs signals, authors have used unstructured notes of EHRs rather than SRSs. Authors have evaluated their suggested approach in terms of sensitivity and specificity using gold stan-

dard of known interactions from DrugBank and the MediSpan® Drug Therapy Monitoring System™ (Wolters Kluwer Health, Indianapolis, IN). In overall, authors' method achieved 81.5% area under the Receiver Operator Characteristic (ROC) curve for signaling known DDIs.

Segura *et al.* [78] proposed the shallow linguistic kernel technique for automatic DDIs extraction from biomedical texts. As the recognition of drug names is an essential prerequisite step for the automatic discovery of DDIs from biomedical texts, authors used UMLS MetaMap Transfer (MMTx) tool to automatically identify drug entities occurring in unstructured textual documents of DDIs from DrugBank database. The MMTx recognizes and annotates drug entities according to UMLS semantic types (e.g., Clinical Drug [clnd], Pharmacological Substance [phsu], and Antibiotic [antb]). Drug-DDI corpus was created, annotated with 3169 DDIs, as the output of MMTx in XML format. Also, MMTx tool was used to extract and label candidates of drug pairs with 1 if interaction exists or labeled 0 if no interaction exists. The performance of proposed shallow linguistic kernel-based technique, with a precision of 51.03%, a recall of 72.82% and an F-measure of 60.01%, confirmed the remarkable impact of automatic entity recognition on the relation extraction task.

In 2013, White *et al.* [79] have demonstrated that the internet search logs that are generated by consumers are able to provide early clues about DIAEs. The authors conducted a large scale study of web search log paying attention on specific interaction between paroxetine (an antidepressant) and pravastatin (a cholesterol-lowering drug), which was recently reported in 2011 to cause hyperglycemia. Yang *et al.* [80] proposed a text mining methodology where they used social media to identify DDI patterns by combining ARM and lexicon-based algorithms. The authors used a data source called "MedHelp" as an example for an online healthcare forums. Only triplets were considered whereas each co-occurrence was counted for a pair of interacting drugs associated with an adverse event. They examined the significance of the constructed drug-drug-adverse-event associations using 4 metrics as follows {support, leverage, lift, and confidence} [81]. Some limitations related to the previous mentioned measures for example, the leverage could be very small or even negative because of the very small of threads related to ADR compared with the total number of threads. For that, authors adopted another measure called interaction ratio to capture DDI signals through which variation of confidence can be calculated. The generated patterns have proven the feasibility of the proposed approach in DDI signal detection from DrugBank database.

5. Machine learning approaches for the prediction of safety signals

In recent years, research efforts have directed from data mining approaches towards machine-learning (ML) to discover DDIs early during the drug discovery process before becoming commercially available. ML methods can be characterized into two broad categories: supervised and unsupervised [82]. In supervised learning, a model is built prior to the analysis where the model learns from labeled input data in order to estimate specific relationships between the features and the labels of the input data. Then, the learned model is applied to unseen data to predict the category or quantity of interest. Classification and regression are very common supervised learning techniques used in biomedical research. It is worth mentioning that regression-based approaches are traditionally adopted for either biostatistics or data mining purposes as illustrated in section 4. However, in several studies either linear or logistic regression are considered as ML models for binary/multi-class prediction of continuous/categorical features. Supervised ML algorithms may be used for feature selection, classification, prediction or evaluating the efficiency of some parameters [83]. In contrast, for unsupervised learning, the algorithm is applied directly on unlabeled input data to find regularities, and then a model will be built according to the identified patterns. Clustering and biclustering algorithms are the most common unsupervised ML techniques [84]. The following subsections from 5.1 through 5.5 summarize the most common ML algorithms adopted by researchers in the field of safety signal detection.

5.1. Support Vector Machine

Support vector machine (SVM) is the most popular non-probabilistic ML algorithm which was first developed by Vapnik and Lerner for binary classification problems [85]. The main concept of SVM is to create the optimal hyperplane/decision boundary. A hyperplane is a line/(n-1) dimensional plane that can separate n-dimensional input space into two classes so that a new data point can be easily categorized into the correct class [84, 86]. There are two types of SVM: linear SVM and non-linear SVM depending on whether data are linearly separable or not. For linear SVM, the distance between the decision boundary and the closest data points is referred to as the margin. As shown in Fig. 2, the optimal hyperplane is the line with the largest margin that can separate the two classes. The margin is calculated as the perpendicular distance from the hyperplane to only the closest points referred to as the support vectors of the hyperplane. Only these points are relevant in defining the line and in the construction of the classifier.

The real-world applications of SVM is subject to non-linearly separable data where a linear decision boundary cannot be used to classify the dataset. This leads to the emerging of a new generation of learning systems called kernel functions. A SVM Kernel is a function where it takes input training dataset with non-linearly separable decision surface and transforms the low-dimensional feature space into an abstract high-dimensional space. The most common kernel functions used in SMV modeling are Gaussian radial basis function (RBF), polynomial, and sigmoid kernels [87].

In this, we can conclude the merits and demerits of SVM. The key advantage is the kernel trick that enables SVM to create non-linear maximum margin decision boundaries in the original non-linearly separable feature space and consequently allows SVM to operate well in high-dimensional spaces. While disadvantages are represented mainly in: (i) Long training time for large datasets; (ii) its black-box nature that makes the final model difficult to interpret.

5.2. K-Nearest Neighbors (KNN)

K-nearest neighbors (KNN) is a popular ML algorithm belonging to the family of instance-based learning for the purposes of classification and regression prediction [88,89]. KNN at its core is a distance-based algorithm where the distance between two objects/data rows is calculated. KNN typically requires the calculation of distances between an example and k samples in the training dataset with the smallest distance to determine the prediction. While in the test set, the class label is predicted by only calculating distances for the nearest k neighbors (see Fig. 3). The most commonly used distance metrics in this regard are "Euclidean distance" and "Hamming distance" [90,91]. "Euclidean distance" metric is computed as the square root of the sum of the squared differences between the two vectors as seen in the following formula:

$$D_{Euclidean}(p,q) = \sqrt{\sum_{i=1}^{k} (q_i - p_i)^2} \tag{9}$$

While the "Hamming distance" metric computes the distance between two binary vectors to the sum of the bit differences between binary strings as seen in the following formula:

$$D_{Hamming}(p,q) = \sum_{i=1}^{k} |q_i - p_i| \tag{10}$$

Here $p = (p_1, \ldots., p_k)$ and $q = (q_1, \ldots., q_k)$ refer to vectors in k-dimensional space. The notations "p & q" are two objects in n-space; "q_i & p_i" are vectors starting from the origin of the space (initial point/object); n denotes n-space.

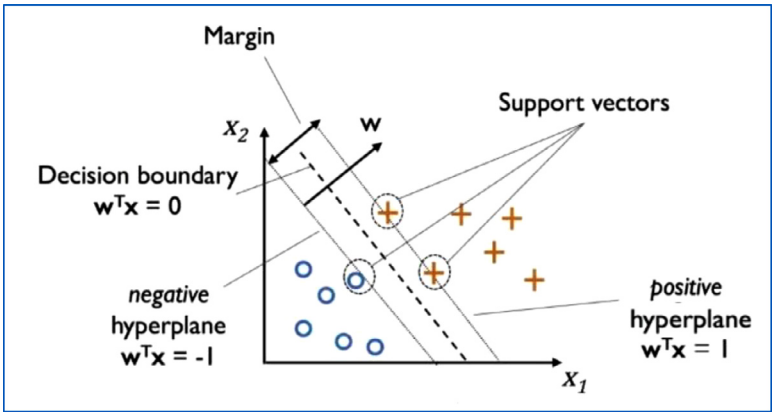

Fig. 2. Schematic representation of linear support vector machine algorithm concept. The ω denotes the normalized vector to the decision hyperplane; Horizontal and vertical axes (X_1 and X_2) symbolize the feature space (feature 1 & feature 2).

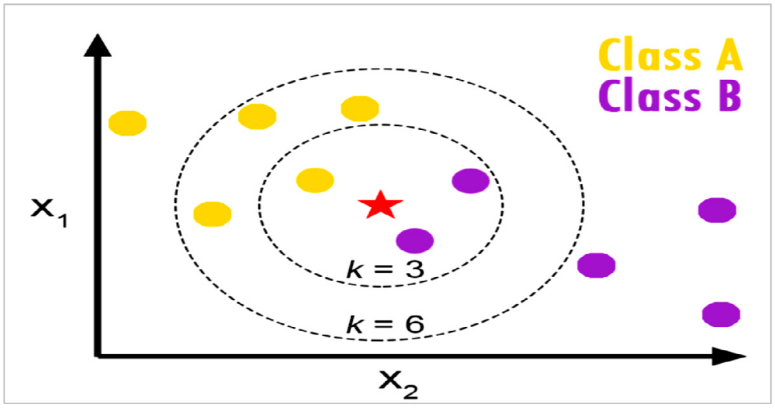

Fig. 3. Schematic representation of k-Nearest Neighbor (KNN) algorithm concept. Where, k = number of neighbors; X_1 & X_2 are the input descriptors; Red star is the query instance [89].

The main advantage of KNN algorithm is its capability to arbitrarily approximate any data distribution. However its demerits can be abridged as follows: (i) Computationally challenging as KNN uses all data points of the loaded training dataset for conducting queries; (ii) the number of data needed to be scaled grows exponentially with dimension by increasing the dataset size. It is known as the "curse of dimensionality".

5.3. Random Forest Classifier

Ensemble learning is a general meta-approach that combines predictions from multiple ML models to achieve better predictive performance. One of the powerful ML algorithm based on bagging ensemble learning is random forest classifier (abbreviated as "RFC"). It was developed by Leo Breiman [92]. Bagging is one of the main standard strategies used in ensemble learning techniques that fits many decision trees on different samples of the same dataset and averages the predictions (see Fig. 4) [93].

RFC is a supervised learning algorithm that is based on the infrastructure of decision tree algorithm and to classify either categorical or continuous features [94]. It estimates numbers of decision tree classifiers which fit on various subsets of a training dataset to augment the predictive performance and manage overfitting compared to "Decision tree algorithm" (refer to Fig. 5). The considered number of features at each sub-tree equals to the square root of the number of features in the input training dataset. Then, it averages the scores of each decision tree to predict the class of the test dataset instance [95,96].

Besides the high variance associated with tree classifiers due to substantial change in the training dataset resulting in very different subtrees and being non-interpretable, RFC is still privileged as a relatively fast training model that requires little data preparation can be adopted for multidimensional problems [97,98].

5.4. Artificial Neural Networks

Artificial Neural Networks (ANN) are a set of algorithms developed to imitate the neurons in human brain that are designed to recognize patterns in data which may be used for classification or clustering problems [99,100]. There are two main categories of ANN: (i) simple neural networks; (ii) deep learning neural networks (refer to Fig. 6). We here outline the general ANN architectures and common activation functions with taking into consideration deep learning neural networks are out of scope of the underlying review where ANNs.

There are two specific architectures for ANNs: (1) Feed-forward neural networks and (2) Feedback/Recurrent neural networks. *Feed-forward neural networks* are the most common architecture of ANN in practical applications. In this architecture, the information travels in one direction towards the output layer with one or multiple hidden layers. While *Feedback/recurrent neural networks (RNNs)* are more flexible and much difficult to analyze than feed-forward networks. RNNs can process variable length inputs by processing them in time steps, allowing the output of deeper layers to be fed back to previous layers in subsequent time steps. Therefore this type of ANN is typically utilized in sequential and time-series tasks.

Fig. 7 shows scheme of an interfacial neuron where a sequence of three operations can be represented as follows: (1) take an input feature i for a vector X and assign it a weight, then sum those inputs to obtain modulated features coming to the body of the neuron, and finally have an activation function f "*Perceptron*" that transforms those modulated features into the ultimate output/class [96].

There are various types of activation functions but the most common ones used in ANN modeling are Heaviside step function [101], Sigmoid function [102], Hyperbolic Tangent function (Tanh) [103], Rectified Linear Unit (ReLU) [104], Softmax function [105]. Heaviside step function is the classic and simplest ANN activation function with bi-

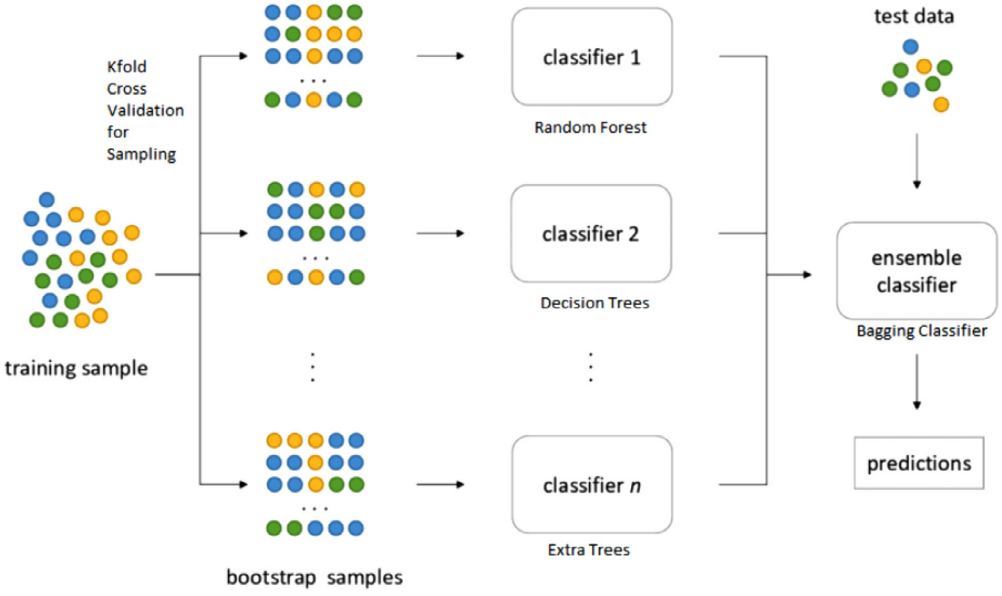

Fig. 4. Schematic representation of bagging ensemble learning concept.

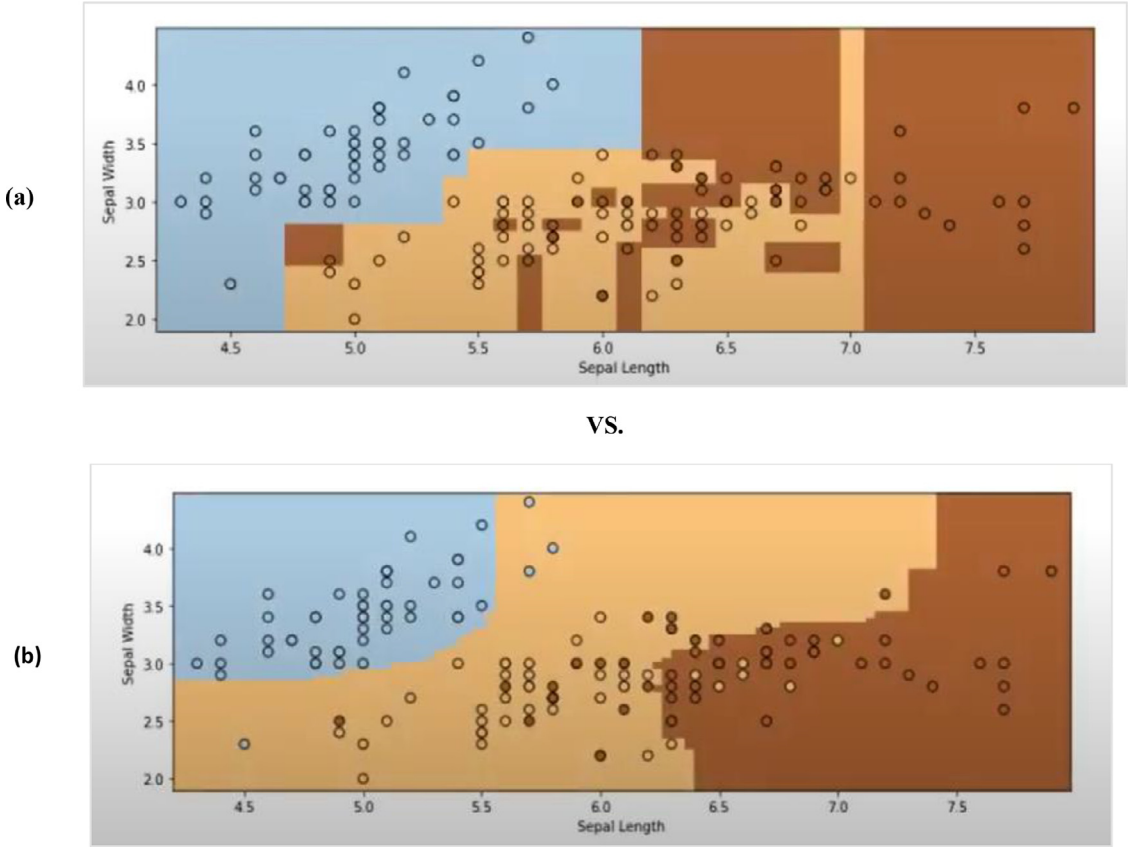

Fig. 5. Comparison of (a) decision boundaries of decision tree classifier versus (b) decision boundaries of random forest classifier on classic example "Iris dataset".

nary outcome that was first developed by Frank Rosenblatt (Refer to the "blue circle" in Fig. 7 above). It outputs 0 if the score is ≤ 0, and 1 otherwise.

Sigmoid is a non-linear activation function which is also known as "Logistic function". It provides a smooth and differentiated gradient curve. Sigmoid is typically used as the activation function in binary classification problems where its prediction output is normalized into the range [0–1]. Like Sigmoid, Hyperbolic tangent activation function is used in binary classification problems. Nonetheless, it is zero centric and the output values range from -1 to 1. B is the most popular activation function used in multiclass classification problems besides deep learning neural networks applications. ReLU allows the elimination of negative units in an ANN which diminish the sparse activation of only about 50%. Unlike the previously mentioned activation functions, Softmax's output is computed by the probabilities of modulated inputs to predict target class with the highest probability. The sum of the entire probabilities must be equal to 1. This function is commonly utilized in multiclass classification problems.

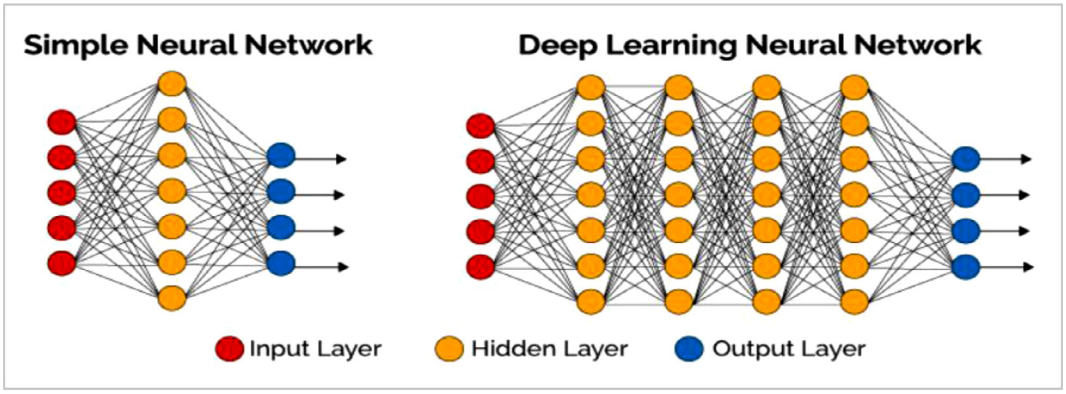

Fig. 6. Main types of artificial neural networks (ANN).

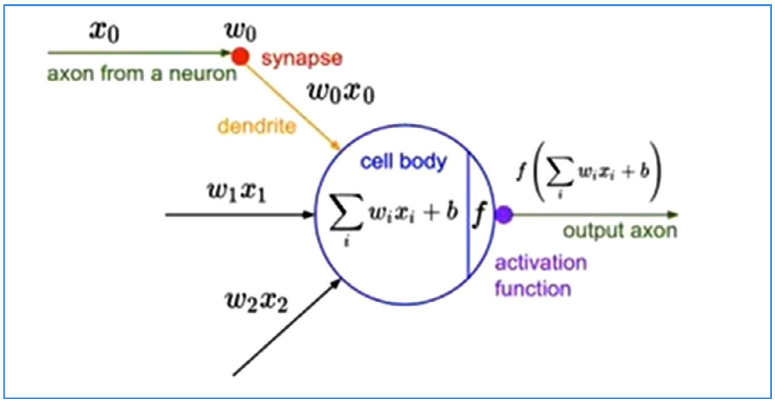

Fig. 7. Schematic representation of "Perceptron learning rule" to imitate an artificial neuron to biological neuron. X_i: Input feature i to a vector X in a single neuron; W_i: assigned weight for each input feature i .

5.5. Naive Bayes

Naive Bayes (NB) is one of the simplest supervised ML algorithms that is suitable for binary and multiclass classification of categorical input features. NB is based on Bayes' theorem to estimate posterior probability with the assumption of independence between features [106]. The comprehension of NB as a probabilistic classifier can be expressed as in the following equation:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{9}$$

Where $P(cx)$ denotes the posterior/conditional probability of class (target) given predictor (feature); $P(c)$ is the prior probability of class; $P(xc)$ symbolizes the likelihood which is the probability of predictor given class and $P(x)$ represents the prior probability of predictor.

The merit of NB modeling is that it is easy to build with no complicated iterative parameter estimation. It is particularly useful to implement on large datasets [107]. However, the main limitation of NB is the assumption that all the predictor features are completely independent.

6. Paradigms of machine learning studies for prediction of DDI signals

In the light of principles of the state-of-art ML algorithms described throughout section 5, in this section, we present paradigms of ML studies reported in biomedical literature. Two main types of ML approaches - (network-based [108–111] and similarity-based [17, 112–120]) - have been adopted by researchers for the purposes of predicting DDI safety signals.

Regarding network-based ML methods, they have been proposed by researchers for different objectives, including DDI prediction, drug-target-interaction (DTI) prediction, adverse drug reaction prediction, or drug repurposing [121]. For discovering unknown DDI signals, the strategies of network-based ML approaches depend on inferring drug similarities between network nodes or learning about topological features of the network structure [122,123]. Those knowledge networks can be constructed by extracting and integrating existing drug knowledge from one or multiple data sources (e.g. chemical, biological, target, genomic, pharmacological databases) leading to various shapes of networks (e.g. drug-drug, drug-target, protein-protein, pharmacodynamic, drug-gene, phenotypic, pharmacokinetic, etc.) [124].

Cami *et al.,* [108] suggested a predictive pharmacointeraction networks (PPIN) framework to predict DDIs by utilizing the network topological structure for all known DDIs of a knowledge database "Multum Vantage Rx", a commercial available database of curated DDIs and related side effects. Based on intrinsic and taxonomic properties of drugs nodes and edges representing interaction between nodes, the PPIN has reported 48% sensitivity, 90% specificity and 81% area under the receiver operating characteristic curve (AUROC). On the other hand, Huang *et al.*, [109] have conducted a study based on protein-protein-interaction (PPI) network for systematic prediction of PD DDIs. They designed two metrics of network topology to integrate phenotypic & genomic similarities between drug pairs. Numerous PD DDIs have been discovered with high accuracy score (0.82) and modest recall score (0.62). In 2014 [111], Cheng *et al.,* adopted a heterogeneous network-assisted inference (HNAI) framework for large-scale prediction of ligand–receptor DDIs. Their work depended on integrating drug phenotypic, therapeutic, chemical, and genomic properties. They tested five ML algorithms: Naive Bayes, Decision Tree, 3-nearest neighbors (or 3NN), logistic regression and support vector machine to predict DDIs. The key finding of the HNAI model was a reasonable area under the receiver operating characteristic (ROC) curve (AUC) (0.67) as evaluated using five-fold cross-validation. In 2019, Cheng *et al.* [110] have reported a network-based method for predicting beneficial drug-combinations discriminated from adverse DDIs for FDA-approved an-

tihypertensive drugs. Three types of experimentally evidenced clinical data sources (FDA-approved drug labels, Database of Clinicaltrials.gov, and published preclinical studies) were utilized. In this study two topological measures were adopted for capturing associations between elements of disease and drug–target networks in the human protein–protein interactome. Network-based proximity measure was used to capture the topological linking between two drug–target modules, whereas, Z-score was used to get closeness among disease and drug networks. As a result, six different relationships of drug-drug-disease modules could be distinguished by combining a number of decision rules.

In recent years, Similarity-based ML frameworks have been greatly adopted for classification tasks in the field of pharmacovigilance, particularly DDI prediction compared to feature vector-based frameworks. Similarity-based methods postulate that similar treatments may adversely interact with the same drug. A Similarity-based ML framework ideally composes of five phases that can be summarized as follows: (i) Feature capturing from one or more data sources; (ii) Constructing feature matrices usually show relationships between a drug and feature descriptors expressing various relationships between interacting drugs from various perspectives. These feature descriptors may reflect the chemical properties of a drug (e.g. 2d chemical substructures, 3d pharmacophores, anatomical therapeutic chemical "ATC" codes, etc.). They may also represent pharmacodynamic/pharmacokinetic properties of a drug (e.g. target proteins, metabolizing enzymes, transporter proteins, and carrier proteins). Moreover, descriptors may reveal some pharmacogenomics/ clinical properties (e.g. gene profile, interaction-profile, adverse-drug-event (ADE) profile, etc.); (iii) Computing feature similarity scores between bit-vectors using various types of similarity indices [125–134]; (iv) Implementing the prediction phase by adopting well-established ML algorithms; (v) Evaluating predictive performance results of the Similarity-based ML framework.

Those similarity-based ML frameworks have a number of notable merits compared to network-based or feature vector-based frameworks: (1) They do not require feature engineering that is usually a complex procedure; (2) Existing similarity indices (e.g. Tanimoto coefficient) are well-developed and broadly used; (3) higher predictive performance outcomes.

Ferdousi *et al.,* [17] have conducted a recent study for computational prediction of DDIs based on the similarities of the biological elements between drug pairs extracted from two publically available databases (DrugBank & KEGG). They compared 12 similarity measures for their ability to recall positive DDIs (i.e. known DDIs). Then, they selected the most reliable binary-similarity measure. They could extract 250,000 potential DDIs categorized as (low, moderate, high, very high) according to their prediction scores. They qualitatively validate few of the predicted DDI associations and no quantitative performance evaluation has been made. Vilar, S. *et al.,* [112–115] have conducted a series of studies built on similarity-based modeling using DrugBank as a data source. The use of drug structural similarity information for DDI prediction gave an overall precision of 0.26, sensitivity of 0.68, and specificity of 0.96 [112]. Vilar, S. *et al.*, [113] have successfully applied a similarity-based model for DDI prediction based on the interaction profile fingerprints (IPFs). A database of 17,230 DDI candidates along with their potential pharmacological effects has been provided for supporting decision in DDI detection and pharmacovigilance (PV) data analysis. This method has reported precision values ranging from 0.4–0.5. Whereas in 2014 [114], Vilar, S. *et al.*, have presented a protocol integrating five similarity fingerprints of drug pairs in purpose of large-scale DDI prediction. The study reported a robust AUROC > 0.95. Vilar, S. *et al.,* [115] have extended their work in 2015 targeting specific type of adverse DDIs using FDA PV database (FAERS). Their similarity-based method depended on capturing chemical and pharmacological features. Their results showed enhanced sensitivity, specificity and precision in comparison to the data mining algorithm proportional reporting ratio "PRR" traditionally applied on PV databases. Likewise, Sornalakshmi *et al.*, [116] included different similarity measures represented in 2D chemical structure, 3D pharmacological, interaction profile fingerprint (IPF), target, and ADE from DrugBank and SIDER data sources as ground truth for DDI labels, then trained SVM models to predict potential DDIs. Gottlieb, A. *et al.,* [117] have proposed the inferring drug interactions (INDI) method which infers both metabolizing enzyme CYP450-associated PK DDIs and PD DDIs based on chemical and side effect similarities. They made use of two publicly available resources to extract known DDIs & their severity. The results showed high AUC value ≥ 0.93, in addition to high sensitivity and specificity levels by using 10-fold cross-validation. Herrero-Zazo *et al.* [118] used Lexicomp and Vidal compendia as gold standards for DDI labels. They adopted supervised learning models using ANN constructed from DrugBank data based on PK and PD properties, along with drug–enzymes relationships. Tatonetti *et al.* have conducted a study where data from two different sources (FAERS and HER's Stanford hospital) have been linked aiming at detecting DDIs spontaneous reporting systems [119]. A predictive logistic regression model using FAERS data source has been constructed where eight classes of serious clinical adverse events (SAEs) profiles were generated. Two datasets were utilized, one for training and the other for testing. Nearly 171 potential DDI signals have been identified which validated using Stanford health records. The AUC values ranged from 0.51 to 0.71 for the eight models. In Vilar S *et al.* study, 2018, NLP has been adopted to extract the embedded representations of DDIs from PV literature, then apply ML to directly make predictions yielding good predictive performance in terms of sensitivity, but modest precision with numerous false signals [120].

7. Discussion

This paper investigates pharmacovigilance signal detection for drug-drug interactions from several aspects. Based on the illustration presented above in sections and subsections, herein we concluded a general framework of informatics-driven modelling that may be used by researchers for the purposes of adverse DDIs discovery (as shown in Fig. 8). One of the main objective of knowledge-based studies is to evaluate the efficiency of a model features in discovering novel DDI signals. Various metrics were used in purpose of evaluating the capability of informatics-based approaches in DDIs prediction. The most common metrics utilized in these studies are precision, sensitivity, specificity, accuracy, F1-measure and AUROC. Sensitivity indicates how well the model distinguishes adverse DDI signals. Specificity measures how well the model can disclose non-DDI signals. Accuracy measures how well the model can predict both DDI signals and non-DDI signals. AUROC metric is mainly used in ML studies to estimate how well specific descriptors can distinguish between two classes (DDI/ non-DDI).

Table 1 displays a comparison of performance metrics for either data mining or machine learning studies – (previously presented in sections 4.0 & 6.0) – in the context of DDI signal detection. There are no state-of-art methods that can be benchmarked. This is due to different subsets size, quality & diverse of data preparation approaches, different validation methods, etc. Henceforth, in Table 1, we focused the comparison on two main items (research context and data source), besides the well-established algorithms per each context (please refer to Fig. 8). From the data mining perspective, hybrid Apriori seems to provide the best performance among data mining algorithms in terms of sensitivity and precision.

Classic PV signal detection practices are focused on adopting DPA algorithms to mine SRS data for constituting hypotheses of single drug-AE combinations that need further investigation to establish evidence–based medicine to confirm or refute causality associations between those pairs. Then, regulatory actions may be taken to protect the public health. However, there is still limitations towards adopting data mining approaches in regular PV practices to generate novel hypotheses for DDI signals. Moreover, modest efforts are in place to mine DDI signals for experimental drugs where there is no or limited data in various data sources. Wherever a novel hypothesis for DDI signal is created, more

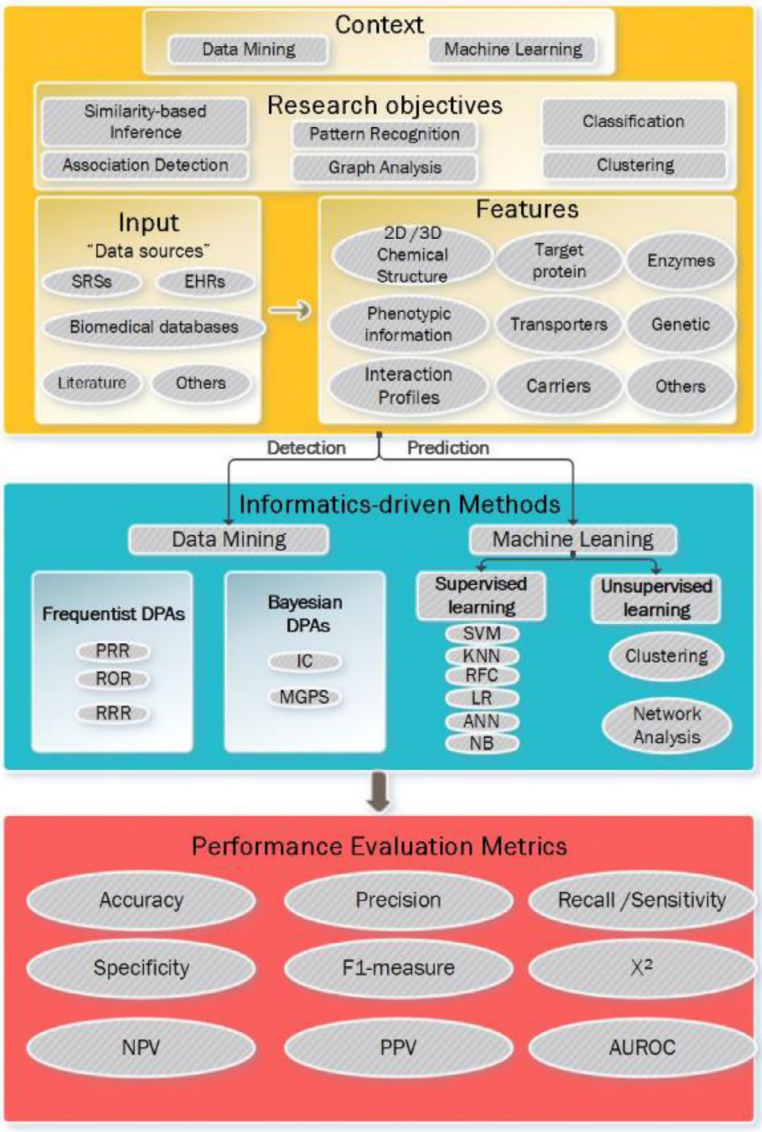

Fig. 8. A multi-layered informatics-driven framework for discovering drug-drug interactions safety signals. SRS: Spontaneous Reporting System; EHR: Electronic Health Records; PRR: Proportional Reporting Ratio; ROR: Relative Odds Ratio; RRR: Relative Reporting Ratio; IC: Information Component; MGPS: Multi-Gamma Poison Shrinker; SVM: Support Vector Machine; KNN: K-nearest neighbor; RFC: Random Forest Classifier; LR: Logistic regression; ANN: Artificial Neural Network; NB: Naïve Bayes; X^2 : Chi-Square; NPV: Negative Predictive Value; PPV: Positive Predictive Value; AUROC: Area under the ROC Curve.

Table 1
Summary of the Performance Metrics of Machine Learning and Data Mining methods adopted in the studies described in sections 4 and 6, respectively. AUCROC: Area under the Receiver Operator Characteristic curve; DDI: drug-drug interaction; SRS: spontaneous reporting system; ARM: Association rule mining; SVM: Support vector machine; IPF: Interaction profile fingerprint; PPIN: Protein-protein interaction network.

Research Objective	Method (Measurements)				
	Best Sensitivity	Best Precision	Best Specificity	Best Accuracy	Best AUROC
Data mining methods for DDI Signal detection implemented to SRS data					
Association pattern recognition (ARM)	Hybrid Apriori (0.81) [72]	Hybrid Apriori (0.85) [72]			
Machine learning methods for DDI Signal detection implemented into biomedical knowledge databases					
Classification					SVM (0.93) [117]
Similarity-based modeling			Molecular Structure Fingerprint (0.96) [112]		IPF (0.96) [113]
Network analysis			PPIN (0.96) [106]	PPIN (0.81) [109]	

experimental/clinical investigations may be needed to provide further evidences for confirming a DDI signal. Several factors have impact on detecting DDI signals where important ones may be missed or spurious associations are more likely to be produced. These influences can be summarized in: minimal demographic data, duplicated reports, no information regarding patient exposure, underreporting that may artificially increase expected risk estimates of the drugs, verbatim medical terms, unstandardized drug names, etc. Accordingly, optimizing current data mining techniques with chosen thresholds aiming at overcoming raw data challenges and capturing DDI signals with higher sensitivity and precision is crucial. Also, researchers have shown ARM and its extended algorithms as good choices for handling the sparse of postmarketing PV data within SRSs where strong association patterns could be generated to easily recognize higher order of Drug-AE associations. On the other side, adopting EHRs as a complementary source for routine PV practices generally and DDI signal detection specifically is challenging. This may be attributed to the ontology problems, heterogeneity of data from various sources and patients' information are mainly being in textual formats, besides confounding by co-medications, risk factors and/or comorbidities. Hence, developing NLP approaches and to be integrated with regression methods can be an option to handle ontology problems along with managing confounders. Another avenue of NLP can be adopted in regular PV data in purposes of reducing the large space of drug-drug-event associations inherently present in SRSs and allowing further processing in the context of DDI signal detection [134].

Whereas from ML perspective, PPIN framework provides a better overall performance concerning both accuracy and specificity compared to other methods. While, SVM appears to have the best overall DDI predictive performance in terms of AUROC compared to other supervised ML in predicting DDI signals [135–140]. However, SVM has inherited limitations represented in the time lapse of model running and the black box nature limiting the interpretation of the prediction outcomes. In recent years, the dawn of adopting supervised ML in PV signal detection has resulted from the evolution of multiple useful biomedical knowledge resources as gold standards of labeled DDIs (e.g. DrugBank [141], KEGG, Micromedex), besides using these resources to assess the predictive performance of supervised ML avenues.

The use of supervised ML has its limits such as class imbalance, sparse features, low overlapping and consistency between different DDIs resources [142]. Therefore, advancements in incorporating unsupervised ML approaches in DDIs discovery studies, improvements of DDI corpora annotation, and establishing standardized guidelines for establishing DDI gold reference datasets are probably essential steps to develop more realistic ML frameworks during drug-development, as well as, premarketing phases. Other directions of developments to be implemented into routine PV practices are represented in how ML may contribute in boosting DDI signal detection from SRS data and/or predicting their types/change in severity, combining more than one data sources (e.g. EHRs, patient support programs, prospective surveys, randomized controlled trials, etc.) rather than restricting to SRS data, assessing ML algorithmic performance by constructing reliable test sets, in addition to advancing frameworks for increasing transparency or explainability of ML outputs.

According to the abovementioned illustrations and beyond the structured PV data sources, the exponential increase of the scientific literature complicates the exploration of such biomedical corpora [143]. Biomedical corpora enables constructing learning datasets of medically related terms that embed prior knowledge and aid in the pragmatic use of certain words within specific contexts [120]. Informatics-based methods, in particular text mining and ML frameworks have great applicability in this regard to aid in extracting, analyzing and classifying biological information designated in scientific publications. During the past years, biomedical corpora have presented a valuable PV data source for the detection and analysis of DDI signals. The unstructured nature of this type of data sources is challenging. Natural Language Processing (NLP) approaches are crucial in this context to annotate, standardize and map relevant medical terminologies and generic drug names is crucial for implementing reliable computational models.

8. Concluding remarks

The availability of freely available safety data sources along with the adoption of novel DM and/or ML methods has advanced the PV domain. In this review, we particularly focus on developments in informatics techniques in the context of DDI signals. We have shown a portfolio of PV resources, DM and ML approaches techniques proposed to discover adverse DDI signals. Each method may re-stimulate interest to evolve DDI surveillance practices by offering different prospects. To the best we know, the underlying paper is the first review combining machine learning and data mining methods to disclose potential signals of DDIs in drug safety surveillance. Possible causal DDIs associations are experimented during drug development cycle, then monitored via postmarketing PV systems after being in market. Developing better predictive methods, to flag potential DDI signals, becomes of great importance to industry and regulatory bodies [144,145].

The black box nature of ML methods is one of their major limitations leading to difficulty in interpreting of what features are important and/or interpreting the model results. However, generally, adopting informatics-driven methods to be implemented in regular drug development process will be valuable. Since this will enable greater linkage between basic experimental platforms and controlled clinical settings early during drug development phases, then empowering the analysis of important safety concerns for investigational drugs at earlier stages leading to more efficient triage for novel drug candidates for further steps in the development process [146]. Boosted predictive approaches can be achieved by integrating structural and biological knowledge with enriched safety datasets [147]. Previous studies have shown benefits from linking drugs' phenotypic, therapeutic, structural, and genomic information in revealing DDI patterns in both drug discovery and postmarketing PV processes [148–150].

Innovative paradigms have arisen from exploiting various data sources compared to conventional PV practices, permitting for the active monitoring of DDI drug profiles. As spontaneous systems are considered the biggest assembly of real world data for distinguishing DDIs [151–154]. The discovery of higher-order drug–event combinations is more challenging than identifying signals of single drug–event pairs due to the limitation of under-reporting rates in large SRSs. Although the DDI detection research shifted away from SRSs utilization into other data sources in the recent years, this couldn't replace the main role of SRSs in regular PV. Accordingly, more efforts are required in purposes of evolving methods for routine DDIs discovery across different safety data sources.

Also, the main challenge in DDI signal detection studies is the lack of well-established guidances for assessing informatics-driven methods performance. Principally, this is due to the lack of reference standard for interaction safety profiles of the whole marketed drug products [155]. Hence, researchers should conduct more studies with purposes for acquiring better comprehension to the landscapes of multivariate association measures along with estimating corresponding predictive performance. Furthermore, developing approaches targeting the estimation of optimal DDI signal detection thresholds that achieve balanced trade-off between sensitivity and specificity with decreasing false signals can be a valuable perspective.

In conclusion, informatics-driven approaches don't prove causality association, but instead they can be valuable complementary tool to flag hypotheses of DDI combinations and support signal assessment. Comprehensive clinical judgment will always endure fundamental for establishing causal relationships between drugs and safety concerns.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this review paper.

References

[1] Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet 2000;356:1255–9.

[2] Hartford CG, Petchel KS, Mickail H, Perez-Gutthann S, McHale M, Grana JM, Marquez P. Pharmacovigilance during the pre-approval phases: an evolving pharmaceutical industry model in response to ICH E2E, CIOMS VI, FDA and EMEA/CHMP riskmanagement guidelines. Drug Saf 2006;29(8):657–73.

[3] Sanson-Fisher RW, Bonevski B, Green LW, D'Este C. Limitations of the randomized controlled trial in evaluating population-based health interventions. Am J Prev Med 2007;33(2):155–61.

[4] Amery WK. Why there is a need for pharmacovigilance. Pharmacoepidemiol Drug Saf 1999;8(1):61–4.

[5] Pirmohamed D, Orrne ML, et al. Drug Interactions of Clinical Importance. In: Davies D, et al., editors. Davies's Textbook of adverse Drug Reactions. London: Chapman & Hall Medical; 1998. p. 888–912. London.

[6] Gomba VK, Enslein K, Blake BW. Assessment of developmental toxicity potential of chemicals by quantitative structure–toxicity relationship models. Chemosphere 1995;31(1):2499–510.

[7] Van der Heijden PG, van Puijenbroek EP, van Buuren S, van der Hofstede JW. On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of underreporting on odds ratios. Stat Med 2002;21(14):2027e44.

[8] Olvey EL, Clauschee S, Malone DC. Comparison of critical drug-drug interaction listings: the Department of Veterans Affairs medical system and standard reference compendia. Clin Pharmacol Ther 2010;87(1):48e51.

[9] Norén G, Sundberg R, Bate A. A statistical methodology for drug-drug interaction surveillance. Stat Med 2008;27:3057–70.

[10] Bui PH, Quesada A, Handforth A, Hankinson O. The Mibefradil Derivative NNC55-0396, a Specific T-Type Calcium Channel Antagonist, Exhibits Less CYP3A4 Inhibition than Mibefradil Drug Metab. Dispos 2008;36(7):1291–9.

[11] Meinertz T. Mibefradil - a drug which may enhance the propensity for the development of abnormal QT prolongation. Eur Heart J 2001;3:K89–92 Suppl.

[12] US Food and Drug Administration [online]. Available from URL: http://www.fda.gov/Drugs/GuidanceComplianceRegulatory. Acessed 20 December 2020.

[13] Robertson S, Penzak S, Arthur JA, Darrell RA, Charles ED, et al.. Drug Interactions. Principles of Clinical Pharmacology. Burlington: Academic Press; 2007. p. 229–47.

[14] Patsalos PN, Perucca E. Clinically important drug interactions in epilepsy: general features and interactions between antiepileptic drugs. The Lancet Neurology 2003;2:347–56.

[15] Opie LH. Adverse cardiovascular drug interactions. Current Problems in Cardiology 2000;25:628–32.

[16] Boobis A, Gundert-Remy U, Kremers P, Macheras P, Pelkonen O. In silico prediction of ADME and pharmacokinetics: Report of an expert meeting organized by COST B15. Eur J Pharm Sci 2002;17:183–93.

[17] Ferdousi R, Safdari R, Omidi Y. Computational prediction of drug-drug interactions based on drugs functional similarities. Journal of Biomedical Informatics 2017;70:54–64.

[18] Alvarez Y, Hidalgo A, Maignen F, Slattery J. Validation of statistical signal detection procedures in EudraVigilance postauthorisation data: a retrospective evaluation of the potential for earlier signalling. Drug Saf 2010;33:475–87.

[19] Lindquist M. VigiBase, the WHO global ICSR database system: basic facts. Drug Inf J 2008;42(5):409–19.

[20] The (FAERS): Quarterly Data Files. Available: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm082193.htm. Acessed 04 November 2020.

[21] Zhan C, Roughead E, Liu L, Pratt N, Li J. Detecting high-quality signals of adverse drug-drug interactions from spontaneous reporting data. J Biomed Inform 2020;112:103603.

[22] EudraVigilance [online]. Available: http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000239.jsp. Acessed 11 December 2020.

[23] Vaccine adverse event reporting systems (VAERS). Available from URL: https://vaers.hhs.gov/data.html . Acessed 16 December 2020.

[24] Ivanov S, Lagunin A, Filimonov D, Poroikov V. Assessment of the cardiovascular adverse effects of drug-drug interactions through a combined analysis of spontaneous reports and predicted drug-target interactions. PLoS Comput Biol 2019;15(7).

[25] International Conference on Harmonisation, ICH E2B http://www.ich.org/products/guidelines/efficacy/article/efficacyguidelines.html. Acessed 24 October 2020.

[26] Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the United States, 1969-2002: the importance of reporting suspected reactions. Arch. Intern. Med. 2005;165(12):1363–9.

[27] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In: Proc AMIA Symp.; 2001. p. 17–21.

[28] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392–402.

[29] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36(6):462–77.

[30] Wang X, Hripcsak G, Markatou M, Friedman C. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. J Am Med Inform Assoc 2009;16(3):328–37.

[31] Segura-Bedmar I, Martínez P, De Pablo-Sánchez C. Combining syntactic information and domain-specific lexical patterns to extract drug–drug interactions from biomedical texts. In: Proceedings of the ACM fourth international workshop on data and text mining in biomedical informatics (DTMBIO'10); 2010. p. 49–56.

[32] Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug events from medical case reports. J Biomed Semantics 2012;20(1):15 3.

[33] Dandala B, Joopudi V, Tsou CH, Liang JJ, Suryanarayanan P. Extraction of Information Related to Drug Safety Surveillance From Electronic Health Record Notes: Joint Modeling of Entities and Relations Using Knowledge-Aware Neural Attentive Models. JMIR Med Inform 2020;10(7):e18417 8.

[34] Bouzillé G, Morival C, Westerlynck R, et al. An Automated Detection System of Drug-Drug Interactions from Electronic Patient Records Using Big Data Analytics. Studies in Health Technology and Informatics 2019;264:45–9. doi:10.3233/shti190180.

[35] Tari L, Anwar S, Liang S, Cai J, Baral C. Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics 2010;26(18):i547–53.

[36] Kostoff R. The extraction of useful information from the biomedical literature. ACADEMIC MEDICINE 2001;76(12):1265–70.

[37] Zhang T, Leng J, Liu Y. Deep learning for drug-drug interaction extraction from the literature: a review. Brief Bioinform 2020;25(5):1609–27 21PMID: 31686105. doi:10.1093/bib/bbz087.

[38] DrugBank website. Available at URL: http://www.drugbank.ca/. Acessed 14 August 2020.

[39] PharmGKB database. Available from URL: https://www.pharmgkb.org/index.jsp. Acessed 09 July 2020.

[40] Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res 2008;36(Database issue):D684–8.

[41] Vanderwall DE, et al. Molecular clinical safety intelligence: a system for bridging clinically focused safety knowledge to earlystage drug discovery - the GSK experience. Drug Discov Today 2011;16(15-16):646–53.

[42] Yang CC, Yang H. Mining heterogeneous networks with topological features constructed from patient-contributed content for pharmacovigilance. Artif Intell Med 2018;90:42–52. doi:10.1016/j.artmed.2018.07.002.

[43] Ting SL, Shum CC, Kwok SK, Tsang AHC, Lee WB. Data Mining in Biomedicine: Current Applications and Further Directions for Research. J Software Engineering & Applications 2009;2:150–9.

[44] Noguchi Y, Aoyama K, Kubo S, Tachi T, Teramachi H. Improved Detection Criteria for Detecting Drug-Drug Interaction Signals Using the Proportional Reporting Ratio. Pharmaceuticals 2021;14(4).

[45] Noguchi Y, Tachi T, Teramachi H. Comparison of Signal Detection Algorithms Based on Frequency Statistical Model for Drug-Drug Interaction Using Spontaneous Reporting Systems. Pharm Res 2020;30(5):86 37. doi:10.1007/s11095-020-02801-3.

[46] Hauben M, Zhou X. Quantitative Methods in Pharmacovigilance: focus on signal detection. Drug Saf 2003;26(3):159–86.

[47] Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001;10(6):483–6.

[48] Evans SJ. Pharmacovigilance: a science or fielding emergencies? Stat Med 2000;19(23):3199–209.

[49] Chee BW, Berlin R, Schatz B. Description: Predicting Adverse Drug Events from Personal Health Messages. AMIA Annu Symp Proc 2011:217–26.

[50] Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De-Freitas RM. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 1998;54(4):315–21.

[51] Bate A, Lindquist M, Edwards IR, Orre R. A data mining approach for signal detection and analysis. Drug Saf 2002;25(6):393–7.

[52] DuMouchel W. Bayesian data mining in large frequency tables, with an application the FDA Spontaneous Reporting System. Am. Stat. 1999;53(3):177–90.

[53] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD international conference on management of data; 1993. p. 207–16.

[54] Iskander J, Pool V, Zhou W, English-Bullard R. Data mining in the US using the vaccine adverse event reporting system. Drug Saf 2006;29(5):375–84.

[55] Carrino JA, Ohno-Machado L. Development of radiology prediction models using feature analysis. Acad Radiol 2005;12(4):415–21.

[56] Sarawagi S, Thomas S, Agrawal R. Integrating association rule mining with relational database systems: alternatives and implications. Data Mining Knowledge Discov 2000;4(2):89–125.

[57] Iyer SV, Lependu P, Harpaz R, Bauer-Mehren A, Shah NH. Learning Signals of Adverse Drug-Drug Interactions from the Unstructured Text of Electronic Health Records. AMIA Jt Summits Transl Sci Proc 2013:216.

[58] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Nat Cancer Inst 1959;22(4):719–48.

[59] Noren GN, Bate A, Orre R, Edwards IR. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. Stat Med 2006;25(21):3740–57.

[60] Hopstadius J, Norén GN, Bate A, Edwards IR. Impact of stratification on adverse drug reaction surveillance. Drug Saf 2008;31(11):1035–48.

[61] Nicholas P. Jewell.Statistics for Epidemiology. Texts for statistical sciences. Chapman and Hall; 2004. Hardcover edition.

[62] Van Puijenbroek EP, Egberts AC, Heerdink ER, Leufkens HG. Detecting drug-drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs. Eur J Clin Pharmacol 2000;56(9-10):733–8.

[63] Van Puijenbroek EP, Egberts AC, Meyboom RH, Leufkens HG. Signalling possible drug–drug interactions in a spontaneous reporting system: delay of withdrawal bleeding during concomitant use of oral contraceptives and itraconazole. Br J Clin Pharmacol 1999;47(6):689–93.

[64] Van Puijenbroek EP, Egberts AC, Heerdink ER, Leufkens HG. Detecting drug–drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs. Eur J Clin Pharmacol 2000;56(9-10):733–8.

[65] Rawlins MD. Spontaneous reporting of adverse drug reactions. II: Uses. Br J Clin Pharmacol 1988;26(1):7–11.

[66] Roux E, Thiessard F, Fourrier A, Bégaud B. Tubert-Bitter P. Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. IEEE Tran INF Technol Biomed 2005;9(4):518–27.

[67] Van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R, Egberts AC. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002;11(1):3–10.

[68] Choi CA, Chang MJ, Choi HD, Chung WY, Shin WG. Application of a drug-interaction detection method to the Korean National Health Insurance claims database. Regul Toxicol Pharmacol; 2013;67(2):294–8.

[69] Yang X, Fram D. Using disproportionality analysis as a tool to explore drug-drug interavtions in AERS database. Pharmacoepidemiol Drug Safety 2004;13:S247.

[70] Schuemie MJ. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. Pharmacoepidemiol Drug Saf 2011;20:292–9.

[71] Harpaz R, Haerian K, Chase HS, Friedman C. Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. Proc of AMIA Annu Symp 2010:281–5128.

[72] Ibrahim H, Saad A, Abdo A, Sharaf EA. Mining association patterns of drug-interactions using post marketing FDA's spontaneous reporting data. J Biomed Inform 2016;60:294–308.

[73] Van Puijenbroek EP, Egberts AC, Heerdink ER, Leufkens HG. Detecting drug–drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs. Eur J Clin Pharmacol 2000;56(9-10):733–8.

[74] Thakrar BT, Grundschober SB, Doessegger L. Detecting signals of drug–drug interactions in a spontaneous reports database. Br J Clin Pharmacol 2007;64(4):489–95.

[75] Iyer SV, Harpaz R, LePendu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug–drug interactions. J. Am. Med. Inform. Assoc. 2014;21(2):353–62.

[76] Duke JD, et al. Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. PLoS Comput Biol 2012;8(8):e1002614.

[77] LePendu P, Iyer SV, Fairon C, Shah NH. Annotation analysis for testing drug safety signals using unstructured clinical notes. J Biomed Semantics 2012;3(Suppl 1):S5.

[78] Segura-Bedmar I, Martínez P, de Pablo-Sánchez C. Using a shallow linguistic kernel for drug–drug interaction extraction. J Biomed Inform 2011;44(5):789–804.

[79] White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc 2013;20(3):404–8.

[80] Yang H, Yang CC. Harnessing Social Media for DrugDrug Interactions Detection. In: Proceedings of IEEE International Conference on Healthcare Informatics; 2013. p. 22–9.

[81] Yang CC, Yang H, Jiang L. Postmarketing Drug Safety Surveillance using Publicly Available Health Consumer Contributed Content in Social Media. ACM Transactions on Management Information Systems (TMIS) 2014;5(1):2–21.

[82] Cattral R, Oppacher F, Deugo D. Supervised and Unsupervised Data Mining with an Evolutionary Algorithm. Proceedings of the Congress on Evolutionary Computation 2001:767–74.

[83] Zheng Y, Peng H, Zhang X, et al. DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions. BMC Bioinformatics 2019;20(661).

[84] Han J, Pei J, Yin Y, Mao R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery 2004;8:53–87.

[85] Vapnik V, Lerner A. Pattern Recognition Using Generalized Portrait Method. Automat. Remote Contr. 1963;24:774–80.

[86] García-Gonzalo E, Fernández-Muñiz Z, García Nieto PJ, Bernardo Sánchez A, Menéndez Fernández M. Hard-Rock Stability Analysis for Span Design in Entry–Type Excavations with Learning Classifiers. Materials (Basel) 2016;9(7):531.

[87] Patle A, Chouhan DS. SVM kernel functions for classification. In: International Conference on Advances in Technology and Engineering (ICATE); 2013. p. 1–9. doi:10.1109/ICAdTE.2013.6524743.

[88] Ayyad SM, Saleh AI, Labib LM. Gene expression cancer classification using modified K-Nearest Neighbors technique. Biosystems 2019;176:41–51.

[89] Yan C, Duan G, Pan Y, Wu FX, Wang J. DDIGIP: predicting drug-drug interactions based on Gaussian interaction profile kernels. BMC Bioinformatics 2019;24(Suppl 15):538 20. doi:10.1186/s12859-019-3093-x.

[90] Zhang Q, Guan X, Wang H, Pardalos PM. Maximum shortest path interdiction problem by upgrading edges on trees under hamming distance. Optimization Letters 2021:1–20.

[91] Wang D, Tan X. Robust distance metric learning via Bayesian inference. IEEE Transactions on Image Processing 2017;27(3):1542–53.

[92] Breiman L. Random Forests. Machine Learning 2001;45:5–32.

[93] Yaman E, Subasi A. Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Automated EMG Signal Classification. Biomed Res Int 2019;2019:9152506.

[94] Wang MWH, Goodman JM, Allen TEH. Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models. Chem Res Toxicol 2021;34(2):217–39.

[95] Azar AT, Elshazly HI, Hassanien AE, Elkorany AM. A random forest classifier for lymph diseases. Comput Methods Programs Biomed 2014;113(2):465–73.

[96] Sahoo AJ, Kumar Y. Seminal quality prediction using data mining methods. Technol Health Care 2014;22(4):531–45.

[97] Mittal M, Balas VE, Goyal LM, Kumar R. Big Data Processing Using Spark in Cloud. Springer; 2018.

[98] Hansen PW, Clemmensen L, Sehested TSG, Fosbol EL, Torp-Pedersen C, Kober L, et al. Identifying Drug-Drug Interactions by Data Mining: A Pilot Study of Warfarin-Associated Drug Interactions. Circ. Cardiovasc. Qual. Outcomes. 2019;9:621–8.

[99] Rosen CA. Pattern classification by adaptive machines. Science 1967;156(3771):38–44.

[100] Shtar G, Rokach L, Shapira B. Detecting drug-drug interactions using artificial neural networks and classic graph similarity measures. PLoS One 2019;14(8):e0219796. doi:10.1371/journal.pone.0219796.

[101] Huang GB, Chen YQ, Babri HA. Classification ability of single hidden layer feedforward neural networks. IEEE Trans Neural Netw 2000;11(3):799–801.

[102] Ranka S, Mohan CK, Mehrotra K, Menon A. Characterization of a Class of Sigmoid Functions with Applications to Neural Networks. Neural Netw 1996;9(5):819–35.

[103] Mathias AC, Rech PC. Hopfield neural network: the hyperbolic tangent and the piecewise-linear activation functions. Neural Netw 2012;34:42–5.

[104] Wang SH, Phillips P, Sui Y, Liu B, Yang M, Cheng H. Classification of Alzheimer's Disease Based on Eight-Layer Convolutional Neural Network with Leaky Rectified Linear Unit and Max Pooling. J Med Syst 2018;42(5):85.

[105] Cao C, Liu F, Tan H, et al. Deep Learning and Its Applications in Biomedicine. Genomics Proteomics Bioinformatics 2018;16(1):17–32.

[106] Zhang Z. Naïve Bayes classification in R. ANN Transl Med 2016;4(12):241.

[107] Huang Y, Li L. Naive Bayes classification algorithm based on small sample set. In: IEEE International Conference on Cloud Computing and Intelligence Systems; 2011. p. 34–9. doi:10.1109/CCIS.2011.6045027.

[108] Cami A, Manzi S, Arnold A, et al. Pharmacointeraction network models predict unknown drug-drug interactions. PLoS ONE 2013;8:e61468.

[109] Huang J, Niu C, Green CD, Yang L, Mei H, Han J-DJ. Systematic Prediction of Pharmacodynamic Drug-Drug Interactions through Protein-Protein-Interaction Network. PLoS Comput Bio 2013;l9(3):e1002998.

[110] Cheng F, Kovács IA, Barabási AL. Network-based prediction of drug combinations. Nat Commun 2019;10(1):1806.

[111] Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. J Am Med Inform Assoc 2014;21:e278–86.

[112] Vilar S, Harpaz R, Uriarte E, Santana L, Rabadan R, Friedman C. Drug-drug interaction through molecular structure similarity analysis. J Am Med Inform Assoc 2012;19(6):1066–74.

[113] Vilar S, Uriarte E, Santana L, Tatonetti NP, Friedman C. Detection of drug–drug interactions by modeling interaction profile fingerprints. PLOS ONE 2013;8(3):e58321.

[114] Vilar S, Uriarte E, Santana L, Lorberbaum T, Hripcsak G, Friedman C, Tatonetti N. Similarity-based modeling in large-scale prediction of drug-drug interactions. Nature Protocols 2014;9(9):2147–63.

[115] Vilar S, Lorberbaum T, Hripcsak G, Tatonetti N. Improving detection of arrhythmia drug-drug interactions in pharmacovigilance data through the implementation of similarity-based modeling. PLoS ONE 2015;10(6).

[116] Sornalakshmi K, et al. A survey on using social media data analytics for pharmacovigilance. Res. J. Pharm. Technol. 2017(10):3474–8.

[117] Gottlieb A, Stein G, Oron Y, Ruppin E, Sharan R. INDI: a computational framework for inferring drug interactions and their associated recommendations. Mol Syst Biol 2012;8:592.

[118] Herrero-Zazo M, Lille M, Barlow D. Application of machine learning in knowledge discovery for pharmaceutical drug-drug interactions. KDWeb; 2016.

[119] Tatonetti NP, Fernald GH, Altman RB. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. J AM Med Inform Assoc 2012;19(1):79–85.

[120] Vilar S, Friedman C, Hripcsak G. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. Brief Bioinform 2018;19(5):863–77.

[121] Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, et al. Predicting new molecular targets for known drugs. Nature 2009;462:175–81.

[122] Dhami DS, Kunapuli G, Das M, Page D, Natarajan S. Drug-Drug Interaction Discovery: Kernel Learning from Heterogeneous Similarities. Smart Health (Amst) 2018 9-10:88-100. doi:10.1016/j.smhl.2018.07.007.

[123] Park K, Kim D, Ha S, Lee D. Predicting pharmacodynamic drug-drug interactions through signaling propagation interference on protein-protein interaction networks. PloS one 2015;10(10).

[124] Berger SI, Iyengar R. Network analyses in systems pharmacology. Bioinformatics 2009;25(19):2466–72.

[125] Willett P. Similarity-based approaches to virtual screening. Biochem Soc Trans 2003;31:603–6.

[126] Cha S, Tappert C, Yoon S. Enhancing Binary Feature Vector Similarity Measures. Journal of Pattern Recognition Research 2006:63–77.

[127] Jiajing Z, Yongguo L, Chuanbiao W. MTMA: Multi-task multi-attribute learning for the prediction of adverse drug–drug interaction. Knowledge-Based Systems 2020;199 105978 (ISSN 0950-7051).

[128] Jackson D, Somers K, Harvey H. Similarity Coefficients: Measures of Co-Occurrence and Association or Simply Measures of Occurrence? The American Naturalist 1989;133(3):436–53.

[129] Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 2008;24(13):i232–40.
[130] Bleakley K, Yamanishi Y. Supervised prediction of drug target interactions using bipartite local models. Bioinformatics 2009;25(18):2397–403.
[131] Jacob L, Vert JP. Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 2008;24(19):2149–56.
[132] Van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics 2011;27(21):3036–43.
[133] Gonen M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics 2012;28(18):2304–10.
[134] Gattepaille L. Using the WHO database of spontaneous reports to build joint vector representations of drugs and adverse drug reactions, a promising avenue for pharmacovigilance. In: IEEE International Conference on Healthcare Informatics (ICHI); 2019. p. 1–6.
[135] Song D, Chen Y, Min Q, Sun Q, Ye K, Zhou C, et al. Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies. J Clin Pharm Ther 2019;18(2):268–75 44.
[136] Mahadevan AA, Vishnuvajjala A, Dosi N, Rao S. A Predictive Model for Drug-Drug Interaction Using a Similarity Measure. In: 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); 2019. p. 1–8. doi:10.1109/CIBCB.2019.8791458.
[137] Song D, Chen Y, Min Q, Sun Q, Ye K, Zhou C, Yuan S, Sun Z, Liao J. Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies. J Clin Pharm Ther 2019;44(2):268–75. doi:10.1111/jcpt.12786.
[138] Wu H-Y, Chiang C-W, Li L. Text mining for drug-drug interaction. Methods Mol Biol 2014;1159:47–75.
[139] Dere S, Ayvaz S. Prediction of Drug-Drug Interactions by Using Profile Fingerprint Vectors and Protein Similarities. Healthc Inform Res 2020;26(1):42–9. doi:10.4258/hir.2020.26.1.42.
[140] Liu N, Chen CB, Kumara S. Semi-Supervised Learning Algorithm for Identifying High-Priority Drug-Drug Interactions Through Adverse Event Reports. IEEE J Biomed Health Inform 2020;24(1):57–68. doi:10.1109/JBHI.2019.2932740.
[141] Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2017. doi:10.1093/nar/gkx1037.
[142] Ayvaz S, Horn J, Hassanzadeh O, et al. Toward a complete dataset of drug–drug interaction information from publicly available sources. J Biomed Inform 2015;55:206–17.
[143] Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? Mol Cell 2006;21:589–94.
[144] Percha B, Garten Y, Altman RB. DISCOVERY AND EXPLANATION OF DRUG-DRUG INTERACTIONS VIA TEXT MINING. Pac Symp Biocomput 2012:410–21.
[145] Bjornsson TD, Callaghan JT, Einolf HJ, Fischer V, Gan L, et al. The conduct of in vitro and in vivo drug-drug interaction studies: A Pharmaceutical Research and Manufacturers of America (PhRMA) perspective. Drug Metabolism and Disposition 2003;31:815–32.
[146] De-Boer A. When to publish measures of disproportionality derived from spontaneous reporting databases? Br J Clin Pharmacol 2011;72(6):909–11.
[147] Vilar S, Uriarte E, Santana L, Tatonetti NP, Friedman C. Detection of drug-drug interactions by modeling interaction profile fingerprints. PLoS One 2013;8(3):e58321.
[148] Vilar S, Harpaz R, Chase HS, et al. Facilitating adverse drug event detection pharmacovigilance databases using molecular structure similarity: application rhabdomyolysis. J Am Med Inform Assoc 2011;18(Suppl 1):i73–80.
[149] Wishart DS, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2008;36(Database issue):D901–6.
[150] Bjornsson TD, Callaghan JT, Einolf HJ, Fischer V, Gan L, et al. The conduct of in vitro and in vivo drug-drug interaction studies: A Pharmaceutical Research and Manufacturers of America (PhRMA) perspective. Drug Metabolism and Disposition 2003;31(7):815–32.
[151] Matthews EJ, et al. Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part B. Use of (Q)SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities. Regul Toxicol Pharmacol 2009;54:23–42.
[152] Frid AA, Matthews EJ. Prediction of drug-related cardiac adverse effects in humans–B: use of QSAR programs for early detection of drug-induced cardiac toxicities. Regul Toxicol Pharmacol 2010;56:276–89.
[153] Greene N, Judsona PN, Langowskia JJ, Marchanta CA. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR, and METEOR. SAR & QSAR. Environ. Res. 1999;10(2-3):299–313.
[154] Song D, Chen Y, Min Q, Sun Q, Ye K, Zhou C, Yuan S, Sun Z, Liao J. Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies. J Clin Pharm Ther 2019;44(2):268–75. doi:10.1111/jcpt.12786.
[155] Lu Z. Information technology in pharmacovigilance: Benefits, challenges, and future directions from industry perspectives. Drug Healthc Patient Saf 2009;1:35–45.