International Journal of Medical, Pharmacy and Drug Research (IJMPD)
[Vol-6, Issue-6, Nov-Dec 2022]
ISSN: 2456-8015
https://dx.doi.org/10.22161/ijmpd.6.6.4

Peer-Reviewed Journal

Application of Machine Learning in Drug Discovery and Development Lifecycle

Geerisha Jain

Department of Computer Science Engineering, Vellore Institute of Technology, Vellore, India

Received: 11 Oct 2022; Received in revised form: 11 Oct 2022; Accepted: 15 Nov 2022; Available online: 20 Nov 2022

Abstract— *Machine learning and Artificial Intelligence have significantly advanced in recent years owing to their potential to considerably increase the quality of life while reducing human workload. The paper demonstrates how AI and ML are used in the drug development process to shorten and enhance the overall timeline. It contains pertinent information on a variety of Machine Learning approaches and algorithms that are used across the whole drug development process to speed up research, save expenses, and reduce risks related to clinical trials. A range of QSAR analysis, hit finding, and de novo drug design applications are used in the pharmaceutical industry to enhance decision-making. As technologies like high-throughput screening and computation analysis of databases used for lead and target identification and development create and integrate vast volumes of data, machine learning and deep learning have grown in importance. It has also been emphasized how these cognitive models and tools may be used in lead creation, optimization, and thorough virtual screening. In this paper, problem statements and the corresponding state-of-the-art models have been considered for target validation, prognostic biomarkers, and digital pathology. Machine Learning models play a vital role in the various operations related to clinical trials embracing protocol optimization, participant management, data analysis and storage, clinical trial data verification, and surveillance. Post-development drug monitoring and unique industrially prevalent ML applications of pharmacovigilance have also been discussed. As a result, the goal of this study is to investigate the machine learning and deep learning algorithms utilised across the drug development lifecycle as well as the supporting techniques that have the potential to be useful.*

***Keywords*—Machine Learning, Artificial Intelligence, Drug Discovery, Drug Development, Pharmacovigilance**

I. INTRODUCTION

Over the last ten years, machine learning (ML) has been more popular in the area of medicine. Since the middle of the 20th century, machine learning has been explored, but recent advances in computing power, data accessibility, cutting-edge techniques, and a broad variety of technical expertise have expedited its use in healthcare.

Machine learning methods have been used by drug companies since 1962. These techniques make it easier to gather pertinent characteristics, which advances our knowledge of complex biological systems. The pharmaceutical industry is increasingly using many prediction models to enhance the drug development process. We can finally acquire answers to topics that present a higher challenge to chemists, all thanks to the algorithms used by various computational methodologies. They aid chemists in accurately modeling, analyzing, and forecasting a variety of biological responses with regard to drug design. With the help of the annotated data, machine learning algorithms learn intricate patterns to predict the annotations of new test data sets [1]. Genome association, protein function prediction, and other tasks involve the application of machine learning. It helps in comprehending a diverse array of drug features such as solubility, binding, and target-related assays. Despite the positive results, it is never easy to apply machine learning to the complex problem of drug development. Drug development, in contrast to other areas, has unique challenges in choosing an appropriate representation for the targets in a medication, such as the molecules and their complexes that are important to the drug's intended effect. The lack of bio-activity descriptions is one of the biggest problems. It is crucial to consider how to use the data at hand to accomplish the desired result.

Therefore, determining the correct representation is always the most difficult task. The training data is particularly important for machine learning techniques. This is made much more difficult by the fact that the data used to make the majority of the forecasts is often inconsistent, noisy, and imprecise. It may become much more challenging due to the scarce and uneven data produced by the chemical experiments that were conducted. Recently, computational approaches to deal with these challenges have been created. Drug discovery and development may be sped up in different ways by increasing the use of machine learning to bioactivity data.

Due to the significant time and financial commitment, the process of developing new drugs is exceedingly challenging. Finding a drug to combat a target often takes a huge number of years and billions of dollars. Even then, regardless of a great deal of effort, the success rate is extremely low. There is a risk that many long-term research endeavors may fail, wasting tremendous effort. Bestseller drugs are those that are frequently prescribed for common conditions like the flu, diabetes, high blood pressure, asthma, cold, etc. They are quite successful in the pharmaceutical sector and generate great annual revenues and daily profits. However, if the drug exhibits any side effects, it might also pose problems for the company. Drugs typically face competition from less-priced substitutes when their patents expire. As a result, finding new drugs is a difficult and risky process that is constantly driven by the potential good it could do for millions of individuals suffering from various ailments. The life cycle of drug development is shown in Figure 1

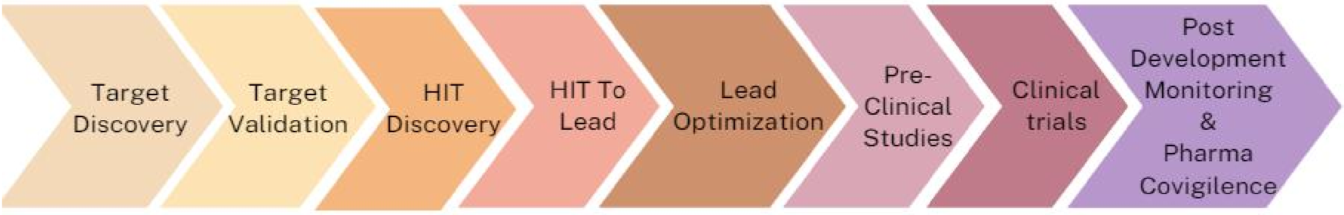

Fig 1. Drug Discovery and Development Life Cycle

1. The first stage is target discovery. We now select the illness target upon which to focus drug development. The target helps us better understand how parasite infection affects genes, proteins, RNA, and other cellular components.

2. Phase two entails verifying the accuracy of the intended target. During this stage, the discovered target is verified to confirm that the drug being developed addresses the right problem.

3. Discovery of HITs is the third phase. In this step, we synthesize and purify the intended target-interacting chemical compounds. Chemists and assay developers work together to test the chosen substances at this step.

4. The fourth stage is the Hit to lead transformation. This phase involves finding prospective lead compounds from the molecules discovered as part of High Throughput Screening (HTS) in the previous stage.

5. The fifth stage is lead optimization. This phase is designed to provide a better and safer scaffold by minimizing structural alteration while eliminating the undesirable effects of the current active analogs.

6. The stage 5 is pre-clinical studies. This comprises identifying medications and comprehending drug mechanisms for reasonable patients, as well as applying biomarkers to increase the effectiveness of clinical trials. It clarifies us on the disease's activity and allows for more precise functional imaging of its response to the drug created to treat it.

7. The following step, clinical trials, involves testing the drug on human subjects. If the medicine achieves its intended results, then the process is complete.

8. Post-development Monitoring and pharmacovigilance are the process's last steps. Medical professionals may clinically prescribe the drug after it has been evaluated and given FDA approval. After that, the drug is put on the market for consumer purchase, and it needs to be monitored continuously.

It takes many years to successfully complete each of these phases. Continuous research is being done to increase the efficiency and speed of this procedure. This paper is aimed at discussing and reviewing the various applications of cognitive sciences and machine learning in order to drive productive benefits for the drug development lifecycle in the pharma industry. The rest of the essay is divided into the following sections: With sections describing the ML models used in each phase and the constraints they impose, Section II sheds insight on the types of ML algorithms employed in different phases of drug development. This section is then followed by a conclusion and a list of references.

II. ML ALGORITHMS USED IN VARIOUS STAGES OF DRUG DEVELOPMENT

Drug development has considerably advanced because of Machine Learning algorithms. Consequently, the use of multiple ML algorithms in drug discovery has significantly benefited pharmaceutical companies. ML algorithms have been used to construct many models for predicting the chemical, biological, and physical characteristics of compounds used in drug discovery. Over the duration of drug discovery, these trained models will become invaluable. Machine learning has been put to use in the pharmaceutical sector for a variety of purposes, such as drug efficacy identification, drug-protein interaction prediction, safety biomarker confirmation, and molecule bioactivity enhancement. Several ML methods have seen extensive usage in the pharmaceutical industry recently. These include the support vector machine (SVM), random forest (RF), and naïve bayesian (NB).

Figure 2 depicts the four main categories of machine learning algorithms: supervised, unsupervised, semi-supervised, and reinforcement learning. [2][3]. Input data must be provided for supervised learning, along with the expected results. During the training phase, it also looks after delivering accuracy rate predictions. Before using the method on new test data, the features, instances, and models to be utilized must be established. Learning can be stopped once performance reaches an acceptable level. The supervised learning framework can be categorized as either classification or regression problems. Any situation where the output is a category falls under the classification problem, for example, YES or NO. A real-valued output for instance height, weight, etc. falls under the category of the regression problem. Unsupervised algorithms, on the other hand, don't need to be trained for the intended result. They model the underlying distribution via an iterative process, giving them the opportunity to understand the data better. These problems are classified as association or clustering problems. We aim to define the rules for understanding the vast data by defining the inherent groupings in the data in clustering and by doing the same in the association. Moreover, semi-supervised learning employs input data with just a subset of labels for training. Many of these issues really occur often in the real world. To solve these issues, researchers use both supervised and unsupervised study methods. In reinforcement learning, the observations derived from environmental interaction are utilised. The reinforcement learning system repeatedly takes up new information from the environment until risk is reduced. To learn the behavior of the environment, it makes use of a feedback signal called a reinforcement signal.

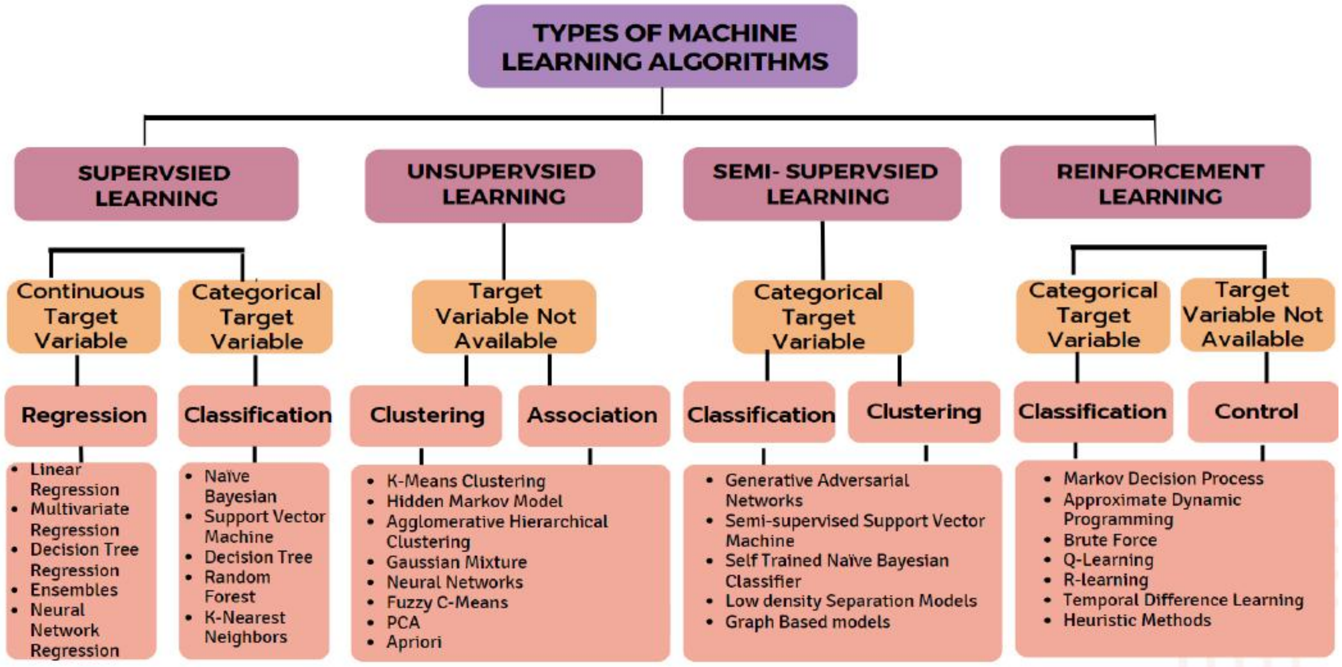
TYPES OF MACHINE LEARNING ALGORITHMS
SUPERVSIED LEARNING
UNSUPERVSIED LEARNING
SEMI- SUPERVSIED LEARNING
REINFORCEMENT LEARNING
Continuous Target Variable
Categorical Target Variable
Target Variable Not Available
Categorical Target Variable
Categorical Target Variable
Target Variable Not Available
Regression
Classification
Clustering
Association
Classification
Clustering
Classification
Control
• Linear Regression • Multivariate Regression • Decision Tree Regression • Ensembles • Neural Network Regression
• Naïve Bayesian • Support Vector Machine • Decision Tree • Random Forest • K-Nearest Neighbors
• K-Means Clustering • Hidden Markov Model • Agglomerative Hierarchical Clustering • Gaussian Mixture • Neural Networks • Fuzzy C-Means • PCA • Apriori
• Generative Adversarial Networks • Semi-supervised Support Vector Machine • Self Trained Naïve Bayesian Classifier • Low density Separation Models • Graph Based models
• Markov Decision Process • Approximate Dynamic Programming • Brute Force • Q-Learning • R-learning • Temporal Difference Learning • Heuristic Methods

Fig 2. Types of Machine Learning Algorithm

There are various machine learning algorithms[4][5], some of the popular approaches are:

• Decision Trees: This model uses data about various decisions and their respective potential outcomes to form a tree-like graph. By eliminating the low-value branches, pruning can help a tree function better. This minimizes both the over-fitting and the tree's complexity.

• Naive Bayes Classification: These Bayes theorem-based classifiers are typically used when the inputs have a large dimensionality. In comparison to other, more complex models, this one has the greatest result.

• K-means Clustering: This technique aids in grouping the data, and K stands for the group number. It uses given features to iteratively assign each data point to a group. The collected data is then clustered in terms of shared characteristics. K-means clustering returns the data's labels and cluster axes' centres.

• Logistic Regression: Methods of statistical analysis used to identify the significance of one or more independent variables in a data collection. It is a way to describe data that is used for prediction, with the goal of learning more about the association between a binary variable and other independent factors.

• Support Vector Machines: In order to maximize the separation between classes, this technique aims to categorize and model the training data into a decision boundary. For cases when linear data separation is not an option, the kernel function is used.

• Neural Networks: These are parameterized non-linear algorithms that classify input data at each layer using a multi-layer perceptron. The accuracy of the model is determined by the perceptron' and hidden layers' numerical values.

2.1. Drug Discovery

A useful classification of the literature review is made possible by the application of ML at every step of drug development, from target identification and validation through hit discovery and hit-to-lead optimization through pre-clinical trials. The drug design methodologies rely on datasets that were created using various ML algorithms. When ML algorithms are properly trained, verified, and used across the drug development phase, they may expedite error-prone, previously difficult procedures and provide insightful findings. The majority of drug design processes now incorporate ML approaches to cut down on time and manual interference, hence leading to optimal results and timelines.

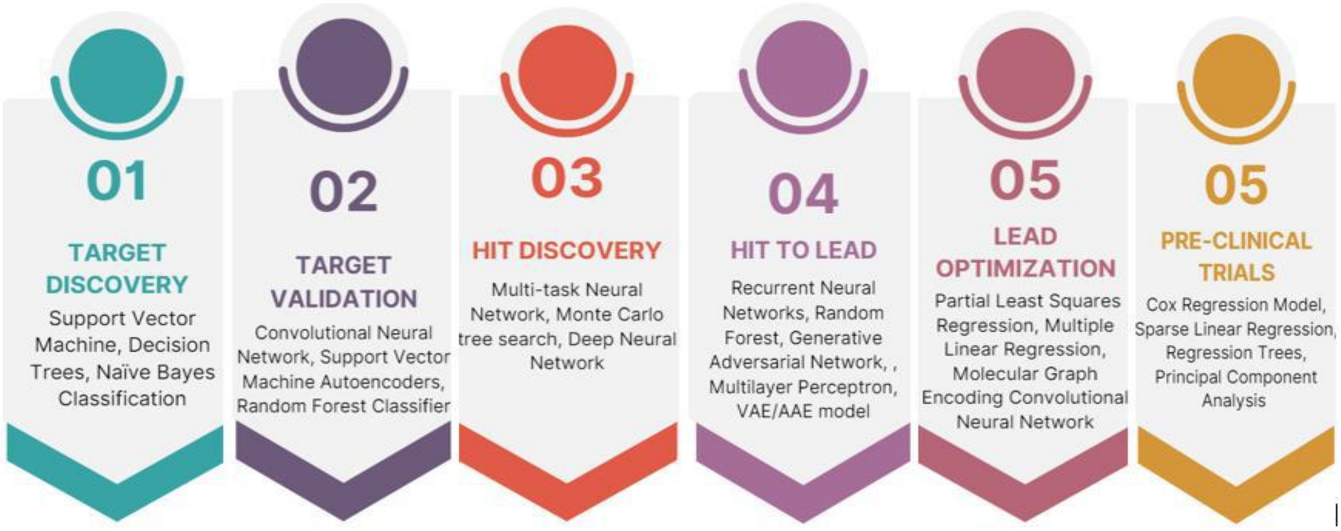
01
TARGET DISCOVERY
Support Vector Machine, Decision Trees, Naïve Bayes Classification
02
TARGET VALIDATION
Convolutional Neural Network, Support Vector Machine Autoencoders, Random Forest Classifier
03
HIT DISCOVERY
Multi-task Neural Network, Monte Carlo tree search, Deep Neural Network
04
HIT TO LEAD
Recurrent Neural Networks, Random Forest, Generative Adversarial Network, , Multilayer Perceptron, VAE/AAE model
05
LEAD OPTIMIZATION
Partial Least Squares Regression, Multiple Linear Regression, Molecular Graph Encoding Convolutional Neural Network
05
PRE-CLINICAL TRIALS
Cox Regression Model, Sparse Linear Regression, Regression Trees, Principal Component Analysis

Fig 3. Machine Learning Models used in various stages of the Drug Discovery Process

2.1.1 Target discovery

The first step in the target identification and characterization procedure is to determine the function and significance of a gene or protein that may be used as a therapeutic target. After a target has been identified, the molecular pathways, it is expected to effect may be described. Effectiveness, safety, and compliance with clinical and business needs are some of the qualities of a good target.

Producing drugs (small molecules, peptides, antibodies, or more advanced techniques like short RNAs or cell therapies) that will change the disease state by modifying the activity of a biological target is the main objective of drug discovery.[6]. The selection of a target with a valid therapeutic hypothesis, that is, that modifying the target would modify the disease state, is still important before beginning a drug development program, despite the recent revival of phenotypic screens. Target identification and prioritization is the process of choosing this target based on the information available. The next stage is to demonstrate the selected target's involvement in the illness using ex vivo and in vivo models that are physiologically relevant (target validation). Early target validation is essential to concentrate on high-probability projects even if clinical trials will eventually confirm the target.

The pharmaceutical sciences devote considerable attention to the study of drug-target interactions (DTI) [7]. The procedure of discovering new medicines is costly and time-consuming. Therefore, the ability to anticipate drug-target interactions is useful to biologists since it allows them to focus their research. The first and most important step in the drug development process is determining the desired effects of the medicine. Medicable proteins that play a role in illness make up the majority of these areas of intervention. Drug-target interaction prediction is used to find novel treatment approaches. Potential targets include proteins with enzymes, ion channels, G protein-coupled receptors (GPCRs), and nuclear receptors. Certain ligands may alter the functioning of these groups. As a result, studying the genomic space generated by these protein classes enables us to precisely predict the likelihood of an interaction. Drug discovery and drug repositioning for novel targets are both possible using DTI. The three main categories of DTI prediction tools are ligand-based, docking-based, and chemogenomic strategies. Similarity between ligands and target proteins is used as a predictor of DTI in the Ligand-based Approach [8], [9]. The target protein's three-dimensional structure may be used to identify the probability of a pharmacological interaction. The Docking-based approach is used for this, which considers the relative stability and binding affinity of the proteins [10, 11]. If the drug's chemical information, protein genomic data, and known DTIs are all considered, the chemogenomic technique is then used.[12][13]. In a ligand-based technique, a target with a small number of binding ligands frequently yields subpar DTI predictions. This is a shortcoming of this approach. Similarly, the docking-based approach is time-consuming and depends on the target proteins' 3D structures. Due to these drawbacks, the chemogenomic technique has recently gained popularity for the identification of DTI. The DTI problem is presented as a machine learning problem using this method, and a classifier is often created and trained using publicly accessible interaction data. In order to forecast the unknown interactions, this classifier is used[13]. The chemogenomic fully utilized a number of techniques. Bipartite graphs[14], recommendation systems[15], and supervised classification problems are some of these[16]. However, when we look at the data, we can see that there will only be a small number of favorable interactions, and the other possible interactions are unknown. For instance, there could only be 7000 favorable drug interactions out of the 35 million potential drug options[17]. The two types of computational chemogenomic methodologies are feature-based and similarity-based strategies. Features are the inputs for a set of instances defined by a specific class label for feature-based methods. In most cases, the targets are the features and the instances are the drugs. The presence of a possible associations is represented by the binary value of the class label. Support vector machines, decision trees, and random forests are a few examples of feature-based classification techniques.[18]. Drug-target interactions are often identified using Support Vector Machine or Random Forest. [19].

Particularly, the target identification portion of the drug development process significantly relies on the categorization of biomedical data. The classification of biomedical data, which is sometimes replete with irrelevant information and data known as noise, has shown excellent potential when using Naïve Bayes classifier (NB) algorithms [20]. Lead discovery might be considerably enhanced by applying NB approaches to predict ligand-target interactions.[21]. In recent years, researchers have been able to use NB strategies in many areas of the drug development process. In a research aimed at finding new breast cancer therapies, Pang et al. [22] employed NB models and other methods to categorize compounds according on their potential efficacy as estrogen receptor antagonists. "The model produced impressive results when used in conjunction with other techniques, such as the extended-connectivity fingerprint-6. In a study by Wei et al. [23], potential drugs that would be effective against the targets of the hepatitis C virus and human immunodeficiency virus type 1 were predicted using a mix of NB and support vector machine methods. Their approach included two distinct descriptor systems, one of which was the extended-connectivity fingerprint-6, with NB as a classifier technique. Utilizing NB in conjunction with other approaches and technologies has proven effective in implementing drug discovery processes.

2.1.2 Target validation

The concept of creating a medicine for a certain target is also an important consideration for the pharmaceutical companies. For instance, identifying targets with properties that imply that these proteins can bind tiny molecules is necessary for small-molecule drugs [24]. These druggable models can be created using various target attributes. Using the physicochemical, structural, and geometric characteristics of 1,187 drug-binding and 99 non-drug-binding cavities in a sample of 99 proteins, Nayal and Honig [25] created a random forest classifier. The most important characteristics were the size and shape of the surface voids. On the basis of the protein sequences of well-known drug and non-drug targets, some studies have used SVMs [26][27] or biassed SVMs with stacked autoencoders, a Deep Learning model [28] to forecast druggable targets to assess different physicochemical properties. Additionally, it has been discovered that druggable proteins tend to be strongly linked and occupy certain areas of protein-protein interaction networks[29][30][31]. These ML methods reduced the search area by generating lists of drug-binding targets, but further research is needed to confirm these forecasts.

The holy grail of target identification or validation, namely the ability to accurately anticipate the outcome of a drug's clinical trials in advance, has not yet been attained. Success indicators have been the subject of several non-ML studies. Rouillard et al.[32] evaluated 332 targets that were either successful or unsuccessful in phase III clinical trials by analyzing their omics data using ML and selecting multivariate characteristics. The gene expression data, which was characterized by low mean RNA expression and considerable heterogeneity across tissues, was shown to be a strong predictor of effective targeting. This work provided more evidence that optimal targets are expressed selectively in diseased tissues [33]. In order to anticipate de novo therapeutic targets, Ferrero et al. [34][35] trained a variety of ML classifiers utilizing target-disease linkages from the open target's platform. It was determined that regardless of indication, the three most essential data categories for therapeutic target prediction are gene expression, genetic data, and the availability of an animal model. . The sparseness of the data and the lack of knowledge regarding the causes of failed programs, however, pose a limitation to this technique. Fundamentally, because it takes years to build a good drug discovery plan and finally bring it to market, successful programs are a reflection of earlier drug development models. Considering the arrival of more recent therapeutic modalities like biologics (including antibodies), it's doubtful that the elements that contributed to the success of small-molecule projects in the past will be relevant in the present. Additional restrictions are imposed by precision medicine's growing importance. For future prediction tools to be effective, large amounts of information on both successful and failed drug development projects must be made accessible with metadata in the public domain.

2.1.3 Hit Discovery

It is essential to execute comprehensive virtual and experimental high-throughput screening of large chemical pools to identify treatment candidates that inhibit or activate the target protein of interest. The pharmacodynamic, pharmacokinetic, and toxicological characteristics of candidate structures are further improved, as well as their target specificity and selectivity. It is important to note, however, that there may be a lack of enough high-quality data in this domain, which might limit the use of ML to new chemistry. This is especially true for macrocycles and proteolysis-targeting chimaeras (PROTACs).

For ligand-based virtual screening, a lot of attention has been paid to the use of Deep Learning models, such as multi-task neural networks. Computational methods may use a particular lead molecule to identify physically comparable compounds with similar chemical characteristics. The use of multi-task DNNs has shown to be more effective than standard statistical approaches for this job, which was previously performed. When it comes to predicting the properties and functionalities of small substances, DNNs may greatly increase predictive power [37]. One-shot learning may significantly reduce the amount of information needed to accurately forecast how a molecule would read out in a new experimental environment. The binding mechanism of opiates to the -opioid receptor was previously unknown, however a Markov State model and Machine Learning approaches were able to pinpoint an allosteric region implicated in this activation [38]. The advantages of multi-task models over single-task models, however, vary depending on the data set. The evaluation of ML algorithms has made use of MoleculeNet [39], a large benchmarking data set produced by Pande et al. to assist in the comparison of different ML algorithms. Data on the characteristics of more than 700,000 molecules can be found in MoleculeNet. The open-source DeepChem package now includes all of these hand-picked data sets along with a number of additional number of advantages.

Planning effective chemical synthesis routes can also be done using DNNs and contemporary tree search techniques. A target molecule is formally deconstructed using reversed processes in order to plan its production (retrosynthesis). In order to synthesise the target, this method creates a sequence of processes that may be carried out in a straightforward way in the laboratory. The systematic application of synthetic chemistry skills to this method is a tremendous task. The exponential growth of chemical knowledge and the inadequate understanding of the range and boundaries of many reactions have made the manual insertion of transformation rules impracticable. A database called Reaxys (with 11 million reactions and 300,000 rules) was utilised by Segler et al.[40] to automatically extract the rules. He used a Monte Carlo tree search (MCTS) to weight the tree's nodes and DNNs in order to determine the most profitable paths for further research. This strategy outperforms the industry norm for best first search in quantitative analyses using two different implementations (heuristic method and neural). Furthermore, for around two-thirds of the examined chemicals, MCTS is 30 times quicker than conventional computer-aided search methods. In a double-blind experiment, qualitative evaluations were also incorporated. Organic chemists were required to pick between expected and literature-based synthesis paths in a blind process. For the first time, chemists agreed that the predicted routes' quality was, on average, on par with routes selected from the literature.

2.1.4 Hit to Lead

This procedure is sometimes referred to as "lead generation" in the early phases of drug development research. Insufficient optimization during the High Throughput Screen (HTS) to find potential lead compounds leads to the discovery of molecules, or "hits." Using a preexisting kinase inhibitor library, the "design layer"/Random Forest regression mapping method is used to construct new chemical spaces with biological activity. This method of optimising hits into leads is a practical use of chemical synthesis. [41]

By adjusting or rebalancing the target interest, de novo drug design produced distinct chemical structures [42]. By starting from scratch, de novo techniques introduce new molecules using a fragment-based methodology. If the molecular structure has impracticalities and complexity at this stage [43], the risk emerges in the structure's development and the evaluation of bioactivity becomes challenging. In order to develop a novel structure with the necessary properties, deep learning models could be used in terms of their extensive knowledge and generative skills [44]. The use of reinforcement learning in molecular de novo design is another significant application of Machine Learning. By modifying a sequence-based generative model to produce molecules with almost ideal values for solubility, pharmacokinetic characteristics, bioactivity, and other factors, researchers at AstraZeneca were able to expand the chemical space. Similar models were created by Kadurin et al. utilizing deep GANs to extract chemical features from very huge data sets [45]. It's important to keep in mind, however, that reinforcement learning may not be helpful when trying to find previously undiscovered synthetic pathways.

Olivecrona et al’s[46] expansion of the use of deep reinforcement learning to the prediction of biological activities in the creation of new drugs included some RNN model modifications. To understand the SMILES syntax, an RNN model must be trained; chemBL compounds may be gathered for training. Agents take engage in activities in reinforcement learning under certain rules. If the agent is rewarded enough, the trend of their actions will be revived. [47]. Use the SVM methodology to improve a few methods based on the ligands concept in the training set in order to achieve a high benefit for activity scoring. Create a few compounds that are antagonistic to the dopamine receptor 2-type before using the RNN and deep reinforcement learning model. Additionally, it was noted that with SVM's scoring capability, predictions for structures in the bioactive region have exceeded 95%. The auto-encoders method can be used to produce unique molecules by employing deep learning algorithms. Then, Gomez-Bombarelli et al. [48] used the multilayer perceptron (MLP) and variational autoencoder (VAE) to automatically produce new molecules with the required characteristics.

Kadurin et al. [49] used on the AAE model, now known as druGAN, to create molecular fingerprints. The AAE approach produced impressive results when applied to the VAE model in terms of power production, reconstruction inaccuracy, and subsequent extraction effectiveness. Coley et al. [50] proposed analysing the synthetic molecule to determine whether it was accessible synthetically. As a result of the great approximation capabilities for producing synthetic complexity measures, he postulated that the neural network was trained in line with the response database. The product complexity score must be higher than the reactant complexity score for a synthetic reaction to be successful. [51]. In order to demonstrate correlation inequality between the complexity of the products and reactants, Coley made several efforts to develop a scoring function by encoding chemical responses into pairs of products and reactants. In order for neural networks to feel at ease with any kind of scoring capability at that time, they must be trained using the reactant and product pairings that Coley utilised across a scope of 22 million. Additionally, the synthesis process's conclusion was established with a great deal of complexity. Finally, generative models reveal both the complexity of the synthetic process caused by eliminating the non-realistic molecules and the pharmacological actions in inverse synthetic planning.

How to adequately explain the chemical structure is a challenge in small molecule design. There are several representations to select from, ranging from simple circular fingerprints like the extended-connectivity fingerprint (ECFP) to intricate symmetry functions [52]. Which structural representation is best for every small-molecule design challenge is yet unclear. It would be fascinating to see whether the extensive body of ML research in cheminformatics provides any new information on the most efficient approach for structural representation.

2.1.5 Lead Optimization

The optimization of potential drug leads is an important step in the drug development process. If a fragment has the potential to be used in medicinal chemistry, it will be evaluated as a potential future step in the research process. Lead optimization aims to offer a better and safer scaffold by reducing structural modification and removing negative effects of existing active analogues. An illustration of this is the advancement of Autotoxin inhibitors, such as the investigational drug GLPG1690, in human clinical trials to treat pulmonary fibrosis. Figure 4 below provides an overview of the factors that can make active analogs more potent by using customized methods. Here, we evaluate a substance's ADME/T features, including its toxicity, chemical makeup, physical attributes, and rates of absorption, distribution, metabolism, and excretion.

Chemical and physical properties
Lead Optimization
ADME Features
Toxicity

Fig 4. Factors Affecting Lead Optimization

- *Chemical and physical properties*

Physical and chemical properties have been used in the drug development process to lessen the number of significant failure. To this end, researchers have turned to lead optimization strategies powered by deep learning models [53]. Due to their dependence on the interpretability principle, Duvenaud et al. [54] directly gathered data from the molecular graph using the CNN-ANN idea to generate a prediction, i.e. (MAE = 0.53+0.07). Duvenaud's study was motivated by Coley et al efforts' to improve molecular aqueous approaches. Additionally, the tensor-based convolutional approach was used, and the improved results were MAE (0.424+0.005).

Clearly defining molecular graph attribution is crucial since tensor-based approaches must incorporate properties like bonds and atom levels. To predict molecular aqueous solutions, Coley's model utilised a lot more atom-level information than Duvenaud's [55]. It was shown that Caco-2 permeability coefficients had a good correlation with oral drug absorption (P app) for predicting the candidate drug while pharmacokinetic parameters were being evaluated.[56, 57]. Using the Caco-2 permeability data, Wang et al's [58] attempt to generate 30 descriptors' worth of prediction templates necessitated the building of 1,272 components, including models like SVM regression and boosting. In the test set, where it also had the maximum expectation capacity, the boosting model fared the best. It conforms with the OECD's (Organization for Economic Co-operation and Development) standards for promoting reliability and logical arguments since it adheres to the QSAR principles set out by the OECD.

- Absorption, distribution, metabolism, and excretion

Injecting pharmaceuticals or treatments into a person's veins is a method of absorption. Bioavailability parameter is used to examine the level of absorptions. Several clinical departments explained how to increase absorption properties using molecular predictions for bioavailability [59]. Tian et al. used 1,014 compounds to predict bioavailability using molecular resources and structural fingerprints using the MLR model. The predicted performance of applying the genetic function approach was excellent, with RMSE = 0.2355 and a correlation value of 0.71. The distribution of medications or treatments within the human body is influenced by intracellular and interstitial fluids as well as specific drug absorption characteristics.[60]. The steady state distribution of a drug is the amount of drug that makes it from the in vivo phase to the plasma reaction (VDss). The steady phase is a crucial indicator for evaluating the drug distribution mechanism. Lombardo and Jing used 1,096 molecules and the PLS and Random Forest methodologies to make predictions about VDss. [61]. The board members in this case are dissatisfied with the prediction findings since 50% of the compounds are accessible in a twofold mistake. Because of the presence of such obscure components, VDss may be affected. The use of VDss in molecular structure data is intentionally put to the test by this issue. Any drug or treatment taken by a person under these circumstances will try to produce the already-existing toxic metabolite as a result of the metabolic system's inbuilt redundancies. It is important to maintain the integrity of the metabolic structure, hence structural optimization methodologies are utilised to motivate the metabolism to make very accurate prediction. Numerous machine learning (ML) methods were used to predict specific metabolic enzymes, such as UDP-glucuronosyltransferases (UGTs), cytochrome P450s, etc., using a vast quantity of drug metabolism data. In addition, Xenosite's platform has UGT-trained neural networks for predicting UGT metabolism [62][63][64]. When a drug is digested, it is eliminated from the body in a process known as excretion. Because certain medications are soluble in water, water may be used to flush them out of the body, or in the absence of metabolism, the metabolites can be eliminated directly. The PCA method was utilised by

Lombardo et al. to get excellent results in innovative approaches, with a predicted rate of accuracy of 84%.[65].

- Toxicity and the ADME/T multi-task neural networks

In clinical and preclinical damage completion, about one-third of the most important compounds utilised in drug localization were shown to be inadequate. Risks were reduced by improved toxicity prediction and molecular optimization [66]. Kidney and liver toxicity profiles are among those that may be predicted using tools like structural warnings and rule-based expert knowledge. To improve the accuracy of toxicity predictions, deep learning models are required. Similarly, Xu et al., anticipating results from CNN molecular graph encoding, created the acute-oral toxicity prediction model (MGE-CNN). When compared to the SVM model, predicted results were shown to be better [67]. The similarity in training neural networks feature extraction, model construction, and molecule encoding resulted in the success of the MGE-CNN model. Due to the adaptability of the MGE-CNN model, the issue was reformulated in terms of molecular fingerprints. To categorise TOX Alerts and collect high-quality data on structural alerts, Xu et al. [68] employed hazardous characteristics for fingerprints. When comparing parameters, multi-task neural networks that have been trained to retrieve comparable characteristics outperform single-task neural networks. [69]. This is because the neural networks are more supportive of multiple tasks and share parameters. The human body receives data after the drug's absorption, distribution, metabolism, and excretion have all been taken care of and prediction has been enhanced using multi-tasking neural networks. In this study, Kearnes et al. examined single-task and multi-task performance using ADME/T experimental data. The results demonstrated that the multi-task approach was superior. [70].

2.1.6 Pre-Clinical Studies

Through the use of ML models, biomarker discovery increases the effectiveness of clinical trials by identifying drugs and understanding how they work for reasonable people [71][72][73]. The completion of a clinical trial requires a lot of money and time. Throughout the first stages of clinical trials, expected models must be used, developed, and validated in order to solve this problem. In preclinical data collection, ML systems enable the prediction of translational biomarkers. Following data validation, corresponding biomarkers and models may examine the patient's symptoms and provide a treatment strategy. Although many scholarly articles proposed predictive models and biomarkers, only a few of those articles were actually implemented in clinical trials. For a clinical situation, it is required to consider model development, design, data access, data quality, software, and model selection. The main problem was how ML methods evaluated the effectiveness of community-driven efforts to create regression and classification models. The US Food and Drug Administration-led (MAQC II) MicroArray Quality Control [74] analyzed ML algorithms for predicting gene expression data in the last step of clinical trials a number of years ago. 36 independent organizations that examined 6 microarray data sets created predictive models for categorizing clinical locations nearing completion of development. Information simulates the best methods for clinical trials by including high-quality data, trained scientists, and control systems. Patients with multiple myeloma had poor prognosis and their treatment was stopped after 24 months owing to an incomplete application. Multiple myeloma and gene expression are continuous variables, hence their future behavior may be predicted using a regression-based method. A gene expression profile may be utilized in combination with Cox regression models to identify a patient's illness risk factors, as has been shown [75]. Here, the advantage of using regression models is emphasized due to the lack of specified classes that might perform prediction in clinical trials. [76][77][78][79]. The National Cancer Institute (NCI) finds it challenging to develop medicine prediction models in order to assess regression models. [80]. The best model with key parameters must be used to acquire training data (for instance, treating 35 breast cancer cells with 31 medicines), and models must be validated using identical blind testing data (i.e., treating 18 breast tumour cells with similar 31 drugs). Using data from six different data profiles—RNA sequencing, RNA microarray, reverse protein phase array, SNP (Single Nucleotide Polymorphism) array, DNA methylation status, and exome sequencing—better prediction algorithms are created". These profiles are used to conduct multivariate statistical analyses on 44 sets of data using a variety of regression models, including sparse linear regression, kernel methods, regression trees, and principal component analysis. The MAQC II findings showed that certain groups performed very well, while other groups utilized similar models. While some teams concentrated on technical issues like feature selection, quality control, data reduction, modifying ML parameters, and splitting strategy, others utilized biological information like gene expression data to set themselves apart from the competition. A huge number of medications are feasible for creating a prediction model when compared to other approaches.

2.2. Clinical Trials

Proper Clinical drug development follows the completion of preclinical research and includes investigations with human volunteers and clinical trials to further perfect the drug for human consumption. The intricacy of clinical trial design, the cost of conducting such a study, and the difficulties inherent in putting it into practise are all factors that may affect trials performed at this level. Trials must be safe and effective, done within the budget given for drug development, and adhere to a certain methodology to guarantee the medicine is useful and practical for its intended application. For this rigorous process to be successful, it needs to be properly set up and involve a significant volunteer base. In order to successfully carry out these various tasks involved in clinical trials as shown in Figure 5; ML algorithms have been extensively used in each sphere thereby aiding the process as a whole.

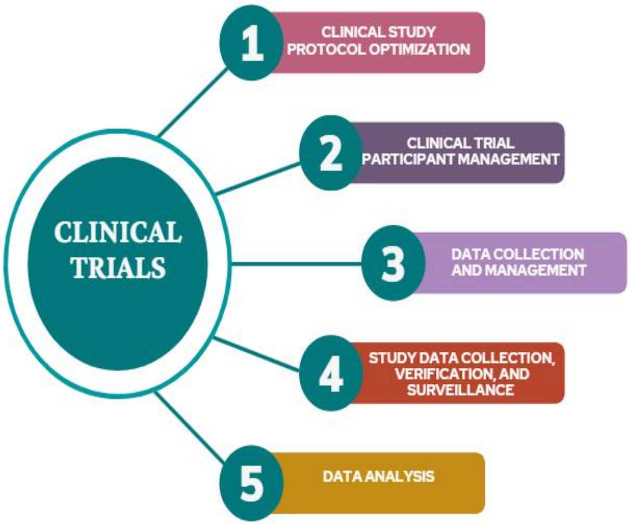
CLINICAL TRIALS
1
CLINICAL STUDY PROTOCOL OPTIMIZATION
2
CLINICAL TRIAL PARTICIPANT MANAGEMENT
3
DATA COLLECTION AND MANAGEMENT
4
STUDY DATA COLLECTION, VERIFICATION, AND SURVEILLANCE
5
DATA ANALYSIS

Fig. 5 Various Tasks associated with Clinical Trials

2.2.1 Clinical study protocol optimization

The success and effectiveness of human clinical trials may be improved by using ML to ease the formulation of trial protocols in advance using simulation techniques on a significant quantity of data from prior studies. As shown in reinforcement learning approaches for Alzheimer's disease and non-small cell lung cancer [81, 82], study simulation, for instance, may optimize the selection of treatment regimens for trials. Researchers may submit protocols using AI, which employs natural language processing, in order to detect potential roadblocks and obstacles to successfully completing trials (such as inclusion/exclusion criteria or outcome indicators). Although the use of ML in research planning may theoretically ensure that a particular trial design is best suited to the needs of the stakeholders, this is just a promise since the effectiveness of these sample models has not been assessed in a peer-reviewed manner. In conclusion, machine learning clearly has the potential to improve the effectiveness and productivity of preclinical research and the planning of clinical trials. However, rather than focusing on the planning of clinical trials, the great majority of peer-reviewed studies on the use of ML in this context are preclinical research and development-focused. This could be because there are more large, high-dimensional datasets available in translational contexts or because using ML in clinical trial settings comes with higher costs, hazards, and regulatory requirements. We require scholarly research on the effectiveness of ML in clinical trial design to solve these challenges.

2.2.2 Clinical trial participant management

Clinical trial participant management involves selecting research populations, enlisting patients, and maintaining their participation. Despite significant investment in participant management, studies often run over budget, take longer than expected, or fail to provide useful data due to patient drop-out and non-adherence. There is a total failure rate of 13.8 percent for medications evaluated in phase I, with estimates indicating that between 33.6 percent and 52.4 percent of clinical studies underpinning drug development that take place during stages 1-3 are unsuccessful.. [83]. ML techniques can help with participant identification recruiting and retention and choosing the study demographics of patients. If individuals were more carefully selected for trials, the sample size required to detect an impact may be less. Alternatively stated, improved techniques of selecting the patient population may lead to fewer individuals being offered therapies for which they are not likely to improve outcomes. Previous studies have indicated that for every expected response, there are anywhere from three to twenty-four non-responders for the most commonly prescribed drugs, making progress in this field a continual challenge. Many people who use these drugs end up having unintended consequences [84]. Unsupervised machine learning of patient populations may quickly analyze large databases of existing research and in turn aid with patient population selection as well as reveal patterns in patient features that may be utilized to choose patient phenotypes that are best suited for treatment [85]. A cross-modal inference learning model technique may more successfully match patients to trials using EHR data by concurrently encoding enrollment criteria (text) and patient records (tabular data) into a shared latent space [86]. The utility of these procedures is questioned by the lack of peer-reviewed documentation of their development and performance measures [87]. Mendel.AI and Deep6AI are two businesses that provide comparable services. This method may have the benefit of not requiring participants to be identified precisely by structured data fields, which has been demonstrated to dramatically skew trial populations. [88, 89]. There are two basic strategies to boost retention and policy adherence using ML models, as shown by the monitoring of participants and their adherence to protocols. The first stage is to use ML to identify, investigate, and penalise participants who are likely to infringe upon the terms of the research. The second approach is to use ML to make the research easier for participants and to improve their overall experiences. AiCure is a company that employs face recognition technology to track whether or not patients really take their prescription. AiCure was shown to be more efficient than directly observed modified therapy in detecting and improving patient adherence in studies on schizophrenia patients and recent stroke survivors using anticoagulation [90, 91]. AiCure's performance may differ across patient subgroups since its model building and validation technique is not publicly known, as has been shown in previous computer vision applications. [92]. Additionally, data obtained during routine clinical care might be analyzed using ML techniques to provide data that can be utilized for study. For instance, rather of exposing all participants to the additional strain and cost of more in-depth and multiplexed imaging, generative adversarial network modeling of typically clinically stained slides with hematoxylin and eosin may identify the ones who need it. [93]. Natural Language Processing may also make it simpler to repurpose clinical data for research purposes by automatically filling out study case report forms when used often with the Unified Medical Language System [94]. There are two examples of how patients produce useful content outside of the clinical trial context that ML can process into study data to lessen the burden of data collection for trial participants: natural language processing of social media posts to identify serious drug reactions with high fidelity [95]. The International Parkinson and Movement Disorders Society's Unified Parkinson's Disease Rating Scale has been found to correlate participant activity with wearable device data, which can also be used to distinguish between neuropsychiatric symptom ontology patterns, identify patient falls, and identify participant activity [96]. In summary, ML and NLP have shown promise for a number of tasks related to improved participant management in clinical trials; nevertheless, additional research comparing various approaches to participant management is required to further improve clinical trial quality and participant experience.

2.2.3 Data collection and management

Applying ML to clinical trials has the potential to enhance the methods used to gather, handle, and analyse trial data. ML techniques can also aid in addressing some of the challenges related to collecting real-world data and dealing with corresponding missing data. Wearable and other mobile/electronic device data on patients' health may supplement or even replace data collected via more conventional means, such as in-person visits for a research. The usage and validation of new, patient-centered biomarkers may be made possible by wearables and other devices. When creating new "digital biomarkers" from the data acquired by the device's numerous sensors, ML processing is often necessary since the data provided by mobile devices might be sparse and inconsistent in quality, accessibility, and synchronisation (such as cameras, audio recorders, accelerometers, and photoplethysmography).Therefore, in order to analyse the massive and complicated data created by wearables and other devices, appropriate data collecting, storage, validation, and analysis procedures are required [97]. Patients with atopic dermatitis had their accelerometer data processed using a recurrent neural network [98], a mobile single-lead electrocardiogram platform's input was processed using a deep neural network, and an audio signal from a Parkinson's disease patient was processed using a random forest model. [99]. These cutting-edge digital biomarkers might make clinical studies run more smoothly and with a focus on patients, but there are risks associated with this strategy. Although this risk exists for all data, regardless of processing technique, using machine learning to evaluate wearable sensor output to define research goals involves the possibility of producing false results, as was shown to happen with an electrocardiogram classification model[100]. Lack of awareness of participant privacy attitudes in relation to the sharing and use of device data, as well as a lack of a precise description of the overlap between authorised clinical aims and patient-centric digital biomarkers, are obstacles to ML processing of device data implementation.

2.2.4 Study data collection, verification, and surveillance

An intriguing use of ML is in automating data collecting into case report forms, which may save time, money, and human error in either prospective trials or retrospective evaluations. Specifically, Natural Language Processing is very important for this kind of data administration. Depression [101], epilepsy [102], and cancer [103] are just a few examples of diseases where this application has showed early promise despite having to overcome varied data formats and provenances. Regardless of the method used for data collection, ML might support risk-based monitoring techniques for clinical trial surveillance. This allows for the avoidance or early detection of issues like site failure, fraud, and inconsistent or nonsensical data that might otherwise delay database lock and subsequent analysis. For instance, when people fill out case report forms, the accuracy of the information acquired for result determination may be evaluated by combining optical character recognition with natural language processing (usually supplied in PDF form). Clinical trials and observational studies may benefit from auto-encoders since they can be used to identify potentially fraudulent data patterns by classifying them as plausible or improbable [104]. Endpoint detection, adjudication, and safety signal detection are all examples of how machine learning may be used in data processing. Currently, events are manually adjudicated by a committee of doctors. However, there may be time, money, and complexity savings with semi-automated endpoint identification and adjudication. While categorising events into useful categories has typically been the domain of semi-automated ML systems, adjudicating endpoints has historically required a significant amount of human labor. Although this technique has not been peer-reviewed, IQVIA Inc. has described the capacity to automatically treat certain adverse events connected to pharmacological therapy utilizing a mix of optical character recognition and natural language processing [105]. A classification model would theoretically need to be retrained for each new experiment due to the fact that endpoint criteria and the data needed to support them often alter across research. This might be a roadblock in the way of fully automated event adjudication (which is not a viable approach). Although not all studies adhere to these objectives, there have been recent attempts to standardize outcomes in the area of cardiovascular research. The majority of areas have not combined trial data to enable model training for cardiovascular endpoints [106]. For this area to go further, stakeholders must establish consensus definitions, really accept the definitions of events, and be prepared to provide the right data from several trials for model training.

The issue of missing data may be solved using different ML applications. This may be accomplished by thinking about the data's context, the assumptions and objectives made about the data, the methods used to acquire the data, and the analyses that will be conducted. Goals could include computing other important quantities by averaging over a large number of potential values from a learning distribution or directly calculating precise estimates of the missing covariate values. Though more modern approaches are still in their infancy and thorough comparisons are needed, preliminary studies show that complex ML methods may not always be superior than simple imputation strategies like the population mean estimate. [107]. One use of missing value algorithms is the analysis of sparse datasets like those found in registries, electronic health records, ergonomic studies, and data collected from wearable devices. [108][109]. Data augmentation solutions may mitigate the effects of missing data or values, but they should be used with caution lest they lead to models that are only partly generalizable to newly collected data that has inherent flaws. Therefore, using ML to enhance data gathering while conducting research itself might be a more fruitful approach.

2.2.5 Data analysis

Rich sources of information for study design, risk modeling, and counterfactual simulation include data collected in clinical trials, registries, and clinical practices. These projects are ideally suited for machine learning. Unsupervised learning, for instance, might find phenotypic clusters in real-world data that can be explored further in clinical research [110]. Additionally, ML has the potential to advance the established practice of secondary trial analysis by more accurately identifying treatment heterogeneity while still providing some (albeit insufficient) protection against false-positive findings, thereby revealing more intriguing areas for further research [111]. Additionally, machine learning may provide risk predictions that may be evaluated in the future with the proper utilization of previous data. For instance, a random forest model in the COMPANION trial data performed better at identifying individuals who might benefit from cardiac resynchronization treatment than a multiple logistic regression [112]. The results demonstrated that random forests may explain feature interactions that are often missed by simpler models.

ML shows considerable promise in this area by increasing the precision with which it can distinguish real-world evidence from real-world data, even if it is still a highly desired (and extremely difficult) objective (i.e., draw causal inferences). A vital and important endeavor is the creation of predictive models that can predict future occurrences. A few of the methods suggested in the literature include optimal discriminant analysis, targeted maximum likelihood estimate, and propensity score weighting made possible by ML [113][114].

The use of ML to provide counterfactual policy estimates, where existing data is used to anticipate outcomes under circumstances that do not now exist or may not, is particularly fascinating. For instance, reinforcement learning suggests better treatment plans based on prior unsuccessful treatments and outcomes, and trees of predictors may provide survival predictions for heart failure patients under the conditions of obtaining or not receiving a heart transplant. [115]. Risky data sharing agreements that restrict the amount of data accessible for model training and a lack of compliance with EHR data systems are the key obstacles to adoption. [116]. In conclusion, there are many efficient ML algorithms for managing, processing, and analyzing data from clinical trials, but there are much less methods for enhancing data quality from the start. High-quality trials must be conducted in order to enable more advanced ML processing since the availability and quality of data are the foundations of ML techniques.

2.3 Post Drug Development Sector and Pharmacovigilance

Once the results of clinical studies have been compiled and the treatment has been developed to achieve maximal effectiveness and safety, the FDA will move it forward for comprehensive assessment. Currently, the FDA examines the drug application that the pharmaceutical company has submitted and decides whether to approve it or not. Once the pharmaceutical company has received permission, it can start selling drugs and continue to manage its products.

There is a completely different sector or area of technology, processes, and advanced improvements that open up once the drug hits the market and is ready for use. Figure 6 below shows some of the practical illustrations of how companies can and have applied AI and ML technologies to the post-drug development and pharmacovigilance arena.

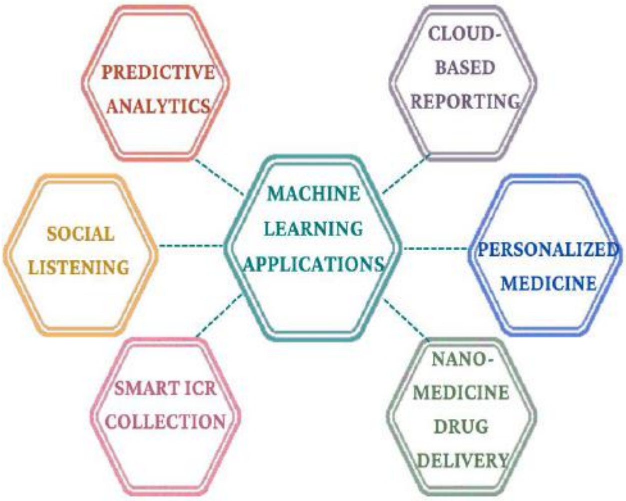

Fig 6. Machine Learning Applications in Pharmacovigilance

- Predictive Analytics

ML aids in making predictions based on that analysis while AI aids in managing enormous amounts of data. The time it takes for new drugs to reach the market has been cut in half with the help of AI and ML. Typically, the lifecycle of a drug design lasts 10–15 years[117]. Artificial intelligence and machine learning allow specialists to use statistical models to learn from the past, present, and future, speeding up the process of discovering and testing new treatments. SciBite makes the most of the predictive analytics that AI and ML offer [118]. The company reduced the amount of time it took for new pharmaceuticals to hit the market by integrating AI into its R&D methodology. According to New York University, 80 percent of clinical data is unstructured [119]. To speed up operations in the post-drug development sector, AI and ML are the tools that can operate with such a vast information segment.

- Social Listening for Accurate Health and Drug-Related Information

Social media sites may provide a wealth of vital information if the correct tools are available. The article that was published following the Pacific Symposium on Biocomputing demonstrates how AI can provide important insights into the effectiveness of antidepressants by analyzing five million posts[120]. The study also emphasizes the value of social listening in identifying drug safety combinations and adverse drug reactions (ADRs). In fact, researchers were able to identify some new side effects of prescription medications. They used artificial intelligence (AI) to examine public posts from users and learned various patterns.

- Smarter Individual Case Safety Report (ICSR) Collection

ICSR report collection constitutes a significant problem. An even bigger challenge is their analysis. Over 20 million ICSRs are said to be stored in the WHO database, which might prove to be an invaluable resource for studying drug side effects and other potential dangers. According to a study published in the journal Clinical Pharmacology & Therapeutics[121], the ICSR collection process is made smarter overall with the use of AI and ML. The experts forecast that ICSR reporting would be significantly more advanced than it is presently by 2030. Massive amounts of unstructured text in ICSRs can be analyzed using AI-based technologies like Natural Language Processing (NLP), resulting in ICSR management that is enhanced by AI.

- Cloud-Based Reporting

AI pharmacovigilance and cloud computing go hand in hand. The experts believe that cloud technology will be used to gather and analyze data. It is anticipated that the cost-efficiency, scalability, and simplicity of Pharmacovigilance will increase with the integration of cloud-based computing with AI and ML.

- Personalized Medicine

In order to create individually customized treatments, personalized drugs can be done by identifying a person's biological, physical, physiological, and genetic markers. Healthcare practitioners may evaluate thousands of markers using artificial intelligence in pharmacovigilance to produce considerably more precise predictions about how particular drugs will affect particular people[122].

Pharmacological therapy will inevitably become more personalized, reducing ADRs and boosting drug effectiveness.

- Nanomedicine and Drug Delivery

Nanomedicine is currently a reality, not just a concept from science fiction. The research, which was published in the academic journal Drug Discovery Today[117], demonstrates how pioneers are using nanotechnology and medicine in tandem to identify, treat, and keep track of a variety of complex illnesses. Specialists deal with asthma, cancer, malaria, and HIV. Although nanomedicine is still in its infancy, advances have been made in the field of medication delivery by nanoparticle modification. According to a recently released study from a scholarly magazine[123], engineers and scientists are striving to construct implantable nanorobots that will improve drug delivery. Fuzzy logic, integrations, and neural networks are examples of AI techniques that can simplify the overall process.

III. CONCLUSION

The pharmaceutical industry is experiencing difficulties with drug development projects due to rising drug development costs and fewer chances of discovering new drug molecules. This finding has led to an increase in the number of pharmaceutical corporations and research institutions investigating the use of ML and robotics techniques to hasten the development of novel therapies and make the exchange of observational data and clinical trial outcomes easier. Multiple points in a drug's life cycle are amenable to ML algorithms. This has been demonstrated in detail in the preceding sections, where we discussed ML applications beginning with the drug discovery phases, such as target prediction and validation, discovering therapeutic and toxicity effect profiles of drugs, for prediction of, structure, bioactivity, and mode of action. More data on high-risk populations, long-term effects, food and drug interactions, and the escalation of known and unknown adverse effects of the drug over time are revealed by post-market drug monitoring. The use of ML in drug post-market monitoring increases compliance adherence and reduces expenditure significantly for each and every ICSR. Pharmacovigilance that uses AI can classify the harmful nature of reported events in addition to evaluating their quality.

Despite the fact that ML, augmented intelligence, and a variety of medical data from around the world are paving the road for unified global healthcare, some challenges in the utilization of Machine Learning algorithms for drug development lifecycle still persist today. For the construction and training of ML models, high-quality, precise, and painstakingly vetted data is necessary. The intricacy of the data type and the problem to be solved dictate the requirements for the necessary data quantity and accuracy. As a result, producing large data sets might be costly. It's important to keep in mind that when training, numerous neural network parameters are adjusted, some theoretical and practical frameworks for enhancing these models are not yet available. Another area where ML models fall short is in the prediction of novel paradigms. Because ML relies on training data to produce usable models, these models can only make predictions within the training data's predefined framework.

Drug research might be sped up and saved money by using AI technology. Although ML might not be a solution for all issues in drug discovery, it is unquestionably a useful tool when used appropriately with the right data. The power of artificial intelligence (AI) technology will undoubtedly be used to complement human intelligence and enhance our capabilities, thereby transforming the way we approach drug development.

REFERENCES

[1] Burbidge, R., Trotter, M.W.B., Buxton, B.F. & Holden, S. (2001) Drug design by machine learning: Support vector machines for pharmaceutical data analysis. Computers and Chemistry, 26, 5–14 [DOI: 10.1016/s0097-8485(01)00094-8] [PubMed: 11765851].

[2] Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, Perrot, Matthieu, Duchesnay, Edouard, Louppe & Gilles (2012). Scikit-learn: Machine Learning in Python Pedregosa, Fabian & Varoquaux, Gael & Gramfort. Journal of Machine Learning Research. Alexandre & Michel: Vincent & Thirion: Bertrand, USA, 12.

[3] Domingos, P. (2012) A few useful things to know about machine learning. Communications of the ACM, 55, 78–87 [DOI: 10.1145/2347736.2347755].

[4] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M. & Zhao, S. (2019) Applications of machine learning in drug discovery and development. Nature Reviews. Drug Discovery, 18, 463–477 [DOI: 10.1038/s41573-019-0024-5] [PubMed: 30976107] [PubMed Central: PMC6552674].

[5] Fakhraei, S., Huang, B., Raschid, L. & Getoor, L. (2014) Network-based drug-target interaction prediction with probabilistic soft logic. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11, 775–787 [DOI: 10.1109/TCBB.2014.2325031] [PubMed: 26356852].

[6] Bolton, E.E., Wang, Y., Thiessen, P.A. & Bryant, S.H. (2008) PubChem: Integrated platform of small molecules and biological activities. Annual Reports in Computational

Chemistry, 4, 217–241 [DOI: 10.1016/S1574-1400(08)00012-1].

[7] Xia, Z., Wu, L.Y., Zhou, X. & Wong, S.T. (2010) Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Systems Biology. BioMed Central, 4 (Supplement 2), S6 [DOI: 10.1186/1752-0509-4-S2-S6] [PubMed: 20840733].

[8] Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J. & Shoichet, B.K. (2007) Relating protein pharmacology by ligand chemistry. Nature Biotechnology, 25, 197–206 [DOI: 10.1038/nbt1284] [PubMed: 17287757].

[9] Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B., Whaley, R., Glennon, R.A., Hert, J., Thomas, K.L., Edwards, D.D., Shoichet, B.K. & Roth, B.L. (2009) Predicting new molecular targets for known drugs. Nature, 462, 175–181 [DOI: 10.1038/nature08506]. Epub 1 November 2009 [PubMed: 19881490] [PubMed Central: PMC2784146].

[10] Cheng, A.C., Coleman, R.G., Smyth, K.T., Cao, Q., Soulard, P., Caffrey, D.R., Salzberg, A.C. & Huang, E.S. (2007) Structure-based maximal affinity model predicts small-molecule druggability. Nature Biotechnology, 25, 71–75 [DOI: 10.1038/nbt1273] [PubMed: 17211405].

[11] Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S. & Olson, A.J. (2009) Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. Journal of Computational Chemistry, 30, 2785–2791 [DOI: 10.1002/jcc.21256] [PubMed: 19399780].

[12] Mousavian, Z. & Masoudi-Nejad, A. (2014) Drug–target interaction prediction via chemogenomic space: Learning-based methods. Expert Opinion on Drug Metabolism and Toxicology, 10, 1273–1287 [DOI: 10.1517/17425255.2014.950222] [PubMed: 25112457].

[13] Ding, H., Takigawa, I., Mamitsuka, H. & Zhu, S. (2014) Similarity-based machine learning methods for predicting drug–target interactions: A brief review. Briefings in Bioinformatics, 15, 734–747 [DOI: 10.1093/bib/bbt056] [PubMed: 23933754].

[14] Bleakley, K. & Yamanishi, Y. (2009) Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics, 25, 2397–2403 [DOI: 10.1093/bioinformatics/btp433] [PubMed: 19605421].

[15] Alaimo, S., Giugno, R. & Pulvirenti, A. (2016). Recommendation Techniques for Drug–Target Interaction Prediction and Drug Repositioning. Data Mining Techniques for the Life Sciences, 441–462.

[16] Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y. & Lu, H. (2017) Deep-learning-based drug-target interaction prediction. Journal of Proteome Research, 16, 1401–1409 [DOI: 10.1021/acs.jproteome.6b00618] [PubMed: 28264154].

[17] Bolton, E.E., Wang, Y., Thiessen, P.A. & Bryant, S.H. (2008) PubChem: Integrated platform of small molecules and biological activities. Annual Reports in Computational Chemistry, 4, 217–241 [DOI: 10.1016/S1574-1400(08)00012-1].

[18] Cristianini, N. & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press: Cambridge.

[19] Yu, H., Chen, J., Xu, X., Li, Y., Zhao, H., Fang, Y., Li, X., Zhou, W., Wang, W. & Wang, Y. (2012) A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. PLOS ONE, 7, e37608 [DOI: 10.1371/journal.pone.0037608] [PubMed: 22666371].

[20] Anagaw, A. & Chang, Y.-L. (2019) A new complement naïve Bayesian approach for biomedical data classification. Journal of Ambient Intelligence and Humanized Computing, 10, 3889–3897 [DOI: 10.1007/s12652-018-1160-1].

[21] Nigsch, F., Bender, A., Jenkins, J.L. & Mitchell, J.B.O. (2008) Ligand-target prediction using winnow and naïve Bayesian algorithms and the implications of overall performance statistics. Journal of Chemical Information and Modeling, 48, 2313–2325 [DOI: 10.1021/ci800079x] [PubMed: 19055411].

[22] Pang, X., Fu, W., Wang, J., Kang, D., Xu, L., Zhao, Y., Liu, A.L. & Du, G.H. (2018) Identification of estrogen receptor α antagonists from natural products via in vitro and in silico approaches. Oxidative Medicine and Cellular Longevity, 2018, 6040149 [DOI: 10.1155/2018/6040149] [PubMed: 29861831].

[23] Wei, Y., Li, W., Du, T., Hong, Z. & Lin, J. (2019) Targeting HIV/HCV coinfection using a machine learning-based multiple quantitative structure–activity relationships (multiple QSAR) method. International Journal of Molecular Sciences, 20, 3572 [DOI: 10.3390/ijms20143572] [PubMed: 31336592].

[24] Lengauer, T. (2007) Bioinformatics—From genomes to therapies. Bioinformatics- from Genomes to Therapies, 1–24.

[25] Nayal, M. & Honig, B. (2006) On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. Proteins, 63, 892–906 [DOI: 10.1002/prot.20897] [PubMed: 16477622]

[26] Li, Q. & Lai, L. (2007) Prediction of potential drug targets based on simple sequence properties. BMC Bioinformatics, 8, 353 [DOI: 10.1186/1471-2105-8-353] [PubMed: 17883836].

[27] Bakheet, T.M. & Doig, A.J. (2009) Properties and identification of human protein drug targets. Bioinformatics, 25, 451–457 [DOI: 10.1093/bioinformatics/btp002] [PubMed: 19164304].

[28] Wang, Q., Feng, Y., Huang, J., Wang, T. & Cheng, G. (2017) A novel framework for the identification of drug target proteins: Combining stacked auto-encoders with a biased support vector machine. PLOS ONE, 12, e0176486 [DOI: 10.1371/journal.pone.0176486] [PubMed: 28453576].

[29] Jeon, J., Nim, S., Teyra, J., Datti, A., Wrana, J.L., Sidhu, S.S., Moffat, J. & Kim, P.M. (2014) A systematic approach

to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. Genome Medicine, 6, 57 [DOI: 10.1186/s13073-014-0057-7] [PubMed: 25165489].

[30] Costa, P.R., Acencio, M.L. & Lemke, N. (2010) A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. BMC Genomics, 11 (Supplement 5), S9–S9 [DOI: 10.1186/1471-2164-11-S5-S9] [PubMed: 21210975].

[31] Kandoi, G., Acencio, M.L. & Lemke, N. (2015) Prediction of druggable proteins using machine learning and systems biology: A mini-review. Frontiers in Physiology, 6, 366–366 [DOI: 10.3389/fphys.2015.00366] [PubMed: 26696900].

[32] Rouillard, A.D., Hurle, M.R. & Agarwal, P. (2018) Systematic interrogation of diverse Omic data reveals interpretable, robust, and generalizable transcriptomic features of clinically successful therapeutic targets. PLOS Computational Biology, 14, e1006142 [DOI: 10.1371/journal.pcbi.1006142] [PubMed: 29782487].

[33] Kumar, V., Sanseau, P., Simola, D.F., Hurle, M.R. & Agarwal, P. (2016) Systematic analysis of drug targets confirms expression in disease-relevant tissues. Scientific Reports, 6, 36205 [DOI: 10.1038/srep36205] [PubMed: 27824084].

[34] Ferrero, E., Dunham, I. & Sanseau, P. (2017) In silico prediction of novel therapeutic targets using gene-disease association data. Journal of Translational Medicine, 15, 182 [DOI: 10.1186/s12967-017-1285-6] [PubMed: 28851378].

[35] Koscielny, G., An, P., Carvalho-Silva, D., Cham, J.A., Fumis, L., Gasparyan, R., Hasan, S., Karamanis, N., Maguire, M., Papa, E., Pierleoni, A., Pignatelli, M., Platt, T., Rowland, F., Wankar, P., Bento, A.P., Burdett, T., Fabregat, A., Forbes, S., Gaulton, A., Gonzalez, C.Y., Hermjakob, H., Hersey, A., Jupe, S., Kafkas, Ş, Keays, M., Leroy, C., Lopez, F.J., Magarinos, M.P., Malone, J., McEntyre, J., Munoz-Pomer Fuentes, A., O'Donovan, C., Papatheodorou, I., Parkinson, H., Palka, B., Paschall, J., Petryszak, R., Pratanwanich, N., Sarntivijal, S., Saunders, G., Sidiropoulos, K., Smith, T., Sondka, Z., Stegle, O., Tang, Y.A., Turner, E., Vaughan, B., Vrousgou, O., Watkins, X., Martin, M.J., Sanseau, P., Vamathevan, J., Birney, E., Barrett, J. & Dunham, I. (2017) Open targets: A platform for therapeutic target identification and validation. Nucleic Acids Research, 45, D985–D994 [DOI: 10.1093/nar/gkw1055] [PubMed: 27899665].

[36] Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R.P. & Pande, V. (2017) Is multitask deep learning practical for pharma? Journal of Chemical Information and Modeling, 57, 2068–2076 [DOI: 10.1021/acs.jcim.7b00146] [PubMed: 28692267].

[37] Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E. & Svetnik, V. (2015) Deep neural nets as a method for quantitative structure–activity relationships. Journal of Chemical Information and Modeling, 55, 263–274 [DOI: 10.1021/ci500747n] [PubMed: 25635324].

[38] Barati Farimani, A., Feinberg, E. & Pande, V. (2018) Binding pathway of opiates to μ-opioid receptors revealed by machine learning. Biophysical Journal, 114, 62a–63a [DOI: 10.1016/j.bpj.2017.11.390].

[39] Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K. & Pande, V. (2018) MoleculeNet: A benchmark for molecular machine learning. Chemical Science, 9, 513–530 [DOI: 10.1039/c7sc02664a] [PubMed: 29629118].

[40] Segler, M.H.S., Preuss, M. & Waller, M.P. (2018) Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 555, 604–610 [DOI: 10.1038/nature25978] [PubMed: 29595767].

[41] Desai, B., Dixon, K., Farrant, E., Feng, Q., Gibson, K.R., van Hoorn, W.P., Mills, J., Morgan, T., Parry, D.M., Ramjee, M.K., Selway, C.N., Tarver, G.J., Whitlock, G. & Wright, A.G. (2013) Rapid discovery of a novel series of abl kinase inhibitors by application of an integrated microfluidic synthesis and screening platform. Journal of Medicinal Chemistry, 56, 3033–3047 [DOI: 10.1021/jm400099d] [PubMed: 23441572].

[42] Hartenfeller, M. & Schneider, G. (2011) De novo drug design. In: Chemoinformatics and computational chemical biology. Methods in Molecular Biology. Springer: Berlin, 672, 299–323 [DOI: 10.1007/978-1-60761-839-3_12] [PubMed: 20838974].

[43] Schneider, G., Funatsu, K., Okuno, Y. & Winkler, D. (2017) De novo drug design-ye olde scoring problem revisited. Molecular Informatics, 36, 1681031 [DOI: 10.1002/minf.201681031] [PubMed: 28124833].

[44] Mullard, A. (2017) The drug-maker's guide to the galaxy. Nature, 549, 445–447 [DOI: 10.1038/549445a].

[45] Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K. & Zhavoronkov, A. (2017) The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. Oncotarget, 8, 10883–10890 [DOI: 10.18632/oncotarget.14073] [PubMed: 28029644].

[46] Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. (2017) Molecular de-novo design through deep reinforcement learning. Journal of Cheminformatics, 9, 48 [DOI: 10.1186/s13321-017-0235-x] [PubMed: 29086083].

[47] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. (2015) Human-level control through deep reinforcement learning. Nature, 518, 529–533 [DOI: 10.1038/nature14236] [PubMed: 25719670].

[48] Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P. & Aspuru-Guzik, A. (2018) Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science, 4, 268–276 [DOI: 10.1021/acscentsci.7b00572] [PubMed: 29532027].

[49] Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K. & Zhavoronkov, A. (2017)

The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. Oncotarget, 8, 10883–10890 [DOI: 10.18632/oncotarget.14073] [PubMed: 28029644].

[50] Coley, C.W., Rogers, L., Green, W.H. & Jensen, K.F. (2018) Scscore: Synthetic complexity learned from a reaction corpus. Journal of Chemical Information and Modeling, 58, 252–261 [DOI: 10.1021/acs.jcim.7b00622] [PubMed: 29309147].

[51] Andras, P. (2018) High-dimensional function approximation with neural networks for large volumes of data. IEEE Transactions on Neural Networks and Learning Systems, 29, 500–508 [DOI: 10.1109/TNNLS.2017.2651985] [PubMed: 28129193].

[52] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M. & Zhao, S. (2019) Applications of machine learning in drug discovery and development. Nature Reviews. Drug Discovery, 18, 463–477. Gale Academic OneFile accessed 2 October, 2022 [DOI: 10.1038/s41573-019-0024-5] [PubMed: 30976107].

[53] Lusci, A., Pollastri, G. & Baldi, P. (2013) Deep architectures and deep learning in chemoinformatics: The prediction of aqueous solubility for drug-like molecules. Journal of Chemical Information and Modeling, 53, 1563–1575 [DOI: 10.1021/ci400187y] [PubMed: 23795551].

[54] Duvenaud, D.K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A. & Adams, R.P. (2015) Convolutional networks on graphs for learning molecular fingerprints. In: Advances in Neural Information Processing Systems 28 (edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama & R. Garnett). Curran Associates, Inc, (2224–2232).

[55] Coley, C.W., Barzilay, R., Green, W.H., Jaakkola, T.S. & Jensen, K.F. (2017) Convolutional embedding of attributed molecular graphs for physical property prediction. Journal of Chemical Information and Modeling, 57, 1757–1772 [DOI: 10.1021/acs.jcim.6b00601] [PubMed: 28696688].

[56] Artursson, P. & Karlsson, J. (1991) Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (caco-2) cells. Biochemical and Biophysical Research Communications, 175, 880–885 [DOI: 10.1016/0006-291x(91)91647-u] [PubMed: 1673839].

[57] Hubatsch, I., Ragnarsson, E.G.E. & Artursson, P. (2007) Determination of drug permeability and prediction of drug absorption in caco-2 monolayers. Nature Protocols, 2, 2111–2119 [DOI: 10.1038/nprot.2007.303] [PubMed: 17853866].

[58] Wang, N.N., Dong, J., Deng, Y.H., Zhu, M.F., Wen, M., Yao, Z.J., Lu, A.P., Wang, J.B. & Cao, D.S. (2016) Adme properties evaluation in drug discovery: Prediction of caco-2 cell permeability using a combination of nsga-ii and boosting. Journal of Chemical Information and Modeling, 56, 763–773 [DOI: 10.1021/acs.jcim.5b00642] [PubMed: 27018227].

[59] Tian, S., Li, Y., Wang, J., Zhang, J. & Hou, T. (2011) Adme evaluation in drug discovery. 9. prediction of oral bioavailability in humans based on molecular properties and structural fingerprints. Molecular Pharmaceutics, 8, 841–851 [DOI: 10.1021/mp100444g] [PubMed: 21548635].

[60] Sim, D.S.M. (2015) Drug distribution. In: Pharmacological Basis of Acute Care. Springer: Berlin, pp. 27–36.

[61] Lombardo, F. & Jing, Y. (2016) In silico prediction of vol of distribution in humans. extensive data set and the exploration of linear and nonlinear methods coupled with molecular interaction fields descriptors. Journal of Chemical Information and Modeling, 56, 2042–2052 [DOI: 10.1021/acs.jcim.6b00044] [PubMed: 27602694].

[62] Matlock, M.K., Hughes, T.B. & Swamidass, S.J. (2015) Xenosite server: A web-available site of metabolism prediction tool. Bioinformatics, 31, 1136–1137 [DOI: 10.1093/bioinformatics/btu761] [PubMed: 25411327].

[63] Zaretzki, J., Matlock, M. & Swamidass, S.J. (2013) Xenosite: Accurately predicting cyp-mediated sites of metabolism with neural networks. Journal of Chemical Information and Modeling, 53, 3373–3383 [DOI: 10.1021/ci400518g] [PubMed: 24224933].

[64] Dang, N.L., Hughes, T.B., Krishnamurthy, V. & Swamidass, S.J. (2016) A simple model predicts ugt-mediated metabolism. Bioinformatics, 32, 3183–3189 [DOI: 10.1093/bioinformatics/btw350] [PubMed: 27324196].

[65] Lombardo, F., Obach, R.S., Varma, M.V., Stringer, R. & Berellini, G. (2014) Clearance mechanism assignment and total clearance prediction in human based upon in silico models. Journal of Medicinal Chemistry, 57, 4397–4405 [DOI: 10.1021/jm500436v] [PubMed: 24773013].

[66] Guengerich, F.P. (2011) Mechanisms of drug toxicity and relevance to pharmaceutical development. Drug Metabolism and Pharmacokinetics, p 1010210090, 26, 3–14 [DOI: 10.2133/dmpk.dmpk-10-rv-062] [PubMed: 20978361].

[67] Xu, Y., Pei, J. & Lai, L. (2017) Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. Journal of Chemical Information and Modeling, 57, 2672–2685 [DOI: 10.1021/acs.jcim.7b00244] [PubMed: 29019671].

[68] Sushko, I., Salmina, E., Potemkin, V.A., Poda, G. & Tetko, I.V. (2012) Toxalerts: A web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. Journal of Chemical Information and Modeling, 52, 2310–2316 [DOI: 10.1021/ci300245q] [PubMed: 22876798].

[69] Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. (2016) Deeptox: Toxicity prediction using deep learning. Frontiers in Environmental Science, 3, 80 [DOI: 10.3389/fenvs.2015.00080].

[70] Kearnes, S., Goldman, B. & Pande, V. (2016). Modeling Industrial ADMET Data with Multitask Networks. arXiv:1606.08793 [DOI: 10.48550/arXiv.1606.08793]

[71] Li, B., Shin, H., Gulbekyan, G., Pustovalova, O., Nikolsky, Y., Hope, A., Bessarabova, M., Schu, M., Kolpakova-Hart, E., Merberg, D., Dorner, A. & Trepicchio, W.L. (2015) Development of a drug-response modeling framework to

identify cell line derived translational biomarkers that can predict treatment outcome to erlotinib or sorafenib. PLOS ONE, 10, e0130700 [DOI: 10.1371/journal.pone.0130700] [PubMed: 26107615].

[72] van Gool, A.J., Bietrix, F., Caldenhoven, E., Zatloukal, K., Scherer, A., Litton, J.E., Meijer, G., Blomberg, N., Smith, A., Mons, B., Heringa, J., Koot, W.J., Smit, M.J., Hajduch, M., Rijnders, T. & Ussi, A. (2017) Bridging the translational innovation gap through good biomarker practice. Nature Reviews. Drug Discovery, 16, 587–588 [DOI: 10.1038/nrd.2017.72] [PubMed: 28450744].

[73] Kraus, V.B. (2018) Biomarkers as drug development tools: Discovery, validation, qualification and use. Nature Reviews. Rheumatology, 14, 354–362 [DOI: 10.1038/s41584-018-0005-9] [PubMed: 29760435].

[74] Shi, L., Campbell, G., Jones, W., Campagne, F., Wen, Z., Walker, S., Su, Z., Chu, T., Goodsaid, F., Pusztai, L. et al. (2010). The maqc-ii Project: A Comprehensive Study of Common Practices for the Development and Validation of Microarray-Based Predictive Models.

[75] Zhan, F., Huang, Y., Colla, S., Stewart, J.P., Hanamura, I., Gupta, S., Epstein, J., Yaccoby, S., Sawyer, J., Burington, B., Anaissie, E., Hollmig, K., Pineda-Roman, M., Tricot, G., van Rhee, F., Walker, R., Zangari, M., Crowley, J., Barlogie, B. & Shaughnessy, J.D. (2006) The molecular classification of multiple myeloma. Blood, 108, 2020–2028 [DOI: 10.1182/blood-2005-11-013458] [PubMed: 16728703].

[76] Shaughnessy, J.D., Jr, Zhan, F., Burington, B.E., Huang, Y., Colla, S., Hanamura, I., Stewart, J.P., Kordsmeier, B., Randolph, C., Williams, D.R., Xiao, Y., Xu, H., Epstein, J., Anaissie, E., Krishna, S.G., Cottler-Fox, M., Hollmig, K., Mohiuddin, A., Pineda-Roman, M., Tricot, G., van Rhee, F., Sawyer, J., Alsayed, Y., Walker, R., Zangari, M., Crowley, J. & Barlogie, B. (2007) A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. Blood, 109, 2276–2284 [DOI: 10.1182/blood-2006-07-038430] [PubMed: 17105813].

[77] Zhan, F., Barlogie, B., Mulligan, G., Shaughnessy, J.D., Jr & Bryant, B. (2008) High-risk myeloma: A gene expression-based risk-stratification model for newly diagnosed multiple myeloma treated with high-dose therapy is predictive of outcome in relapsed disease treated with single-agent bortezomib or high-dose dexamethasone. Blood J Am. Blood, 111, 968–969 [DOI: 10.1182/blood-2007-10-119321] [PubMed: 18182586].

[78] Decaux, O., Lodé, L., Magrangeas, F., Charbonnel, C., Gouraud, W., Jézéquel, P., Attal, M., Harousseau, J.L., Moreau, P., Bataille, R., Campion, L., Avet-Loiseau, H., Minvielle, S. & Intergroupe Francophone du Myélome (2008) Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosomal instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients: A study of the intergroupe francophone du myelome. Journal of Clinical Oncology, 26, 4798–4805 [DOI: 10.1200/JCO.2007.13.8545] [PubMed: 18591550].

[79] Mulligan, G., Mitsiades, C., Bryant, B., Zhan, F., Chng, W.J., Roels, S., Koenig, E., Fergus, A., Huang, Y., Richardson, P., Trepicchio, W.L., Broyl, A., Sonneveld, P., Shaughnessy, J.D., Bergsagel, P.L., Schenkein, D., Esseltine, D.L., Boral, A. & Anderson, K.C. (2007) Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. Blood, 109, 3177–3188 [DOI: 10.1182/blood-2006-09-044974] [PubMed: 17185464].

[80] Costello, J.C., Heiser, L.M., Georgii, E., Gönen, M., Menden, M.P., Wang, N.J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S.A., Mpindi, J.P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., NCI DREAM Community, Collins, J.J., Gallahan, D., Singer, D., Saez-Rodriguez, J., Kaski, S., Gray, J.W. & Stolovitzky, G. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. Nature Biotechnology, 32, 1202–1212 [DOI: 10.1038/nbt.2877] [PubMed: 24880487].

[81] Romero, K., Ito, K., Rogers, J.A., Polhamus, D., Qiu, R., Stephenson, D., Mohs, R., Lalonde, R., Sinha, V., Wang, Y., Brown, D., Isaac, M., Vamvakas, S., Hemmings, R., Pani, L., Bain, L.J., Corrigan, B., Alzheimer's Disease Neuroimaging Initiative & Coalition Against Major Diseases (2015) The future is now: Model-based clinical trial design for Alzheimer’s disease. Clinical Pharmacology and Therapeutics, 97, 210–214 [DOI: 10.1002/cpt.16] [PubMed: 25669145].

[82] Zhao, Y., Zeng, D., Socinski, M.A. & Kosorok, M.R. (2011) Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. Biometrics, 67, 1422–1433 [DOI: 10.1111/j.1541-0420.2011.01572.x] [PubMed: 21385164].

[83] Wong, C.H., Siah, K.W. & Lo, A.W. (2019) Estimation of clinical trial success rates and related parameters. Biostatistics, 20, 273–286 [DOI: 10.1093/biostatistics/kxx069] [PubMed: 29394327].

[84] Schork, N.J. (2015) Personalized medicine: Time for one-person trials. Nature, 520, 609–611 [DOI: 10.1038/520609a] [PubMed: 25925459].

[85] Vasudev, Naveen & Selby, Peter & Banks, Rosamonde. (2012). Renal cancer biomarkers: The promise of personalized care. BMC medicine. 10. 112. 10.1186/1741-7015-10-112.

[86] Zhang, X., Xiao, C., Glass, L. M., & Sun, J. (2020). DeepEnroll: Patient-Trial Matching with Deep Embedding and Entailment Prediction. In The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020 (pp. 1029-1037). (The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020). Association for Computing Machinery, Inc. https://doi.org/10.1145/3366423.3380181

[87] Calaprice-Whitty, D., Galil, K., Salloum, W., Zariv, A. & Jimenez, B. (2020) Improving clinical trial participant prescreening with artificial intelligence (AI): A comparison of the results of AI-assisted vs standard methods in 3 oncology trials. Therapeutic Innovation and Regulatory

Science, 54, 69–74 [DOI: 10.1007/s43441-019-00030-4] [PubMed: 32008227].

[88] Vassy, J.L., Ho, Y.L., Honerlaw, J., Cho, K., Gaziano, J.M., Wilson, P.W.F. & Gagnon, D.R. (2018) Yield and bias in defining a cohort study baseline from electronic health record data. Journal of Biomedical Informatics, 78, 54–59 [DOI: 10.1016/j.jbi.2017.12.017] [PubMed: 29305952].

[89] Weber, G.M., Adams, W.G., Bernstam, E.V., Bickel, J.P., Fox, K.P., Marsolo, K., Raghavan, V.A., Turchin, A., Zhou, X., Murphy, S.N. & Mandl, K.D. (2017) Biases introduced by filtering electronic health records for patients with 'complete data'. Journal of the American Medical Informatics Association, 24, 1134–1141 [DOI: 10.1093/jamia/ocx071] [PubMed: 29016972].

[90] Bain, E.E., Shafner, L., Walling, D.P., Othman, A.A., Chuang-Stein, C., Hinkle, J. & Hanina, A. (2017) Use of a novel artificial intelligence platform on mobile devices to assess dosing compliance in a phase 2 clinical trial in subjects with schizophrenia. JMIR mHealth and uHealth, 5, e18 [DOI: 10.2196/mhealth.7030] [PubMed: 28223265].

[91] Labovitz, D.L., Shafner, L., Reyes Gil, M., Virmani, D. & Hanina, A. (2017) Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy. Stroke, 48, 1416–1419 [DOI: 10.1161/STROKEAHA.116.016281] [PubMed: 28386037].

[92] Adamson, A.S. & Smith, A. (2018) Machine learning and health care disparities in dermatology. JAMA Dermatology, 154, 1247–1248 [DOI: 10.1001/jamadermatol.2018.2348] [PubMed: 30073260].

[93] Burlingame, E.A., Margolin, A.A., Gray, J.W. & Chang, Y.H. (2018) SHIFT: Speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks. Proceedings of SPIE–The International Society for Optical Engineering, 10581 [DOI: 10.1117/12.2293249] [PubMed: 30283195].

[94] Han, J., Chen, K., Fang, L., Zhang, S., Wang, F., Ma, H., Zhao, L. & Liu, S. (2019) Improving the efficacy of the data entry process for clinical research with a natural language processing-driven medical information extraction system: Quantitative field research. JMIR Medical Informatics, 7, e13331 [DOI: 10.2196/13331] [PubMed: 31313661].

[95] Gavrielov-Yusim, N., Kürzinger, M.L., Nishikawa, C., Pan, C., Pouget, J., Epstein, L.B., Golant, Y., Tcherny-Lessenot, S., Lin, S., Hamelin, B. & Juhaeri, J. (2019) Comparison of text processing methods in social media-based signal detection. Pharmacoepidemiology and Drug Safety, 28, 1309–1317 [DOI: 10.1002/pds.4857] [PubMed: 31392844].

[96] Barnett, I., Torous, J., Staples, P., Sandoval, L., Keshavan, M. & Onnela, J.P. (2018) Relapse prediction in schizophrenia through digital phenotyping: A pilot study. Neuropsychopharmacology, 43, 1660–1666 [DOI: 10.1038/s41386-018-0030-z] [PubMed: 29511333].

[97] Yurtman, A., Barshan, B. & Fidan, B. (2018) Activity recognition invariant to wearable sensor unit orientation using differential rotational transformations represented by quaternions. Sensors, 18, 2725 [DOI: 10.3390/s18082725] [PubMed: 30126235].

[98] Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P. & Ng, A.Y. (2019) Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25, 65–69 [DOI: 10.1038/s41591-018-0268-3] [PubMed: 30617320].

[99] Ozkanca, Y., Öztürk, M.G., Ekmekci, M.N., Atkins, D.C., Demiroglu, C. & Ghomi, R.H. (2019) Depression screening from voice samples of patients affected by Parkinson's disease. Digital Biomarkers, 3, 72–82 [DOI: 10.1159/000500354] [PubMed: 31872172].

[100] Han, X., Hu, Y., Foschini, L., Chinitz, L., Jankelson, L. & Ranganath, R. (2020) Deep learning models for electrocardiograms are susceptible to adversarial attack. Nature Medicine, 26, 360–363 [DOI: 10.1038/s41591-020-0791-x] [PubMed: 32152582].

[101] Vaci, N., Liu, Q., Kormilitzin, A., De Crescenzo, F., Kurtulmus, A., Harvey, J., O'Dell, B., Innocent, S., Tomlinson, A., Cipriani, A. & Nevado-Holgado, A. (2020) Natural language processing for structuring clinical text data on depression using UK-CRIS. Evidence-Based Mental Health, 23, 21–26 [DOI: 10.1136/ebmental-2019-300134] [PubMed: 32046989].

[102] Fonferko-Shadrach, B., Lacey, A.S., Roberts, A., Akbari, A., Thompson, S., Ford, D.V., Lyons, R.A., Rees, M.I. & Pickrell, W.O. (2019) Using natural language processing to extract structured epilepsy data from unstructured clinic letters: Development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open, 9, e023232 [DOI: 10.1136/bmjopen-2018-023232] [PubMed: 30940752].

[103] Savova, G.K., Danciu, I., Alamudun, F., Miller, T., Lin, C., Bitterman, D.S., Tourassi, G. & Warner, J.L. (2019) Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Research, 79, 5463–5470 [DOI: 10.1158/0008-5472.CAN-19-0579] [PubMed: 31395609].

[104] Estiri, H. & Murphy, S.N. (2019) Semi-supervised encoding for outlier detection in clinical observation data. Computer Methods and Programs in Biomedicine, 181, 104830 [DOI: 10.1016/j.cmpb.2019.01.002] [PubMed: 30658851].

[105] Glass & LMS (2019) G; Patil, R. Ai in Clinical Development: Improving Safety and Accelerating Results [White paper].

[106] Hicks, K.A., Mahaffey, K.W., Mehran, R., Nissen, S.E., Wiviott, S.D., Dunn, B., Solomon, S.D., Marler, J.R., Teerlink, J.R., Farb, A., Morrow, D.A., Targum, S.L., Sila, C.A., Hai, M.T.T., Jaff, M.R., Joffe, H.V., Cutlip, D.E., Desai, A.S., Lewis, E.F., Gibson, C.M., Landray, M.J., Lincoff, A.M., White, C.J., Brooks, S.S., Rosenfield, K., Domanski, M.J., Lansky, A.J., McMurray, J.J.V., Tcheng, J.E., Steinhubl, S.R., Burton, P., Mauri, L., O'Connor, C.M., Pfeffer, M.A., Hung, H.M.J., Stockbridge, N.L., Chaitman, B.R., Temple, R.J. & Standardized Data Collection for Cardiovascular Trials Initiative (SCTI)

(2018) 2017 Cardiovascular and stroke endpoint definitions for clinical trials. Circulation, 137, 961–972 [DOI: 10.1161/CIRCULATIONAHA.117.033502] [PubMed: 29483172].

[107] Liu, Y. & Gopalakrishnan, V. (2017) An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. Data, 2, 8 [DOI: 10.3390/data2010008] [PubMed: 28243594].

[108] Feng, T. & Narayanan, S. (2019) Imputing missing data in large-scale multivariate biomedical wearable recordings using bidirectional recurrent neural networks with temporal activation regularization. In:. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, 2529–2534 [DOI: 10.1109/EMBC.2019.8856966] [PubMed: 31946412].

[109] Qiu, Y.L., Zheng, H. & Gevaert, O.J. (2018). A Deep Learning Framework for Imputing Missing Values in Genomic Data.

[110] Tomic, A., Tomic, I., Rosenberg-Hasson, Y., Dekker, C.L., Maecker, H.T. & Davis, M.M. (2019) SIMON, an automated machine learning system, reveals immune signatures of influenza vaccine responses. Journal of Immunology, 203, 749–759 [DOI: 10.4049/jimmunol.1900033] [PubMed: 31201239].

[111] Rigdon, J., Baiocchi, M. & Basu, S. (2018) Preventing false discovery of heterogeneous treatment effect subgroups in randomized trials. Trials, 19, 382 [DOI: 10.1186/s13063-018-2774-5] [PubMed: 30012181].

[112] Kalscheur, M.M., Kipp, R.T., Tattersall, M.C., Mei, C., Buhr, K.A., DeMets, D.L., Field, M.E., Eckhardt, L.L. & Page, C.D. (2018) Machine learning algorithm predicts cardiac resynchronization therapy outcomes: Lessons from the companion trial. Circulation. Arrhythmia and Electrophysiology, 11, e005499 [DOI: 10.1161/CIRCEP.117.005499] [PubMed: 29326129].

[113] Linden, A. & Yarnold, P.R. (2016) Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. Journal of Evaluation in Clinical Practice, 22, 871–881 [DOI: 10.1111/jep.12610] [PubMed: 27421786].

[114] Schuler, M.S. & Rose, S. (2017) Targeted maximum likelihood estimation for causal inference in observational studies. American Journal of Epidemiology, 185, 65–73 [DOI: 10.1093/aje/kww165] [PubMed: 27941068].

[115] Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C. & Faisal, A.A. (2018) The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine, 24, 1716–1720 [DOI: 10.1038/s41591-018-0213-5] [PubMed: 30349085].

[116] Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y. & Ranganath, R. (2019) Practical guidance on artificial intelligence for health-care data. Lancet. Digital Health, 1, e157–e159 [DOI: 10.1016/S2589-7500(19)30084-6] [PubMed: 33323184].

[117] Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K. & Tekade, R.K. (2021) Artificial intelligence in drug discovery and development. Drug Discovery Today, 26, 80–93 [DOI: 10.1016/j.drudis.2020.10.010] [PubMed: 33099022].

[118] SciBite Limited (2022). Available at: https://www.scibite.com/

[119] Hauben, M. (2020) The potential of artificial intelligence in pharmacovigilance. Journal of the Faculty of Pharmaceutical Medicine. Available at: https://www.fpm.org.uk/journals/the-potential-of-artificial-intelligence-in-pharmacovigilance/. 10 November 2020.

[120] Correia, R.B., Li, L. & Rocha, L.M. (2016) Monitoring potential drug interactions and reactions via network analysis of Instagram user timelines. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 21, 492–503 [PubMed: 26776212] [PubMed Central: PMC4720984].

[121] Arlett, P., Straus, S. & Rasi, G. (2020) Pharmacovigilance 2030: Invited commentary for the January 2020 ‘futures’ edition of Clinical Pharmacology and therapeutics. Clinical Pharmacology and Therapeutics, 107, 89–91 [DOI: 10.1002/cpt.1689]. Epub 22 November 2019 [PubMed: 31758540] [PubMed Central: PMC6977396].

[122] Mishra, V., Chanda, P., Tambuwala, M.M. & Suttee, A. (2019)Personalized medicine: An overview. International Journal of Pharmaceutical Quality Assurance, 10, 290–294 [DOI: 10.25258/ijpqa.10.2.13].

[123] Fu, J. & Yan, H. (2012) Controlled drug release by a nanorobot. Nature Biotechnology, 30, 407–408 [DOI: 10.1038/nbt.2206] [PubMed: 22565965].