Published in final edited form as:

Clin Pharmacol Ther. 2021 September ; 110(3): 552–554. doi:10.1002/cpt.2329.

Machine Learning Challenges in Pharmacogenomic Research

Wei-Qi Wei, MD, PhD[1], **Juan Zhao, PhD**[1], **Dan Roden, MD**[1,2,3], **Josh F. Peterson, MD, MPH**[1,4]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

[2]Division of Cardiovascular Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[3]Oates Institute for Experimental Therapeutics, Vanderbilt University Medical Center, Nashville, TN, USA

[4]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Keywords

Machine learning; EHR; Pharmacogenomics; Omics data

Introduction:

Machine Learning (ML) is a potent technique for discovering and integrating big data associations. In this perspective, we outline the current opportunities, challenges, and limitations of applying ML to pharmacogenomic research.

Pharmacogenomics (PGx) studies the interaction between drug exposure and the human genome, including the impact of genetic variants on pharmacodynamics (PD), pharmacokinetics (PK) and subsequent clinical outcomes.[1] This relatively new field is rapidly growing in the past few decades, primarily due to the affordability of genotyped data and the exponential increase and availability of phenotypic data. With advances in genotyping technologies and improved phenotyping methods using electronic health records (EHR), researchers can study millions of genetic variants linked with thousands of disease phenotypes and drug treatments. Today, the amount of data generated in genomics and PGx phenotypes fits the definition of big data, which is characterized by high volume (the dataset size), wide variety (heterogeneity of data types), high velocity (the speed of accumulation of data), and inconsistent quality (the veracity and reliability of data). Analyzing such very large datasets may uncover novel trends and complex drug response patterns that are otherwise hidden from smaller, more controlled experiments. However, the task of sifting through these data, formalizing data representations, reconciling datasets across multiple sources, and distilling observed associations into knowledge that can personalize drug therapy often exceeds the capacity of human cognition. New methods and

Correspondence to: Wei-Qi Wei, MD PhD, wei-qi.wei@vumc.org, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave., Suite 1500, Nashville, TN 37203, Tel: (615) 343-1956.

CONFLICT OF INTEREST
The authors declared no competing interests for this work.

computing technologies to automate analytical model building for PGx are needed and could significantly accelerate the discovery of new PGx relationships.

As one of the foundations of artificial intelligence, Machine Learning (ML) represents a set of methods that can automatically uncover patterns in data and use the detected patterns to predict future data. ML is more capable of finding hidden patterns among multivariable than traditional association analyses; therefore, it is well-suited for high-dimensional data analysis needed for PGx research. Supervised, unsupervised, and reinforcement learning are the three basic ML paradigms (Figure 1). Supervised learning aims to solve classification/ regression problems. It formulates a model from pre-labeled (training) data and uses the model to predict unseen (testing) data, e.g., to estimate an individual's risk of developing a severe reaction after taking a drug. Unsupervised learning, i.e., knowledge discovery, aims to recognize patterns in unlabeled data through cluster analysis. A good example is automated grouping of individuals with a similar response to a drug. Reinforcement learning has gained momentum recently after reaching human-level performance in playing the ancient game of Go[2]. It focuses on sequential decision problems, in which a model interactively learns by using trial and error and using feedback from its own actions and experiences. A notable implementation of reinforcement learning is to dynamically optimize treatment for sepsis.[3]

ML has made great strides in the last decade and found use widely. Traditionally, investigators start with feature engineering to preselect relevant features such as comorbidities. Deep learning (DL), an emerging subtype of ML defined by greater layer depth of the underlying neural networks, offers an end-to-end learning ability that can automatically extract features. Recently, DeepMind used DL to accurately determine a protein's structure from its amino-acid sequence, demonstrating the real-world impact of ML in biomedical research.[4] A similar approach could be applied to biological data from highly controlled drug discovery experiments that carefully collect drug exposure, PK, and PD data over time, e.g., accelerating therapeutics for opportunities in medicine (ATOM). That type of systems biology analysis could unearth the network of physiologic and molecular influences on drug response. Few studies focus exclusively on PGx data. Two recent studies used ML to predict drug response using a combination of perturbational data, clinical information, and pharmacogenomic biomarkers.[5,6] Several manuscripts have provided a systematic review of ML for clinical medicine.[7,8] This paper will discuss some current opportunities, challenges, and limitations of applying ML within PGx research.

Applying ML techniques to pharmacogenomics faces several significant challenges. First, sufficient PGx data for ML model development is difficult to obtain, partially because the use of PGx in practice is still limited. Advanced ML models are 'greedy' because they demand massive training data, including labeled cases (individuals with a target phenotype) and controls (individuals without the phenotype). However, many PGx-mediated events (e.g., Steven-Johnson syndrome) are much rarer than healthy controls; therefore, most datasets annotated for PGx research are inherently imbalanced. ML cannot effectively learn from severely skewed data until adequate cases are provided. Additionally, identifying PGx phenotypes often requires gathering details about drug exposure (e.g., dose and adherence), clinical manifestations (e.g., vancomycin and red-man syndrome), and temporal sequence (e.g., events occurring within a specific time-frame following drug exposure). Such fine-grained details are often embedded in unstructured clinical text or dispersed across multiple types of EHRs or pharmacy systems; therefore, they remain challenging to extract from EHRs.[9] An ideal dataset would have both longitudinal drug exposure and response phenotypes precisely defined in addition to population-based pharmacogenetic results. To be widely generalizable, PGx data from diverse populations is needed to account for the differences in pharmacogene allele frequency among individuals of different ancestries. Without this type of data, ML could compound existing inequities through algorithmic bias. If used to find patterns of drug response related to genetic diversity, ML will need to account for existing healthcare disparities associated with poor drug outcomes that may confound the clinical-genetic relationships. Currently, these types of datasets and the associated deep phenotyping techniques to accurately extract PGx phenotypes remain underdeveloped.

Lack of sufficient data also impedes the generalizability of ML models. Even when large genetics studies exist, the number of participants is typically much smaller than the number of variants studied. Considering genomics has millions of variants while most are not related to prediction, such high dimensionality significantly increases the model complexity and quickly leads to overfitting, a type of error that occurs when a mathematical function is too closely aligned to a limited set of data points. Overfitting leads to poor reproducibility of an ML model when applied to another dataset, particularly from a health system external to the one that produced the test data. Standard technical solutions include compensation by imposing a penalty and feature reduction. However, methods to efficiently learn data representation for genomic data are still in their infancy. Data augmentation techniques like cropping and padding aid in optimizing the training dataset's potential, but their help is limited. As many large-scale data for genomics analyses are becoming publicly available, researchers can combine datasets from multiple studies. Techniques like federated learning and cloud computing enable multiple study sites to collaboratively learn a shared prediction model on the cloud while keeping all the training data for their own to reduce data security and privacy concerns. Nevertheless, considerable preprocessing and data harmonization are needed to generate a representative cohort for ML training.

The quality of the training data is crucial for ML's performance as a model needs to estimate the feature distribution of the target population based on these preselected samples. Yet, obtaining high-quality data requires both substantial effort and expense. Although genotyping is more affordable than ever before, extracting desired drug-relevant phenotypes from EHR remains complicated, typically demanding complex natural language processing and text mining techniques. We need to explore novel ways to formalize knowledge so a computer can understand the connections between drugs and other clinical terminology concepts, such as linking RxNorm and their indication diagnoses. Besides, data entry errors, outdated information, missing values, and inconsistent text in EHR generate loads of errors for using phenotypic data. Insufficient understanding of such challenges leads to failure in an ML task. Although several data quality dimensions have been proposed, there still lacks a consistent and generalizable method to assess EHR data quality.[10] Understanding and cleaning data properly are probably the most critical step of applying ML to biomedical data.

Another prominent issue of ML models is interpretability. Traditional ML models, e.g., decision tree, are interpretable. However, advanced ML models such as convolutional neural networks suffer from the black-box issue due to their complex innate structure. It remains challenging to use derived models to explain the mechanism behind and establish the causality relationship between the data and the outcome. This observation is more evident in the PGx domain, considering that researchers prefer scientific rationale of what actionable biomarkers are essential for translating clinical practice knowledge. Network-based approaches such as the deep graph network represent a promising alternative to elucidate the interaction among data. Existing efforts in decoding a black-box include SHAP (SHapley Additive exPlanations) and LIME. Besides, interpretation of the analysis results requires multi-field collaborations.

Considering that the drug response is often polygenic, researchers need to conduct large-scale gene analysis in carefully assessed cohorts before applying prediction models to real-world data. As PGx research moves quickly and irreversibly into the era of big data analytics, ML will almost certainly be used more frequently. We anticipate that scientists will discover many PGx associations, unveil the integrated effect of drug-drug interactions, and develop helpful ML models to predict drug responses. A standard flow to compare ML-enhanced treatment versus clinical-guideline-driven management will be necessary for future work before implementing ML-derived knowledge in clinic flow. To fulfill the power of advanced ML, we would hope to establish evaluation standards and create PGx benchmark datasets for assessing model performance. Investigators need to be careful when selecting data for ML training to avoid bias and ensure the clinical utilities. Tools that can assess the data quality, such as the cohort's diversity and the missing data, are essential. To accurately predict an individual's drug response requires understanding multi-view data, including the complex interactions among genetic, phenotypic, environmental, and lifestyle risk factors. It would also be critical to developing a data standard for collecting these data from various sources. Finally, the effect sizes of genetic variants are usually smaller than clinical risk factors. Applicable ML models need to be robust in handling heterogeneous data with disparate effect sizes.

FUNDING

The manuscript was supported by National Institute of General Medical Studies (P50 GM115305), National Heart, Lung, and Blood Institute (R01 HL133786), National Institute of General Medical Sciences (R01 GM139891), and National Institute on Aging (R01 AG069900).

References

1. Roden DM, McLeod HL, Relling MV, et al.Pharmacogenomics. Lancet. 2019;394(10197):521–532. [PubMed: 31395440]
2. Silver D, Huang A, Maddison CJ, et al.Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529(7587):484–489. [PubMed: 26819042]
3. Saria SIndividualized sepsis treatment using reinforcement learning. Nat Med. 2018;24(11):1641–1642. [PubMed: 30397359]
4. Senior AW, Evans R, Jumper J, et al.Improved protein structure prediction using potentials from deep learning. Nature. 2020;577(7792):706–710. [PubMed: 31942072]

5. Athreya AP, Neavin D, Carrillo-Roa T, et al.Pharmacogenomics-Driven Prediction of Antidepressant Treatment Outcomes: A Machine-Learning Approach With Multi-trial Replication. Clin Pharmacol Ther. 2019;106(4):855–865. [PubMed: 31012492]

6. Yuan H, Paskov I, Paskov H, Gonzalez AJ, Leslie CS. Multitask learning improves prediction of cancer drug sensitivity. Sci Rep. 2016;6:31619. [PubMed: 27550087]

7. Darcy AM, Louie AK, Roberts LW. Machine Learning and the Profession of Medicine. JAMA. 2016;315(6):551–552. [PubMed: 26864406]

8. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380(14):1347–1358. [PubMed: 30943338]

9. Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med. 2015;7(1):41. [PubMed: 25937834]

10. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144–151. [PubMed: 22733976]

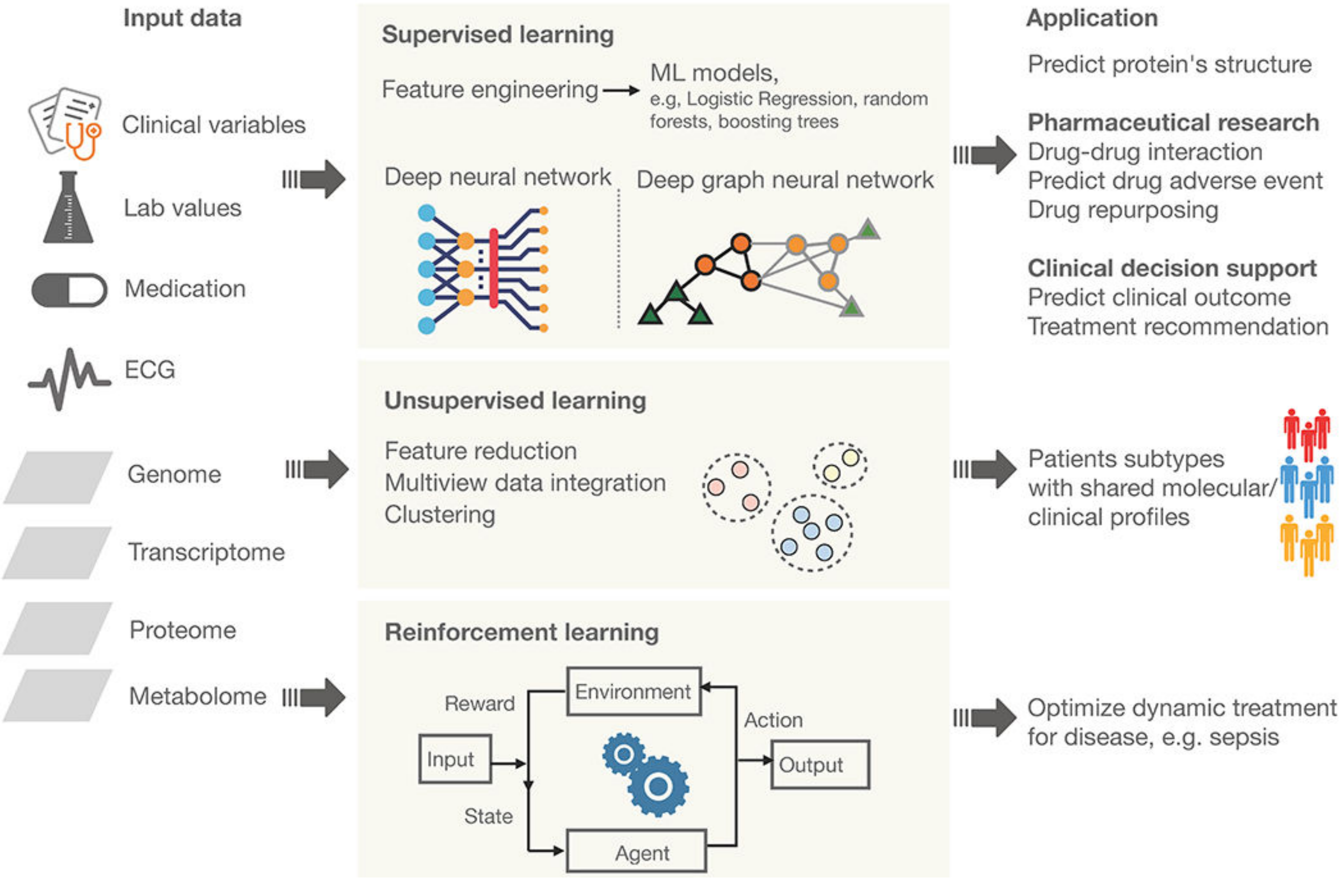

Figure 1.

Machine learning and PGx research.