Large scale patient pooling for drug discovery, pharmacovigilance investigations and precision medicines.

Klevinda Fili, MSc
Faculty of food and biochemical technology
University of Chemistry and Technology
Prague, Czech Republic
filik@vscht.cz

Kanishk Dwivedi, BE
Dept. Computer Science and Engineering, SIRT
University of Technology of Madhya Pradesh
Bhopal, Republic of India
kanishk.dwivedi@yahoo.com

***Abstract*— Patient pooling has been a major problem in the field of drug discovery and drug investigation. Even what is more daunting, is to provide a large scale solution for the classification of diseases and find side effects of personalised or precision medicine by clustering the pool and find similar investigations for pharmacovigilance, drug discovery and precision medicine. This can be solved by generating patterns through machine learning and deep learning models to find the common pools of similar pattern and diagnosis from clusters and distribute it by mobile application for the large scale patients clustering.This method is presented for Precision medicine, Pharmacovigilance and Drug discovery. Patients raw data is processed for classification and for personalised medicine. Patients collective information stored in database warehouses for clustering and applying advanced machine learning models on it will help in pharmacovigilance and early information regarding demographic disease epidemics. Patients diagnosis clustering can help to find out the pattern for drug discovery with respect to the geographical location and similar characteristics which have been found effective and will reduce time in drug discovery.**

Keywords—precision medicine, machine learning, AI, drug discovery, pharmacovigilance, disease epidemics, drug discovery, mobile application.

I. INTRODUCTION

The demand of the medical analysis for pharmaceuticals, drug discovery, pharmacovigilance, personalised and precision medicine requires more refined accurate and processed data for better treatments now and collection of data for near future. Patients, today have a discreet pattern of their treatment findings which makes it more and more difficult to discover drugs and medicines, also making it all time consuming. Another fact is pharmacovigilance which requires help to monitor the patients and find out any side effects of the drugs for a particular region. Real Time Patient Cluster (RTPP) is a collective solution which helps in many cases and provides real time analysis for healthcare solutions.

Precision medicine is one of the most promising initiatives for the medical future. It is patient-centric, where his genome decides the successful treatments and therapeutics. The patient places his treatment through the genetic information, lifestyle and environmental conditions in which he lives. Personalised medicine shifts the tailoring of medical approach and treatment to the characteristics and diagnosis of each patient and is expected to become the paradigm of future healthcare. Precision medicine is an initiative that includes a diverse field of study from biotechnology, pharmacogeonomics, pharmacoproteomics and computer science. For every, precise methods, there requires to have a relationship between all the diagnosis of a particular human being which is called Inter Diagnosis Relationships (IDRs) which is sub clustering of an individuals data to have accurate and precise information respective of the diagnosis.

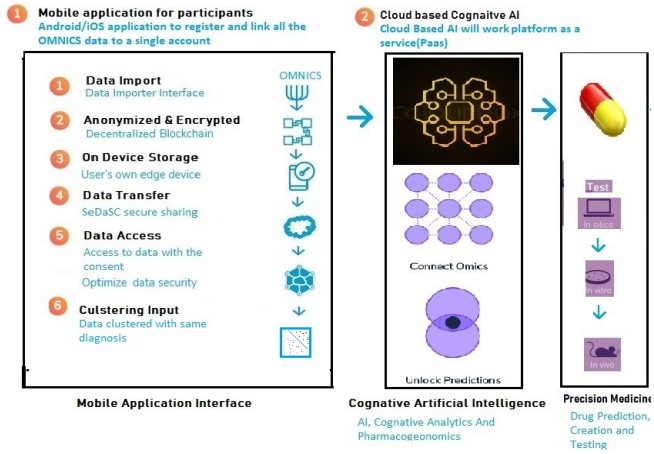

Fig 1

Pharmacovigilance has been challenging for pharmaceutical industries as the population grows and so does the medicine. Also, with sudden weather and geographic changes, it's getting hard for teams to monitor the products and asses them. With Real Time Patient Pooling, data is carefully monitored and can be refined by more advance machine learning algorithms. This clustered data will help pharmacovigilance teams to collect, detect, assess, monitor and prevent any pharmaceutical product before any big harm[1].

Drug discovery has been a more daunting challenge due its time consumption in discovering molecules and finding out more information about new molecule or drug. Drug discovery requires a lot of time because of data gathering and finding pattens in a clustered group for the epidemic diseases.

II. Related Work

There has been many ongoing research and breakthroughs in the field of machine learning, artificial intelligence and mass data gathering in healthcare sectors. Many researchers have done work in disease prediction and detection. approaches for these challenges are more streamlined, exposing new digital clinical endpoints and treatment response of biomarkers with close and efficient monitoring (such as circulating tumour DNA), improving safety and efficacy while reducing toxicity and adverse effects and greater insights into the patient journey via sensors, and low cost imaging. Securing, standardizing, and enhancing routinely collected electronic health data as a source of credible medical evidence can facilitate the organization of clinical trials at the point-of-care and should serve to improve the clinical development process [2]. Machine learning to predict accurate and precise pharmaceutical properties of molecular compounds and targets for drug discovery [3] . deep-learning models on multidimensional data sources such as combining genomic and clinical data to predict new predictive models [4].MIT Workshops discussed new pathways set up by regulatory agencies for evaluation and adoption of AI and ML in clinical development. For example, in 2016, the 21st Century Cures Act was signed into law, a significant bipartisan legislative achievement aimed at accelerating the discovery, development, and delivery of new cures and treatments was highlighted. FDA's current strategic policy places works on leveraging innovation and research, advancing digital health technologies, and developing next-generation artificial intelligence approaches to improve health care, it's access, and provide public health goals.

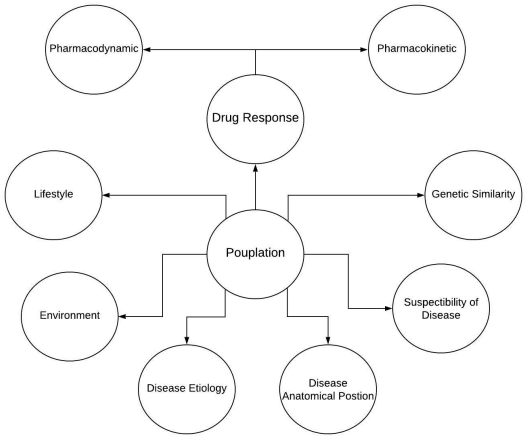

Fig 2

Major clinical guidelines, while advocating evidence-based suggestions, are not highly personalized to the pathophysiology of individual patients and fundamentally use a standard universal approach at the individual level for the clusters of the data clustered. For pharmacovigilance, It can be stated as :

"It is based on processed data from many sources (including observations and experiments), which suggests an association (either adverse or beneficial) between a drug or intervention and set of related events (e.g., a syndrome); (ii) it represents an association that is new and important, and has not been previously investigated and refuted. It demands investigation, being judged to be of sufficient likelihood to justify verify and, when necessary, act on remedial actions" [7].

pharmacogeonomics and omics are integrated by technologies will have the potential to identify genetic signals and patterns that are predictive of response or shows adverse outcome to particular drugs, and guide selection for a given individual. research in this field will enhance our understanding of how to maximise the deploy of various other drug classes to optimize the effects response at the individual level [8]. The machine learning and information pooling for detecting adverse reactions is explored and tested in pharmacovigilance generally but may have limited application to drugs due to lack of cluttered data. Sources may be infrequently used as communication channels by patients with rare disease, and adverse drug reactions or their caregivers or by health care providers; any adverse reactions identified are likely to reflect is known about safety of the drug from the network that grows up around these patients [9].

III. Proposed Model

A. Mobile application

Mobile applications are getting more easily accessible throughout the world by an estimation that there is 2.71 billion smartphone users in the world today – This data means

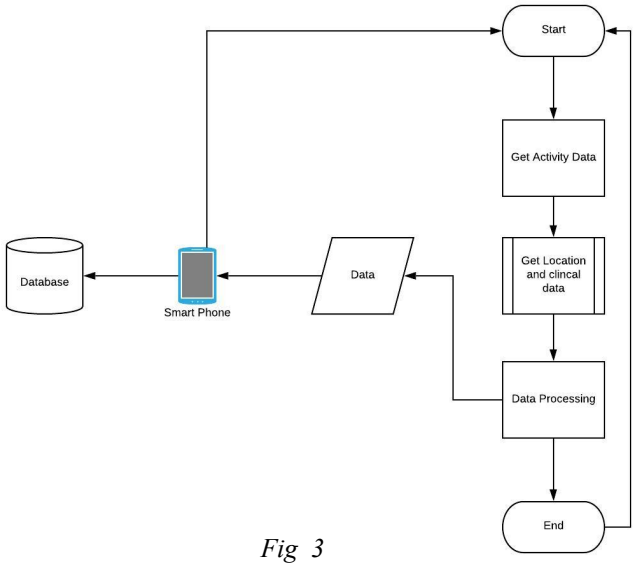

Fig 3

that in the world of communication, 35.13% of the world's population have a smartphone today [11]. Accessibility for research and medicine has been increased exponentially. Mobile applications are more in reach and collects more precise data like locations, sleeping and sitting activity.

These applications can be integrated with other applications to get clinical as well as molecular data.The security for data collection can be based on a centralised system as well as on a decentralised quorum based blockchain system.

Fig 3 explains the working of applications for collecting data and send them to cloud based infrastructure for processing.

B. Data Preprocessing

Data considered for preprocessing to remove the curse of dimensionality by selecting a selected number of data variables and finding out dimensions which are considered for further evaluation. Every individuals variable is collected and at first processed for optimising dimensions and identifying major dimensions which have more weight for investigative studies in Precision medicine, Drug discovery and Pharmacology.
Dataset of particular individual will have *n* dimensions which will be determine according to the diagnosis and the variables which shows more effect. The dataset is reduced using *P value*. Usually, if the P value of a data set is below a certain pre-determined amount (like, for instance, 0.05), we reject the "null hypothesis" of the experiment. We rule out the hypothesis that the variables of

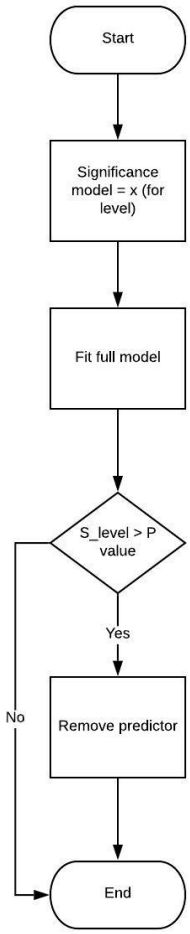

Fig 4

their experiment had no meaningful effect on the results. Today, p values are usually found on a reference table by first calculating a *chi square value.*

p values is calculated by :

$$x^2 = \sum ((o - e)^2 / e)$$

e = expected values
o = original values

Algorithm for the variable elimination is done by the backward elimination method.

C. Architecture

The modules of the model were trained in a joint simulation. This architecture follows a 3 stage data modelling and processing and training datasets with regards to the preprocessed model selected.

Drug Discovery : Drug discovery follows up a with variables which are important for work in drug discovery. Datasets are encoded with regards to the model and the hypothetical and dummy variables will be removed from the dataset. Data are more focused on the development of drugs and molecule reactions on the body.We have integrated different techniques and data sources in order to build a classifier whose outcome is a therapeutic class for a given drug[10]. The classified data is sent for more accurate and precise clustering of data. The clustering algorithm used as classification for the development of the model. This model classify similar clinical data for the molecular drug discovery or for early disease epidemics.

Precision Medicine : PM is very related to Pharmacogeonomics. Pharmacogeonomics is the study of genes affect on response to drugs. By using information about your genetic makeup, we can avoid the errors and trials in our patients. Using Pharmacogeonomics we manage a better way in using medicaments towards the current disease. So we lead the current medicaments into precise medicaments. the right drug, in the right time for the right patients, helps beating medicaments resistance.
Our understanding of disease mechanisms and gene function, information from patients' DNA can now also be used to provide diagnostic testing, inform therapeutic strategies and design preventative interventions tailored to the individual.9 DNA analysis has also led to expansion of pharmacogeonomics, through which drugs are prescribed based on information from a patient's genome, which tells clinicians how a patient will respond to a particular treatment.
The preprocessed will solely focus on the clustering of classified data for more accurate clusters. These cluster depends upon the variables based on the pharmacogeonomics and environmental based variables. These data variables can be filtered out by individuals mobile application data filter.

Pharmacovigilance : Automatic monitoring of Adverse Drug Reactions (ADRs), are for adverse patient outcomes caused by medicaments, is a challenging problem that is currently receiving attention from the medical informatics community [11]. Variables in pharmacovigilance model are separated during data preprocessing and uses variables that are mostly in datasets having various drug reactions and cluster the similar reactions to find out the a similar pattern in the patients and help to find similar information among the clusters. This will help in detection of side effects of drugs in a local environment also by identifying similar patterns.

Fig. 5 explains the flowchart of the architecture of the following pipeline of the model. This model can be modified in

deep neural networks for acquiring more precise information.

l = landmark

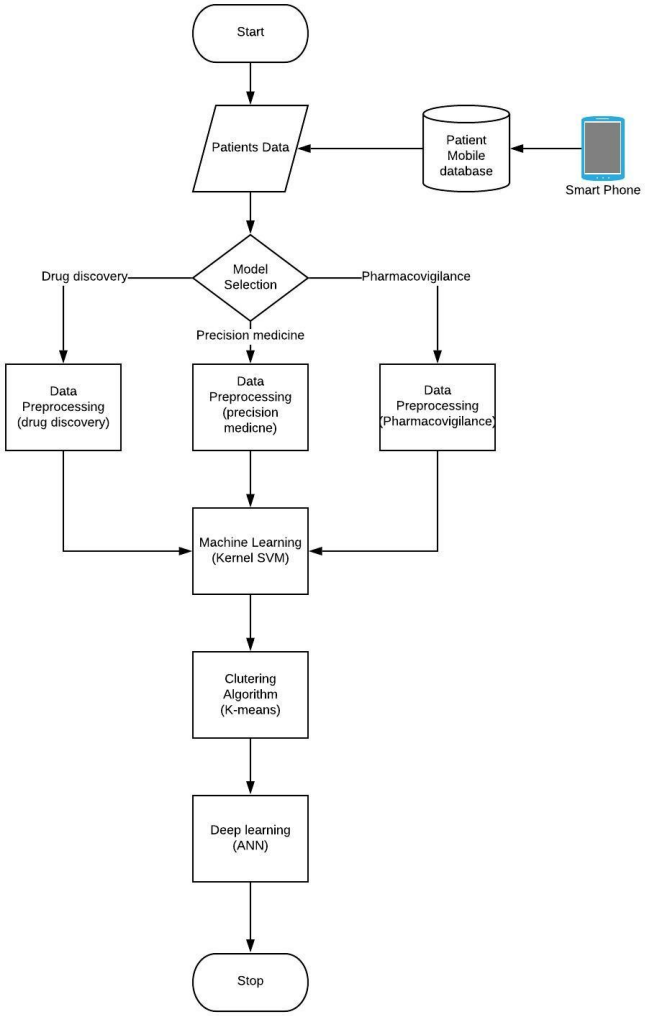

Fig 5

D. Machine Learning Models

Machine learning models are using the following models for the Classification and pooling of individuals data.

Kernel SVM : Kernel SVM (Support Vector Machine) is a classification algorithm which is non linear version of SVM using the 3D space and then after classification, it is inverse transformed into the 2D plane with linear hyperline as the SVM.
Kernel SVM uses Gaussian RBF kernel which is given as :

$$K(\vec{x}, \vec{l^i}) = e^{-\frac{||\vec{x} - \vec{l^i}||^2}{2\sigma^2}}$$

With RBF kernel, the information is given more accurate and precise with respect to SVM.

Kernel SVM is given on the following equations which are similar in every model, with respect to change of variables.

$$K(\vec{x}, \vec{l^{i1}}) + K(\vec{x}, \vec{l^{i2}})$$

$$K(\vec{x}, \vec{l^{i1}}) + K(\vec{x}, \vec{l^{i2}}) + \ldots > 0$$

$$K(\vec{x}, \vec{l^{i1}}) + K(\vec{x}, \vec{l^{i2}}) + \ldots = 0$$

$$K(\vec{x}, \vec{l^{i1}}) + K(\vec{x}, \vec{l^{i2}}) + \ldots < 0$$

The conditions for kernel SVM changes with respect to the preprocessed data and the pooling threshold of data. Also, the preprocessed data only verifies if the individuals clinical data is true for the clustering for the pool of sub-population. [22]

K-Means : using clustering algorithms on your dataset helps in unexpected things to show up like structures, clusters and groupings we never thought of otherwise.
K-Means clustering algorithm is the main algorithm which creates pools of individuals on the basis of location, environment and similar patterns found in patients. It is given by :

$$WCSS = \sum_{k=1}^{n} (\sum_{P_i : P_i \in k}^{k} distance\ (P_i, C_k)^2))$$

This equation determines the number of clusters required on the basis of multidimensional variables in the data. Also, It helps in finding optimal cluster for every data.

These datasets are clustered by finding similar patterns in clinical, pharmacogeonomics, environment and location data. Each variables differentiate in every model with accordance to the need of the information, which are regarding drug discovery, precision medicine and pharmacovigilance. [22]

IV. Experiments & Results

A. Experiment

We experimented with 1 variant of modular architecture-one where we used pre-trained models to generate inputs for our main prediction model. To test these models, last 1000 samples from the WHO dataset were combined as single test dataset. This combined dataset tries to simulate a pooling for similar pattern in individuals for drug discovery with instances of every type. Due to lack of samples in the pharmacovigilance and precision medicine dataset, it was not considered for testing.
The dataset was uploaded from mobile application developed on MIT app invented and called from PostrgeSQL database where a data was saved and called in python programming interface. The database was a report of HIV bacteria present in Tuberculosis patients in 217 countries including their environmental and location data.
In the modular architecture, we try to obtain an accuracy higher than each of the independent modules. The main prediction model is expected to learn to look at the predictions of each independent module, analyze them and make a single final prediction.
We used significant value of .5 and for the variables to be in the datasets. The p value was more than significant level, hence many variables were deleted to avoid curse of dimensionality. After, data preprocessing, the data is filtered and cleaned by removing 12 dummy variables out of 29 total variables. This preprocessing was done on the countries which had nil patients out of total countries. The final dataset had 60 countries with 17 variables. This model was fitted in kernel SVM algorithm to be processed and again filter out data which is not of use. The Dataset were split in training models and testing models with a split ratio of 0.2. These Datasets after modification were fitted into the K-means clusters. The K-means clusters applied the elbow method for determining the number of countries and clusters which share similar pattern. The model was then developed on the given clusters with a random value of 101.

B. Results

Mobile application uploaded the file to PostgreSQL and then was sent to the algorithm as input. The dataset was preprocessed with having *p values* higher than 0.5 in 17 variables out of total of 29 variables and 60 countries which had patients having both HIV as well as TB bacteria.
The data when fitted in Kernel SVM and then the refined database provided 11 out of 12 predictions true with respect to the testing data.

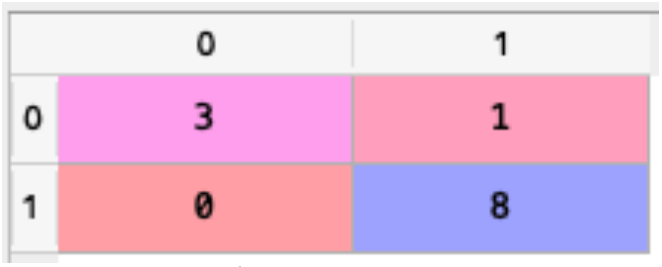

Fig 6

Fig 6. Explains the confusion matrix of predicted model to have an accuracy of 83.3%.

The Trained dataset is sent to the K-Means clusters with more refined data. The algorithms predicts 6 countries to have patients with HIV and TB bacteria together.

This was followed by using the elbow method to predict the number of clusters for the dataset.

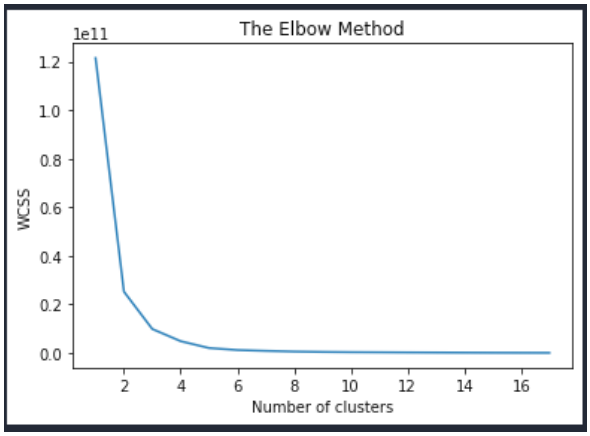

Fig 7

The algorithm successfully clustered and predicted 6 countries to have patients with HIV and TB bacteria. More data defined data shows to have Belarus and Ukraine having patients with same condition.
Fig 8 is the predicted clusters which shows Belarus, India, Pakistan, Russia, Ukraine, USA have patients with HIV and TB bacterias. More specific Belarus and Pakistan had more patients with HIV and TB bacteria respectively.

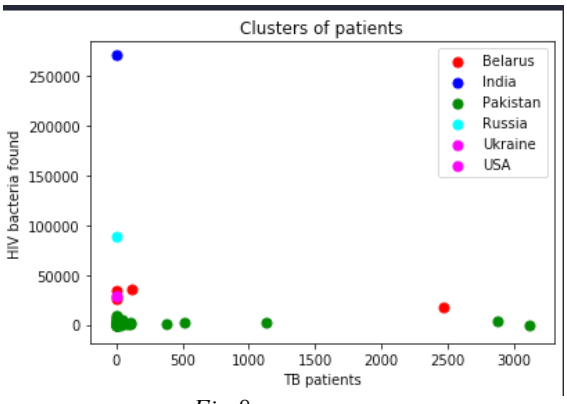

Fig 8

V. CONCLUSION

In conclusion, The proposed modular architecture was successful in achieving highest accuracy with all kinds of different dataset. The model successfully integrated all data in a dataset and uploaded to database and downloaded the same database to make a single final prediction. Discovering other variants of the modular architectures and improving hyper-parameters of the model to increase its existing accuracy can be progressed. Further, when we get more datasets with more variables in precision medicine, pharmacovigilance and drug discovery of different types of clinical, molecular, clinical and environmental data will help the model learn better and more variant features. The modular architecture can also be used in other fields of machine learning where there is a need to combine predictions from multiple independent variables to make a single prediction.

REFERENCES

1. Pratik Shah, Francis Kendall, Sean Khozin, Ryan Goosen, Jianying Hu, Jason Laramie, Michael Ringel & Nicholas Schork "Artificial intelligence and machine learning in clinical development: a translational perspective".
2. Jan Trøst Jørgensen, Maria Hersom "Companion diagnostics—a tool to improve pharmacotherapy" (Vol 4, No 24 (December 2016))
3. Shyamal Patel, Hyung Park, Paolo Bonato, Leighton Chan & Mary Rodgers, "A review of wearable sensors and systems with application in rehabilitation".
4. Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado & Jeffrey Dean, "Scalable and accurate deep learning with electronic health records".
5. Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review".
6. FDA, "Digital Health" (https://www.fda.gov/medical-devices/digital-health)
7. Carmine Savoia, Massimo Volpe, Guido Grassi, Claudio Borghi, Enrico Agabiti Rosei, and Rhian M. Touyz "Personalized medicine—a modern approach for the diagnosis and management of hypertension."
8. Pharmacovigilance, (https://en.wikipedia.org/wiki/Pharmacovigilance#Rest_of_Europe,_including_non-EU)
9. C. Lee Ventola, "Big Data and Pharmacovigilance: Data Mining for Adverse Drug Events and Interactions."
10. Francesco Napolitano, Yan Zhao, Vânia M Moreira, Roberto Tagliaferri, Juha Kere, Mauro D'Amato & Dario Greco, "Drug repositioning: a machine-learning approach through data integration."
11. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, Upadhaya T, Gonzalez G,"Utilizing social media data for pharmacovigilance: A review."
12. Allison B. Chambliss and Daniel W. Chan, "Precision medicine: from pharmacogenomics to pharmacoproteomics."
13. Genetic Home Reference, "What is the difference between precision medicine and personalized medicine? What about pharmacogenomics?" (https://ghr.nlm.nih.gov/primer/precisionmedicine/precisionvspersonalized).
14. Allison B. Chambliss and Daniel W. Chan, "Precision medicine: from pharmacogenomics to pharmacoproteomics".
15. National Cancer Institute, "Precision Medicine in Cancer Treatment". (https://www.cancer.gov/about-cancer/treatment/types/precision-medicine).
16. Thomas Ferkol and Paul Quinton "Precision Medicine: At What Price?"(Vol. 192, No. 6 | Sep 15, 2015).
17. Georgetown College of bioinformatics, Rice University (2017), "Demystifying biomedical Big data".
18. Matt Might, "Algorithm of Precision Medicine".Nancy Brown, Jay Flatley, Scott Gottlieb, Vasant Narasimhan, Tan Chorh Chuan, "The promise of precision medicine". (World Economic Forum 2018)
19. Jeanette McCarthy, "Geonomics and Precision medicine".
20. Vasant Narasimhan, "Precision medicine".(World Medical Innovation Forum 2018)
21. Victor Dzau, Geoffrey S Ginsburg, Elizabeth Finkelman, Celynne Balatbat, Kelsey Flott, Jessica Prestt, "PRECISION MEDICINE A GLOBAL ACTION PLAN FOR IMPACT".
22. Kirill Eremenko, Hadelin de Ponteves, Machine learning (https://www.udemy.com/course/machinelearning/#overview)