ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 9 N. 1 (2020), 5-14
eISSN: 2255-2863
DOI: https://dx.doi.org/10.14201/ADCAIJ202091514

Machine Learning techniques and Polygenic Risk Score application to prediction genetic diseases

Nibeth Mena Mamani[ab]

[a] Faculty of Science, University of Salamanca, 1 Escuelas St., Salamanca, 37008
[b] Master of Intelligent Systems, University of Salamanca, 1 Escuelas St., Salamanca, 37003
nmena@usal.es, nibeth.mena@gmail.com

KEYWORD

Machine Learning; Polygenic Risk Score; Genomic Data; Risk Prediction

ABSTRACT

For the last 10 years and after important discoveries such as DNA sequence of the entire human genome, there has been a considerable increase in the interest on researches risk prediction models associated with genetic originated diseases through two principal approaches: Polygenic Risk Score and Machine Learning techniques. The aim of this work is the literature review on Machine Learning techniques applied to obtaining the polygenic risk score, highlighting the most relevant researches and applications at present. The application of these techniques has provided many benefits in the prediction of diseases mot it is evident that the challenges of the use and optimization of these two approaches are still being discussed and investigated in order to have a greater precision in the prediction of genetic diseases.

1. Introduction

Advances in medicine and technology have enabled researchers to gather a wide range of information on the prediction of genetic diseases at a rapid rate. Complex human diseases such as cancer, cardiovascular, respiratory diseases and neurological disorders have caused enormous public health problems and economic loads (Ferlay *et al.*, 2013). Environmental factors such as smoking exposure, nutrient intake, physical exercise and genomic factors are believed to contribute to the development of complex human diseases (World Health Organization, 2018).

The aim of the work was to investigate the applicability of Machine Learning (ML) techniques and the Polygenic Risk Score (PRS) for prediction genetic diseases. In order to achieve the objective searches were conducted to identify relevant literature using the terms “machine learning and

polygenic risk score". First, a literature search was conducted through health and technology related research databases like PubMed and Science Direct. The Figure 1 is reflecting the considerable growth of articles published that contain the terms "machine learning and polygenic risk score" in the last 10 year ago.

Study selection was on base of the publication range for articles from 2015 to 2019 where the articles were included in the review if the following criteria were met: 1) The article had both terms "machine learning and polygenic risk score" in the title, abstract or content. 2) The article reported of ML techniques to address prediction genetic diseases and 3) The article was available in English. The articles were excluded if the following criteria were met: 1) The article report ML techniques but not focused to prediction genetic diseases using PRS and 2) The full text of the article was not available.

The search strategies identified 630 articles, with 50 of these articles meeting the inclusion criteria and 580 articles excluded for not met the inclusion criteria. From 50 articles reviewed with inclusion criteria applied 29 articles were excluded for met the exclusion criteria and finally there are 21 articles included in the literature review.

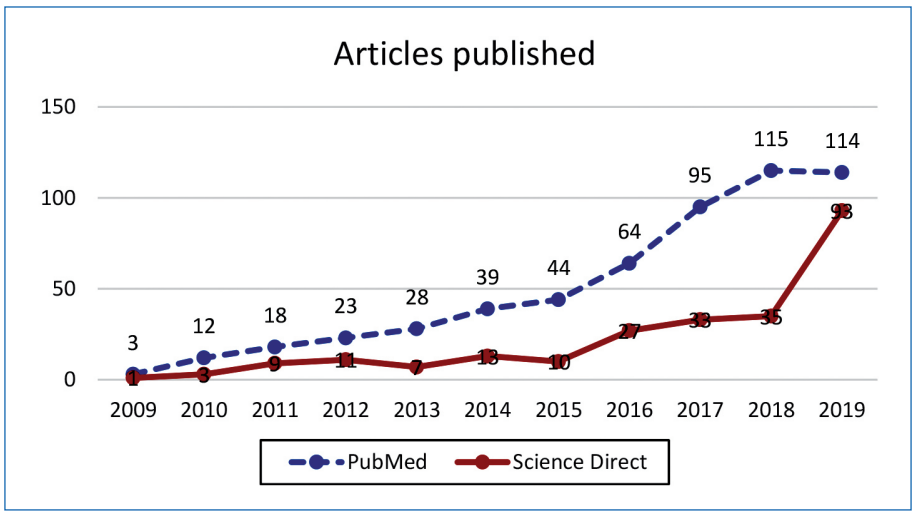

Figure 1: Articles published with "machine learning and polygenic risk score" in the title or content

2. Background

The great advances in technology have allowed scientists to make significant discoveries. A robust technique that has emerged to analyze this data is machine learning (ML), which aims to construct systems that can automatically improve through experience using advanced statistical and probabilistic techniques (Jordan *et al.,* 2015). ML has provided significant benefits to a range of fields including artificial intelligence, computer vision, speech recognition and natural language processing allowing researchers and developers to extract vital information from data, provide personalized experiences and develop intelligent systems (Jordan *et al.,* 2015). Within the field of medicine, the ML has brought significant advances allowing predictions of genetic diseases with complex data and variables (Ho *et al.,* 2019).

Nowadays, most of the diseases originate from genetic mutations in the DNA. In order to predict a disease, there are two main approaches. The first is the polygenic risk score known as a traditional approach which is a probabilistic data obtained through statistical techniques that predict the probability of contracting any genetic disease. The second approach is Machine Learning techniques which through a set of supervised and unsupervised algorithms predict the probability of contracting the genetic disease. Both approaches continue to be studied after the appearance of dependencies or factors that are external to genetics such as the environment, healthy habits, etc. which influence the development of the disease. Each study has a particular case where either PRS or ML are applied to predict a particular disease; There are studies where both approaches where applied to finally make the comparison of precision in the prediction. The objective of these approaches is to achieve maximum precision in the prediction of polygenic disorders. However, both approaches have challenges and limitations that are currently being investigated by scientists and doctors.

2.1. Genome-Wide Association Studies – GWAS

Thanks to the Next Generation Sequencing (NGS), which allows sequencing of the entire genome, a big leap from gene association study to Genome-wide Association Study known as GWAS was made. GWAS are studies that examine all the existing mutations in the genome by increasing the probability of finding the gene or genes that cause a certain disease. There are millions of single nucleotide polymorphisms (SNPs, also known as genetic variants). GWA identify SNPs that mark genomic regions that are strongly associated with phenotypes in a population. These genomic regions must contain the variant that is causally associated with the phenotype. However, it does not follow that the SNP identified by the GWA study is causal. In particular many common and complex diseases (e.g., type 2 diabetes (T2D) and obesity) are influenced by multiple SNPs each with small SNP effect sizes. GWA studies define SNPs according to their association with a disease / phenotype at the population level (Ho *et al.,* 2019).

3. Polygenic Risk Score – PRS

The polygenic risk score uses a fixed model approach to add the contribution of a set of risk alleles to a specific complex disease (Amin *et al.,* 2009).

In 2007 a method was proposed for examining the aggregate influence of multiple genetic markers by Wray *et al.* (2007). The method involved generating a PRS based on the results of a GWAS. After running a GWAS on a discovery sample, SNPs are selected for inclusion in the PRS based on their association with the phenotype. Using a validation sample, the PRS can be calculated as a sum of the alleles associated with the phenotype (often weighted by the specific SNP coefficients of the GWAS). With that score the joint association of multiple SNPs with the given trait can be evaluated. In general the PRS techniques have become increasingly popular facilitating genetic discoveries for complex traits. However since they are based on linear combinations of markers, traditional PRSs may not capture nonlinearity among SNPs (Levie *et al.,* 2017).

The best known and used software for estimating PRS in studies on the prediction of risk of genetic diseases is PRSice (https://www.prsice.info). That software allows to calculate of the PRS, evaluate and analyze the results of the PRSs (Choi *et al.,* 2019; Euesden *et al.,* 2019).

4. Machine Learning Techniques

Machine learning approaches adapt a set of sophisticated statistical and computational algorithms. The application of these techniques based on their set of supervised and unsupervised algorithms offers several application approaches such as profile analysis, disease diagnosis, drug development (Kristy *et al.,* 2018; Vamathevan *et al.,* 2019), disease prediction, detect epistasis within the human genome (McKinney *et al.,* 2006). These techniques: support vector machine (SVM) (Cortes *et al.,* 1995), Random forests (Breiman *et al.,* 2001) and k-nearest neighbors (Altman *et al.,* 1992), have been successfully applied in the prediction of risk of complex diseases according to clinical data (Zhang *et al.,* 2016; Gao *et al.,* 2018; Wu *et al.,* 2018; Antonucci *et al., 2019*).

In Gao's study (Gao *et al.,* 2018) of prediction and classification of Parkinson disease, for example, several methods are used without models like Random Forest, AdaBoost, XGBoost, Support Vector Machines, Neural Network and SuperLearner. Like the development of the PRS, the ML algorithms have the development and validation phase. In the development phase the algorithm will be trained with a dataset selected, in the validation phase as well as the PRS evaluates the prediction power in an independent dataset, adding a final validation called Test using another independent dataset so that the prediction power is finally confirmed (Ho *et al*., 2019).

These algorithms benefit from constant learning or retraining, since they do not guarantee optimized classification / regression results. However, when properly and effectively trained, maintained and reinforced, automatic learning methods without models have great potential for solving real-world problems (prediction and data mining) (Gao *et al.,* 2018).

Machine learning covers an extensive class of algorithms widely used to solve complex prediction problems. Regression trees are ideal for updating SNP weights in polygenic risk scores. This gradient-powered technique is powerful and versatile methods for continuous prediction of results. Gradient Boosting (Shapire *et al.,* 2012) is an efficient algorithm that sequentially combines a large number of weakly predictive models to optimize performance (Paré *et al.,* 2017).

4.1. Deep Learning

Deep learning algorithms developed from neural network algorithms have gained much interest after their successful implementation in image recognition and natural language processing applications. In genomics, deep learning applications are helping to identify functional DNA sequences, protein binding motifs, epigenetic markers (Ho *et al.,* 2019), model the complex dependencies of the genome to provide predictors (Telenti *et al.,* 2018), also has been used for discovery of sites for regulation or splicing (Leung *et al.,* 2014; Xiong *et al.,* 2015) and prediction of variant functions (Zhou *et al.,* 2015). However, the performance of deep learning in predicting disease status is in an early stage (Wu *et al.,* 2018).

An important recent advance in machine learning is the rapid development of deep learning algorithms that can efficiently extract meaningful characteristics from complex high-dimensional datasets through a hierarchical stacked learning process. There are few studies like of Wu *et al.* (2018) and Telenti *et al,* (2018) where they applied the Deep Learning approach to predict disease status based on genomic data indicating. However, to predict disease status is still in its infancy.

5. Machine Learning ang Polygenic Risk Score

Genetic risk prediction models are generally constructed by: (1) polygenic risk rating; or (2) Machine learning. The predictive performance of both types of models is evaluated by the receiver's operational characteristic curves (ROC) where the sensitivity and specificity of the predictions are classified into several cut-off values (Ho *et al.,* 2019).

The use of Machine Learning and PRS techniques will depend very much on the type of genetic disease being studied. For example, for psychiatric diseases the PRS is generated by PRSice Software (https://prsice.info/) (Choi *et al.,* 2019; Euesden *et al.,* 2019) and genome-wide association studies as a complement to polygenic risk prediction. Other techniques are used such as multivariate relevant vector regression (RVR) where RVR is a nucleus-based probabilistic pattern recognition method that uses Bayesian inference to obtain dispersed regression models and allows the extraction of patterns within a high-dimensional characteristic space (Ranlund *et al.,* 2018) a multivariate automatic learning method.

A new heuristic based on automatic learning techniques (GraBLD) is proposed to increase performance and improve prediction of PRSs (Paré *et al.,* 2017). This study demonstrated the use of ML techniques and GWAS to improve the prediction of polygenic traits where it proposes to take advantage of the large number of SNPs and the statistics available at the summary level of genome-wide association study to calibrate the weights of SNPs that contribute to the polygenic risk score based on pruning the linkage disequilibrium (LD) of the SNPs, prioritizing the most significant associations up to an empirically determined p-value threshold and pruning the remaining LD-based SNPs (Paré *et al.,* 2017).

Weighted analysis of the genetic correlation network (WGCNA) has been used repeatedly for the successful identification of epigenetic and transcriptomic networks which relate to a number of physical behavioral and disease traits. Morgan's study (Levine *et al.,* 2017) demonstrates a WGCNA-based method that can be applied to SNP data called weighted SNP correlation network analysis (WSCNA). In addition to taking into account the influence of LD this method also incorporates a semi-supervised ML approach that will facilitate the detection of modules that are specific to each trait where GWAS, PRS and heritability analysis have been extensively studied.

Despite recent improvements, the results of the polygenic risk score remain limited due to the approaches currently in use. In contrast, ML algorithms have increased predictive capabilities for the risk of complex diseases. This increase in predictive capabilities results from the ability of ML to handle multidimensional data (Ho *et al.,* 2019).

5.1. Comparations between Machine Learning and Polygenic Risk Score

Polygenic risk scoring and machine learning are two main approaches to predicting disease risk. Despite recent improvements the polygenic risk score results remain limited due to the approaches currently used like the Pisanu's study where said "the prediction accuracy shown by the PRS in the study is still in-sufficient to support the implementation of the model into the clinical practice" (Pisanu *et al.,* 2019). In contrast, machine learning algorithms have increased predictive capabilities for the risk of complex diseases (Ho *et al.,* 2019). However, there are other studies that indicate that PRS remains the best option for classifying cases of diseases such as schizophrenia (Griffiths *et al.,* 2019; Doan *et al.,* 2017; Cao *et al.,* 2018).

There are studies such as that of Ho *et al.* (2019) which provides a general description of the polygenic risk score and machine learning in predicting the risk of complex diseases. Highlighting recent

developments in machine learning applications and how machine learning approaches can lead to better prediction of complex diseases.

Figure 2 summarizes the strengths and weaknesses of both the polygenic risk score and the automatic learning predictive models (Ho *et al.,* 2019).

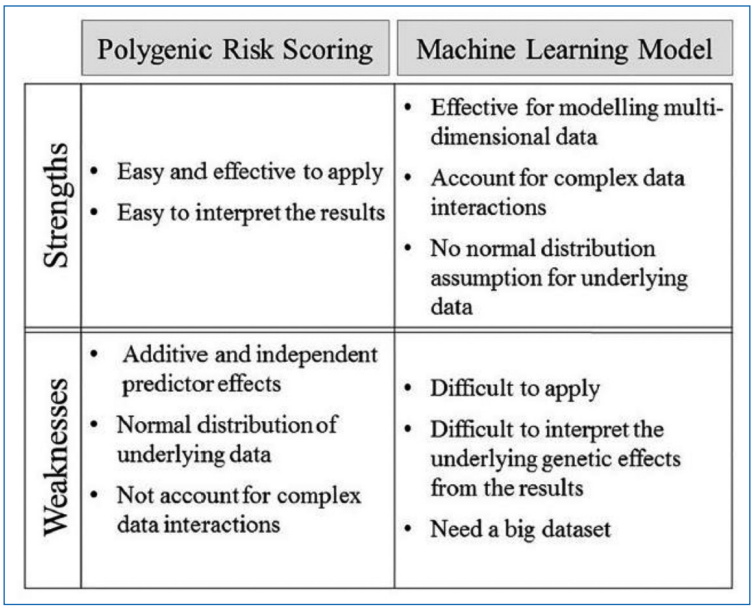

Figure 2: The strengths and weaknesses of polygenic risk scoring and machine learning model (Ho et al.*, 2019).*

6. Discussion

Since genomic sequencing was achieved GWAS and PRS have become a powerful tool for predicting the genetic predisposition of diseases. According to the reviewed articles using of PRS in peculiar diseases such as breast cancer, diabetes and obesity is very high (Shieh *et al.,* 2017; Kuchenbaecker *et al.,* 2017; Reisberg *et al.,* 2017; Torkamani *et al.,* 2019) but some articles identified that PRS is affected by environmental risk factors, heritability, diet, exercise, even sex and age. On the other hand, these factors are better handled by the ML tools but the limiting factor in it as already mentioned in other sections is the data for training, validation and tests.

Machine learning techniques are used after previous selection of the set of existing techniques based on the characteristics of variables and parameters in the data to be analyzed, for example if you have images as a data set you will have to use SVM. It would not be too much to have justifications and orientations of why to choose a particular technique. ML techniques have helped to solve a wide range of prediction problems but they are not widely used to construct polygenic risk scores for predicting complex traits (Paré *et al.,* 2017).

A study has three methods selected for Parkinson's diseases (Gao *et al.,* 2018). These three techniques will not have the same impact on schizophrenia diseases. PRS and ML are unique for each

type of disease. That means that once obtained they cannot be reused in other diseases. It is also very important to know the type of information what type of data you have as a sample and to know all the variables AND parameters so the interpretation of results is easy and significant. In addition to be able to select a group of techniques that complement and provide better prediction accuracy likewise a group of techniques should be used because it is necessary to make comparisons of the results of the experiment and evaluate the performance of each technique to see which one of them allows us to have a better precision of the risk of a disease.

Results from studies by Paré *et al.* (2017) show that ML techniques together with large meta-analysis data from the entire genome and the large number of genetic variants reported in GWAS to train gradient powered regression tree models through genome division improve the prediction of polygenic traits (Paré *et al*., 2017) It means that they are in favor of ML techniques for disease prediction.

Deep learning as a proposal for disease prediction too, because in the studies reviewed they mentioned the number of hidden layers and the number of nodes in each hidden layer used for automatic encoders but no justifications and guidance were given on why to choose those specific numbers of hidden layers and those specific numbers of nodes in each hidden layer. This is probably one of the main reasons why deep learning has not been widely used in genomic research (Wu *et al.,* 2018).

The main limitation of the studies is related to the sample size that may be relatively small for the purposes of genetic studies (Ranlund *et al.,* 2018; Telenti *et al.,* 2018) where the number of genomic factors is much greater than the number of samples, which results in an excessive adjustment of the model and computational inefficiency (Wu *et al.,* 2018) and it could be a strong reason why we don't even trust Machine learning prediction methods to clinically make the corresponding decision making. Applying machine learning or Deep learning techniques not only entails having great data for the training and testing of the proposed models but also requires super computations (https://hpcc.usc.edu/) of high performance.

7. Conclusion

PRS is a statistical method to predict diseases with multiple genetic variations and ML tools are other computational techniques that also predict diseases with genetic variations. There are studies that show that ML can improve the prediction of a specific disease as there are also studies that indicate that traditional PRS is better than ML to predict. What will it depend on? it could depend a lot on the type of disease. The values that GWAS provides or other genetic and external factors that are under study. It is evident that both go in parallel paths but aiming at the same objective: To predict more accurately the risk of contracting a genetic disease. Finally, for both approaches the research and experiments must be continued by the scientific community facing all the challenges that arise and take full advantage of the full potential of Machine Learning to achieve high precision in the predictions and thus be able to make clinical decisions about detection and / or early intervention of genetic diseases.

8. Future work

The ML techniques predicting, diagnose the disease, personalize the treatment and develop new medicines as report the Kristy *et al.* (2018) and Wu *et al.* (2018) articles. It is important to make full use of existing prediction techniques from traditional PRS, ML and data mining techniques if

intelligently bringing together all the techniques in the corresponding phases and with required data in terms of quality and quantity them could have a greater approximation to the high precision of predictions of polygenic diseases. Something very interesting is the following article by Ripke *et al.* (2019) which was recently published where it refers to a path based on the Polygenic Risk Score using prediction techniques and classification of Machine Learning (Ripke *et al.,* 2019).

This literature review was conducted at a general level. It would be very interesting to have a study on ML algorithms used to predict a particular disease or to have an area of focus, for example ML techniques to predict diseases like oncology, neurological, schizophrenia, parkinson, alzheimer, prostate cancer, breast cancer, etc. Having the techniques used in different studies for the same disease, let's say 3 to 4 ML techniques we should look the benefits that ML techniques offer and how much precision they have in predicting the disease under study in addition to the comparison these techniques of ML and traditional PRSs have been defined in other studies with the help of GWAS of course under the same disease.

9. References

Amin, N., van Duijn, C. M., & Janssens, A. C. 2009. Genetic scoring analysis: a way forward in genome wide association studies. *European journal of epidemiology,* 24(10), 585–587. Springer.

Antonucci L, Pergola G, Dwyer D, Torretta S, Romano R, …, et al. 2019 Classification of Schizophrenia Using Machine Learning with Multimodal Markers. *Biological Psychiatry, Elsevier,* Vol. 85, p. S107.

Altman N, 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46: 175–185.

Breiman L, 2001. Random forests. *Machine learning* 45: 5–32.

Cao, H., Meyer-Lindenberg, A., & Schwarz, E. 2018. Comparative Evaluation of Machine Learning Strategies for Analyzing Big Data in Psychiatry. *International journal of molecular sciences*, 19(11), 3387.

Choi SW, and O'Reilly PF. 2019. PRSice-2: Polygenic Risk Score Software for Biobank-Scale Data. GigaScience 8. PRSice

Cortes C, Vapnik V, 1995. Support-vector networks. *Machine* learning 20: 273–297.

Doan, N. T., Kaufmann, T., Bettella, F., Jørgensen, K. N., Brandt, C. L., Moberget, T., Alnæs, D., Douaud, G., Duff, E., Djurovic, S., Melle, I., Ueland, T., Agartz, I., Andreassen, O. A., & Westlye, L. T. 2017. Distinct multivariate brain morphological patterns and their added predictive value with cognitive and polygenic risk scores in mental disorders. *NeuroImage. Clinical*, *Elsevier,* Vol. 15, pages 719–731.

Euesden, J., Lewis, C. M., & O'Reilly, P. F. 2015. PRSice: Polygenic Risk Score software. *Bioinformatics (Oxford, England)*, 31(9), pages 1466–1468.

Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, 2013. Cancer Incidence and Mortality World GLOBOCAN 2012 v1.0, wide: IARC Cancer Base. International Agency for Research on Cancer: Lyon, France.

Gao C, Sun H, Wang T, Tang M, Bohnen Nl, et al. 2018. Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease. *Scientific Reports*, 8(1): 7129.

Griffiths, T., Baker, E., Schmidt, K. M., Bracher-Smith, M., Walters, J., Artemiou, A., … Escott-Price, V. 2019. Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score

with kernel support vector machines approach. *American journal of medical genetics*. 180(1): pages 80–85.

Ho, D., Schierding, W., Wake, M., Saffery, R., & O'Sullivan, 2019. Machine Learning SNP Based Prediction for Precision Medicine. *Frontiers in genetics*, 10: 267.

Jordan MI, Mitchell TM., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245): pages 255-60.

Kristy A. Carpenter, Xudong Huang, 2018. Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. *A Review. Current pharmaceutical design*, 24(28): pages 3347–3358.

Kuchenbaecker, K. B., McGuffog, L., Barrowdale, D., Lee, A., Soucy, P., Dennis, J., Domchek, S. M., Robson, M., Spurdle, A. B., Ramus, S. J., Mavaddat, N., Terry, M. B., Neuhausen, S. L., Schmutzler, R. K., Simard, J., Pharoah, P., Offit, K., Couch, F. J., Chenevix-Trench, G., Easton, D. F., … Antoniou, A. C. 2017. Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *Journal of the National Cancer Institute*, 109(7): djw302.

Leung, M. K., Xiong, H. Y., Lee, L. J., & Frey, B. J. 2014. Deep learning of the tissue-regulated splicing code. *Bioinformatics (Oxford, England)*, 30(12): i121–i129.

Levine, M. E., Langfelder, P., & Horvath, S. 2017. A Weighted SNP Correlation Network Method for Estimating Polygenic Risk Scores. *Methods in molecular biology (Clifton, N.J.),* 1613: pages 277–290.

McKinney, B. A., Reif, D. M., Ritchie, M. D., & Moore, J. H. 2006. Machine learning for detecting gene-gene interactions: a review. *Applied bioinformatics,* 5(2): pages 77–88.

Paré G, Mao S, Deng W Q, 2017. A machine-learning heuristic to improve gene score prediction of polygenic traits, *Scientific reports, 7(1): 12665.*

Pisanu, C., & Squassina, A. 2019. Treatment-Resistant Schizophrenia: Insights from Genetic Studies and Machine Learning Approaches. *Frontiers in pharmacology*, 10: 617.

Ranlund S, Joao M, Jong S, James H, Kyriakopoulos M, Cynthia H, Mitul A, Dima D. 2018. Associations between polygenic risk scores for four psychiatric illnesses and brain structure using multivariate pattern recognition. *Neuroimage Clinical, Elsevier,* Vol 20, pages 1026-1036.

Reisberg, S., Iljasenko, T., Läll, K., Fischer, K., & Vilo, J. 2017. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PloS one*, 12(7): e0179238.

Ripke S, Baker E, Escott V, et al. 2019. T22INVESTIGATION OF PATHWAY-BASED POLYGENIC RISK SCORES USING MACHINE LEARNING PREDICTION AND CLASSIFICATION SCHEMES. *European Neuropsychopharmacology*, Vol 29, Supplement 5, pages S229-S230.

Shapire, R. E. & Freund, Y. 2012. Boosting: Foundations and algorithms MIT Press, Cambridge (2012)

Shieh, Y., Hu, D., Ma, L., Huntsman, S., Gard, C. C., Leung, J., Tice, J. A., Ziv, E., Kerlikowske, K., & Cummings, S. R. 2017. Joint relative risks for estrogen receptor-positive breast cancer from a clinical model, polygenic risk score, and sex hormones. *Breast cancer research and treatment,* 166(2): pages 603–612.

Telenti, A., Lippert, C., Chang, P. C., & DePristo, M. 2018. Deep learning of genomic variation and regulatory network data. *Human molecular genetics*, 27(R1): R63–R71.

Torkamani A., Topol E., 2019. Polygenic Risk Scores Expand to Obesity. *Cell*, Vol 177, Issue 3, pages 518-520.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., … Zhao, S. 2019. Applications of machine learning in drug discovery and development. *Nature reviews. Drug discovery*, 18(6): pages 463–477.

World Health Organization, 2018. Genes and noncommunicable diseases. Genes and human diseases.

Wray, N. R., Goddard, M. E., & Visscher, P. M. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome research*, 17(10): pages 1520–1528.

Wu, Q., Boueiz, A., Bozkurt, A., Masoomi, A., Wang, A., DeMeo, D. L., Qiu, W. 2018. Deep Learning Methods for Predicting Disease Status Using Genomic Data. *Journal of biometrics & biostatistics*, 9(5): 417.

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., Hua, Y., Gueroussov, S., Najafabadi, H. S., Hughes, T. R., Morris, Q., Barash, Y., Krainer, A. R., Jojic, N., Scherer, S. W., Blencowe, B. J., & Frey, B. J. 2015. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science (New York, N.Y.),* 347(6218): 1254806.

Zhang YD, Wang J, Wu CJ, Bao ML, Li H, et al. 2016. An imaging-based approach predicts clinical outcomes in prostate cancer through a novel support vector machine classification. *Oncotarget*, Vol. 7(47): pages 78140–78151.

Zhou, J., & Troyanskaya, O. G. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods,* 12(10): pages 931–934.