For reprint orders, please contact reprints@futuremedicine.com

Commentary

Future Medicine AI

Role of genome-wide association studies, polygenic risk score and AI/ML using big data for personalized treatment to the patients with cardiovascular disease

Habiba Abdelhalim[1], Rachel-Mae Hunter[1], William DeGroat[1], Dinesh Mendhe[1], Saman Zeeshan[2] & Zeeshan Ahmed*,[1,3]

[1]Rutgers Institute for Health, Health Care Policy & Aging Research, Rutgers University, 112 Paterson St, New Brunswick, NJ 08901, USA

[2]Rutgers Cancer Institute of New Jersey, Rutgers University, 195 Little Albany St, New Brunswick, NJ 08901, USA

[3]Department of Medicine/Cardiovascular Disease & Hypertension, Robert Wood Johnson Medical School, Rutgers Biomedical & Health Sciences, 125 Paterson St, New Brunswick, NJ 08901, USA

*Author for correspondence: zahmed@ifh.rutgers.edu

> “we need portable AI/ML pipelines integrating genomics and healthcare data for discovering novel biomarkers and predicting CVD with high accuracy to support clinical diagnostics and decision-making processes.”

Tweetable abstract: Role of GWAS, polygenic risk score and AI/ML using big data for personalized treatment to the patients with cardiovascular disease

First draft submitted: 11 September 2023; Accepted for publication: 31 October 2023; Published online: 9 November 2023

Keywords: cardiovascular disease • GWAS • polygenic risk score • precision medicine • single nucleotide variants

Big data, GWAS, polygenic risk scores & AI/ML

Big data is a term used to describe extensive and wide data sets that are very intricate, convoluted, multifaceted and cannot be manually analyzed due to human error. Before big data, information was traditionally stored in small copious amounts, making it localized to the individual who inputs these datasets into their system [1]. The principles that define big data can be used in quality improvement under the guise of individual genomic data, such as the sequencing of DNA, RNA and their characteristics, and are therefore tied to clinical decision-making in examples such as personalized medicine [2]. Advancements in technology are driving the growth of personalized medicine, resulting in a parallel expansion of genomic medicine. This expansion aims to enhance individual diagnostics, ultimately leading to a reduction in personal side effects [3,4]. Previously, we reported the significance of an integrated approach that combines gene variant and clinical data [5]. We employed analyses of functional mutations, splice variants, variant distribution and divergence to uncover the importance and prevalence of variants linked to well-studied genes associated with heart failure (HF) and cardiovascular disease (CVD) [5]. Additionally, we have conducted comparative studies where we explored gene identification through multi-ethnic and ancestry-specific studies, how multiple single nucleotide polymorphisms (SNPs) are related to single disease-associated genes, and the correlation of specific biomarkers to both HF and atrial fibrillation (AF) [6]. To advance cardiovascular genomic medicine toward a predictive and preventive paradigm, it is imperative to precisely evaluate disease risk, effectively communicate variant findings, and establish clinical interventions aimed at averting or mitigating the associated ailments. Through a deep understanding of an individual's entire genome, we can leverage artificial intelligence (AI) and machine learning (ML) models to create a more refined approach for managing patients with CVD.

To understand big data in fields of health professions and research, a consensus for the definition that was widely accepted was coined by the Health Directorate of the Directorate-General for Research and Innovation at

Future Medicine

the European Commission (EC). A vast array of biological, clinical, environmental, and lifestyle data is collected from individuals at various time points to highlight their health status leading to the emergence of big data [2,7]. Through the usage of supercomputers, big data is understood to have the capability to recognize various trends and patterns into digestible knowledge for genomics for disease prediction in the field of precision medicine [3]. The utilization of big data is further identified to be useful in understanding the pathogenesis of CVD [8]. While big data offers numerous benefits for advancing genomics in healthcare, and precision medicine, it is also associated with significant limitations including but not limited to the inability for it to be standardized and transferred within other databases, claims that it can be "*incomplete, inaccurate or missing*" [3,8], and the idea that big data is not user-friendly, meaning that the average clinician would not be able to readily interpret it without training beforehand, privacy concerns, as well as owning a supercomputer in their practice which can be very expensive, which therefore leads to more precise techniques practiced using genome-wide association studies (GWAS) [9].

GWAS supports precision medicine by aiding in disease prediction of CVD [9]. GWAS is an observational study that is used to evaluate single nucleotide variants (SNV) throughout an entire genome in an individual [9,10]. This technology is used to compare the potential risk factors of an individual based on a reference genome and observe if there are any variants associated with the traits throughout the entire genome [9,10]. Additionally, GWAS identifies single nucleotide polymorphisms (SNPs) and different phenotypes to further explain CVD heritability. The significance of the SNVs can be analyzed and weighed to account for the effect of alleles at varied loci within the genome, which is also known as its linkage disequilibrium [9]. Through the utilization of GWAS, there have been a multitude of SNVs that can be linked to CVD [11]. By rapidly scanning genetic markers throughout an individual's genome, finding those variations can aid in calculating a polygenic risk score (PRS) to make predictions based on those mutations. Essentially, PRS is the weighted estimate of disease associated SNVs, is used as a tool to predict common, complex diseases such as CVD [11]. PRS integrates and aggregates the effects of multiple SNPs and SNVs across the genome into a single composite score to predict disease risk outcomes [11]. The SNPs that are derived from GWAS are calculated based on the weights of each nucleotide with a specific risk allele in an individual. PRS can be utilized in predicting drug efficiency, predicting cardiovascular reactions to certain drugs in individuals, and personalized drug therapy [3,11]. While these state-of-the-art technologies have greatly enhanced the field of precision medicine, there are some limitations associated with them. GWAS can inadvertently implicate genes that lack biological relevance to disease predisposition [10]. GWAS reveals additional limitations, including challenges related to addressing only a portion of missing heritability, difficulty in identifying complex traits, and pinpointing precise SNVs [10]. PRS has its own set of challenges, primarily in its inherent simplicity [12]. This risk estimation technique, due to its specificity, struggles to predict complex traits like rare variants in SNPs and exhibits a limited range of transferability across different populations [11].

AI/ML offers multiple supervised and unsupervised algorithms to analyze genomic data with the potential for learning from a continuum of dataset displaying heterogeneous levels of granularity. Recently, we have conducted and published important review studies, where we reported evaluation and comparative analysis of various bioinformatics and AI/ML approaches using the genomic data for state-of-the-art statistical and predictive analysis [6,13]. Conclusions of our studies included the support vector machine (SVM) and random forest (RF) as the most applied and successful AI/ML algorithms in the last few years [13]. Among the frequently employed AI/ML algorithms in genomics for bioinformatics, statistics and predictive analyses across a broad spectrum of diseases, RF and SVM stand out. We also established that a multitude of other predictive ML algorithms are also employed but less utilized including but not limited to artificial neural networks (ANN), k-nearest neighbors (k-NN) and gradient boosting [13]. Alternative AI/ML approaches exist, however, their adoption for the analysis of multi-genomic data remains limited. Most of these algorithms do not accommodate unprocessed sequence data for predictive analysis. Instead, these algorithms employ outcomes derived from enrichment, annotation, and pathway analysis to enable predictive modeling and risk stratification for variable diseases such as inflammatory bowel disease [14], colon cancer [15] and hypertension [16].

SVMs demonstrate prowess in managing datasets with a substantial number of features, rendering them apt for intricate data analyses. Their propensity for mitigating overfitting surpasses that of other ML algorithms, achieved by skillfully striking an equilibrium between optimizing the margin and reducing classification errors. SVMs boast versatility, being adaptable to diverse data types encompassing classification, regression and the identification of outliers. Moreover, SVMs have garnered a reputation for their robust generalization capabilities when faced with novel, unobserved data, thus endowing them with dependability in predictive modeling scenarios [13]. However, adjusting SVM hyperparameters is essential to prevent both overfitting and underfitting. Additionally, they may exhibit suboptimal performance when dealing with overlapping target classes or when the number of features per data point exceeds the quantity of training data samples [13]. RF is preferred over SVM when it comes to small datasets as it can provide predictions without the need for hyperparameter tuning [13]. RF is renowned for its proficiency in achieving remarkable accuracy across both classification and regression tasks, showcasing its strength in navigating intricate data relationships. This exceptional performance is fortified by its ensemble composition, featuring multiple decision trees that collectively diminish the susceptibility to overfitting, a common pitfall in ML. Through the amalgamation of predictions from these diverse trees, RF effectively reduces variance, consequently mitigating the peril of overfitting. Moreover, while the interpretability of RF models may not be as immediate as that of individual decision trees, they still offer valuable insights into feature importance, thereby enriching the overall interpretability of the model [13].

Gradient boosting and neural networks are also among the commonly used AI/ML algorithms in big data and predictive analyses for a wide variety of diseases. Deep neural network algorithms like k-NN and ANN are usually preferred over SVM when it comes to larger datasets [17,18]. This preference is rooted in the fact that these neural network algorithms are not only capable of handling vast amounts of data but also exhibit superior adaptability to diverse, high-dimensional data structures [19,20]. Additionally, their inherent ability to capture intricate, non-linear patterns makes them well-suited for complex real-world applications, including image recognition, natural language processing and recommendation systems [17,19]. Additionally, gradient boosting is preferred over RF due to its superior computational efficiency and predictive accuracy [21,22]. It excels in handling large and complex datasets, as a result of its parallel processing and optimization techniques, resulting in faster training and prediction times. This algorithm's superior predictive accuracy stems from its ability to optimize decision trees and combines the strengths of weak learners to create robust ensemble models, making it particularly suitable for tasks like regression, classification and ranking that demand precise predictions [23,24]. However, a limitation of neural networks and gradient boosting ML algorithms is their susceptibility to overfitting [17,22]. These algorithms can be tuned to work with the smaller and larger datasets, as well as also being implemented in tandem with CVD patients based on certain subgroups such as their age, gender, race and diagnosis [25]. While different CVD phenotypes such as hypertension, atherosclerosis, cardiomyopathy and HF can have variable clinical data, multi-genomic data exhibits a sustained consistency. The issue of data variability can be mitigated by adapting data into a standardized format compatible with AI/ML models. Recently, we have successfully integrated clinical and genomics data to accurately predict two separate clinical manifestation of CVDs such as HF and AF based on impactful biomarkers [25]. Additionally, we have concluded that an integrated approach to analyze multi-genomic, clinical and demographic data presented in an AI/ML-ready dataset with the combined utilization of a specialized ensemble of AI/ML algorithms will enable precise predictions regarding disease etiology through the identification of disease-associated biomarkers [26].

Going forward, we need portable AI/ML pipelines integrating genomics and healthcare data for discovering novel biomarkers and predicting CVD with high accuracy to support clinical diagnostics and decision-making processes. The potential implication will accelerate our ability to use AI/ML for discoveries and important breakthroughs in medical and life sciences with broad impact [26,27].

Author contributions

H Abdelhalim, R-M Hunter, W DeGroat, D Mendhe and S Zeeshan, participated in research. Z Ahmed proposed and led the study. H Abdelhalim and Z Ahmed drafted and revised manuscript. All authors have participated in writing and have approved manuscript for publication.

Acknowledgments

We appreciate great support by the Department of Medicine/Cardiovascular Disease and Hypertension, Robert Wood Johnson Medical School (RWJMS), Rutgers Biomedical and Health Sciences (RBHS), and Institute for Health, Health Care Policy and Aging Research (IFH), at Rutgers, The State University of New Jersey.

We thank members and collaborators of Ahmed Lab at Rutgers (RWJMS and IFH) for their support, participation and contribution to this study.

Financial disclosure

The authors have no financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Competing interests disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Writing disclosure

No writing assistance was utilized in the production of this manuscript.

Biographical note

H Abdelhalim is the Senior Research Assistant at the Ahmed lab, Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University-New Brunswick, NJ.

R-M Hunter is the Research Assistant at the Ahmed lab, Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University-New Brunswick, NJ.

D Mendhe is the lead software engineer at the Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, New Brunswick.

W DeGroat is the Research Assistant at the Ahmed lab, Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University-New Brunswick, NJ.

S Zeeshan is the visiting post-doctoral researcher at Rutgers Cancer Institute of New Jersey.

Z Ahmed is the Assistant Professor at the Department of Medicine/Cardiovascular Disease and Hypertension, Division of General Internal Medicine, Rutgers Robert Wood Johnson Medical School, which is the part of Rutgers Biomedical and Health Sciences. Dr. Ahmed is a Core Faculty Member at the Rutgers Institute for Health, Health Care Policy and Aging Research, at Rutgers, The State University of New Jersey.

Open access

References

1. Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. *J. Integr. Bioinform.* 15(3), 20170030 (2018).
2. Hassan M, Awan FM, Naz A *et al.* Innovations in Genomics and Big Data Analytics for Personalized Medicine and Health Care: A Review. *Int. J. Mol. Sci.* 23(9), 4645 (2022).
3. Ahmed Z, Zeeshan S, Lee D. Editorial: artificial intelligence for personalized and predictive genomics data analysis. *Front. Genet.* 14, 1162869 (2023).
4. Canzoneri R, Lacunza E, Abba MC. Genomics and bioinformatics as pillars of precision medicine in oncology. Genómica y bioinformática como pilares de la medicina de precisión en oncología. *Medicina (B Aires).* 79(Spec 6/1), 587–592 (2019).
5. Mhatre I, Abdelhalim H, Degroat W, Ashok S, Liang BT, Ahmed Z. Functional mutation, splice, distribution, and divergence analysis of impactful genes associated with heart failure and other cardiovascular diseases. *Sci. Rep.* 13(1), 16769 (2023).
6. Patel KK, Venkatesan C, Abdelhalim H *et al.* Genomic approaches to identify and investigate genes associated with atrial fibrillation and heart failure susceptibility. *Hum. Genomics.* 17(1), 47 (2023).
7. Auffray C, Balling R, Barroso I *et al.* Making sense of big data in health research: towards an EU action plan [published correction appears in Genome Med. 2016 Nov 7;8(1):118]. *Genome Med.* 8(1), 71 (2016).
8. Abdelhalim H, Berber A, Lodi M *et al.* Artificial Intelligence, Healthcare, Clinical Genomics, and Pharmacogenomics Approaches in Precision Medicine. *Front. Genet.* 13, 929736 (2022).
9. Wagle AA, Isakadze N, Nasir K, Martin SS. Strengthening the Learning Health System in Cardiovascular Disease Prevention: Time to Leverage Big Data and Digital Solutions. *Curr. Atheroscler. Rep.* 23(5), 19 (2021).
10. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20(8), 467–484 (2019).
11. O'Sullivan JW, Raghavan S, Marquez-Luna C *et al.* Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation* 146(8), e93–e118 (2022).
12. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19(9), 581–590 (2018).
13. Vadapalli S, Abdelhalim H, Zeeshan S, Ahmed Z. Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. *Brief Bioinform.* 23(5), bbac191 (2022).
14. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. *Nature* 474(7351), 307–317 (2011).

15. Cappell MS. Pathophysiology, clinical presentation, and management of colon cancer. *Gastroenterol. Clin. North Am.* 37(1), 1-v (2008).

16. Held E, Cape J, Tintle N. Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data. *BMC Proc.* 10(Suppl. 7), 141–145 (2016).

17. Zou J, Han Y, So SS. Overview of artificial neural networks. *Methods Mol. Biol.* 458, 15–23 (2008).

18. Zhang Z. A gentle introduction to artificial neural networks. *Ann. Transl. Med.* 4(19), 370 (2016).

19. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117 (2015).

20. Zhang Z. Introduction to machine learning: k-nearest neighbors. *Ann. Transl. Med.* 4(11), 218 (2016).

21. González-Recio O, Jiménez-Montero JA, Alenda R. The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. *J. Dairy Sci.* 96(1), 614–624 (2013).

22. Ying J, Wang Q, Xu T, Lu Z. Diagnostic potential of a gradient boosting-based model for detecting pediatric sepsis. *Genomics* 113(1 Pt 2), 874–883 (2021).

23. Liu K, Chen W, Lin H. XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites [published correction appears in Mol Genet Genomics. 2021 Nov;296(6):1357]. *Mol. Genet. Genomics* 295(1), 13–21 (2020).

24. Parente DJ. PolyBoost: an enhanced genomic variant classifier using extreme gradient boosting. *Proteomics Clin Appl.* 15(2–3), e1900124 (2021).

25. Venkat V, Abdelhalim H, DeGroat W, Zeeshan S, Ahmed Z. Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine. *Genomics* 115(2), 110584 (2023).

26. Degroat W, Venkat V, Pierre-Louis W, Abdelhalim H, Ahmed Z. Hygieia: AI/ML pipeline integrating healthcare and genomics data to investigate genes associated with targeted disorders and predict disease. *Software Impacts* 16, 100493 (2023).

27. Degroat W, Abdelhalim H, Patel K, Mendhe D, Zeeshan S, Ahmed Z. Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. *bioRxiv* doi: 10.1101/2023.09.08.553995 (2023).