Multimodal data for systolic and diastolic blood pressure prediction: The hypertension conscious artificial intelligence

Quincy A. Hathaway,[a]* Naveena Yanamala,[b] and Partho P. Sengupta[b]

[a]Heart and Vascular Institute, West Virginia University School of Medicine, Morgantown, WV, USA
[b]Division of Cardiovascular Diseases, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ, USA

eBioMedicine 2022;84: 104261
Published online xxx
https://doi.org/10.1016/j.ebiom.2022.104261

Commentary for "Machine learning integration of multimodal data identifies key features of blood pressure regulation"

In 2010, it was estimated that 31.1% of adults worldwide had hypertension, with the majority residing in low- and middle-income countries (1.04 billion people) and disproportionately affected compared with those in high-income countries.[1] As of 2020, more than 670,000 deaths were linked to hypertension as a primary or contributing cause in the United States alone.[2] Hypertension-related deaths most commonly involve major cardiovascular events, such as cerebrovascular accidents and heart disease, two of the most common causes of death in the developed world.[2] The increasing morbidity and mortality of hypertension has resulted in stricter guidelines for blood pressure control, evidenced by the 2017 American College of Cardiology (ACC) and American Heart Association (AHA) guidelines; these guidelines categorize normal blood pressure as less than 120 mmHg systolic and 80 mmHg diastolic. Further, treatment goals by the International Society of Hypertension in 2020 and the World Health Organization in 2021 have been updated to reflect better management strategies. Ultimately, there remains a substantial need for new investigations aimed at understanding the underlying causes and contributing factors of hypertension.

In this issue of *eBioMedicine*, Louca *et al.* utilized an array of genetic, metabolic, laboratory, and demographic information to better predict systolic and diastolic blood pressure measurements in two separate populations (i.e., the TwinsUK and Qatari biobank cohorts).[3] This cross-sectional study evaluated features that included 891 single nucleotide polymorphisms, metabolomics (i.e., amino-acids, peptides, carbohydrates, energy intermediates, lipids, 127 nucleotides, cofactors and vitamins, and xenobiotics), blood chemistry (e.g., sodium, potassium, chloride), food frequency questionnaires, and basic demographic information. To effectively assess each feature and its contribution in predicting systolic and diastolic blood pressure measurements, the authors utilized a XGBoost algorithm, with 5-fold cross validation, splitting the TwinsUK dataset into a 80% training and 20% testing set.[3]

The study by Louca et al. highlights the utility of applying multiple, comprehensive biological and/or molecular assessments (multi-omics or multimodal data) to address a clinical problem.[3] To date, only a few studies have utilized multimodal data to understand blood pressure measurements,[4,5] with these approaches utilizing advanced statistical methods, not machine/deep learning models. The application of multimodal data has a distinct advantage when it comes to hypothesis generating research. Studies that assess only a single component of the disease have included those identifying single nucleotide polymorphisms and their estimated contribution toward developing hypertension and future comorbidities.[6] Comparatively, when providing a more holistic approach, such as through multimodal data, it is possible to delineate which factors (e.g., demographic, metabolic, genetic, etc.) are correlated with hypertension and to what degree each feature explains hypertension. For example, in the current study, the most important non-demographic factors were metabolic (i.e., dihomo-linolenate, cis-4-decenoyl carnitine) and serum chemistry based (i.e., lactate, chloride, urate, and creatinine), not genetic.

The use of machine and deep learning models for predicting blood pressure have included regularizing gradient boosting frameworks (i.e., the current study), random forest, support vector machines, and deep learning multi-layer perceptron.[7] Regardless of the machine or deep learning approach utilized, the most import factor is proper study design. The use of standardized guidelines for the development and implementation of machine learning algorithms are

DOI of original article: http://dx.doi.org/10.1016/j.ebiom.2022.104243

*Corresponding author at: PO Box 9227, Room 3033-A HSC North, 1 Medical Center Drive, Morgantown, WV 26506, USA.
E-mail address: qahathaway@mix.wvu.edu (Q.A. Hathaway).

becoming more prominent in both data[8] and imaging[9] driven approaches. The framework for properly developing and implementing machine and deep learning algorithms follows a general formula, including 1) *preprocessing*, such as imputation of data and transformation to logarithmic scales, 2) *feature selection*, correcting for overfitting if necessary and removing colinear variables, 3) *algorithm design*, through hyperparameter optimization, 4) *model training*, utilizing appropriate training, testing, and holdout datasets with variable weighting for underrepresented classes, and 5) *evaluation of the model*, through measures of accuracy, sensitivity, specificity, area under the curve, F1 score, etc.

The authors appropriately managed problems unique to their dataset. For instance, the machine learning algorithm uniquely handled the family clustered IDs to avoid introducing bias from twin subjects. As twin subjects share genetically identical information, the two genetically identical individuals were both assigned to either the training or testing set, to avoid potential data leakage if separated. Additionally, median imputation can become a problem for machine learning-based studies. For each study participant, data may not have been collected for each of the tested features in the study. If data for a feature is missing, imputation of data, either through the mean or median of the training cohort, can be applied to artificially assign values where needed. The top 50 features identified by Louca *et al.* only required imputation of less than 3.8% on average, which would be assumed to present minimal, if any, bias to the study. The authors also provided an appropriate external validation cohort (i.e., the Qatar biobank), as it contained 2,807 participants who were both racial/ethnically and gender diverse from the training cohort. This is important, as assessing the generalizability of a machine learning model is needed to justify persistence of feature performance and scalability to broader investigations in the future.[3]

What remains to be understood is whether the current multimodal data study, and future investigations, can better integrate information on factors influencing blood pressure into a risk score. This has been studied with single nucleotide polymorphisms,[10] but not in the context of large, aggregated datasets. While it is known that demographic-based features such as age and BMI significantly influence increases in blood pressure, the integration of other features, as highlighted in the current study, can enhance our understanding of predicting progression toward hypertension. Such as the pooled cohort equations for atherosclerotic cardiovascular disease (ASCVD), a risk score for calculating likelihood of progression toward hypertension, or those more susceptible to hypertensive-related comorbidities, would be useful in modifying screening and treatment practices.

Contributors

Literature search (QAH), writing-original draft (QAH), writing, review & editing (QAH, NY, PPS). All authors read and approved the final manuscript.

Declaration of interests

QAH is the Chief Science Officer for Aspirations LLC. PPS serves a consultant to Heart Sciences, Ultromics, and Kencor Health. The other authors have nothing to disclose.

References

1 Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. *Nat Rev Nephrol*. 2020;16(4):223–237.

2 *About Multiple Cause of Death, 1999-2020*. Atlanta, GA: Centers for Disease Control and Prevention; 2022. [cited August 19, 2022].

3 Louca P, Tran TQB, Toit C, et al. Machine learning integration of multimodal data identifies key features of blood pressure regulation. *eBiomedicine*. 2022. https://doi.org/10.1016/j.ebiom.2022.104243. In Press.

4 Drouard G, Ollikainen M, Mykkanen J, et al. Multi-omics integration in a twin cohort and predictive modeling of blood pressure values. *OMICS*. 2022;26(3):130–141.

5 Liu J, de Vries PS, Del Greco MF, et al. A multi-omics study of circulating phospholipid markers of blood pressure. *Sci Rep*. 2022;12(1):574.

6 Lip S, Padmanabhan S. Genomics of blood pressure and hypertension: extending the mosaic theory toward stratification. *Can J Cardiol*. 2020;36(5):694–705.

7 Nguyen TM, Le HL, Hwang KB, Hong YC, Kim JH. Predicting high blood pressure using DNA methylome-based machine learning models. *Biomedicines*. 2022;10(6):1–14. https://doi.org/10.3390/biomedicines10061406.

8 Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. *Circ Cardiovasc Qual Outcomes*. 2020;13(10):e006556.

9 Sengupta PP, Shrestha S, Berthon B, et al. Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): a checklist: reviewed by the American College of Cardiology Healthcare Innovation Council. *JACC Cardiovasc Imaging*. 2020;13(9):2017–2035.

10 Niiranen TJ, Havulinna AS, Langen VL, Salomaa V, Jula AM. Prediction of blood pressure and blood pressure change with a genetic risk score. *J Clin Hypertens*. 2016;18(3):181–186.