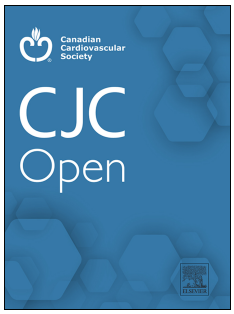
Journal Pre-proof



Assessing Machine Learning for Diagnostic Classification of Hypertension Types Identified by Ambulatory Blood Pressure Monitoring

Tran Quoc Bao Tran, MSc, Stefanie Lip, MBChB, Clea du Toit, MSc, Tejas Kumar Kalaria, MRCP, Ravi K. Bhaskar, MS, Alison Q. O'Neil, EngD, Beata Graff, MD PhD, Michał Hoffmann, MD PhD, Anna Szyndler, MD PhD, Katarzyna Polonis, PhD, Jacek Wolf, MD PhD, Sandeep Reddy, MBBS PhD, Krzysztof Narkiewicz, MD PhD, Indranil Dasgupta, DM, Anna F. Dominiczak, MD FMedSci, Shyam Visweswaran, MD PhD, Linsay McCallum, PhD, Sandosh Padmanabhan, MD PhD

PII: S2589-790X(24)00136-7

DOI: https://doi.org/10.1016/j.cjco.2024.03.005

Reference: CJCO 820

To appear in: *CJC Open*

Received Date: 14 December 2023

Revised Date: 12 February 2024

Accepted Date: 11 March 2024

Please cite this article as: T.Q. Bao Tran, S. Lip, C. du Toit, T.K. Kalaria, R.K Bhaskar, A.Q O'Neil, B. Graff, M. Hoffmann, A. Szyndler, K. Polonis, J. Wolf, S. Reddy, K. Narkiewicz, I. Dasgupta, A.F. Dominiczak, S. Visweswaran, L. McCallum, S. Padmanabhan, Assessing Machine Learning for Diagnostic Classification of Hypertension Types Identified by Ambulatory Blood Pressure Monitoring, *CJC Open* (2024), doi: https://doi.org/10.1016/j.cjco.2024.03.005.

Assessing Machine Learning for Diagnostic Classification of Hypertension Types Identified by Ambulatory Blood Pressure Monitoring

Tran Quoc Bao Tran MSc,*[1] Stefanie Lip MBChB,*[1] Clea du Toit MSc,*[1] Tejas Kumar Kalaria MRCP,[2] Ravi K Bhaskar MS,[3] Alison Q O'Neil EngD,[4] Beata Graff MD PhD,[5] Michał Hoffmann MD PhD,[5] Anna Szyndler MD PhD,[5] Katarzyna Polonis PhD,[5] Jacek Wolf MD PhD,[5] Sandeep Reddy MBBS PhD,[6] Krzysztof Narkiewicz MD PhD,[5] Indranil Dasgupta DM,[2] Anna F. Dominiczak MD FMedSci,[1] Shyam Visweswaran MD PhD,[7] Linsay McCallum PhD,[1] Sandosh Padmanabhan[1] MD PhD

*Equal contribution

Short title: Machine Learning in ABPM classification

Affiliations

[1]School of Cardiovascular and Metabolic Health, University of Glasgow, Glasgow, G12 8TA, UK

[2]University Hospitals Birmingham NHS Foundation Trust, Heartlands Hospital, Bordesley Green East, Birmingham B9 5SS, UK

[3]Mabgenex, Hyderabad, TA 500039, India

[4]Canon Medical Research Europe, Bonnington Bond, 2 Anderson Place, Edinburgh, EH6 5NP, UK

[5]Department of Hypertension and Diabetology, Medical University of Gdańsk, Mariana Smoluchowskiego 17, 80-214 Gdańsk, Poland

[6]School of Medicine, Deakin University, Australia

[7]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

Address for correspondence

Sandosh Padmanabhan,

School of Cardiovascular and Metabolic Health,

University of Glasgow,

Glasgow, G12 8TA, UK

Email: sandosh.padmanabhan@glasgow.ac.uk

Telephone: +44 141 330 2228

Abstract

Background

Inaccurate blood pressure classification results in inappropriate treatment. We tested if machine learning (ML), using routine clinical data, can serve as a reliable alternative to Ambulatory Blood Pressure Monitoring (ABPM) in classifying blood pressure status.

Methods

This study employed a multi-centre approach involving three derivation cohorts from Glasgow, Gdańsk, and Birmingham, and a fourth independent evaluation cohort. ML models were trained using office BP, ABPM, and clinical, laboratory, and demographic data, collected from patients referred for hypertension assessment. Seven ML algorithms were trained to classify patients into five groups: Normal/Target, Hypertension-Masked, Normal/Target-White-Coat, Hypertension-White-Coat, and Hypertension. The 10-year cardiovascular outcomes and 27-year all-cause mortality risks were calculated for the ML-derived groups using the Cox proportional hazards model.

Results

Overall XGBoost showed the highest AUROC of 0.85-0.88 across derivation cohorts, Glasgow (n=923; 43% females; age 50.7±16.3 years), Gdańsk (n=709; 46% females; age 54.4±13 years), and Birmingham (n=1,222; 56% females; age 55.7±14 years). But accuracy (0·57-0·72) and F1 scores (0·57-0·69) were low across the three patient cohorts. The evaluation cohort (n=6213, 51% females; age 51.2±10.8 years) indicated elevated 10-year risks of composite cardiovascular events in the Normal/Target-White-Coat and Hypertension-White-Coat groups, with heightened 27-year all-cause mortality observed in all groups except Hypertension-Masked, compared to the Normal/Target group.

Conclusions

Machine learning has limited potential in accurate blood pressure classification when ABPM is unavailable. Larger studies including diverse patient groups and different resource settings are warranted.

Introduction

Clinical guidelines now recommend out-of-office blood pressure (BP) measurements using ambulatory BP monitoring (ABPM) to screen and diagnose hypertension and to monitor on-treatment BP control, as clinic or office (oBP) is prone to error.[1-3] Compared to oBP, ABPM is a superior predictor of hypertension-mediated organ damage, cardiovascular disease (CVD) morbidity and mortality,[3] and can additionally identify patients with sustained hypertension, white-coat (WC) hypertension (BP overestimation by oBP), and masked hypertension (BP underestimation by oBP). WC and masked hypertension comprise at least a third of the at-risk population and masked hypertension is more prevalent in populations of African ancestry.[4] WC and masked hypertension are associated with an increased risk of progression to sustained hypertension,[5] highlighting the importance of early identification for more intensive follow-up even if pharmacological therapy is not initiated. However, ABPM has several drawbacks, including limited access to devices (often only available in secondary care); significantly higher device costs compared to oBP devices; staff training; patient discomfort; and sleep disruption. [6,7] Any method that can infer true hypertension status without incurring the expense and discomfort of using ABPM will enable effective and timely management in both primary and secondary care. We hypothesised that machine learning applied to routine clinical data can replace the need for ambulatory blood pressure monitoring. Our primary objective was to develop an ML algorithm capable of accurately classifying patients without ABPM measurements into distinct BP risk groups that are as informative as classifying with an ABPM. Our secondary objective was to demonstrate clinical utility of the ML classification by comparing the risk of CVD morbidity and mortality between these groups. The more extensive evaluation of all the additional information that an ABPM provides (such as BP variability, circadian variation) was beyond the scope of this study. A validated ML tool that

performs well in different settings and free from algorithmic bias will help hypertension management in both resource-poor and resource-rich settings allowing clinicians and nurses to effectively manage hypertension by risk-stratifying hypertensive patients, without incurring the added cost or inconvenience to both clinic and patients.

Methods

Study datasets

This is a retrospective study of patients referred for hypertension assessment to the Glasgow Blood Pressure Clinic (GBPC) [8] at the Queen Elizabeth University Hospital, Glasgow, UK between 2017 and 2019 (Glasgow cohort), patients participating in the CARE NORTH study, a prospective study of hypertensive patients from the outpatient specialist clinic at the Medical University of Gdańsk, Poland (Gdańsk cohort), and patients attending hypertension clinics at Birmingham Heartlands Hospital, Birmingham, UK between 2001 and 2020 (Birmingham cohort). A temporally distinct cohort for ML model evaluation was extracted from the GBPC database comprising patients referred to the clinic between 1985 and 2011 and followed up until 2013 (Glasgow non-ABPM cohort; non-overlapping with the Glasgow Cohort). Details of the cohorts and ethics statements are provided in the **Supplemental Appendix S1**.

BP Groups

Based on office systolic BP (oSBP) and the average systolic BP from 24-hr ABPM (aSBP), we defined five clinically relevant BP groups using BP oSBP and aSBP thresholds obtained from current US, European and UK guidelines[1-3,9]. These are described in **Table 1** and in the **Supplemental Appendix S1**.

If both oSBP and aSBP meet non-hypertensive thresholds, this indicates normal SBP in the absence of treatment or at target SBP if on treatment; these patients are grouped together as ‘Normal/Target’. The group ‘Hypertension’ defines sustained hypertension.

The group ‘Hypertension-Masked’ defines hypertensive aSBP in the presence of normal oSBP. Those with a WC effect are divided into two groups: normal aSBP in the presence of hypertensive oSBP are labelled ‘Normal/Target-WC’, whereas those with hypertensive aSBP and oSBP where oSBP is >=15 mmHg higher than aSBP are labelled ‘Hypertension-WC’. Any difference <15mmHg between the oSBP and aSBP of an individual with hypertensive oSBP and aSBP qualifies the individual for the group ‘Hypertension’. We assigned each patient to one of five BP groups based on the definitions in **Table 1** for the Glasgow, Gdańsk, and Birmingham cohorts. ML models were derived using these labels.

Clinical Features

Demographics, cardiovascular disease, antihypertensive therapy, and blood chemistry test results were used to develop the model (see **Supplemental Table S1**). At the time of ABPM or within a 3-month window preceding ABPM, features were obtained from standard clinical assessment of patients with hypertension in primary or secondary care who underwent ABPM. However, there were cohort-specific variations in the available features, reflecting differences in clinical practice in the real world. The Glasgow cohort has 14 features, the Gdańsk cohort has 12 features, and the Birmingham cohort has 10 features (see **Supplemental Tables S1-S4**).

Machine Learning

All data analysis and ML were performed in Python 3.8 using Scikit-learn, XGBoost, and Matplotlib software libraries. Missing data were imputed using the *k*-nearest neighbor (*k*NN) algorithm. The Shapiro-Wilk test was utilized to examine normality, and the Chi-square test was utilized to examine the independence of categorical variables.

We derived seven ML models from clinical features (excluding ABPM but including oBP measurements) including multinomial logistic regression (MLR), support vector machine

(SVM), *k*-nearest neighbor (*k*NN), naïve Bayes (NB), decision tree (DT), random forest (RF), and tree-based eXtreme Gradient Boosting (XGBoost).[10] Using stratified five-fold cross-validation, separate ML models were developed for each of the three cohorts due to the distinct features of each and performance metrics reported for each (see **Supplemental Appendix S1**).

Performance Metrics

Performance of models was reported using the area under the receiver operating characteristic curve (AUROC) and measures calculated from confusion matrices generated for each BP group: accuracy, precision (positive predictive value (PPV); proportion of relevant instances among the retrieved instances), recall (sensitivity), specificity, F1-score (the harmonic mean of precision and recall), negative predictive value (NPV), and number needed to misdiagnose (the number of patients who need to be tested in order for one to be misdiagnosed by the test; NNM). Calibration (the degree of similarity between observed and predicted probability) was assessed by calibration plot and Brier score as recommended by TRIPOD guidelines (see **Supplemental Appendix S1**).[11] The performance of a model was obtained by averaging the performance across all five folds. Results were reported as mean ± standard deviation (SD). In all cases, the significance level was $p<0.05$.

Predicting BP Group in the Glasgow non-ABPM Cohort

Models derived from the Glasgow, Gdańsk and Birmingham cohorts were used to predict model-specific BP groups for the 7,812 patients in the Glasgow non-ABPM cohort. Hypertension duration and number of antihypertensive medications were not available for the Glasgow non-ABPM cohort, so these features were omitted from the models.

Survival Analysis

Survival analysis was conducted on the Glasgow non-ABPM cohort by applying models derived from Glasgow, Gdańsk and Birmingham ABPM cohorts. The outcomes analyzed were all-cause mortality and CVD composite outcome (defined as time to first CVD admission or CVD death from myocardial infarction, ischemic heart disease, cerebrovascular accident, heart failure, or peripheral vascular disease) (see **Supplemental Appendix S1**). The patients were followed from their first BP clinic visit until death, emigration, or April 1, 2011 (the end of follow-up). Multivariable Cox proportional hazards models were used to assess the prognostic effect of the ML-derived BP groups on all-cause mortality and composite CVD events after adjustment for baseline variables, age, sex, body mass index (BMI), cholesterol, smoking status, Charlson comorbidity index, and a variable on year of the first visit strata (epochs) to adjust for secular trends in mortality. Schoenfeld residuals were used to test the proportional hazards assumption. Multiple imputation by chained equations (MICE) was performed for variables with <10% of values missing (BMI and cholesterol). Ten imputation datasets were generated, and pooled estimates from Cox regression are reported.

Results

Table 2 and **Supplemental Tables S2-S4** summarizes the clinical, laboratory, and demographic characteristics of the three study datasets. The proportion of patients on antihypertensive medication varied across cohorts (54% in Glasgow, 96% in Gdańsk, and 42% in Birmingham). Glasgow's cohort had the highest oBP and ABPM readings, while Gdańsk's cohort had the lowest. In **Supplemental Tables S2-S4**, the distribution of patients across the five BP groups and a summary of the clinical features of the three cohorts are presented. The BP groups with the smallest proportion of patients were Hypertension-Masked in the Birmingham (2.1%) and Glasgow (5.6%) cohorts and

Hypertension-WC in the Gdańsk cohort (6.8%). Normal/Target-WC in Glasgow (37.7%), Normal/Target-WC in Gdańsk (46.3%), and Normal/Target-WC in Birmingham (44.8%) cohorts had the highest proportion of patients.

Model Performance

The performance measures are presented in **Table 3 and Supplemental Table S5**. Table 3 shows the performance of the XGBoost machine learning model in classifying patients into distinct blood pressure groups, as defined ABPM data, which we consider the gold standard in this context. Overall, XGBoost and RF models with AUROC values between 0·85 and 0·88 had the highest performance (see **Supplemental Figures S1 and S2**). For XGBoost, accuracy ranged between 0·57 and 0·72 and F1 between 0·57 and 0·69 across the three patient cohorts. The simplest *k*NN (Glasgow, Gdańsk) and NB (Glasgow, Birmingham) models performed the worst. The confusion matrices and the corresponding performance metrics for XGBoost's classification of patients into the five BP groups in the three cohorts are depicted in **Table 3 and Supplemental Figure S3**. Across all three cohorts, XGBoost correctly classified 79% -99% of Normal/Target patients and 60% -80% of Normal/Target-WC patients. The classification accuracies for Hypertension and Hypertension-Masked were variable across the three cohorts with misclassification as Normal/Target-WC and Normal/Target respectively. **Supplemental Figure S4** depicts XGBoost calibration.

Feature Importance

The XGBoost model revealed that in all cohorts, oSBP, age, cholesterol, and creatinine were the most influential predictors (see **Supplemental Figure S5**). There were substantial variations in influential characteristics between models and between cohorts (see **Supplemental Figure S6**). XGBoost demonstrated a balanced contribution from all

features, whereas other models exhibited a predominance of one to three features in determining model performance (**Supplemental Figure S6**).

Survival Analysis

The demographic characteristics of the Glasgow non-ABPM cohort are summarized in **Supplemental Table S6**. The results of Cox regression analysis are displayed in **Table 4** and **Supplemental Figure S7**. A significantly higher 10-year risk of composite CVD events was evident for the Normal/Target-WC (H.R. 1·3 [1·09;1·55], P=0·003) and Hypertension-WC (1·38 [1·14;1·67], P=0·001) groups compared to Normal/Target when predicted by the model derived from the Glasgow cohort. When heterologous models, i.e., Birmingham or Gdańsk XGBoost models, were applied to the Glasgow non-ABPM cohort, the outcomes were largely consistent. Hypertension-WC groups predicted by Birmingham (1·42 [1·21;1·66] P<0·0001) and Gdańsk (1·33 [1·12;1·57] P=0·001) XGBoost models were associated with higher 10-year CVD risk compared to Normal/Target. The predicted Normal-Target WC groups did not show an increased risk of 10-year cardiovascular events for the Birmingham (1·16 [0·99;1·36] P=0·059) and Gdańsk (1·16 [0·96;1·39] P=0·127) cohorts. Interestingly, the Glasgow and Gdańsk models did not reveal an increased risk of CVD outcomes in the Hypertension group. Normal/Target-WC and Hypertension-WC groups showed significantly higher all-cause mortality compared to Normal/Target for all three models. The Hypertension group showed significantly increased mortality risk with the Glasgow and Gdańsk models. Hypertension-Masked showed significantly increased mortality risk only with the Glasgow model. For all-cause mortality outcomes, Normal/Target-WC (Glasgow: 1·42 [1·17;1·73], P=0·0004; Birmingham: 1·2[1·01;1·42], P=0·036; Gdańsk:1·25 [1·03;1·52], P=0·025) and Hypertension-WC (Glasgow:1·69[1·38;2·07], P<0·0001; Birmingham:1·5 [1·27;1·76], P<0·0001; Gdańsk: 1·39 [1·17;1·66], P=0·0002) showed significant associations

regardless of the classification model used. All-cause mortality for the Hypertension group did not attain statistical significance for the Birmingham model (Glasgow:1·33 [1·02;1·74], P=0·034; Birmingham: 1·44 [0·99;2·07], P=0·052; Gdansk: 1·23 [1·0;1·51], P=0·049) and Hypertension-Masked group showed an increased risk of death only with the Glasgow model (1·35 [1·01;1·82], P=0·046).

Discussion

In our proof-of-concept investigation, we explored the viability of using machine learning (ML) applied to routine clinical data as a substitute for Ambulatory Blood Pressure Monitoring (ABPM) in the clinical management of hypertension. Despite the modest accuracy of the machine learning classifications, our analysis of associations with clinical outcomes revealed that certain ML-derived categories correlate with a higher risk when compared to the reference category of normal/target BP. This observation suggests that even with current accuracy limitations, ML classifications can provide preliminary insights into risk stratification that merit further investigation. The implications for our results, despite the acknowledged low accuracies are two-fold. First, it demonstrates the potential clinical relevance of ML classifications in identifying risk groups, which could be refined and validated in larger, more diverse datasets. Second, it underscores the importance of enhancing accuracy and reliability in future research. The observed associations, despite the current limitations, indicate a promising direction for leveraging ML in hypertension management, especially in scenarios where ABPM is not feasible. An interesting insight from our study is the potential generalizability of the ML model regardless of the originating cohort in consistently identified individuals exhibiting the white coat effect with consequent increased cardiovascular risk. Our approach to classifying individuals into five risk groups differs from previous efforts, which primarily focused on deducing ABPM BP values from oBP and other clinical parameters.[12]

Our ML algorithms exhibited some degree of misclassification across all three cohorts, particularly between the Normal/Target and Hypertension-Masked groups and within the trio of Normal/Target-WC, Hypertension-WC, and Hypertension. While the models showed high AUROCs, these misclassifications resulted in suboptimal calibration metrics. XGBoost model performed better in the Glasgow cohort (40%) in the detection of Hypertension and Hypertension-Masked (54%) compared to Gdańsk and Birmingham cohorts (20% and 9% for Hypertension and 20% and 0% for Hypertension-Masked respectively). The relatively higher detection rate for the smaller Hypertension-Masked and Hypertension groups in the Glasgow cohort may be attributable to the Glasgow cohort having more input features (**Supplemental Table S1**) and a more balanced distribution of groups between the two cohorts (**Supplemental Tables S2-S4**). The advantage of XGBoost models is that they harmonized a range of covariates, including demographic and lab values, to construct the prediction with no single covariate predominating in contrast to logistic regression (see **Supplemental Figure S6**). Nevertheless, the development of ML models from unbalanced data can result in predictions that are skewed toward the majority group and deceptively high performance. Typically, this phenomenon is referred to as the accuracy paradox. Consequently, despite the Glasgow cohort's superior predictive ability for smaller subgroups, its model produced lower accuracy, F1, and AUROC scores than those of the Gdańsk and Birmingham cohorts. Performance matrices (**Supplemental Table S5**), which provide a more in-depth analysis of the model's classification accuracy, demonstrated that the Glasgow cohort's model has the best performance in detecting clinically significant smaller subgroups. Oversampling implicitly increases the prior probability of smaller subgroups, which could lead to misclassifications of patients from the majority group.

ABPM is considered the gold standard for diagnosing hypertension.[2,3,9,13] ABPM provides a more precise hypertension diagnosis, which benefits both the individual patient and the

healthcare system. Avoiding unnecessary antihypertensive treatment in patients has resulted in a 3-14% decrease in treatment costs. However, due to the requirements for technology and other resources, the immediate cost of ABPM is significantly higher than that of routine clinic BP measurement.[6,7] ABPM is also more taxing on patients since the BP cuff inflates and deflates at least twice or thrice hourly: a third of patients report pain or bruising during the 24-hour monitoring process, and two-thirds report sleep disturbances. The primary strengths of our study are the development of multiple ML models in three temporally and geographically distinct cohorts which strengthen the study's conclusions, the use of cross-validation to generate robust confidence intervals for our estimates, the demonstration of an impact on long-term outcomes, and adherence to the current TRIPOD-AI guidelines for reporting prediction models.[11] However, our study did not come without limitations. In addition to the misclassification and poor calibration metrics noted above, some of the BP groups were underrepresented due to the small sample size, most notably Hypertension-Masked, and the majority of patients were white Europeans with little ethnic diversity. We were unable to account for secular effects between cohorts in the survival analyses and used only the year of inclusion as a covariate to address this. The Glasgow non-ABPM cohort comprised hypertensive patients attending the Glasgow BP clinic – thus all were managed nearly similarly at a specialist clinic. The discrepancy in the CV outcomes risk seen in the hypertensive group and the white-coat groups compared to the normal/target BP group on applying models derived from the three cohorts on outcomes in the Glasgow non-ABPM cohort may reflect the existing unmet need where those patients with white coat effect may still be under-treated. We surmise further research using larger datasets will help determine the predictive power of the ML models and clarify the implications on long-term outcomes.

This study serves as a foundational proof of concept, demonstrating the potential of ML to categorize patients into clinically pertinent risk groups when ABPM is unavailable, though it underscores that ML cannot completely supplant ABPM at this juncture. Our findings advocate for the conduct of more extensive prospective studies, with the inclusion of a diverse array of racial and ethnic groups and a variety of resource settings, alongside ensuring adequate representation across all blood pressure classifications. Furthermore, the successful and ethical deployment of ML in healthcare, particularly in resource-poor settings, requires a participatory approach that includes patients, healthcare providers, and other stakeholders from these communities from the outset. In conclusion, it is essential to conduct a comprehensive evaluation to determine the true potential of machine learning as either a direct substitute for or a significant complement to ambulatory blood pressure monitoring. By integrating ML into the clinical pathway, we have the opportunity to revolutionize the management of hypertension, enabling timely and effective interventions across diverse healthcare environments.

Acknowledgements

The authors thank the patients of the Glasgow, Birmingham and Gdańsk cohorts.

Funding

SP and AFD are supported by the British Heart Foundation Centre of Excellence Award (RE/18/6/34217) and the UKRI Strength in Places Fund (SIPF00007/1).

Disclosures

AQO'N is employed by Canon Medical, SP and SV are co-founders of Kvatchii Ltd, SV is co-founder of READE.ai Inc, SR is the founder of Medi-AI, and RKB is on the Technical Advisory Board of Mabgenex.

Patient Consent Statement

The authors confirm that patient consent is not applicable to this article. This is a retrospective analysis using de-identified data; therefore, there is no ethics committee requirement to obtain individual consent from the patient.

Data availability

The training data from this study are accessible through the data access procedures of each study center upon request. Access may be granted to those who meet the criteria for confidential access, but data governance regulations prevent the data from being accessible to the public. The leads for each cohort (SP, KN, and ID) can provide guidance on data access procedures, which may necessitate separate applications to the respective health boards. The three XGBoost models, corresponding to three cohorts used in this study have been made freely available at https://github.com/Tran031194/abpmML.

References

1. NICE. *Hypertension in adults: diagnosis and management NICE Guideline [NG136]*. NICE; 2019.
2. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension*. Jun 2018;71(6):e13-e115. doi:10.1161/HYP.0000000000000065
3. Banegas JR, Ruilope LM, de la Sierra A, et al. Relationship between Clinic and Ambulatory Blood-Pressure Measurements and Mortality. *N Engl J Med*. Apr 19 2018;378(16):1509-1520. doi:10.1056/NEJMoa1712231
4. Diaz KM, Veerabhadrappa P, Brown MD, Whited MC, Dubbert PM, Hickson DA. Prevalence, Determinants, and Clinical Significance of Masked Hypertension in a Population-Based Sample of African Americans: The Jackson Heart Study. *Am J Hypertens*. Jul 2015;28(7):900-8. doi:10.1093/ajh/hpu241

5. Siven SS, Niiranen TJ, Kantola IM, Jula AM. White-coat and masked hypertension as risk factors for progression to sustained hypertension: the Finn-Home study. *J Hypertens*. Jan 2016;34(1):54-60. doi:10.1097/HJH.0000000000000750
6. Martin-Rioboo E, Perula de Torres LA, Banegas JR, et al. Knowledge, availability, and use of ambulatory and home blood pressure monitoring in primary care in Spain: the MAMPA study. *J Hypertens*. May 2018;36(5):1051-1058. doi:10.1097/HJH.0000000000001673
7. Kronish IM, Kent S, Moise N, et al. Barriers to conducting ambulatory and home blood pressure monitoring during hypertension screening in the United States. *J Am Soc Hypertens*. Sep 2017;11(9):573-580. doi:10.1016/j.jash.2017.06.012
8. Lip S, Tan LE, Jeemon P, McCallum L, Dominiczak AF, Padmanabhan S. Diastolic Blood Pressure J-Curve Phenomenon in a Tertiary-Care Hypertension Clinic. *Hypertension*. Oct 2019;74(4):767-775. doi:10.1161/HYPERTENSIONAHA.119.12787
9. Williams B, Mancia G, Spiering W, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology and the European Society of Hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology and the European Society of Hypertension. *J Hypertens*. Oct 2018;36(10):1953-2041. doi:10.1097/HJH.0000000000001940
10. Krittanawong C, Johnson KW, Rosenson RS, et al. Deep learning for cardiovascular medicine: a practical primer. *Eur Heart J*. Jul 1 2019;40(25):2058-2073. doi:10.1093/eurheartj/ehz056
11. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. Jan 6 2015;162(1):W1-73. doi:10.7326/M14-0698
12. Sheppard JP, Stevens R, Gill P, et al. Predicting Out-of-Office Blood Pressure in the Clinic (PROOF-BP): Derivation and Validation of a Tool to Improve the Accuracy of Blood Pressure Measurement in Clinical Practice. *Hypertension*. May 2016;67(5):941-50. doi:10.1161/HYPERTENSIONAHA.115.07108
13. O'Brien E, White WB, Parati G, Dolan E. Ambulatory blood pressure monitoring in the 21st century. *J Clin Hypertens (Greenwich)*. Jul 2018;20(7):1108-1111. doi:10.1111/jch.13275

Table Legends

Table 1: Definition of five BP groups. aSBP: ABPM 24-hr systolic blood pressure average; oSBP: office systolic blood pressure; BP: blood pressure; WC: white-coat.

Table 2: Summary of patient characteristics presented as mean ±SD, unless presented as proportion (%) of total cohort. Distribution of patients across five BP groups is also shown. Grey shading indicates absence of data. ABPM: ambulatory blood pressure monitoring; ALT: alanine aminotransferase; CVD: cardiovascular disease; DBP: systolic blood pressure; oBP: office blood pressure; SBP: systolic blood pressure; WC: white-coat.

Table 3: Summary of XGBoost performance in the three cohorts for each BP group. AUROC: area; PPV: positive predictive value (precision); WC: white coat. Sensitivity, Specificity, and PPV are reported for each fold.

Table 4: Results of Cox regression analysis for 27-year all-cause mortality and for 10-year composite cardiovascular events.

Table 1: Definition of five BP groups. aSBP: ABPM 24-hr systolic blood pressure average; BP: blood pressure; oSBP: office systolic blood pressure; WC: white-coat.

BP Group	oSBP (mmHg)	aSBP (mmHg)	oSBP-aSBP (mmHg)
Normal/Target	≤140	≤135	..
Hypertension-Masked	≤140	>135	..
Normal/Target-WC	>140	≤135	..
Hypertension-WC	>140	>135	≥15
Hypertension	>140	>135	..

Table 2: Summary of patient characteristics presented as mean ±SD, unless presented as proportion (%) of total cohort. Distribution of patients across five BP groups is also shown. Grey shading indicates absence of data. ABPM: ambulatory blood pressure monitoring; ALT: alanine aminotransferase; CVD: cardiovascular disease; DBP: systolic blood pressure; oBP: office blood pressure; SBP: systolic blood pressure; WC: white-coat.

	Glasgow	**Gdańsk**	**Birmingham**
	n=923	**n=709**	**n=1,222**
Demographics			
Age (years)	50·7±16·3	54·4±13·1	55·7±13·9
Sex (% female)	43·1	45·6	56·2
BP Measurements & Hypertension Status			
oBP – SBP (mmHg)	163·1±22·5	139·9±19·5	157·4±19·9
oBP – DBP (mmHg)	95·4±14·3	82·2±11·2	94·6±12·0
ABPM - SBP (mmHg)	138·0±15·9	129·1±12·3	132·3±13·1
ABPM - DBP (mmHg)	81·5±11·5	77·3±8·9	78·5±9·7
Hypertension Duration (years)	3·7±4·7	12·1±8·9	··
Prevalent CVD (%)	56·6	16·2	7·8
Antihypertensive Treatment (% treated)	54·0	96·1	42·2
BP Groups			
Normal/Target	96 (10·4%)	328 (46·3%)	210 (17·2%)
Hypertension-Masked	52 (5·6%)	75 (10·6%)	26 (2·1%)
Normal/Target-WC	348 (37·7%)	184 (26·0%)	547 (44·8%)
Hypertension-WC	249 (27·0%)	48(6·8%)	326 (26·7%)
Hypertension	178 (19·3%)	74 (10·4%)	113 (9·2%)
Blood Chemistry			
Sodium (mmol/L)	139·4+2·5	140·0±2·5	140·1±2·7
Potassium (mmol/L)	4·4±0·4	4·2±0·4	4·3±0·4
Creatinine (µmol/L)	75·7±19·3	73·6±13·1	80·7±17·7
Cholesterol (mmol/L)	5·4±1·2	5·0±1·1	5·4±1·1

Triglyceride (mmol/L)	1·9±1·2	1·5±0·8	..
ALT (U/L)	28·5±18·6	..	..
Bilirubin (μmol/L)	10·5±5·3	..	..

Table 3: Summary of XGBoost performance in the three cohorts for each BP group. AUROC: area; PPV: positive predictive value (precision); WC: white coat. Sensitivity, Specificity, and PPV are reported for each fold.

Hypertension Classification	AUROC	Accuracy	F1 Score	Sensitivity	Specificity	PPV
Glasgow (n=923)						
Normal/Target	0·877 ± 0·008	0·574 ± 0·032	0·570 ± 0·032	0·792 ± 0·116	0·971 ± 0·009	0·761 ± 0·048
Hypertension-Masked	0·877 ± 0·008	0·574 ± 0·032	0·570 ± 0·032	0·535 ± 0·182	0·977 ± 0·013	0·587 ± 0·141
Normal/Target-WC	0·877 ± 0·008	0·574 ± 0·032	0·570 ± 0·032	0·604 ± 0·113	0·69 ± 0·014	0·54 ± 0·048
Hypertension-WC	0·877 ± 0·008	0·574 ± 0·032	0·570 ± 0·032	0·582 ± 0·086	0·84 ± 0·045	0·574 ± 0·092
Hypertension	0·877 ± 0·008	0·574 ± 0·032	0·570 ± 0·032	0·399 ± 0·142	0·915 ± 0·031	0·532 ± 0·185
Gdańsk (n=709)						
Normal/Target	0·938 ± 0·004	0·664 ± 0·036	0·729 ± 0·031	0·969 ± 0·051	0·816 ± 0·027	0·82 ± 0·017
Hypertension-Masked	0·938 ± 0·004	0·664 ± 0·036	0·729 ± 0·031	0·067 ± 0·119	0·984 ± 0·026	0·45 ± 0·714
Normal/Target-WC	0·938 ± 0·004	0·664 ± 0·036	0·729 ± 0·031	0·913 ± 0·054	0·819 ± 0·028	0·639 ± 0·038
Hypertension-WC	0·938 ± 0·004	0·664 ± 0·036	0·729 ± 0·031	0·371 ± 0·466	0·983 ± 0·02	0·654 ± 0·523
Hypertension	0·938 ± 0·004	0·664 ± 0·036	0·729 ± 0·031	0·107 ± 0·16	0·991 ± 0·025	0·69 ± 0·623
Birmingham (n=1,222)						
Normal/Target	0·937 ± 0·010	0·692 ± 0·047	0·724 ± 0·045	0·986 ± 0·038	0·974 ± 0·004	0·888 ± 0·014
Hypertension-Masked	0·937 ± 0·010	0·692 ± 0·047	0·724 ± 0·045	0·0 ± 0·0	0·997 ± 0·007	0·0 ± 0·0
Normal/Target-WC	0·937 ± 0·010	0·692 ± 0·047	0·724 ± 0·045	0·812 ± 0·081	0·716 ± 0·085	0·699 ± 0·09
Hypertension-WC	0·937 ± 0·010	0·692 ± 0·047	0·724 ± 0·045	0·687 ± 0·134	0·884 ± 0·038	0·684 ± 0·069
Hypertension	0·937 ± 0·010	0·692 ± 0·047	0·724 ± 0·045	0·089 ± 0·101	0·989 ± 0·016	0·521 ± 0·538

Table 4: Results of Cox regression analysis for 27-year all-cause mortality and for 10-year composite cardiovascular events.

	Glasgow XGBoost Model	**Birmingham XGBoost Model**	**Gdańsk XGBoost Model**
27-year all-cause mortality			
Normal/Target	1	1	1
Normal/Target-WC	1·42 [1·17;1·73] P=0·0004	1·2 [1·01;1·42] P=0·036	1·25 [1·03;1·52] P=0·025
Hypertension-Masked	1·35 [1·01;1·82] P=0·046	0·4 [0·06;2·86] P=0·362	0·85 [0·59;1·22] P=0·365
Hypertension - WC	1·69 [1·38;2·07] P<0·0001	1·5 [1·27;1·76] P<0·0001	1·39 [1·17;1·66] P=0·0002
Hypertension	1·33 [1·02;1·74] P=0·034	1·44 [0·99;2·07] P=0·052	1·23 [1·0;1·51] P=0·049
10-year composite cardiovascular events			
Normal/Target	1	1	1
Normal/Target-WC	1·3 [1·09;1·55] P=0·003	1·16 [0·99;1·36] P=0·059	1·16 [0·96;1·39] P=0·127
Hypertension-Masked	1·03 [0·77;1·37] P=0·853	1·47 [0·47;4·6] P=0·509	0·83 [0·60;1·14] P=0·24
Hypertension - WC	1·38 [1·14;1·67] P=0·001	1·42 [1·21;1·66] P<0·0001	1·33 [1·12;1·57] P=0·001
Hypertension	1·05 [0·81;1·35] P=0·708	1·4 [0·99;1·98] P=0·058	1·13 [0·93;1·37] P=0·223