Predicting hypertension onset using logistic regression models with labs and/or easily accessible variables: the role of blood pressure measurements

Chiara Roversi[1], Martina Vettoretti[2], Barbara Di Camillo[3], and Andrea Facchinetti[4]
[1,2,3,4]*Department of Information Engineering,* [4]*Department of Comparative Biomedicine and Food Science*
University of Padova
Padova, Italy
[1]roversic@dei.unipd.it, [2]martina.vettoretti@unipd.it, [3]barbara.dicamillo@unipd.it, [4]andrea.facchinetti@unipd.it

***Abstract*—Hypertension is a critical condition that represents a leading risk factor for mortality. The identification of subjects at risk of developing hypertension is important to improve life expectancy and reduce the burden of healthcare systems. Available models to predict hypertension onset in some years in the future mainly include blood pressure (BP) measurements as well as blood test and lifestyle variables. However, systolic and diastolic BP are inevitably strong predictors of the disease and their presence in such models may hide a possible key role of other covariates. The aim of this work is to develop predictive models of hypertension onset both with and without the use of BP measurements to investigate if and how BP variables influence the feature selection process. By involving a large dataset on individuals socio-economic status, demographics, wellbeing, lifestyle, medical history and blood exams, logistic regression models (w/ and w/o BP) have been trained using a stepwise selection procedure to select only highly predictive variables. The model with systolic and diastolic BP selected as important variables HDL cholesterol, hemoglobin, marital status, depression scale and alcohol drinking, achieving an area under the receiver-operating characteristic curve (AU-ROC) of 0.80. The model without BP variables exploits heart rate, waist, age and marital status, and achieves AU-ROC=0.74. As expected, the model employing BP measurements performs better than the one that does not consider them. However, also without BP, it was possible to develop a model with satisfactory performance involving only easily accessible information that do not require laboratory tests.**

***Index Terms*—hypertension, risk factors, preventive medicine, predictive model, logistic regression**

I. Introduction

Hypertension is one of the major economic and public health burden worldwide due to its high prevalence and concomitant risk of cardiovascular problems including stroke, coronary heart disease, cardiac failure, and renal disease. Moreover, it has been identified as the leading risk factor for mortality [1]. Avoiding or even delaying the incidence of hypertension with, e.g., lifestyle modification has been widely demonstrated [2], [3].

Being able to prevent such condition is an important goal. Multivariable predictive models of hypertension onset can be useful to understand which factors can play a role in the disease development, helping healthcare providers and clinicians in designing prevention strategy and identifying individuals at high risk. Several literature models predicting near-term and/or long-term incidence of hypertension were developed [4], which mainly include variables related to blood test exams and lifestyle behaviours, together with measurements of systolic and diastolic blood pressure (BP) collected at the baseline visit. However, BP measurements, being very strong predictors of hypertension onset due to their central role in defining the disease diagnosis, can hide the importance of other predictor variables.

In this work, we will develop models to predict the incidence of hypertension 4 years after the baseline in two main scenarios: Scenario 1 that includes, among the pool of possible covariates, BP measurements; Scenario 2, in which BP measurements are not present. The aim of our work is two-fold: investigating the different relationships that occur among risk factors and the outcome, as well as the possibility to predict with reasonable accuracy, when systolic and diastolic BP are not available. In the specific, to address our questions we will employ logistic regression models with a robust stepwise selection procedure based on a bootstrap sampling. To achieve our aims, a large longitudinal dataset on English adults aged 50 and older (the English Longitudinal Study of Ageing, ELSA) will be employed, allowing to investigate, in the same analysis, the predictive power of not only lifestyle indicators, blood test biomarkers and physical measurements, widely used in the literature models, but also of variables related to the socio-economic, wellbeing and home-environmental status.

II. Material

A. The English Longitudinal Study of Ageing

The ELSA is an ongoing longitudinal study of health, quality of life and socio-economic status in the English population aged 50 years and older [5]. Since the study start in 2002, participants undergo an interview about every 2 years and a clinical examination about every 4 years. Currently, data of eight interviews (hereafter labelled as "waves"), covering a period of 15 years (2002-2017), are available. At waves 3, 4,

This research was partially funded by the initiative "Departments of Excellence" of the Italian Ministry of Education, University and Research (Law 232/2016).

TABLE I
VARIABLES SELECTED FOR THE PREDICTION OF HYPERTENSION ONSET

Category	Variable	Values
Demographics	Sex	1=male, 0=female
	Age	Continuous [years]
	Marital status	0=married or living as married 1=divorced, separated or windowed 2=never married
	Immigrant status	0=born in living country 1=born outside living country
Lifestyle	Smoking	0=no, 1=past, 2=current smoker
	Alcohol drinking	0=never, 1=moderate, 2=frequent
	Moderate or vigourous physical activity	0=hardly ever or never 1=1-3 times/week 2=once/week 3= >once/week
Physical measurements	Waist circumference	Double [cm]
	Systolic blood pressure	Double [mmHg]
	Diastolic blood pressure	Double [mmHg]
	Heart rate	Double [beats/min]
Blood test biomarkers	Ferritin	Double [L/mL]
	Hemoglobin	Double [g/dL]
	Fibrinogen	Double [g/L]
	Triglycerides	Double [mg/dL]
	Total cholesterol	Double [mg/dL]
	Hdl cholesterol	Double [mg/dL]
	C-reactive protein	Double [mg/L]
Medical history	History of diabetes	0=no, 1=yes
	History of arthritis	0=no, 1=yes
Socio-economic	Employment status	0=homemaker, 1=employed, 2=unemployed/retired
	Deprivation	0=never, 1=rarely, 2=sometimes, 3=often, 4=most of the time
Wellbeing	Depression score	Integers, range 1-8
	Self-reported health	Integers, range 0-4
	Life expectation	Integers, range 1-100
Home environment	Accomodations problems	Integers, range 0-10

6 and 7 the study was replenished with new participants, to maintain the size and representativeness of the panel.

B. Dataset preprocessing and variables selection

Since data on participants' clinical examinations are available only at even waves, for each subject we considered as baseline wave the first participated visit among waves 2, 4 and 6 (not wave 8 since no follow-up after this visit would be present). The final selected sample includes subjects without hypertension at baseline wave and without missing values, i.e. only subjects with complete baseline visit are considered. A subject developed hypertension if she/he responded "yes" to the question "Has a doctor ever told you that you have high BP or hypertension?". According to this information, a binary variable, indicating the incidence of hypertension, was created. In details, such variable was set equal to "1" if the subject developed hypertension during the 4-years observation period and equal to "0" if the subject did not report hypertension in the same period. A total of 3399 subjects were considered, 435 of whom developed the disease during the follow-up period after the baseline. The age of the selected sample has median [interquartile range] of 60 [55, 68] years.

From the set of variables collected in ELSA, 26 of them, potentially predictive for hypertension onset, were selected. Such variables can be grouped in 8 different categories: demographics, lifestyle habits, physical measurements, blood test biomarkers, medical history, socio-economic aspect, well-being and home environment status. The full list of variables is reported in Table I. In particular, economic deprivation was measured by the question "How often you have too little money to spend on personal and household needs?"; depression level was quantified by the CESD scale, where higher values represent higher levels of depression symptoms; self-reported health status varies from 0 to 4, where 0 means excellent and 4 means poor; life expectation represents the self-reported probability of living to a specific age (e.g. to 75 years if the respondent is under 65, etc); accomodation problems were defined as the number of problems in the current house (e.g. damp, noise, condensation, shortage of space, etc).

III. METHODS

A. Model development

Data were randomly split into a training and test set, including the 80% and 20% of selected subjects, respectively. The split, performed by stratifying for incidence of hypertension, allowed to obtain balanced values between training and test sets for all the considered variables. The training set contained 2711 subjects, with 356 positive cases of hypertension during follow-up; whereas the test set contained 688 subjects, with 79 positive cases.

Then, a logistic regression model was fit on the training set:

$$log\big(\frac{p}{1-p}\big) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \qquad (1)$$

where p is the probability of hypertension onset during the observation period, $X_1, ...X_p$ the p predictor variables, $\beta_1 ..., \beta_p$ the corresponding coefficients and β_0 the model intercept [6]. A maximum likelihood estimation procedure was used to identify the model coefficients β_i. Then, for each coefficient, a t-statistic was computed to test if the corresponding coefficient was statistically different from zero, adopting 5% as significance level.

To reduce the complexity of the model and select the most important predictors, a stepwise variable selection with bidirectional elimination [6] was applied. The stopping rule was based on the p-value of an F-test or chi-squared test of the change in the deviance that results from adding (p-value threshold set at 0.05) or removing (p-value threshold set at 0.1) the term. To make the selection process more robust, we

TABLE II
COEFFICIENTS (AND P-VALUE) FOR THE STEPWISE MODELS OF SCENARIO 1 AND 2, OBTAINED WITH CUT-OFF THRESHOLD (THR) OF 50 AND 30. THE VARIABLES ARE LISTED IN DECREASING ORDER W.R.T. THE NUMBER OF TIMES THEY WERE SELECTED IN THE 100 STEPWISE ITERATIONS.

Scenario 1			**Scenario 2**		
Variable	*Coefficient (p-value) for thr=50*	*Coefficient (p-value) for thr=30*	*Variable*	*Coefficient (p-value) for thr=50*	*Coefficient (p-value) for thr=30*
Intercept	-8.117 (**<0.0001**)	-7.931 (**<0.0001**)	Intercept	-6.096 (**<0.0001**)	-5.649 (**<0.0001**)
Systolic BP	0.051 (**<0.0001**)	0.053 (**<0.0001**)	Heart rate	0.060 (**<0.0001**)	0.060 (**<0.0001**)
Diastolic BP	0.024 (**0.0016**)	0.020 (**0.0350**)	Waist	0.026 (**<0.0001**)	0.024 (**<0.0001**)
Hdl	-0.017 (**0.0002**)	-0.013 (**0.0094**)	Age	-0.030 (**0.0002**)	-0.033 (**<0.0001**)
Hemoglobin	-0.106 (**0.0312**)	-0.146 (**0.0046**)	Marital status, class 1	0.556 (**0.0002**)	0.490 (**0.0001**)
Marital status, class 1	0.304 (**0.0499**)	0.383 (**0.0168**)	Marital status, class 2	0.259 (0.3493)	0.217 (0.4336)
Marital status, class 2	0.052 (0.8529)	0.055 (0.8483)	Alcohol drinking	-	-0.242 (**0.0402**)
Depression	0.091 (**0.0068**)	0.085 (**0.0125**)	Depression	-	0.067 (**0.0422**)
Alcohol drinking	-0.191 (0.1137)	-0.255 (**0.0382**)	Ferritin	-	0.001 (0.0663)
Ferritin	-	0.001 (**0.0256**)			
Waist	-	0.010 (0.0941)			
Age	-	-0.012 (0.1726)			

generated 100 bootstrap samples on the training set and repeated the stepwise selection on each of them, generating 100 stepwise models with possible different selected variables. The number of times each variable is selected by the 100 stepwise procedures gives a level of confidence of the importance of that variable in predicting the outcome. Final logistic regression models were trained by including only the variables with a reasonable confidence level, i.e. selecting those that appear in the stepwise models a number of times greater than two investigated cut-off thresholds, 30 and 50 over 100, where higher the threshold fewer the selected variables.

B. Definition of Scenarios 1 and 2

To investigate the possibility of hypertension prediction with and without BP measurements, we trained the models by repeating the procedure explained in Section III-A for two different scenarios. Scenario 1 considers BP measurements but not the heart rate, being it strongly correlated to systolic BP. Scenario 2 includes the heart rate, but not systolic and diastolic BP measurements. The other variables listed in Table I are shared between the two scenarios.

C. Assessment of model performance

Performance of the developed models were assessed in terms of their discrimination ability by computing the receiver-operating characteristic (ROC) curve and the corresponding Area Under the ROC curve (AU-ROC) [7]. The ROC curve is a plot showing the relationship between model sensitivity and false positive rate (i.e. 1-specificity), while the AU-ROC is a number ranging between 0 and 1, where 1 represents perfect discrimination, while 0.5 represents random score assignment. The following performance metrics were computed on both test set and on the 100 out-of-bag samples extracted from the training bootstrap samples derived in Section III-A.

IV. RESULTS

Four different models to predict hypertension onset are identified: two of them include the variables of Scenario 1, with a stepwise cut-off threshold of 30 and 50 (over 100), respectively; the other two models exploit the variables of Scenario 2, with a stepwise cut-off threshold of 30 and 50 (over 100), respectively. The estimated models coefficients, with the corresponding p-value, are reported in Table II.

The first column of Table II highlights the variables selected for models developed in Scenario 1, ordered from the most to the less impactful based on the number of times the variables were selected on the 100 bootstrap samples. As expected, the strongest predictor is the systolic BP, followed by the diastolic BP, HDL cholesterol and hemoglobin level, all significantly associated with the outcome. Specifically, BP measurements have a positive association with hypertension risk (i.e. a positive model coefficient), while HDL and hemoglobin have a negative association with the outcome (i.e. a negative model coefficient). Among the social factors, being divorced, separated or windowed compared to being married and having an higher depression score lead significantly to a higher disease risk. Also a lifestyle factor appears in the model, i.e. alcohol drinking, whose coefficient, however, is not significantly different from zero. By lowering the stepwise cut-off threshold to 30, three new variables are included in the model: ferritin level, waist circumference and age. Among them, only ferritin has a significant (positive) association with the disease development.

Scenario 2 with stepwise cut-off threshold of 50 highlights a small set of variables needed to predict hypertension onset in absence of BP measurements: heart rate, waist, marital status, and age. In particular, higher heart rate and waist, together with being divorced, separated or windowed compared to being married lead to a higher risk of disease onset. Instead, age appears in the model with a negative coefficient, meaning that being older decrease the risk of hypertension. This is probably due to the age of subjects involved in the study, ranging from 40 to 90 years: new onset of hypertension might arise at a stage of age between 40 and 70 years rather than after 70 years. It is interesting to remark that marital status plays a significant

TABLE III
AU-ROC FOR THE STEPWISE MODELS WITH CUT-OFF THRESHOLD (THR) OF 50 AND 30, IN BOTH SCENARIOS 1 AND 2, COMPUTED ON TEST SET AND ON 100 OUT-OF-BAG TRAINING SAMPLES (IN THE LAST CASE, MEDIAN [95% CONFIDENCE INTERVAL] ARE SHOWN).

Model	**Scenario 1**		**Scenario 2**	
	Training out-of-bag	*Test set*	*Training out-of-bag*	*Test set*
Stepwise thr=50	0.780 [0.744, 0.810]	0.804	0.726 [0.690, 0.759]	0.742
Stepwise thr=30	0.778 [0.740, 0.808]	0.802	0.727 [0.696, 0.763]	0.721

role in predicting the outcome in both scenarios. Instead, waist and age were selected in Scenario 1 only when more variables were included in the model and without a significant effect. Finally, by lowering the stepwise cut-off threshold to 30, alcohol drinking, depression and ferritin appears in the model, with significant coefficients only for the first two variables.

The plot of the ROC curve and the corresponding AU-ROC of the four developed stepwise models are reported in Figure 1 and Table III, respectively. All the models perform well in terms of discrimination, with AU-ROC values around 0.8 for Scenario 1 and around 0.72 for Scenario 2. The obtained performances are in line with the ones reported in the other literature works (AU-ROC varying between 0.70 and 0.85) [4]. Models related to Scenario 2, which do not consider systolic and diastolic BP, perform slightly worse than ones of Scenario 1, but still reasonably well. This demonstrates that it is possible to predict the risk of hypertension incidence also without BP measurements, but using the heart rate.

In general, more selective models (i.e. the ones with cut-off threshold equal to 50) performed slightly better that the others including more variables (i.e. the ones with cut-off threshold equal to 30). This is particularly evident for Scenario 2, by looking at the ROC curves. This means that relaxing the stepwise cut-off threhsold, thus including additional variables in the model, does not give a valuable contribution in predicting the outcome.

V. CONCLUSIONS

In this work we developed four models for the prediction of hypertension onset, which differ for the number and type of variables involved. Two scenarios were examined to investigate the main predictors and performance of models developed with (Scenario 1) and without (Scenario 2) BP measurements.

The best model of Scenario 1 includes measurements of diastolic and systolic BP, together with some blood test biomarkers (HDL cholesterol and hemoglobin), and variables related to the demographic (marital status), wellbeing (depression scale) and lifestyle (alcohol drinking level) status. For Scenario 2, the best model includes as predictors two physical measurements, i.e. heart rate and waist, together with two variables related to the subject's demographic, i.e. age and marital status. As expected, models of Scenario 1, employing BP measurements, perform better than the ones of Scenario 2. However, in Scenario 2 we obtained a very simple and usable model that allows to predict hypertension onset with acceptable accuracy by using easily accessible information. Indeed, the only physical measurements needed are easily self-reportable or obtainable by common smart-watches and do not require laboratory tests for the collection of clinical data, which generally might not be feasible to obtain outside of clinical trials or research studies.

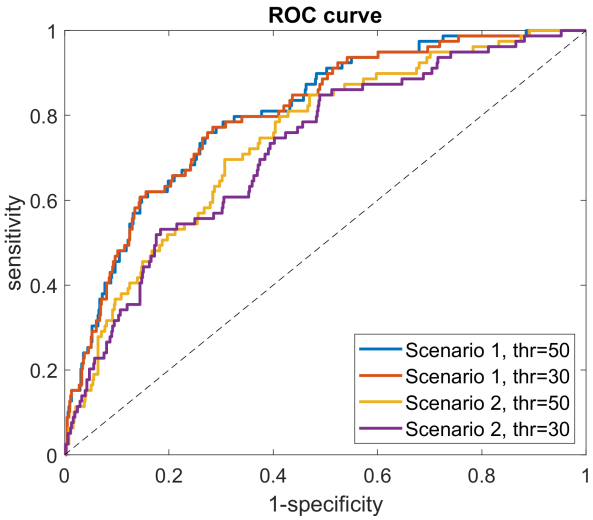

Fig. 1. ROC curve on test set for the four developed stepwise models: models of Scenario 1 with cut-off threshold (thr) of 50 (blue) and 30 (orange); models of Scenario 2 with threshold of 50 (yellow) and 30 (violet).

Finally, it is interesting to remark the importance of marital status to predict hypertension onset in both scenarios. Further investigations are needed to understand the reason leading to this kind of association.

ACKNOWLEDGMENT

ELSA was developed at Univ. College London, NatCen Social Research, the Institute for Fiscal Studies, the Univ. of Manchester and the Univ. of East Anglia. Funding is provided by the US National Institute on Aging, a consortium of UK government departments coordinated by the National Institute for Health Research and the Economic and Social Research Council.

REFERENCES

[1] P. M. Kearney, M. Whelton, K. Reynolds, P. Muntner, P. K. Whelton, and J. He, "Global burden of hypertension: analysis of worldwide data," *The lancet*, vol. 365, no. 9455, pp. 217–223, 2005.

[2] J. He, P. K. Whelton, L. J. Appel, J. Charleston, and M. J. Klag, "Long-term effects of weight loss and dietary sodium reduction on incidence of hypertension," *Hypertension*, vol. 35, no. 2, pp. 544–549, 2000.

[3] M. Slama, D. Susic, and E. D. Frohlich, "Prevention of hypertension," *Current opinion in cardiology*, vol. 17, no. 5, pp. 531–536, 2002.

[4] J. B. Echouffo-Tcheugui, G. D. Batty, M. Kivimäki, and A. P. Kengne, "Risk models to predict hypertension: a systematic review," *PloS one*, vol. 8, no. 7, p. e67370, 2013.

[5] A. Steptoe, E. Breeze, J. Banks, and J. Nazroo, "Cohort profile: the english longitudinal study of ageing," *International journal of epidemiology*, vol. 42, no. 6, pp. 1640–1648, 2013.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[7] A. C. Alba, T. Agoritsas, M. Walsh, S. Hanna, A. Iorio, P. Devereaux, T. McGinn, and G. Guyatt, "Discrimination and calibration of clinical prediction models: users' guides to the medical literature," *Jama*, vol. 318, no. 14, pp. 1377–1384, 2017.