Original article

Modelling of South African Hypertension: Application of Classical Quantile Regression

Anesu Gelfand Kuhudzai[1], Guido Van Hal[2], Stefan Van Dongen[3], Muhammed Ehsanul Hoque[4]

Abstract:

Background: High blood pressure, medically known as hypertension is the major risk factor for cardiovascular diseases (CVDs) and premature death globally. The aim of the present study was to explore possible interactions amongst systolic blood pressure`s (SBP) and diastolic blood pressure`s (DBP) risk factors in South Africa. ***Methods****:*A retrospective study was conducted using data acquired from the South African National Income Dynamics Study Wave 5, Household Survey which was carried out in 2017-2018.A final data set of 21 180 adults was utilized for data analysis. An application of the hierarchical group-lasso approach to detect interactions between SBP`s and DBP`s risk factors and classical quantile regression analysis were performed in this study. ***Results***: By using only upper quantilesbody mass index (BMI), age, race, never exercised, and the following nine interactions: BMI and age, BMI and gender male, age and never exercised, gender male and race African, race coloured and depression some or little of the time, BMI and cigarette consumption, age and race white, gender male and employment status, never exercised and cigarette consumptionwere found to be significantdeterminantsof hypertension in South Africa. ***Conclusion***: The evidence of this study suggests that it is ideal to consider interactions amongst risk factors when modelling hypertension.

***Keywords*:** Hierarchical Interactions; Group-lasso approach;Classical Quantile Regression; Hypertension; South Africa

Bangladesh Journal of Medical Science Vol. 21 No. 04 October'22 Page : 772-781
DOI: https://doi.org/10.3329/bjms.v21i4.60238

Introduction

High blood pressure, medically known as hypertension is the major risk factor for cardiovascular diseases (CVDs) and premature death globally. CVDs are conditions that affect the structures or functions of the heart. These are abnormal heart rhythms or arrhythmias, aorta disease and Marfan syndrome, congenital heart disease, coronary artery disease (narrowing of the arteries), deep vein thrombosis and pulmonary embolism, heart attack, heart failure, heart muscle disease (cardiomyopathy), heart valve disease, pericardial disease, peripheral vascular disease, rheumatic heart disease,stroke and vascular disease (blood vessel disease). CVDs are the number 1 cause of death worldwide and an estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths [1]. Hypertension is responsible for 7.6 million deaths per annum globally [2].

The prevalence and burden of hypertension is rising across the world, especially in low and middle-income countries including South Africa [3]. Approximately, 27.4% of men and 26.1% of women in South Africa have raised blood pressure in 2015 [4]. Based on these high prevalence rates of hypertension in South Africa, this study sought to establish the prevalence of hypertension amongst adults in South

1. Anesu Gelfand Kuhudzai, Ph.D. Candidate. University of Antwerp, Department of Social Epidemiology and Health Policy, Belgium & Statistical and Data Science Consultant. University of Johannesburg, Statistical Consultation Services, South Africa.
2. Prof Guido Van Hal, University of Antwerp, Department of Social Epidemiology and Health Policy, Belgium
3. Prof Stefan Van Dongen, University of Antwerp, Department of Evolutionary Ecology and Biology, Belgium
4. Prof Muhammed Ehsanul Hoque, Senior Research Associate Research Department Management College of South Africa, Durban, South Africa.

Correspondence: Anesu Gelfand Kuhudzai, Ph.D. Candidate. University of Antwerp, Department of Social Epidemiology and Health Policy, Belgium & Statistical and Data Science Consultant. University of Johannesburg, Statistical Consultation Services, South Africa. (gelfand9@yahoo.com) & 0027 787689666

Africa attributable to high systolic and diastolic blood pressure. Thus, systolic blood pressure (SBP) readings greater than or equal to 140 mmHg and/or diastolic blood pressure (DBP) greater than or equal to 90 mmHg [5].

Most previous studies on hypertension have used descriptive statistics to study the prevalence and awareness of hypertension in South Africa[6,7,8]. Some have applied mean regression [7,9,10]. In addition to descriptive statistics, this study shall also apply classical quantile regression. Modelling hypertension using quantile regression (QR) is more appropriate than using descriptive statistics and mean regression only in that it provides flexibility to estimate the influence of potential risk factors on the upper quantiles (75% or 95%) of the conditional distribution of hypertension. When modelling hypertension, it makes more sense to model high values of systolic and diastolic blood pressure which corresponds to the upper distribution of either SBP or DBP[11]. Hence, the aim of the present study was to explore possible interactions amongst SBP`s and DBP`s risk factors. Modelling interactions is conceivable to play an important role when predicting diseases [12].

Materials and Methods

In this section, the data and variables, theoretical models and data analysis techniques applied in this paper are presented.

Data and Variables

This was a retrospective study conducted using data acquired from the South African National Income Dynamics Study (NIDS) Wave 5, Household Survey which was carried out in 2017-2018.The South African National Income Dynamics Study provides good quality anthropometric, sociodemographic and behavioural data sampled across the South African population[7].

From this particular secondary data, data for South African male and female adults aged 18 years and above was extracted for analysis. NIDS was embarked by the South African Presidency in order to track changes in the well-being of South Africans citizens in the entire country [13]. Hence, this survey provided nationally representative data.

The target population for NIDS was private households in all nine provinces of South Africa, and residents in workers' hostels, convents and monasteries. The frame excludes other collective living quarters, such as student hostels, old age homes, hospitals, prisons and military barracks. Fieldworkers were trained and instructed to interview and collect data on subjects residing at selected households.

Data cleaning was conducted before analyzing the data. The data cleaning process involved dropping observations with missing data (7 127) for any of the variables used in the study and participants aged below 18 (1 803). A final data set of 21 180 adults was obtained from 30 110 adults originally observed.

The variables used in this study are systolic blood pressure, diastolic blood pressure, non-modifiable risk factors (age, gender and race) and modifiable risk factors (BMI, exercises, cigarette consumption, depression and employment status). Systolic blood pressure and diastolic blood pressure are the dependent variables whilst age, body mass index, gender, race, exercises, cigarette consumption, depression and employment status are the independent variables. Age was computed by subtracting *date of birth* from *date of interview* and body mass index was calculated by dividing *weight* (kg) by *height in meters squared* (m^2).

Classical Quantile Regression and Computational Methods

In statistical modelling, regression analysis is one of the most widely used and powerful multivariate techniques in order to assess the impact of a set of variables X on a certain outcome variable Y. Classical or Standard linear regression centers on the expectation of variable Y conditional on the values of a set of variables X, thus $E(Y|X)$[14]. This is called the regression function. Since this function focuses on a specific location which is the mean, quantile regression extends this approach to allow the conditional distribution of Y on X at different locations to be established [15]. In this regard, quantile regression provides a global view on the interrelations between Y and X. Quantile Regression Model equation for the $\tau\, th$ quantile is given by,

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \cdots + \beta_p(\tau)x_{ip}$$

Where p is the number of regressor variables.

The quantile regression model estimates can be obtained by considering the unconditional quantiles as an optimization problem. Like the mean which is obtained as the solution to the problem of minimizing the sum of squared deviations:

$$\min_{\mu\in\square} \sum^{n} (y_i - \mu)^2$$

The median can be obtained as the solution to the problem of minimizing the sum of absolute deviations:

$$\min_{\xi \in \mathbb{R}} \sum_{i=1}^{n} |y_i - \xi| \quad (3)$$

The τ th sample quantile $\hat{\xi}_\tau$ can be then obtained as the solution to the problem of minimizing an asymmetric weighted absolute deviations. The optimization problem is defined as:

$$\min_{y_i \in \mathbb{R}} \left(\sum_{i:y_i \geq \xi} \tau |y_i - \xi| + \sum_{i:y_i < \xi} (1-\tau) |y_i - \xi| \right) \quad (4)$$

Equivalent to:

$$\min_{y_i \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau (y_i - \xi) \quad (5)$$

Where $\rho_\tau(u) = \tau |u| I(u \geq 0) + (1-\tau) |u| I(u < 0)$ is called the pinball loss function [16]. Linear programming methods can then be utilised to obtain quantile regression estimates [17].

These linear programming methods include the simplex algorithm of Barrodale and Roberts (1973), the Sparse Frisch-Newton algorithm described in Portnoy and Koenker (1997) and the Sparse Frisch-Newton algorithm with pre-processing. The Barrodale and Roberts (1973) simplex algorithm is the default method implemented in the r package called quantreg[20]. The implementation of the simplex method and further developments in linear programming have made quantile regression to better than classical linear regression methods [21].

Pairwise Hierarchical Interactions

$$y = \beta_0 + \sum_{l=1}^{L_i} \sum_{i=1}^{m} \beta_{i,l,\ i,l} + \sum_{1 \leq i < j \leq m} \sum_{l=1}^{L_i} \sum_{k=1}^{L_j} \theta_{i,j,l,k} X_{i,l} X_{j,k}$$

Now interaction between say two independent variables X_i and X_j, occur when the effect of X_i will vary depending on the level or value of X_j. Pairwise hierarchical interactions in this study shall be conducted in a manner that satisfies strong hierarchy and then parameter selection is applied via the group-lasso. A model is said to obey strong hierarchy whenever an interaction is estimated to be nonzero and both main effects are included in the model [12]. Weak hierarchy is obtained as long as either of its main effects are present [12]. Interactions amongst variables can play an important role in predicting diseases such as hypertension [23].

Basically, there are three possible cases of interaction:

- Interaction between two continuous variables.
- Interaction between two categorical variables.
- Interaction between a categorical variable and a continuous variable.

Interaction between two continuous variables

Let Z_1 and Z_2 be two continuous variables, then the interaction between these continuous variables is

$$Z_{1:2} = [1 \quad Z_1] * [1 \quad Z_2]$$

$$= [1 \quad Z_1 \quad Z_2(Z_1 * Z_2)]$$

Interaction between two categorical variables

Let X_1 and X_2 be two categorical variables, then the interaction between these categorical variables is

$$X_{1:2} = [1 \quad X_1] * [1 \quad X_2]$$

$$= [1 \quad X_1 \quad X_2(X_1 * X_2)]$$

Whenever there is an interaction between two categorical variables, interactions are taken at each level of the variable.

Interaction between a categorical variable and a continuous variable

Let Z be a categorical variable with L levels and X be a continuous variable, then the interaction between these two variables can result into one of the following [12]:

- $\mu_{ij} = \mu$ (main effects, no interaction),
- $\mu_{ij} = \mu + \theta_1^i$ (Z main effect or X),
- $\mu_{ij} = \mu + \theta_1^i + \theta_2 z$ (X main effect or Z),
- $\mu_{ij} = \mu + \theta_1^i + \theta_2 z + \theta_{1:2}^i z$ (main effects and interaction).

Lasso

Lasso defined as the least absolute shrinkage and selection operator has emerged as a critical tool for variable selection. It is quite convenient to apply lasso in estimating the quantile regression models so as to improve the prediction accuracy by eliminating irrelevant variables [24]. The lasso includes an ℓ_1 -penalty term that constraints the minimum size of the estimated model coefficients, forcing the model to have fewer parameters. The lasso coefficient estimates solve the following problem [25]:

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2,$$

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

e following function:

s s

where has to be greater than zero. is the tuning parameter that controls the amount of shrinkage.

Group-lasso is an extension of lasso that performs variable selection on non-overlapping groups of variables and sets groups of coefficients to zero [26].

Data Analysis

Descriptive statistics were analysed by use of IBM SPSS version 27. Frequencies on demographic and lifestyle characteristics of participants and summary statistics on continuous variables such as SBP, DBP, BMI and age were produced. To explore possible interactions amongst SBP`s and DBP`s risk factors, the Least Absolute Shrinkage and Selection Operator (Lasso) via pairwise hierarchical interactions technique the R packages namely hierNet[27] and glmnet[28] were utilised. The classical quantile regression model wasfitted using the quantreg R package [20].

Ethical Consideration

The South African National Income Dynamics Survey was conducted after ethical approval was granted by the University of Cape Town, Faculty of Commerce Ethics Committee. Informed consent was also obtained from each study participant.

Results

This section presents the empirical resultsof the study. These results are presented in form of tables and figures. Also, interpretation of the results is given in this section.

***Table 1:* Biographical Details**

Characteristic	Category	n	Percentage
Gender	Male	8 616	40.7%
	Female	12 564	59.3%
Race	African	16 999	80.3%
	Coloured	2 792	13.2%
	Asian/Indian	338	1.6%
	White	1 051	5.0%
Age	18 – 29 years	7 658	36.2%
	30 – 39 years	4 434	20.9%
	40 – 49 years	3 192	15.1%
	50 and above years	5 896	27.8%

The study results in Table 1 show that 8 616 (40.7%) of the respondents were males and 12 564 (59.3%) were females. Most of the participants were African and they were 16 999 (80.3%) and the least number of participants were Asian/Indian and they were 338 (1.6%). In regard to the age distribution, 7 658 (36.2%) were between 18-29 years, followed by the 50 years and above age group who were 5 896 (27.8%). The least number of participants by age were 3 192 (15.1%) and they were aged between 40 to 49 years.

***Table 2:* Life Style Characteristics**

Life Style Characteristics	Levels	n	Percentage
Exercises	Never	14 595	68.9%
	Once or Twice a week	3 861	18.2%
	Three or More times a week	2 724	12.9%
Cigarette Consumption	Yes	4 046	19.1%
	No	17 134	80.9%
Depression	Rarely or none of the time (Less than 1 day)	12 152	57.4%
	Some or Little of the time (1-2 days)	6 271	29.6%
	Occasionally or All of the time (3-7 days)	2 757	13.0%
Systolic Blood Pressure	Normal (Less than 120)	10 988	51.9%
	Pre-Hypertension (120 – 139)	6 872	32.4%
	High Blood Pressure Stage 1 (140 – 159)	2 250	10.6%
	High Blood Pressure Stage 2 (160 or higher)	752	3.6%
	Hypertensive Crisis (Higher than 180)	318	1.5%
Diastolic Blood Pressure	Normal (Less than 80)	12 015	56.7%
	Pre-Hypertension (80 – 89)	5 400	25.5%
	High Blood Pressure Stage 1 (90 – 99)	2 577	12.2%
	High Blood Pressure Stage 2 (100 or higher)	814	3.8%
	Hypertensive Crisis (Higher than 110)	374	1.8%

Life Style Characteristics	Levels	n	Percentage
Body Mass Index	Underweight (< 18.50)	1 311	6.2%
	Healthy (18.50 – 24.99)	8 608	40.6%
	Overweight (25.00 – 29.99)	5 100	24.1%
	Obese (30.00 – 34.99)	3 304	15.6%
	Very Obese (35.00 – 39.99)	1 709	8.1%
	Morbidly Obese (≥ 40.00)	1 148	5.4%
Employment Status	Yes	6 772	32.0%
	No	14 408	68.0%

It is apparent from Table 2 that very few respondents 2 724 (12.9%) do exercise regularly, i.e. (three or more times a week). Majority of respondents 14 595 (68.9%) indicated that they do not exercise. A total of 4 046 (19.1%) participants do smoke whilst 17 134 (80.9%) do not smoke.

Respondents were asked to indicate the number of times in a week they are likely to suffer from depression. 12 152 (57.4%) respondents revealed that they rarely suffer from depression and 2 757 (13%) indicated that they are likely to be affected by a depression between 3 to 7 days a week. The study also considered employment status as a possible risk factor of raised blood pressure. It can be seen from the results in Table 4 that 14 408 (68%) of the study participants are not employed whilst 6 772 (32%) are employed.

It is indicated in Table 2 that, 3 320 (15.7%) of the total respondents had high SBP (more than 140 mmHg) and 3 765 (17.8%) participants had abnormal DBP (more than 90 mmHg). Finally yet importantly, 5 100 (24.1%) study participants were overweight (25 – 29.9 kg/m^2) and 6 161 (29.1%) were obese, thus 30 kg/m^2 and above.

Table 3: Summary Statistics of Continuous Variables

	SBP (mmHg)	DBP (mmHg)	Age	BMI (kg/m^2)
Mean	121.85	78.99	39.56	26.95
Median	119.00	77.50	35.00	25.61
Standard Deviation	20.06	12.30	16.75	7.05
Minimum	42.00	28.50	18	11.25
Maximum	237.50	146.00	93	86.98
Range	187.50	117.50	75	75.74
Interquartile Range	24.00	16.00	27	9.65

The average age of the respondents was 39.56 years (Table 3). The average BMI for both female and male participants was 26.95 kg/m^2 , more than the normal level of between 18.50 – 24.99 kg/m^2 according to World Health Organization, 2000 classification table. The mean SBP and DBP were 121.85 mmHg and 78.99 mmHg respectively.

Pairwise Interactions for SBP

This section presents the main effects and interactions retrieved for the SBP model. The analysis was conducted using 8 explanatory variables which consists of 2 continuous and 6 categorical variables. Categorical variables were utilised in the model as per their respective levels using dummy coding. The results obtained illustrate that 14 main effects and 21 interactions were deduced for the SBP model. It is apparent that the model satisfies the concept of strong hierarchy as the main effects of all 21 interactions are present. These results indicate that systolic blood pressure can be predicted by the 14main effect variables and the 21 interactions detected. Applying the group-lasso technique would be necessary to predict SBP more accurately by eliminating uninformative variables.

Lasso Model Selection for SBP

Table 4: Coefficient Extraction for SBP Model

Main Effects	Interactions Detected
Age	BMI * Age
Race(Coloured)	BMI * Gender(Male)
	BMI * Cigarette Consumption
	Age * Excercises(Never)
	Gender(Male) * Race(African)
	Race(Coloured) * Cigarette Consumption
	Race(Coloured) * Depression (Some or little of the time)

Results of the lasso model selection for SBP after fitting 14 main effect variables and 21 interactions are summarised in Table 4. Nine non-zero coefficients representing 2 main effects and 7 interactions were extracted in the sparse matrix, as possible strong predictors of systolic blood pressure.

Figure 1 illustrates the cross-validation curve with dotted lines and error bars. The left vertical line in the plot shows the value at which the minimal mean squared error is achieved and the right vertical line shows the most regularized model whose mean squared error is within 1 standard deviation of the minimum. It is evident from Figure 1 that the errors increase substantially when the number of variables decreases, but they remain constant between 9 to 34 variables.

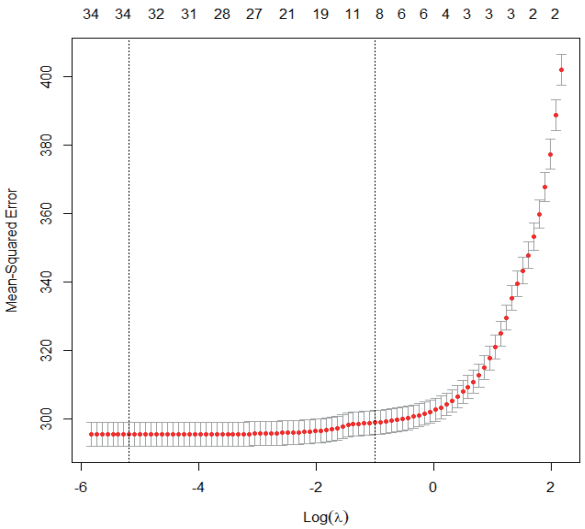

***Figure 1:** Cross Validation Plot for SBP Model*

These results suggests that the model has optimally chosen 9 variables to be the possible best predictors of SBP, confirming the findings of the sparse matrix represented in Table 5.

***Table 5:** Classical Quantile Regression Estimates for SBP`s Risk Factors*

τ	Q(0.75)	Q(0.95)
Age	0.20****	0.46***
Race(Coloured)	5.07***	7.41***
BMI * Age	0.01***	0.02***
BMI * Gender(Male)	0.31***	0.30***
BMI * Cigarette Consumption	0.02	0.01
Age * Exercises (Never)	0.03**	0.09***
Gender(Male) * Race(African)	4.08***	5.07***
Race(Coloured) * Cigarette Consumption	1.80	0.45
Race(Coloured) * Depression (Some or little of the time)	2.30*	0.08
* p - value < 0.05; ** p - value < 0.01; *** p - value < 0.001		

Table 5 presents the upper classical quantile regression estimated coefficients for SBP`s risk factors. Only the upper quantiles (75% or 95%) were estimated in order to examine how blood pressure risk factors affects individuals most at risk for hypertension. It can be seen from Table 5 that, in all upper quantiles ($\tau \in \{0.75, 0.95\}$), age and race coloured had positive statistically significant effects on SBP. The interactions between BMI and age, BMI and gender male, age and exercises never & gender male and race African also presented statistically significant relations across all upper quantiles. The interactions between BMI and cigarette consumption & race coloured and cigarette consumption did not present statistically significant coefficients for both high quantiles. Interaction between race coloured and depression for some or little of the time presented a significant effect on the 75th quantile and did not present a significant effect on the 95th quantile.

Similarly to SBP, the pairwise interactions for DBP were conducted using 8 explanatory variables which consists of 2 continuous and 6 categorical variables. Categorical variables were also treated in the model as per their respective levels.It can be deducted from the analysis that 15 main effects and 29 interactions obeying strong hierarchy concept were extracted for the SBP model. It is ideal to fit a group-lasso model on the variables extracted so as to eliminate irrelevant variables when predicting SBP.

Lasso Model Selection for DBP

***Table 6:** Coefficient Extraction for DBP Model*

Main Effects	Interactions Detected
BMI	BMI * Age
Age	BMI * Gender(Male)
Race(Coloured)	BMI * Race(Coloured)
Excercises(Never)	BMI * Cigarette Consumption
	BMI * Employment Status
	Age * Race(White)
	Gender(Male) * Employment Status
	Excercises(Never) * Cigarette Consumption

It can be seen from Table 6, that after fitting the group-lasso model, 4 main effects and 8 interactions were extracted from the sparse matrix as possible strong predictors of DBP.

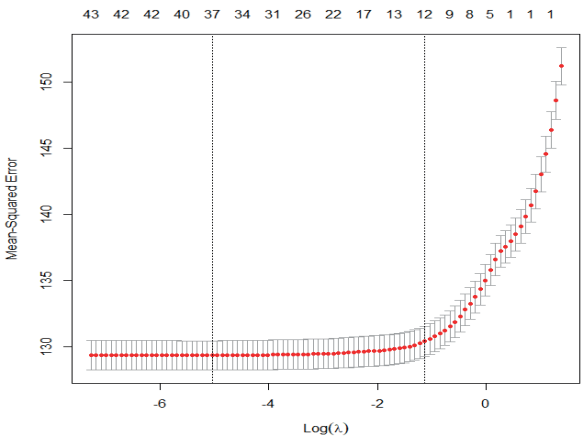

***Figure 2:** Cross Validation Plot for DBP Model*

Figure 1 illustrates that the mean squared errors substantially increase when the number of variables decreases, but they remain constant between 12 to 43 variables. These results implies that the model has optimally chosen 12 variables to be the possible strong predictors of DBP, confirming the findings of the sparse matrix represented in Table 8.

Table 7: Classical Quantile Regression Estimates for DBP`s Risk Factors

τ	Q(0.75)	Q(0.95)
BMI	0.53***	0.50***
Age	0.29***	0.36***
Race(Coloured)	3.74**	8.32***
Exercises(Never)	0.96***	1.62**
BMI * Age	-0.002	-0.001
BMI * Gender(Male)	0.13***	0.12***
BMI * Race(Coloured)	-0.03	-0.19
BMI * Cigarette Consumption	0.06***	0.05
BMI * Employment Status	0.005	0.02
Age * Race(White)	-0.06***	-0.13***
Gender(Male) * Employment Status	1.38**	0.34
Exercises(Never) * Cigarette Consumption	1.17*	1.07

* p - value < 0.05; ** p - value < 0.01; *** p - value < 0.001

Table 7 illustrates the upper classical quantile regression estimated coefficients for DBP`s risk factors. BMI, age, race coloured, exercises never,the interaction between BMI and gender male & age and race white presented statistically significant effects on DBP across all higher quantiles. Interaction effects between BMI and age, BMI and race coloured & BMI and Employment Status did not present any statistically significant relations with DBP. The interactions between BMI and cigarette consumption, gender male and employment status &exercises never and cigarette consumption displayed statistically significant association with DBP, only at the 75th quantile.

Discussion

This study revealed statistically significant risk factors of hypertension based on the classical quantile regression models estimated. Quantile regression was more helpful in this study because it appropriately captured the effects of the observed risk factors on the upper quantiles of both SBP and DBP.

Study results illustrated that age had positive statistically significant estimated coefficients with both SBP and DBP respectively. The magnitude of the association increased from the 75th quantile to the 95th quantile. These findings suggests that prevalence of hypertension increase with age increase.The combination of BMI and age had positive statistically significant effects with SBP only across the upper quantiles. These results imply that the increase in both BMI and age is likely to influence the occurrence of raised blood pressure. The present findings seem to be consistent with other research which found that hypertension increases with age, possibly because age is mostly associated with structural changes in the arteries and especially with large artery stiffness[29].

BMI and gender male presented positive significant relations with SBP and DBP on both higher quantiles ($\tau \in \{0.75,0.95\}$). These findings indicate that among males, an increase in BMI is associated with an increase in SBP. BMI presented positive statistically significant impact on DBP only. These findings imply that an increase in BMI is related with the increase in prevalence of elevated blood pressure. These results are consistent with previous studies which suggests that men are likely to be more hypertensive as compared to women [9] and that BMI is significantly associated with hypertension and individuals who are overweight and obese are at high risk of developing high blood pressure [30].

Exercises never was found to be positively significant with DBP only on both higher quantiles. This indicates that South African individuals who do not exercise are vulnerable to hypertension. Positive empirical estimated coefficients for the interaction between age and exercises never were statistically significant with only SBP across both quantiles. This finding implies that among South Africans who do not take part in exercises, every additional year of age is associated with an increase in SBP. These results are quite in line with other studies which revealed that the incidences of high blood pressure are most common in individuals with sedentary lifestyle [29].

Race coloured had a positive impact with both BP measures across all the upper quantiles. These results suggests that the prevalence of raised blood pressure is likely to increase among the coloured people as compared to other racial groups. Also, negative statistically significant effect on DBP only was found on the interaction between age and race white, suggesting that among South Africans who are not white, an increase in age is likely to influence the occurrence of raised blood pressure. In regard to racial differences in hypertension, this finding is coherent with prior studies that blacks do develop hypertension at an earlier age than whites [31].

Interaction effects on DBP only between gender male and employment status &exercises never and cigarette consumption were statistically significant on the 75th quantile only. These findings imply that employed males are prone to suffer from high blood pressure possibly due to work pressure and stress. The interaction between exercises never and cigarette consumption suggests that individuals who smoke as well as do not take part in physical exercises are prone to hypertension. In regard to cigarette consumption, these results are similar to past studies which also revealed that cigarette smoking is modestly associated with an increased risk of developing hypertension [32].

A positive statistically significant interaction effect on SBP between gender male and race African was found across both upper quantiles. This indicates that the occurrence of high blood pressure is likely to increase among African males. A finding noted by other studies that high prevalence of blood pressure is experienced more among black males [31].

The interaction between race coloured and depression for some or little of the time was found to be positively significant with SBP only for the 75th quantile. This outcome indicates that coloured individuals who sometimes suffer from depression are more likely to suffer from hypertension. Similarly, previous studies have suggested that depression increases the risk of suffering from uncontrolled hypertension [33].

Other risk factors extracted after fitting the SBP and DBP group-lasso models were not statistically significant after conducting the classical quantile regression models as indicated in Table 6 and 9 respectively. This may be attributed to very few participants with such lifestyle characteristics in this study.

Conclusion

This study presented an application of the hierarchical group-lasso approach to detect possible interactions between SBP`s and DBP`s risk factors and perform variable selections whilst obeying the concept of strong hierarchy. Also, classical quantile regression analysis was conducted in order to estimate the influence of potential risk factors on the upper quantiles (75% or 95%) of the conditional distribution of hypertension. The results derived from the group-lasso interaction model were considered to be conclusive given their ability to capture linear and non-linear effects while performing variable selection.

The application of the techniques identified some important variables as risk factors of hypertension in South Africa.The evidence of this study suggests that it is ideal to consider interactions amongst risk factors when modelling hypertension and possibly other diseases instead of only considering main effect variables. Repeated surveys of this nature are ideal to be administered regularly across South Africa so as to continuously monitor and manage the risk factors of hypertension.

Acknowledgements

The authors are quite grateful to the research team of the South African National Income Dynamics Study 2017-2018 (NIDS) for their permission to use their data.

Funding

There was no organisation that sponsored this study.

Competing Interests

The authors declare that they have no competing interests.

Availability of data and materials

The dataset analysed during the current study are available in the South African National Dynamics Study repository,

https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/712/download/9579

Authors's contribution

Data gathering and idea owner of this study: AnesuGelfandKuhudzai

Study design: AnesuGelfandKuhudzai

Data gathering: AnesuGelfandKuhudzai

Writing and submitting manuscript: AnesuGelfandKuhudzai

Editing and approval of final draft: Prof Guido Van Hal, Prof Stefan Van Dongen andProf MuhammedEhsanulHoque

References:

1. World Health Organisation. Cardiovascular Diseases (CVDs), https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (17 May 2017, accessed 11 March 2021).
2. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. *Nat Rev Nephrol* 2020; **16**: 223-237. https://doi.org/10.1038/s41581-019-0244-2
3. Jongen VW, Lalla-Edward ST, Vos AG, et al. Hypertension in a rural community in South Africa: what they know, what they think they know and what they recommend. *BMC Public Health* 2019; **19**: 341. https://doi.org/10.1186/s12889-019-6642-3
4. World Health Organisation. Raised blood pressure (SBP ≥ 140 OR DBP ≥ 90), age-standardized (%): Estimates by country, https://apps.who.int/gho/data/node.main.A875STANDARD?lang=en (17 November 2017, accessed 12 March 2021).
5. World Health Organisation. Hypertension, https://www.who.int/news-room/fact-sheets/detail/hypertension (13 September 2019, accessed 12 March 2021).
6. Rayner B. Hypertension: Detection and Management in South Africa. **Nephron ClinPract** 2010; **116**: c269-c273. https://doi.org/10.1159/000318788
7. Cois A, Ehrlich R. Analysing the socioeconomic determinants of hypertension in South Africa: a structural equation modelling approach. *BMC Public Health* 2014; **14**: 414. https://doi.org/10.1186/1471-2458-14-414
8. Lloyd-Sherlock P, Beard J, Minicuci N, et al. Hypertension among older adults in low- and middle-income countries: prevalence, awareness and control. *International Journal of Epidemiology* 2014; **43**: 116-128. https://doi.org/10.1093/ije/dyt215
9. Berry KM, Parker W, Mchiza ZJ, et al. Quantifying unmet need for hypertension care in South Africa through a care cascade: evidence from the SANHANES, 2011-2012. *BMJ Glob Health* 2017; **2**: e000348. https://doi.org/10.1136/bmjgh-2017-000348
10. Bhimma R, Naicker E, Gounden V, et al. Prevalence of Primary Hypertension and Risk Factors in Grade XII Learners in KwaZulu-Natal, South Africa. *International Journal of Hypertension* 2018: 1-9. https://doi.org/10.1155/2018/3848591
11. Fenske N, Kneib T, Hothorn T. Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression. *Journal of the American Statistical Association* 2011; **106**: 494-510. https://doi.org/10.1198/jasa.2011.ap09272
12. Lim M, Hastie T. Learning Interactions via Hierarchical Group-Lasso Regularization. Journal of *Computational and Graphical Statistics* 2015; **24**: 627-654. https://doi.org/10.1080/10618600.2014.938812
13. Leibbrandt M, Woolard I, de Villiers L. National Income Dynamics Study Methodology: Report on NIDS Wave 1 Technical Paper no. 1, http://www.nids.uct.ac.za/publications/technical-papers/108-nids-technical-paper-no1/file (July 2009, accessed 16 April 2017).
14. Weisberg S. Applied linear regression. Fourth edition. Hoboken, New Jersey: Wiley, 2014.
15. Davino C, Furno M, Vistocco D. Quantile regression: theory and applications. 1. ed. Chichester: Wiley, 2014. https://doi.org/10.1002/9781118752685
16. Koenker R, Bassett GJ. Regression Quantiles. *Econometrica* 1978; **46**: 33-50. https://doi.org/10.2307/1913643
17. Koenker R, Hallock KF. Quantile Regression An Introduction. *Journal of EconomicPerspectives* 2001; **15**: 143-156. https://doi.org/10.1257/jep.15.4.143
18. Barrodale I, Roberts FDK. An Improved Algorithm for Discrete $l_1 $ Linear Approximation. *SIAM J Numer Anal* 1973; **10**: 839-848. https://doi.org/10.1137/0710069
19. Portnoy S, Koenker R. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. Statist Sci; 12. Epub ahead of print 1 November 1997. DOI: 10.1214/ss/1030037960. https://doi.org/10.1214/ss/1030037960
20. Koenker R. Package 'quantreg', https://cran.r-project.org/web/packages/quantreg/quantreg.pdf (2017, accessed 27 June 2017).
21. Koenker R. Quantile Regression IN R: A VIGNETTE, https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf (2017, accessed 16 September 2017).
22. Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions. Ann Statist; 41. Epub ahead of print 1 June 2013. DOI: 10.1214/13-AOS1096. https://doi.org/10.1214/13-AOS1096
23. Schwender H, Ickstadt K. Identification of SNP interactions using logic regression. *Biostatistics* 2008; **9**: 187-198. https://doi.org/10.1093/biostatistics/kxm024
24. Jiang L, Bondell HD, Wang HJ. Interquantile shrinkage and variable selection in quantile regression. *Computational Statistics & Data Analysis* 2014; **69**: 208-219. https://doi.org/10.1016/j.csda.2013.08.006
25. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal StatisticalSociety* 1996; **58**: 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
26. Simon N, Friedman J, Hastie T, et al. A Sparse-Group Lasso. *Journal of Computational*

andGraphical Statistics 2013; **22**: 231-245. https://doi.org/10.1080/10618600.2012.681250

27. Bien J, Tibshirani R. A Lasso for Hierarchical Interactions, https://cran.r-project.org/web/packages/hierNet/hierNet.pdf (2015, accessed 19 October 2017).
28. Hastie T, Qian J. Glmnet_Vignette, https://web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf (2016, accessed 11 October 2017).
29. Princewel F, Cumber SN, Kimbi JA, et al. Prevalence and risk factors associated with hypertension among adults in a rural setting: the case of Ombe, Cameroon. *Pan Afr MedJ* 34. Epub ahead of print 14 November 2019. DOI: 10.11604/pamj.2019.34.147.17518. https://doi.org/10.11604/pamj.2019.34.147.17518
30. Hamano T, Shiotani Y, Takeda M, et al. Is the Effect of Body Mass Index on Hypertension Modified by the Elevation? A Cross-Sectional Study of Rural Areas in Japan. *IJERPH* 2017; **14**: 1022. https://doi.org/10.3390/ijerph14091022
31. Lackland DT. Racial Differences in Hypertension: Implications for High Blood Pressure Management. *The American Journal of the Medical Sciences* 2014; **348**: 135-138. https://doi.org/10.1097/MAJ.0000000000000308
32. Bowman TS, Gaziano JM, Buring JE, et al. A Prospective Study of Cigarette Smoking and Risk of Incident Hypertension in Women. *Journal of the American College of Cardiology* 2007; **50**: 2085-2092. https://doi.org/10.1016/j.jacc.2007.08.017
33. Meng L, Chen D, Yang Y, et al. Depression increases the risk of hypertension incidence: a meta-analysis of prospective cohort studies. *Journal of Hypertension* 2012; **30**: 842-851. https://doi.org/10.1097/HJH.0b013e32835080b7