International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 13, 2013, pp 55-72

Prevalence Rates of Hypertension and Related Risk Factors in Palestine

Mahmoud K. Okasha
Department of Applied Statistics
Al-Azhar University - Gaza, Palestine

[Received January 20, 2013; Revised September 9, 2013; Accepted December 23, 2013]

Abstract

The prevalence of hypertension is increasing rapidly around the globe, including Palestine. This paper aimed to estimate the prevalence rates and to identify the risk factors that have the greatest effects on the prevalence of hypertension in Palestine through the application of relevant statistical models for classifying and identifying cases. Logistic regression analyses were applied to two data sets on the prevalence of hypertension in Palestine. Two stepwise logistic regression models were selected using the Akaike Information Criterion (AIC) and fitted to the two data sets and tested in terms of accuracy and correct classification rates. The results indicated that the overall prevalence rate of hypertension in Palestine is 3.7% in the total population, 7.6% in the adult population, and 24.4% in the population aged 45 and over. Moreover, significant risk factors for hypertension are age, sex, smoking status, peptic ulcer disease, arthritis rheumatism, high cholesterol, hyperthyroidism (nonmalignant), fasting blood sugar, microalbuminurea, and type of locality (rural and refugee camps). Estimation of the odds ratios for the various risk factors for hypertension in Palestine showed that the odds ratio for gender was 3.012, indicating that women were 3.012 times more likely to have hypertension compared to men and smokers were 5.16 times more likely to have hypertension compared to nonsmokers.

Keywords and Phrases: AIC, Hosmer and Lemeshow tests, Hypertension, Logistic regression model, Odds ratio, Prevalence rate, Risk factors, ROC curves.

AMS Classification: Primary 62J02; Secondary 62J20.

1 Introduction

High blood pressure represents one of the main health risks in developed and developing countries. High blood pressure is a disease and a risk factor that can increase the chance of developing heart disease, stroke, and other serious conditions. Blood pressure is the pressure of the blood in arteries and measured in millimeters of mercury (mm Hg) and recorded as two figures. They are the systolic blood pressure (SBP) and the diastolic blood pressure (DBP), which measure the pressure in the arteries when the heart contracts and when the heart rests between each heartbeat, respectively. The higher the blood pressure, the greater the risk of hypertension. A person is considered hypertensive if his or her DBP is greater than or equal to 90 mm Hg, his or her SBP is greater than or equal to 140 mm Hg, or he or she is currently taking antihypertensive medications, regardless of the actual BP measurement. Treatment includes a change in lifestyle risk factors, such as losing weight, taking up regular physical activity, consuming a healthy diet and stopping smoking, and having a low salt and caffeine intake. A strong relationship between high BP in the initial phases of life and hypertension in adulthood has been demonstrated (Parati et al., 1998). This finding has attracted the interest of researchers in investigating the prevalence of high BP in childhood and adolescence, as well as the associated risk factors. Hypertension is estimated to cause 4.5% of the global disease burden. The prevalence of hypertension among Palestinian adults at present exceeds 20.0%. High BP values have been associated with excessive body weight and cardiovascular diseases. However, few studies have investigated the contribution of other risk behaviors, such as an unsuitable diet and socio-economic factors. In addition, few studies have simultaneously analyzed three or more risk factors for high BP.

1.1 Background

Yang, Wang, Zhi, Zhu, and Liu (2011) conducted a study to identify the prevalence rates of hypertension and associated risk factors among the Chinese population of Tianjin and estimated the prevalence rates of hypertension among adults. The results of the logistic regression analyses indicated that the odds ratio (ORs) for the combined risk factors for hypertension in Tianjin were associated with factors such as a lower level of education, living in rural areas, age, alcohol assumption, overweight, obesity, and impaired fasting glucose. The researchers concluded that the prevalence rate of the different subtypes of hypertension was high in the population of Tianjin and that different measurements of prevention and treatment should be taken according to the different subtypes of hypertension. In a different cross-sectional study on the prevalence rates of hypertension and associated risk factors in other Chinese subpopulations, Ruixing et al. (2006) examined different demographic characteristics, health-related behaviors, and lifestyle factors collected by questionnaire. The overall prevalence rates of hypertension and isolated systolic hypertension among the subpopulations were significantly different. The SBP levels among the subpopulations

were also significantly different. The prevalence of hypertension was positively correlated with triglycerides, male gender, age, total cholesterol, and alcohol consumption. Sarraf-Zadegan, Boshtam, Mostafavi, and Rafiei (1999) studied the prevalence rates of hypertension and associated risk factors in Isfahan in the Islamic Republic of Iran and found that, overall, 18.0% (16.8% of men and 19.4% of women) had systemic hypertension. The most important risk factors for the mean SBP and DBP levels were age and body mass index (BMI). There was a high prevalence of obesity, hyperlipidemia, and diabetes mellitus among individuals with hypertension compared with individuals without hypertension.

Mishra and Kumar (2011) studied risk factors for hypertension in a rural location, Varanasi in India. They found that high- and intermediate-level educational status, upper-middle and high socio-economic status, and anxiety were significant risk factors for hypertension. They also found that a negative energy balance had a significant protective effect on the occurrence of hypertension. In a study of the prevalence of hypertension and its associated risk factors in two rural communities in Malaysia, Tee et al. (2010) applied logistic regression analysis and found that age, history of alcohol consumption, and BMI were independently associated with hypertension. The researchers concluded that age, education level, alcohol consumption, and BMI are important risk factors associated with the prevalence of hypertension among the villagers. These risk factors are comparable to those reported by Yadav et al. (2008) in India. Nakanishi et al. (2003) assessed major risk factors associated with hypertension among Japanese male office workers aged 23–59 years, including a family history of hypertension, obesity, diabetes mellitus, hypercholesterolemia, hypertriglyceridemia, hyperuricemia, and increased white blood cell counts. The authors concluded that the accumulation of risk factors was highly associated with an increased risk of hypertension in Japanese men.

In Palestine, few studies involving only simple analyses have been conducted to determine the exact magnitude of the prevalence of hypertension rates and the awareness, control, and presence of hypertension risk factors. Identifying differences in the prevalence rates of hypertension between different groups in the population may help to identify particular factors—genetic, behavioral, and/or environmental—that might be related to the development of hypertension. One study was conducted by Abed and Abu Haddaf (2013) to identify the most common hypertension risk factors but only among refugees living in the Gaza Strip and registered with the United Nations Relief and Works Agency (UNRWA) primary health care as adult patients suffering from essential hypertension. The researchers collected data on exposure to selected risk factors from the adult Palestinian population living in the Gaza Strip. The most common risk factors identified in this study were a lack of physical activity, obesity, family history, high cholesterol level, smoking, low high-density lipoprotein (HDL), and high triglycerides level (TG). Abed and Abu Haddaf (2013) also found a higher prevalence rate of hypertension among women than men and claimed that a similar result has been found by many national and international researchers. In contrast,

men seemed to have a higher risk of developing hypertension than women among a Portuguese adult population (Macedo et al., 2005). Another study (Wilkins, Campbell, & Joffres, 2010) reported that the rate of hypertension was nearly equal between men and women.

The goal of this paper was to study an emerging public health problem, hypertension, in Palestinian society and to provide an example of a statistical methodology that can be followed in similar public health problems in the region. Therefore, we analyzed two data sets that involved all the necessary variables to assess possible risk factors for hypertension in Palestine. Possible classification methods that could be applied to such data sets to achieve similar goals were discussed and compared in detail by Okasha and Abu Samra (2013). The results indicated that the logistic regression model is the most appropriate for such data sets. Thus, we mainly applied the logistic regression model to the two data sets to assess all possible risk factors of the disease in the Palestinian Territories. All computations in this article were performed using the *R* statistical computing software (R Core Team, 2013).

1.2 Data Description

In this study, we used two data sets to investigate the prevalence rates and the risk factors for hypertension in Palestine because we were interested in identifying all possible risk factors of the disease. No single data set exists that covers all areas of Palestine and all socio-economic and health statuses, as well as medical records of Palestinian individuals with hypertension and without hypertension. The first data set (Palestinian Central Bureau of Statistics [PCBS], 2006) was a survey that covered all Palestinian localities, including the West Bank, the Gaza Strip, and East Jerusalem, and was highly representative of the distribution of the Palestinian population. The Palestinian Family Health Survey was conducted in 2006 by the PCBS (2006) in cooperation with the Pan Arab Project for Family Health (PAPFAM), UNICEF, and UNFPA and was based on a representative household sample survey of all Palestinian localities. The survey was designed to collect demographic and health data pertaining to the Palestinian population living in the Palestinian Territories, with a focus on demography, fertility, family planning, and maternal and child health, in addition to youth and senior citizens. The survey included supplementary sections and information on, for example, the basic health and socio-economic status of different groups within the population, including children less than five years and elderly individuals aged 60 years and over. The survey covered 13,238 Palestinian households from all Palestinian governments: 8,781 in the West Bank and 4,457 in the Gaza Strip. The survey's methodology was designed to take into account the conditions in Palestine, international standards, data processing requirements, and the comparability of the outputs with other related surveys conducted in Palestine. The second data set is a sample obtained from Jebril (2012). It comprised a random sample drawn from the clinical records of patients with chronic diseases, mainly hypertension and diabetes, in Palestine. The data were originally collected to study the characteristics of patients

with diabetes, but the records contain many variables related to hypertension and diabetes. The data also included a control group. The response variable was binary and took the value "yes or 1" if the patient suffered from hypertension and "no or 0" if not. A set of causes, including gender, age, smoking status, BMI, fasting blood sugar, glycated hemoglobin, microalbumin, urea, total cholesterol, and HDL, were available and used as independent variables.

We were interested in applying relevant statistical models that could correctly predict hypertensive patients with the greatest probability and based on significant risk factors in Palestine. The aim was to identify the most influential variables that could best classify these patients and to determine the association of these variables with the prevalence of hypertension in Palestine. Thus, all relevant variables of interest in both data sets were separated and split into two data sets. They were independent of each other because they were gathered from completely different sampling frames. However, the data sets complemented each other, as the first set included the socio-economic and environmental factors, and the second set contained the health and medical factors.

1.3 Statement of the Problem

Currently, there is insufficient information and studies conducted on the prevalence of hypertension in the Palestinian Territories. The problem of this study was to assess and apply the logistic regression model, as a classification model, to classify patients with high hypertension in Palestine using available data published by different sources, mainly the Palestinian Central Bureau of Statistics, and to propose the use of a powerful statistical model that can classify patients with hypertension and identify the most influential risk factors for the disease in Palestine.

1.4 Methodology

In this study, two data sets that involve all possible risk factors for the prevalence of hypertension in Palestine were made available for the analysis. The purpose was to estimate the prevalence rates and to identify significant risk factors for hypertension in Palestine. To achieve this goal, stepwise logistic regression analysis was conducted on the two data sets to classify cases of hypertension in Palestine (Yamashita, Yamashita & Kamimura, 2007). The identified models were assessed for goodness of fit to the data, and the model diagnoses were checked. The odds ratios were estimated and assessed for all the risk factors for hypertension in Palestine. The resulting models were also assessed in terms of their accuracy and their error rates to arrive at the best models for our data.

Table 1. Percentages of the variables of the first survey data.

Variables	**Percentage (N= 18701)**
Hypertension (diseased)	7.6
Sex of household member (male)	50.2
Age in complete years (older than 45 years old)	26.8
Educational status (illiterates)	8.9
Position in labor force (employed)	32.1
Smoking status (smokers)	24.4
Diabetes	5.9
Peptic ulcer disease	2.3
Cardiac disease	2.5
Renal disease	0.9
Hepatic disease	0.3
Arthritis (rheumatism)	4.4
Osteoporosis	0.7
Epilepsy	0.2
Asthma	1.2
High cholesterol	0.7
Depression	0,2
Hyperthyroidism (nonmalignant)	0.5
Glaucoma	0.7
Chronic back pain problems	1.9
Impaired vision	1.2
Impaired movement	1.9
Region (Gaza Strip)	37.8
Locality type (urban)	53.4
Wealth index quintiles (very poor)	20.6

2 Basic Indicators and Prevalence Rates

In this study, we used both data sets to assess all possible risk factors for hypertension in Palestine. The first data set was the Palestinian Family Health Survey published by the PCBS (2006). This data set contained more than 300 demographic and health indicators pertaining to the Palestinian population living in the Palestinian Territories, with special focus on demography, fertility, family planning, and maternal and child health, in addition to youth and senior citizens. The survey included a sample of 38,654 individuals of different ages from 13,238 randomly selected Palestinian households in all Palestinian governorates: 8,781 in the West Bank and 4,457 in the Gaza Strip. Whether the individual suffered from hypertension was selected as the response variable in the present study, in addition to 40 relevant variables. These variables included age, sex, smoking, governorate, years of schooling, position in the labor force,

region, wealth index, presence of diseases, including diabetes and chronic diseases, and disabilities. Table 1 contains percentages of the important variables in this data set. There were many binary variables, and all variables in Table 1 were transferred to binary variables and coded as 0 and 1. The averages of these variables were used to provide estimates for the proportion of cases that took the value 1. This is because the average of a binary variable equals the proportion of cases that take the value 1.

The second data set contained 21 independent variables and 180 records. The variables are sex, smoking, BMI, fasting blood sugar, microalbumin, urea, total cholesterol, glycated hemoglobin, High-density lipoprotein, triglyceride, age, LDL, CR, wt1, ht, GA, T.P, Alb, GP, STAT, and region. All independent variables were investigated; however, only 11 variables were relevant to hypertension. Table 2 contains the descriptive statistics of these variables.

Okasha and Abu Samra (2013) examined different possible classification techniques that can be applied to similar data sets and found that the logistic regression model is the best model for analyzing data sets with mixed types of independent variables and a binary response variable in terms of their accuracy and error rates. Other studies, including those by Kuan (2006), Arana, Mart-Bonmart, Paredes, and Bautista (1998), Yarmohammadi, Abdolmaleki, and Gity (2004), Kumari and Godara (2011), and Kurt, Ture, and Kurum (2008), arrived at a similar conclusion. In Section 3, we discuss the results of fitting the logistic regression model to both data sets. We also assess the accuracy of each model and discuss the diagnostic checks of the model as described by Agresti (2007).

Table 2. Descriptive statistics of the second sample data.

Variables	**Description**	**Min.**	**Max.**	**Mean**	**S.D.**
Age	Age of the patient	39	70	54.75	7.04
BMI	Body mass index	19	47.1	30.24	5.34
FBS	Fasting blood sugar	10	338	142.60	67.44
MAU	Microalbuminurea	5	1920	171.47	372.69
UR	Urea	17	184	49.86	40.32
T.CH	Total cholesterol	79	289	159.09	44.84
HBA1C	Glycated hemoglobin	4.1	8	6.12	1.0814
HDL	High-density lipoprotein	23	83	47.855	11.358
T.G	Triglyceride	10	461	153;089	60.624
Smoking	Smoking status	Yes		20.6%	
		No		79.4%	
Sex	Sex of patient	Male		46.7%	
		Female		53.3%	
Hypertension	Hypertension status	Yes		24.4%	
		No		75.6%	

The first data set was composed of 38,668 individuals from randomly selected

households. To make it representative of all the Palestinian population, the PCBS assigned each observation a weight representing its size in the whole population, and the weights were provided as an extra variable. Therefore, all observations were weighted by this variable, and the weighted averages were computed when the means and proportions were estimated. The estimated overall prevalence rate of hypertension in the Palestinian community was 3.7% of the total population, with 1414 individuals found to have hypertension at all ages and in all types of localities in the sample. The distribution of all hypertensive cases in Palestine in the different types of localities seemed to be symmetric. The estimated prevalence rate in each type of locality is given in Table 3. Examining the age distribution of those with hypertension, there was only one case less than 18 years old, despite more than half the Palestinian population (51.6%) under the age of 18 years. The age distribution of hypertensive cases in Palestine is shown in Table 4. Therefore, we decided to limit our study only to the adult population aged 18 years and over, yielding a sample size of 18,701. Among these individuals, there were 1413 cases of hypertension, giving an estimated prevalence rate of hypertension among the Palestinian adult population of 7.6%. If we considered only the older population aged 45 years of age and above, the estimated rate was 24.4%. The subsequent analysis was based on the adult population aged 18 years and over.

Table 3. Distribution of the Palestinian population in the sample according to the type of locality and the presence of hypertension.

Locality type		**Presence of hypertension**
Urban	*N*	814
	% within locality type	4.0%
	% with hypertension	57.6%
Rural	*N*	332
	% within locality type	2.9%
	% with hypertension	23.5%
Refugee camp	*N*	268
	% within locality type	4.0%
	% with hypertension	19.0%
Total	*N*	1414
	% within locality type	3.7%
	% with hypertension	100.0%

3 Fitting the Logistic Regression Model to the First Data Set

One advantage of the logistic regression model is that there are no assumptions made about the distributional properties of the independent variable or the predictor vari-

Table 4. Distribution of the Palestinian population in the sample according to age and the presence of hypertension.

Age group		**Presence of hypertension**
0-17	*N*	1
	% within age group	0.0%
	% with hypertension	0.1%
18-24	*N*	49
	% within age group	0.5%
	% with hypertension	3.5%
35-44	*N*	141
	% within age group	3.7%
	% with hypertension	10.0%
45+	*N*	1223
	% within age group	24.4%
	% with hypertension	86.5%
Total	*N*	1414
	% within age group	3.7%
	% with hypertension	100.0%

ables. The independent variables can be numeric or categorical or in any form. They do not have to be normally distributed, linearly related, or of equal variance within each group (Tabachnick & Fidell, 2013). The relationship between the predictor and the response variables is not a linear function in logistic regression: Instead, the logistic regression function is used, which is the logit transformation of π the probability of success and takes the following form:

$$\pi(\underline{x}) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k\right)} \tag{1}$$

where β is the coefficient of the predictor variables. An alternative form of the logistic regression equation can be expressed as:

$$\begin{aligned} \text{logit}\left[\pi(\underline{x})\right] &= \text{logit}\ \left[P\left(Y = 1 | \, X_1, ..., X_k\right)\right] = \log\left[\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right] \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \end{aligned} \tag{2}$$

where the *odds* $(\mathrm{Y} = 1) = \frac{\pi(\underline{x})}{1-\pi(\underline{x})} = \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k\right)$.

The goal of logistic regression is to predict the category of outcome for individual cases using the probability of a specific case to fall into each category and to identify the most parsimonious model. To accomplish this goal, a model is created that includes all the predictor variables that are useful in predicting the response variable. The variables can be entered into the model in the order specified by the researcher, or the logistic regression can test the fit of the model after each coefficient is added or deleted

in a procedure called a stepwise regression analysis (Hosmer & Lemeshow, 2000 and Yamashita, Yamashita & Kamimura, 2007).

The first step of the analysis of the hypertension data involved examining and filtering the first data set and omitting the invalid cases according to the definition of the study population, particularly their age, as our study population comprised only Palestinians living inside the Palestinian Territories aged 18 years and over. The aim was to identify the logistic regression model that best fits the data. Using the null model, which does not include a predictor variable, the overall percentage of correctly classified cases was 92.4%. This is the percentage of cases in the sample who did not have high hypertension. When we fitted the full model, which contained all 40 available predictors, the correct classification rate became 94.2%. We then applied the stepwise selection procedure using the AIC to find the best subset of predictors that could correctly classify hypertensive cases in the sample. Some 20 variables only were kept in the model, and this did not affect the ability of the model to correctly classify hypertensive cases in the sample. The variables shown to have significant effects on the classification results and that were included in the stepwise selected model, together with the estimates of the model parameters, are shown in Table 5. The AIC of the full model was 4198.4 and reduced in the final model to only 4177.8. In the final model, only eight predictors had coefficients significantly different from zero. Those predictors were age, sex, smoking status, peptic ulcer disease, arthritis rheumatism, high cholesterol, hyperthyroidism (nonmalignant), and locality type (rural and refugee camps). The final model can be expressed using equation (2) and the estimates of its parameters that appear in Table 5.

Table 5. Estimates of the parameters of the final logistic regression model, odds ratios, and the significance of hypertension in Palestine.

Variables	**Estimate**	**Std. error**	***z* value**	**Pr (>|z|)**	**Odds ratio**	**Lower 95% conf. limit**	**Upper 95% conf. limit**
Constant	-6.123	0.170	-36.02	0.000	0.002	0.002	0.003
Sex (male)	-0.747	0.080	-9.30	0.000	(2.111)	1.804	2.472
Age	0.080	0.002	41.34	0.000	1.083	1.079	1.087
Smoking	0.063	0.027	2.28	0.022	1.065	1.009	1.124
Peptic ulcer disease	0.510	0.145	3.51	0.001	1.666	1.253	2.215
Arthritis rheumatism	0.298	0.101	2.96	0.003	1.348	1.106	1.642
High cholesterol	1.865	0.217	8.58	0.000	6.453	4.214	9.879
Hyperthyroidism (nonmalignant)	1.040	0.293	3.55	0.000	2.830	1.594	5.024
Locality type (rural)	-0.402	0.077	-5.22	0.000	(1.495)	1.286	1.739
Locality type (refugee camp)	0.159	0.085	1.88	0.060	1.173	0.994	1.384

The predictors in Table 5 are the most important risk factors affecting the preva-

lence of hypertension in Palestine. They had a significant effect on the classification accuracy and statistically significant coefficients. The overall classification accuracy of the model was 92.1%. However, the model classified only 212 of 1413 hypertensive cases in Palestine, whereas the full model successfully classified 756 cases. The stepwise model successfully classified 744 cases, with an overall classification accuracy of 94.14%. This means that although some coefficients were not significant, the classification accuracy significantly increased with the inclusion of the 20 variables. To further assess the goodness of fit of the model, the Hosmer and Lemeshow test was applied, and an analysis of variance table was constructed for all the significant risk factors (Tabachnick & Fidell, 2013). The results are shown in Table 5. They show that the model is highly representative of the data.

These results indicate that among the 40 predictors examined as possible risk factors for the prevalence of hypertension in the first data set, only eight were highly significant predictors, and 12 variables showed a less significant relation with the prevalence of hypertension in Palestine. The 12 variables were educational status, cardiac disease, osteoporosis, epilepsy, asthma, depression, nervous illness, glaucoma, chronic back pain/spinal cord problems, chronic disease, visual impairment, and the wealth index. These variables may have indirect effects on the prevalence rate of hypertension in Palestine through their relation with other more important risk factors, such as the cholesterol level or hyperthyroidism. Therefore, we concluded that eight of the underlying predictors were significant risk factors for the prevalence rate of hypertension in Palestine. The other 12 variables may be indirectly related to hypertension through other unknown risk factors.

4 Fitting Logistic Regression Model to the Second Data Set

The second data set was a random sample drawn from the clinical records of patients with chronic diseases in Palestine. The data were originally amassed to study patients with diabetes. Although the records contained many variables related to hypertension and diabetics, as well as a control group, due to the small sample size and the set of predictors, the prevalence rates of hypertension in Palestine cannot be estimated. However, we fitted a logistic regression model to the data and identified the risk factors for hypertension in the region. Analogous to the analysis of the first data set, when we fitted the full model containing all 20 available predictors, the rate of hypertension was 94.4%. We then applied the stepwise selection procedure using AIC to find the best subset of predictors that could correctly classify hypertensive cases in the sample. Only six variables were kept in the model, without much loss of accuracy of the ability of the model to correctly classify hypertensive cases in the sample. The AIC of the full model was 93.31 and reduced in the final reduced model to only 70.33. The correct classification rate of hypertensive cases of the model was also reduced from 94.4% for the full model to 92.2% in the final model. The model classified 37

of 44 individuals as having hypertension, whereas the full model with all 20 variables classified 34 of 44 patients with hypertension. The variables that had significant effects on the classification results that were included in the stepwise selected model, together with the estimates of the model parameters, are shown in Table 6. In the final model, only four predictors had coefficients significantly different from zero: sex, fasting blood sugar, microalbuminurea, and smoking status. The final model can be expressed using equation (2) and the estimates of its parameters that appear in Table 6. Note that the variables sex and smoking status appeared to be significant in the analysis of both data sets.

Again, the four predictors are the most important risk factors affecting the prevalence of hypertension in Palestine, and they had a significant effect on the classification accuracy and statistically significant coefficients. The overall classification accuracy of the model was 92.8%, which is slightly higher than the overall correct classification rate of the stepwise model. However, the model classified only 31 of 44 hypertensive cases in Palestine, whereas the full model classified 37 cases. The stepwise model classified 34 cases, with an overall classification accuracy 92.2%. This means that although some coefficients were not significant, the classification accuracy increased significantly with the inclusion of two additional variables, HDL and glycated hemoglobin, in the model. To further assess the goodness of fit of the model, the Hosmer and Lemeshow test was applied, and the results are shown in Table 6.

Table 6. Estimates of the parameters of the full logistic regression model, their significance, and the odds ratios to the second set of hypertension data in Palestine.

Variables	**Estimate**	**Std. error**	***z* value**	**Pr (>|z|)**	**Odds ratio**	**Lower 95% conf. limit**	**Upper 95% conf. limit**
Constant	-2.783	0.923	-3.014	0.003	16.164	2.647	98.715
Sex (male)	-1.103	0.727	-1.516	0.130	(3.012)	1.381	12.526
Fasting blood sugar (FBS)	0.011	0.004	2.549	0.011	1.011	1.003	1.020
Microalbuminurea (MAU)	0.018	0.015	1.167	0.243	1.018	0.988	1.048
Smoking status (smoker)	-1.641	0.699	-2.347	0.019	5.158	1.310	20.306

These results indicate that among the 20 predictors examined as possible risk factors for the prevalence of hypertension in Palestine, four were highly significant, and two were less significant. The variables HDL and glycated hemoglobin may have indirect effects on the prevalence rate of hypertension in Palestine through their relation

with other more important risk factors. Therefore, we concluded that there are four significant risk factors for the prevalence rate of hypertension in Palestine among all the underlying predictors. The other two variables may be indirectly related to hypertension through other unknown risk factors. Other variables such as the BMI, fasting blood sugar, and total cholesterol did not contribute significantly to the second model. The odds of a patient having hypertension were estimated for each risk factor using the invariance property of the odds ratio with respect to interchanging the categories of some risk factors and indicated between brackets. The odds according to gender were 3.0119 for women, meaning women were 3.0119 more likely to have hypertension compared to men. Smokers were 5.1582 times more likely to have hypertension compared to nonsmokers, all other factors being equal.

5 Goodness of Fit Tests and Diagnostic Checks of the Model

For both models for the two data sets, we checked the accuracy of the model and its goodness of fit by computing the Hosmer and Lemeshow test of the goodness of the model fit and the Nagelkerke R-square values. In the first model, the Hosmer and Lemeshow test indicated that the value of the chi-square test statistic equaled 5.778, with 8 degrees of freedom, and that the p-value equaled 0.672, suggesting that the first model provided a good fit for the first data set. The Nagelkerke R-Square value provides an indication of the amount of variation in the dependent variable explained by the model. For this first model, it was 0.6524, meaning that 65.24% of the variability was explained by the set of variables entered in the first model.

For the second model of the second data set, to test the goodness of fit, we performed the Hosmer and Lemeshow goodness of fit test. The value for the chi-square test statistic was 10.642, with 8 degrees of freedom. The p-value was 0.223. These findings suggest that the model was appropriate and that it provided a good fit for the second data set. The Nagelkerke R-square value was 0.7371. This means that 73.71% of the variability in the response variable was explained by that set of independent variables. In other words, 73.71% of the amount of variation in the dependent variable was explained by the model that included only four risk factors.

Using the rates of correct and false classification of each logistic regression model fitted for the two data sets, we computed additional statistics frequently reported in the medical literature, such as by Vuk and Curk (2006). The confusion matrix is a matrix representation of the classification results. The indicators that can be computed from the confusion matrix of the logistic regression and other classification techniques are sensitivity, specificity, accuracy, precision, and error rates. They indicate the accuracy of the fit of the models to each data set. The receiver operating characteristic (ROC) curve, a technique for visualizing, organizing, improving, and assessing models based on their performance, can be drawn from those indicators. This representation has a practical purpose because it enables a binary classifier to be constructed for each point

on the convex hull. A convex hull is just one approach for constructing a ROC curve from a given set of points (Centor, 1991). The area under the ROC curve is often used as a measure of the quality of a probabilistic classifier. It is close to the perception of classification quality that most people have. The indicators were computed for each model, and the ROC curve was drawn.

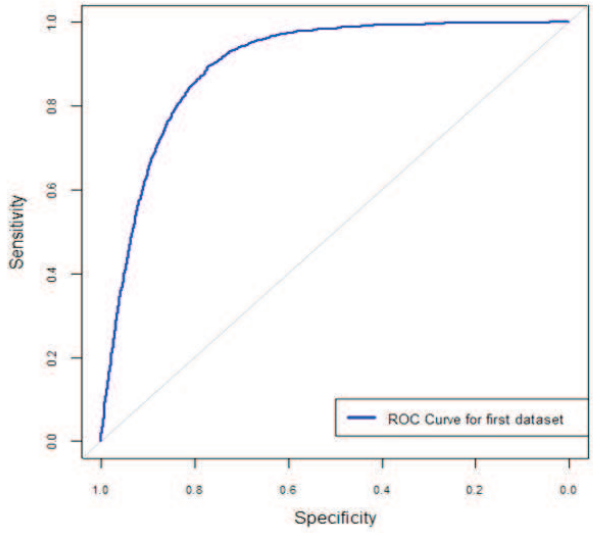

Figure 1: ROC curve of the logistic regression model for the first set of data.

The accuracy of the final logistic regression model for the first data set was 92.09%. As shown in Figure 1, the area under the ROC curve for the logistic regression was 89.64% (with a 95% confidence interval 88.98% to 90.31%). The error rate for the logistic regression was 7.81%. However, the accuracy of the final logistic regression model for the second data set was 93.95%, and the area under the ROC curve for the logistic regression was 93.95% (with a 95% confidence interval 90.23% to 97.67%) (Figure 2). The error rate for the logistic regression was 6.05%. We can conclude from the two ROC curves and the accuracy of the logistic regression model of both data sets that the logistic regression model was very efficient in depicting the prevalence of hypertension in Palestine.

The results discussed in sections 3 and 4 indicate that both models fit very well for both data sets and that they could efficiently be used to identify the most effective independent variables in the models and the most serious risk factors for the prevalence of hypertension in Palestine. The results also show that sex and smoking status were common and important risk factors in both models. The type of locality was less important, although it appeared from the first model that living in a rural area or a refugee camp were associated with a greater probability of having hypertension in

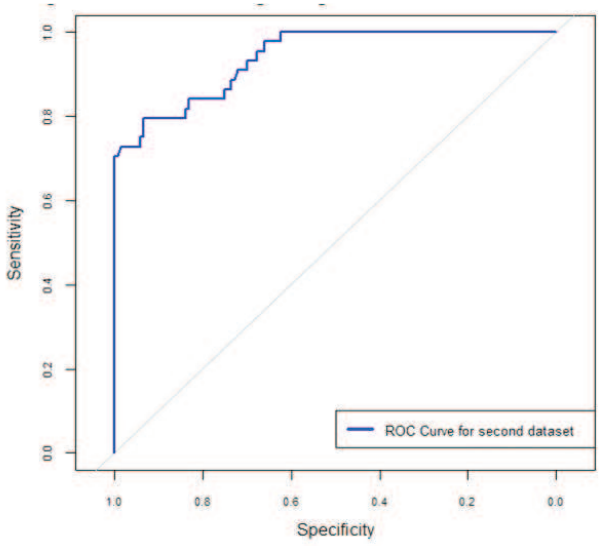

Figure 2: ROC curve of the logistic regression model for the second set of data.

Palestine. However, there was no evidence that the prevalence rates in different Palestinian governorates were significantly different, and no significant difference between the rates of hypertension in the West Bank and the Gaza Strip.

6 Conclusion

Our results indicate that the estimated overall prevalence rate of hypertension in Palestine is 3.7%. There appears to be no significant difference in the distribution of hypertensive cases in Palestine in different localities. In terms of the age distribution of those with hypertension, there was only one case less than 18 years old in the sample. Therefore, only individuals aged 18 years and older were included in the analysis, yielding a sample size 18,701, of whom 1413 had hypertension. This gives an estimated prevalence rate of hypertension among the Palestinian adult population of 7.6%. If we include only older individuals aged 45 years of age and above, the estimated rate is 24.4%.

According to the main results of the first logistic regression model, the major risk factors for hypertension in Palestine were age, sex, smoking status, peptic ulcer disease, arthritis rheumatism, high cholesterol, hyperthyroidism (nonmalignant), and locality type (rural and refugee camps). Other less significant risk factors included educational status, cardiac disease, osteoporosis, epilepsy, asthma, depression, nervous illness, glaucoma, chronic back pain/spinal cord problems, chronic disease, visual impairment, and the wealth index. However, the second logistic regression model gen-

erated from the second data set indicated that the main risk factors for hypertension in Palestine were sex, fasting blood sugar, microalbuminurea, and smoking status. Two other less important risk factors were HDL and glycated hemoglobin. Thus, in addition to other factors, there are ten significant risk factors for hypertension in Palestine: age, sex, smoking status, peptic ulcer disease, arthritis rheumatism, high cholesterol, hyperthyroidism (nonmalignant), fasting blood sugar, microalbuminurea, and locality type (rural and refugee camps).

According to the odds ratios for various risk factors for hypertension in Palestine, the odds ratio for the prevalence of hypertension of gender is 3.0119, indicating that women were 3.0119 times more likely to have hypertension compared to men. The odds ratio of smoking is 5.1582, indicating that smokers were 5.1582 times more likely to have hypertension compared to nonsmokers.

References

[1] Abed, Y., & Abu-Haddaf, S. (2013). Risk factors of hypertension at UNRWA primary health care centers in Gaza governorates. *ISRN Epidemiology*, 1-9.

[2] Agresti, A. (2007). *Building and applying logistic regression models; An introduction to categorical data analysis*. Hoboken, NJ: Wiley.

[3] Arana, E., Mart-Bonmat, L., Paredes, R., & Bautista, D. (1998). Focal calvarial bone lesions. Comparison of logistic regression and neural network. *Investigative Radiology, 33,* 738–745.

[4] Centor, R. M. (1991). Signal detectability: The use of ROC curves and their analyses. *Medical Decision Making, 11*(2), 102-106.

[5] Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.

[6] Jebril, M. A. (2012). *Glycated albumin as a sensitive indicator of glycemic control compared with HbA1c among type 2 diabetes mellitus patients* (Unpublished doctoral dissertation). Islamic University, Bangladesh.

[7] Kuan, C.-M. (2006). *Artificial neural networks* (IEAS Working Paper No. 06-A010).

[8] Kumari, M., & Godara, S. (2011). Review of data mining classification models in cardiovascular disease diagnosis. *International Journal of Computer Science and Technology, 2*(2), 304-305.

[9] Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications: An International Journal, 34*(1), 366-374.

[10] Macedo, M. E., Lima, M. J., Silva, A. O., Alcantara, P., Ramalhinho, V., & Carmona, J. (2005). Prevalence, awareness, treatment and control of hypertension in Portugal: The PAP study. *Journal of Hypertension, 23*(9), 1661-1666.

[11] Mishra, C. P., & Kumar, S. (2011). Risk factors of hypertension in a rural area of Varanasi. *Indian Journal of Preventive and Social Medicine, 42*(1), 101-111.

[12] Nakanishi, N., Li, W., Fukuda, H., Takatorige, T., Suzuki, K., & Tatara, K. (2003). Multiple risk factor clustering and risk of hypertension in Japanese male office workers. *Industrial Health, 41*(4), 327-331.

[13] Okasha, M. K., & Abu Samra, A. I. (2013). Comparison between logistic regression and neural networks in classifying hypertension patients in Palestine. *Proceedings of the 24th Annual International Conference on Statistics and Computer Modeling in Human and Social Sciences, To appear.*

[14] Palestinian Central Bureau of Statistics. (2006). *Palestinian family health survey.* Ramallah, Palestine: Author.

[15] Parati, G., Di Rienzo, M., Ulian, L., Santucciu, C., Girard, A., Elghozi, J.-L., & Mancia, G. (1998). Clinical relevance of blood pressure variability. *Journal of Hypertension, 16*(suppl 3), 25-33.

[16] R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/.

[17] Ruixing, Y., Limei, Y., Yuming, C., Dezhai, Y., Weixiong, L., Muyan, L., Fengping, H., . . . Zhenbiao, N. (2006). Prevalence, awareness, treatment, control and risk factors of hypertension in the Guangxi Hei Yi Zhuang and Han populations. Hypertension Research, *29*(6), 423-432.

[18] Sarraf-Zadegan, N., Boshtam, M., Mostafavi, S., & Rafiei, M. (1999). Prevalence of hypertension and associated risk factors in Isfahan, Islamic Republic of Iran. Eastern Mediterranean Health Journal, *5*(2), 992-1001.

[19] Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Allyn and Bacon.

[20] Tee, S. R., Teoh, X. Y., Aiman, W. A. R. W. M., Aiful, A., Har, C. S. Y., Tan, Z. F., & Khan, A. R. (2010). The prevalence of hypertension and its associated risk factors in two rural communities in Penang, Malaysia. *IeJSME, 4*(2), 27-40.

[21] Vuk, M., & Curk, T. (2006). ROC curve, lift chart and calibration plot. *Metodolo ki zvezki, 3*(1), 89-108.

[22] Wilkins, K., Campbell, N. R., & Joffres, M. R. (2010). Blood pressure in Canadian adults. *Health Reports, 21*(1), 37-46.

[23] Yadav, S., Boddula, R., Genitta, G., Bhatia, V., Bansal, B., Kongara, S., . . . Bhatia, E. (2008). Prevalence & risk factors of pre-hypertension & hypertension in an affluent north Indian population. *Indian Journal of Medical Research, 128,* 712-720.

[24] Yang, J., Wang, J. H., Zhi, X. Y., Zhu, H., & Liu, X. M. (2011). Prevalence rate of hypertension and related risk factors in populations of Tianjin. *Zhonghua Liu Xing Bing Xue Za Zhi, 32*(3), 239-243.

[25] Yamashita, T., Yamashita, K. & Kamimura, R. (2007). A Stepwise AIC Method for Variable Selection in Linear Regression", *Communications in Statistics - Theory and Methods*, 36 (13), 2395-2403.

[26] Yarmohammadi, M., Abdolmaleki, P., & Gity, M. (2004). Comparison of logistic regression and neural network models in predicting the outcome of biopsy in breast cancer from MRI findings. *Iranian Journal of Radiation Research, 1*(4), 217-228.