RISK FOR DIABETES AMONG TEXANS

THESIS

Presented to the Graduate College of
Texas State University – San Marcos
In Partial Fulfillment of
The Requirements

For the Degree of
Master of SCIENCE

By

Jie Li, B.S.

San Marcos, Texas
December 2003

ACKNOWLEDGEMENTS

I would like to thank the following people: Ram Shanmugam, Charles Johnson, Jean Brender, Wen Chui and Weihua Li.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

RISK FOR DIABETES AMONG TEXANS

By

Jie Li, B.S.
Texas State University
December 2003

Supervising Professor: Ram Shanmugam

Diabetes is a disease in which a person's body fails to properly use and store glucose. Glucose is retained in the bloodstream, and consequently causes the person's blood glucose or "sugar" to rise too high. There are two major types of diabetes: Type 1 diabetes and Type 2 diabetes. With Type 1 diabetes, the body completely stops producing insulin. People with Type 1 diabetes must take daily insulin injections. With Type 2 diabetes, the body produces insufficient amount of insulin to convert food into energy, or in the case of insulin-resistance, the body can't properly use the insulin it does produce. The American Diabetes Association reports about 17 million people (6.2% of the population) suffer from diabetes and many of them are not even aware that they have the disease. Diabetes was the sixth-leading cause of death according to analysis of 1999 U.S. death certificates (Diabetes Week, February 3, 2003). Each year, an estimated 12,000–24,000 people become blind because of diabetic eye disease. In addition, more than 38,000 people with diabetes begin treatment for kidney failure each year, and about

86,000 undergo diabetes-related lower extremity amputations. The total of direct and indirect costs of diabetes in the US is nearly $132 billion a year. In Texas, diabetes contributed 13,553 deaths in 1998 and 15,130 deaths 2000. There were 911,039 diagnosed diabetes patients (about 6.2% of the adult population in the age of 18 years or older) and 450,504 people with undiagnosed diabetes (about 3.6% of the adult population in the age of 20 years or older) in Texas in 2001 (TDH Diabetes Council Report, 2001). Because diabetes is a serious, costly, and increasingly common chronic disease that can cause devastating complications and often result in disability and death, appropriate predictors need to be identified. The probability that a person will develop diabetes needs to be estimated, so that people can formulate healthy lifestyles to reduce their risk for developing diabetes and its complications.

Many studies have been done to determine the risk factors for diabetes (Anastasia C. Thanopoulou, et al., 2003; Karin M. Nelson, et al., 2002; Bahman P. Tabaei, William H. Herman, 2002; HU FB Sigal RJ, Rich-Edwards JW, et al., 1999; Hu FB, van Dam RM, Liu S., 2001). Identified non-modifiable risk factors include age, ethnicity and family history of diabetes; identified modifiable risk factors include diet, obesity, physical inactivity, alcohol consumption, tobacco smoking and hypertension. These genetic, environmental, and metabolic risk factors are interrelated and contribute to the development of type 2 diabetes. Multivariate logistic regression has been used in previous diabetes studies (O. Rolandsson, et al. 2001; Philip S. Mehler, et al., 1998; Bahman P. Tabaei, William H. Herman, 2002). Findings from these studies indicated that age, race, body mass index, physical activity, alcohol consumption and family history of diabetes were significant predictors for diabetes.

The purpose of this study was to identify predictors of diabetes and to estimate the probability of diabetes prevalence in Texas. This study used existing data made available by the Texas Department of Health (TDH). Data were collected by Texas Behavioral Risk Factor Surveillance System (BRFSS) in a 2001 survey. Texas BRFSS used the Centers for Disease Control and Prevention "2001 Behavioral Risk Factor Surveillance System Questionnaire" with disproportionate stratified random sampling (DSS) to collect data. There were 5916 participants older than 18 years. Seventeen variables listed in Appendix A were entered into logistic regression model. Of these 5916 participants, 1221 had no missing values in the 17 variables used as analysis sample to generated final logistic model. For the purpose of validation of the final logistic regression equation, the final model was applied to TDH BRFSS survey data of year 1999. Fourteen variables listed in Appendix B were entered into the validation model. Of the 4990 survey participants in year 1999, 1633 had no missing values in the above 14 variables and were used as a validation sample. Stepwise selection identified age, race, blood pressure, blood cholesterol, and body mass index as significant predictors. The final model for prediction of probability of diabetes presence was:

Log $[\hat{P} / (1 - \hat{P})] = -3.3389 + 1.1776$ (Age_55-64 years) + 1.1505 (Age_64+ years) + 1.1763 (Non-Hispanic black) + 0.9279 (Hispanic) + 0.5611 (High blood pressure) + 0.5691 (High blood cholesterol) – 0.8746 (Alcohol drink status_yes) – 0.6198 (Leisure time physical activity_yes) + 0.6968 (Obese),

Where as $\hat{P}$ = estimated probability of presence of diabetes

This study showed that diabetes presence was more likely associated with age groups older than 55 years than with 18 – 34 years group (adjusted odds ratio = 3.247, with 95% CI: 1.793 – 5.878 for age 55 - 64 years group; adjusted odds ratio = 3.160, with 95% CI: 1.731 – 5.769 for 65+ age group), with being non-Hispanic black (adjusted odds ratio = 3.242, with 95% CI: 1.637 – 6.422) and Hispanic (adjusted odds ratio = 2.529, with 95% CI: 1.417 – 4.513) rather than being non-Hispanic white; with high blood pressure (adjusted odds ratio = 1.753, with 95% CI: 1.005 – 2.911) and high blood cholesterol (adjusted odds ratio = 1.767, with 95% CI: 1.079 – 2.892) rather than normal blood pressure and normal blood cholesterol; with obesity (adjusted odds ratio = 2.007, with 95% CI: 1.237 – 3.256) rather than body mass index less than 25. This study also found that adjusted odds ratios were 0.417 (95% CI: 0.249 – 0.700) for alcohol drinkers and 0.538 (95% CI: 0.330 – 0.878) for people who did leisure time physical activity, indicating that appropriate consumption of alcohol and physical activity protect people from diabetes.

By applying the estimated coefficients in the final equation, a person's probability to be a diabetic can be calculated. For example, a physically active non-Hispanic white person who is a moderate level alcohol drinker aged younger than 55 years, with normal blood pressure, normal blood cholesterol, BMI less than 25, only has a 0.79% probability to be a diabetic. For a physically inactive non-Hispanic black person who is alcohol non-drinker older than 65 years, with high blood pressure, high blood cholesterol, and body mass index greater than 30, the predicted probability to be a diabetic will increase to 69.32%. However, this study suggests that when apply this prediction equation to population, prevalence of diabetes will be underestimated because lack of other important predictors in the final model. Therefore, a threshold of predicted probability of 0.10 generated from ROC curve analysis should be considered for the purpose of prevention

and early detection. Individuals who had predicted probability greater than this threshold could be identified as a group of people at high risk, and they need further medical diagnostic investigation. When this model is applied to the population, a lower (higher) cutpoint of predicted probability would be considered to meet expectation of a higher (lower) sensitivity and a lower (higher) specificity.

In this sample, more than 70% of participants had a low intake (less than 5 servings per day) of fruit and vegetable, leading to a crude odds ratio of 0.77 with 95% confidence interval: 0.49 – 1.21 for group of people who had a low intake of fruit and vegetable, compared to group of people who had more than 5 servings of fruit and vegetable per day. This result could be generated from survey information bias. Because the survey questionnaire only considered information 'in the past 30 days', some diabetics diagnosed before this period, might have changed their habits and eaten more fruits and vegetables after being diagnosed with diabetes. This finding implies that the data should include information before and after diabetes had been diagnosed to minimize the information bias.

The final logistic regression model for this study yields a relatively low coefficient of determination (0.22). This result suggests that variables not accounted for in this study, such as family history, vitamin supplements, and other related chronic diseases, might explain a significant proportion of the variance in diabetes presence. A recommendation for future study is that information about family history, vitamin supplements and comorbidities should be collected and analyzed in future studies. A more accurate predicted probability of diabetes presence will be generated by the enrichment of the information.

CHAPTER 1

INTRODUCTION

1.1 Diabetes

Diabetes is a disease in which a person's body fails to properly use and store glucose (a form of sugar). Glucose is retained in the bloodstream, and consequently causes the person's blood glucose or "sugar" to rise too high. There are two major types of diabetes: Type 1 diabetes (also called juvenile-onset or insulin-dependent diabetes) and Type 2 diabetes (also called adult-onset or non insulin-dependent diabetes). With Type 1 diabetes, the body completely stops producing insulin, a hormone that enables your body to use the glucose found in foods. People with Type 1 diabetes must take daily insulin injections. This form of diabetes typically develops in children or young adults, but it can occur at any age. With Type 2 diabetes, the body produces insufficient amount of insulin to convert food into energy, or in the case of insulin-resistance, the body can't properly use the insulin it does produce. This form of diabetes usually occurs in people who are over 40, overweight, and have a family history of diabetes.

Diabetes is a serious, costly, and increasingly common chronic disease that can cause devastating complications including nerve damage, kidney failure, cardiovascular (heart and lung) disease, eye damage and blindness, poor healing of infections and wounds, periodontal (tooth and gum) disease, and impotence in men. The complications from diabetes often result in disability and death. The mortality rate among diabetic patients is four times higher than the rate in non-diabetic subjects (Morgan C, Currie C, Peters J, 2000). Diabetes may lead to an increased risk of developing and dying from an infectious disease. A retrospective study showed that nearly half of all people with diabetes had at least one hospitalization or physician claim for an infectious disease. The risk ratio for diabetic versus non-diabetic people was 1.21 (99% confidence interval: 1.20-1.22) (Baiju R. Shah, Janet E. Hux, 2003). Twenty-four cohort studies from Asia, Australia, and New Zealand showed that the rapidly growing prevalence of diabetes in Asia heralds a large increase in the incidence of diabetes-related death in the coming decades (The hazard ratio associated with diabetes was 1.97, 95% confidence interval: 1.72-2.25) (Asia Pacific Cohort Studies Collaboration, 2003).

The American Diabetes Association reports about 17 million people (6.2% of the population) suffer from diabetes and many of them are not even aware that they have the disease. Diabetes was the sixth-leading cause of death according to analysis of 1999 U.S. death certificates (Diabetes Week, February 3, 2003). The diagnosed cases of diabetes (including gestational diabetes) increased from 7.3% to 7.9% during period of 2000-2001. This increase prevailed across sex, age, race and educational status. (Diabetes Week, January 20, 2003). Each year, an estimated 12,000–24,000 people become blind because of diabetic eye disease. In addition, more than 38,000 people with diabetes begin treatment for kidney failure each year, and about 86,000 undergo diabetes-related lower extremity amputations. The total of direct and indirect costs of diabetes in the US is nearly $132 billion (Indirect costs: $40.2 billion (disability, work loss, premature mortality), direct medical costs: $91.8 billion) a year (National Diabetes Statistics, 2003).

In Texas, diabetes contributed 13,553 deaths in the year 1998 and 15,130 deaths in the year 2000, based on Texas death certificate data, and it is believed to be under-reported on death certificates, both as a condition and as a cause of death. The average mortality rate per county was 23.0 per 100,000 during 1990 through 1998. There were 911,039 diagnosed diabetes patients (about 6.2% of the adult population in the age of 18 years or older) and 450,504 people with undiagnosed diabetes (about 3.6% of the adult population in the age of 20 years or older) in Texas in 2001 (TDH Diabetes Council Report, 2001).

Type 2 diabetes results from genetic, behavioral and environmental interactions and is a preventable disease. The prevention idea was originally recommended by Dr. Elliott Joslin, a renowned diabetologist of the early 20th century, as early as 1921. Prevention is important because by the time diabetes is diagnosed by traditional methods, a person might have already developed heart disease, the number one killer in this country. Therefore, early detection, improved delivery of care, and better self-management are key strategies for preventing diabetes. Self-awareness of the risk of getting diabetes is important to protect people from developing diabetes and diabetes-caused death. The importance of risk factor identification is to initiate prevention measures and to improve outcomes in diabetes by screening, early diagnostics and treatment. Appropriate predictors need to be identified. The probability that a person will develop diabetes needs to be estimated, so that people can formulate healthy lifestyles to reduce their risk for developing diabetes and its complications.

1.2 Multivariate Logistic Regression

Multivariate logistic regression analysis introduced by McCullagh, P. and Nelder, J. A. (1989), also known as logit analysis, is a suitable statistical technique for analyzing diabetes. In multivariate logistic regression model, the dependent variable is dichotomous. The independent variables can be either categorical or continuous variables. Multivariate logistic regression equation is a linear equation, predicting the log odds. The odds is defined as the ratio of the probability that an event occurs to the probability that it fails to occur. The predictors do not have to be normally distributed, linearly related, or of equal variance within each group.

Multivariate logistic regression was employed for data analysis in this study. The multivariate logistic regression equation was:

$$\text{Logit } P = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

With the presence of diabetes scored as 1, P is the probability of a person was a diabetic and this probability is a function of the k independent variables $X_1, X_2, \ldots, X_k$. The parameter β_0 represents the log odds ratio of diabetes risk for a person with a standard ($X_1 = X_2 = \ldots X_k = 0$) set of independent variables, while β_i is a fraction by which the risk is increased (or decreased) for a unit change in X_i (D.W. Hosmer, Jr. and S. Lemeshow, 1988).

1.3 Purpose of the Study

The purpose of this study was to identify predictors of diabetes and estimate the probability of diabetes prevalence in Texas.

1.4 Research Questions

The following research questions were addressed in the topic:

1. What demographic, medical conditions and health behavioral characteristics can best predict that an individual is a diabetic?
2. What is the probability that a person is a diabetic based on the independent variables entering into a multivariate logistic regression model?

The secondary research questions that were examined prior to the final logistic analysis are:

1. What is the diabetes prevalence in each of demographic characteristics, medical conditions and health behavioral variables among Texans in 2001?
2. What are the crude odds ratios and multiple adjusted odds ratios for any significant associations between predictor variables and diabetes?
3. Are there associations that exist between a person's demographic characteristics, medical conditions or health behaviors and diabetes?
4. Does the diabetes prevalence significantly increase (or decrease) as a person's age, education level, household income, or body mass index increase?

1.5 Research Hypotheses

H_{01}: There is no association between each of a person's demographic characteristics, medical conditions or health behaviors and diabetes presence. Diabetes presence is independent from demographic, medical condition and health behavior variables.

$$H_{01}: \quad \pi_{ij} = \pi_{i.}\ \pi_{.j}$$

Where as i = row variable, demographic, medical condition or behavioral variable, for i=1, 2, · · ·, i

j = column variable, diabetic status, for j =1, 2. 1=diabetic, 2=non-diabetic

π_{ij} = the probability of a person observed on state ith row (demographic, medical conditions or behavioral group) and state jth column (diabetic status)

$\pi_{i.}$ = the probability of a person observed on state ith row (demographic, medical conditions or behavioral group)

$\pi_{.j}$ = the probability of a person observed on state jth column (diabetic status)

***H**$_{A1}$*: There are associations between each of a person's demographic characteristics, medical conditions or health behaviors and diabetes. Diabetes presence is not independent from demographics, medical conditions and health behaviors.

$$H_{A1}: \; \pi_{ij} \neq \pi_{i} \; \pi_{.j}$$

The *Chi*-square tests for independence of diabetes presence and demographic, medical conditions and health behavior variables were performed by analyzing two-way crosstabulations. Crude odds ratios and the 95% confidence intervals were calculated.

***H**$_{02}$*: The probability of diabetes presence remained the same as a person's age or body mass index increased.

$$H_{02}: \quad \pi_{1} = \pi_{2} = \cdot \cdot \cdot = \pi_{i}$$

Where as π_{i} = the conditional probability of a person observed on state ith age or body mass index category of a diabetes response.

***H**$_{A2}$:* The probability of diabetes presence increased as a person's age or body mass index increased.

$$H_{A2}: \quad \pi_1 < \pi_2 < \cdots < \pi_\imath$$

H₀₃: The probability of diabetes presence remained the same as a person's education level or household income increased.

$$H_{03}: \quad \pi_1 = \pi_2 = \cdots = \pi_\imath$$

Where as $\pi_\imath$ = the conditional probability of a person observed on state *i*th education level or household income category of a diabetes response.

H_{A3}*:* The probability of diabetes presence decreased as a person's education level or household income increased.

$$H_{A3}: \quad \pi_1 > \pi_2 > \cdots > \pi_\imath$$

The Cochran Armitage test is a method of directing *Chi*-Square tests toward narrow alternatives and is sensitive to the linearity between response variables and experimental variables, and detects trends that would otherwise be missed by more crude methods (Armitage P. 1955, Cochran WG. 1954). The Cochran Armitage test (based on df = 1) was used to detect these diabetes presence trends among age, education level, household income and body mass index categories.

H_{04} There is no significant relationship between diabetes presence and each one of demographic, medical conditions or health behavior variables.

$$H_{04} \quad \beta_\imath = 0$$

Where as $\beta_\imath$ = coefficient of *i*th independent variable in multivariate logistic regression equation.

If $\beta_\imath = 0$, it means that *i*th independent variable is not a significant predictor of diabetes presence.

$\boldsymbol{H_{A4}}$ There is a significant relationship between diabetes presence and each one of demographic, medical conditions or health behavior variables.

$$\boldsymbol{H_{A4}} \quad \beta_i \neq 0$$

If $\beta_i \neq 0$, it indicates that *i*th independent variable is a significant predictor of diabetes presence. Diabetes presence is significantly related to *i*th independent variable.

Multivariate logistic regression was used to evaluate these significant relationships. The dependent variable was diabetes presence status. Demographic characteristics, medical conditions, and health behaviors were the independent variables.

1.6 Study Limitations

1. The TDH 2001 survey data collected by the Texas Behavioral Risk Factor Surveillance (BRFSS) was used. The prevalence of diabetes and risk exposures were measured; however, no information is available on whether exposures occurred before or after diabetes had been diagnosed. The survey questionnaire only considered information 'in the past 30 days'. Some diabetics diagnosed before 'past 30 days' period, might have quit their harmful habits, such as smoking or heavy alcohol consumption. They may have become more physically active, or eaten more fruit and vegetables, in the 'past 30 days period' after they knew they have diabetes. This part of the data might mislead the detection of the associations between risk behaviors and diabetes occurrence.
2. At the time the data were collected, there could be some misclassified cases because their disease had not been diagnosed yet. The diabetes prevalence could be underestimated.

3. No questions were asked about family history. This important predictor of diabetes was not entered into logistic regression to predict probability of diabetes presence.

CHAPTER 2

REVIEW OF THE LITERATURE

2.1 Risk Factors of Diabetes

Many studies have been done to determine the risk factors for diabetes (Anastasia C. Thanopoulou, et al., 2003; Karin M. Nelson, et al., 2002; Bahman P. Tabaei, William H. Herman, 2002; HU FB Sigal RJ, Rich-Edwards JW, et al., 1999; Hu FB, van Dam RM, Liu S., 2001; etc.). There are a number of non-modifiable and modifiable risk factors. The following are non-modifiable risk factors:

Ethnicity

Ethnicity is an established risk factor of diabetes. The prevalence of diabetes is higher in African Americans than in non-Hispanic whites for all ages (American Diabetes Association. Diabetes, 2001). Data from the Third National Health and Nutrition Examination Survey, 1988-1994, showed that diabetes prevalence for non-Hispanic whites, non-Hispanic blacks, and Mexican Americans at 20 years of age or older were 4.8%. 8.2% and 9.3%, respectively (Harris MI, Flegal KM, Cowie CC, et al, 1998). According to National Diabetes Statistics (2003) report, on average, non-Hispanic black, Hispanic/Latino Americans and America Indians/Alaska Natives are 2 times, 1.9 times and 2.6 times, respectively, more likely to have diabetes than non-Hispanic whites of similar age. The Texas BRFSS 1996-1999 survey data report showed that Hispanics (6.6%) and African Americans (7.4%) had higher prevalence of diabetes than non-Hispanic white (4.7%) (Weihua Li, et al., 2001).

Family History

It is well accepted that Type 2 diabetes is an inherited condition. A positive family history strongly predisposes to the development of Type 2 diabetes (American Diabetes Association, 2003; S. Bo, P. Cavallo-Perin, L. Gentile, E. Repetti and G. Pagano, 2000). In a case-control study conducted by Anastasia C. Thanopoulou, et al. (2003), the study found that diabetics had a family history of diabetes more frequently than non-diabetics (49% vs 14.2%; p <0.001). Results from a cohort study (which followed 1,947 nondiabetic men for 22.5 years) indicated that maternal family history of diabetes showed a relative risk of 2.51 for diabetes (95% confidence interval: 1.55-4.07); paternal family history was associated with a relative risk of 1.41 (95% confidence interval: 0.657-3.05); and a combined maternal and paternal family history was associated with a relative risk of 3.96 (95% confidence interval: 1.22-12.9). The results indicated that family history appears to be an important risk factor for Type 2 diabetes (Bjornholt, Jorgen V. et al., 2000). The Framingham Offspring Study found that the risk for Type 2 diabetes among offspring of a single parent with diabetes was 3.5 times greater, and for those with two diabetic parents was 6 times greater when compared with offspring without parental diabetes (Meigs JB, Cupples LA, Wilson, PW. 2000).

Age

Age was identified as a risk predictor for diabetes (Bahman P. Tabaei; William H. Herman, 2002). Type 2 diabetes has been known for years as “adult onset,” or “maturityonset," emphasizing that the prevalence of Type 2 diabetes increases with age. In the year 2000, 20.1% of all people in the age group of 65 years or older have diabetes (National Diabetes Statistics, 2003). The Texas BRFSS 1966-1999 survey data report showed that prevalence of diabetes increases with age: the youngest age group (18-24 years old) had the lowest prevalence of 0.4% (95% confidence interval: -0.4-1.2), and the oldest age group (65 years or older) had the highest prevalence of 13.1% (95% confidence interval: 6.9-9.3) (Weihua Li, et al., 2001).

The following are modifiable risk factors:

Diet

A case-control study by analyzing Third National Health and Nutrition Examination Survey data showed associations between diets both high in saturated fat and low in fruit and vegetable intake and the presence of diabetes (Karin M. Nelson, et al., 2002). Another case-control study compared the distribution of diabetics and non-diabetics among the quartiles of animal fat intake in grams. The results showed that diagnosed and undiagnosed diabetics significantly clustered in the upper quartiles and non-diabetics in the lower quartiles in the distribution. The relative risk for having diabetes is 1.8 for recent diagnosed diabetes and 3.1 for undiagnosed diabetes in the two upper quartiles of animal fat intake compared with the two lower quartiles ($p < 0.01$ for both) (Anastasia C. Thanopoulou, et al., 2003). However, a high intake of vegetable fat was inversely associated with the risk of type 2 diabetes in the Iowa Women's Study (Meyer K. Jacobs D Jr. Kushi 1., Folsom A., 2001).

Obesity

Body mass index (BMI) was identified as a predictor of diabetes (Bahman P.

Tabaei; William H. Herman, 2002). The Texas BRFSS 1966-1999 survey data report showed that the percentage of obesity (BMI greater than 30 kg/m^2) population is higher in diabetics (40.2%, with a 95% confidence interval: 34.8-45.6) than in non-diabetics (18.8%, with a 95% confidence interval: 17.3-20.2) (Weihua Li, et al., 2001). There appears to be an association between the amount of caloric intake and body weight with the development of diabetes. The upper body (android) obesity is associated with greater insulin resistance than lower body (gynoid) obesity (Kissebah AH., 1996). The previous study also showed that overweight individuals are more susceptible to developing insulin resistance on high-saturated fat diets (Jennifer C. Lovejoy, et al., 2002). When obesity is compounded by physical inactivity, the risk for Type 2 diabetes dramatically increases (A joint editorial statement. 1999).

Physical Inactivity

A number of previous studies demonstrated that physical inactivity is associated with diabetes (Anastasia C. Thanopoulou, et al., 2003; Karin M. Nelson, et al., 2002). Lack of exercise increases risk of diabetes, even after adjustment for BMI (Hu FB, van Dam RM, Liu S., 2001). The Texas BRFSS 1966-1999 survey data report showed that men with diabetes are more sedentary (9% more) than those without diabetes, although this difference was not statistically significant. Similarly, Texas women with diabetes are more sedentary (13% more) than those without diabetes ($p < 0.05$) (Weihua Li, et al., 2001). A cohort study examined the association between total, moderate, and vigorous exercise and the incidence of Type 2 diabetes; 70,102 females were followed for 8 years. The subjects were divided into five quintiles of physical activity based on metabolic equivalent task hours. This study found that 2.1 to 4.6 metabolic equivalent task hours per week resulted in a 16% reduction in Type 2 diabetes risk, and 21.8 or larger metabolic equivalent task hours per week associated with a 26% reduction in Type 2 diabetes risk. This study also found that equivalent energy expenditures from either walking or vigorous activity were associated with comparable risk reductions in the risk of developing Type 2 diabetes (HU FB Sigal RJ, Rich-Edwards JW, et al., 1999).

Alcohol Consumption

Alcohol consumption represents a potentially important, modifiable risk factor of Type 2 diabetes. In a cohort study conducted among Japanese men with BMIs equal to or greater than 22.1 kg/m^2, moderate alcohol consumption (29.1-50.0 ml/day) was associated with a significantly reduced risk of Type 2 diabetes compared with nondrinkers (adjusted RR = 0.58, 95% confidence interval: 0.39-0.87). However, among lean men with BMIs equal to or less than 22.0 kg/m^2, heavy alcohol consumption ($\geq$ 50.1 ml/day) was strongly associated with an increased risk of Type 2 diabetes compared with nondrinkers (adjusted RR = 2.48, 95% confidence interval: 1.31-4.71) (Tsumura, Kei, Hayashi, et al., 1999). In a population-based cross-sectional study consisting of 3,128 Swedish men, the adjusted odds ratio of diabetes was 2.1 (95% CI: 1.0-4.5) in men with high consumption of alcohol (corresponding to over 12 drinks per week) and 0.7 (95% CI: 0.3-1.8) in moderate consumers (7-12 drinks per week), indicating that high consumption of alcohol increases the occurrence of Type 2 diabetes and regular alcohol consumption was associated with a reduced prevalence, particularly at moderate level (S. Carlsson, et al., 2000).

Tobacco Smoking

Tobacco smoking has been found to be an important predictor of diabetes and a contributor to the incidence of complications associated with diabetes. Avoidance of tobacco smoking is an important component in the management of diabetes (Haire-Joshu D, Glasgow RE. Tibbs TL. 1999). A population-based cross-sectional study of glucose intolerance and tobacco use in Stochholm, Sweden during 1992-1994 showed that the odds ratio (OR) of Type 2 diabetes was increased for persons who smoked more than 25 cigarettes per day (OR=2.6, 95% confidence interval: 1.1-5.9) as well as for moist snuff dippers of greater than three boxes per week (OR=2.7, 95% confidence interval: 1.3-5.5). The results indicate that heavy users of cigarettes or moist snuff have an increased risk of Type 2 diabetes (P. –G Persson, S. Carlsson, et al., 2000).

Hypertension

People with diabetes are as much as three times more likely to have high blood pressure than people without diabetes, and therefore are at a substantially greater risk for heart disease than non-diabetics (Peterson, Kevin, 2003). The Texas BRFSS 1966-1999 survey data report showed that Texans with diabetes were five times more likely to have hypertension than those without diabetes (OR = 5.23). With combination of hypertension and diabetes, the risk of stroke can be 2 to 4 times higher than it is for those with hypertension only (Weihua Li, et al., 2001). In a cohort study conducted between 1981 and 1997 among 7,594 Japanese men, both high normal blood pressure and hypertension were associated with risk of Type 2 diabetes. Compared with normotensive men, men with high normal blood pressure had an adjusted relative risk of 1.39 for diabetes (95% confidence interval: 1.14-1.69), and men with hypertension had an adjusted relative risk of 1.76 for diabetes (95% confidence interval: 1.43-2.16). Even among lean men (BMI less than 22.7 kg/m^2), men with high normal blood pressure had a relative risk of 1.71 for diabetes (95% confidence interval: 1.20-2.42), while men with hypertension had a relative risk of 2.02 for diabetes (95% confidence interval: 1.34-3.04) compared with normotensive men (Tomoshige Hayashi, et al., 1999).

The following figure illustrates the interrelationship of risk factors for developing type 2 diabetes mellitus and the metabolic abnormalities associated with insulin resistance (Barbara Fletcher, Meg Gulanick, Cindy Lamendola. 2002).

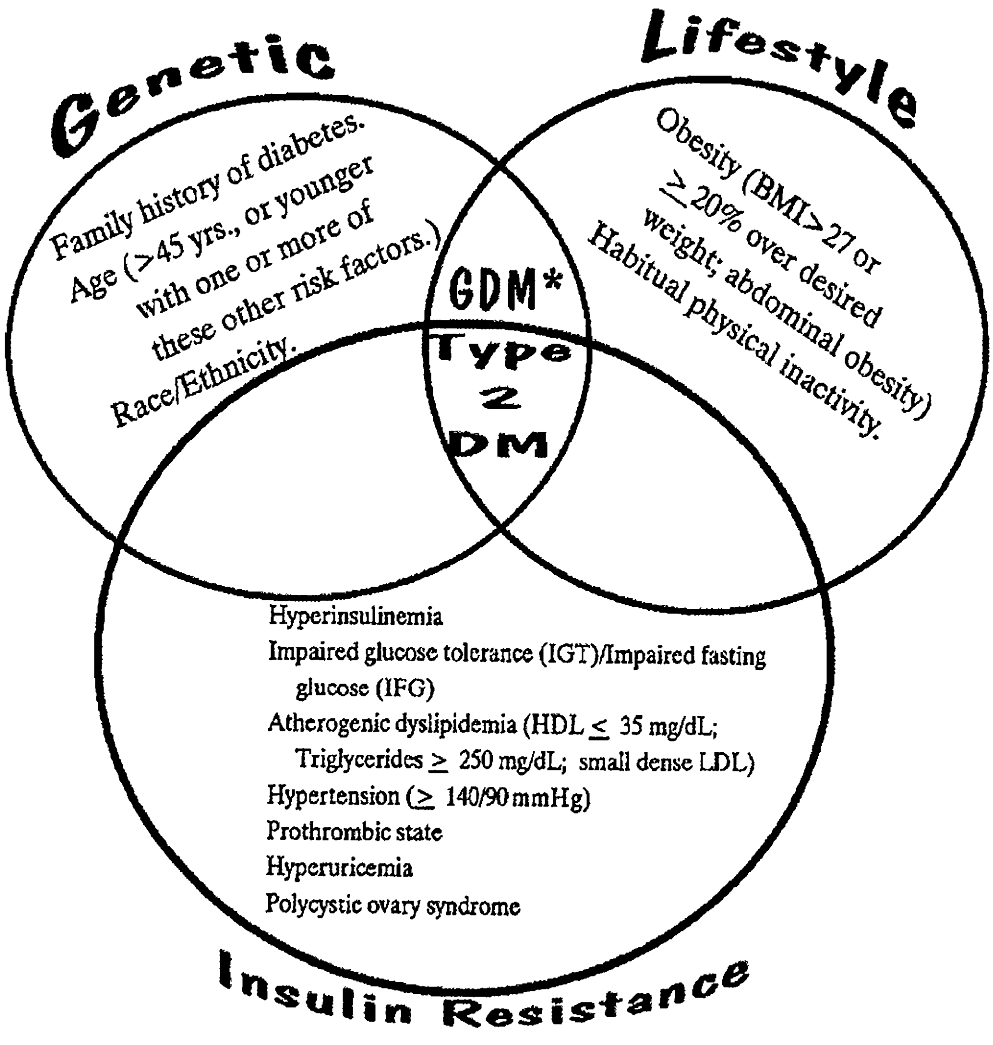
Genetic
Family history of diabetes.
Age (>45 yrs., or younger with one or more of these other risk factors.)
Race/Ethnicity.
Lifestyle
Obesity (BMI>27 or ≥ 20% over desired weight; abdominal obesity)
Habitual physical inactivity.
GDM*
Type 2 DM
Hyperinsulinemia
Impaired glucose tolerance (IGT)/Impaired fasting glucose (IFG)
Atherogenic dyslipidemia (HDL ≤ 35 mg/dL; Triglycerides ≥ 250 mg/dL; small dense LDL)
Hypertension (≥ 140/90 mmHg)
Prothrombic state
Hyperuricemia
Polycystic ovary syndrome
Insulin Resistance

Figure 2.1: Risk factors for developing Type 2 diabetes mellitus and the metabolic abnormalities associated with insulin resistance

Note: * GDM (gestatıonal dıabetes mellıtus) may be ınfluenced by genetıcs, lıfestyle, ınsulın resistance, or a combınatıon thereof. *Source:* Data from Amerıcan Dıabetes Assocıatıon, Clınıcal Practıce Recommendatıons 2001, vol. 24: S21–S24.

Overall, genetic, environmental, and metabolic risk factors are interrelated and contribute to the development of Type 2 diabetes. A strong family history of diabetes, advanced age, obesity, and physical inactivity identify those individuals at highest risk. Minority populations are also at higher risk, not only because of family history and genetics, but also because of adaptation to American environmental influences of poor dietary and exercise habits (Barbara Fletcher, Meg Gulanick, Cindy Lamendola. 2002).

2.2 The Application of Multivariate Logistic Regression to the Diabetes Study

Multivariate logistic regression has been applied to many previous diabetes studies (O. Rolandsson, et al. 2001; Philip S. Mehler, et al., 1998; Bahman P. Tabaei, William H. Herman, 2002; etc.)

To assess the likelihood of previously undiagnosed diabetes, multiple logistic regression analysis was employed to develop a predictive equation by using data from 1,032 Egyptian subjects between 1992 and 1993. Age, sex, BMI (kg/m^2), postprandial time (self-reported number of hours since last food or drink other than water) as 0 to ≥ 8 hours, and random capillary plasma glucose (mg/dl) were included in the equation:

Logit P = -10.0382 + 0.0331 (age in year) + 0.0308 (random plasma glucose in mg/dl) + 0.2500 (postprandial time assessed as 0 to ≥ 8 hours) + 0.5620 (if female) + 0.0346 BMI

This recommended multivariate logistic regression was developed for undiagnosed diabetes screening. It can be easily implemented to predict previously undiagnosed diabetes (Bahman P. Tabaei, William H. Herman, 2002).

Multivariate logistic regression was used in a case-control study in 20 hospitals in seven German cities and counties to quantify the relationship between diabetes and amputations. Male sex, grouped age, diabetes, and interaction terms age by diabetes and sex by diabetes were entered into the logistic regression model as predictor variables to predict the presence of amputation. In the model adjusting for age and sex, diabetes was associated with an odds ratio of 18.2 (CI: 14.2-23.6). This study demonstrated a strong association between the risk of amputation and diabetes. The researchers concluded that the reduction of amputations in the general population would be achieved by improving foot care in people with diabetes (C. Trautner, B. Haastert, G. Giani and M. Berger, 2002).

To assess factors predicting perceptions of diabetes risk, multivariate logistic regression was employed in the screening test data analysis in the areas of the Oxford and Northampton Health Authorities, 1997. Multivariate logistic regression suggested that predictors of diabetes were female sex (*p*=0.003), age 35-54 years rather than 55-74 years (*p*=0.003), and having a parent with diabetes (*p*<0.00001). Body mass index did not affect perception of likelihood (A. J. Farmer, J. C. Levy and R. C. Turner, 1999).

In a population-based cross-sectional study Stockholm, Sweden, researchers calculated the association between tobacco use and glucose intolerance and estimated odds ratios with the use of multiple logistic regression analysis (P. –G Persson, S. Carlsson, et al., 2000). Age, BMI, family history of diabetes, physical activity and alcohol consumption also were entered into a logistic regression model as independent variables. The result showed that the odds ratio of Type 2 diabetes was increased for smokers of 25 or more cigarettes per day (odds ratio=2.6, 95% confidence interval: 1.1-

5.9) as well as for moist snuff dippers of 3 or more boxes per week (OR= 2.7, 95% confidence interval: 1.3-5.5). The odds ratio of relatively high fasting insulin levels in subjects with impaired glucose tolerance associated with cigarette smoking of 25 or more cigarettes per day was 1.5 (95% confidence interval: 0.7-3.6). The corresponding estimated odds ratio of a relatively low 2-hour insulin response was 2.5 (95% confidence interval: 0.9-7.1). The results indicate that heavy users of cigarettes or moist snuff have an increased risk of Type 2 diabetes and also suggest that tobacco use is associated with a low insulin response.

CHAPTER 3

METHODS

3.1 Research Design

This proposed thesis research used a population-based cross-sectional design. Prevalence of diabetes and potential risk factors were measured at the same time. The dependent variable was defined by the question "Have you ever been told by a doctor that you have diabetes?" Diabetes status was dichotomous answers: "Yes" or "No" (excluding answer of 'yes, but only during pregnancy'). The possible predictors were demographic, medical conditions (high blood pressure and high blood cholesterol) and health behavior characteristics. The final logistic regression procedure "stepwise selection" selected significant predictors and computed the probability that an individual is a diabetic.

3.2 Data Source and Data Collection

This study used existing data made available by the Texas Department of Health (TDH). Data was collected by Texas Behavioral Risk Factor Surveillance System (BRFSS) from a 2001 survey. BRFSS consists of a monthly survey of 500 randomly selected Texans who are 18 years or older. Texas BRFSS used Centers for Disease Control and Prevention "2001 Behavioral Risk Factor Surveillance System Questionnaire" and disproportionate stratified random sampling (DSS), which is a special type of probability cluster sampling. For sampling design, information obtained from a previous survey was used to classify 100-number blocks of telephone numbers into strata that are either likely or unlikely to yield residential numbers (Centers for Disease Control and Prevention: BRFSS User's Guide, 1998, Survey samples and sampling methods. pp 3-2; Data Management, 00.8-2.). When the 2001 survey was conducted, participants were randomly divided into two groups: 3031 participants completed Survey A and 2885 participants completed Survey B of the Texas Behavioral Risk Factor Surveillance System. Some of survey questionnaire sections were only conducted in Survey A or Survey B, and some were in both.

3.3 Variable Selection

There are mixed variables including original questions and BRFSS summated variables in this data set. Ninety-four original variables were collected from the following CDC 2001 survey questionnaire sections:

Section 1: Health Status

Section 2: Health Care Access

Section 3: Exercise

Section 4: Hypertension Awareness

Section 5: Cholesterol Awareness

Section 7: Diabetes (Survey A and Survey B)

Section 9: Immunization

Section 10: Tobacco Use

Section 11: Alcohol Consumption

Section 13: Demographics

Section 15: Physical Activity

Optional modules 1: Diabetes

Optional module 8: Heart Attack and Stroke (Survey B only)

Optional module 9: Cardiovascular disease (Survey B only)

Optional module 10: Fruits and Vegetables (Survey A only)

Optional module 11: Weight Control (Survey A only)

Optional module 14: Other Tobacco Products (Survey A and Survey B)

Texas BRFSS created 51 summated variables based on the answers of the questionnaire in the raw data set. Some of them are duplicated, such as three different ways to group age, three different household income groups, and two different education groups. There are a total of 145 variables and 5916 observations in this data set.

For the purpose of a reduction in the number of variables, some BRFSS summated health behavior variables were used instead of original health behavior variables. The BRFSS substitute variables provided the same information as original questions did, because these summated variables were calculated based on the variables of the raw data set. Other BRFSS summated variables were also used because they provided information that could not be obtained from the original questionnaire, such as person's body mass index. Six BRFSS summated behavior variables; two original medical condition variables and three original behavior variables were used. Among the five demographic variables, the only BRFSS summated variable will be age groups for sake of convenience of chi-square test; other demographic variables will be the original ones.

According to risk factors of diabetes identified in previous studies, the following final 17 variables were selected. They are all categorical variables:

<u>Dependent variable</u>:

1. Diabetes Status (Have you ever be told by a doctor that you have a diabetes?)

   code: 1 = “Yes”; 2 = “Yes, but only during pregnancy”; 3 = “No”; 7 = “DK/NS”; 9 = “Refused”

(Excluding participants who have missing values on this question and who have answers of ‘Yes, but only during pregnancy’)

<u>Demographic indicators:</u>

2. Age

   code: 1 = “18-34”; 2 = “35-44”; 3 = “45-54”; 4 = “55-64”; 5 = “64+”

3. Race/Ethnicity

   code: 1 = ‘Non-Hispanic White”; 2 = “Non-Hispanic Black”; 3 = “Hispanic”; 4 = “Others”; 7 = “DK/NS”; 9 = “Refused”

4. Education

   code: 1 = “Elementary or kindergarten”; 2 = “Some high school”; 3 = “High school graduate”; 4 = “Some college or technical school”; 5 = “College graduate”; 9 = “Refused”

5. Household Income

   code: 1 = “less than $10,000”; 2 = “$10,000-$14,999”; 3 = “$15,000-19,999”; 4 = “$20,000-$24,999”; 5 = “$25,000-34,999”; 6 = “$35,000-$49,999”; 7 = “$50,000-$74,999”; 77 = “DK/NS”; 8 = “$75,000”; 99 = “Refused”

6. Sex

   code: 1 = “Male”; 2 = “Female”; 7 = “DK/NS”; 9 = “Refused”

<u>Possible medical condition and health behavior indicators:</u>

7. High Blood Pressure (Have you ever been told by a doctor, nurse, or other health professional that you have high blood pressure?)

   code: 1 = “Yes”; 2 = “No”; 7 = “DK/NS”; 9 = “Refused”

8. High Blood Cholesterol (Have you ever been told by a doctor, nurse or other health professional that your blood cholesterol is high?)

   code: 1 = “Yes”; 2 = “No”; 7 = “DK/NS”; 9 = “Refused”

9. Smoking Risk

   code: 1 = “Not at risk”; 2 = “At risk”; 7 = “DK/NS”; 9 = “Refused”

   (At risk was defined as smoked at least 100 cigarettes in entire life and smoked during some or all of the past 30 days.)

10. Alcohol Drink Status

    code: 1 = “Yes”; 2 = “No”; 7 = “DK/NS”; 9 = “Refused”

11. Acute Alcohol Risk

    code: 1 = “Not at risk”; 2 = “At risk”; 7 = “DK/NS”; 9 = “Refused”

    (At risk was defied as having 5 or more alcohol drinks on at least one occasion in the past 30 days.)

12. Moderate Activity (Do you do moderate activities for at least 10 minutes at a time?)

    code: 1 = “Yes”; 2 = “No”; 7 = “DK/NS”; 9 = “Refused”

13. Vigorous Activity (Do you do vigorous activities for at least 10 minutes at a time?)

    code: 1 = “Yes”; 2 = “No”; 7 = “DK/NS”; 9 = “Refused”

14. Met Recommendations for Physical Activity

    code: 1 = “Yes”; 2 = “No”; 9 = “DK/NS or Refused”

(CDC recommendations for physical activity: Adults should engage in moderate-intensity physical activities for at least 30 minutes on 5 or more days of the week or in vigorous-intensity physical activity 3 or more days per week for 20 or more minutes per occasion.)

15. Leisure Time Physical Activity

    code: 1 = “Leisure time activity in the past month”;

    2 = “No leisure time activity in the past month”; 9 = “DK/NS or Refused”

16. Low Intake of Fruits and Vegetables?

    code: 1 = “Not at risk”; 2 = “At risk”; 7 = “DK/NS”; 9 = “Refused”

(At risk was defined as intake fruit or vegetable less than 5 serving per day. A serving was defined as ‘about a handful of fruit or vegetables, or half cup’, which is equivalent to about 150 g of fruit or 75 g of vegetables.)

17. Overweight / Obese

    code: 1 = “Neither overweight nor obese”; 2 = “Overweight”; 3 = “Obese”;

    7 = “DK/NS”; 9 = “Refused”

(Obese person was defined as body mass index equal to or greater than 30 kg / m^2; overweight person was defined as body mass index 25-29.9 kg/m^2. Body mass index was calculated as weight in kilograms divided by height in meters squared.)

Because summated variables ‘Low Intake of Fruits and Vegetables’ and ‘Overweight/Obese’ were created based on survey questionnaire optional module 10 and 11 (both conducted in Survey A only), literally, only Survey A participants had answers for these questions. Of 5916 survey participants, 1221 observations that had no missing values on these 17 selected variables were selected and used in sample data analyses.

3.4 Statistical Analyses

Statistical analyses were performed by using SAS statistical software (version 8.1; SAS Institute Inc. Cary NC.). The following statistical procedures were used:

1. Prevalence of diabetes, Crude odds ratios and 95% confidence intervals were calculated for each demographic, medical condition and health behavior variable.
2. Pearson *Chi*-square test for independence of diabetes presence and each of the demographic, medical condition and health behavior variables was performed by analyzing two-way crosstabulations.
3. Cochran Armitage *Chi*-Square Tests for Trend was used to detect that prevalence of diabetes was significantly different by stratum of age, education, household income and overweight/obese variables.
4. Logistic regression model was performed to determine demographic characteristics, medical conditions and health behaviors most significantly related to predicting diabetes. The dependent variable was diabetes status, which is the dichotomous answer of the question “Have you ever been told by a doctor that you have diabetes?” Independent variables were demographic, medical condition

and health behavior variables. Stepwise selection method was used to select significant predictors of diabetes. Probability value for diabetes presence given the independent variables was determined from the multivariate logistic regression model. Multiple adjusted odds ratios were calculated for all significant variables in the logistic regression model. An α level of 0.05 was used for all statistical tests. The –2 log-likelihood ratio test was used to test the overall significance of the predictive equation. The significance of the variables in the model was assessed by the Wald *Chi*-Square test and 95% confidence intervals. The fit of the model was assessed by the Hosmer-Lemeshow goodness of fit *Chi*-Square test and Nagelkerke R^2. Concordance (the closer to 100 the better) and discordance values, derived from the logistic regression analysis, were used to measure the association of predicted probabilities and to check the ability of the model to predict diabetes presence. The higher the value of the concordance and the lower the value of discordance, the greater the ability of the model to predict diabetes presence. Somers'D, Gamma, and c (when the response is dichotomous, these values should be above 0.5. The closer to 1 the better) were also used to assess the model fit. The number (and percent) of correctly and incorrectly classified responses for different cutpoints, sensitivity, specificity, false positives, and false negatives were generated from logistic regression analysis. To evaluate the overall predictive performance of the logistic equation, discrimination was considered. Discrimination was defined as the ability of the equation to distinguish high-risk subjects from low-risk subjects and was quantified by the area under the receiver-operating characteristic (ROC) curve. To select the optimal cut point to define a high-risk individual, ROC curve was constructed by plotting sensitivity against the false-positive rate (1-specificity) over a range of cut-point values. The closer an ROC curve is to the upper left corner of the graph (as true-positive rate approaches 1 and false-positive rate approaches 0), the larger the area under the curve, and more accurate the prediction model. Each point on the curve represents a cutoff probability. A lower cutoff typically gives more false positives. A high cutoff gives more false negatives, a low sensitivity, and a high specificity. Generally, the best cut point is at or near the shoulder of the ROC curve. Individuals who had a predicted probability generated by applying the estimated coefficients in the final equation higher than this optimal cutpoint were classified as diabetics by the model even they reported themselves as non-diabetics. These individuals could be identified as a group of people at high risk and need further diagnostic investigation. To assess outliers and detect extreme points in the design space, logistic regression diagnostics were performed by plotting the diagnostic statistic against the observation number using hat matrix diagonal and Pearson and Deviance residuals analyses. To validate the final logistic regression equation, the model was applied to TDH BRFSS survey data of year 1999 that had not been used to generate the equation. (Lora D. Delwiche and Susan J. Slaughter,1998; Lloyd D. Fisher, Gerald Van Belle,1993; B. S. Everitt and G. Der, 1997; SAS Institute Inc. 1987; Stanto a Glantz , Bryan K. Slinker, 2001; Hair Anderson, Tatham Black, 1998; Steuerberg EW, et al., 2001).

CHAPTER 4

RESULTS

4.1 Descriptive Analyses of the Sample Population

4.1.1 Demographics

The Texas BRFSS 2001 data set consisted of 5916 participants. The participants ranged in age from 18 to 99 years with a mean age of 46 years. The sampled population consisted of 41.2% male, 58.8% female, 61.68% Non-Hispanic white, 9.69% Non-Hispanic black, 24.17% Hispanic and 3.79% others. Of these 5916 participants, 1221 had

Table 4.1: Sample Demographics

Variable	Number	%	Variable	Number	%
Age			**Education**		
18-34	245	20.07	College graduate	459	37.59
35-44	315	25.80	Some college or technical school	340	27.85
45-54	266	21.79	High school graduate	302	24.73
55-64	170	13.92	Some high school	85	6.96
65+	225	18.43	Elementary or kindergarten	35	2.87
Race/Ethnicity			**Household Income**		
Non-Hispanic white	851	69.70	$75,000+	305	24.98
Non-Hispanic black	108	8.85	$50,000-$74,999	215	17.61
Hispanic	208	17.04	$35,000-$49,999	214	17.53
Others	54	4.42	$25,000-$34,999	166	13.6
Sex			$20,000-$24,999	110	9.01
Female	678	55.53	$15,000-$19,999	100	8.19
Male	543	44.47	$10,000-$14,999	62	5.08
			Less than $10,000	49	4.01

no missing values in 17 selected variables that were selected and used in the analyses. The demographic description of these 1221 subjects is displayed in Table 4.1.

4.1.2 Medical Condition and Health Behavior Variables

Table 4.2 contains the medical condition and health behavior variable proportions. Of these 1221 participants, about 68% had normal blood pressure and normal blood cholesterol. Over 16% were current or past smokers, 13% were at acute alcohol risk, and more than 70% had a low intake of fruits and vegetables. More than 52% drank alcohol, and 26.62% were obese. About 82%, 42%, 80% did moderate activity, vigorous activity

Table 4.2: Medical Condition and Health Behavioral Proportions

Variable	Number	%	Variable	Number	%
High blood pressure			**Vigorous activity**		
No	836	68.47	No	704	57.66
Yes	385	31.53	Yes	517	42.34
High blood cholesterol			**Met recommendations for physical activity**		
No	838	68.63	No	1079	88.37
Yes	383	31.37	Yes	142	11.63
Smoking risk			**Leisure time physical activity**		
Not at risk	1017	83.29	No	249	20.39
At risk	204	16.71	Yes	972	79.61
Alcohol drink status			**Low intake of Fruits and vegetable**		
No	581	47.58	Not at risk	365	29.89
Yes	640	52.42	At risk	856	70.11
Acute alcohol risk			**Overweigh/obese**		
Not at risk	1062	86.98	Neither overweight nor obese	433	35.46
At risk	159	13.02	Overweight	463	37.92
Moderate activity			Obese	325	26.62
No	217	17.77			
Yes	1004	82.23			

and leisure time physical activity, respectively. But, only 11.63% of participants met CDC recommendations for physical activity.

4.1.3 Diabetes Status

Table 4.3 contains the sample proportion of the diabetes status. The sample excluded participants who had diabetes only during their pregnancies. More than 92% of participants reported themselves as non-diabetics and 7.45% had been told by doctors that they were diabetics.

Table 4.3: Diabetes Status Proportion

Diabetes Status	Number	%
Diabetic	91	7.45
Non-diabetic	1130	92.55
Total	1221	100.00

4.2 Analysis of Diabetes Prevalence

4.2.1 Diabetes Prevalence Among Sample Demographics

Table 4.4 contains self-reported stratified prevalence of diabetes for each demographic variable. Among the five age groups, the age group 18-34 years had the lowest diabetes prevalence, and the highest diabetes prevalence was found among persons aged 55-64 years. Diabetes prevalence increased as participants' age increased. In race groups, non-Hispanic blacks had an approximately doubled prevalence of diabetes compared to non-Hispanic whites. Diabetes prevalence was higher in Hispanics than in non-Hispanic whites. None of 54 'Others' race group people had been told by a doctor that they had diabetes at the time 2001 survey was conducted. As education level increased, diabetes prevalence decreased. The highest education level (college graduated) group had the lowest diabetes prevalence, and the highest diabetes prevalence of was found in the lowest education level (elementary or kindergarten). A similar pattern was seen in household income categories. As income increased, diabetes prevalence

Table 4.4: Diabetes Prevalence Among Sample Demographics

Variable	Number of Diabetes	Total Population	Prevalence of Diabetes per 1000
Age			
18-34	4	245	16.33
35-44	13	315	41.27
45-54	16	266	60.15
55-64	27	170	158.82
65+	31	225	137.78
Race/Ethnicity			
Non-Hispanic white	52	851	61.10
Non-Hispanic black	16	108	148.15
Hispanic	23	208	110.58
Others	0	54	0.00
Education			
College graduate	19	459	41.39
Some college or technical school	25	340	73.53
High school graduate	32	302	105.96
Some high school	10	85	117.65
Elementary or kindergarten	5	35	142.86
Household Income			
$75,000+	8	305	26.23
$50,000-$74,999	9	215	41.86
$35,000-$49,999	13	214	60.75
$25,000-$34,999	16	166	96.39
$20,000-$24,999	11	110	100.00
$15,000-$19,999	14	100	140.00
$10,000-$14,999	12	62	193.55
Less than $10,000	8	49	163.27
Sex			
Female	46	678	67.85
Male	45	543	82.87

decreased. The lowest diabetes prevalence was found in the highest household income ($75,000+), and the second lowest level of income ($10,000-$14,999) had the highest diabetes prevalence. Diabetes prevalence was higher in males than in females.

The self-reported diabetes prevalence among sample demographics are displayed in the following Figure 4.1 – Figure 4.5.

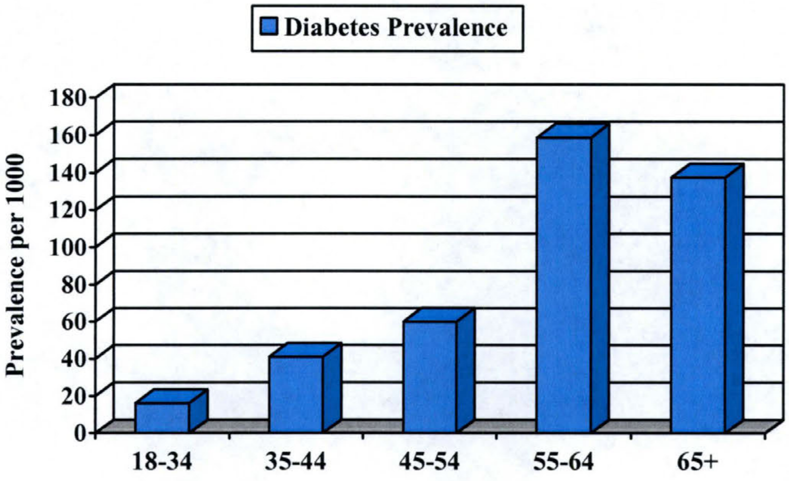

Figure 4.1: Prevalence of Diabetes in Age

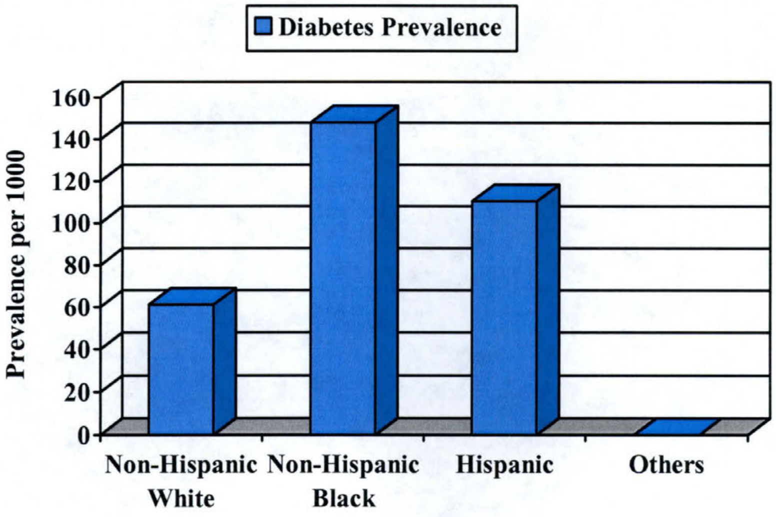

Figure 4.2: Prevalence of Diabetes in Race

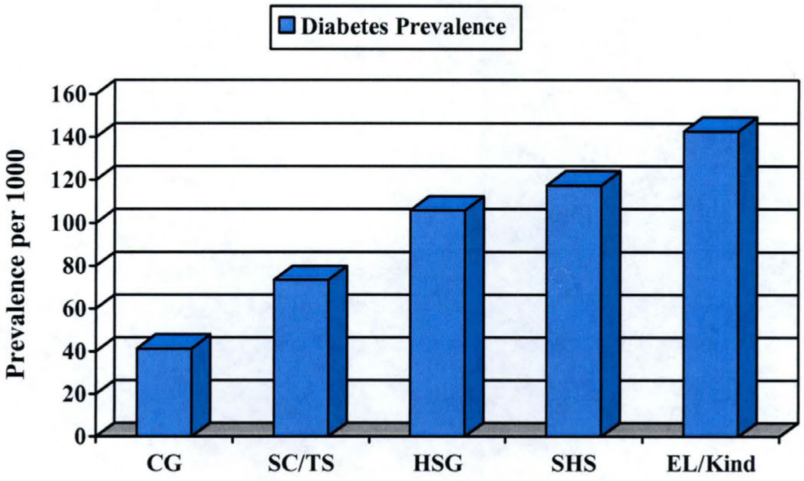

Figure 4.3: Prevalence of Diabetes in Education

Note: CG = College Graduate; SC/TS = Some College/Technical School; HSG = High School Graduate; SHS = Some High School; EL/Kind = Elementary/Kindergarten.

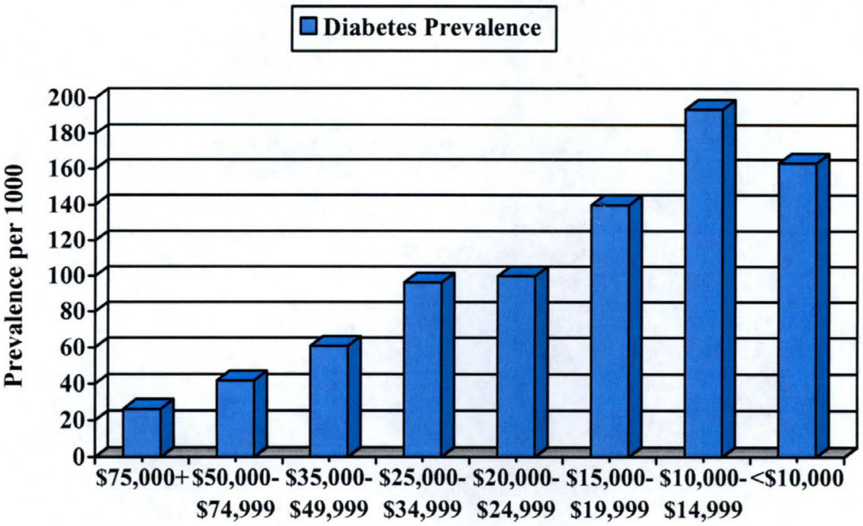

Figure 4.4: Prevalence of Diabetes in Household Income

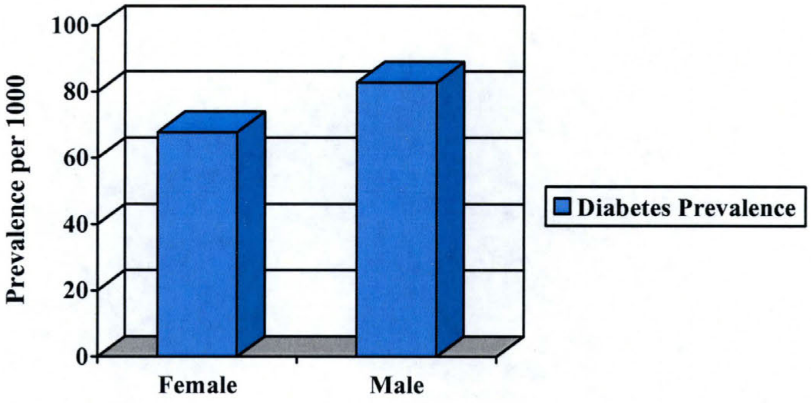

Figure 4.5: Prevalence of Diabetes in Sex

4.2.2 Diabetes Prevalence Among Medical Conditions and Health Behaviors

Table 4.5 displays diabetes prevalence in each type of medical conditions and health behaviors. Among persons with medical conditions, diabetes prevalence was found higher in persons with high blood pressure and high blood cholesterol compared to persons with normal blood pressure and normal blood cholesterol. For health behaviors, there is not much difference of diabetes prevalence between group of participants who were current or past smokers and these participants who were not current or past smokers. The non-drinkers of alcohol had about three times diabetes prevalence compared to alcohol drinkers. People who were 'at acute alcohol risk' had lower diabetes prevalence than people who were 'not at acute alcohol risk'. Reduced diabetes prevalence was seen in both groups of people who did moderate physical activity and people who did vigorous physical activity compared to group of people who did not engaged in moderate physical activity or vigorous physical activity. People who did not report leisure time physical activity in the past month had 2.5 times the diabetes prevalence compared to people who did report leisure time physical activity in the past month. Diabetes prevalence was also higher in people who did not meet recommendations for physical activity than in people who met recommendations for physical activity. People who reported a low intake of fruits and vegetables had lower prevalence of diabetes than people who reported a higher intake of these foods. Increased diabetes prevalence was found among persons in the 'overweight' and 'obese' categories as compared to 'neither overweight nor obese' category. As body mass index increased, diabetes prevalence increased. Table 4.2_2 summarizes diabetes prevalence in each category among medical conditions and health behaviors. The following Figure 4.6 – Figure 4.16 display the diabetes prevalence in categories among medical conditions and health behaviors.

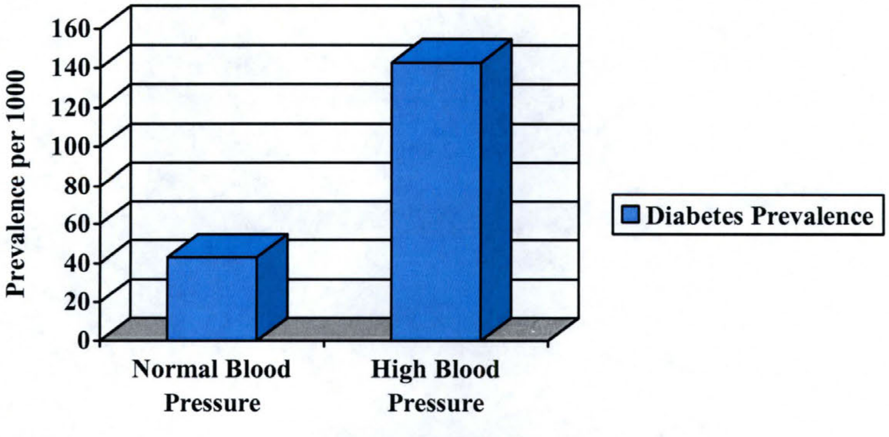

Figure 4.6: Prevalence of Diabetes by Blood Pressure Status

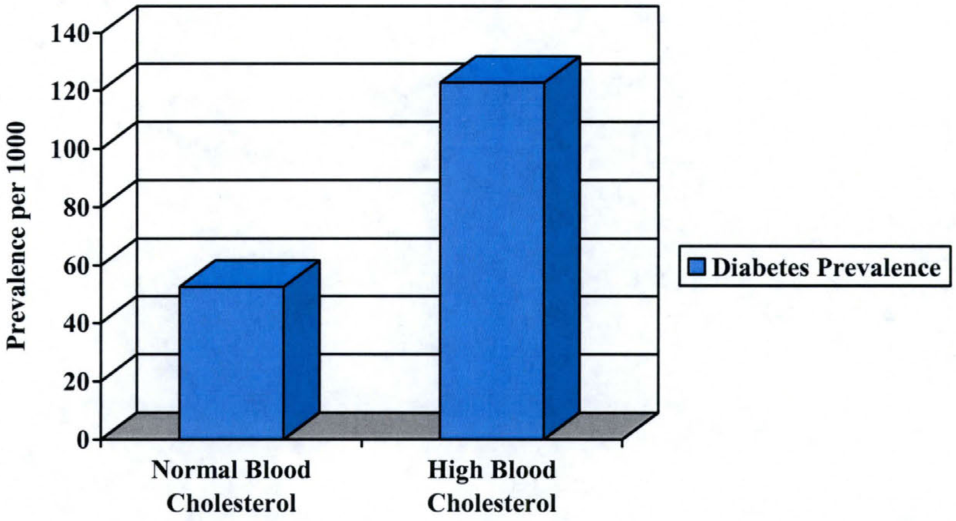

Figure 4.7: Prevalence of Diabetes by Blood Cholesterol Status

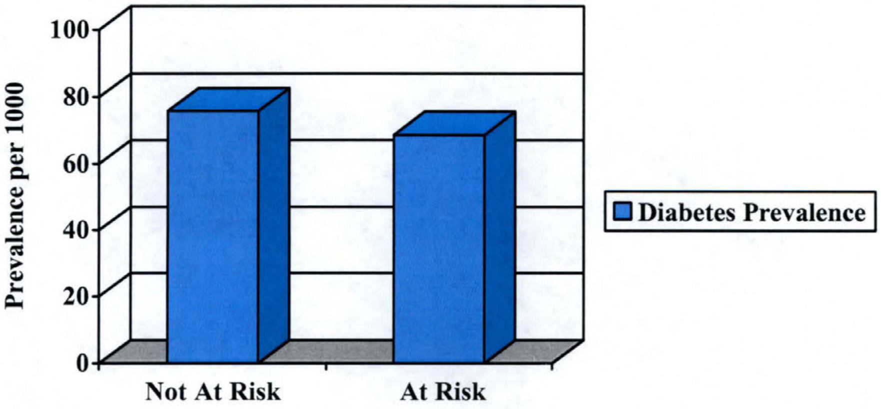

Figure 4.8: Prevalence of Diabetes in Smoking Risk

Note: 'At Risk' was defined as smoked at least 100 cigarettes in entire life and smoked during some or all of the past 30 days.

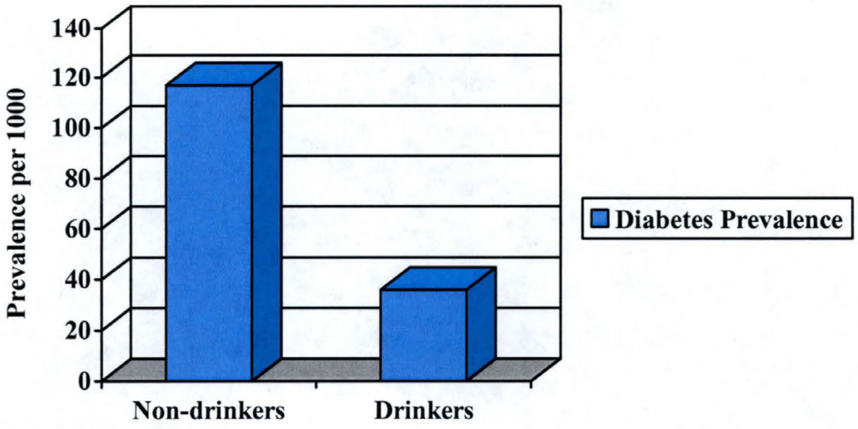

Figure 4.9: Prevalence of Diabetes by Alcohol Consumption

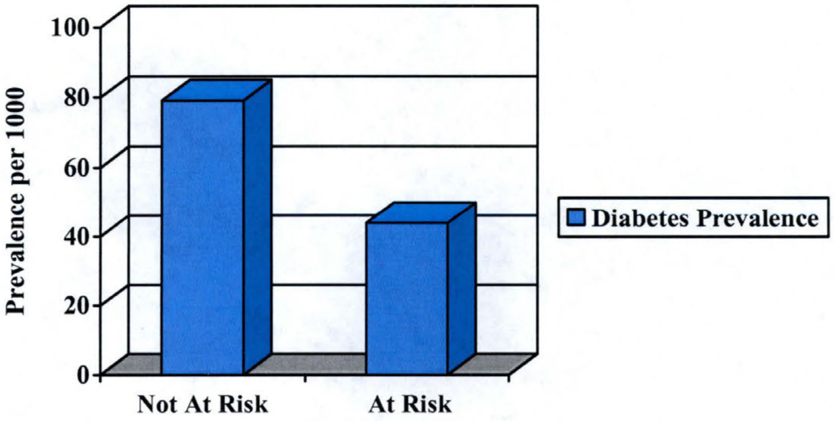

Figure 4.10: Prevalence of Diabetes in Persons with Acute Alcohol Risk

Note: 'At Risk' was defined as having 5 or more alcohol drinks on at least one occasion in the past 30 days.

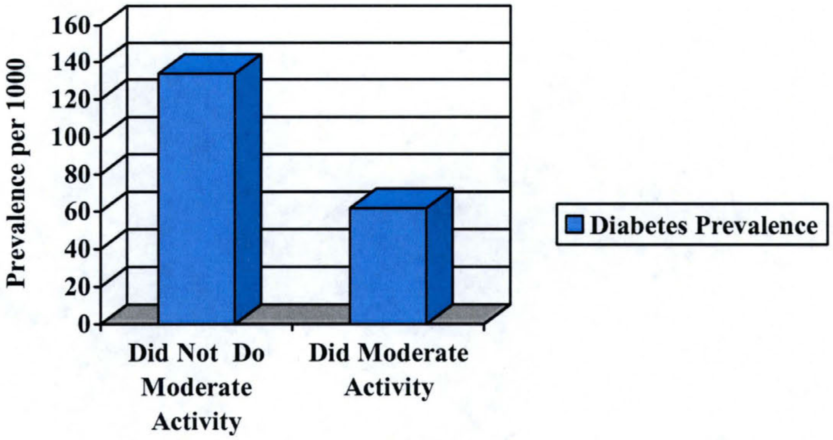

Figure 4.11: Prevalence of Diabetes in Moderate Activity

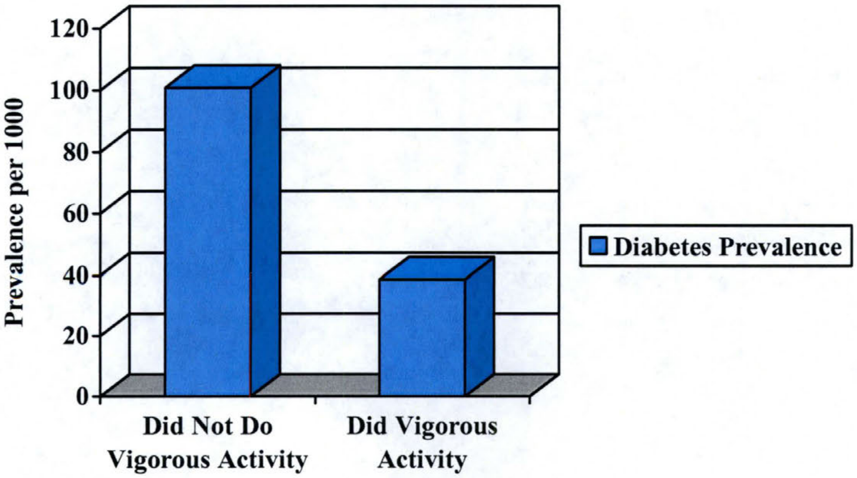

Figure 4.12: Prevalence of Diabetes in Vigorous Activity

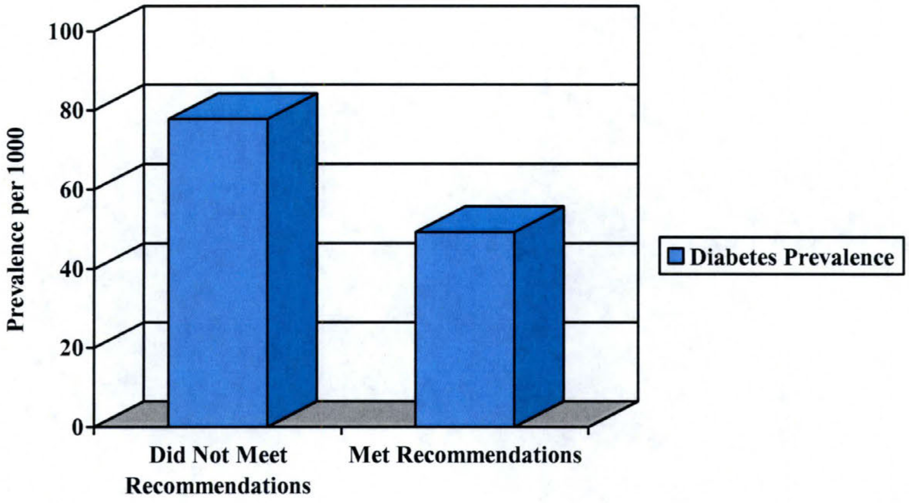

Figure 4.13: Prevalence of Diabetes

by Whether Met Recommendations for Physical Activity

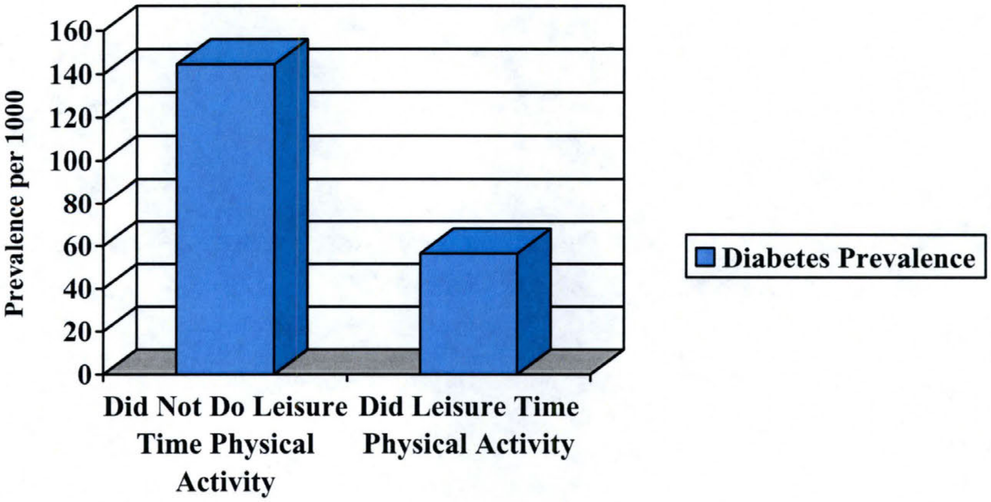

Figure 4.14: Prevalence of Diabetes

by Reported Leisure Time Physical Activity

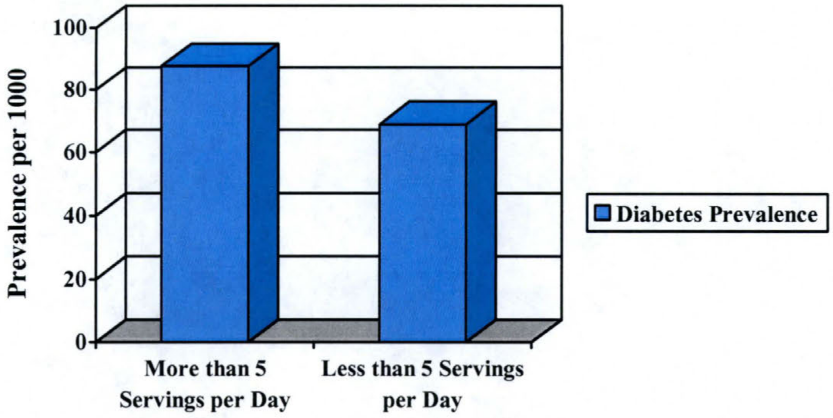

Figure 4.15: Prevalence of Diabetes by Fruit and Vegetable Intake

Note: A serving was defined as 'about a handful of fruit or vegetables, or half cup', which is equivalent to about 150 g of fruit or 75 g of vegetables.

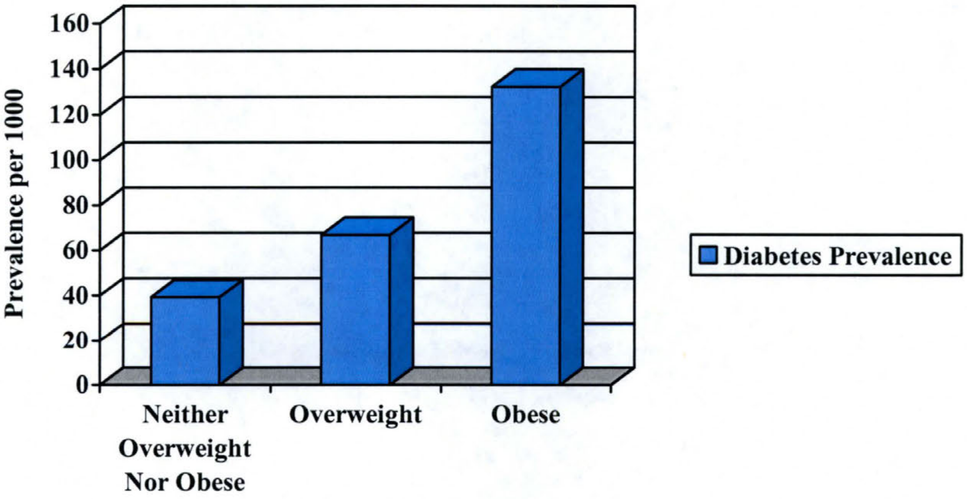

Figure 4.16: Prevalence of Diabetes by Body Mass Index

Note: 'Obese' person was defined as BMI $\geq$ 30 kg / m^2; 'Overweight' person was defined as BMI = 25 — 29.9 kg / m^2. BMI was calculated as weight in kilograms divided by height in meters squared.)

Table 4.5: Diabetes Prevalence Among Medical Conditions and Health Behaviors

Variable	Number of Diabetics	Total Population	Prevalence of Diabetes per 1000
High blood pressure			
No	36	836	43.06
Yes	55	385	142.86
High blood cholesterol			
No	44	838	52.51
Yes	47	383	122.72
Smoking risk			
Not at risk	77	1017	75.71
At risk	14	204	68.63
Alcohol drink status			
No	68	581	117.04
Yes	23	640	35.94
Acute alcohol risk			
Not at risk	84	1062	79.10
At risk	7	159	44.03
Moderate activity			
No	29	217	133.64
Yes	62	1004	61.75
Vigorous activity			
No	71	704	100.85
Yes	20	517	38.68
Met recommendations for physical activity			
No	84	1079	77.85
Yes	7	142	49.30
Leisure time physical activity			
No	36	249	144.58
Yes	55	972	56.58
Low Intake of fruits and vegetables			
Not at risk	32	365	87.67
At risk	59	856	68.93
Overweigh/Obese			
Neither overweight nor obese	17	433	39.26
Overweight	31	463	66.95
Obese	43	325	132.31

4.3 Analysis of Associations Between Diabetes and Potential Risk Factors

4.3.1 Crude Odds Ratios (OR) and 95% Confidence Interval (95% CI) for Diabetes Associated with Sample Demographics

Table 4.6 contains crude odds ratios and 95% confidence interval for diabetes associated with sample demographics. As age increased, the crude odds ratio increased indicating that an increased age is positively associated with an increased prevalence of diabetes. On the contrary, as education level and household income decreased, the crude odds ratio increased indicating that education level or household income was negatively associated with prevalence of diabetes. The odds ratios for non-Hispanic blacks and Hispanics, compared to non-Hispanic whites, indicated that being non-Hispanic black or Hispanic was associated with an increased prevalence of diabetes. The odds ratio could not be calculated for 'Others' race group of people, as there were no case of diabetes among the 'Others'. The results also showed the non-significance of the association between prevalence of diabetes and sex.

4.3.2 Crude Odds Ratios and 95% Confidence Interval for Diabetes Associated with Medical Conditions and Health Behaviors

Table 4.7 contains crude odds ratios and 95% confidence interval for diabetes associated with various medical conditions and health behaviors. Results indicated that high blood pressure, high blood cholesterol and obesity were related to an increased prevalence of diabetes. Drinking alcohol; doing moderate activity; doing vigorous activity and doing physical activity at leisure time in the past month were associated with a decreased prevalence of diabetes. However, no significant association was found between prevalence of diabetes and smoking risk, acute alcohol risk, low intake of fruits and vegetables, meeting recommendations for physical activity, or being overweight.

Table 4.6: Crude ORs and 95% CI for Diabetes Associated with Sample Demographics

Variable	Odds Ratio	95% Confidence Interval	
Age			
18-34	1.00		
35-44	2.59	0.84	8.06
45-54	3.86	1.27	11.70
55-64	11.38	3.90	33.17
65+	9.63	3.34	27.74
Race/Ethnicity			
Non-Hispanic white	1.00		
Non-Hispanic black	2.67	1.47	4.87
Hispanic	1.91	1.14	3.20
Others			
Education			
College graduate	1.00		
Some college or technical school	1.84	0.99	3.40
High school graduate	2.74	1.53	4.94
Some high school	3.09	1.38	6.90
Elementary or kindergarten	3.86	1.35	11.05
Household Income			
$75,000+	1.00		
$50,000-$74,999	1.62	0.62	4.27
$35,000-$49,999	2.40	0.98	5.90
$25,000-$34,999	3.96	1.66	9.46
$20,000-$24,999	4.13	1.61	10.55
$15,000-$19,999	6.04	2.45	14.88
$10,000-$14,999	8.91	3.47	22.89
Less than $10,000	7.24	2.58	20.35
Sex			
Female	1.00		
Male	1.24	0.81	1.90

Table 4.7: Crude ORs and 95% CI for Diabetes

Associated with Various Medical Conditions and Health Behaviors

Variable	Odds Ratio	95% Confidence Interval	
High blood pressure			
No	1.00		
Yes	3.70	2.39	5.75
High blood cholesterol			
No	1.00		
Yes	2.52	1.64	3.88
Smoking risk			
Not at risk	1.00		
At risk	0.90	0.50	1.62
Alcohol drink status			
No	1.00		
Yes	0.28	0.17	0.46
Acute alcohol risk			
Not at risk	1.00		
At risk	0.54	0.24	1.18
Moderate activity			
No	1.00		
Yes	0.43	0.27	0.68
Vigorous activity			
No	1.00		
Yes	0.36	0.22	0.60
Met recommendations for physical activity			
No	1.00		
Yes	0.61	0.28	1.36
Leisure time physical activity			
No	1.00		
Yes	0.35	0.23	0.55
Fruit and vegetable intake risk			
Not at risk	1.00		
At risk	0.77	0.49	1.21
Overweigh/obese			
Neither overweight nor obese	1.00		
Overweight	1.76	0.96	3.22
Obese	3.73	2.09	6.67

4.3.3 Pearson *Chi*-Square Tests for Independence

Table 4.8 contains results of Pearson *Chi*-Square tests for independence. The results showed prevalence of diabetes was significantly (at α = 0.05 level) associated with the following variables: age, race, education, household income, high blood pressure, high blood cholesterol, alcohol drink status, moderate activity, vigorous activity, leisure time physical activity, and overweight/obese. The following variables: sex, smoking risk, acute alcohol risk, meeting recommendations for physical activity, and low fruit and vegetable intake were found to be insignificant.

Table 4.8: Pearson *Chi*-Square Tests for Independence

Variable	*Chi*-Square Statistic	DF	*P* Value
Age	48.44	4	<0.0001
Race/Ethnicity	18.98	3	0.0003
Education	16.30	4	0.0026
Household Income	40.96	7	<0.0001
Sex	0.99	1	0.3205
High blood pressure	38.06	1	<0.0001
High blood cholesterol	18.79	1	<0.0001
Smoking risk	0.12	1	0.7251
Alcohol drink status	29.04	1	<0.0001
Acute alcohol risk	2.47	1	0.1163
Moderate activity	13.37	1	0.0003
Vigorous activity	16.70	1	<0.0001
Met recommendations for physical activity	1.48	1	0.2233
Leisure time physical activity	22.25	1	<0.0001
Low Intake of Fruits and vegetables	1.30	1	0.2535
Overweigh/Obese	23.92	2	<0.0001

4.3.4 Cochran Armitage *Chi*-Square Tests for Trend

Cochran Armitage test is a method of directing *Chi*-Square Tests to detect trends that should otherwise not be noticed. The SAS system provides **Proc Freq**, specifying 'trend' within the **Tables** statement generated the statistic, based on one degree of freedom. Table 4.9 contains the results of Cochran Armitage test for trend for age, education, household income and overweight/obese variables that had more than two strata. The probability of being a diabetic significantly increased ($P < 0.0001$) as age or body mass index increased. The probability of being a diabetic significantly decreased ($P < 0.0001$) as education or income increased.

Table 4.9: Cochran Armitage Tests for Trend

Variable	Statistic [Z]	One-sided Pr < Z	One-sided Pr > Z	Two-Sided Pr > |Z|
Age	−6.4420	<0.0001		<0.0001
Education	3.9832		<0.0001	<0.0001
Income	6.2451		<0.0001	<0.0001
Overweight/Obese Risk	−4.7389	<0.0001		<0.0001

4.4 Logistic Regression to Determine Risk Factors Best Predicting Presence of Diabetes

4.4.1 Initial Logistic Regression Model

First, all demographic variables, medical conditions and health behaviors were considered in the building logistic regression model. The SAS stepwise selection option within the **Proc logistic** statement was used to determine the variables that best predict presence of diabetes. The 'Lackfit' option was used to do 'Hosmer and Lemeshow Test' for measurement of overall goodness-of-fit statistical test. The 'CTABLE' option also was used to generate the number (and percent) of correctly and incorrectly classified responses for different cutpoints. By default, SAS 'Proc logistic' predicts the probability of the smallest ordered value, which means it would be modeling the probability that 'diabetes status = 1', namely probability of participant being a diabetic.

The stepwise procedure selected the following variables as significant predictors of presence of diabetes:

1. HBP_1: High Blood Pressure (Yes)
2. ALCDRS_1: Alcohol Drink Status (Yes)
3. LTPA_1: Leisure Time Physical Activity (Yes)
4. AGE_4: Age (55-64)
5. AGE_5: Age (65+)
6. BMI_3: Obese
7. RACE_2: Non-Hispanic Black
8. RACE_3: Hispanic
9. HBC_1: High Blood Cholesterol (Yes)

The following Table 4.10 to Table 4.12 display the outputs of SAS 'Proc logistic'. The value of − 2 log likelihood for the fitted logistic model is 531.714, and Nagelkerke R^2 is 0.22 (22%). AIC and SC in Table 4.10 are two criteria to assess competing models. When comparing models, lower values of AIC and SC indicate the ones to be preferred. For testing null hypothesis $\beta = 0$, *Chi*-Squares for Wald, Score and Likelihood Ratio with small P values rejected the null hypothesis of $\beta = 0$, indicating these selected predictors are significant. 'Hosmer and Lemeshow Goodness-of-Fit Test' resulted *Chi*-Square of 4.5894, with P value of 0.8004 (df = 8), indicating that there is no statistically significant difference between the observed and predicted classifications. Therefore, the model fits this data set.

Table 4.13 displays measures of association for assessing the predictive ability of the model. The values of Somers's D, Goodman and Kruskal's gamma and *c* were above 0.5 (the closer to 1 the better). The percent of concordant observations, which is the percent of total number of pairs of observations with different outcomes, was 81.3, close to 100 (That is, in this study, one participant could have diabetic outcome and non-diabetic outcome, with predicted probability for observation with observed diabetic outcome higher than predicted probability for observation with observed non-diabetic outcome.). The higher the value of the concordance and the lower the value of discordance, the greater the ability of the model to predict outcome. Therefore, the results indicate the model has a reasonable predictive ability.

Table 4.10: SAS Output of Model Fit Statistics for the Initial Model

```
            Model Fit Statistic
                                   Intercept
                   Intercept          and
Criterion            Only          Covariates

AIC                 649.618           551.714
SC                  654.725           602.788
-2 Log L            647.618           531.714
```

Table 4.11: SAS Output of Hosmer and Lemeshow Test for the Initial Model

```
Hosmer and Lemeshow Goodness-of-Fit Test

   Chi-Square       DF      Pr > ChiSq

       4.5894        8          0.8004
```

Table 4.12: SAS Output of Testing Null Hypothesis $\beta = 0$

and Nagelkerke R^2 of the Initial Model

```
                     The LOGISTIC Procedure

    R-Square    0.0906    Max-rescaled R-Square    0.2200

             Testing Global Null Hypothesis: BETA=0

    Test                 Chi-Square       DF     Pr > ChiSq

    Likelihood Ratio       115.9038        9         <.0001
    Score                  122.5330        9         <.0001
    Wald                    95.3278        9         <.0001
```

Table 4.13: SAS Output of Model Prediction Accuracy for the Initial Model

```
Association of Predicted Probabilities and Observed Responses

     Percent Concordant      81.3    Somers' D    0.640
     Percent Discordant      17.3    Gamma        0.649
     Percent Tied             1.3    Tau-a        0.088
     Pairs                 102830    c            0.820
```

Table 4.14: SAS Output of Estimate of β Coefficients, Standard Error, *Chi*-Square

Statistics and *P* Values for Variables in the Initial Logistic Regression Model

```
                Analysis of Maximum Likelihood Estimates

                                   Standard
Parameter     DF      Estimate        Error      Chi-Square      Pr > ChiSq

Intercept      1       -3.3389       0.3481         91.9779          <.0001
AGE_4          1        1.1776       0.3029         15.1175          0.0001
AGE_5          1        1.1505       0.3072         14.0307          0.0002
RACE_2         1        1.1763       0.3487         11.3774          0.0007
RACE_3         1        0.9279       0.2954          9.8645          0.0017
HBP_1          1        0.5611       0.2588          4.6985          0.0302
HBC_1          1        0.5691       0.2515          5.1230          0.0236
ALCDRS_1       1       -0.8746       0.2639         10.9836          0.0009
LTPA_1         1       -0.6198       0.2497          6.1599          0.0131
BMI_3          1        0.6968       0.2468          7.9700          0.0048
```

Table 4.15: SAS Output of Multiple Adjusted ORs and 95% CI

for Variables in the Initial Model

```
                 Odds Ratio Estimates

                    Point          95% Wald
Effect            Estimate      Confidence Limits

AGE_4                3.247       1.793       5.878
AGE_5                3.160       1.731       5.769
RACE_2               3.242       1.637       6.422
RACE_3               2.529       1.417       4.513
HBP_1                1.753       1.055       2.911
HBC_1                1.767       1.079       2.892
ALCDRS_1             0.417       0.249       0.700
LTPA_1               0.538       0.330       0.878
BMI_3                2.007       1.237       3.256
```

The estimated coefficients of the nine selected predictors, standard error, *Chi*- Square statistics and *P* values are displayed in Table 4.14. Table 4.15 contains multiple adjusted odds ratios and 95% confidence intervals for these selected predictors.

4.4.2 Logistic Regression Model with Interaction Terms

The Pearson *Chi*-Square statistics indicated that the following selected predictors were significantly associated with each other: for demographic predictors, age was associated with race (χ^2 = 70.7273, df = 12, P < 0.0001). For medical condition predictors, high blood pressure was associated with high blood cholesterol (χ^2 = 105.1131, df = 1, P < 0.0001). Among health behavior predictors, overweight/obese was associated with high blood pressure (χ^2 = 56.0826, df = 2, P < 0.0001), high blood cholesterol (χ^2 = 29.3432, df = 2, P < 0.0001) and leisure time physical inactivity (χ^2 = 14.5719, df = 2, P = 0.0007); Leisure time physical activity was associated with alcohol drink status (χ^2 = 18.8356, df = 1, P < 0.0001). For each of these significant associations, a total of ten interaction terms were created, including a three-way interaction term (high blood pressure by high blood cholesterol by obese). The above nine predictors were selected by initial logistic regression model and these ten interaction terms were entered into the stepwise logistic regression again to select best predictors of presence of diabetes.

The second stepwise logistic regression selection procedure generated a very similar model as the initial model. The minor difference was that the interaction term high blood pressure by high blood cholesterol was selected into the second model instead of main effect of high blood pressure and high blood cholesterol variables in the first initial model. Although AIC (Akaike's information criterion, Akaike, 1974) and SC (Schwartz's criterion) were slightly lower in the second model (When comparing models, lower values of AIC and SC indicates the ones to be preferred), and Nagelkerke R^2 was slightly higher (22.13%). The percent concordant was slightly lower compared to initial model. Therefore, the first initial model was preferred as the final model. The following Table 4.16 and Table 4.17 display the SAS output of second model with interaction term.

Table 4.16: SAS Output of Model Fit Statistics for the Model with Interaction Term

```
                 Model Fit Statistics

                                        Intercept
                        Intercept          and
        Criterion          Only         Covariates

        AIC              649.618          548.968
        SC               654.725          594.935
        -2 Log L         647.618          530.968
```

Table 4.17: SAS Output of Model Prediction Accuracy

for the Model with Interaction Term

```
Association of Predicted Probabilities and Observed Responses

     Percent Concordant       80.9    Somers' D    0.641
     Percent Discordant       16.8    Gamma        0.655
     Percent Tied              2.3    Tau-a        0.088
     Pairs                  102830    c            0.820
```

4.4.3. The Receiver-Operating Characteristic Curve Analysis for the Final Logistic Model

The final model for prediction of probability of diabetes presence was the following:

Log [$\hat{P}$ / (1 – $\hat{P}$)] = –3.3389 + 1.1776 (Age_55-64 years) + 1.1505 (Age_64+ years) + 1.1763 (Non-Hispanic black) + 0.9279 (Hispanic) + 0.5611 (High blood pressure) + 0.5691 (High blood cholesterol) – 0.8746 (Alcohol drink status_yes) – 0.6198 (Leisure time physical activity_yes) + 0.6968 (Obese),

Where as $\hat{P}$ = estimated probability of presence of diabetes

To evaluate the overall performance of the above final logistic regression equation, the measurement of discrimination was considered. Discrimination was defined as the ability of the equation to distinguish high-risk subjects from low-risk subjects and is quantified by the area under the receiver-operating characteristic (ROC) curve. (Steuerberg EW, et al., 2001). Table 4.18 displays SAS output of classification table. It contains the number (and percent) of correctly and incorrectly classified responses for different cutpoints. It also provides the percentage of correct (100*(sum of diabetic and non-diabetic correctly classified / total number of observations)), sensitivity (100*number of correctly classified diabetic / total number of diabetic), specificity (100*number of correctly classified non-diabetic / total number of non-diabetic), false positive (100* the number of observations classified incorrectly as diabetic / total number of observations classified as diabetic), false negative (100* the number of observations classified incorrectly as non-diabetic / total number of cases classified as non-diabetic).

Figure 4.17 displays the ROC curve. A ROC curve was constructed by plotting sensitivity against the false-positive rate (1-specificity) over a range of cut-point values. As mentioned in Chapter 3, the area under the ROC curve quantifies how well the model correctly distinguishes a diabetic from a non-diabetic. The larger the area under the curve, and more accurate the prediction model (A perfect model has a value of 1). Each point on the curve represents a cutoff probability. A lower cutoff typically gives more false positive. A higher cutoff gives more false negatives, a low sensitivity, and a high specificity. The best cut point is at or near the shoulder of the ROC curve. If the predicted probability of presence of diabetes exceeds the optimal cutpoint then the model classifies this person as a diabetic, otherwise the person is classified as a non-diabetic. The area under the ROC curve is given by the statistic *c* in the Table 4.13, which is 0.82 (82%). Therefore, the final model has a reasonable predictive ability. The probability level that provided an optimal cutpoint was 0.10. Based on the classification table, at this optimal cutpoint, sensitivity was 59.3%; specificity was 81.7%; the final logistic regression model correctly predicted 80% of participants; and this fitted model produced the confusion matrix in Table 4.19. Individuals who had predicted probability generated by applying the estimated coefficients in the final equation higher than this threshold (0.10) were classified as diabetics by the model even they reported themselves as non-diabetics. These individuals could be identified as group of people at high risk and need further diagnostic investigation. When this model is applied to the population, a lower (higher) cutpoint would be considered to meet expectation of higher (lower) sensitivity and lower (higher) specificity.

Table 4.18: SAS Output of Classification Table for the Fitted Initial Model

```
                          Classification Table

            Correct         Incorrect                    Percentages
 Prob              Non-             Non-             Sensi-   Speci-   False    False
 Level  Event     Event    Event   Event   Correct   tivity   ficity    POS      NEG

 0.000    91         0     1130       0       7.5    100.0      0.0    92.5       .
 0.020    87       393      737       4      39.3     95.6     34.8    89.4     1.0
 0.040    79       605      525      12      56.0     86.8     53.5    86.9     1.9
 0.060    71       768      362      20      68.7     78.0     68.0    83.6     2.5
 0.080    66       842      288      25      74.4     72.5     74.5    81.4     2.9
 0.100    54       923      207      37      80.0     59.3     81.7    79.3     3.9
 0.120    53       981      149      38      84.7     58.2     86.8    73.8     3.7
 0.140    50       994      136      41      85.5     54.9     88.0    73.1     4.0
 0.160    38      1027      103      53      87.2     41.8     90.9    73.0     4.9
 0.180    32      1049       81      59      88.5     35.2     92.8    71.7     5.3
 0.200    31      1063       67      60      89.6     34.1     94.1    68.4     5.3
 0.220    29      1066       64      62      89.7     31.9     94.3    68.8     5.5
 0.240    27      1068       62      64      89.7     29.7     94.5    69.7     5.7
 0.260    19      1074       56      72      89.5     20.9     95.0    74.7     6.3
 0.280    15      1088       42      76      90.3     16.5     96.3    73.7     6.5
 0.300    14      1099       31      77      91.2     15.4     97.3    68.9     6.5
 0.320    12      1099       31      79      91.0     13.2     97.3    72.1     6.7
 0.340    11      1102       28      80      91.2     12.1     97.5    71.8     6.8
 0.380    10      1107       23      81      91.5     11.0     98.0    69.7     6.8
 0.420     3      1109       21      88      91.1      3.3     98.1    87.5     7.4
 0.440     3      1118       12      88      91.8      3.3     98.9    80.0     7.3
 0.460     2      1120       10      89      91.9      2.2     99.1    83.3     7.4
 0.520     2      1124        6      89      92.2      2.2     99.5    75.0     7.3
 0.540     2      1125        5      89      92.3      2.2     99.6    71.4     7.3
 0.580     2      1127        3      89      92.5      2.2     99.7    60.0     7.3
 0.600     2      1128        2      89      92.5      2.2     99.8    50.0     7.3
 0.620     2      1128        2      89      92.5      2.2     99.8    50.0     7.3
 0.640     1      1128        2      90      92.5      1.1     99.8    66.7     7.4
 0.660     1      1128        2      90      92.5      1.1     99.8    66.7     7.4
 0.680     1      1129        1      90      92.5      1.1     99.9    50.0     7.4
 0.700     0      1129        1      91      92.5      0.0     99.9   100.0     7.5
 0.720     0      1129        1      91      92.5      0.0     99.9   100.0     7.5
 0.740     0      1130        0      91      92. 5     0.0    100.0      .       7.5
```

Table 4.19: Prediction of the Final Model at the Optimal Cutpoint

	True Diabetes	True Non-Diabetes	Total
Predicted Diabetes	54	207	261
Predicted Non-Diabetes	37	923	960
Total	91	1130	1221

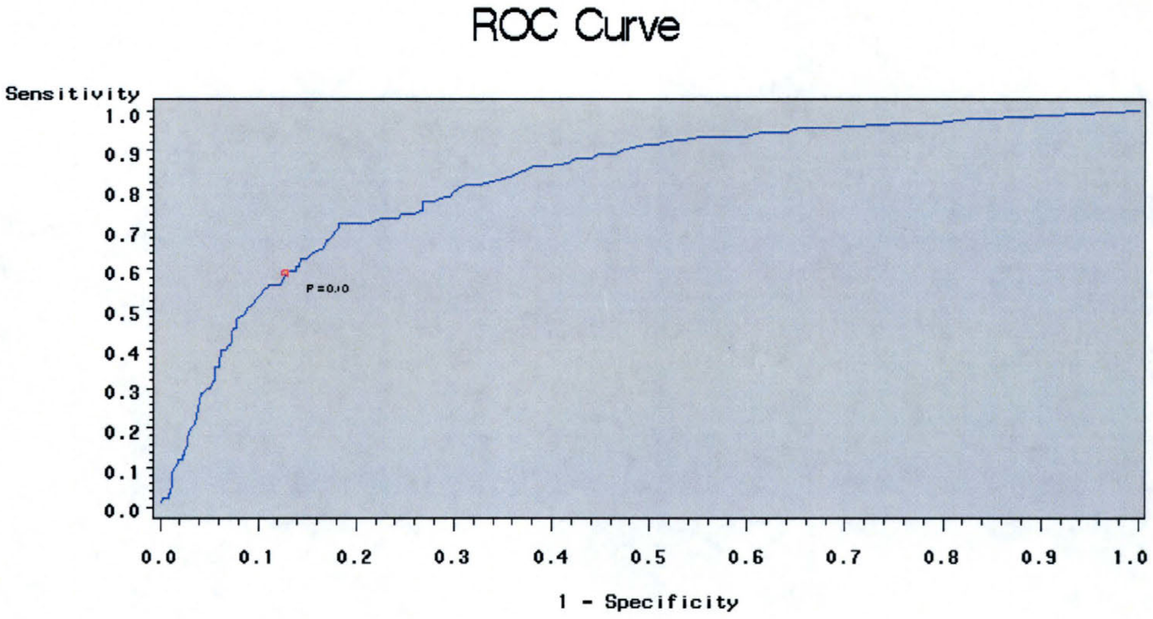

Figure 4.17: The Receiver-Operating Characteristic Curve

4.4.4 Casewise Diagnostics for the Fitted Model

To detect potential outliers, the following values were calculated: devres (deviance residuals for identifying poorly fitted observations), chires (Pearson residuals also useful for identifying observations that are not well explained by the fitted model), and phat (the predicted probability obtained by substituting the estimated regression coefficient in equation). Any observation with absolute value of devres or chires greater than 2, were considered as potential outliers. Table 4.20 displays SAS output of potential outliers and their values of devres, chires and phat. There were 39 observations had both deviance and Pearson residuals greater than 2. The positive values of the residuals in each case indicate that the predicted value of diabetes presence for these observations is far smaller than the observed values. The fitted model misclassified them as non-diabetics when they reported themselves as diabetics. The result indicates that diabetes prevalence would be underestimated if using this fitted logistic model to predict diabetes presence. The index plots of Deviance and Pearson residuals are displayed in Figure 4.18 and Figure 4.19, respectively, indicating the same suggestion. These extreme potential outliers need closer investigation.

Table 4.20: Casewise Diagnostics for the Fitted Model

```
            Casewise Diagnostics for Fitted Model

 Obs    DIABETES      devres       chires       phat

   15       1        2.48110      4.55122    0.04605
   39       1        2.26856       3.4796    0.07629
   49       1        2.71339       6.2206    0.02519
   62       1        2.09835       2.8353    0.11063
   71       1        3.11167      11.2083    0.00790
  128       1        2.10980       2.8739    0.10800
  134       1        2.38380       4.0171    0.05835
  141       1        3.11167      11.2083    0.00790
  175       1        2.73037      6.36970    0.02405
  241       1        2.14225      2.98676    0.10080
  302       1        2.11839      2.90327    0.10606
  328       1        2.09835      2.83530    0.11063
  339       1        2.34872      3.84346    0.06340
  506       1        2.32847      3.74733    0.06648
  509       1        2.61639      5.44551    0.03262
  563       1        2.48357      4.56586    0.04577
  564       1        2.00049      2.52909    0.13520
  622       1        2.40247      4.11343    0.05580
  623       1        2.14225      2.98676    0.10080
  631       1        2.13143      2.94855    0.10316
  692       1        2.92511      8.43244    0.01387
  794       1        2.25873      3.43790    0.07801
  846       1        2.50557      4.69889    0.04333
  849       1        2.59661      5.30221    0.03435
  936       1        2.29240      3.58328    0.07225
  940       1        2.56884      5.10868    0.03690
  942       1        2.09583      2.82687    0.11122
  944       1        2.15078      3.01729    0.09897
  950       1        2.26269      3.45462    0.07731
  953       1        2.39447      4.07180    0.05688
  969       1        2.45385      4.39333    0.04926
 1006       1        2.01108      2.56030    0.13236
 1047       1        2.82051      7.23810    0.01873
 1143       1        2.13089      2.94665    0.10328
 1152       1        2.54879      4.97424    0.03885
 1186       1        2.22928      3.31649    0.08334
 1198       1        2.16683      3.07576    0.09560
 1203       1        2.33063      3.75744    0.06614
 1211       1        2.14225      2.98676    0.10080
```

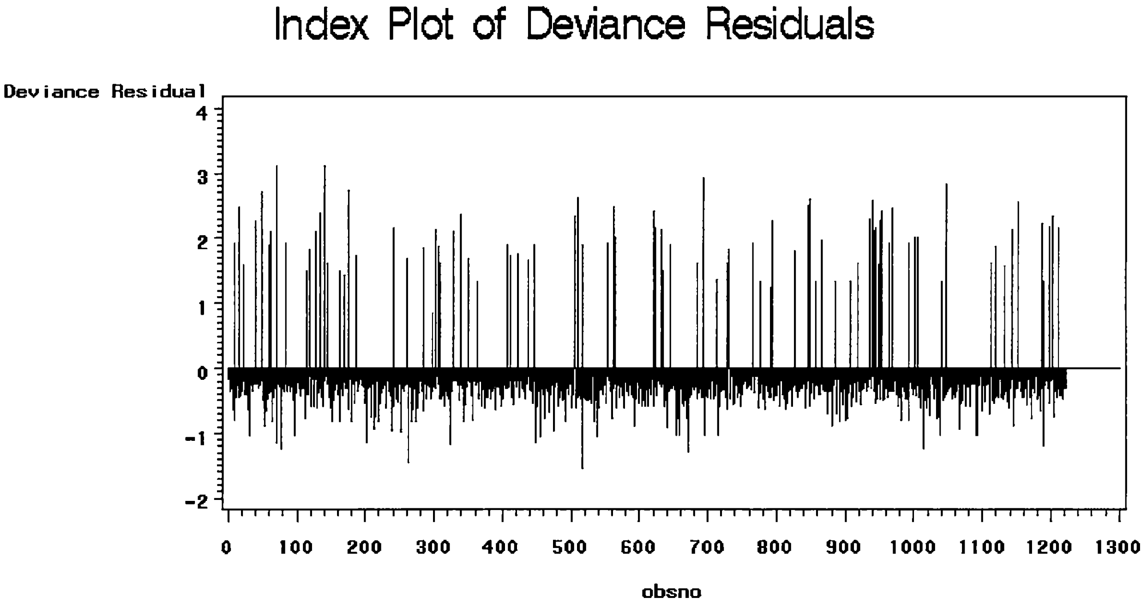

Figure 4.18: Index Plot of Deviance Residuals

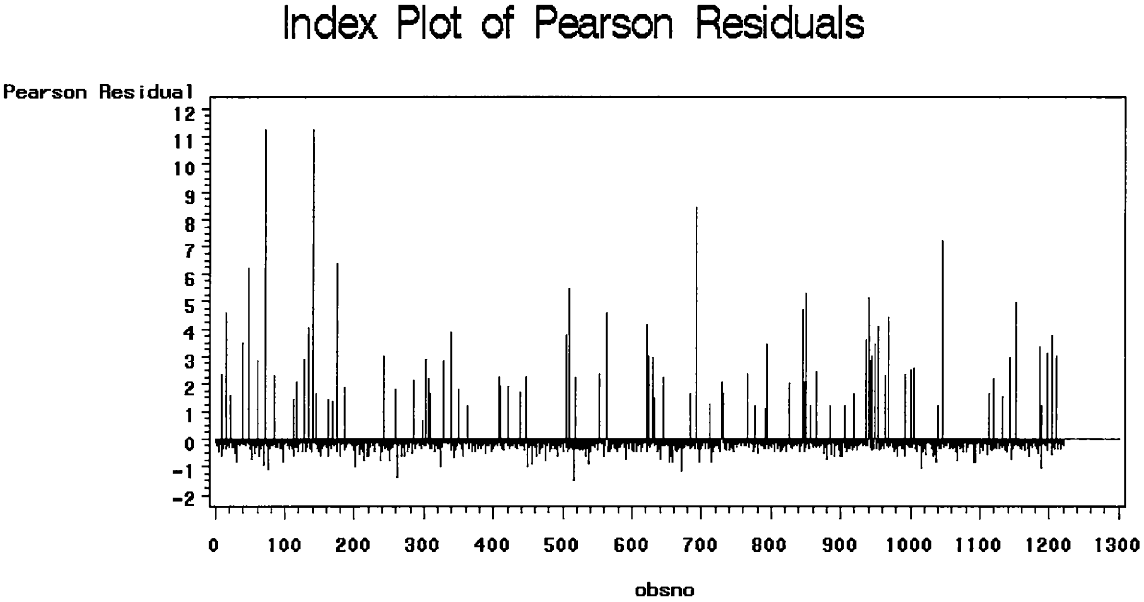

Figure 4.19: Index Plot of Pearson Residuals

4.4.5 Validation of the Fitted Model

To validate the equation, the final fitted model was applied to TDH BRFSS survey data of year 1999 that had not been used to generate the equation. There was 4990 Texans participated TDH BRFSS survey in year 1999. Of the 4990 participants, 296 (5.9%) were diabetics, excluding women (52 women, 1.0%) who had diabetes only during their pregnancies. The whole sample was randomly split into A and B group as in the survey of year 2001. The following variables were available and used in validation. Variable number nine and number ten were only answered by survey group B.

1. Diabetes Status (1 = Diabetes; 2 = Diabetes only during pregnancy; 3 = Non-diabetes)
2. Age (5 groups same as in TDH BRFSS survey of year 2001)
3. Race (4 groups same as in TDH BRFSS survey of year 2001)
4. Education (1 = Grades 1-8; 2 = Grades 9-11; 3 = Grades 12 or GED; 4 = College 1-3; 5 = College graduate)
5. Household Income (8 groups same as in TDH BRFSS survey of year 2001)
6. Sex (1 = Male; 2 = Female)
7. High Blood Pressure (1 = Yes; 2 = No)
8. High blood Cholesterol (1 = Yes; 2 = No)
9. Diet (Are you eating fewer high fat or high cholesterol foods? 1 = Yes; 2 = No)
10. Exercise (Are you exercising more? 1 = Yes; 2 = No)
11. Smoking Risk (1 = Not At Risk; 2 = At Risk)
12. Chronic Alcohol Risk (>/= 60 drinks in the past month. 1 = Not At Risk; 2 = At Risk)

13. Acute Alcohol Risk (same definition as in TDH BRFSS survey of year 2001)
14. Overweigh/Obese (same definition as in TDH BRFSS survey of year 2001)

The following variables were not included in survey of year 1999, and could not be found:

- Alcohol Drink Status
- Moderate Activity
- Vigorous Activity
- Met Recommendations for Physical Activity
- Leisure Time Physical Activity

Of these 4990 participants, 1633 had no missing values in the above 14 variables and were used as validation sample. The logistic regression stepwise selection generated the following predictors similar as in the final fitted logistic regression model by using survey data of year 2001:

1. High Blood Pressure (Yes)
2. Obese
3. High Blood Cholesterol (Yes)
4. Overweight
5. Household Income ($10,000 - $14,000)
6. Age (55-64)
7. Age (65+)
8. Others (Race)
9. Age (45-54)
10. Age (35-44)

11. Hispanic

The following Table 4.21 to Table 4.24 display outputs of SAS logistic procedure. These results were very similar as the final fitted logistic model. This model generated from validation sample had a –2 likelihood value of 698.354 and a Nagelkerke R^2 of 0.2273 (22.73%). *Chi*-Squares for Wald, Score and Likelihood Ratio with small *P* values rejected the null hypothesis of $\beta = 0$, indicating these selected predictors are significant. ‘Hosmer and Lemeshow Goodness-of-Fit Test’ resulted in *Chi*-Square of 7.4923, with *P* value of 0.4846 (df = 8), indicating that there is no statistically significant difference between the observed and predicted classifications. The percent of concordant (81.8), Somer’s D (0.648), Kruskal’s gamma (0.655) and *c* (0.824) were also about same as those for final model. But this model had larger values of AIC (722.354) and SC (787.132) than these of values (AIC = 548.968, SC = 594.935) for final model, indicating that the model generated from survey data of year 2001 is preferred. It also suggests that some of these variables that survey of year 1999 did not collected are very important predictors of diabetes presence and should be included in the future surveys.

Table 4.21: SAS Output of Model Fit Statistics

for the Model Generated from the Validation Sample

```
                  Model Fit Statistics

                                          Intercept
                         Intercept           and
        Criterion          Only          Covariates

        AIC              859.521            722.354
        SC               864.920            787.132
        -2 Log L         857.521            698.354
```

Table 4.22: SAS Output of Testing Null Hypothesis $\beta = 0$

and Nagelkerke R^2 of the Model Generated from the Validation Sample

```
                  The LOGISTIC Procedure

 R-Square    0.0929    Max-rescaled R-Square    0.2273


           Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio       159.1673       11         <.0001
Score                  157.7093       11         <.0001
Wald                   113.2740       11         <.0001
```

Table 4.23: SAS Output of Hosmer and Lemeshow Test

for the Model Generated from the Validation Sample

```
Hosmer and Lemeshow Goodness-of-Fit Test

   Chi-Square       DF     Pr > ChiSq

       7.4923        8         0.4846
```

Table 4.24: SAS Output of Model Prediction Accuracy

for the Model Generated from the Validation Sample

```
Association of Predicted Probabilities and Observed Responses

     Percent Concordant       81.8    Somers' D    0.648
     Percent Discordant       17.1    Gamma        0.655
     Percent Tied              1.1    Tau-a        0.088
     Pairs                  181560    c            0.824
```

The estimated coefficients of the eleven selected predictors, standard error, *Chi-Square* statistics and *P* values generated for the validation sample are displayed in table 4.25. Table 4.26 contains multiple adjusted odds ratios and 95% confidence intervals.

These positive estimated coefficients with small *P* values for *Chi*-Square statistics indicate that these selected predictors are positively associated with diabetes. Multiple adjusted odds ratios for selected age groups increased as participant's age increased compared to youngest age group (18-34) indicating that age is a significant risk factor for diabetes. The odds ratio for Hispanic (1.675, 95% CI: 1.027 – 2.733) and 'others' race group (4.391, 95% CI: 1.612 – 11.959 indicate that Hispanic and Others race groups are higher risk ethnicities for diabetes, compared to Non-Hispanic white people. High blood pressure (adjusted OR = 1.844, 95% CI: 1.206 – 2.820) and high blood cholesterol (adjusted OR = 1.830, 95% CI: 1.1.216 – 2.753) also are risk factors for diabetes. Increased adjusted odds ratios were seen in 'body mass index categories' (adjusted OR = 2.421, 95% CI: 1.354 – 4.328 for overweight group; adjusted OR = 5.083, 95% CI: 2.837 – 9.105 for obese group), indicating that risk for diabetes increased as body mass index increased. The other predictor that was not selected by the final fitted model was low household income category $10,000 - $14,000 (adjusted OR = 2.219, 95% CI: 1.179 – 4.174, compared to category of $75,000+). These results were similar to the results generated by analyzing the survey sample of year 2001, which is that age, race, blood pressure, blood cholesterol and body mass index are important predictors for diabetes. The validation results lead to the conclusion that the final fitted model is generalizable to whole Texas population and not specific to the survey sample of year 2001 that was used in model estimation.

Table 4.25: SAS Output of Estimate of β Coefficients, Standard Error, *Chi*-Square Statistics and *P* Values for Variables in the Logistic Regression Model Generated from the Validation Sample

```
              Analysis of Maximum Likelihood Estimates

                                  Standard
Parameter    DF     Estimate       Error     Chi-Square    Pr > ChiSq

Intercept     1      -6.1603      0.6301        95.5927        <.0001
AGE_2         1       1.6057      0.6353         6.3888        0.0115
AGE_3         1       2.0110      0.6244        10.3732        0.0013
AGE_4         1       2.5877      0.6220        17.3085        <.0001
AGE_5         1       2.6564      0.6295        17.8055        <.0001
RACE_3        1       0.5157      0.2498         4.2633        0.0389
RACE_4        1       1.4795      0.5112         8.3762        0.0038
INCOME_2      1       0.7969      0.3225         6.1066        0.0135
HBP_1         1       0.6120      0.2166         7.9809        0.0047
HBC_1         1       0.6042      0.2085         8.3997        0.0038
BMI_2         1       0.8842      0.2964         8.8980        0.0029
BMI_3         1       1.6258      0.2974        29.8784        <.0001
```

Table 4.26: SAS Output of Multiple Adjusted ORs and 95% CI for Variables in the Model Generated from the Validation Sample

```
                    The LOGISTIC Procedure

                     Odds Ratio Estimates

                 Point          95% Wald
Effect        Estimate      Confidence Limits

AGE_2            4.981       1.434      17.302
AGE_3            7.471       2.197      25.400
AGE_4           13.300       3.930      45.007
AGE_5           14.246       4.148      48.927
RACE_3           1.675       1.027       2.733
RACE_4           4.391       1.612      11.959
INCOME_2         2.219       1.179       4.174
HBP_1            1.844       1.206       2.820
HBC_1            1.830       1.216       2.753
BMI_2            2.421       1.354       4.328
BMI_3            5.083       2.837       9.105
```

4.4.6 Conclusion

By using logistic technique, age (55-64 years group), age (65+ years group), race (Non-Hispanic black), race (Hispanic), high blood pressure (Yes), high blood cholesterol (Yes), alcohol drink status (Yes), leisure time physical activity (Yes) and obese were selected as significant predictors. The final model for prediction of probability of diabetes presence is found to be

Log [$\hat{P}$ / (1 – $\hat{P}$)] = –3.3389 + 1.1776 (Age_55-64 years) + 1.1505 (Age_64+ years) + 1.1763 (Non-Hispanic black) + 0.9279 (Hispanic) + 0.5611 (High blood pressure) + 0.5691 (High blood cholesterol) – 0.8746 (Alcohol drink status_yes) – 0.6198 (Leisure time physical activity_yes) + 0.6968 (Obese),

Where as $\hat{P}$ = estimated probability of presence of diabetes

A ROC curve analysis selected an optimal cutpoint of predicted probability of 0.10 as a threshold. Individuals who had a predicted probability generated by applying the estimated coefficients in the final equation higher than this threshold were classified as diabetics by the model even if they reported themselves as non-diabetics. These individuals could be identified as a group of people at high risk and need further diagnostic investigation. Based on the classification table, at this optimal cutoff value, the sensitivity was 59.3%, the specificity was 81.7%, and the final logistic regression model correctly predicted 80% of participants. When this model is applied to the population, a lower (higher) cutpoint would be considered to meet expectation of higher (lower) sensitivity and lower (higher) specificity.

For the purpose of validation of the final logistic regression equation, the final model was applied to TDH BRFSS survey data of year 1999. Age (35-44 years), Age

(45-54 years), Age (55-64 years), Age (65+ years), Rave (Hispanic), Race (Others), Income ($10,000 - $14,000), High Blood Pressure (Yes), High Blood Cholesterol (Yes), Overweight and Obese were selected as significant predictors by analyzing validation sample. The validation results led to the conclusion that the final fitted model is generalizable to the whole Texas population and not specific to the survey sample of year 2001 that was used in model estimation. Because the model generated from the validation sample had larger values of AIC (722.354) and SC (787.132) than these of values (AIC = 548.968, SC = 594.935) for the final model, indicating that the model generated from survey data of year 2001 is preferred. It also suggests that some of these variables that the survey of year 1999 did not collect are very important predictors of diabetes presence and should be included in the future surveys.

Multiple adjusted odds ratios for age (55-64 years group) and age (65+ years group) were 3.247 (95% CI: 1.793 – 5.878) and 3.160 (95% CI: 1.731 – 5.769), respectively, indicating that participants who were 55 years or older were about three times as likely to be a diabetic than participants who were just 18-24 years old. This result suggests that age of 55 years could be used as cut-off age point. Individuals who are 55 years or older should pay more attention to prevention of diabetes. These Non-Hispanic black and Hispanic had an adjusted odds ratio of 3.242 (95% CI: 1.637 – 6.422), 2.529 (95% CI: 1.417 – 4.513), respectively, indicating non-Hispanic blacks and Hispanics are at higher risk for diabetes, compared to non-Hispanic white people. High blood pressure and high blood cholesterol also increased the risk for presence of diabetes (adjusted odds ratio was 1.753, with 95% CI: 1.005 – 2.911 for high blood pressure and 1.767, with 95% CI: 1.079 – 2.892 for high blood cholesterol). An adjusted odds ratio of

2.007 (95% CI: 1.237 – 3.256) suggested that people whose body mass index was 30 or higher had a doubled risk for diabetes compared to people whose body mass index was less than 25. Adjusted odds ratios were 0.417 (95% CI: 0.249 – 0.700) for alcohol drinkers and 0.538 (95% CI: 0.330 – 0.878) for people who did leisure time physical activity, indicating that appropriate consumption of alcohol and physical activity may protect people from diabetes.

It is important to note that the percent concordant for prediction is 81.3% and Nagelkerke R^2 is 0.22. Although this generated final logistic regression model has reasonable prediction ability, other risk factors not measured in this study also predict risk of diabetes, such as family history, vitamin supplements and other related chronic diseases.

CHAPTER 5

DISCUSSION

In this study, diabetes prevalence in Texas in year 2001 was higher than the 1996 – 1999 yearly average diabetes prevalence (5.4%) (Weihua Li, et al. 2001) and national estimated prevalence of diabetes (6.2%) in year 2000 (National Diabetes Statistics. 2003). Similar trends of diabetes prevalence in age, household income, and education categories in this study were found as those in analysis of Texas 1996 – 1999 survey data (Weihua Li, et al. 2001), that is, older individuals are at increased risk for diabetes while persons with higher education levels and higher household incomes have a decreased risk for diabetes. Non-Hispanic blacks and Hispanics had a higher prevalence of diabetes than non-Hispanic whites. The odds of diabetes in non-Hispanic blacks and Hispanics are 3.242 times and 2.529 times, respectively greater than that of non-Hispanic whites. A higher prevalence of diabetes was seen in obese individuals, alcohol non-drinkers and leisure time inactive persons than in individuals with normal BMI, alcohol drinkers and who engaged in leisure time physical activities. These results are consistent with the results of Texas 1996 – 1999 survey data (Weihua Li, et al. 2001).

The independent predictors of diabetes selected and significant risk factors identified in this study are similar to those found in previous studies. A study by Bahman P. Tabaei, William H. Herman (2002) generated a prediction model for diabetes.

Age and body mass index were selected as independent predictors in the logistic regression equation. In another study conducted by S. Carlsson, et al., (2000), multiple logistic regression analysis also included body mass index, age, physical activity, smoking, a family history of diabetes, and alcohol consumption as predictors for diabetes.

Although smoking risk, low intake of fruits and vegetables and the interaction term of Obesity by physical inactivity were not selected into the final logistic regression model, it is widely accepted that lack of exercise, a poor diet, current smoking, and abstinence from alcohol use are all associated with a significantly increased risk of diabetes, even after adjustment for BMI (Hu FB, Manson JE, Stampfer MJ, et al. 2001). It is also established that the risk for Type 2 diabetes dramatically increases when obesity is compounded by physical inactivity (A joint editorial statement. 1999).

In this sample, more than 70% of participants reported a low intake of fruits and vegetables. This finding should be interpreted with caution. It does not mean that lower fruit and vegetable intake protects people from diabetes. The finding could be a result of information bias. Because the survey questionnaire considered the time period 'in the past 30 days', some diabetics diagnosed before the 'past 30 days' period, might have changed their habits and ate more fruit and vegetables, in the 'past 30 days period' after they knew they had diabetes. A future recommendation for this study is that the data should include information before and after diabetes had been diagnosed to minimize the information bias.

In summary, the nine predictors of diabetes presence in this study were age (55-64 years group), age (65+ years group), race (Non-Hispanic black), race (Hispanic), high blood pressure (Yes), high blood cholesterol (Yes), alcohol drink status (Yes), leisure time physical activity (Yes) and obesity. A physically active non-Hispanic white person, who is a moderate level alcohol drinker aged younger than 55 years, with normal blood pressure, normal blood cholesterol, BMI less than 25, has a 0.79% probability to be a diabetic. The predicted probability of being diabetic will rise to 69.317%, if a person is physically inactive, non-Hispanic black, who is a non-drinker, older than 65 years, with high blood pressure, high blood cholesterol, and body mass index greater than 30. However, the final logistic regression model misclassified 39 diabetics in sample as non-diabetics. The predicted probability of diabetes for these observations is far smaller than the observed values. When this prediction equation is applied to the population, prevalence of diabetes will be underestimated because of lack of other important predictors in the final model. Therefore, a threshold of predicted probability of 0.10 generated from ROC curve analysis should be considered for the purpose of prevention and early detection. Individuals who had predicted probability greater than this threshold could be identified as group of people at high risk, and they need further medical diagnostic investigation. When this model is applied to the population, a lower (higher) cutpoint of predicted probability would be considered to meet the expectation of higher (lower) sensitivity and lower (higher) specificity.

The final logistic regression model for this study yields a relatively low coefficient of determination (0.22). This result suggests that variables not accounted for in this study, such as family history, vitamin supplements and other related chronic diseases, might explain a significant proportion of the variance. Another recommendation for the future study is that information about these missed variables should be collected and considered into future studies. A more accurate predicted probability of diabetes presence will be generated by the enrichment of the information.

REFERENCES

A. J. Farmer, J. C. Levy and R. C. Turner (1999). Knowledge of Risk of Developing Diabetes Mellitus among Siblings of Type 2 Diabetic Patients 1999 British Diabetic Association. Diabetic Medicine, 16, 233-237

A joint editorial statement by the American Diabetes Association; The National Heart, Lung, and Blood Institute; The Juvenile Diabetes Foundation International; The National Institute of Diabetes and Digestive and Kidney Disease; and The American Heart Association. (1999). Diabetes mellitus: a major risk factor for cardiovascular diseases. Circulation. 1999; 100: 1132-1133.

American Diabetes Association. Diabetes (2001). 2001 Vital Statistics. Alexandria, VA: American Diabetes Association.

American Diabetes Association (2003). Screening for Type 2 Diabetes. Diabetes Care, volume 26, Supplement, January 2003

Anastasia C. Thanopoulou, et al (2003). Dietary Fat Intake as Risk Factor for the Development of diabetes Diabetes Care volume 26, Number 2, February 2003

Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics.* 1955; 11:375–386.

Asia Pacific Cohort Studies Collaboration (2003). The Effects of Diabetes on the Risks of Major Cardiovascular Diseases and Death in the Asia-Pacific Region Diabetes Care 26:360-366, 2003

Bahman P. Tabaei, William H. Herman, (2002). A Multivariate Logistic Regression Equation to Screen for Diabetes. Diabetes Care, volume 25, Number 11, November 2002.

Baiju R. Shah, Janet E. Hux, (2003). Quantifying the Risk of Infectious Diabetes for People with Diabetes. Diabetes Care volume 26, Number 2, February 2003

Barbara Fletcher, Meg Gulanick, Cindy Lamendola. (2002). Risk Factors for Type 2 Diabetes Mellitus. *J* Cardiovasc Nurs 2002; 16(2):17-23

Bjornholt, Jorgen V. et al. (2000). Type 2 Diabetes and Maternal Family History Diabetes Care, 01495992, Sep2000, Vol. 23, Issue 9

Brad McCulloch, BSC. Gdipcompsci et al. (2003). Self-Reported Diabetes and Health Behaviors in Remote Indigenous Communities in Northern Queensland, Australia. Diabetes Care, volume 26, Number 2, February 2003.

B. S. Everitt and G. Der (1997). A Handbook of Statistical Analyses Using SAS. New York, NY. Chapman & Hall press.

Centers for Disease Control and Prevention. (1998). BRFSS User's Guide, 1998, Survey samples and sampling methods.pp3-2; Data Management, 00.8-2.

Centers for Disease Control and Prevention. Recommendations http://www.cdc.gov/nccdphp/dnpa/physical/recommendations/index.htm

Cochran WG. Some methods for strengthening the common χ^2 tests. *Biometrics.* 1954; 10:417–451.

C. Trautner, B. Haastert, G. Giani and M. Berger (2002). Amputations and Diabetes: A Case-Control Study 2002 Diabetes UK. Diabetic Medicine, 19, 35-40

D. W. Hosmer, Jr. and S. Lemeshow (1988). Applied Logistic Regression, New York, NY. Wiley Press.

Hair Anderson, Tatham Black (1998). Multivariate Data Analysis Fifth Edition. New Jersey, Prentice-Hall Inc. Press.

Haire-Joshu D, Glasgow RE, Tibbs TL. (1999). Smoking and Diabetes. Diabetes Care 22: 1887-1898

Harris MI, Flegal KM, Cowie CC, et al. (1998). The Third National Health and Nutrition Examination survey, 1988-1994; prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in US adults. Diabetes Care. 1998; 21:518-524.

Hu FB, Manson JE, Stampfer MJ, et al. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. N Engl J Med. 2001;345:790-797

Hu FB Sigal RJ, Rich-Edwards JW, et al. (1999). Walking Compared With Vigorous Physical Activity and Risk of Type 2 Diabetes in Women: A Prospective Study. JAMA. 1999; 282: 1433-1439.

Hu FB, van Dam RM, Liu S. (2001). Diet and risk of Type II Diabetes: the Role of Types of Fat and Carbohydrate. Diabetologia 44:805-817,2001.

Jennifer C. Lovejoy, et al. (2002). Effects of Diets Enriched in Saturated (Palmitic), Monounsturated (Oleic), or Tran (Elaidic) Fatty Acids on Insulin Sensitivity and substrate Oxidation in Healthy Adults. Diabetes Care, volume 25, Number 8, August 2002.

Joslin E (1921). The Prevention of Diabetes Mellitus. JAMA. 1921; 76:76-84.

Karin M. Nelson, et al. (2002). Diet and Exercise Among Adults With Type 2 Diabetes. Diabetes Care, volume 25, Number 10, October 2002.

Kissebah AH. (1996). Intra-abdominal fat: Is It A Major Factor in Developing Diabetes and Coronary Artery disease? Diabetes Res Clin Pract. 1996; 30 (Suppl): 25-30.

Lloyd D. Fisher, Gerald Van Belle (1993). Biostatistics A Methodology for the Health Sciences. New York, NY. John Wiley & Sons, Inc press.

Lora D. Delwiche and Susan J. Slaughter (1998). The Little SAS Book. A Primer Second Edition Cary NC. SAS Institute Inc. press.

McMullagh, P. and Nelder, J. A. (1989). Generalized Linear Models. London. Chapman & Hall Press.

Meigs JB, Cupples LA, Wilson, PW. (2000). Parental transmission of type 2 diabetes: the Framingham Offspring Study. Diabetes. 2000; 49:2201-2207

Meyer K. Jacobs D Jr. Kushi 1., Folsom A. (2001). Dietary Fat and Incidence of Type 2 Diabetes in Older Iowa Women. Diabetes Care 24:1528-1535, 2001

Morgan C, Currie C, Peters J (2000). Relationship between diabetes and mortality: a population study using records linkage. Diabetes Care 23:1103-1107

National Diabetes Statistics. (2003) NIH Publication No. 03-3892; May 2003. http://diabetes.niddk.nih.gov/dm/pubs/statistics/index.htm

O. Rolandsson, et al. (2001). Prediction of Diabetes with Body Mass Index, Oral Glucose Tolerance Test and Islet Cell Autoantibodies in a regional Population. Journal of Internal Medicine 2001; 249: 279-288

Peterson, Kevin. (2003). The Diabetes ABC's Diabetes Forcast, 00958301, Feb2003, Vol.56 Issue 2.

P. –G. Persson, S. Carlsson, et al. (2000). Cigarette Smoking, Oral Moist Snuff Use and Glucose Intolerance. Journal of Internal Medicine 2000; 248: 103-110

Philip S. Mehler, et al., (1998). Smoking as a Risk Factor for Nephropathy in Non-Insulin –Dependent Diabetes. J GEN INTERN MED 1998; 13:842-845

SAS Institute Inc. (1987). SAS System for Elementary Statistical Analysis Cary NC, SAS Institute Inc. press.

S. Bo, P. Cavallo-Perin, L. Gentile, E. Repetti and G. Pagano, (2000). Influence of a family History of Diabetes on the Clinical Characteristics of Patients with Type 2 Diabetes Mellitus. 2000 Diabetes UK. Diabetic Medicine, 17, 538-542

S. Carlsson, et al., (2000). Alcohol Consumption, Type 2 diabetes mellitus and impaired glucose tolerance in middle-aged Swedish. 2000 Diabetes UK. Diabetic Medicine, 17, 776-781

Stanto a Glantz , Bryan K. Slinker, (2001). Primer of Applied Regression & Analysis of Variance Second Edition, New York, McGraw-Hill Inc. Press.

Steuerberg EW, et al., (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 54:774-781, 2001

TDH Diabetes Council Report (2001). Texas Diabetes Fact Sheet 2001. http://www.tdh.state.tx.us/diabetes/data/facts.pdf

The Burden of Heart Disease, Stroke, Cancer, and Diabetes, United States http://www.cdc.gov/nccdphp/burdenbook2002/02_diabetes.htm

Tomoshige Hayashi, et al. (1999). High Normal Blood Pressure, Hypertension, and the Risk of Type 2 Diabetes in Japanese Men. Diabetes Care, 01495992, Oct. 99, Vol.22, Issue 10.

Tsumura, Kei, Hayashi, et al.(1999). Daily Alcohol Consumption and the Risk of Type 2 Diabetes in Japanese Men. Diabetes Care, 01495992, Sep. 99, Vol.22, Issue 9.

Weihua Li, et al. (2001). Texas BFRSS Diabetes in Texas: A Risk Factor Report 1996-1999 Survey Data Publication No. 16-11164 June 2001

APPENDIX A – Variables Entered into the Initial Logistic Regression Model in Alphabetical Order

ACALCR_2	Acute alcohol drinking risk_at risk
AGE_2	Age_35-44 years group
AGE_3	Age_45-54 years group
AGE_4	Age_55-64 years group
AGE_5	Age_65+ years group
ALCDRS_1	Alcohol drinking status_yes
BMIR_2	Overweight (BMI: 25 – 29.9 kg / m^2)
BMIR_3	Obesity (BMI > 30 kg / m^2)
EDU_1	Education level_elementary / kindergarten
EDU_2	Education level_some high school
EDU_3	Education level_high school graduated
EDU_4	Education level_some college / technical school
FVINT_2	Low intake of fruits and vegetables_at risk
HBC_1	High blood cholesterol_yes
HBP_1	High blood pressure_yes
INCOME_1	Household income_< $10,000
INCOME_2	Household income_$10,000 - $14,999
INCOME_3	Household income_$15,000 - $19,999
INCOME_4	Household income_$20,000 - $24,999
INCOME_5	Household income_$25,000 - $34,999
INCOME_6	Household income_$35,000 - $49,999

INCOME_7	Household income_$50,000 - $74,999
LTPA_1	Leisure time physical activity_yes
MACT_1	Moderate physical activity_yes
MPAREC_1	Met CDC recommendations for physical activity_yes
RACE_2	Race_non-Hispanic black
RACE_3	Race_Hispanic
RACE_4	Race_others
SEX_1	Sex_male
SMK_2	Smoking risk_at risk
VACT_1	Vigorous physical activity_yes

APPENDIX B – Interaction Terms Added into the Initial Logistic Regression Model in Alphabetical Order

X1	AGE_55-64 years group by RACE_ non-Hispanic black
X2	AGE_65+ years group by RACE_ non-Hispanic black
X3	AGE_55-64 years group by RACE_ Hispanic
X4	AGE_65+ years group by RACE_ Hispanic
X5	High blood pressure_yes by High blood cholesterol_yes
X6	Obesity by High blood pressure_yes
X7	Obesity by High blood cholesterol_yes
X8	Leisure time physical activity_yes by Alcohol drinking status_yes
X9	High blood pressure_yes by High blood cholesterol_yes by Obese
X10	Leisure time physical inactivity by Obesity

APPENDIX C – Variables Entered into the Validation Logistic Model in Alphabetical Order

ACALCR_2	Acute alcohol drinking risk_at risk
AGE_2	Age_35-44 years group
AGE_3	Age_45-54 years group
AGE_4	Age_55-64 years group
AGE_5	Age_65+ years group
BMIR_2	Overweight (BMI: 25 – 29.9 kg / m^2)
BMIR_3	Obesity (BMI > 30 kg / m^2)
CHALCR_2	Chronic alcohol risk_at risk (>/= 60 drinks in the past month)
DIET_1	Eating fewer high fat or high cholesterol foods_yes
EDU_1	Grades 1 - 8
EDU_2	Grades 9 - 11
EDU_3	Grades 12 or GED
EDU_4	College 1 - 3
EXERCISE_1	Exercise more_yes
HBC_1	High blood cholesterol_yes
HBP_1	High blood pressure_yes
INCOME_1	Household income_< $10,000
INCOME_2	Household income_$10,000 - $14,999
INCOME_3	Household income_$15,000 - $19,999
INCOME_4	Household income_$20,000 - $24,999

INCOME_5	Household income_$25,000 - $34,999
INCOME_6	Household income_$35,000 - $49,999
INCOME_7	Household income_$50,000 - $74,999
RACE_2	Race_non-Hispanic black
RACE_3	Race_Hispanic
RACE_4	Race_others
SEX_1	Sex_male
SMK_2	Smoking risk_at risk

VITA

Jie Li was born in Jiangan, Sichuan Province, China on November 21, 1962, the daughter of Changge Li and Shirong Liu. She entered The Agriculture University of Hunan in Changsha, Hunan Province, China, in July 1979 and received the degree of Bachelor of Science in Agronomy in July 1983. During the following years, she was employed as a Biology teacher with No. 26 High School and No. 28 High School in Changsha, Hunan Province, China. In August 2001, she entered the Graduate College of Texas State University, San Marcos, Texas.

Permanent address: 7607 Monona Avenue
Austin, Texas 78717

This thesis was typed by Jie Li