**Assessing Diabetes risk by Insulin Serum Concentration after 2hours, Plasma Glucose concentration, Age Group and Smoking Status**

Sandra Ovuegbe

Department of Epidemiology, UT Health Science Center School of Public Health

PH1700: Intermediate Biostatistics

Dr. Adriana Perez

November 17, 2021

Sandra Ovuegbe

PH\_1700 Fall 2021

Final Project, Fall 2021.

**ABSTRACT**

Diabetes is a serious health issue of the 21st century due to our reduced physical activity and high calorie diet. Several available research articles on diabetes point to one or more of the following risk factors: high blood pressure, race, age, smoking status etc. In this paper, we will test out the relationship between diabetes and insulin levels after 2 hours, plasma glucose concentration, age and smoking status. Specifically, our first hypothesis is whether or not there is a difference between the mean plasma glucose concentration of participants that have diabetes and those that do not have diabetes. Our second hypothesis is whether or not there is a difference in the diastolic blood pressure groups between those who have diabetes and those that do not. First, we will compute summary statistics for each key variable, test for normality and then test our first and second hypothesis. We will then determine how predictive the risk factors are for diabetes by conducting a regression test to find regression coefficients (slope and intercept). We determined in our first hypothesis that plasma glucose concentration was different by diabetes status, specifically the higher the plasma glucose concentration means an increased chance for diabetes. In our second hypothesis we concluded that the probability of having high diastolic blood pressure was different for participants who had diabetes vs those who didn’t. Lastly, we concluded that a positive linear relationship existed between explanatory variables (plasma glucose concentration, age group and high blood pressure) and outcome variable (2 hour insulin serum levels), although the predictiveness was weak. Our results and conclusions were comparable to past research, specifically the ones assessing the predictiveness of our key variables for diabetes risk.

**INTRODUCTION**

Diabetes, one of the top 10 leading causes of death in developed countries has continued to rise in frequency over the years. CDC reports that positive diagnosis for adult diabetes has doubled in the last 20 years. Onset of diabetes is generally known to be caused by a range of factors such as (but not limited to): sedentary lifestyle, obesity, smoking status, age or genetics. Typically, type 2 diabetes used to be considered an adult disease that presents at age 45 and above. However, in recent times, the frequency of diagnosis among younger adults and children are beginning to catch up. A 2018 study on age and diabetes found that diabetes diagnosed earlier in life presents worse health complications than adult diabetes (Nanayakkara et al., 2020). Another cohort study on diabetes and plasma glucose concentration at 30, 60 and 120 minutes found that plasma glucose concentration during all durations were significant predictors of type 2 diabetes, however the concentration at 60 minutes was the strongest predictor. The relationship was determined by plotting sensitivity against false positive rate. The larger the area under the concentration curve, the stronger the predictive effect (Abdul-Ghani et al., 2009). Serum insulin level at 2hours was also found to be a significant predictive factor of type 2 diabetes in a prospective cohort study of black vs white participants. Black women that had significantly higher insulin level after fasting were more likely to be diagnosed with type 2 diabetes than white women (Carnethon et al., 2002). Another study used logistic regression to determine that diabetes pedigree is a strong predictive factor for diabetes with a 78% accuracy rate. Using a pedigree function to assess diabetes risk involves utilizing a person’s family history of diabetes to predict how likely they are to be diagnosed. A high pedigree function increases the risk for diabetes (Joshi, R. D., & Dhakal, C. K., 2021). Lastly, a longitudinal cohort study on diabetes mellitus and hypertension found that participants who developed hypertension- measured by systolic and diastolic blood pressure levels were more likely to develop diabetes later on in the study. The participants who already had diabetes at baseline were at higher risk for developing hypertension. Therefore, diabetes mellitus and diastolic blood pressure are co-morbidities that track each other over time (Tsimihodimos et al., 2018). The objective of this study is to assess the onset of diabetes, relative to age group, plasma glucose concentration, diabetes pedigree function, smoking status and 2-hour insulin level. Our first hypothesis is to determine if there is a difference between the mean plasma glucose concentration of participants that have diabetes and those that do not have diabetes. Our second hypothesis is to assess whether there is a difference in the diastolic blood pressure groups between those who have diabetes and those that do not. Finally, we evaluate if there is a linear relationship between 2hour insulin levels and our key variables with a regression model and determine how strong their predictive powers are.

**METHODS**

All statistical analysis were conducted using SataBE-64 version 17 statistical software (Stata Corp, 2019). To familiarize with our data set, we conducted appropriate descriptive statistics for our key continuous and categorical variables. First, key continuous variables: plasma glucose and concentration and 2-hour serum insulin were assessed for normality using box plots, quintile-quantile plot and Shapiro wilk test. Mean and Standard deviation were determined for those that yielded a normal distribution as seen in table 1. To get these values, we used the summarize command. For those that were non-normal, we used the summarize detail command to get values for median and IQR. Frequency statistics (percentage and count) were reported for key categorical variables as shown in table 1. Next, we plotted a boxplot of 2-hour serum insulin concentration by diastolic bp groups (high and not high) to assess the relationship between them. In part 2 of our project, we conducted a two-sided 2 independent sample t test with unequal variance using a 95% confidence interval to test our first hypothesis. We used this test given that the assumptions for independence was satisfied, CLT was satisfied for plasma glucose concentration, and the assumption of equal variance was met. Background information on the data collection methods does not point to a paired or matched data, so we assumed independence for diabetes and non-diabetes participants who had their plasma glucose concentrations measured. Total sample size for the two groups was greater than 30 and there were no particular extreme outliers, so CLT was satisfied. We conducted an f-test using alpha 0.01 to check the equal variance assumption for plasma glucose concentration among diabetes and no diabetes groups. If the resulting p value was >0.05, we concluded equal variance, but if the p value was <0.05, we conclude unequal variance. Our parameter of interest was the difference in mean plasma glucose concentration between diabetes vs non diabetes participants. The null hypothesis is that there is no difference in mean plasma glucose concentration among participants diagnosed with diabetes and participants not diagnosed. The alternative hypothesis is that there is a mean difference. Hnull: mean difference=0, Halt: mean difference=/ 0. If the resulting t statistic is less than the critical value on the left tail, we reject null. Also, if the t statistic is greater than the critical value on the right tail, we reject null.

To tests our second hypothesis for the difference in diastolic blood pressure groups between those who have diabetes vs no diabetes, we constructed a 2x2 contingency table with expected values using STATA. We chose to perform a chi square test of independence large sample since all expected cells were >= 5 and there was nothing in the data that suggested that samples were paired, so independence assumption was satisfied. Our parameter of interest was the difference in proportion of diastolic bp groups between those with/ without diabetes. The distribution was Chi square X2 with degrees of freedom 1. Null hypothesis was H0: P1= P2 (The probability of having high bp and not high bp is the same for participants who have diabetes vs those who don’t have) and alternative hypothesis was Halt: P1=/P2 (Probability of having high bp and not high bp is not the same for participants who have diabetes vs those who don’t have). Rejection region was defined as when X2 > X21, 0.95 or when p value <0.05

Next, we tested the association between the dependent variable 2hour insulin levels and the independent variables: Age groups, Plasma glucose concentration, Diastolic blood pressure groups and smoking status. We performed a simple linear regression to determine the regression coefficients Slope (beta) and Intercept (alpha). The distribution used was an F distribution with degrees of freedom (1, 1498) and alpha=0.05. The regression yielded an ANOVA table output and t-test table with predicted regression coefficients. For independent categorical variables with more than 2 groups, (e.g. smoking status), we created dummy variables to represent their categories and then performed a regression on those categories. The hypothesis for our F test (ANOVA) was Hnull: (There is not a linear association between the dependent and independent variable) while H alternative: (There is a linear association between the dependent and independent variable). For our t-test, null hypothesis was that Intercept and Slope were equal to zero and the alternative was that the Intercept and Slope were different from zero.

We also performed a multiple linear regression to assess the linear association between the same predictor variables and outcome variable we used in our simple linear regression. The distribution used was an F distribution with degrees of freedom (7, 1492) and alpha=0.05. The reason for the multiple linear regression was to see how each predictor variable simultaneously affect the outcome variable when the other predictor variables are constant with. After performing our regression, we checked the assumptions for normality, independence, linearity and homoscedasticity. We plotted residual plots (unadjusted, standardized, studentized) and fitted the regression line against our outcome variable insulin2h to check assumptions. The normality assumption was not satisfied for our final model, but for the sake of this project, we assumed it was.

**RESULTS**

The box plot for plasma glucose by dm produced multiple outliers at the upper and lower whisker for the no diabetes group. Meanwhile, the diabetes group produced no outliers but was left skewed. QQ-plots were also left skewed and the resulting SWILK test gave p value less than 0.01. 2hr insulin by dm gave a right skewed box plot with multiple outliers at the upper whiskers of both no diabetes and diabetes groups. The quintile quantile plots also showed a right-skewed distribution, and the resulting Shapiro Wilk Test gave p value<0.01. Even with our skewed plots and significant SWILK tests for plasgluc and 2hr insulin, we still concluded normality because we have a sufficient sample size in both variables (n>30).

Table 1 below reports summary statistics (mean and standard deviation) for normally distributed continuous variables. We can see that the reported mean and standard deviations for our continuous variable groups are different for their diabetes vs non diabetes participants. Diabetes positive participants have higher mean 2hr serum insulin than non-diabetes participants. The same goes for plasma glucose concentration. However, we cannot conclude on the relationship from just these figures. This is why we performed a t test to assess if the means are truly different. Additionally, the sample size for non-diabetes is about twice the size of the diabetes sample, this imbalance can result in possible errors for our tests. While assessing normality for our categorical variables we found both plasma glucose concentration and 2h insulin to be non-normal with Shapiro Wilk tests yielding p values <0.01. However, even with the non-normal distributions, sample sizes for the two variables are well above 30, therefore we can assume normality for our statistical tests moving forward.

**Table 1**: Summary statistics for key variables:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Does not have diabetes (dm=0) | Has Diabetes (dm=1) | Total sample |
| Age group | 100% (1007) | 100% (493) | (100%)1500 |
| <35 years old **(% (n))** | 75.97% (765) | 52.94% (261) | (128.9%)1026 |
| >35 years old **(% (n))** | 24.03% (242) | 47.06% (232) | (71.09%) 474 |
| Diastolic bp group | 100% (1007) | 100% (493) | (100%) 1500 |
| Not high dbp **group (% (n))** | 76.56% (771) | 69.98% (345) | 146.54% (1116) |
| High dbp group **(% (n))** | 23.44% (236) | 30.02% (148) | 53.42% (384) |
| Plasma glucose concentration at 2 hours in an oral glucose tolerance test **(Mean (SD)** | 111.69 (28.11) | 139.21(32.02) | 120.74(32.16) |
| 2-Hour serum insulin (µU/ml, **Mean (SD))** | 99.14 (111.87) | 143.73 (131.58) | 113.80 (120.50) |
| Smoking status | 100% (1007) | 100% (493) | 100%(1500) |
| cigarettes**(%(n))** | 29.79% (300) | 29.01% (143) | 58.8% (443) |
| e-cigarettes**(%(n))** | 25.82% (260) | 28.80% (142) | 54.62% (402) |
| Cigars **(%(n))** | 26.42% (266) | 25.76% (127) | 52.18% (393) |
| Non-smoker **(%(n))** | 10.43% (105) | 8.72% (43) | 19.15% (148) |
| Dual tobacco products **(%(n))** | 7.55% (76) | 7.71% (38) | 15.26% (114) |
|  |  |  |  |



OUTPUT 1: Boxplot of 2-hour insulin concentration by high and not high diastolic bp groups

In our plot for 2hr insulin level by diastolic bp group above, both boxplots for high diastolic blood pressure and not high diastolic blood pressure groups appeared to be right skewed. The not high bp group had multiple outliers at the upper whisker from 400 µU/ml up to approximately 800 µU/ml. The high bp group also had multiple outliers at the upper whisker from 350 µU/ml up to approximately 700µU/ml. The median insulin conc after 2hours for participants in both not high bp and high bp groups were the same at 100 µU/ml

Results from our F-test of equal variance yielded a significant p value <0.05 with F-statistic 0.7708 and degrees of freedom 1006,492. Therefore, we proceeded with a 2 independent sample t-test of unequal variance. Our 2 independent sample t test with unequal variance, using degrees of freedom 872.6 resulted in a test statistic of -16.26 and a p value less than 0.001. The mean difference of plasma glucose between non diabetes and diabetes participants was between (-30.84, -24.19) which does not contain 0. Therefore, given our confidence interval and p value results, we reject the null hypothesis and conclude that there is a mean difference between plasma glucose concentration for diabetes vs non diabetes test groups. Due to our confidence interval being negative values when mean difference is = (no diabetes – diabetes), we can say that the diabetes group has higher plasma glucose concentration than the non-diabetes group.



OUTPUT 2: Two sample t test of unequal variance comparing mean plasma glucose concentration by diabetes vs non-diabetes group

The chi square test for independence (large sample test) for our 2nd hypothesis gave a X2 statistic of 7.5335 with p value 0.006. Since we indicated that pvalue < 0.05 is the rejection region, we reject our null hypothesis. Therefore, we conclude that there is sufficient evidence that the probability of having high bp and no high bp by is different for participants who have diabetes vs no diabetes.

**Table 2**: Regression table showing results of univariate analysis with insulin 2h as the outcome variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SBP outcome mm HG | Regression Coefficient | SE Regression Coefficient | T test | Pvalue |
| **Model 1 Univariate Analysis** |  |  |  |  |
| Intercept | 120.517 | 5.53 | 21.78 | <0.001 |
| Less than 35 **(reference group)** | -9.824 | 6.690 | -1.47 | 0.142 |
| 35 above | 9.824 | 6.690 | 1.47 | 0.142 |
| F statistic (1, 1498) = 2.16 |  |  |  |  |
| R2=0.0014 |  |  |  |  |
| **Model 2 Univariate Analysis** |  |  |  |  |
| Intercept | -80.286 | 10.926 | -7.35 | <0.05 |
| Plasma glucose concentration | 1.607503 | 0.0874511 | 18.38 | <0.05 |
| F statistic (1, 1498)=337.89 |  |  |  |  |
| R2= 0.1840 |  |  |  |  |
| **Model 3 Univariate Analysis** |  |  |  |  |
| Intercept | 115.45 | 6.151 | 18.77 | <0.001 |
| Not High diastolic bp (**reference group)** | -2.219 | 7.131 | -0.31 | 0.756 |
| high diastolic bp | 2.219 | 7.131 | 0.31 | 0.756 |
| F statistic (1, 1498)= 0.10 |  |  |  |  |
| R2=0.0001 |  |  |  |  |
| **Model 4**  **Univariate Analysis** |  |  |  |  |
| Intercept | 106.868 | 11.295 | 9.46 | <0.001 |
| Smoking Status |  |  |  |  |
| Cigarettes | 7.524 | 12.666 | 0.59 | 0.553 |
| e-cigarettes | 9.032 | 12.797 | 0.71 | 0.480 |
| Cigars | 3.073 | 12.829 | 0.24 | 0.811 |
| Non-smoker | 15.010 | 15.029 | 1.00 | 0.318 |
| Dual tobacco products **(reference group)** | -7.50 | 11.74 | -0.64 | 0.523 |
| F statistics (1, 1498)= 0.39 |  |  |  |  |
| R2= 0.0011 |  |  |  |  |

In table 2, results were reported for regression coefficients, their standard errors, t-test statistic, p value, R^2 value and F statistic. The only significant regression coefficient was for model 2, plasma glucose concentration (p value <0.05). However, even with the significant test, only 18.4% of the variation of 2hr insulin can be explained by the model. Regression coefficients for models tested can be interpreted thus:

The 2hour serum Insulin baseline for a person is 120.5 µU/ml

9.824 µU/ml is the expected difference of 2hour insulin serum levels when comparing participants above 35 to participants less than 35

If plasma glucose concentration increases by one unit, insulin serum levels will increase by 1.608 µU/ml

2.219 µU/ml is the expected difference of 2-hour serum insulin levels when comparing participants with high blood pressure to those with not high diastolic bp

7.52 µU/ml is the expected difference of 2hour insulin levels when comparing and participants who smoke cigarettes to those who use dual tobacco products

9.03 µU/ml is the expected difference of 2hour insulin levels when comparing and participants who smoke e-cigarettes to those who use dual tobacco products

3.073 µU/ml is the expected difference of 2hour insulin levels when comparing and participants who smoke cigars to those who use dual tobacco products

15.010 µU/ml is the expected difference of 2hour insulin levels when comparing and participants who do not smoke to those who use dual tobacco products

**Table 3:** Table showing Multiple linear regression results with inslin2h as the outcome variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SBP outcome | Regression coefficient | SE Regression coefficient | T test | Pvalue |
| Multivariable Analysis |  |  |  |  |
| Intercept | -81.05 | 11.76 | -6.89 | <0.001 |
| Plasma glucose | 1.68 | 0.09 | 18.63 | <0.001 |
| High Diastolic bp group | -15.77 | 6.66 | -2.37 | 0.018 |
| Low diastolic bp group **(reference group)** |  |  |  |  |
| Age Less than 35 (**reference group)** |  |  |  |  |
| 35 above | -12. 647 | 6.32 | -2.00 | 0.046 |
| Smoking status: cigarettes **(reference group)** |  |  |  |  |
| e-cigarettes | 1.78 | 7.50 | 0.24 | 0.813 |
| cigars | -1.95 | 7.54 | -0.26 | 0.796 |
| Nonsmokers | 3.73 | 10.33 | 0.36 | 0.718 |
| Dual tobacco products | -5.69 | 11.42 | -0.50 | 0.618 |
| F statistic (7, 1492)= 50.25 |  |  |  | <0.001 |
| R2= 0.1908 |  |  |  |  |

Our ANOVA test was highly significant with p value <0.0001, leading us to reject our null and conclude that a linear relationship exists between our predictor variables and outcome variable (insulin2h). This model, however, can explain only 19% of the variation of 2hour serum insulin. This is very weak. In our t-tests, we found exactly three significant predictors in our final model: Plasma glucose, Diastolic blood pressure and Age group with p values <0.05. For age groups, however, the p value was was borderline (0.046) which could represent a potential issue with our power. Our linear equation will resemble:

Insulin2h= -81.05 + (1.68plasgluc) + (-15.77dbpgroup) + (-12.647\*agegroup) + [1.78\*e-cigarettes + -1.95\*cigars + 3.73\*nonsmokers + -5.69\*dual tobacco products]

The regression coefficients can be interpreted thus:

The 2hour serum Insulin baseline for a person is -81.05

If plasma glucose concentration increases by one unit, insulin serum levels will increase by 1.68 µU/ml while agegroup, diastolic bp and smoking status remain constant

-15.77 µU/ml is the expected difference of 2-hour serum insulin levels when comparing participants with high blood pressure to those with not high diastolic bp

-12.647 µU/ml is the expected difference of 2hour insulin serum levels when comparing participants above 35 to participants less than 35.

1.78 µU/ml is the expected difference of 2hour insulin levels when comparing and participants who smoke e-cigarettes to those who smoke cigarettes

-1.95 µU/ml is the expected difference of 2hour insulin levels when comparing and participants who smoke cigars to those who smoke cigarettes

3.73 µU/ml is the expected difference of 2hour insulin levels when comparing and participants who do not smoke to those who smoke cigarettes

-5.69 µU/ml is the expected difference of 2hour insulin levels when comparing and participants who smoke dual tobacco products to those who smoke cigarettes

**DISCUSSION**

We concluded in our first hypothesis test that mean plasma glucose concentration is greater among participants diagnosed with diabetes mellitus than those not diagnosed. This is consistent with findings in Abdul-Ghani et al 2019, in which plasma glucose concentration at 30,60, and 120 minutes were greater in diabetes mellitus participants than non-diabetes participants. The only difference is that Abdul-Ghani et al found the 60-minute concentration to be the strongest predictor with cutoff point 155mg/dl, where >155 means higher risk for diabetes. However, for our tests, we used the concentration at 120mins to test its predictiveness, using cut off point >=200 to diagnose diabetes.

In our second hypothesis test, we concluded that there is sufficient evidence that the probability of high blood pressure and no high blood pressure is different for participants who have diabetes vs those who do not. This is in tandem with our literature review on Tsimihodimos et al in which participants from the study where more likely to develop diabetes if they developed high blood pressure (computed as high diastolic and systolic blood pressure) prior.

We determined a weak positive linear relationship between plasma glucose concentration and insulin 2-hour serum insulin in our univariate analysis. Given that plasma glucose concentration was the defining factor for a positive diabetes status, a positive linear relationship between plasma glucose and 2-hour serum insulin suggests that increasing diabetes risk correlates with increasing 2-hour serum insulin levels. Although we determined this relationship was weak, Abdul-Ghani et al in his study deemed plasma glucose level at 1hour and 2hours to be strong predictors of diabetes risk, while and Carnethon et al found a similar positive relationship for 2 hours insulin serum levels and diabetes risk.

Again, in our multiple linear regression we found a weak linear relationship between plasma glucose concentration and 2hour insulin when age group, diastolic hbp group and smoking status was controlled for. There were also other significant predictors in our results: older than 35 age group and high diastolic blood pressure. The significant predictiveness for high blood pressure partly confirms the relationship we observed between high blood pressure and diabetes status in our second hypothesis which was consistent with study observations in Tsimihodimos et al. On the predictiveness of age group, we found p value to be borderline (almost close to 0.50). Comparing this to Nanayakkara et al. which suggests that the risk of diabetes for the younger and older demographic are almost levelling off, we observed something similar given the p value was borderline. Therefore, although type 2 diabetes generally affects the older population, the disease risk has become more spread out among the younger populations.

References

Nanayakkara, N., Curtis, A. J., Heritier, S., Gadowski, A. M., Pavkov, M. E., Kenealy, T., Owens, D. R.,

Thomas, R. L., Song, S., Wong, J., Chan, J. C., Luk, A. O., Penno, G., Ji, L., Mohan, V., Amutha, A., Romero-Aroca, P., Gasevic, D., Magliano, D. J., Teede, H. J., … Zoungas, S. (2021). Impact of age at type 2 diabetes mellitus diagnosis on mortality and vascular complications: systematic review and meta-analyses. *Diabetologia*, *64*(2), 275–287. https://doi.org/10.1007/s00125-020-05319-w

Abdul-Ghani, M. A., Lyssenko, V., Tuomi, T., DeFronzo, R. A., & Groop, L. (2009). Fasting versus

postload plasma glucose concentration and the risk for future type 2 diabetes: results from the Botnia Study. *Diabetes care*, *32*(2), 281–286. [https://doi.org/10.2337/dc08-1264](about:blank)

Carnethon, M. R., Palaniappan, L. P., Burchfiel, C. M., Brancati, F. L., & Fortmann, S. P. (2002, August 1). *Serum insulin, obesity, and the incidence of type 2 diabetes in black and white adults*. Diabetes Care. Retrieved October 19, 2021, from https://care.diabetesjournals.org/content/25/8/1358.

Joshi, R. D., & Dhakal, C. K. (2021). Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. *International journal of environmental research and public health*, *18*(14), 7346. [https://doi.org/10.3390/ijerph18147346](about:blank)

Tsimihodimos, V., Ferrannin, E., Meigs, J. B., & Gonzalez-Villalpando, C. (2018, March 3). *JavaScript disabled*. Hypertension and Diabetes Mellitus. Retrieved October 19, 2021, from https://www.ahajournals.org/doi/epub/10.1161/HYPERTENSIONAHA.117.10546.

StataCorp. 2019. *Stata Statistical Software: Release 17*. College Station, TC: StataCorp LLC.