The Effect of Diabetes and Biological Markers on Blood Pressure Readings through both a Predictive and Traditional Statistical Lens

Team 6: Ente Kang, Jingwen Ji

Link to Github: https://github.com/entekang/6-FinalProject

Abstract

The objective of our research was to investigate the effects of Diabetes and other biological markers on 1) Hypertension (HTN) and 2) systolic blood pressure readings (sBP). The dataset used for the analysis contained 10,000 observations with 8,602 unique patients. It was found that individuals with diabetes had a higher odds of developing hypertension than those without, and an increase in age was associated with increasing sBP. From a prediction perspective, our Logistic Regression model was able to achieve a 85% recall rate on the test set. Our results indicate the importance of biological markers and medical conditions on individuals' blood pressure readings, and provide a pathway for how future researchers can implement predictive models while understanding the contributions of their predictors.

Introduction

Diabetes is a chronic condition that affects millions of people, and the number of adults diagnosed with diabetes has more than doubled over the past 20 years [1]. Diabetes also has many complications such as increased risk of developing cardiovascular diseases. A study done by the Johns Hopkins Medicine indicated Patients with diabetes were twice as likely to have high blood pressure (hypertension) [2]. As a matter of fact, in 2021, hypertension was a primary or contributing cause of 691,095 deaths in the United States [3]. From a numbers perspective, hypertension is defined as having a systolic blood pressure > 130-139 and a diastolic blood pressure > 80-89. Research from the American Heart Association suggests that one's systolic blood pressure readings are best for predicting future cardiovascular diseases and death [4].

These facts motivated us to address the following questions. What are the effects of Diabetes and other biological markers on 1) Hypertension (HTN) status and 2) systolic

blood pressure readings (sBP)?. We will be examining this through a predictive and traditional statistical inference lens. Doing so, we hope that our results can shine light on some contributing factors of cardiovascular diseases and pave way for more accurate prediction models.

Data

Our dataset was obtained from CPCSSN (www.cpcssn.ca) under research project 2015SRSC51. There were a total of 10,000 observations, with 8,602 unique patients, and a study end date of June 30, 2015. Patient level information includes their age at examination, sex, blood pressure readings etc. Individuals in our data had ranging from 1 - 8 repeated measurements of their hypertension status and systolic blood pressure readings. It is worth noting that one's diabetes (binary 0,1) or hypertension status (binary 0,1) did not change over time within this dataset. Looking at unique patients, 4,460 did not have diabetes while 4,142 did, which is fairly balanced. On the other hand, 3,291 did not have hypertension while 5,311 did.

Our processing of the data consisted of two parts, one for each research question. In predicting and modeling hypertension status, we chose to not use a patient's repeated measurements as their hypertension status did not change, so incorporating a longitudinal response measure would not be beneficial in terms of prediction accuracy. Instead, we used the most recent measurement for each patient, resulting in a dataset of exactly 8,602 individuals. Predictor variables were chosen based on clinical associations with hypertension status, which resulted in the selection of age at examination, BMI, sex, systolic blood pressure and diabetes status. For the purpose of modeling, sex was converted to a binary (0=male, 1=female) variable. A quick check was conducted to see that these predictors were not highly correlated with each other, which would cause complications in model fitting and assessments. Only systolic blood pressure (sBP) had 4 missing values, which were then imputed using the median, as it is more robust against outlying observations in comparison to the mean imputation strategy.

As for modeling and predicting one's longitudinal systolic blood pressure (sBP) readings, predictors were selected in a similar fashion as above, which resulted in

diabetes, age, sex and BMI being selected. The four missing observations for sBP were dropped from the analysis.

Methods

Since our goal is to be able to make predictions and understand the contributions of our predictor variables, Logistic Regression was our choice for predicting hypertension status. We acknowledge the fact that there are more advanced methods such as Decision Trees, Support Vector Machines and Neural Networks which are better at handling more complex relationships and may provide better predictive accuracy. But only Logistic Regression would allow us to see how changes in our predictors variables affect the odds of developing hypertension.

Our logistic regression was defined as follows:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 Diabetes + \beta_2 sex + \beta_3 Age + \beta_4 BMI + \beta_5 sBP \quad (1)$$

Where p = the probability of having hypertension, diabetes (0=no, 1=yes), sex (0=male, 1=female), age (at examination, in years), BMI (numerical), sBP (numerical).

Assumptions of logistic regression include 1) Linearity of the logit function with respect to the explanatory variables, 2) independent response variable measurements, 3) predictors do not have multicollinearity and 4) no extreme outliers.

In modeling and predicting longitudinal measurements of systolic blood pressure, we adopted a Linear Mixed Effects model (LMM) approach, which takes the following form.

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i \quad (2)$$

Where Y_i = systolic blood pressure readings for individual i (i=1, …, N) with n_i being the number of repeated measurements for person i. X_i is a n_i x p covariate matrix containing information on our p predictors. β is a p x 1 vector of our fixed effects estimates, Z_i is the design matrix of random effects of dimension n_i x q, where $q \leq p$.

b_i is a vector of random effect estimates of size n_i x 1 and ϵ_i is a n_i x 1 vector of residual errors.

Assumptions of the model are that 1) b_i are independent of X_i with $b_i \sim MVN(0, G)$, where G = cov(b_i) and 2) ϵ_i are independent of b_i, with $\epsilon_i \sim MVN(0, \Sigma)$, and Σ = cov(ϵ_i).

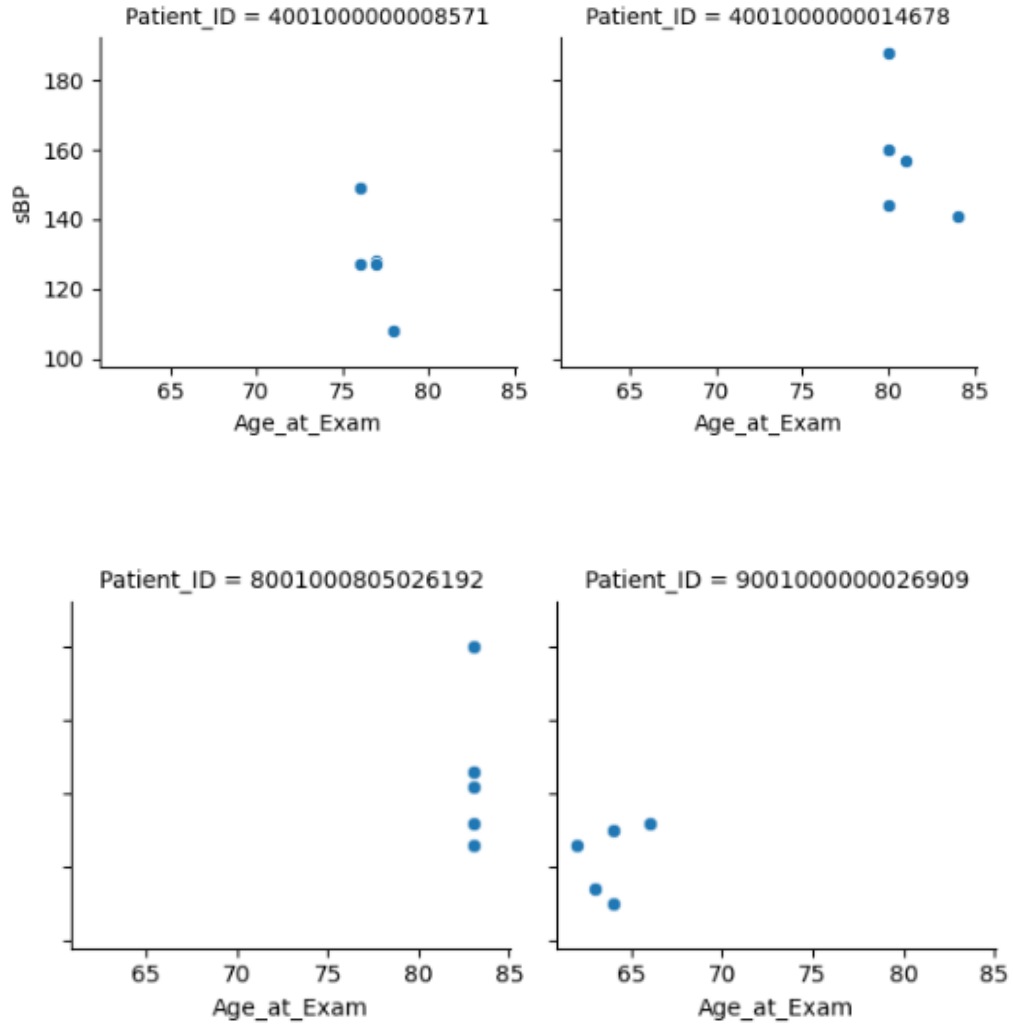
The LMM allows for us to account for individual level heterogeneity by introducing random effects in terms of a random intercept and (or) random slope. We can imagine that individuals will all have different baseline blood pressure readings, hence the need for a random intercept. At the same time, the effect of aging on systolic blood pressure may be different for everybody (***figure 1***), where we can then introduce a random slope for age. In other words, there will be a marginal or population-averaged mean of systolic blood pressure, and we allow for individuals to fluctuate from this mean through the adoption of random intercepts and slopes. More formally,

$$E(Y_i) = X_i\beta \quad (3)$$

$$E(Y_i|b_i) = X_i\beta + Z_i b_i \quad (4)$$

Where $X\beta$ is the marginal or population-averaged mean response, and we allow it to vary by $Z_i b_i$ for each individual i. Another benefit of LMMs is that we can use it to model longitudinal measurements, where the number of repeated measurements may differ for each person. From a pure prediction perspective, we acknowledge that the LSTM (Recurrent Neural Network) may perform better, as here we are imposing an assumption of linearity between response and explanatory variables with the LMM.

Figure 1: Age vs sBP for a random sample of patients with 5 repeated measurements



Results

For Logistic Regression, train-test splits were used with a testing proportion of 0.3. Table 1 shows the model fit on the training set.

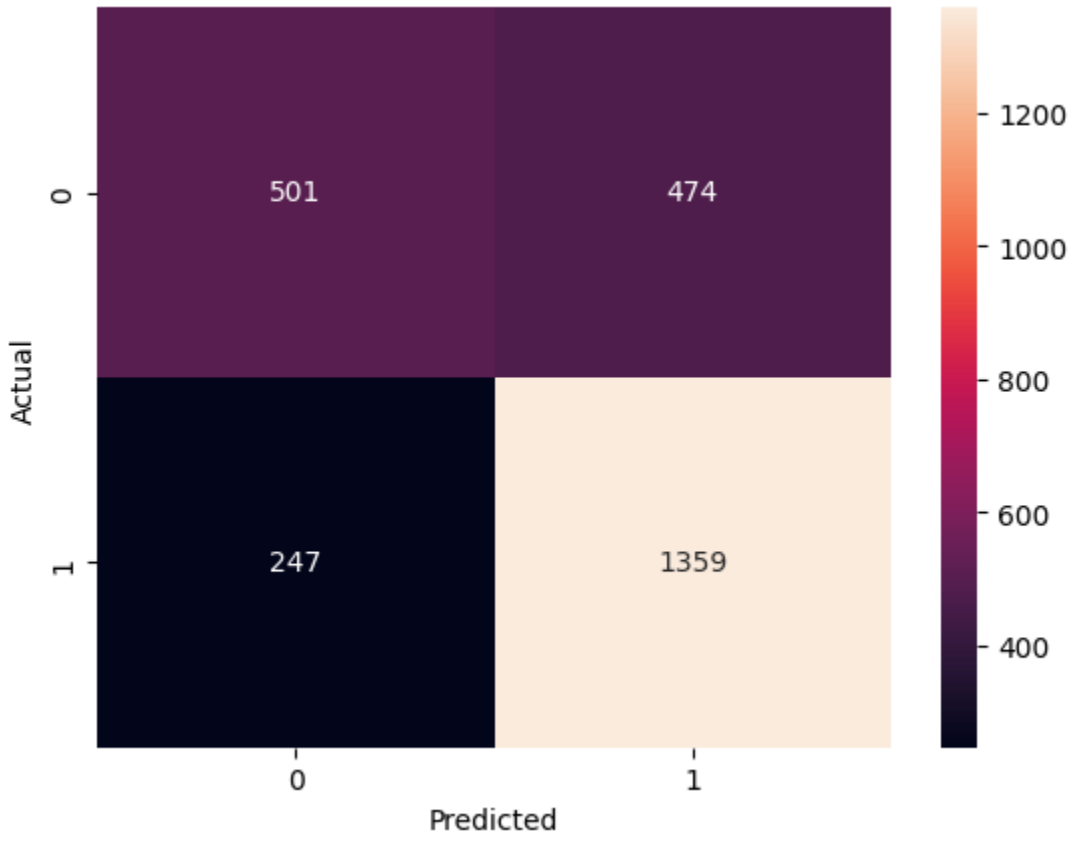
Table 1: Logistic Regression model fit

	coef	std err	z	P>|z|	[0.025	0.975]
const	-8.5903	0.329	-26.102	0.000	-9.235	-7.945
Age_at_Exam	0.0506	0.002	20.457	0.000	0.046	0.056
BMI	0.0419	0.005	8.906	0.000	0.033	0.051
sBP	0.0347	0.002	17.407	0.000	0.031	0.039
Diabetes	0.4831	0.061	7.945	0.000	0.364	0.602
Sex_Male	-0.0812	0.059	-1.372	0.170	-0.197	0.035

We see that age at examination, bmi, sBP, and diabetes have statistically significant effects at the 0.05 level, however, with a negligible sex effect. In particular, those with diabetes had exp(0.4831) = 1.62 times the odds of those without diabetes of having hypertension.

Figure 2 below details the confusion matrix based on the predictions from the test set. Our model was able to achieve a recall score of 85%. To guard against overfitting, we also looked at the performance under the training set and verified that it performed similarly under the test set. As a note, common indicators of overfitting would be having a training accuracy that was much higher than the test accuracy, which is not the case here. We were specifically interested in the recall metric because it can be very costly to say that someone does not have hypertension when they in fact do. We saw that out of all individuals with hypertension, 85% were correctly predicted, which is fairly good.

Figure 2: Confusion matrix for Logistic Regression on the test set



For our LMM, train-test splits were done using a 0.2 testing proportion, however, one big difference is that the splitting was done by patient ids and not systolic blood pressure. The justification is that we wanted all repeated observations of an individual to be in either the training or test sets. If we had a patient whose observations were in both sets, then the test set would no longer be considered as unseen data, which would bias our testing performance.

Table 2 shows the results of the LMM fit on the training data. We saw that a one unit increase in the age at examination was associated with a 0.25 increase in the patient's systolic blood pressure, and is significant at the 0.05 level. This finding highlights the

importance of age as a marker for cardiovascular measurements. A similar finding was seen for the effect of one's bmi on blood pressure readings.

Table 2: LMM model fit

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	104.372	1.391	75.011	0.000	101.645	107.099
Sex[T.Male]	0.763	0.390	1.956	0.051	-0.002	1.528
Age_at_Exam	0.249	0.015	16.587	0.000	0.220	0.279
Diabetes	0.271	0.403	0.673	0.501	-0.519	1.062
BMI	0.315	0.029	10.899	0.000	0.259	0.372
Group Var	108.878	0.835				

In terms of prediction accuracy for the test set, we assessed the mean-squared error, which is the squared deviation between the true value and predicted value for all individuals in the test set. We found a mse of 280, which is quite high, but expected due to the high levels of individual heterogeneity.

Discussion

With an increasing repertoire of available methods for high-performance prediction, it is just as essential to understand the effects of the predictors that you are using for prediction, which speaks of the generalizability.
In our model for predicting hypertension - which is a status that did not change for individuals in the dataset - we took the patients' most recent observation and modeled it using logistic regression. The recall score was 85% and we were able to see that having diabetes, a high bmi or systolic blood pressure, or an older age was associated with an increase in the odds of developing hypertension. We believe that this information will be

invaluable for doctors, as many of these variables are modifiable. In other words, steps can be taken to reduce ones' BMI or blood pressure.

In predicting and evaluating systolic blood pressure - which is believed to be the most important factor for determining future cardiovascular events or death by the American Heart Association - we used a Linear Mixed Effects Model (LMM). It was observed that increasing age or bmi was associated with higher systolic blood pressure. From a prediction perspective, the performance was not something to be overly excited about as the mse was 280.

With the LMM, our train-test split was done at the patient-level, meaning all the observations for a specific patient belonged exclusively to the train or test sets, and not both. This presents a limitation, as we are essentially predicting using the population-averaged or marginal effects, and not fully utilizing the power of random effects. Recall that the random effects are what allows each and every individual to differ from the population average. Since we are predicting using unseen patients, there are no estimates for the random effects as these belong to the patients in the training data. So it makes no sense to incorporate another person's random effects into someone else's predictions. A potential work-around will be to have patients earlier blood pressure readings in the training set, and more recent readings to be in the test set. This way we can incorporate each person's random effects into the predictions. However, this would no longer count as an unseen test set.

Despite the limitation, our research identified risk factors associated with having hypertension or high systolic blood pressure, while being able to make predictions. Models that allow one to see the effects of predictors in addition to being able to provide predictions allow for good generalizability, as future users are able to see the effects of contributing risk factors.

Last but not least, when going through past related work, we had discovered that not many authors included diversity, equity, equality concepts into their work. Most have focused on obtaining high prediction accuracies or utilizing fancy models. This is an area that more researchers should be aware of, since if one is not using data that represents the diversity in the population, the results - regardless how good - will lead to health inequities.

Individual Contributions

Ente Kang: Data processing, modeling, report writing

Jingwen Ji: Literature review

References

1. https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=With%20diabetes%2C%20your%20body%20either,releases%20it%20into%20your%20bloodstream.
2. https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure#:~:text=High%20blood%20pressure%20is%20twice,to%20heart%20disease%20and%20stroke.
3. https://www.cdc.gov/bloodpressure/facts.htm
4. https://www.heart.org/en/news/2021/03/01/which-blood-pressure-number-matters-most-the-answer-might-depend-on-your-age
5. Ambrish G. et al. Logistic Regression Technique for Prediction of Cardiovascular Disease. *Global Transitions Proceedings* 2022;127-130
6. Fitzmaurice G et al. Applied Longitudinal Analysis. Wiley 2004
7. Seabold, Skipper, and Josef Perktold. “statsmodels: Econometric and statistical modeling with python.” *Proceedings of the 9th Python in Science Conference.* 2010
8. Agresti A. Categorical Data Analysis. *Wiley* 2013
9. Nematollahi M. et al. Body composition predicts hypertension using machine learning methods: a cohort study. *Scientific Reports* 2023;13
10. Silva G. Machine Learning for Hypertension Prediction: a Systematic Review. *Springer* 1999
11. Chowdhury M. et al. A comparison of machine learning algorithms and traditional regression-based statistical modeling for predicting hypertension incidence in a Canadian population. *Nature* 2011
12. Chatrati S. et al. Smart home health monitoring system for predicting type 2 diabetes and hypertension. *Journal of King Saud University* 2022;34;862-870
13. Nath T. et al. DXA measured body composition predicts blood pressure using machine learning methods. *Journal of Clinical Hypertensio*n 2020;22;1098-1100
14. Hung M.H. et al. Prediction of Masked Hypertension and Masked Uncontrolled Hypertension Using Machine Learning. *Frontiers in cardiovascular medicine* 2021;8

15. Yewen S. et al. Prediction model of obstructive sleep apnea-related hypertension: Machine learning-based development and interpretation study. *Frontiers in cardiovascular medicine* 2022;9
16. Islam S. et al. Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries. *Frontiers in cardiovascular medicine* 2022;9
17. Huanhuan Z. et al. Predicting the Risk of Hypertension Based on Several Easy-to-Collect Risk Factors: A Machine Learning Method. *Frontiers in public health* 2021;9
18. Hsieh C. et al. Machine learning of home blood pressure to predict short-term and long-term cardiovascular outcomes. *Blood pressure monitoring* 2022;27
19. Bou-Matar R. et al. Machine learning models to predict post-dialysis blood pressure in children and young adults on maintenance hemodialysis. *Nature* 2011
20. Lacson R. et al. Use of machine-learning algorithms to determine features of systolic blood pressure variability that predict poor outcomes in hypertensive patients. *Clinical Kidney Journal* 2019;12;206-212
21. Zhang B. et al. Predicting blood pressure from physiological index data using the SVR algorithm. *BMC Bioinformatics* 2019;20;109