The Impacts of Physiological and Socioeconomic Parameters on the Likelihood of Heart Disease Using a Statistical Model

COLLEGE OF
OSTEOPATHIC MEDICINE
at the Cherokee Nation

Alexander Eddy, M.S. & Micah Hartwell, Ph.D.

BACKGROUND

Heart disease has many predisposing factors. Genetics, lifestyle, socio-economic status have all been shown to play a role.[1,2] The National Health and Nutrition Examination Survey (NHANES) combines data from interviews and physical examinations from approximately 5000 people each year in the United States. It is an excellent source for acquiring nationally representative data on known cardiovascular risk factors. By its nature, survey data, such as from NHANES, frequently has missing entries.

Multiple imputation with chained equations (mice)[3] is a robust statistical method available in the R programming language that is designed to handle missingness. It involves an iterative approach in which missing entries in one variable are predicted by non-missing entries in other variables. The algorithm uses a Bayesian model that considers uncertainty about the missing data and produces several datasets, each with different possible values for the missing data. Each dataset is analyzed individually and then re-combined to form a complete dataset.[4]

METHODS

We used the R statistical programming language to download and process anonymized NHANES data from the 2017-2018 data acquisition cycle. Several parameters known to have a bearing on cardiac health were analyzed. Multiple imputation with chained equations was implemented by the mice package[3] in R to handle missingness in the data (N=9254) by generating five possible datasets (total N=46270) for each variable. Logistic regression was carried out as follows:

Dependent Variable:

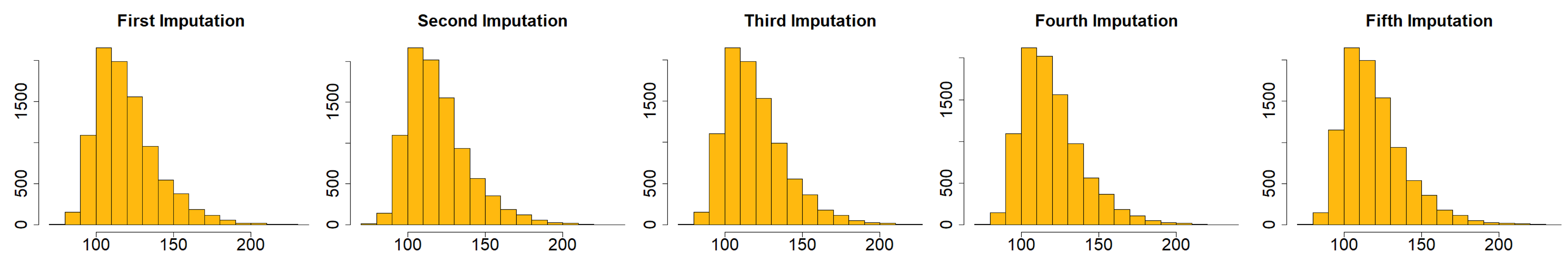
- Presence of Heart Disease (0 if no; 1 if yes. Defined by diagnosis of congestive heart failure, coronary artery disease, angina, and/or heart attack.)

Independent Variables:

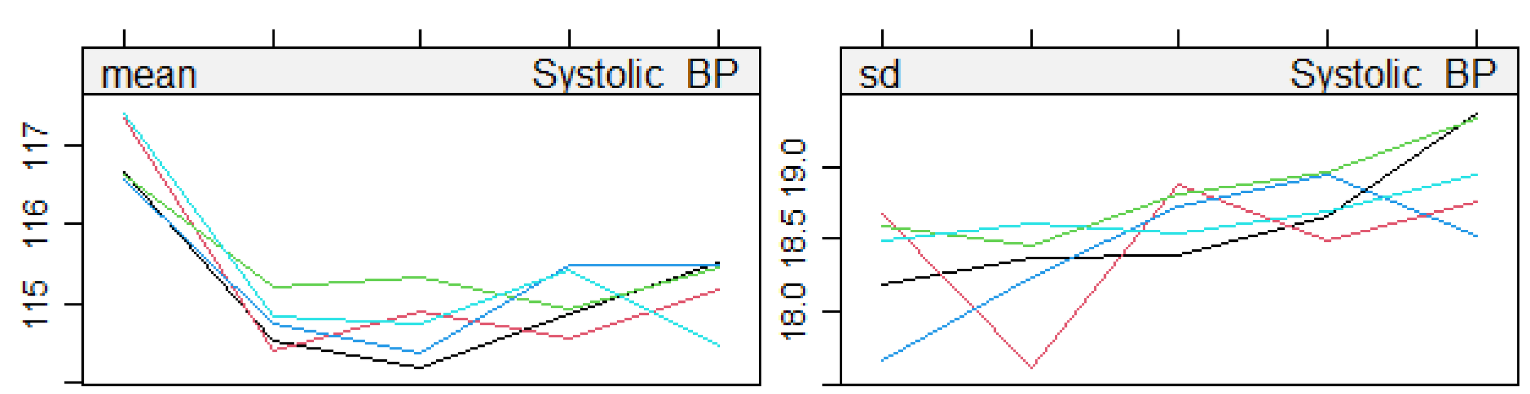
- Systolic BP (mmHg)
- Diastolic BP (mmHg)
- BMI (Body Mass Index, kg/m^2)
- CRP (High sensitivity C-reactive Protein, mg/L)
- HDL (High-Density Lipoprotein) Cholesterol (mg/dL)
- LDL (Low-Density Lipoprotein) Cholesterol (mg/dL)
- Triglycerides (mg/dL)
- Hgb A1c (Percent Glycosylated Hemoglobin)
- Monthly Income ($)
- Family Size (number of family members in residence)
- Weekend Sleep Hrs (average nightly hours slept)
- Weekday Sleep Hrs (average nightly hours slept)
- Relative Had MI (1st degree relative with myocardial infarction, 0 if no; 1 if yes)

RESULTS

Histograms of Systolic BP (x-axis) Frequency (y-axis) in each of the Five Imputations



Iterations (x-axis) of MICE Algorithm Showing Means (left) and Standard Deviations (right) of each of the Five Imputed Datasets for Systolic BP



Post-Imputation Mean Variables for Those with and without Heart Disease			
	With (N=3952)	Without (N=42318)	Combined (N=46270)
Systolic BP	126.73	119.28	119.92
Diastolic BP	62.43	67.40	66.97
BMI	28.09	26.38	26.52
CRP	4.57	3.20	3.32
HDL Cholesterol	50.83	54.42	54.11
LDL Cholesterol	87.18	106.21	104.58
Triglycerides	106.20	100.19	100.70
Hgb A1C	6.15	5.64	5.68
Monthly Income	2910.78	3599.82	3540.95
Family Size	2.93	3.65	3.59
Weekend Sleep Hrs	8.22	8.52	8.50
Weekday Sleep Hrs	7.85	7.72	7.73
Relative Had MI	0.28	0.10	0.12

Variables with higher values in the heart disease group are orange. Those with lower values are blue.

Logistic Regression Formalism

$$p = \frac{1}{1+e^{-t}}, where\ t = \sum_{i=1}^{m} \beta_i x_i$$

p is probability of event, β represents each coefficient, x represents each predictor variable

Logistic Regression Results			
Term	Estimate	Std. Error	p-value
(Intercept)	-8.72E-01	0.50839053	8.92E-02
Family Size	-2.09E-01	0.0409547	2.66E-04
Weekend Sleep Hrs	-1.14E-01	0.03651219	4.35E-03
Diastolic BP	-2.47E-02	0.00290187	1.64E-10
LDL Cholesterol	-1.96E-02	0.00177253	2.80E-11
HDL Cholesterol	-1.87E-02	0.00437651	4.15E-04
Triglycerides	-2.14E-04	0.00049451	6.66E-01
Monthly Income	-4.45E-05	2.828E-05	1.45E-01
Body Mass Index	2.28E-03	0.00749918	7.64E-01
C reactive Protein	6.78E-03	0.00581992	2.57E-01
Systolic BP	2.09E-02	0.00248793	1.28E-11
Weekday Sleep Hrs	7.42E-02	0.04328378	1.09E-01
Hgb A1C	2.32E-01	0.03734322	1.37E-08
Relative Had MI	1.11E+00	0.11748119	3.93E-11

Variables negatively predictive of heart disease are in blue. Variables positively predictive of heart disease are in orange. Those not statistically significantly related (p>0.05) are in gray. Degree of positive impact is ordered from lowest (top) to highest (bottom).

CONCLUSION

Greater family size, sleeping for a longer duration on the weekends, greater diastolic BP, greater LDL cholesterol, and greater HDL cholesterol are negatively predictive of the presence of heart disease. A negative predictive relationship between LDL and heart disease as well as between diastolic BP and heart disease may seem counterintuitive, but their mean values are lower in the group with known heart disease. This could be partially explained by a significant portion of this group adhering to a strict medication regimen of anti-hypertensives and statins.

Greater systolic BP, greater HgbA1c, and having a first degree relative with a history of myocardial infarction are positive predictors for the presence of heart disease.

Of all predictors, family size is the strongest negative predictor for heart disease. It is plausible that individuals with larger families receive greater social support, fostering a healthier lifestyle. In contrast, having a first degree relative with a history of myocardial infarction is the strongest positive predictor for heart disease, likely due to shared genetic and environmental factors.

REFERENCES

1. Menotti A, Puddu PE, Lanti M, Maiani G, Fidanza F. Cardiovascular risk factors predict survival in middle-aged men during 50 years. *Eur J Intern Med*. 2013;24(1):67-74. doi:10.1016/j.ejim.2012.08.004
2. Sesso, HD, Stampfer, MJ, Rosner, B, Hennekens, CH, Gaziano, JM, Manson, JE, Glynn, RJ. Systolic and Diastolic Blood Pressure, Pulse Pressure, and Mean Arterial Pressure as Predictors of Cardiovascular Disease Risk in Men. *Hypertension*. 2000;36(5):801-807. https://doi.org/10.1161/01.HYP.36.5.801
3. van Buuren, S, Groothuis-Oudshoorn, K mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 2011;45(3):1-67. https://doi.org/10.18637/jss.v045.i03
4. Rubin, DB, Schenker, N. Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. *Journal of Official Statistics*. 1987;3(4):375-387.