Predicting Heart Disease Using Diagnostic Tests

David Nemirovsky (dn2501)

Columbia University Mailman School of Public Health
Department of Biostatistics

ABSTRACT

Heart disease (HD) is the leading cause of death in the world, taking approximately 18 million lives each year.[1] Understanding the risk factors of HD can not only help in preventing deaths globally, but also screening for such risk factors may help improve health of at-risk individuals in other ways.[2] Using data obtained from two U.S. hospitals and two European hospitals, this study attempted to predict heart disease by assessing different variables using diagnostic testing. It was found that the primary outcome, presence of heart disease, was found to be significantly impacted by sex, type of chest pain, exercise-induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, presence of the blood disorder, thalassemia, number of major vessels colored by fluoroscopy, and which hospital the patient stayed in. Presence of heart disease was assessed by number of narrowing blood vessels, and then dichotomized in the final model. The final model, using the above covariates, demonstrated excellent ability to predict heart disease, with an area under the curve (AUC) of 0.94.

OBJECTIVES

The purpose of this study was to identify the risk factors most associated with developing heart disease, and then use those to create a predictive model. The desired model will demonstrate excellent predictive ability, goodness-of-fit, and interpretability, as the study aim is to better screen for HD using the given covariates.

METHODS

Data cleaning, exploratory data analysis (EDA), hypothesis testing, and modeling were all done using SAS® Studio Version 3.8. Before EDA, two new variables were created ('country' and 'hosp') to account for possible differences between U.S. and Europe hospital treatments, along with variability between hospitals within countries. An additional variable, 'nar_vess', was created to dichotomize the 'diag' variable, which assesses presence of heart disease based on number of majoring blood vessels with > 50% diameter narrowing. If 'diag' had a value of 0 (meaning no heart disease present), then 'nar_vess' would also have a value of 0 (also meaning no heart disease present). However, if 'diag' had a value of 1-3, then 'nar_vess' was given a value of 1 (meaning heart disease is present).

During EDA, all categorical and continuous covariates were assessed for their differences in the primary outcome (number of majoring blood vessels with > 50% diameter narrowing). Bar charts and cross-tab frequency tables were produced for categorical variables and box plots with summary statistics (mean, median, standard deviation, min, max) were produced for continuous variables to assess difference in distribution of number of majoring blood vessels with > 50% diameter narrowing.

For hypothesis testing, chi-squared tests were performed to assess differences in the primary outcome among categorical covariates, with all assumptions being met for chi-squared testing (less than 20% of expected cells had values of 5 or less). ANOVA tests were performed to assess independence of the outcome (with no heart disease as the reference) among continuous covariates, however, normality and equal variance assumptions were not met for three covariates ('trestbps', 'chol', and 'oldpeak'), so non-parametric tests were performed to assess their independence.

EDA and hypothesis testing showed that most of the variables could be useful for the model. However, stepwise model selection was performed for both outcomes ('diag' and 'nar_vess'), and both final models included the same covariates: 'cp', 'ca', 'thal', 'hosp', 'exang', 'sex', 'oldpeak', and 'slope'.

An ordinal logistic regression model was fit on the 'diag' outcome, however, it did not pass the proportional odds test, so the final model obtained for that outcome was a polytomous logistic regression model.

A logistic regression model was fit on the 'nar_vess' outcome. A receiver operating characteristic (ROC) curve was then generated using this final model. Model predictability was assessed using the AUC of the ROC curve and model fit was assessed using the Hosmer-Lemeshow test for goodness-of-fit. All tests were conducted using a significance level of 5%.

RESULTS

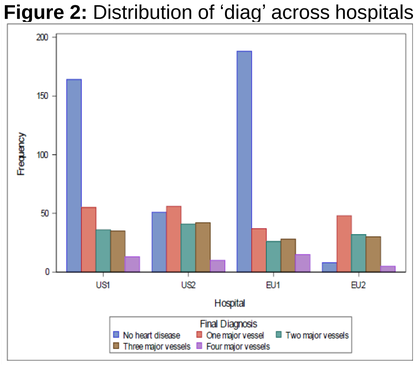
Figure 1: Descriptive statistics for number of majoring blood vessels with > 50% diameter narrowing across covariates used in final model

	Final Diagnosis									
	No heart disease		One major vessel		Two major vessels		Three major vessels		Four major vessels	
	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent
Sex										
Female	144	15.65	23	2.50	11	1.20	11	1.20	5	0.54
Male	267	29.02	173	18.80	124	13.48	124	13.48	38	4.13
Chest Pain Type										
Typical angina	26	2.83	9	0.98	5	0.54	5	0.54	1	0.11
Atypical angina	150	16.30	14	1.52	2	0.22	7	0.76	1	0.11
Non-anginal pain	131	14.24	31	3.37	16	1.74	21	2.28	5	0.54
Asymptomatic	104	11.30	142	15.43	112	12.17	102	11.09	36	3.91
Exercise Induced Angina										
No	350	38.04	94	10.22	53	5.76	42	4.57	13	1.41
Yes	61	6.63	102	11.09	82	8.91	93	10.11	30	3.26
Slope of the Peak Exercise ST Segment										
Upsloping	224	24.35	57	6.20	26	2.83	23	2.50	3	0.33
Flat	159	17.28	118	12.83	87	9.46	90	9.78	30	3.26
Downsloping	28	3.04	21	2.28	22	2.39	22	2.39	10	1.09
Number of Major Vessels (0-3) Colored by Flourosopy										
0	327	35.54	98	10.65	27	2.93	37	4.02	11	1.20
1	54	5.87	63	6.85	50	5.43	36	3.91	8	0.87
2	21	2.28	22	2.39	35	3.80	47	5.11	7	0.76
3	9	0.98	13	1.41	23	2.50	15	1.63	17	1.85
Thalassemia Blood Disorder										
Normal	314	34.13	57	6.20	31	3.37	27	2.93	6	0.65
Fixed defect	21	2.28	17	1.85	27	2.93	11	1.20	9	0.98
Reversable defect	76	8.26	122	13.26	77	8.37	97	10.54	28	3.04
Hospital										
US1	164	17.83	55	5.98	36	3.91	35	3.80	13	1.41
US2	51	5.54	56	6.09	41	4.46	42	4.57	10	1.09
EU1	188	20.43	37	4.02	26	2.83	28	3.04	15	1.63
EU2	8	0.87	48	5.22	32	3.48	30	3.26	5	0.54

Analysis Variable : oldpeak ST Depression Induced by Exercise Relative to Rest

Final Diagnosis	N Obs	Mean	Median	Std Dev	Minimum	Maximum	N Miss
No heart disease	411	0.42	0.00	0.72	-1.10	4.20	0
One major vessel	196	0.82	0.70	0.98	-2.60	3.60	0
Two major vessels	135	1.30	1.50	1.14	-2.00	4.00	0
Three major vessels	135	1.54	1.50	1.22	0.00	6.20	0
Four major vessels	43	2.20	2.50	1.30	0.00	5.00	0

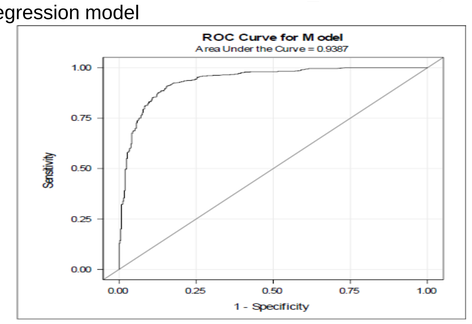
Figure 2: Distribution of 'diag' across hospitals



The above tables in Figure 1 show descriptive summary statistics (count, mean, median, standard deviation, min, max, and missing) of number of majoring blood vessels with > 50% diameter narrowing across the covariates used in the final model. Most appear to show differences between the "no heart disease" group and the groups that had 1+ number of majoring blood vessels with > 50% diameter narrowing. Figure 2 shows how there could have been a huge influence on which hospital they were staying at on developing heart disease.

Figure 3 (left): Parameter estimates using final model (outcome is dichotomous 'nar_vess' variable)
Figure 4 (right): ROC curve of the final logistic regression model

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-3.987191	0.540351	54.4481	<.0001
sex Male	1	0.903118	0.286687	9.9237	0.0016
cp Asymptomatic	1	1.003834	0.407167	6.0783	0.0137
cp Atypical angina	1	-0.734528	0.470550	2.4367	0.1185
cp Non-anginal pain	1	-0.322605	0.425459	0.5749	0.4483
exang Yes	1	1.005086	0.242479	17.1814	<.0001
oldpeak	1	0.511076	0.120820	17.8933	<.0001
slope Downsloping	1	0.309395	0.405553	0.5820	0.4455
slope Flat	1	0.838154	0.242434	11.9525	0.0005
ca	1	0.936670	0.130856	51.2375	<.0001
thal Fixed defect	1	1.125823	0.377362	8.9007	0.0029
thal Reversable defect	1	1.657541	0.228914	52.4303	<.0001
hosp EU2	1	3.023173	0.472651	40.9114	<.0001
hosp US1	1	0.312146	0.270303	1.3336	0.2482
hosp US2	1	0.669332	0.307869	4.7266	0.0297



The final logistic regression model using the dichotomous 'nar_vess' variable as the outcome had much better interpretability than the polytomous model using 'diag'. The following are interpretations of the model, noting only the covariates that were significant at the 5% level: for males, the odds of developing heart disease was 2.467 times the odds of developing heart disease for females, adjusting for all other covariates. For those with asymptotic chest pain, the odds of developing heart disease was 2.729 times the odds of developing heart disease for those suffering from typical angina chest pain, adjusting for all other covariates. For those with exercise induced angina, odds of developing heart disease was 2.732 times the odds of developing heart disease for those without exercise induced agina, adjusting for all other covariates. For every one unit increase in ST depression induced by exercise relative to rest, the odds of developing heart disease increased by 67%, adjusting for all other covariates. For those with flat slopes from the peak exercise ST segment, the odds of developing heart disease was 2.312 times the odds of developing heart disease for those with upward slopes, adjusting for all other covariates. For every increase by one major vessel colored by fluoroscopy, the odds of developing heart disease increased by 155%, adjusting for all other covariates. For those having fixed defect thalassemia blood disorder, the odds of developing heart disease was 3.083 times the odds of developing heart disease for those having no blood disorder, adjusting for all other covariates. For those having reversable defect thalassemia blood disorder, the odds of developing heart disease was 5.246 times the odds of developing heart disease for those having no blood disorder, adjusting for all other covariates.

RESULTS (Cont.)

For those in the European hospital #2, the odds of developing heart disease was 20.555 times the odds of developing heart disease for those in the European hospital #1, adjusting for all other covariates. For those in the U.S. hospital #2, the odds of developing heart disease was 1.953 times the odds of developing heart disease for those in the European hospital #1, adjusting for all other covariates. The AUC of the ROC curve displayed in Figure 4 is roughly 0.94, which shows excellent predictability. Lastly, the Hosmer-Lemeshow test for goodness-of-fit showed that the model fit well.

CONCLUSION

The final predictive model using presence of heart disease as the outcome variable demonstrated excellent predictability, interpretability, and a good fit. The final model included patients' sex, type of chest pain, exercise-induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, presence of the blood disorder, thalassemia, number of major vessels colored by fluoroscopy, and which hospital the patient stayed in. Females tended to be less at-risk of heart disease than men. Those with asymptotic chest pain were significantly more likely to suffer from heart disease than those with angina chest pains. Patients with exercise-induced angina were found to be at higher risk of heart disease than patients without exercise-induced angina. ST depression induced by exercise relative to rest and number of major vessels colored by fluoroscopy were found to be risk factors for heart disease. Those with flat slopes from peak exercise ST segment were at a higher risk f heart disease than those with upward slopes. Those having fixed defect and reversable defect thalassemia were at much higher risk than those having no blood disorder. Lastly, and maybe the most notably, staying at one of the European hospitals seemed to significantly impact the presence of heart disease. This large difference in patients developing heart disease across the different hospitals could be attributed to geographic differences. These geographic differences could then lead to differences in socioeconomic status, which is a significant factor for heart health (diet, stress, exercise, etc.).

Possible limitations of this study could include not taking into account major differences between the hospitals and care/testing done within them. For example, testing could have been done differently in assessing the aforementioned covariates and their relationship to heart disease. One hospital could have done a better job of monitoring or even providing care for those who did seem at-risk in this study. Another possible limitation could have been possible interactions between some of the covariates, although this would have caused any final models to lack interpretability.

Overall, the final logistic regression model presented in this study does a great job in predicting the possible risk factors of heart disease, in order to better screen for them. This improvement in screening in the future will allow for better diagnoses of patents being at-risk, which could prevent them from ever developing heart disease, which is by far the world's leading cause of death.[3]

Future studies could examine possible interactions between some of the covariates studied here, assess patients' socioeconomic status, and factor other things in such as geographic location differences, which were present here using the different hospitals.

BIBLIOGRAPHY

1 The World Health Organization: Cardiovascular diseases. Accessed on 12/20/21. URL: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

2 Mohammad H. Forouzanfar, Andrew E. Moran, Abraham D. Flaxman, Gregory Roth, George A. Mensah, Majid Ezzati, Mohsen Naghavi, Christopher J.L. Murray, Assessing the Global Burden of Ischemic Heart Disease: Part 2: Analytic Methods and Estimates of the Global Epidemiology of Ischemic Heart Disease in 2010, Global Heart, Volume 7, Issue 4, 2012, Pages 331-342, ISSN 2211-8160, https://doi.org/10.1016/j.gheart.2012.10.003.

3 S.Chellammal, R. Sharmila. Recommendation of Attributes for Heart Disease Prediction using Correlation Measure. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S3, July 2019