Statistical Study of Heart Failure

Research team

1. Pattradanai Chairach student id 6442080026
2. Natdanai Nakwongwankul student id 6342029026
3. Supanat Srirod student id 6542115526
4. Chananon Wongkondee student id 6542023426
5. Piraya Yingpisit student id 6542086026
6. Takorn Nithiyakorn student id 6542054926
7. Akarawit Uttamavetin student id 6542131526

Subject GLMS 2603311

Section 1

Term 2/2566

Present

Associate Professor Dr. Vitara Pungpapong

Abstract

This report describes the development of a novel statistical method for identifying patients at high risk for developing heart failure. The method utilizes patient data to uncover statistically significant associations between specific factors and the onset of heart failure. By identifying these high-risk patients, healthcare providers can proactively implement early intervention strategies. This approach has the potential to significantly improve patient outcomes by preventing the progression of heart failure or mitigating its severity.

Keywords: Identifying patients, Heart failure, High risk

Introduction

Cardiovascular diseases (CVDs) are a group of disorders affecting your heart and blood vessels. The most common CVDs are coronary heart disease, stroke, and peripheral arterial disease. CVDs reign as the global killer, claiming millions of lives annually. Despite advancements in medical science, CVDs still account for a staggering 31% of all deaths worldwide, with heart attacks and strokes being the most common culprits.

CVDs, or cardiovascular diseases, develop due to various factors damaging your heart and blood vessels. The most common culprit is atherosclerosis, a gradual build-up of fatty deposits (plaque) lining the arteries. This plaque narrows the arteries, reducing blood flow to vital organs like the heart and brain. There are two main ways CVDs can occur due to atherosclerosis. First, is blocked arteries, which means the plague has buildup and completely blocked arteries, leading to a heart attack (coronary artery) or stroke (carotid or cerebral artery). The second way is blood clots, which result from a piece of plaque breaking off and triggering blood clot formation. This clot then travels and lodges in a narrowed artery, causing similar problems as a complete blockage. Other factors can also contribute to CVDs such as high blood pressure, high blood cholesterol, and diabetes.

This study delves into a dataset containing 11 variables that hold the potential to predict heart failure from the potential cause mentioned above, a crucial step in early detection and management. Using various statistical methods, we aim to build a model that can identify individuals at risk, paving the way for preventative measures and potentially saving lives.

Objectives

1. Developing a statistical method to identify patients at risk for heart failure.
2. Identifying statistically significant risk factors associated with heart failure.

Research Methodology

1. Sample

This dataset comprises secondary data consisting of 918 observations with various variables including Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG MaxHR, ExerciseAngina, Oldpeak, ST_Slope, and HeartDisease.

2. Description of the dataset

This study utilizes a novel, secondary dataset constructed by combining five independent heart disease datasets. These datasets, originating from institutions like the Hungarian Institute of Cardiology and University Hospitals in Zurich and Basel, Switzerland, offer a comprehensive view of heart disease across diverse populations. This combined dataset boasts 11 shared features, making it one of the largest collections of its kind currently available for heart disease research. Furthermore, the dataset comprises 918 observations encompassing a broad age range (28-77 years) and is stratified into two distinct categories: patients with confirmed heart disease and those who did not.

3. Definitions

- **Age:** age of the patient [years]
- **Sex:** sex of the patient [M: Male, F: Female]
- **ChestPainType:** chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- **RestingBP:** resting blood pressure [mm Hg]
- **Cholesterol :** serum cholesterol [mm/dl]
- **FastingBS:** fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- **RestingECG:** resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- **MaxHR:** maximum heart rate achieved [Numeric value between 60 and 202]
- **ExerciseAngina:** exercise-induced angina [Y: Yes, N: No]
- **Oldpeak:** old peak = ST [Numeric value measured in depression]
- **ST_Slope:** the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- **HeartDisease:** output class [1: heart disease, 0: Normal]

4. Exploratory data analysis

In the data exploration process, our initial step involves examining the data for errors, primarily by scrutinizing the summary statistics.

```
> summary(heart.df)
      Age         Sex     ChestPainType   RestingBP      Cholesterol      FastingBS       RestingECG      MaxHR       ExerciseAngina    Oldpeak
 Min.   :28.00   F:193   ASY:496       Min.   :  0.0   Min.   :  0.0   Min.   :0.0000   LVH   :188   Min.   : 60.0   N:547          Min.   :-2.6000
 1st Qu.:47.00   M:725   ATA:173       1st Qu.:120.0   1st Qu.:173.2   1st Qu.:0.0000   Normal:552   1st Qu.:120.0   Y:371          1st Qu.: 0.0000
 Median :54.00           NAP:203       Median :130.0   Median :223.0   Median :0.0000   ST    :178   Median :138.0                  Median : 0.6000
 Mean   :53.51           TA : 46       Mean   :132.4   Mean   :198.8   Mean   :0.2331                  Mean   :136.8                  Mean   : 0.8874
 3rd Qu.:60.00                         3rd Qu.:140.0   3rd Qu.:267.0   3rd Qu.:0.0000                  3rd Qu.:156.0                  3rd Qu.: 1.5000
 Max.   :77.00                         Max.   :200.0   Max.   :603.0   Max.   :1.0000                  Max.   :202.0                  Max.   : 6.2000
 ST_Slope    HeartDisease
 Down: 63   Min.   :0.0000
 Flat:460   1st Qu.:0.0000
 Up  :395   Median :1.0000
            Mean   :0.5534
            3rd Qu.:1.0000
            Max.   :1.0000
```

Figure 1 Shows the summary of raw data

After analyzing the dataset using descriptive statistics, as depicted in Figure 1, several errors were identified among the 918 observations. These anomalies necessitated corrective measures to ensure the integrity of the data for subsequent analysis.

1. RestingBP Column: An aberrant value of 0 was detected in one observation, which is physiologically implausible as it suggests a state incompatible with life. Consequently, this observation was deemed erroneous and removed from the dataset.

2. Cholesterol Column: Approximately 171 observations were found to have a cholesterol value of 0, a biologically unrealistic occurrence in humans. To address this issue, Multiple Imputation by Chained Equations (MICE) methodology was employed to substitute these erroneous values with more plausible estimates, thereby enhancing the dataset's reliability for further analysis.

3. Oldpeak Column: Thirteen observations were noted to contain negative values in the Oldpeak column, likely attributable to measurement inaccuracies. To conform to statistical conventions and facilitate subsequent analyses, these negative values were uniformly transformed to 0.

By rectifying these discrepancies and applying appropriate data preprocessing techniques, the dataset is now primed for rigorous analysis, ensuring the validity and accuracy of subsequent findings, as shown in Figure 2 below.

```
> summary(heart)
      Age            Sex         ChestPainType   RestingBP      Cholesterol    FastingBS  RestingECG           MaxHR       ExerciseAngina
 Min.   :28.00   F:193   ASY:496       Min.   : 80.0   Min.   : 85.0   0:704     Length:918         Min.   : 60.0   N:547
 1st Qu.:47.00   M:725   ATA:173       1st Qu.:120.0   1st Qu.:208.0   1:214     Class :character   1st Qu.:120.0   Y:371
 Median :54.00           NAP:203       Median :130.0   Median :239.0             Mode  :character   Median :138.0
 Mean   :53.51           TA : 46       Mean   :132.5   Mean   :246.3                              Mean   :136.8
 3rd Qu.:60.00                         3rd Qu.:140.0   3rd Qu.:276.0                              3rd Qu.:156.0
 Max.   :77.00                         Max.   :200.0   Max.   :603.0                              Max.   :202.0
    Oldpeak        ST_Slope     HeartDisease
 Min.   :0.0000   Down: 63   Min.   :0.0000
 1st Qu.:0.0000   Flat:460   1st Qu.:0.0000
 Median :0.6000   Up  :395   Median :1.0000
 Mean   :0.9013              Mean   :0.5534
 3rd Qu.:1.5000              3rd Qu.:1.0000
 Max.   :6.2000              Max.   :1.0000
```

Figure 2 Showcases the summary of clean data

After employing descriptive statistics to obtain an overall summary of the cleaned data, we will utilize visualization techniques, such as bar charts and heatmaps, to explore the variable, including the target variable.

- Bar chart

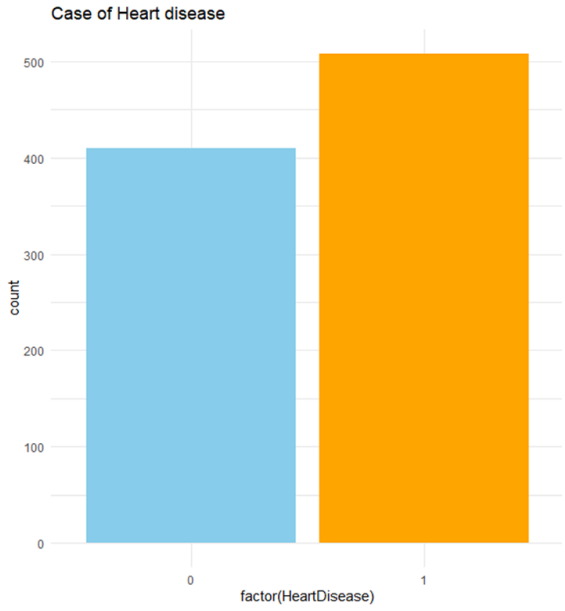

Figure 3 Shows the number of heart disease case by using bar chart

According to Figure 3, the data portrays a relatively balanced distribution, with 508 individuals (55%) having heart disease and 410 individuals (45%) without.

- Heatmap

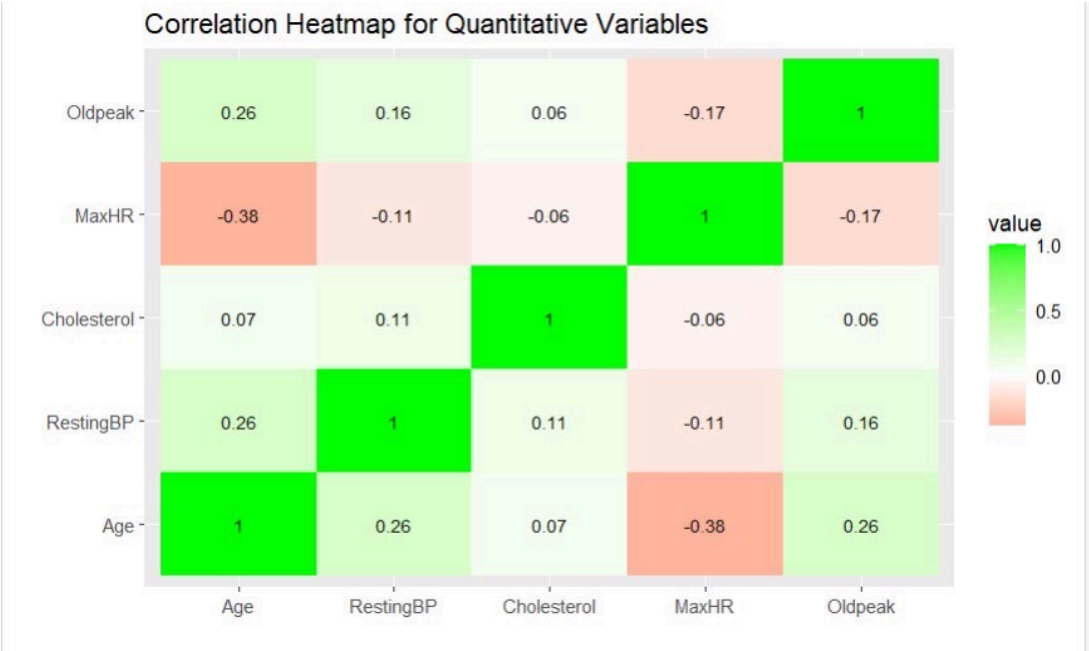

Figure 4 Indicate the correlation heatmap for quantitative variables

Upon meticulous inspection of each variable concerning the target outcome, the correlation heatmap elucidates the interdependencies among the continuous variables within the cleaned dataset.

Inter-variable Relationships:

1. Age and RestingBP: A discernible positive correlation exists between these variables. This pattern suggests a tendency for blood pressure to elevate concomitantly with increasing age, which aligns with the established medical understanding that blood pressure can escalate as part of the aging process.
2. MaxHR and Age: There is an inverse correlation observed, signifying that the peak heart rate during exertion tends to diminish as individuals age. This is a physiological consequence of aging, as the maximum heart rate is known to decrease over the lifespan.

5. Method for Fitting Model and Results

We use the binomial family in GLMS to fit our model, as the target variable represents a binary response. Additionally, we employ backward elimination with a significance level of alpha=0.05.

- Model selection - Backward elimination

```
> fullbinmodel <- update(glm(HeartDisease ~ ., data = heart, family = binomial))
> summary(fullbinmodel)

Call:
glm(formula = HeartDisease ~ ., family = binomial, data = heart)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.95749    1.45665   -1.34  0.17900
Age               0.01868    0.01312    1.42  0.15444
SexM              1.62352    0.27889    5.82  5.8e-09 ***
ChestPainTypeATA -1.89495    0.32435   -5.84  5.2e-09 ***
ChestPainTypeNAP -1.60462    0.25917   -6.19  6.0e-10 ***
ChestPainTypeTA  -1.47386    0.43042   -3.42  0.00062 ***
RestingBP         0.00159    0.00617    0.26  0.79654
Cholesterol       0.00176    0.00186    0.95  0.34352
FastingBS         1.31358    0.26677    4.92  8.5e-07 ***
RestingECGNormal  0.01746    0.26749    0.07  0.94795
RestingECGST     -0.00736    0.34308   -0.02  0.98289
MaxHR            -0.00765    0.00489   -1.56  0.11778
ExerciseAnginaY   0.84915    0.24078    3.53  0.00042 ***
Oldpeak           0.40545    0.12122    3.34  0.00082 ***
ST_SlopeFlat      1.27569    0.42999    2.97  0.00301 **
ST_SlopeUp       -1.03849    0.44888   -2.31  0.02069 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1262.1  on 917  degrees of freedom
Residual deviance:  606.7  on 902  degrees of freedom
AIC: 638.7

Number of Fisher Scoring iterations: 5
```

Figure 5 illustrates the outcome of the initial stage of Backward Elimination

Hypothesis

H0: RestingECG is not a significant predictor

H1: RestingECG is a significant predictor

We apply an alpha of 0.05. Since the p-value of both RestingECGNormal and RestingECGST are 0.94795 and 0.98289 respectively, which is greater than 0.05, we fail to reject the null hypothesis (H0). Therefore, we can conclude that there is insufficient evidence to conclude that RestingECG is a significant predictor at the 0.05 significance level.

```
> mod2<- update(glm(HeartDisease ~. -RestingECG, data=heart, family=binomial))
> summary(mod2)

Call:
glm(formula = HeartDisease ~ . - RestingECG, family = binomial,
    data = heart)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.93122    1.36957   -1.41  0.15851
Age               0.01843    0.01272    1.45  0.14735
SexM              1.62200    0.27776    5.84  5.2e-09 ***
ChestPainTypeATA -1.89389    0.32387   -5.85  5.0e-09 ***
ChestPainTypeNAP -1.60412    0.25911   -6.19  6.0e-10 ***
ChestPainTypeTA  -1.47523    0.42996   -3.43  0.00060 ***
RestingBP         0.00157    0.00616    0.25  0.79913
Cholesterol       0.00175    0.00185    0.95  0.34357
FastingBS         1.31314    0.26549    4.95  7.6e-07 ***
MaxHR            -0.00764    0.00476   -1.60  0.10870
ExerciseAnginaY   0.84902    0.24020    3.53  0.00041 ***
Oldpeak           0.40512    0.12105    3.35  0.00082 ***
ST_SlopeFlat      1.27753    0.42956    2.97  0.00294 **
ST_SlopeUp       -1.03732    0.44858   -2.31  0.02075 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1262.14  on 917  degrees of freedom
Residual deviance:  606.71  on 904  degrees of freedom
AIC: 634.7

Number of Fisher Scoring iterations: 5
```

Figure 6 illustrates the outcome of the second stage of Backward Elimination

Hypothesis

H0: RestingBP is not a significant predictor

H1: RestingBP is a significant predictor

We apply an alpha of 0.05. Since the p-value is 0.79913, which is greater than 0.05, we fail to reject the null hypothesis (H0). Therefore, we can conclude that there is insufficient evidence to conclude that RestingBP is a significant predictor at the 0.05 significance level.

```
> mod3<- update(glm(HeartDisease ~. -RestingECG - RestingBP, data=heart, family=binomial))
> summary(mod3)

Call:
glm(formula = HeartDisease ~ . - RestingECG - RestingBP, family = binomial,
    data = heart)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.78981    1.25164   -1.43  0.15272
Age              0.01912    0.01244    1.54  0.12409
SexM             1.62016    0.27740    5.84  5.2e-09 ***
ChestPainTypeATA -1.88861    0.32310   -5.85  5.1e-09 ***
ChestPainTypeNAP -1.60236    0.25903   -6.19  6.2e-10 ***
ChestPainTypeTA  -1.46480    0.42801   -3.42  0.00062 ***
Cholesterol      0.00181    0.00184    0.98  0.32479
FastingBS        1.31285    0.26534    4.95  7.5e-07 ***
MaxHR           -0.00764    0.00476   -1.60  0.10857
ExerciseAnginaY  0.85633    0.23849    3.59  0.00033 ***
Oldpeak          0.40724    0.12063    3.38  0.00074 ***
ST_SlopeFlat     1.28635    0.42845    3.00  0.00268 **
ST_SlopeUp      -1.02671    0.44695   -2.30  0.02161 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1262.14  on 917  degrees of freedom
Residual deviance:  606.77  on 905  degrees of freedom
AIC: 632.8
```

Figure 7 illustrates the outcome of the third stage of Backward Elimination

Hypothesis

H0: Cholesterol is not a significant predictor

H1: Cholesterol is a significant predictor

We apply an alpha of 0.05. Since the p-value is 0.32479, which is greater than 0.05, we fail to reject the null hypothesis (H0). Therefore, we can conclude that there is insufficient evidence to conclude that Cholesterol is a significant predictor at the 0.05 significance level.

```
> mod4<- update(glm(HeartDisease ~. -RestingECG - RestingBP - Cholesterol, data=heart, family=binomial))
> summary(mod4)

Call:
glm(formula = HeartDisease ~ . - RestingECG - RestingBP - Cholesterol,
    family = binomial, data = heart)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.39421    1.18465   -1.18  0.23924
Age              0.01943    0.01242    1.56  0.11775
SexM             1.57878    0.27347    5.77  7.8e-09 ***
ChestPainTypeATA -1.89127    0.32283   -5.86  4.7e-09 ***
ChestPainTypeNAP -1.62746    0.25805   -6.31  2.8e-10 ***
ChestPainTypeTA  -1.46719    0.42562   -3.45  0.00057 ***
FastingBS         1.31573    0.26485    4.97  6.8e-07 ***
MaxHR            -0.00736    0.00474   -1.55  0.12058
ExerciseAnginaY   0.86147    0.23837    3.61  0.00030 ***
Oldpeak           0.40173    0.12050    3.33  0.00086 ***
ST_SlopeFlat      1.33320    0.42650    3.13  0.00177 **
ST_SlopeUp       -1.01257    0.44723   -2.26  0.02357 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1262.14  on 917  degrees of freedom
Residual deviance:  607.76  on 906  degrees of freedom
AIC: 631.8

Number of Fisher Scoring iterations: 5
```

Figure 8 illustrates the outcome of the fourth stage of Backward Elimination

Hypothesis

H0: MaxHR is not a significant predictor

H1: MaxHR is a significant predictor

We apply an alpha of 0.05. Since the p-value is 0.12058, which is greater than 0.05, we fail to reject the null hypothesis (H0). Therefore, we can conclude that there is insufficient evidence to conclude that MaxHR is a significant predictor at the 0.05 significance level.

```
> mod5<- update(glm(HeartDisease ~. -RestingECG - RestingBP - MaxHR - Cholesterol, data=heart, family=binomial))
> summary(mod5)

Call:
glm(formula = HeartDisease ~ . - RestingECG - RestingBP - MaxHR -
    Cholesterol, family = binomial, data = heart)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.7367     0.8149   -3.36  0.00078 ***
Age               0.0259     0.0117    2.20  0.02751 *
SexM              1.6201     0.2727    5.94  2.8e-09 ***
ChestPainTypeATA -1.9410     0.3209   -6.05  1.5e-09 ***
ChestPainTypeNAP -1.6828     0.2553   -6.59  4.4e-11 ***
ChestPainTypeTA  -1.5500     0.4272   -3.63  0.00029 ***
FastingBS         1.3251     0.2642    5.01  5.3e-07 ***
ExerciseAnginaY   0.9321     0.2333    3.99  6.5e-05 ***
Oldpeak           0.3754     0.1186    3.17  0.00154 **
ST_SlopeFlat      1.3343     0.4231    3.15  0.00161 **
ST_SlopeUp       -1.0889     0.4413   -2.47  0.01362 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1262.14  on 917  degrees of freedom
Residual deviance:  610.17  on 907  degrees of freedom
AIC: 632.2

Number of Fisher Scoring iterations: 5
```

Figure 9 illustrates the outcome of the last stage of Backward Elimination

Hypothesis

H0: Age is not a significant predictor

H1: Age is a significant predictor

We apply an alpha of 0.05. Since the p-value is 0.02751, which is less than 0.05, we reject the null hypothesis (H0). Therefore, we can conclude that there is sufficient evidence to conclude that Age is a significant predictor at the 0.05 significance level.

After applying backward elimination, our resulting model includes Age, Sex, ChestPainType, FastingBS, ExerciseAngina, Oldpeak, and ST_Slope as the key predictors deemed significant, consistent with the Chi-Square test.

After obtaining the initial model, Our investigation suggested the inclusion of interactions between Age and RestingBP, as well as ChestPainType and RestingECG. However, upon performing Wald's test, only the interaction between ChestPainType and RestingECG emerged as statistically significant, warranting its inclusion in the model.

```
> summary(heart.binmodel)

Call:
glm(formula = HeartDisease ~ Age + Sex + ChestPainType + FastingBS +
    ExerciseAngina + RestingECG + ChestPainType:RestingECG +
    +Oldpeak + ST_Slope, family = binomial, data = heart)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -2.5751     0.8980   -2.87  0.00413 **
Age                               0.0284     0.0122    2.33  0.01985 *
SexM                              1.7858     0.2839    6.29  3.2e-10 ***
ChestPainTypeATA                 -1.1493     0.6408   -1.79  0.07287 .
ChestPainTypeNAP                 -2.8381     0.5583   -5.08  3.7e-07 ***
ChestPainTypeTA                  -3.2879     0.7322   -4.49  7.1e-06 ***
FastingBS                         1.3160     0.2683    4.91  9.3e-07 ***
ExerciseAnginaY                   0.9078     0.2406    3.77  0.00016 ***
RestingECGNormal                 -0.3309     0.3714   -0.89  0.37291
RestingECGST                     -0.7340     0.4496   -1.63  0.10257
Oldpeak                           0.3908     0.1227    3.19  0.00144 **
ST_SlopeFlat                      1.2762     0.4383    2.91  0.00360 **
ST_SlopeUp                       -1.2212     0.4583   -2.66  0.00771 **
ChestPainTypeATA:RestingECGNormal -1.2697     0.7753   -1.64  0.10148
ChestPainTypeNAP:RestingECGNormal  1.2008     0.6388    1.88  0.06013 .
ChestPainTypeTA:RestingECGNormal   2.5855     0.9423    2.74  0.00607 **
ChestPainTypeATA:RestingECGST     -0.2488     0.9993   -0.25  0.80341
ChestPainTypeNAP:RestingECGST      2.4025     0.8384    2.87  0.00416 **
ChestPainTypeTA:RestingECGST       2.8063     1.2669    2.22  0.02676 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1262.14  on 917  degrees of freedom
Residual deviance:  588.65  on 899  degrees of freedom
AIC: 626.6

Number of Fisher Scoring iterations: 6
```

Figure 10 Illustrates the summary of the model fitted by using full data

- Model Evaluation

Following the acquisition of the model for our data, we partition it into training(70%) and testing(30%) datasets to ensure that our model performs well on unseen data, generalizes effectively, and avoids overfitting.

```
> summary(heart_train.binmodel)

Call:
glm(formula = HeartDisease ~ Age + Sex + ChestPainType + FastingBS +
    +ExerciseAngina + RestingECG + ChestPainType:RestingECG +
    Oldpeak + ST_Slope, family = binomial, data = train_part)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.83495    1.03486  -2.739 0.006154 **
Age                              0.02998    0.01433   2.092 0.036480 *
SexM                             1.46113    0.33117   4.412 1.02e-05 ***
ChestPainTypeATA                -0.73486    0.78242  -0.939 0.347626
ChestPainTypeNAP                -2.07078    0.63819  -3.245 0.001176 **
ChestPainTypeTA                 -2.90283    0.91274  -3.180 0.001471 **
FastingBS                        1.37526    0.31872   4.315 1.60e-05 ***
ExerciseAnginaY                  0.95202    0.27294   3.488 0.000487 ***
RestingECGNormal                -0.11141    0.42705  -0.261 0.794178
RestingECGST                    -0.68547    0.51842  -1.322 0.186092
Oldpeak                          0.41211    0.13494   3.054 0.002258 **
ST_SlopeFlat                     1.19005    0.50405   2.361 0.018227 *
ST_SlopeUp                      -1.05518    0.53048  -1.989 0.046688 *
ChestPainTypeATA:RestingECGNormal -1.39472    0.91449  -1.525 0.127227
ChestPainTypeNAP:RestingECGNormal  0.57644    0.73645   0.783 0.433787
ChestPainTypeTA:RestingECGNormal   2.34934    1.15533   2.033 0.042005 *
ChestPainTypeATA:RestingECGST     -0.04429    1.14306  -0.039 0.969089
ChestPainTypeNAP:RestingECGST      2.11581    0.96944   2.183 0.029072 *
ChestPainTypeTA:RestingECGST       2.94538    1.50199   1.961 0.049882 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 883.53  on 642  degrees of freedom
Residual deviance: 445.84  on 624  degrees of freedom
AIC: 483.84

Number of Fisher Scoring iterations: 5
```

Figure 11 Illustrates the summary of the model fitted by using training data

```
> roc_heart

Call:
roc.default(response = test_part$HeartDisease, predictor = predictedprob,     plot = TRUE)

Data: predictedprob in 124 controls (test_part$HeartDisease 0) < 151 cases (test_part$HeartDisease 1).
Area under the curve: 0.9589
```

Figure 12 Illustrates the area under the receiver operating characteristics curve (ROC Curve) indicating the performance of a classification model

```
> best_cutoff_heart
  threshold
1 0.5248212
```

Figure 13 Illustrates the best cutoff for our model on the test dataset

```
> table(test_part$HeartDisease, yhat)
   yhat
      0   1
  0 115   9
  1  18 133
```

Figure 14 Illustrates the confusion matrix for our model on the test dataset

```
> hosmerlem(train_part$HeartDisease, heart_train.binmodel$fitted.values)
     X^2       Df   P(>Chi)
5.147667 8.000000 0.741683
```

Figure 15 Illustrates the results from the Hosmer-Lemeshow test for our model on the test dataset.

Hypothesis

H0: Model fits the data

H1: Model does not fit the data

We apply an alpha of 0.05. Since the p-value is 0.741683, which is greater than 0.05, we fail to reject the null hypothesis (H0). Therefore, we can conclude that there is insufficient evidence to suggest that the model does not fit the data at the 0.05 significance level.

Upon partitioning the dataset into two subsets, training, and testing, we evaluated the predictive performance of our model. This evaluation resulted in a p-value of0.741683 for the Hosmer-Lemeshow test, surpassing the significance level ($\alpha = 0.05$). Consequently, we did not reject the null hypothesis (H0), signifying inadequate evidence to suggest that our model does not fit the data at the 95% significance level. Thus, we conclude that our model fits the data well.

Figure 14 displays the results of our confusion matrix. We applied the model to the test dataset (test_part) to assess its predictive performance. Predicted probabilities of heart disease were derived using the predict function, with a threshold of 0.5248212 chosen through receiver operating characteristic (ROC) curve analysis to optimize sensitivity and specificity. The area under the ROC curve (AUC) was approximately 0.9589, indicating the excellent discriminatory power of the model.

Conclusion

This research utilized a secondary heart failure prediction dataset obtained from Kaggle. The dataset contained 918 observations, split into a training set (642 observations, 70%) and a test set (275 observations, 30%) by using a clean dataset. Before analysis, data cleaning was performed. This included excluding the cholesterol level column due to a majority of missing data. Other problematic values, such as invalid entries, were also removed.

Following data cleaning, exploratory data analysis (EDA) was conducted to understand the dataset's attributes and elements relevant to our research objectives. The next stage involved model fitting using various methods. First, we employed a chi-square test, backward elimination with AIC, and Wald test to eliminate irrelevant independent variables, resulting in an initial model. Further analysis revealed statistically significant interaction terms between Age and RestingBP, which were then incorporated into the model.

Once the final model was established, its performance was evaluated using ROC analysis to determine the optimal cutoff point. Additionally, a confusion matrix was generated to assess the model's performance on the test set. Finally, the Hosmer-Lemeshow goodness-of-fit test confirmed that the model adequately fits the data, and all independent variables within the final model were statistically significant.

In summary, we have reached both objectives of the study including developing a statistical method to identify patients at risk for heart failure and identifying statistically significant risk factors associated with heart failure. By looking at Figure 9, It is established that statistically significant factors in identifying heart failure are Age, Sex, ChestPainType, FastingBS, ExerciseAngina, Oldpeak, and ST_Slope. Additionally, after considering interaction terms, the interaction between ChestPainType and RestingECG is included in the list of statistically significant factors. Moreover, by looking at Figure 11, a model for predicting heart failure is established. To evaluate this model's performance, we employed statistical tests and methods like the ROC Curve, Hosmer-Lemeshow test, and confusion matrix. These tests assessed the model's predictive ability, goodness-of-fit, and discriminatory power. The positive results from these tests demonstrate the model's effectiveness in accurately predicting heart failure.

Lastly, Our study offers a valuable tool for predicting heart failure and identifying risk factors for heart failure. Therefore, Our study can provide great benefit by being used to develop better screening tools and raise awareness about modifiable risk factors for heart failure.

Suggestion

This research paper analyzes a sample created by combining data from five heart disease datasets originating from different locations. However, the combined sample may not be entirely representative of the target population.

In our analysis of the Kaggle heart failure prediction dataset, we discovered that the cholesterol column contained a higher proportion of error values (171 out of 918 observations) than we could tolerate. This high number of errors could significantly impact the model's performance.

Reference

1.FEDESORIANO(2021).heart failure prediction dataset.Available at: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
2.Flávio D. Fuchs and Paul K. Whelton(2020).Hypertension.Available at:https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.119.14240
3.Julian M Aroesty, MDJoseph P Kannam, MD(2022).Chest pain.Available at:https://www.uptodate.com/contents/chest-pain-beyond-the-basics
4.MountSinai.Hardening of the arteries.Available at:https://www.mountsinai.org/health-library/diseases-conditions/hardening-of-the-arteries

Appendix

```
#1 Data Preparation
setwd("C:/Users/Ppeachary/Downloads")
heart.df <- read.csv('heart3.csv')
heart.df$Sex <- as.factor(heart.df$Sex)
heart.df$ChestPainType <- as.factor(heart.df$ChestPainType)
heart.df$RestingECG <- as.factor(heart.df$RestingECG)
heart.df$ExerciseAngina <- as.factor(heart.df$ExerciseAngina)
heart.df$ST_Slope <- as.factor(heart.df$ST_Slope)
summary(heart.df)
heart.df$Cholesterol[heart.df$Cholesterol == 0] <- NA
heart.df$RestingBP[heart.df$RestingBP == 0] <- NA
#Import ข้อมูล + ทำให้ข้อมูล 0 เป็น NA
#install.packages("mice")
library(mice)
heart <- complete(mice(heart.df, method = "pmm")) #MICE Imputation
heart$Oldpeak[heart$Oldpeak < 0] <- 0 #แทนที่ค่า Negative Oldpeak
summary(heart)
#backward elimination
fullbinmodel <- update(glm(HeartDisease ~ ., data = heart, family = binomial))
summary(fullbinmodel)
mod2<- update(glm(HeartDisease ~. -RestingECG, data=heart, family=binomial))
summary(mod2)
mod3<- update(glm(HeartDisease ~. -RestingECG - RestingBP, data=heart, family=binomial))
summary(mod3)
mod4<- update(glm(HeartDisease ~. -RestingECG - RestingBP - Cholesterol, data=heart,
family=binomial))
summary(mod4)
mod5<- update(glm(HeartDisease ~. -RestingECG - RestingBP - MaxHR - Cholesterol,
data=heart, family=binomial))
summary(mod5)
#model
heart.binmodel<- glm(HeartDisease ~ Age + Sex + ChestPainType + FastingBS
                +ExerciseAngina +RestingECG + ChestPainType:RestingECG + Oldpeak +
ST_Slope, data=heart, family=binomial)
summary(heart.binmodel)
plot(heart.binmodel, 4)
heart.binmodel<-glm(HeartDisease ~ Age + Sex + ChestPainType + FastingBS
```

```
              +ExerciseAngina +RestingECG + ChestPainType:RestingECG + Oldpeak +
ST_Slope, data=heart, family=binomial)
redmodel<-glm(HeartDisease ~ Age + Sex + ChestPainType + FastingBS
          +ExerciseAngina +RestingECG + Oldpeak + ST_Slope, data=heart, family=binomial)
LRT<-deviance(redmodel)-deviance(heart.binmodel)
1-pchisq(LRT, df.residual(redmodel)-df.residual(heart.binmodel))
#model train test
heart_train.binmodel<- glm(HeartDisease ~ Age + Sex + ChestPainType + FastingBS +
                    +ExerciseAngina + RestingECG + ChestPainType:RestingECG + Oldpeak +
ST_Slope, data=train_part, family=binomial)
summary(heart_train.binmodel)
#ROC
library(pROC)
roc_heart<- roc(test_part$HeartDisease, predictedprob, plot=TRUE)
roc_heart
#หา cutoff ดีสุด
best_cutoff_heart <- coords(roc_heart, "best", ret = "threshold")
best_cutoff_heart
#predict
predictedprob<-predict(heart_train.binmodel, test_part, type="response")
yhat <- rep(0, nrow(test_part))
yhat[predictedprob>=0.5248212]<-1
table(test_part$HeartDisease, yhat)
hosmerlem <-function (y, yhat, g = 10)
{
  cutyhat <- cut(yhat, breaks = quantile(yhat,
                                probs = seq(0,1, 1/g)), include.lowest = T)
  obs <- xtabs(cbind(1 - y, y) ~ cutyhat)
  expect <- xtabs(cbind(1 - yhat, yhat) ~ cutyhat)
  chisq <- sum((obs - expect)^2/expect)
  P <- 1 - pchisq(chisq, g - 2)
  c("X^2" = chisq, Df = g - 2, "P(>Chi)" = P)
}
hosmerlem(train_part$HeartDisease, heart_train.binmodel$fitted.values)
```