1. Identify an e-copy of a Healthcare or Biomedical Analytics paper from a peer reviewed journal with experiments in the last 5 years (real world or computational) (Done)
2. **Describe, in the student's own words, the contributions that the paper make to the field of health**
3. **Discuss issues, challenges and limitations relating the project described in the paper**
4. **Replicate one experimental plot with R or Python using data provided in the experiment or via simulation**
5. Report including source code and data are to be submitted via github (Done)

4.

There are 3 main experiments in the paper, namely Fisher’s Exact Test, Linear Regression Analysis and Logistic Regression Analysis. These methods follow a rational logic process. Firstly, the researchers want to know whether there is relationship between high PIRI score and high prevalence of Metabolic Syndrome, by using Fisher’s Exact Test. Secondly, after proving the positive relationship between PIRI and MES, they are willing to investigate more on PIRI score related to the number of MES components, by using Linear Regression. After determining higher PIRI scores have positive correlation with number of MES components, Logistic Regression is used to analyse whether psychological injury is significantly related to the presence of at least 1 MES component. The 3 experiments are conducted with top-down approach to investigate the association between work-related psychological injury and metabolic syndrome components in apparently healthy workers.

We choose to replicate experiment of logistic regression analysis with the data provided, by using R. The reason why we choose this experiment is that Fisher’s Exact Test and Linear Regression Analysis give us only a rough idea about the association between psychological injury and metabolic syndrome, although they are clear enough to understand. Only the Logistic Regression Analysis can provide us a more profound understanding of the topic.

Basically this experiment is to investigate whether there is significant relationship between high PIRI score and at least 1 MES syndrome as well between combining high PIRI score and biographic data and at least 1 MES syndrome, called model1 and model2 respectively. In both situations, we focus more on the following 3 index, namely p-value, odds ratio and R-square. Those 3 measures are the main ways to quantify how strongly the presence or absence of independent variables is associated with the presence or absence of dependent variables in a given population[i].

In R, we use *glm()* function to build logistic regression model. Both models use at least 1 MES syndrome as dependent variables. For model1, only high PIRI score (whether greater than 25) is used as independent variables. For model 2, high PIRI score and biological data (sex, age, smoker, alcohol, exercise and sleep) are treated as independent variables. *summary()* function is used to check the p-value of each independent variables. From Table XXX, we can see in model1, high PIRI score is significant as p-value is much smaller than 0.05. In model 2, besides high PIRI score, although there are 6 more independent variables, only age is significant with p-value smaller than 0.001. That shows only high PIRI score and age are the main factors which are statistically significant to MES syndrome.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **model1** | | | | **model2** | | | |
| **OR** | **95% CI(L&U)** | | **p-value** | **OR** | **95% CI(L&U)** | | **p-value** |
| **High PIRI** | 1.779 | 1.194 | 2.641 | **0.004** | 1.596 | 1.023 | 2.481 | **0.038** |
| **Sex** | - | - | - | - | 0.710 | 0.478 | 1.056 | 0.090 |
| **Age** | - | - | - | - | 1.057 | 1.037 | 1.078 | **0.000** |
| **Smoke** | - | - | - | - | 0.868 | 0.589 | 1.272 | 0.470 |
| **Alcohol** | - | - | - | - | 1.227 | 0.638 | 2.320 | 0.533 |
| **Exercise** | - | - | - | - | 1.056 | 0.710 | 1.582 | 0.788 |
| **Sleep** | - | - | - | - | 1.020 | 0.571 | 1.794 | 0.945 |
| **C&U R^2** | **0.019** | | | | **0.111** | | | |

Table XXX: Replicated Experiment Result by R

For odds ratio, we use R function exp(cbind(OR=coef(model), confint(model))) to calculate, subjected to different model names. Before biographic data added in the model, High PIRI has OR of 1.779 with 95% confidence interval (1.194-2.641). After biographic data added, High PIRI has OR of 1.596 with 95% C.I. (1.023-2.481) instead. For R-squared value which determines the fitness of the model, model after biographic data added has greater value 0.111 than model before biographic data added (0.019). Here, Cragg and Uhler’s Pseudo R-square is used which can be found in *pR2()* function in library(pscl). However, although R-squared value has improved more than 5 times after biographic data introduced to the model, it is still a relatively low value. In order words, the current model is not perfect as the small R-squared value does not explain too much variation of the model.

[i] Frank L. Schmidt, Methods of Meta-Analysis, Correcting Error and Bias in Research Findings