Università
di Catania

University of Catania

Master Degree Course in Data Science for Management

"Super Size Me"
Depression Analysis throw Dietary Habits

Author:
Alessia Simone

Professor:
Andrea Giuseppe Maugeri

Final Exam Project

1000037243

June 5, 2023

Contents

1 Abstract

Inspired by the thought-provoking documentary "Super Size Me," this report delves into an extensive data analysis focusing on patients with arthritis.

Like Morgan Spurlock's experiment, where he consumed a diet exclusively from McDonald's for an entire month, my study aims to explore the intricate relationship between arthritis, food habits, and depression levels among patients.

With Spurlock's journey as a cautionary tale, I investigate the potential effects of dietary choices on health outcomes, drawing valuable insights from a dataset encompassing patients with measured indicators of arthritis, food habits, and depression.

By examining these factors, we seek to shed light on the multifaceted nature of arthritis and its associated health consequences, aiming to contribute to public health interventions and strategies.

Furthermore, this analysis prompts us to critically examine broader societal factors that contribute to the soaring obesity rates in the United States, such as food accessibility, advertising influence, and their impact on individuals' well-being, particularly in the context of public health and young populations.

The dataset [Minamino et al., 2021] was collected through a self-reported questionnaire in addition to the KURAMA dataset.

2 Dataset Description

The dataset [Minamino et al., 2021] was collected through a self-reported questionnaire in addition to the KURAMA dataset. and was in ".xlsx" format where each column is separated into distinct cells.

Some variables' names have been modified as the "-" symbol can be misleading while performing linear regression analysis. Then, discrete variables have been converted as integer to allow the program to recognize them, "ID" and "Sex" column has been deleted as useless and null values have been omitted.

ID	Patient ID
Age	Age in years
Disease Duration	Duration of disease in years
Sex	Female patients only
BMI	Body Mass Index
Depression Score	HADS Questionnaire
Anxiety Score	HADS Questionnaire
VAS	Visual Analog Scale
DAS28-CRP	Disease Activity Score-28 with CRP
RA Stage	Theumatoid Arthritis stage
HAQ-DI	Health Assessment Questionnaire Disability Index
Biologics use	Biologics drugs
MTX use	Methotrexate drugs
Prednisolone use	Prednisolone drugs
Food Consumptions Variable	Food consumption frequencies
HGB	Haemoglobin
ALB	Albumin
CRP	C-Reactive Protein (inflammation mark)

Table 1: Variables description

3 Data Manipulation

Some additional variables have been extrapolated from the original variables to perform various types of analysis. As suggested in the official paper, the depression score has been divided into values from 0 to 9 indicating a low depression score and from 10 to 21 indicating a high depression score.

The plot 1 below shows that the majority of patients have a depression score lower than 10:

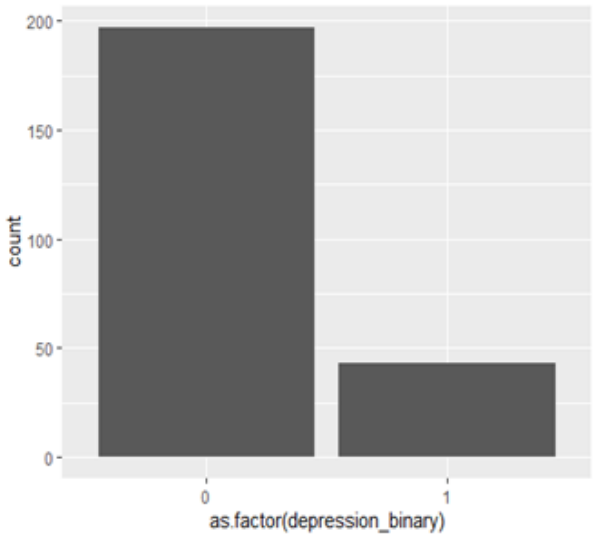

Figure 1: Barplot of binary depression variable

Then, a variable identifying the healthy or unhealthy diet habits of each patient has been created by counting the frequencies of healthy and unhealthy food types. As healthy food types staple foods, meat, fish, tofu, vegetable and fruits have been selected, as unhealthy food types fried food, cake, juice, snack food, sweets, miso soup, pickles, ham, frozen food and alcohol have been selected. The variable indicates 0 if the patient has a healthy diet habit, 1 instead.

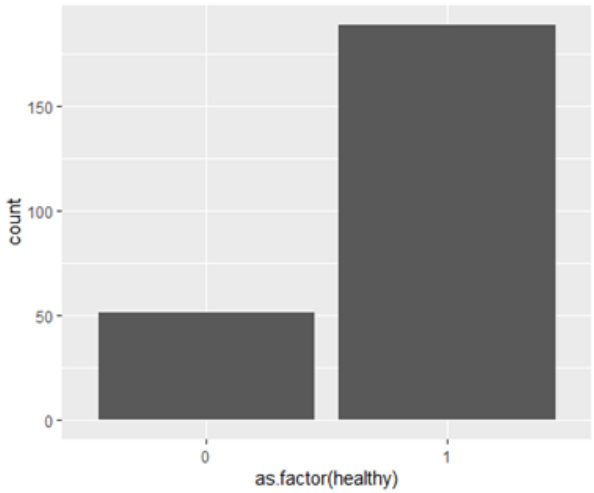

Figure 2: Barplot of binary diet habits

Finally, the third binary variable artificially created identify the stage of the arthritis with 0 meaning low level stage and 1 meaning high level stage.

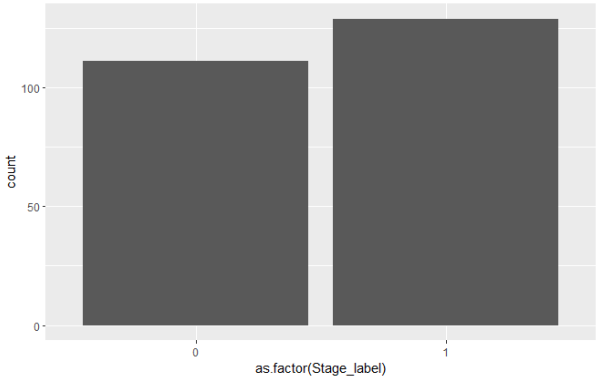

Figure 3: Barplot of arthritis stage

4 Data Analysis

Various types of analysis has been conducted to a deeply inspection of the behaviour of patients:

- **Descriptive Analysis:** descriptive statistics, such as distribution, correlation and standard deviation, has been computed and visualized in order to have a general understanding of the data.
- **Prevalence Analysis:** prevalence analysis helps estimate the proportion of the population affected by arthritis, depression and diet habits.
- **Risk Factor Analysis:** exploration the association between potential risk factors such as various types of food and the outcome of interest which is depression score in this case. This has been done using regression analysis to assess the impact of these factors on the outcome.
- **Dietary Analysis:** investigation the relationship between dietary factors (e.g., Staple food at breakfast, Meat, Vegetable) and health outcomes such as disease severity or inflammation markers (e.g., CRP). This involves analyzing dietary patterns, examining nutrient intake and exploring associations between specific food groups and health outcomes.
- **Mediation Analysis:** due to the suspect that dietary habits might mediate the relationship between depression and arthritis, conduction of mediation analysis has been applied to assess the indirect effects and quantify the extent of mediation.
- **Survival Analysis:** taking into consideration the disease duration, survival analysis has been performed using variables related to the disease only. This allows to estimate survival probabilities and identify factors influencing disease progression or mortality.

4.1 Descriptive Analysis

The dataset occupies 54 Kb and missing observations has been removed for this purpose. The dataset contains 37 measures of 240 patients. The variables have very different mean between each other which indicates different scales of measures and there aren't variables with a very low variance (see picture 5). The plots 4 represent the bar plots of the binary variables: the majority of patients don't use biologics and prednisolone drugs instead of MTX drugs. In addition, the majority of patients had a score on HAD questionnaire lower than 9 and have an unhealthy diet habit.

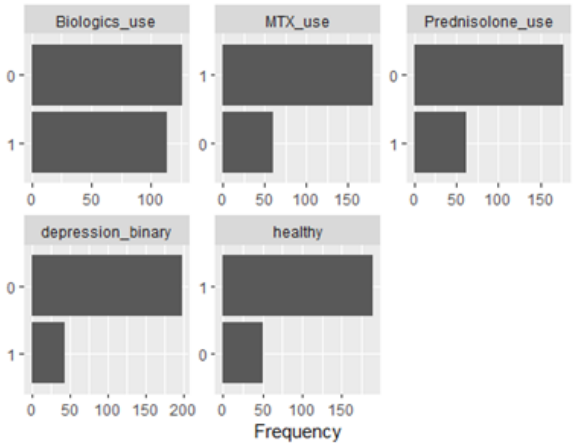

Figure 4: Barplots of binary variables

Then, the figure 5 represents the probability density function and, as we can see, almost all continuous variables are positively skewed.

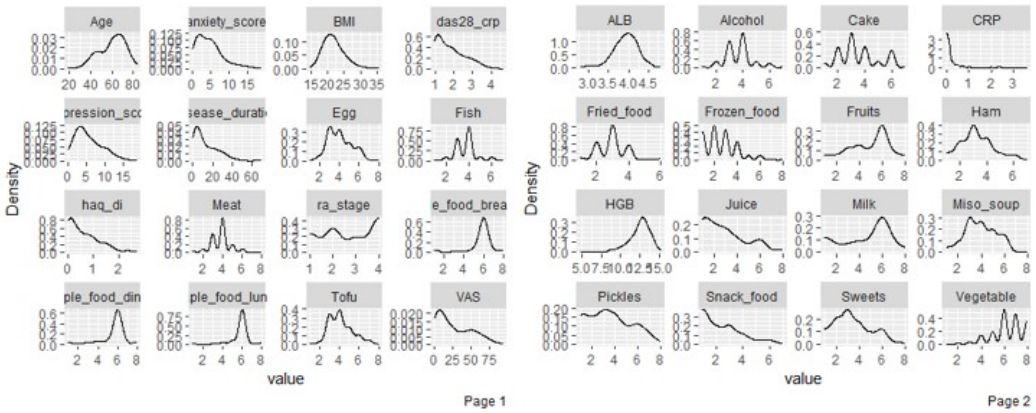

Figure 5: Density functions of continuous variables

4.2 Prevalence Analysis

For this purpose, three artificial binary variables has been taken into consideration: diet habits, RA stage and depression: 53.75% of patients have high stage of rheumathoids

arthritis, 17.92% of patients have high depression score to be attentioned of follow-up and 78.75% of patients have unhealthy diet habits.

4.3 Risk Factor Analysis

In this section, the linear regression analysis has been computed to understand the correlation between the depression score as a continuous variable and food consumption variables.

The variables have been selected with a step-wise variable selection in terms of the AIC index (1313.3):

- Eggs negatively contributes with a weight of 0.32
- Fish negatively contribute with a weight of 0.73
- Miso soup positively contribute with a weight of 0.27
- Pickles negatively contributes with a weight of 0.23

Throw this result it's possible to assume that eggs, fish and pickles are good candidates to reduce depression within people.

Inspecting the residuals (see figure 6) are a very important approach to understanding further details about the model: the residuals seem correctly normally distributed and leverage patients which negatively contributes to the prediction are patient 74, 176, 225.

Figure 6: Residuals

4.3.1 Measures of Association

```
             Outcome +    Outcome -      Total                 Inc risk *
Exposed +           32           11         43     74.42 (58.83 to 86.48)
Exposed -          157           40        197     79.70 (73.39 to 85.08)
Total              189           51        240     78.75 (73.03 to 83.75)

Point estimates and 95% CIs:
------------------------------------------------------------------------
Inc risk ratio                                  0.93 (0.77, 1.13)
Odds ratio                                      0.74 (0.34, 1.60)
Attrib risk in the exposed *                    -5.28 (-19.48, 8.92)
Attrib fraction in the exposed (%)             -7.09 (-29.36, 11.34)
Attrib risk in the population *                 -0.95 (-8.58, 6.69)
Attrib fraction in the population (%)          -1.20 (-4.51, 2.00)
------------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 0.587 Pr>chi2 = 0.443
Fisher exact test that OR = 1: Pr>chi2 = 0.419
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

Figure 7: Measures of association

The provided results shown in figure 7 are presented in a table format, displaying depressions related to exposure and non-exposure to healthy diet habits, along with their corresponding totals. The point estimates and 95% confidence intervals (CIs) for various measures of association, such as the incidence risk ratio and odds ratio, are also provided. Additionally, the table includes measures of attributable risk and attributable fraction both in the exposed population and the overall population.

- Incidence Risk Ratio: The incidence risk ratio of 0.93 (95% CI: 0.77, 1.13) suggests that there is a slight decrease in the risk of depression among individuals exposed to unhealthy diet habits compared to those not exposed. However, the confidence interval includes 1, indicating that the observed difference could be due to chance.

- Odds Ratio: The odds ratio of 0.74 (95% CI: 0.34, 1.60) indicates lower odds of depression among the exposed group compared to the non-exposed group. However, the wide confidence interval suggests substantial uncertainty in this estimate.

- Attributable Risk in the Exposed: The attributable risk in the exposed population is -5.28 (95% CI: -19.48, 8.92), which suggests a negative association between unhealthy diet habits and depression. However, the confidence interval includes 0, indicating that the observed difference could be due to random variation.

- Attributable Fraction in the Exposed: The attributable fraction in the exposed population is -7.09% (95% CI: -29.36, 11.34), indicating that a negative proportion of the depression can be attributed to the exposure. However, the confidence interval includes 0, suggesting the possibility of chance findings.

- Attributable Risk in the Population: The attributable risk in the overall population is -0.95 (95% CI: -8.58, 6.69), indicating a negligible difference in the outcome between the exposed and non-exposed populations. The confidence interval includes 0, suggesting no substantial impact of the exposure on the outcome at the population level.

- Attributable Fraction in the Population: The attributable fraction in the population is -1.20% (95% CI: -4.51, 2.00), indicating a minimal proportion of the outcome that can be attributed to exposure at the population level. The confidence interval includes 0, suggesting the absence of a significant association.

In summary, based on the provided results, there is limited evidence to suggest a significant association between unhealthy diet habits and depression. The point estimates and confidence intervals indicate that the observed differences could be due to chance, and no substantial impact of unhealthy diet habits on depression is detected at both the individual and population levels. Further research or a larger sample size may be necessary to obtain more conclusive findings.

4.4 Dietary Analysis

The analysis in this section aims to investigate the relationship between dietary habits and health outcomes such as disease severity of inflammation markers (C-Reactive Protein). This involves the analysis of dietary patterns, exploring associations between dietary habits and health outcomes.

The graph 8 shows that patients who eat lots of food included in the unhealthy criteria are at higher risk of inflammation, this leads to affirming that an healthy diet habit leads to lower inflammation mark.

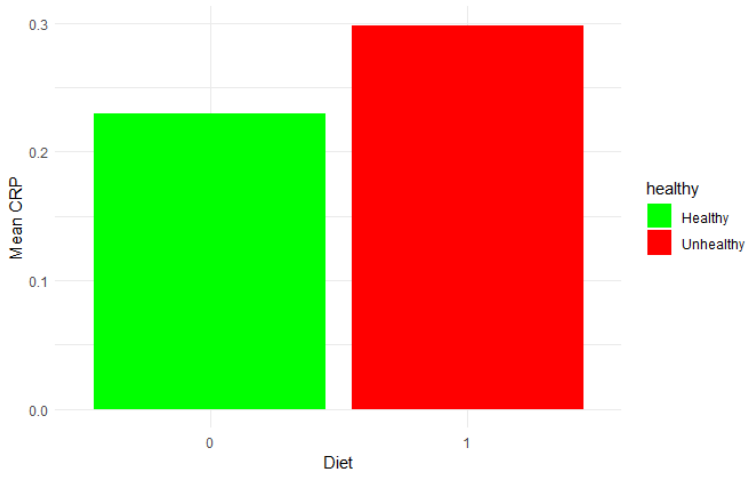

Figure 8: Measures of association

4.5 Mediation Analysis

In this part of the analysis, the aim was to investigate the possible existence of an indirect influence of healthy diet habits on the RA stage and depression score. The image 9 shows the results, based on 1000 simulations, of the nonparametric bootstrap confidence intervals with the percentile method:

- ACME (Average Causal Mediation Effect): The estimated effect of depression score on the outcome disability index through the RA stages is -0.000964. The

95% confidence interval suggests that the true effect could range from -0.015126 to 0.01. The p-value associated with the ACME is 0.93, indicating that the effect is not statistically significant.

- ADE (Average Direct Effect): The estimated direct effect of the RA stages on the disability index without considering depression as a mediator is -0.018696. The 95% confidence interval ranges from -0.055126 to 0.02. The p-value associated with the ADE is 0.31, indicating that the effect is not statistically significant.
- Total Effect: The estimated total effect of the RA stages on the disability index considering both the direct and indirect effects, is -0.019660. The 95% confidence interval ranges from -0.056088 to 0.02. The p-value associated with the Total Effect is 0.33, indicating that the effect is not statistically significant.
- Proportion Mediated: The proportion of the total effect that is mediated by the depression is estimated to be 0.049038. The 95% confidence interval ranges from -2.074409 to 1.98. The p-value associated with the Proportion Mediated is 0.85, indicating that the proportion is not statistically significant.

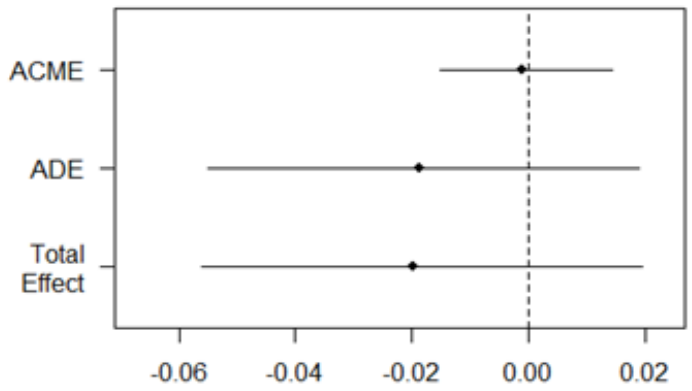

Figure 9: Mediation plot

4.6 Survival Analysis

The survival analysis has been performed by taking into consideration the duration, in terms of a year, of the rheumatoid arthritis stage to inspect the survival probabilities according to different factors. For this purpose, a Cox proportional hazards regression analysis has been applied with the following factors:

- VAS: The hazard ratio for a one-unit increase in VAS is 1.0129, indicating a slight increase in the hazard of the event. The coefficient is statistically significant at the 0.05 level (p = 0.042851).

- Disability Index: The hazard ratio for a one-unit increase in the disability index is 0.5360, suggesting a decrease in the hazard of the event. The coefficient is highly significant (p = 0.000813), indicating that the disability index has a strong impact on survival.
- Disease activity score, Biologics, MTX and Prednisolone drugs: These variables do not show a significant association with the hazard of the event based on their respective p-values.

In addition, an increase in VAS by one unit is associated with a 1.0129 times higher hazard of the event.

The concordance index (C-index) measures the predictive accuracy of the model, with a value of 0.658. The likelihood ratio test, Wald test, and score (log-rank) test assesses the overall significance of the model and indicates that it is statistically significant ($p < 0.05$).

Overall, the results suggest that VAS and disability index have significant associations with survival time, while the other variables in the model do not show significant associations.

5 Discussion and Conclusion

The dataset was composed by higher percentage of patients with arthritis and unhealthy diet habits compared to patients with high depression score. Further analysis can be conducted by balancing the three target variables in order to conduce a binary classification taking into consideration the proper factors.

Then, the p-values of food frequencies consumption in the linear regression and the percentage of measures of occurrences related to the depression score leads to consider as acceptable the thesis that a healthy dietary habits is able to reduce the feelings of depression and that unhealthy dietary habits can influence bad feelings of people.

In addition, unhealthy lifestyle can also leads to some other types of problems in addition to arthrites as the inflammation mark tends to be higher on people which assumes unhealthy food with higher frequency.

Finally, the RA stage also influence the depression of people, but the type of food consumption doesn't mediate in this relation by enforcing the results. This can lead to confirm that food and RA stage are strictly correlated to the depression score without intermediating.

6 Appendix

This report has been developed in R code, available at https://github.com/alessiasimone/epidemiology_depression

References

[Minamino et al., 2021] Minamino, H., Katsushima, M., Hashimoto, M., Fujita, Y., Torii, M., Ikeda, K., Isomura, N., Oguri, Y., Yamamoto, W., Watanabe, R., Murakami, K., Murata, K., Nishitani, K., Tanaka, M., Ito, H., Uda, M., Nin, K., Arai, H., Matsuda, S., Morinobu, A., and Inagaki, N. (2021). Influence of dietary habits on depression among patients with rheumatoid arthritis: A cross-sectional study using kurama cohort database. *PLOS ONE*, 16(8):1–15.