Exploring the Relationship Between β-carotene Levels, BCO1 Variants, and Atherosclerosis Progression through Data Analysis

Insights into Plasma Lipid Profiles and Immune Cell Dynamics for Cardiovascular Health

Mutsa Nyamuranga

2024-04-10

# 1. Summary/Abstract

# 2. Introduction

## 2.1 General Background Information

Atherosclerosis, a chronic inflammatory disease of the arterial wall, remains a leading cause of cardiovascular morbidity and mortality worldwide. Despite significant advancements in understanding its pathophysiology, effective therapeutic interventions are still in demand. Recent research has shed light on the potential role of dietary β-carotene and its conversion to vitamin A in modulating atherosclerosis progression and resolution.

Several studies have demonstrated the beneficial effects of β-carotene in delaying atherosclerosis progression in experimental models, along with its association with reduced cardiovascular risk in humans. Moreover, genetic variants in the BCO1 gene, responsible for converting β-carotene to vitamin A, have been linked to alterations in circulating carotenoid levels and plasma lipid profiles in individuals. However, the precise mechanisms underlying these effects, particularly regarding the interplay between β-carotene, BCO1 activity, and atherosclerosis progression or resolution, remain unclear.

## 2.2 Description of data and data source

The data I will be using is based on:

Pinos, I., et al. (2024). “β-Carotene accelerates the resolution of atherosclerosis in mice.” eLife 12: RP87430.

The data consists of measurements of circulating β-carotene and retinoic acid levels, along with relevant biomarkers and clinical characteristics, collected from a cohort of individuals with varying degrees of atherosclerosis progression and resolution. Genetic information on BCO1 activity, including genetic variants associated with BCO1 activity, may also be included in the dataset. The data was collected through blood sample analysis and genetic testing.

# Load necessary packages  
library(renv)

Attaching package: 'renv'

The following objects are masked from 'package:stats':  
  
 embed, update

The following objects are masked from 'package:utils':  
  
 history, upgrade

The following objects are masked from 'package:base':  
  
 autoload, load, remove

library(readxl) #for loading Excel files  
library(dplyr) #for data processing/cleaning

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(tidyr) #for data processing/cleaning  
library(skimr) #for nice visualization of data   
library(here) #to set paths  
  
# Load necessary package  
# path to data  
# note the use of the here() package and not absolute paths  
data\_location <- here::here("data/raw-data/elife-87430-data1-v1.xlsx")  
rawdata <- readxl::read\_excel(data\_location)

New names:  
• `` -> `...2`  
• `` -> `...3`  
• `` -> `...4`  
• `` -> `...5`  
• `` -> `...6`  
• `` -> `...7`  
• `` -> `...8`  
• `` -> `...9`  
• `` -> `...10`  
• `` -> `...11`  
• `` -> `...12`  
• `` -> `...13`  
• `` -> `...14`  
• `` -> `...15`  
• `` -> `...16`  
• `` -> `...17`  
• `` -> `...18`  
• `` -> `...19`  
• `` -> `...20`  
• `` -> `...21`  
• `` -> `...22`  
• `` -> `...23`  
• `` -> `...24`  
• `` -> `...25`  
• `` -> `...26`  
• `` -> `...27`  
• `` -> `...28`  
• `` -> `...29`  
• `` -> `...30`  
• `` -> `...31`  
• `` -> `...32`  
• `` -> `...33`  
• `` -> `...34`  
• `` -> `...35`  
• `` -> `...36`  
• `` -> `...37`  
• `` -> `...38`  
• `` -> `...39`

## 2.3 Questions/Hypotheses to be addressed

The question I want to answer using the data is whether there is a correlation between circulating β-carotene levels, BCO1 genetic variants, and the progression or resolution of atherosclerosis. Specifically, I am interested in examining how these factors relate to plasma lipid profiles (e.g., cholesterol, triglycerides). The main outcome of interest will be the degree of atherosclerosis progression or resolution, potentially measured by imaging techniques or clinical assessments of plaque burden. Additionally, plasma lipid profiles and immune cell populations, particularly Treg levels, would be of interest as secondary outcomes. The specific predictors of interest would include circulating β-carotene levels, genetic variants associated with BCO1 activity, plasma lipid profiles, and immune cell populations. The data analysis will likely involve statistical methods such as correlation analysis to assess the relationships between predictors and outcomes. Regression modeling could be used to examine the predictive value of circulating β-carotene levels and BCO1 genetic variants on atherosclerosis progression or resolution, controlling for potential confounding variables. Further exploration may involve visualization techniques to identify patterns and trends in the data.

# 3. Methods

## 3.1 Data Acquisition

The data was acquired online from Pinos et. al. in the additional files section. The data was formated in tables within an excel file which had to be processed and cleaned to identify target data. The excel file was loaded in through readxl and downloaded into a processing quarto file.

## 3.2 Data Import and Cleaning

The provided R code is structured to process and clean raw data extracted from an Excel file. Initially, the data is loaded into R using the read\_excel function, and a specific range of cells is extracted from the Excel sheet “Figure 1-suppl figure 1”. The subsequent cleaning steps involve removing columns containing standard deviation data and ensuring clarity and consistency by labeling conditions with “Female” and “Male” appropriately. Following this, basic exploration functions such as print, str, summary, and glimpse are utilized to examine the cleaned data frame’s structure and content. Finally, the cleaned data is saved as an RDS file for future analysis.

## 3.3 Statistical analysis

|  |
| --- |
| Condition Distrobution |

|  |
| --- |
| Condition Distrobution |

### 3.3.1 Time Series Regression Analysis

A time series regression analysis was conducted to investigate temporal trends and relationships between variables over time. The time series data were examined for stationarity and seasonality using appropriate statistical tests. The primary variables of interest were body weight over weeks. A linear regression model was fitted to the time series data to quantify the relationship between the predictor variables and the outcome variable over the study period. Model diagnostics, including residual analysis and goodness-of-fit measures, were performed to assess model adequacy.

### 3.3.2 Basic Linear Regression Analysis

A basic linear regression analysis was performed to examine the association between predictor variables and the outcome variable. The predictor variables included control, baseline and beta-carotene, while the outcome variable was body weight. Prior to analysis, multicollinearity diagnostics were conducted to identify and mitigate issues arising from collinear predictor variables. The linear regression model was fitted to the data, and model assumptions were assessed through residual analysis and goodness-of-fit tests.

### 3.3.3 Train/Test Linear Regression Analysis

To assess the generalizability and predictive performance of the linear regression model, a train/test approach was employed. The dataset was randomly divided into a training set (80% of the data) and a testing set (remaining 20 of the data). The linear regression model was fitted to the training set, and model performance was evaluated using the testing set. Performance metrics such as mean squared error (MSE) and R-squared were calculated to assess the predictive accuracy of the model.

### 3.3.4 Statistical Software

All statistical analyses were conducted using [name of statistical software or programming language, e.g., R, Python]. Statistical significance was determined at the conventional alpha level of 0.05.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

[Table 1](#tbl-summarytable) shows a summary of the data.

Table 1: Data summary table.

| skim\_type | skim\_variable | n\_missing | complete\_rate | character.min | character.max | character.empty | character.n\_unique | character.whitespace | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| character | BaselineF | 3 | 0.8421053 | 2 | 18 | 0 | 16 | 0 | NA | NA | NA | NA | NA | NA | NA | NA |
| character | ControlF | 0 | 1.0000000 | 2 | 18 | 0 | 19 | 0 | NA | NA | NA | NA | NA | NA | NA | NA |
| character | BetaCaroteneF | 0 | 1.0000000 | 2 | 18 | 0 | 19 | 0 | NA | NA | NA | NA | NA | NA | NA | NA |
| character | BaselineM | 3 | 0.8421053 | 4 | 18 | 0 | 16 | 0 | NA | NA | NA | NA | NA | NA | NA | NA |
| character | ControlM | 0 | 1.0000000 | 2 | 18 | 0 | 19 | 0 | NA | NA | NA | NA | NA | NA | NA | NA |
| character | BetaCaroM | 0 | 1.0000000 | 2 | 18 | 0 | 19 | 0 | NA | NA | NA | NA | NA | NA | NA | NA |
| numeric | Week | 0 | 1.0000000 | NA | NA | NA | NA | NA | 10 | 5.627314 | 1 | 5.5 | 10 | 14.5 | 19 | ▇▇▆▇▇ |

## 4.2 Basic statistical analysis

## 4.3 Full analysis

Linear model fit table.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 24.9531250 | 1.420566 | 17.5656258 | 0.0000000 |
| CategoryBaselineM | 8.6291675 | 2.008983 | 4.2952909 | 0.0000399 |
| CategoryBetaCaroM | 9.8465239 | 1.928051 | 5.1069829 | 0.0000015 |
| CategoryBetaCaroteneF | -0.3559324 | 1.928051 | -0.1846073 | 0.8539036 |
| CategoryControlF | 0.1689803 | 1.928051 | 0.0876430 | 0.9303322 |
| CategoryControlM | 9.3995066 | 1.928051 | 4.8751336 | 0.0000040 |

Linear Regression: The results reveal the estimated effects of different treatment categories on the response variable, body weight, along with their statistical significance. The intercept, estimated at 25.0, represents the baseline level of the response variable when all predictors are zero. Relative to a reference category, presumably “BaselineM”, coefficients associated with each treatment category provide insights into their effects on the response variable. For instance, “BaselineM” and “BetaCaroM” are associated with increases in the response variable by approximately 8.63 and 9.85 units, respectively, compared to the reference category. Conversely, “BetaCaroteneF”, “ControlF”, and “ControlM” exhibit smaller or even negative effects, but the statistical significance varies among these categories, with “BetaCaroteneF” showing a non-significant effect and others demonstrating statistically significant effects with low p-values.

Time Series model fit table.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 14.9633916 | 0.6518010 | 22.956992 | 0.0000000 |
| Week | 1.1046852 | 0.2519033 | 4.385354 | 0.0000289 |
| I(Week^2) | 0.0350423 | 0.0292789 | 1.196844 | 0.2342255 |
| I(Week^3) | -0.0023153 | 0.0009802 | -2.362000 | 0.0201334 |
| CategoryBaselineM | 8.6291675 | 0.4469748 | 19.305712 | 0.0000000 |
| CategoryBetaCaroM | 8.6330722 | 0.4325931 | 19.956567 | 0.0000000 |
| CategoryBetaCaroteneF | -1.5693841 | 0.4325931 | -3.627853 | 0.0004541 |
| CategoryControlF | -1.0444715 | 0.4325931 | -2.414443 | 0.0175957 |
| CategoryControlM | 8.1860549 | 0.4325931 | 18.923223 | 0.0000000 |

Time Series Model: The linear regression model was constructed to investigate the relationship between body weight and time (Week) while considering treatment variables. The model revealed several key findings. Firstly, the estimated intercept of 14.9634 suggests that the average body weight at the beginning of the observation period, when all other predictors are zero, is approximately 14.9634 units. Additionally, the coefficient for Week, estimated at 1.1047, indicates that body weight tends to increase by approximately 1.1047 units for each additional week, assuming a linear trend. However, the quadratic and cubic terms of Week were not statistically significant, suggesting that the relationship between body weight and time may not exhibit quadratic or cubic trends.

Furthermore, the model uncovered notable effects of treatment variables on body weight. Specifically, compared to the reference level “BaselineM”, various treatment categories exhibited significant impacts on body weight. For instance, “BetaCaroM” and “ControlM” categories were associated with increased body weight by 8.6331 and 8.1861 units, respectively, compared to “BaselineM”. Conversely, “BetaCaroteneF” and “ControlF” categories were associated with decreases in body weight by 1.5694 and 1.0445 units, respectively, compared to “BaselineM”. These findings collectively suggest that both time and treatment variables play crucial roles in influencing body weight variations over the observed period.

$pre  
$actions  
$actions$recipe  
$recipe

── Recipe ──────────────────────────────────────────────────────────────────────

── Inputs

Number of variables by role

outcome: 1  
predictor: 2

── Operations

• Dummy variables from: all\_nominal() and -all\_outcomes()

• Centering for: all\_predictors()

• Scaling for: all\_predictors()

$blueprint  
Recipe blueprint:   
   
# Predictors: 0   
 # Outcomes: 0   
 Intercept: FALSE   
Novel Levels: FALSE   
 Composition: tibble   
  
attr(,"class")  
[1] "action\_recipe" "action\_pre" "action"   
  
  
$mold  
$mold$predictors  
# A tibble: 89 × 6  
 Week Category\_BaselineM Category\_BetaCaroM Category\_BetaCaroteneF  
 <dbl> <dbl> <dbl> <dbl>  
 1 -1.65 -0.430 -0.430 -0.430  
 2 -1.65 -0.430 -0.430 2.30   
 3 -1.65 2.30 -0.430 -0.430  
 4 -1.65 -0.430 -0.430 -0.430  
 5 -1.47 -0.430 -0.430 -0.430  
 6 -1.47 -0.430 -0.430 -0.430  
 7 -1.47 -0.430 -0.430 2.30   
 8 -1.28 -0.430 -0.430 -0.430  
 9 -1.28 -0.430 -0.430 -0.430  
10 -1.10 -0.430 -0.430 -0.430  
# ℹ 79 more rows  
# ℹ 2 more variables: Category\_ControlF <dbl>, Category\_ControlM <dbl>  
  
$mold$outcomes  
# A tibble: 89 × 1  
 BodyWeight  
 <dbl>  
 1 16.6  
 2 17.5  
 3 21.7  
 4 22.9  
 5 17.8  
 6 18.1  
 7 18.4  
 8 18.3  
 9 19.2  
10 20   
# ℹ 79 more rows  
  
$mold$blueprint  
Recipe blueprint:   
   
# Predictors: 2   
 # Outcomes: 1   
 Intercept: FALSE   
Novel Levels: FALSE   
 Composition: tibble   
  
$mold$extras  
$mold$extras$roles  
NULL  
  
  
  
$case\_weights  
NULL  
  
attr(,"class")  
[1] "stage\_pre" "stage"   
  
$fit  
$actions  
$actions$model  
$spec  
Linear Regression Model Specification (regression)  
  
Computational engine: lm   
  
  
$formula  
NULL  
  
attr(,"class")  
[1] "action\_model" "action\_fit" "action"   
  
  
$fit  
parsnip model object  
  
  
Call:  
stats::lm(formula = ..y ~ ., data = data)  
  
Coefficients:  
 (Intercept) Week Category\_BaselineM   
 30.2850 5.5194 3.1501   
 Category\_BetaCaroM Category\_BetaCaroteneF Category\_ControlF   
 3.1304 -0.6204 -0.5052   
 Category\_ControlM   
 3.1990   
  
  
attr(,"class")  
[1] "stage\_fit" "stage"   
  
$post  
$actions  
named list()  
  
attr(,"class")  
[1] "stage\_post" "stage"   
  
$trained  
[1] TRUE  
  
attr(,"class")  
[1] "workflow"

The predictive model demonstrated favorable performance, with an RMSE of 1.564 and an R-squared value of 0.954. The low RMSE indicates minimal deviation between predicted and actual body weight values, suggesting the model’s predictive accuracy. Furthermore, the high R-squared value indicates that the model explains approximately 95.4% of the variance in body weight, indicating strong explanatory power. Interpretation of model coefficients revealed insights into the relative importance of predictors in influencing body weight.

The results highlight the effectiveness of the predictive model in capturing the relationship between predictors and body weight. The preprocessing steps, including dummy encoding, centering, and scaling, played a crucial role in enhancing model performance by standardizing and transforming predictors. The high explanatory power of the model suggests that the included predictors effectively explain variations in body weight.

# 5. Discussion

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 6. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>