# An Analysis of the Relationship Between Heart Disease and Osteoarthritis

*Inference through Propensity Scores*


**Matthew Berkowitz[1], Coco Liu [1], Barinder Thind[1], & Jiahao Tian[1]**

1. Simon Fraser University

**Contact Information:**

**Matthew Berkowitz** matthew_berkowitz@sfu.ca

**Barinder Thind** bthind@sfu.ca

**Coco Liu** sla214@sfu.ca

**Jiahao Tian** jtian_3@sfu.ca



**Abstract**

The goal of this analysis was to isolate the effect of osteoarthritis on the likelihood of developing cardiovascular disease. The provided data was characterized by mild sparsity, so the analysis was carried out on a complete-cases subset and on an imputed data set. The imputation was done through a combination of Logistic and Polytomous Bayesian Regression. There were a number of covariates that had the potential to confound this effect; in order to circumvent this problem, propensity scores were used [1] in conjunction with the LASSO [3]. More specifically, the propensity scores for all three cycles were computed for each of the two types of data sets and then another analysis was done involving the ensemble of propensity scores and the LASSO. No definitive causal relationship was uncovered although their was evidence of a statistically significant effect.


## Introduction

It is important to identify as many of the factors associated with any particular condition in an attempt to minimize or remove their confounding effect. The purpose of this analysis is to understand whether this kind of antibiotic relationship exists between osteoarthritis and heart disease; the approach is summarized as follows:

- We first explore the data set in an effort to understand the nature of the missing values
- An imputation is implemented on a large and relevant subset of the data
- A propensity score approach is used for inference and the results are provided
- The LASSO was used on half the data sets (for each cycle) for variable selection, after which propensity score analysis was used on the other half of the data sets to provide odds ratio estimates

## Data Description

- The data is filtered out according to study eligibility criteria so that the participants who are not 20-64 years of age, and the participants who were diagnosed as either *Rheumatoid Arthritis* or *Other* are excluded from the analysis
- There are a number of variables that include information on dietary habits, age, location, general health, marital status, substance use, blood pressure, stress levels, and income; in total, 23 covariates are used
- There are separate observations for three "cycles" which reference varying time periods and have about 130,000 observations each.

## Pre-Processing & Imputation

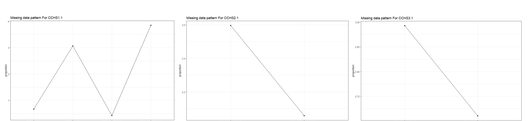
### Recoding

A number of variables were recoded, for example:

- The daily consumption of vegetables was transformed from the number of servings into a categorical variable
- The geographical variable was recoded so that a distinction was only made between the territories and the provinces
- The BMI variable was recoded so that it was not a raw number, but rather categorical in nature

### Polytomous Bayesian Regression for Imputation

The underlying assumption with the imputation implemented here is that the missing data is *missing at random (MAR)*. However, since the validity of this claim can easily be questioned. we decided to carry out the analysis on the complete-cases data set as well.

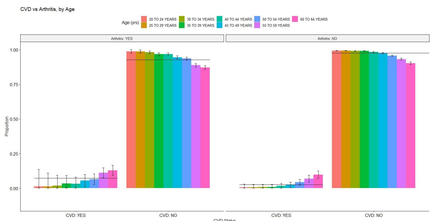



- The proportions of people who have osteoarthritis was calculated for each group. If the MCAR assumption holds, we would not expect to see any pattern of proportions associated with each group in the figure
- The plots shown indicate that the MCAR assumption does not hold in our data set, thus making analyzing the complete cases potentially inappropriate

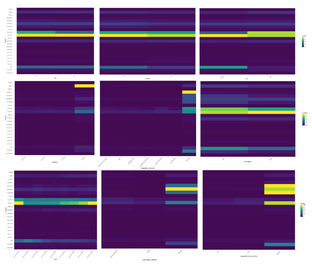
The imputation technique used here was Polytomous (Multinomial) Logistic Regression. This models how multinomial response variable $Y$ depends on a set of $m$ explanatory variables, $X$=$(X_1, X_2, ...X_m)$; in essence, this is a generalized linear model where the random component assumes that the distribution of $Y$ is multinom$(n, \pi)$, where $\pi$ is a vector with probabilities of "success" for each category [2]. Note that the responses here are all of the case where ordinality is not of particular importance.

## Exploratory Data Analysis

In this section, some plots of interest are presented which helped guide the decisions made later in the analysis.




- The above plot shows the relationship between osteoarthritis and cardiovascular disease (CVD), broken down by age. As can clearly be seen visually, age appears to be a significant predictor of CVD, while osteoarthritis status (yes / no) appears to matter much less.



- The variable with the most missing values is the total household income from all sources, and the second is physical activity index.

## Methodology

### Logistic Regression

The response variable here is dichotomous and hence, a logistic regression is the simple and obvious first approach to such a problem. The generalized linear model in this problem is defined as:

$$\Pr(\text{CCCA-121} = 1 \mid \theta) = \frac{\exp(\beta_0 + \beta_1 \text{CCCA-051} + \cdots + \beta_{21} \text{Province})}{1 + \exp(\beta_0 + \beta_1 \text{CCCA-051} + \cdots + \beta_{21} \text{Province})} \tag{1}$$

where $\theta$ is the set of parameters defining the model. This model is then expanded on with propensity scores.

### LASSO

The least absolute shrinkage and selection operator (LASSO) is a least squares technique that has the effect of minimizing the coefficients for particular covariates down to 0. It is mathematically defined as:

$$\hat{\beta}^{lasso} = \underset{\beta \in R^m}{argmin} ||y - \mathbf{X}\beta||_2^2 + \lambda \sum_{i=1}^{m} |\beta_i| \tag{2}$$

$$= \underset{\beta \in R^m}{argmin} \underbrace{||y - \mathbf{X}\beta||_2^2}_{\text{Loss}} + \lambda \underbrace{|\beta_i|_1}_{\text{Penalty}} \tag{3}$$

Here, this was used to lower the dimensionality of our covariates so that we could compare the results between the non-LASSO and LASSO propensity score approaches.

### Propensity Scores

A propensity score, $p(x_i = 1|\theta)$ is the conditional probability, for subject $i, (i = 1, ..., n)$, of being assigned to some particular treatment (in this case, 1) given some covariates, $\theta$. In other words:

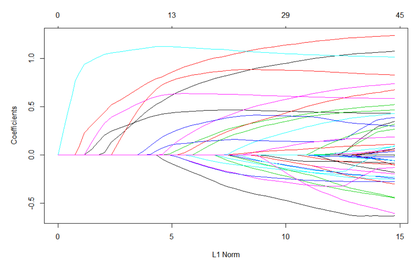
$$P(l) = P(x_i = 1|\theta) \tag{4}$$

Naturally, it follows that if a randomized design were used, we would expect the conditional probability here to be 0.5. For this case study, we applied as follows:

- We used Propensity Score Matching to estimate the effect of having osteoarthritis by accounting for an array of covariates: Propensity score, P(L)=P(O=1|L), where O is having osteoarthritis, and L are the covariates. This attempts to deconfound the effect of osteoarthritis on cardiovascular disease.
- The final model uses the propensity score matched sample and incorporates survey weights into a quasibinomial logistic regression. The matched sample sizes from each cycle, the OR estimates and associated CIs are summarized in the results tables.

## Results & Inference

### LASSO Path

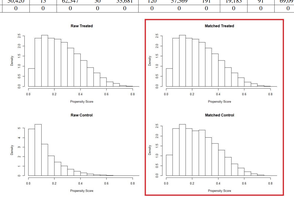



- The variables selected by this regularization were: age, sex, income, type of smoker, blood pressure status, diabetic status, emphysema or chronic obstructive pulmonary disease (COPD) status, self-perceived stress.

### Propensity Scores

The following tables/plots provide information on our results.

| SAMPLE SIZES | Cycle 1 - Complete Cases | | Cycle 1 - Imputed | | Cycle 2 - Complete Cases | | Cycle 2 - Imputed | | Cycle 3 - Complete Cases | | Cycle 3 - Imputed | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Control | Treated | Control | Treated | Control | Treated | Control | Treated | Control | Treated | Control | Treated |
| All | 61,403 | 10,996 | 75,639 | 13,322 | 46,754 | 11,193 | 72,065 | 14,687 | 25,227 | 6,135 | 83,316 | 14,393 |
| Matched | 10,983 | 10,983 | 13,292 | 13,292 | 11,073 | 11,073 | 14,496 | 14,496 | 6,044 | 6,044 | 14,225 | 14,225 |
| Unmatched | 50,420 | 13 | 62,347 | 30 | 35,681 | 120 | 57,569 | 191 | 19,183 | 91 | 69,091 | 168 |
| Discarded | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |



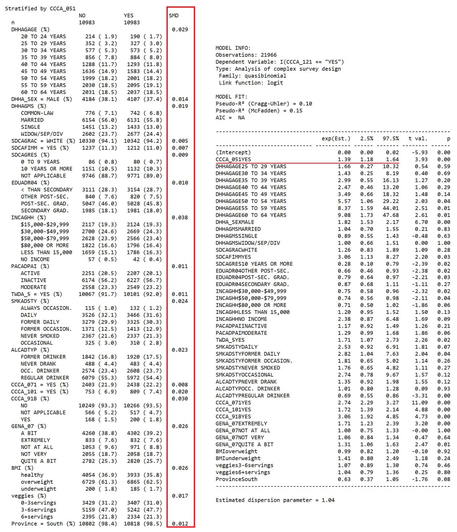

**Table/Figure:** Covariate balance in matched sample

| SUMMARY ESTIMATES (PSA) | Cycle 1 - Complete | Cycle 1 - Imputed | Cycle 2 - Complete | Cycle 2 - Imputed | Cycle 3 - Complete | Cycle 3 - Imputed |
|---|---|---|---|---|---|---|
| OR Estimate | 1.39 | 1.45 | 1.59 | 1.55 | 1.55 | 1.48 |
| 95% Confidence Interval - Lower Bound | 1.18 | 1.25 | 1.34 | 1.33 | 1.23 | 1.26 |
| 95% Confidence Interval - Upper Bound | 1.64 | 1.68 | 1.90 | 1.80 | 1.95 | 1.73 |

**Table 2:** Odds ratio results using propensity scores

| SUMMARY ESTIMATES (PSA / LASSO) | Cycle 1 - Lasso | Cycle 2 - Lasso | Cycle 3 - Lasso |
|---|---|---|---|
| OR Estimate | 1.49 | 1.70 | 1.67 |
| 95% Confidence Interval - Lower Bound | 1.18 | 1.40 | 1.43 |
| 95% Confidence Interval - Upper Bound | 1.88 | 2.07 | 1.96 |

**Table 3:** Propensity score OR results after the LASSO regularization



```
Stratified by CCCA_051
                                NO               YES              SMD
n                               10983            10983
DHHAGAGE (%)                                                      0.029
   20 TO 24 YEARS                214 ( 1.9)       190 ( 1.7)
   25 TO 29 YEARS                352 ( 3.2)       327 ( 3.0)
   30 TO 34 YEARS                577 ( 5.3)       573 ( 5.2)
   35 TO 39 YEARS                856 ( 7.8)       884 ( 8.0)
   40 TO 44 YEARS               1288 (11.7)      1293 (11.8)
   45 TO 49 YEARS               1636 (14.9)      1583 (14.4)
   50 TO 54 YEARS               1999 (18.2)      2001 (18.2)
   55 TO 59 YEARS               2030 (18.5)      2095 (19.1)
   60 TO 64 YEARS               2031 (18.5)      2037 (18.5)
DHH4_SEX = MALE (%)             4184 (38.1)      4207 (38.3)   0.014
DHHADMS (%)                                                       0.019
   COMMON-LAW                    776 ( 7.1)       742 ( 6.8)
   MARRIED                      6154 (56.0)      6133 (55.8)
   SINGLE                       1461 (13.2)      1433 (13.0)
   WIDOW/SEP/DIV                2682 (23.7)      2677 (24.4)
SDCAGRAC = WHITE (%)           10330 (94.1)     10342 (94.2)   0.005
SDCAFIMM = YES (%)              1237 (11.3)      1212 (11.0)   0.007
SDCAGRES (%)                                                      0.009
   0 TO 9 YEARS                   86 ( 0.8)        80 ( 0.7)
   10 YEARS OR MORE             1151 (10.5)      1132 (10.3)
   NOT APPLICABLE               9746 (88.7)      9771 (89.0)
EDUADR04 (%)                                                      0.010
   < THAN SECONDARY             3111 (28.3)      3154 (28.7)
   OTHER POST-SEC.               840 ( 7.6)       820 ( 7.5)
   POST-SEC. GRAD.              5047 (46.0)      5028 (45.8)
   SECONDARY GRAD.              1985 (18.1)      1981 (18.0)
INCAGHH (%)                                                       0.038
   $15,000-$29,999              2117 (19.3)      2124 (19.3)
   $30,000-$49,999              2700 (24.6)      2660 (24.2)
   $50,000-$79,999              2628 (23.9)      2566 (23.4)
   $80,000 OR MORE              1822 (16.6)      1796 (16.4)
   LESS THAN 15,000             1659 (15.1)      1786 (16.3)
   NO INCOME                      57 ( 0.5)        42 ( 0.4)
PACADPAI (%)                                                      0.011
   ACTIVE                       2251 (20.5)      2207 (20.1)
   INACTIVE                     6174 (56.2)      6227 (56.7)
   MODERATE                     2558 (23.3)      2549 (23.2)
TWDA_5 = YES (%)               10067 (91.7)     10101 (92.0)   0.011
SMKDSTY (%)                                                       0.024
   ALWAYS OCCASION.              115 ( 1.0)       132 ( 1.2)
   DAILY                        3526 (32.1)      3466 (31.6)
   FORMER DAILY                 3279 (29.9)      3325 (30.3)
   FORMER OCCASION.             1371 (12.5)      1413 (12.9)
   NEVER SMOKED                 2367 (21.6)      2337 (21.3)
   OCCASIONAL                    325 ( 3.0)       310 ( 2.8)
ALCDTYP (%)                                                       0.023
   FORMER DRINKER               1842 (16.8)      1920 (17.5)
   NEVER DRANK                   488 ( 4.4)       483 ( 4.4)
   OCC. DRINKER                 2574 (23.4)      2608 (23.7)
   REGULAR DRINKER              6079 (55.3)      5972 (54.4)
CCCA_071 = YES (%)             2403 (21.9)      2438 (22.2)   0.008
CCCA_101 = YES (%)              753 ( 6.9)       809 ( 7.4)   0.020
CCCA_91B (%)                                                      0.030
   NO                          10249 (93.3)     10266 (93.5)
   NOT APPLICABLE                566 ( 5.2)       517 ( 4.7)
   YES                           168 ( 1.5)       200 ( 1.8)
GENA_07 (%)                                                       0.026
   A BIT                        4260 (38.8)      4302 (39.2)
   EXTREMELY                     833 ( 7.6)       832 ( 7.6)
   NOT AT ALL                   1053 ( 9.6)       971 ( 8.8)
   NOT VERY                     2055 (18.7)      2058 (18.7)
   QUITE A BIT                  2782 (25.3)      2820 (25.7)
BMI (%)                                                           0.026
   healthy                      4054 (36.9)      3933 (35.8)
   overweight                   6729 (61.3)      6865 (62.5)
   underweight                   200 ( 1.8)       185 ( 1.7)
veggies (%)                                                       0.017
   0-3servings                  3429 (31.2)      3407 (31.0)
   3-6servings                  5159 (47.0)      5242 (47.7)
   6+servings                   2395 (21.8)      2334 (21.3)
Province = South (%)           10802 (98.4)     10818 (98.5)   0.012
```

```
MODEL INFO:
Observations: 21966
Dependent Variable: I(CCCA_121 == "YES")
Type: Analysis of complex survey design
      Family: quasibinomial
      Link function: logit

MODEL FIT:
Pseudo-R² (Cragg-Uhler) = 0.10
Pseudo-R² (McFadden) = 0.15
AIC = NA

-------------------------------------------------------------------------
                                exp(Est.)   2.5%   97.5%   t val.      p
--------------------------- ------------ ------ ------- -------- -------
(Intercept)                          0.00   0.00    0.02    -5.93   0.00
CCCA_051YES                          1.39   1.18    1.64     3.93   0.00
DHHAGAGE25 TO 29 YEARS               1.66   0.27   10.32     0.54   0.59
DHHAGAGE30 TO 34 YEARS               1.43   0.25    8.19     0.40   0.69
DHHAGAGE35 TO 39 YEARS               2.99   0.55   16.13     1.27   0.20
DHHAGAGE40 TO 44 YEARS               2.47   0.46   13.20     1.06   0.29
DHHAGAGE45 TO 49 YEARS               3.49   0.66   18.32     1.48   0.14
DHHAGAGE50 TO 54 YEARS               5.57   1.06   29.22     2.03   0.04
DHHAGAGE55 TO 59 YEARS               8.37   1.59   44.01     2.51   0.01
DHHAGAGE60 TO 64 YEARS               9.08   1.73   47.68     2.61   0.01
DHH4_SEXMALE                         1.82   1.53    2.17     6.70   0.00
DHHADMSMARRIED                       1.04   0.70    1.55     0.21   0.83
DHHADMSSINGLE                        0.89   0.55    1.43    -0.48   0.63
DHHADMSWIDOW/SEP/DIV                 1.00   0.66    1.51     0.00   1.00
SDCAGRACWHITE                        1.26   0.83    1.89     1.09   0.28
SDCAFIMYES                           1.06   1.13    8.27     2.20   0.03
SDCAGRES10 YEARS OR MORE             0.28   0.10    0.79    -2.39   0.02
EDUADR04OTHER POST-SEC.              0.66   0.46    0.93    -2.38   0.02
EDUADR04POST-SEC. GRAD.              0.79   0.64    0.97    -2.21   0.03
EDUADR04SECONDARY GRAD.              0.87   0.68    1.11    -1.11   0.27
INCAGHH$30,000-$49,999               0.75   0.58    0.96    -2.32   0.02
INCAGHH$50,000-$79,999               0.74   0.56    0.98    -2.11   0.04
INCAGHH$80,000 OR MORE               0.71   0.50    1.02    -1.86   0.06
INCAGHHLESS THAN 15,000              1.20   0.95    1.52     1.50   0.13
INCAGHHNO INCOME                     2.38   0.87    6.48     1.69   0.09
PACADPAINACTIVE                      1.17   0.92    1.49     1.26   0.21
PACADPAIMODERATE                     1.29   0.99    1.68     1.86   0.06
TWDA_5YES                            1.71   1.07    2.73     2.26   0.02
SMKDSTYDAILY                         2.53   0.92    6.91     1.81   0.07
SMKDSTYFORMER DAILY                  2.82   1.04    7.63     2.04   0.04
SMKDSTYFORMER OCCASION.              1.81   0.65    5.02     1.34   0.26
SMKDSTYNEVER SMOKED                  1.76   0.65    4.82     1.11   0.27
SMKDSTYOCCASIONAL                    2.74   0.78    9.67     1.57   0.12
ALCDTYPNEVER DRANK                   1.35   0.92    1.98     1.55   0.12
ALCDTYPOCC. DRINKER                  1.01   0.80    1.28     0.09   0.93
ALCDTYPREGULAR DRINKER               0.69   0.55    0.86    -3.31   0.00
CCCA_071YES                          2.74   2.29    3.27    11.09   0.00
CCCA_101YES                          1.72   1.39    2.14     4.88   0.00
CCCA_91BYES                          3.06   1.92    4.85     4.73   0.00
GENA_07EXTREMELY                     1.71   1.23    2.39     3.20   0.00
GENA_07NOT AT ALL                    1.00   0.75    1.33    -0.00   1.00
GENA_07NOT VERY                      1.06   0.84    1.34     0.47   0.64
GENA_07QUITE A BIT                   1.31   1.06    1.63     2.47   0.01
BMIOverweight                        0.99   0.82    1.20    -0.10   0.92
BMIUnderweight                       1.41   0.80    2.49     1.18   0.24
veggies3-6servings                   1.07   0.89    1.30     0.74   0.46
veggies6+servings                    1.04   0.79    1.36     0.25   0.80
ProvinceSouth                        0.63   0.37    1.05    -1.76   0.08
-------------------------------------------------------------------------

Estimated dispersion parameter = 1.04
```

**Figure:** Regression results using propensity scores

## Conclusions & Future Considerations

The goal of this analysis was to isolate the effect of osteoarthritis on cardiovascular disease. We did the analysis on two sets of data and used an ensemble of the LASSO and propensity scores. Through this, our OR estimates of the effect of osteoarthritis on cardiovascular disease are all similar, with the LASSO-derived estimates producing greater uncertainty, as seen with the wider confidence intervals. There appears to be a statistically significant impact of osteoarthritis on the risk of cardiovascular disease; however, it is doubtful that the effect is causative. Further research is needed to study the mediating variables responsible for this relationship.

## References


[1] Eshan Karim. Ps-survey workshop (participant).

[2] Penn State. 8.1 - polytomous (multinomial) logistic regression.

[3] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc.*, 58:267–288, 1996.


## Acknowledgements


Thanks to Joan Hu and Rachel Altman for their helpful comments. Thanks to SFU's department of Statistics & Actuarial Science for their support.