Data Analysis Report

Project: P1363Demourelle  
PI: Kristen Demourelle  
Prepared By: Jaron Arbet  
Date: 09/25/2018

Table of Contents

[1. Exploratory Data Analysis 1](#_Toc525734573)

[1.1 Quantify Missing Data 1](#_Toc525734574)

[1.2 Demographic Covariates 2](#_Toc525734575)

[1.3 anti-CCP Outcome 3](#_Toc525734576)

[1.4 Cyotokine Exposure Variables 4](#_Toc525734577)

[Density Plots 4](#_Toc525734578)

[Summary Statistics 5](#_Toc525734579)

[Correlations 6](#_Toc525734580)

[Composite PCA Cytokine Summary measure 6](#_Toc525734581)

[1.5 RF-IgA Exposure Variable 8](#_Toc525734582)

[1.6 NET Mediator Variables 9](#_Toc525734583)

[1.7 Data Transformations? 10](#_Toc525734584)

[2. Simple Linear Regression Analysis 10](#_Toc525734585)

[2.1 Residual Plots to Check for Outliers 11](#_Toc525734586)

[3. References 15](#_Toc525734587)

# Exploratory Data Analysis

## Quantify Missing Data

Number and Proportion of Subjects missing data for each variable

|  |  |  |
| --- | --- | --- |
| Variable | Num\_Miss | Prop\_Miss |
| ID | 0 | 0.00 |
| labID | 0 | 0.00 |
| Sputum\_CCP\_level | 4 | 0.09 |
| Sputum\_CCP\_outcome | 4 | 0.09 |
| Ln\_Sputum\_DNA\_HNE | 12 | 0.26 |
| Ln\_Sputum\_DNA\_MPO | 12 | 0.26 |
| Ln\_Sputum\_DNA\_CitH3 | 12 | 0.26 |
| Sputum\_cytokine\_TNFa\_level | 12 | 0.26 |
| Sputum\_cytokine\_MCP1\_level | 12 | 0.26 |
| Sputum\_cytokine\_IL10\_level | 12 | 0.26 |
| Sputum\_cytokine\_IL1B\_level | 12 | 0.26 |
| Sputum\_cytokine\_IL6\_level | 12 | 0.26 |
| Sputum\_cytokine\_IL8\_level | 12 | 0.26 |
| Sputum\_cytokine\_MIP1a\_level | 12 | 0.26 |
| Sputum\_cytokine\_MIP1b\_level | 12 | 0.26 |
| Sputum\_RF\_IgA\_outcome | 9 | 0.20 |
| Sp\_RFIgA\_level | 9 | 0.20 |
| Age | 0 | 0.00 |
| Gender\_female1\_male0 | 0 | 0.00 |
| EverSmoke\_yes\_1\_no\_0 | 0 | 0.00 |
| RA\_genetic\_risk\_allele\_Shared\_Epitope\_yes\_1\_no\_0 | 9 | 0.20 |

Number of Subjects with no missing exposure, mediator, or outcome data (i.e. excluding covariates):

## 34

Number of Subjects with no missing data (including covariates):

## 31

The 3 additional subjects with missing covariate data are solely due to the “RA\_genetic\_risk\_allele” covariate.

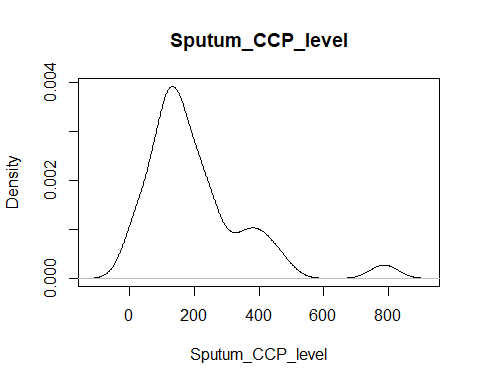
For the remainder of this report, I will only use the 34 subjects who have no missing exposure, mediator, or outcome data.

## Demographic Covariates

Mean(SD) for continuous variables; Counts (%) for categorical variables

|  |  |  |
| --- | --- | --- |
| Variable |  | Summary |
| Age (mean (sd)) |  | 60.41 (13.38) |
| Gender\_female1\_male0 (%) | 0 | 10 (29.4) |
|  | 1 | 24 (70.6) |
| EverSmoke\_yes\_1\_no\_0 (%) | 0 | 24 (70.6) |
|  | 1 | 10 (29.4) |
| RA\_genetic\_risk\_allele | neg | 14 (45.2) |
|  | Pos | 17 (54.8) |

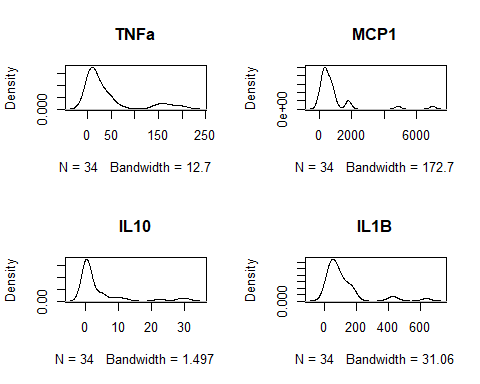
## anti-CCP Outcome

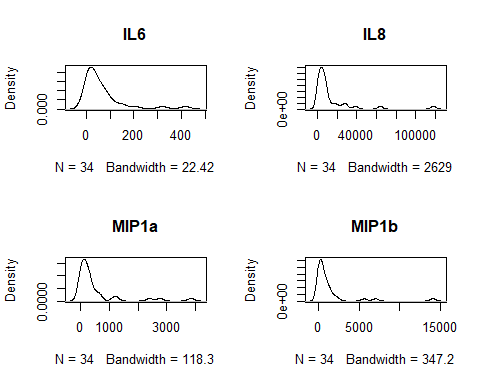


|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Min. | Q1 | Median | Mean | Q3 | Max | SD |
| 14.9 | 113.1 | 160.3 | 202.5 | 243.6 | 786.1 | 155.7 |

## Cyotokine Exposure Variables

### Density Plots



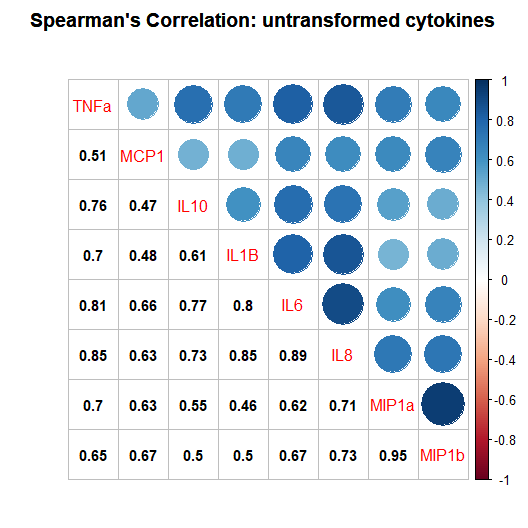


### Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cytokine | Min. | 1st Qt. | Median | Mean | 3rd Qt. | Max | Std. Dev. |
| TNFa | 1.9 | 5.2 | 18.4 | 41.6 | 43.5 | 202.8 | 57.3 |
| MCP1 | 0.1 | 293.0 | 499.7 | 909.4 | 813.6 | 6990.2 | 1379.9 |
| IL10 | 0.1 | 0.3 | 0.6 | 4.3 | 4.8 | 30.7 | 8.0 |
| IL1B | 6.0 | 45.7 | 74.8 | 116.9 | 139.3 | 635.6 | 133.0 |
| IL6 | 1.6 | 14.0 | 38.0 | 67.8 | 81.6 | 415.7 | 90.6 |
| IL8 | 712.3 | 2489.7 | 6158.9 | 13400.0 | 10412.9 | 118110.6 | 22638.3 |
| MIP1a | 3.6 | 76.5 | 230.1 | 508.9 | 433.1 | 3838.7 | 860.0 |
| MIP1b | 0.5 | 80.7 | 467.2 | 1301.7 | 1127.1 | 14110.3 | 2717.8 |

Notice the cytokines are on vastly different scales. We should standardize the cytokines to have mean 0 and variance of 1 during all analyzes, this way they will be on the same scale so their effect sizes will be more comparable.

### Correlations



**Notes on Correlation among Cytokines**

* all pairwise correlations have pvalues<0.01
* Decided to use Spearman’s correlation instead of Pearsons because Spearmans is less sensative to outliers and can account for non-linearity. Also, Spearman’s correlation gives identical results for both untransformed and log-transformed data since it is “invariant under monotone transformations.” Thus there is no reason to log-transform when using Spearman’s correlation.
* plot created using *corrplot* R package [1]

### Composite PCA Cytokine Summary measure

When creating a single “composite variable” to represent multiple variables, Song et al 2013 [2] recommend standardizing all variables to have mean of zero and variance of one, then either take a “simple average” of all variables, or use the first principle component from PCA (principle component analysis). The “weights” in PCA are chosen to to “account for as much variability in the original variables as possible.” I think PCA seems more reasonable compared to taking a “simple average.”

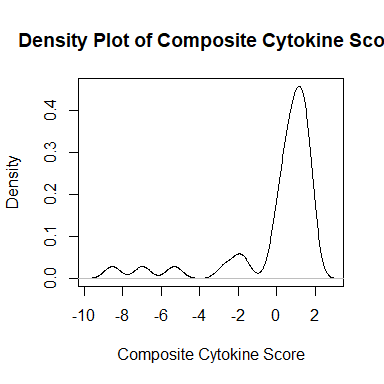
## Importance of first k=5 (out of 8) components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 2.4537 0.9036 0.82202 0.56443 0.27100  
## Proportion of Variance 0.7526 0.1021 0.08446 0.03982 0.00918  
## Cumulative Proportion 0.7526 0.8546 0.93908 0.97890 0.98808

Notice the first principal component explains 75% of the total variance of all 8 cytokine variables, thus it should be a sufficient “composite-measure” of all the cytokines. Lastly, note that if we log-transform the cytokine variables, the first PC only explains 65% of the total variance.

PC1: weights assigned to each cytokine

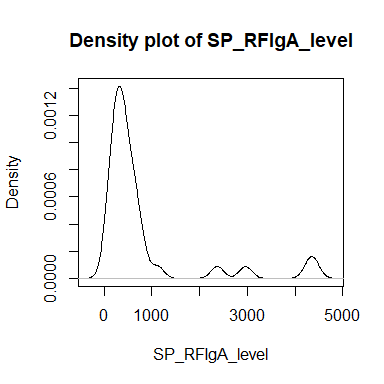
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TNFa | MCP1 | IL10 | IL1B | IL6 | IL8 | MIP1a | MIP1b |
| -0.38 | -0.3 | -0.35 | -0.34 | -0.38 | -0.34 | -0.38 | -0.36 |

Notice nearly all of the cytokines are given equal weight ().



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | SD |
| -8.5 | 0.1 | 0.8 | 0 | 1.4 | 1.7 | 2.5 |

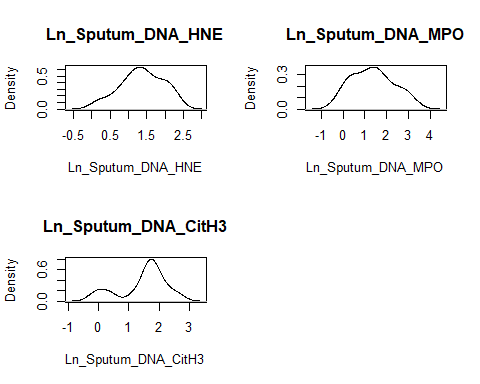
## RF-IgA Exposure Variable



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | SD |
| 77 | 245 | 402 | 774 | 652 | 4405 | 1080 |

## NET Mediator Variables

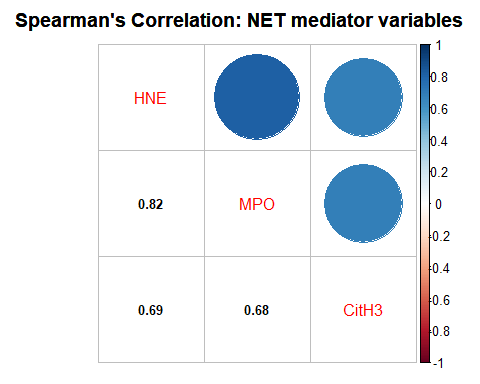
#### Density Plots



#### Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| NET Variable | Min. | 1st Qt. | Median | Mean | 3rd Qt. | Max | Std. Dev. |
| Ln\_Sputum\_DNA\_HNE | 0.2 | 1.0 | 1.4 | 1.4 | 1.8 | 2.2 | 0.57 |
| Ln\_Sputum\_DNA\_MPO | -0.2 | 0.5 | 1.3 | 1.4 | 1.8 | 3.2 | 0.95 |
| Ln\_Sputum\_DNA\_CitH3 | -0.2 | 1.2 | 1.7 | 1.5 | 1.9 | 2.7 | 0.78 |

#### Correlations



All pairwise correlation pvalues<0.01.

## Data Transformations?

* **sputum\_ccp\_level**: I do not think we need to log-transform, it would complicate interpretation and is unnecessary: we do not need the outcome to be normally distributed, we can use methods that allow for non-normal residuals and outliers.
  + what units is sputum\_ccp\_level measured on? Perhaps we could divide by 10 or 100 to put on a smaller scale. On it’s current large-scale, some of the regression coefficients are >100.
* **cytokines**: When not log-transforming, the first principal compoment accounts for 75% of the total variance among all standardized cytokine variables. In contrast, when log-transforming, the first PC only accounts for 65%. I do think we should standardize all cytokines to have mean 0 and variance of 1. This will put all the cytokines on the same scale and make the effect sizes of interest more comparable.
* **RFIGa**: standardize to mean 0 and variance of 1 (makes effect sizes comparable with cytokines)
* **NET variables**: is it necessary to log-transform?

# Simple Linear Regression Analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | Outcomes |  |
| Predictor | CCP\_level | HNE | MPO | CitH3 |
| composite\_cytokine\_score | **-37.16** (9.09) | **-0.13** (0.03) | **-0.29** (0.04) | **-0.16** (0.05) |
| TNFa | **105.31** (20.26) | **0.33** (0.08) | **0.70** (0.11) | **0.39** (0.12) |
| MCP1 | 25.32 (27.15) | 0.20 (0.10) | **0.51** (0.14) | 0.27 (0.13) |
| IL10 | **108.08** (19.80) | **0.29** (0.09) | **0.61** (0.13) | **0.34** (0.12) |
| IL1B | 57.08 (25.60) | **0.34** (0.08) | **0.70** (0.11) | **0.37** (0.12) |
| IL6 | **66.41** (24.89) | **0.32** (0.08) | **0.70** (0.11) | **0.43** (0.12) |
| IL8 | 49.05 (26.12) | **0.31** (0.09) | **0.64** (0.12) | **0.37** (0.12) |
| MIP1a | **103.94** (20.49) | **0.25** (0.09) | **0.59** (0.13) | **0.33** (0.13) |
| MIP1b | **106.45** (20.08) | **0.22** (0.09) | **0.55** (0.14) | 0.29 (0.13) |
| Sp\_RFIgA\_level | **99.41** (21.18) | 0.11 (0.10) | 0.21 (0.16) | 0.28 (0.13) |
| Age | -0.97 (2.05) | 0.00 (0.01) | 0.01 (0.01) | -0.00 (0.01) |
| Gender\_female1\_male0 | 41.05 (59.05) | -0.23 (0.22) | -0.18 (0.36) | -0.12 (0.30) |
| EverSmoke\_yes\_1\_no\_0 | 12.47 (59.46) | **0.47** (0.20) | 0.68 (0.34) | 0.51 (0.28) |
| RA\_genetic\_risk\_allele | 93.52 (53.37) | 0.13 (0.21) | 0.33 (0.35) | 0.55 (0.27) |

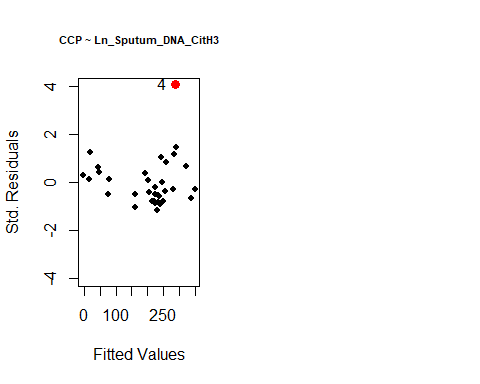
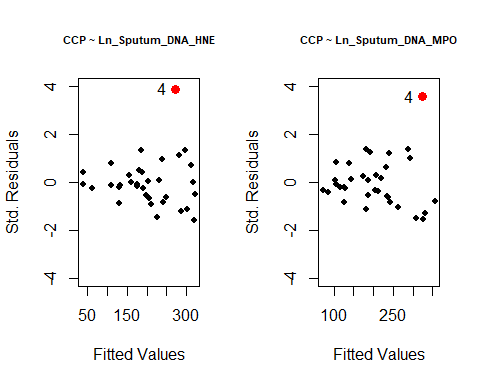
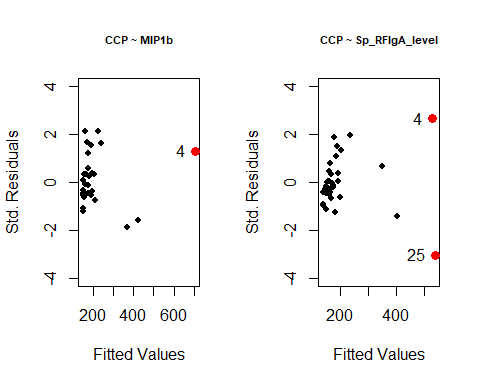
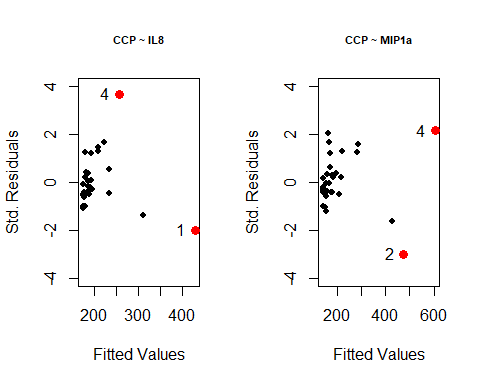
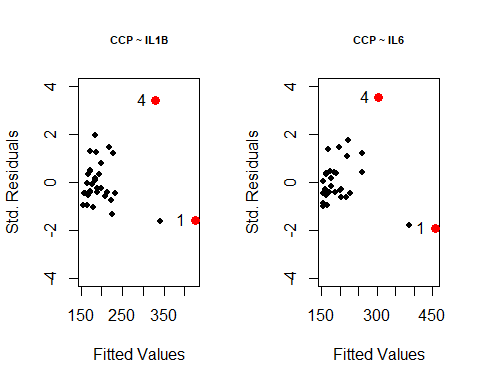
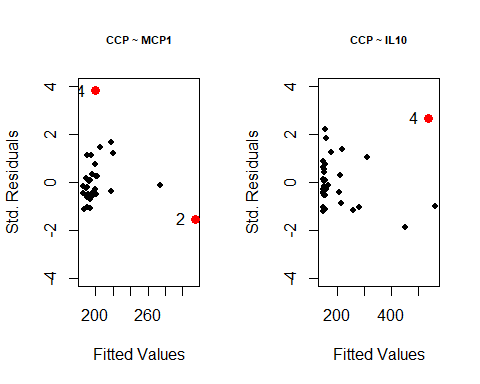
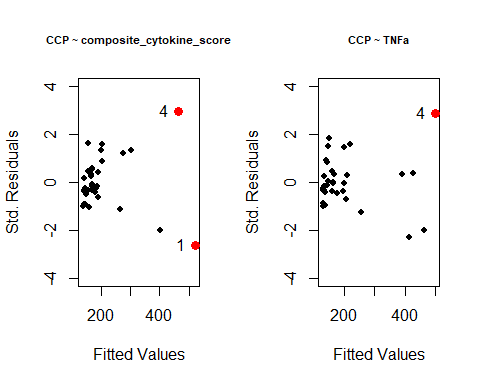
**Bold**: significant while controlling false-discovery-rate (FDR) at level 0.05

#### Notes

* None of the covariates (age, gender, smoke, RA\_genetic\_risk) are significantly associated with the outcome or NET mediators
* What units is Sputum\_CCP\_level measured in? Perhaps we can put on a smaller scale that is easier to interpret (e.g. divide by 10 or 100)

## Residual Plots to Check for Outliers

Subjects in Red are considered potential outliers (Cook’s Distance or std. residual )



Lab Id’s for potential outlier subjects 1,2,4,25:

## [1] 117413 116870 116743 117492

# References

[1] Wei, T. and Simko, V. (2017). R package “corrplot”: Visualization of a Correlation Matrix. Accessed: <https://github.com/taiyun/corrplot>.

[2] Song, M.-K., Lin, F.-C., Ward, S. E. and Fine, J. P. (2013). Composite variables: When and how. *Nursing research* **62** 45.