ABSTRACT

Metabolic syndrome is a cluster of conditions which when present increases one's risk for heart disease, stroke and diabetes and presents a major economic burden on the healthcare system. Utilising publish patient data we aim to explore visually the relationship between risk factors and how they relate to patients with metabolic syndrome and those without metabolic syndrome. Understanding the risk factors and implementing preventative medicine through educating patients is one way to reduce the burden on the economic costs associated with metabolic syndrome. In order to understand the data various methodologies were applied, the risk factors were grouped and summarised to compare the difference between median values, secondly the risk factor underwent a hypothesis testing and the distribution was presented visually after finding a significant relationship between the risk factors, and finally patient data was manipulated and arranged in a way to distinguish based on metabolic syndrome status and a dissimilarity matrix was produced. Overall, it appears that the risk factors present in the patient data are all statistically significantly different. The dissimilarity matrix suggests we are also able to discriminate dissimilarity based on metabolic syndrome status. Future studies could benefit at utilising machine learning to predict one's risk of metabolic syndrome as a preventative method, reducing the economic burden of this condition, and in addition explore non risk factors.

INTRODUCTION

Metabolic syndrome is a cluster of conditions which when present increase the risk of heart disease, stroke, and diabetes through the development of atherosclerosis and insulin resistance. The diagnosis of metabolic syndrome requires that a patient presents with three risk factors out of five risk factors. The risk factors for diagnosis are (1) elevated waist circumference, (2) elevated triglyceride levels (or drug treatment for elevated triglycerides), (3) reduce high density lipoprotein (HDL), elevated blood pressure (or drug treatment for hypertension) and (5) elevated fasting blood glucose (or treatment for elevated glucose), with the diagnostic criteria being slightly different for each gender and race due to variations in physiology (Rochlani et al., 2017). Being able to better detect, treat and prevent metabolic syndrome through proactive measurers such is a valid method to reduce the economic burden this condition has on the healthcare system (Scholze et el., 2010).

This paper aims to explore patient data of those with a positive diagnosis of metabolic syndrome and those diagnosed without metabolic syndrome to look at the trends between these two groups and if any similarities exist between patients through looking at a variety of risk factors and their corresponding values.

DATA METHOD

Patient data was sourced externally online from https://data.world/. It was sourced by searching on the following term 'metabolic syndrome'. Patient data was supplied by the user **Robert Hoyt, MD**. The data was subsequently downloaded and saved into a specified folder and later imported into R studio for data processing. This data was collected and published to https://data.world/ through SQL query sourced from the Centers for Disease Control and Prevention National Centre for Health Statistics published data, no further information could be found pretraining to this data set and how it was initially collected prior to the SQL query.

In preparation for the data to be analysed a simple function (equation 1) was created in R studio to identify the proportion of NA values in each variable.

$$naprop \leftarrow function(x) \sum (is.na(x))/\sum ¿$$

Equation 1: Used to calculate the proportion of NA values in the original data set.

This function (equation 1) in conjunction with apply() identified four variables with missing observations. Two different types of imputations were performed in the following order by creating a new data frame at each step, preserving the original for comparison:

- **Categorical variable imputation:** Marital statues was missing in 9.48% of all observations, it was deemed inappropriate based on the data to attempt to predict this value. Instead all NA values where imputed with a new observation category of 'unknown' utilising replace_na().

- **Imputation of mixed data set:** Due to multiple variables requiring an imputation across a mixed-type data frame a k-NN imputation was deployed utilising Gower's distance with a k value equal to 4 calling on knn() of the ***VIM*** package. This method has been used in clinical research, though a multiple method imputation is preferred (Stevens et al., 2016).

Boxplots were produced for comparison between the imputed variables in the new data frame against the original data frame to see how the shape and distribution of the data had changed, with no major changes occurring calling on the boxplot() function.

The final stage in the data processing step was the coercion of all variables as summarised in table 1 by use of the ***dplr*** package and the mutate() function. The subsequent coercion function is summarised in column two of table 1. This was required due to some functions in the preliminary stage of data exploration encountering errors in R Studio due to the nature of the variables and how they were originally imported.

The complete data set after pre-processing comprised of 2401 observations and 15 variables as summarised in table 1. A subsequent data set was created by taking a random sample of 500 observations without replacement using sample_n() of the ***dplr*** package, this allowed for a smaller sample for visual analysis which requires substantial computing power, the number of variables remained unchanged.

Table 1: Variables of interest post pre-processing.

Variable Name	Datatype/R Coercion	**Other Information**
Seqn	Integer/*as.integer()*	Unique identifier of patient.
Age	Integer/*as.integer()*	
Sex	Factor/*as.factor()*	
Marital	Factor/*as.factor()*	
Income	Integer	
Race	Factor/*as.factor()*	
WaistCirc	Numeric/*as.numeric()*	Waist circumference in CM
BMI	Numeric/*as.numeric()*	
Albuminuria	Integer/*as.integer()*	Indicator of kidney diseases
UrAlbCr	Numeric/*as.numeric()*	
UricAcid	Numeric/*as.numeric()*	Indicator of metabolic syndrome when elevated.
BloodGlucose	Integer/*as.integer()*	Indicator of diabetes when elevate.
HDL	Integer/*as.integer()*	High-density lipoproteins, low levels associated with metabolic syndrome
Triglycerides (Tri)	Integer/*as.integer()*	Elevated levels associated with metabolic syndrome
MetabolicSyndrome	Factor/*as.factor()*	

METHOD

Three analytical tests were conducted on either the complete data set or the sample data set depending on computational demands. The analysis done include, (1) summary of the risk factors by looking at the median value between metabolic syndrome statues and sex, (2) hypothesis testing of the significance of the risk factors and plotting their distribution, and (3) arranging the data in a way which allows to discriminate metabolic syndrome statues between patients and producing a (di)similarity matrix and presenting this visually. All there test were conducted with

R and Rstudio(version 4.0.4) on a x86_64-w64-mingw32 platform. Additional data processing steps are conducted as required in addition to the data section.

Risk factor summary

As confirmed in the exploratory analysis stage many of the observations and risk factors are associated with outliers due to the variation in normally physiology. A pivot table was generated to look at the median values of the diagnostic risk factors present in the complete data set. This was generated utilising the ***dplr*** package. First the data was grouped by metabolic syndrome statues and sex using the group_by() function. The data was then summarised utilising the summarise() function producing the median value of the following risk factors, (1) waist circumference, (2) Triglycerides, (3) HDL and (4) blood glucose, this was achieved by using the median() function internally to the summarise() function.

The statistical significance of risk factors

Based on the risk factors four boxplots where produced using the boxplot() function to determine if a parametric or non-parametric test was suitable. This resulted in the undertaking of four Wilcoxon non-parametric tests due to outliers in the variables associated with risk factors, and a test which is used on clinical data (Kim, 2014). The next step was to develop a generic hypothesis which could be applied across to all risk factors. The following hypothesis was conducted.

- Null hypothesis (H0): That the median value of a risk factor comes from the same population (no difference between those with metabolic syndrome and those without metabolic syndrome).
- Alternative hypothesis (HA): That the median value of the risk factors come from different populations (There is a significance between those with metabolic syndrome and those without metabolic syndrome).

On rejection of the null hypothesis the alternative is adopted and that there is evidence to suggest that the value between risk factors is different between metabolic syndrome statues, assuming a significance level of 5%. Four tests where conducted looking at each risk factor individually, this was done by calling the wilcox.test() function and a two sided test was conducted. The distribution of the risk factors was then generated visually in ggplot2 by calling on the geom_boxplot() function and the graphs aggregated with grid.arrange() which is offered in with the **gridExtra** package.

(Di)similarity between patients with metabolic syndrome and without metabolic syndrome

As the data frame contained mixed variable types as summarised in table 1 and we wanted to differentiate patients based on metabolic syndrome statues we utilised Gower's method to produce a (di)similarity matrix. This was achieved through using a random sample of patient data. We first created a new data set from the subsample and called on the **dplr** package to arrange the new data set by metabolic syndrome statues using the arrange() function. With the new arranged data set we called on the ***cluster*** package, specifically the daisy() function using the 'gower' method and saved it in the global environment as a matrix with the as.matrix() function for further analysis.

To analysis the matrix utilising ggplot2 we first called on the ***Reshape2*** package and used the melt() function to arrange the matrix in pairs longwise, representing var1 and var2 as the seqn variable and the value corresponding to the (di)similarity value between those patients. We then called on ggplot() assigning x var1 and y var2, ggplot was layered with geom_raster() and assigning fill=value in the longwise table, finally the graph was labelled and other

characteristics assigned.

RESULTS

Risk factor summary

The generation of the pivot table as summarised in table 2 suggests that the median value of risk factors associated with metabolic syndrome are different for those with metabolic syndrome and those without metabolic syndrome which holds true regardless of gender. The median values are higher for those with metabolic syndrome for (1) waist circumference, (2) triglycerides and (3) blood glucose, whilst the median value of HDL is lower in those without metabolic syndrome, this is normally referred to as good cholesterol and why it is lower.

Table 2: Median values of BMI, triglycerides (Tri), HDL, blood glucose and a complete count grouping the data by metabolic syndrome statues and gender to allow for variation in physiology by genders. Evidence to suggest that a possible difference exists between those with and without metabolic syndrome.

Metabolic Syndrome	Sex	Waist	Tri	HDL	Glucose	count
Positive	F	104	148	47	110	403
Positive	M	109	167	40	110	419
Negative	F	88	85.5	61	93	808
Negative	M	93	92	49	98	771

The statistical significance of risk factor

In conjunction with table 2 a boxplot was produced to visually study the effects of metabolic syndrome of the distribution against all risk factors present in the data set. It is evident in figure 1 that those with metabolic syndrome tend to have higher waist circumference, blood glucose concentration, triglycerides levels and a lower HDL compared to those without metabolic syndrome, with a large amount of physiological variance as indicated by multiple outliers for all risk factors

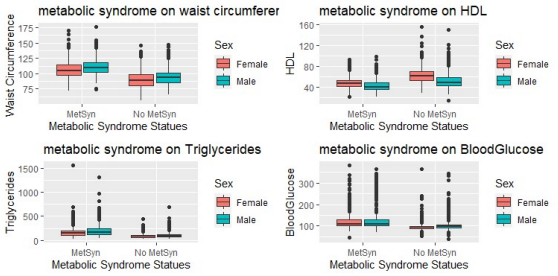

Figure 1: Boxplot distribution of all risk factors present in the data set against those with and without metabolic syndrome. We tend to observe a difference in the distribution of certain risk factors depending on metabolic syndrome status.

Based on the four individual Wilcoxon non-parametric tests as summarised in table 3 there is sufficient evidence to reject the null hypothesis and adopt the alternative hypothesis. This suggest that the median value between risk factors is different between metabolic syndrome statues, assuming a significance level of 5%. There is enough evidence to suggest that those with metabolic syndrome have higher levels of (1) blood glucose, (2) triglycerides and (3) a larger waist circumference, though in addition have lower levels of (1) HDL then those without metabolic syndrome ($P<0.05$ for all tests), complementing table 2.

Table 3: Based on the Wilcoxon non-parametric tests we observe that all risk values present in the data set are statistically significant ($P<0.05$) and that there is enough evidence to reject the null hypothesis. This suggest that the median values in table 2 for each risk factor are significant and distinguishable based on metabolic syndrome statues.

Risk factor	P-Value	Interpretation
Waist circumference	2.2e-16	Reject null hypnosis and adopt alternative.
HDL	2.2e-16	Reject null hypnosis and adopt alternative.
Triglycerides	2.2e-16	Reject null hypnosis and adopt alternative.
Glucose	2.2e-16	Reject null hypnosis and adopt alternative.

(Di)similarity between patients with metabolic syndrome and without metabolic syndrome

The levelplot presented in figure 2 is patient data which has been arranged by metabolic syndrome statues. Looking at the arranged raw data in the matrix we note that patient 0-151 (order of appearance in new data frame) has metabolic syndrome and patient 152-500 (order of appearance in

new data frame) does not have metabolic syndrome. We can see by the x and y axis in figure 2 that there is more similarity as indicated by the lighter shade in each group, and that there is dissimilarity as indicated by the darker shade. For example, the following point 100,400 is comparing patient 100 (metabolic positive) against patient 400 (metabolic syndrome negative) and suggest that there is greater dissimilarity between the patients, which holds true for most observations as indicated by the shade.

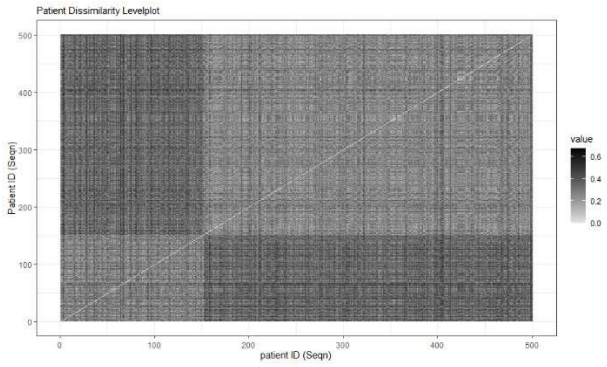

Figure 2: Levelplot comparing the dissimilarity between patients after data was arranged by metabolic syndrome. Patient ID 1-151 (order of appearance) is associated with metabolic syndrome whilst patient 152-500 is associated without metabolic syndrome. The darker the shade of black the higher the dissimilarity score

DISSCUSION

Some limitations in this study are that we were unable to locate how the data was originally sourced prior to the SQL query, in addition another risk factor associated with metabolic syndrome is blood pressure and that was a variable not present in the original data set. The method used to impute missing values, whilst has been used in clinical research there are more appropriate methods which can reduce bias. Though there is evidence to suggest that there is a difference in risk factors between those with metabolic syndrome and those without, with the gower's similarity matrix suggesting that when arranging data by metabolic statues there is clear dissimilarity between the two groups, possible eluding to future research utilising machine learning as a predictive tool in clinical diagnostic. The gower's matrix also suggest future studies may also benefit at looking at other variables which may be cofounding such as age, gender, income and even marital statues which was not explored here, whilst not directly linked to health statues there may be a confounding relationship due to income inequality or that older people are more prone to metabolic syndrome, as an example.

CONCLUSION

Despite the limitations previously mentioned there tends to be consensus between the literature and this study that there is statistical significance between patients with metabolic syndrome and those without. Due to the variation as indicated in physiological variation between patient's risk factors the application of the gower's dissimilarity matrix suggests that we are able to discriminate based between metabolic syndrome based on both risk factors and possibly non risk factors which were not analysed here. The future direction for this paper could suggest on expending on non-risk factors and the application of machine learning to predict and prevent the development of metabolic syndrome and subsequent decline of patient's health.

REFERENCE

Kim, H. (2014). Statistical notes for clinical researchers: Nonparametric statistical methods: 1. Nonparametric methods for comparing two groups. Restorative Dentistry & Endodontics, 39(3), 235. doi: 10.5395/rde.2014.39.3.235

Rochlani, Y., Pothineni, N., Kovelamudi, S., & Mehta, J. (2017). Metabolic syndrome: pathophysiology, management, and modulation by natural compounds. Therapeutic Advances In Cardiovascular Disease, 11(8), 215-225. doi: 10.1177/1753944717711379

Stevens, J., Suyundikov, A., & Slattery, M. (2016). Accounting for Missing Data in Clinical Research. JAMA, 315(5), 517. doi: 10.1001/jama.2015.16461

Scholze, J., Alegria, E., Ferri, C., Langham, S., Stevens, W., Jeffries, D., & Uhl-Hochgraeber, K. (2010). Epidemiological and economic burden of metabolic syndrome and its consequences in patients with hypertension in Germany, Spain and Italy; a prevalence-based model. *BMC Public Health*, *10*(1). doi: 10.1186/1471-2458-10-529

R-Studio Code

```
# Packages used


Install and call for ggplot2

# install.packages("ggplot2")

library(ggplot2)

library(dplyr)

library(Hmisc)

library(tidyr)

library(VIM)

library(cluster)

library(lattice)

library(reshape2)

#########################################
#########################################
###

# Load into data and conduct basic exploratory analysis

# data sourced from data word

#########################################
#########################################
###


metabolic_1 <- read.csv("metabolic.csv", header = T, na.strings=c("","NA"))

summary(metabolic_1)

str(metabolic_1$Sex)


#########################################
#########################################
###

#                - Imputation START -

#

# Determine the percentage of na values in data set

# In this section we will do the following steps to impute the missing values

# Boxplot to look for skewness/outliers - depending if we do univariate of

# predicvie or knn

#

#                - Imputation START -

#########################################
#########################################
###


na_prop <- function(x) sum(is.na(x))/sum(!is.na(x)+is.na(x))

apply(metabolic_1, MARGIN=2, FUN=na_prop)
```

```
#########################################
#########################################
###

# The output suggest that the following variables contain missing values

# We will produce boxplot for the following:

# BMI

# Income

# Waist circ

#########################################
#########################################
###


boxplot(metabolic_1$BMI) # Multiple outliers the higher values

boxplot(metabolic_1$WaistCirc) # Multiple outliers for the higher values

boxplot(metabolic_1$Income) # No outliers, median towards lower value


#########################################
#########################################
###

# As we have multiple missing values accros 3 variables I will employ the K-nn

# function by calling the VIM package

# If martial status is not recorded we will replace it with unknown and then

# conduct a knn imputation

#########################################
#########################################
###


metabolic_2 <- metabolic_1 %>%

  replace_na(list(Marital = "unkown"))



metabolic_3 <- kNN(metabolic_2, k=4, imp_var = FALSE)

apply(metabolic_3, MARGIN=2, FUN=na_prop) # Confrim no more NA values

apply(metabolic_3, MARGIN=2, FUN=na_prop)

boxplot(metabolic_3$BMI, metabolic_2$BMI) # No major chnage

boxplot(metabolic_3$WaistCirc, metabolic_2$WaistCirc) # No major chnage

boxplot(metabolic_3$Income, metabolic_2$Income) # No major change


#########################################
#########################################
###

# change variable types due to errors preventing data loading into certain

# functions

#########################################
#########################################
###


metabolic_4 <- metabolic_3 %>%

  mutate(Sex = as.factor(Sex),

         seqn = as.integer(seqn),

         Age = as.integer(Age),

         Marital = as.factor(Marital),

         Income = as.integer(Income),

         WaistCirc = as.numeric(WaistCirc),

         BMI = as.integer(BMI),

         Race = as.factor(Race),

         Albuminuria = factor(Albuminuria, levels = c("0", "1")),

         UrAlbCr = as.numeric(UrAlbCr),

         UricAcid = as.numeric(UricAcid),

                                  BloodGlucose = as.integer(BloodGlucose),

         HDL = as.integer(HDL),

         Triglycerides = as.integer(Triglycerides),

                                  MetabolicSyndrome = as.factor(MetabolicSyndrome))
```

```
#########################################
#########################################
###

# Data set created by taking a sample due to computational power

#########################################
#########################################
###


# Take a random sample each time code is ran

metabolic_sample <- metabolic_4 %>%

  sample_n(size=500)



#########################################
#########################################
###

#                    - Imputation END -

#

# We have now removed all the na values from the original data set utilsing two

# different methods to imputate.

# We replaced NA in maritial status with unkown status

# we replaceed NA in all the other variables utilsing k-NN imputation

#

#                    - Imputation END -

#########################################
#########################################
###


#########################################
#########################################
###

#                    - Subgrouping and GGPLOT START -

#

# We are curious how the data looks by comparing groups metabolic syndrome

# vs those without metabolic syndrome accross all our risk factors

#

#

#                    - Subgrouping and GGPLOT START -

#########################################
#########################################
###


Risk_factor <- metabolic_4 %>%

  group_by(MetabolicSyndrome, Sex) %>%

  summarise(Med_Waist = median(WaistCirc),

            Med_Tri = median(Triglycerides),

            Med_HDL = median(HDL),

            Med_Glucose = median(BloodGlucose),

            count = n())



#########################################
#########################################
####

# The above summary shows some relationship on face value between the met

# /non-metabolic conditions. We will explore further with boxplots in GGPLOT

#########################################
#########################################
###


#########################################
#########################################
###

# Are these value(s) significant?

# Only sample 1 for now due to limitation in word length of paper
```

```
# Test if waist circumference differes between those with metabolic syndrome

# and those without metabolic syndrome and overlaying the stats ontop

#########################################################################

require(gridExtra) # allows for ggplot grpahs to be condensed as one


plot1 <- ggplot(data = metabolic_4) +

                geom_boxplot(mapping = aes(x=MetabolicSyndrome, y=WaistCirc, fill=Sex)) +

      labs(x="Metabolic Syndrome Statues", y="Waist Circumference ", title="metabolic syndrome on waist circumference")


plot2 <- ggplot(data = metabolic_4) +

geom_boxplot(mapping = aes(x=MetabolicSyndrome, y=HDL, fill=Sex)) +

      labs(x="Metabolic Syndrome Statues", y="HDL", title="metabolic syndrome on HDL")


plot3 <- ggplot(data = metabolic_4) +

                geom_boxplot(mapping = aes(x=MetabolicSyndrome, y=Triglycerides, fill=Sex)) +

      labs(x="Metabolic Syndrome Statues", y="Triglycerides", title="metabolic syndrome on Triglycerides")


plot4 <- ggplot(data = metabolic_4) +

                geom_boxplot(mapping = aes(x=MetabolicSyndrome, y=BloodGlucose, fill=Sex)) +

      labs(x="Metabolic Syndrome Statues", y="BloodGlucose", title="metabolic syndrome on BloodGlucose")


grid.arrange(plot1, plot2, plot3, plot4) # Plot all grpahs on same greid


wtest_waist =wilcox.test(WaistCirc~MetabolicSyndrome, data = metabolic_4, alternative="two.sided",mu=0)

wtest_hdl =wilcox.test(HDL~MetabolicSyndrome, data = metabolic_4, alternative="two.sided",mu=0)

wtest_tri =wilcox.test(Triglycerides~MetabolicSyndrome, data = metabolic_4, alternative="two.sided",mu=0)

wtest_glucose =wilcox.test(BloodGlucose~MetabolicSyndrome, data = metabolic_4, alternative="two.sided",mu=0)


wtest_waist

wtest_hdl

wtest_tri

wtest_glucose


#########################################################################

# Evidence supports the finding that waist circ differs between those with

# metabolic syndrome and those without (p<0.05)

# THis test due to larghe amount of outliers

#########################################################################



#########################################################################

#                    - Similarity START -

#

# The data set provided is a mixed data set with characters, integers and
```

```
# numbers. The method used for this section will be a similairty matrix

# that caters for mixed variable data types. A common method used is the

# Gowers method.

#

#                    - Similarity START -

##########################################################################
###

# Arrange data by metabolic syndrome to ID any obvious patters in the di(similairty)

# matrix when producing a levelplot

metabolic_sample_ordered <- metabolic_sample %>%

  arrange(MetabolicSyndrome)

gower_disimilarity <- (daisy(metabolic_sample_ordered, metric="gower", type = list(symm = c(9))))

disimilarity_matrix <- as.matrix(gower_disimilarity)

# To use GGPLOT we must first convert to long data using reshape package

longData <- melt(disimilarity_matrix) # Reshape2 package: Convert into pairs

ggplot(longData, aes(x = Var2, y = Var1)) +

  geom_raster(aes(fill=value)) +

  scale_fill_gradient(low="grey90", high="black") +

    labs(x="patient ID (Seqn)", y="Patient ID (Seqn)", title="Patient Dissimilarity Levelplot") +

                    theme_bw() + theme(axis.text.x=element_text(size=9, angle=0, vjust=0.3),

                axis.text.y=element_text(size=9),

                plot.title=element_text(size=11))

# Presenting the gowers (di)similairty matrix after arranging the data into metabolic

# syndrome +/- we observe that there is

# This tells us that as the d(x,y) score increase there is greater disimialirty

# between the observational points. We observe that there is a greater dismilairy

# between metabolic + patients and metabolic - patients as indicated by the blue

# regions. Possible machine learning could be applied to predict patient outcome
```