Emerging Patterns of Lifestyle Impact on Health and Wellness

Tanmay Gupta[1]**, Dr. Purav Gandhi**[2]

[1]Department of Business Analytics and Intelligence, Indian Institute of Management, Bangalore

[2]Remedy Social (Healthark Wellness Solutions LLP), C-602, Tulip Citadel, Ambawadi, Ahmedabad, India

Abstract: ***We conducted an analysis on health related information related to demographics, lifestyle, common complains, medical history and family history, reported by over 1,100 people. The population has an average age of 31.5 years, with 80% males and 20% females. A correlation analysis identified multiple clusters related to life style and health problems including daily activity or exercise, regular stress and lack of sleep being closely associated with a combination of heartburn, gastritis, headaches, back-pain, and allergies, amongst other correlations.***

Keywords: lifestyle related health risks, cluster analysis, health risk assessment, correlation, disease patterns

1. Introduction

Lifestyle is changing at a very rapid pace as we enter the internet era. Pace of evolution in terms of technology, lifestyle, work environment, etc. is more rapid than ever before and has resulted in how our lifestyle and health has changed. Health is largely impacted by a combination of lifestyle, environmental factors, genetic make-up an individual and the kind of care the individual receives

We wanted to understand the new health and wellness patterns emerging to be able to design improved solutions to help people preserve and protect health. We conducted a small study using the Remedy Social [1]online health risk assessment platform, to understand some of the relationships between lifestyle, common medical complaints faced by people and their inter-relationships.

2. Methodology

To assess the correlation effect we conducted an online survey using the Remedy Social platform, which comprised of nearly 40 questions, requesting for an individual's demographic, lifestyle, medical complaints, family history, and gynecology details. These questions were selected based on detailed literature review to ensure that key lifestyle risk elements are covered.

The data mining and graph generating exercises were done using open source statistical software R(v 3.2.2) [2] with RStudio as front end development interface.

3. Data Set

The survey was taken by over 1,100 people of whom 945 filled the complete survey. The gender distribution of the survey population is 80% males and 20% females. The mean age of the people who took the survey is 31.5 years with population age ranging from 15 years to 65years.The survey was taken by people from all parts of the country, including few people from overseas, so that the results are not biased and present a true picture of the lifestyle impact.

4. Correlation Analysis

As most of the variables generated from the response were binary in nature, therefore, to estimate the correlation starting from medical condition co-occurrence, we used phi correlation coefficient. The phi coefficient (also referred to as the "mean square contingency coefficient") denoted by ϕ is a measure of association for two binary variables. Introduced by Karl Pearson, this measure is similar to the Pearson correlation coefficient in its interpretation [3].

We can define the correlation coefficient associated with a pair of medical conditions i and j as:

$$\phi_{ij} = \frac{C_{ij}\,N - P_i P_j}{\sqrt{(P_i P_j - C_{ij})(N-P_i)(N-P_j)}} \qquad (1)$$

where:
C_{ij} = the number of patients affected by both the conditions
N = total number of patients in the studied population
P_i = patients with medical condition i
P_j = patients with medical condition j

The value of ϕ ranges from −1 to +1, where ±1 indicates perfect agreement or disagreement, and 0 indicates no relationship. The significance of $\phi \neq 0$ can be determined by calculating the associated p-value and rejecting the null hypothesis where p value < 0.05.

For our analysis, we picked 44 dichotomous variables which resulted in $^{44}C_2$ i.e. a total of 946 pairs. The phi coefficient was calculated for each of the pairs, and for each phi-coefficient value, p-value was calculated using χ^2(Chi-Square) Test.

Please note that phi-coefficient bears a relationship with Chi-Square statistic, χ^2, [4] as:

$$\phi^2 = \frac{\chi^2}{N} \qquad (2)$$

where N is same as define in eq (1) above.

Out of the possible 946 pairs, 770 came to be statistically significant using eq (2). The value of correlation coefficient

varied from -1 to +1 and the thumb rule [5] to interpret the value is:
-1.0 to -0.7 strong negative association.
-0.7 to -0.3 good negative association.
-0.3 to +0.3 little or no association.
+0.3 to +0.7 good positive association.
+0.7 to +1.0 strong positive association.

Using the above rule, there were 195 pairs which had correlation coefficient, ϕ, less than -0.3 and greater than 0.3. That is, from a total of 770 significant correlation pairs, 195 showed greater association and thus, could be considered for drawing insights.

5. Insights

A network plot was drawn to visualize the 195 correlation pairs and make inferences about the same, as shown in Figure 1.

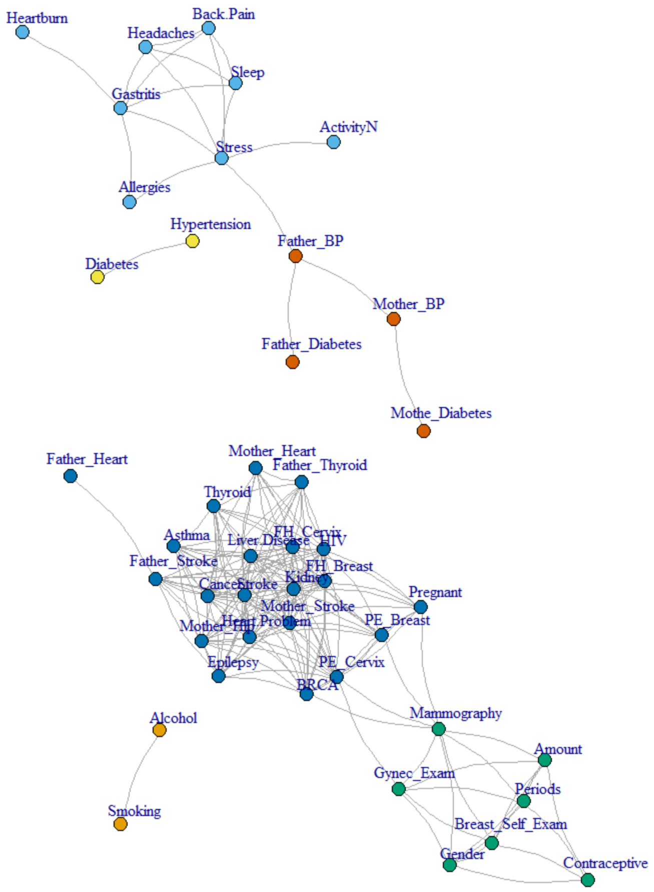

Figure 1: Visualizing correlation pairs using network diagram

Each node in the network graph above shows a clinical condition (demographic, lifestyle, medical complaints, family history, and gynecology details) and an edge between two nodes depicts the existence of correlation. The different colors of the nodes shows presence of different communities (or clusters) formed inside the network. The communities are defined using modularity index [6], where networks with high modularity have dense connections between the nodes within cluster but sparse connections between nodes in a different cluster.

The blue network at the top depicts a close connection between comorbidities viz. Heartburn, Gastritis, Head Aches, Back Pain, Sleep, Stress, Allergies and Activity performed. All these comorbidities are connected to each other which shows a clear association between long working hours, unhealthy food habits and low physical activity. This shows a classic example of present day IT work force which sits and work at their place for a long period of time while having junk food to satisfy their appetite.

Similarly, the red network confirms the finding that if in a family both the parents are diabetic and have instances of high blood pressure, there are greater chances of their kin to develop stress and its related comorbidities.

The dense cluster in the middle shows greater correlation between different comorbidities i.e. Heart Problem, Stroke, Liver Disease, Kidney, Thyroid, Cancer, HIV, Epilepsy and Asthma. There was larger correlation between heart problem & stroke which completely makes sense clinically. Then, a correlation of 0.74 between cancer & HIV suggests that people infected with HIV have a substantially higher risk of cancer compared with uninfected people of the same age. At the same time, there were few correlations which weren't in tune perfectly with clinical literature and seemed biased e.g. stroke & HIV, HIV & epilepsy. The reasons could be filling of inaccurate information or disinterest to fill the complete survey etc. In fact we should consider these correlations as an indication of a relationship or starting point of further research, and not something as a medically proven fact.

Alcohol and smoking share a greater association, that is, if one tend to have any of the habits then there are greater chances of him/her to develop the second habit as well. Although this pair didn't show a greater correlation with any other medical condition but it confirmed a good correlation between the two unhealthy habits.

The green cluster shows association between different medical conditions related to women.

6. Conclusion

The study was able to identify a series of correlation between various lifestyle patterns and clusters in the young population, to be able to develop more effective preventive / mitigation strategies for some of the new age health problems. It also helps identify at risk population in a more effective manner, to be able to bring population under surveillance at an early stage.

Moreover, the analysis creates a framework for further analysis and developing a healthcare learning system on larger datasets to develop effective health prevention strategies.

We are planning to leverage the finding of this analysis to act as a basis for further studies to: (a) validate these patterns and refine them over a larger sample set, and (b) conduct a causality analysis by tracking the same patients and understanding their lifestyle in more detail.

References

[1] [Online]. www.remedysocial.com.

[2] W. Venables, D. Smith, R Core Team,“An Introduction to R”, Version 3.2.2, 14-Aug-2015. [Online]. https://cran.r-project.org/.

[3] O.B. Chedzoy, “Encyclopedia of Statistical Sciences”, Wiley, New York (2006).

[4] T. Colignatus,“Correlation and regression in contingency tables. A measure of association or correlation in nominal data (contingency tables), using determinants,” Item#3394, Thomas Cool Consultancy & Econometrics (2007).

[5] M. Mukaka, “A guide to appropriate use of Correlation coefficient in medical research,” Malawi Med J. 2012 Sep; 24(3): 69–71.

[6] M. E. J. Newman, “Modularity and communitystructure in networks,” Proceedings of theNationalAcademy of Sciences, vol. 103, no. 23, pp. 8577–8582, 2006.