# COMP20008 Project Phase 2: Concept Formulation and Preliminary Investigation

## *1 Community chronic disease risk and health risk: correlation analysis and community feature investigations*

**Author:** XuLin Yang, 904904

**2 Domain:** communities, health

## 3 Questions:

(1) Identify the correlation between community citizens' chronic disease risks and citizens' health risk factors.

(2) What features in communities causes citizens to have these health risks and increase disease risks?

## 3 Question interests, aim & potential benefits:

Australian Government of Human and Health Service and Local government will be interested in these questions. For government health department, they can know how healthy citizens are in each lga[1]. For each local government, they can have a clear idea about what the kind of policies or community activities they can make or organize to help citizens to have a healthier life and lower the health risk in lga. With the discussion and analysis on the data below and questions above, the result derived or observed can be used to suggest the local government doing something to make citizens healthier. These data sets below can be used to reflect the regional incidence of disease risks and give suggestions about improving the local health level of some communities. As a result, government can gain the information about the citizen health state for each geographical area. Citizens can enjoy a healthier life with the open data which are used to benefit Victorians.

## 4 Question innovation

Firstly, the data sets below focus on lga. By wrangling data in groups of geographical areas, characteristic suggestions for improving citizen health can be provided for each local government council. Secondly, how are chronic disease risks related to citizens' life styles can be explored with various community development levels (such as social and economical factor, population eating habits, etc). In "The Chronic Care Model (Wagner et al.1999) [2]", it states that the chronic disease risk is correlated to community health service delivery rate and the personal health system. Compared to what discussed in "The Chronic Care Model (Wagner et al.1999)". This investigation explores a wider range of community features and look for relation between these features and disease risks. With the above exploration, some characteristic community development advice can be obtained and provided for particular communities.

## 5 Datasets used:

All listed below has .csv data file and .json metadata file. And can be downloaded from aurin. All of these data are statistically counted in 2011.

1) LGA11 Chronic Disease Modeled Estimate
2) LGA11 Health Risk Factors Modeled Estimate
3) LGA11 Psychological Distress Modeled Estimate

These three data sets have the similar structure. All of them are labelled by LGA area code. As its name described, they describe features in each cases. See detailed metadata table at the end. (1) contains ten chronic diseases and data confidence level ('xxx - PRMSE'). (2) contains four health risks and data confidence level. (3) contains high psychological distress experienced and data confidence level. These data sets have noise data in about 0%-5% specified by confidence columns.

4) Local Government Area LGA profiles data 2011

This data set is a big data set. It is labelled by lga code. It illustrates facilitate service planning and policy development by enabling access to a broad range of data. It is used to analyze community features. See detailed metadata table at the end. There are something missing in the table. Although

[1] Australian Statistical Geography Standard (ASGS): Volume 3, July 2011 http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1270.0.55.003~July%202016~Main%20Features~Local%20Government%20Areas%20(LGA)~7

[2] Bulgaru-Iliescu, Diana & Oprea, Liviu & Cojocaru, Daniela & Sandu, Antonio. (2013). The Chronic Care Model (CCM) and the Social Gradient in Health. Revista de Cercetare şi Intervenţie Socială. 41. 176-189. https://www.researchgate.net/figure/The-Chronic-Care-Model-Wagner-2004-Wagner-et-al-2001_fig1_271585144

there are some outliers in the table, they can be used in later investigations.

Additional data sets might be used later.

## 6 Data Preprocessing

After reading data, columns with meaningful metadata is replaced. By doing this, real meaning for attributes can be used and all dataframes can have a common key 'Local_Government_Area_Code' when merging tables. Data in data sets (1)-(4) are going to be split in three pd.DataFrame() for disease_risks, health_risks and health_services.

Errors in column name like in 'LGA11 Health Risk Factors Modeled Estimate.csv' {'obese_m_me_3_rrmse_3_11_7_13' : Obese Females 18 Years and Over - RRMSE } is wrong and corrected to 'Obese Males 18 Years and Over - RRMSE'.

Some useless rows (e.g. row.lga_name = 'Unincorporated Vic') are dropped. In consequence, higher efficiency and more accuracy can be achieved for analysis. Similarly, useless columns are not selected from raw data.

When splitting into dataframes studied later, several of this columns(e.g. 'GP_attendence' for 1000 people normalized to average citizen's 'GP_attendence') not normalized are normalized into the same range in the stage. These features are divided by population in that lga. As a result, the purpose to explore relations between features is achievable.

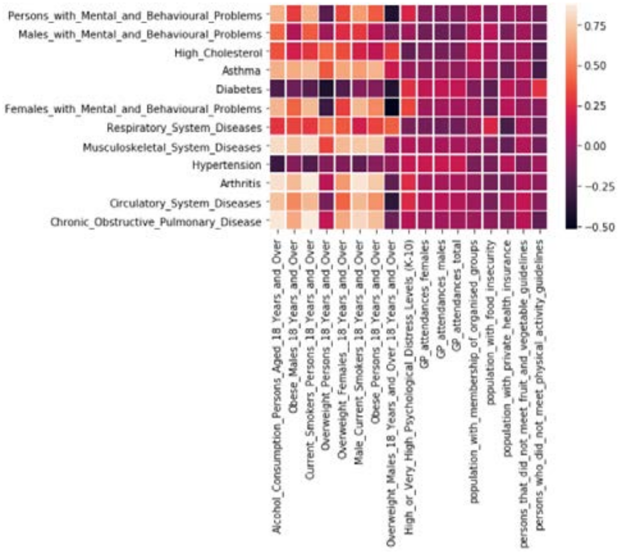
There are some missing data in tables (e.g. lga_profile.LGA_land_use(rural/residential/industrial)) has some missing data as it has something like '<1%' in .csv and can't be read by pd.read_csv(). So using fillna(0) method can make data consistent with the original data set with losing minimum information.

Some data needed to be derived from one columns. For example, LGA_name is a string with 'Cities (C), Rural Cities (RC), Boroughs (B) and Shires (S)' and lga name and type and be transferred. This group lga by its geographical location.

Unreliable data will be dropped if 'xxx - PRMSE' == 1

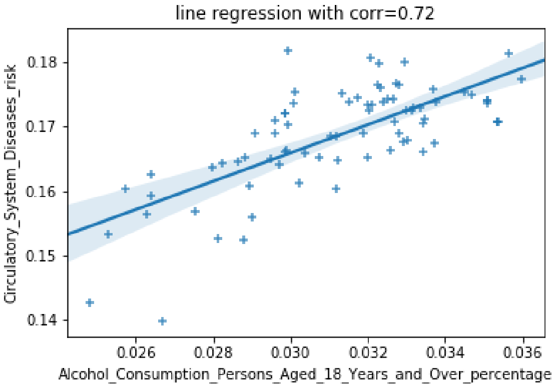
## 7 Initial Investigation

Initially, use Pearson correlation to calculate correlation matrix. Then use heatmap() from seaborn library to plot heatmap and investigate relationship between citizens' chronic disease risk and health risk.

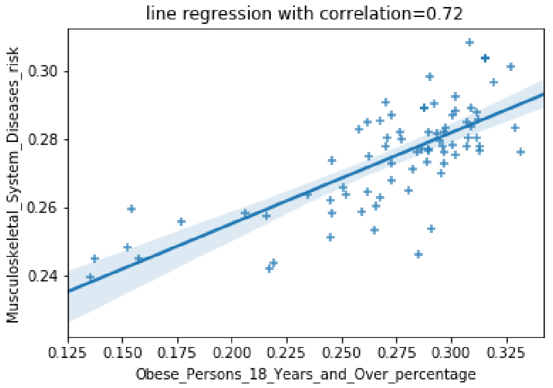


*figure(1) heat map against disease risk and health risk*

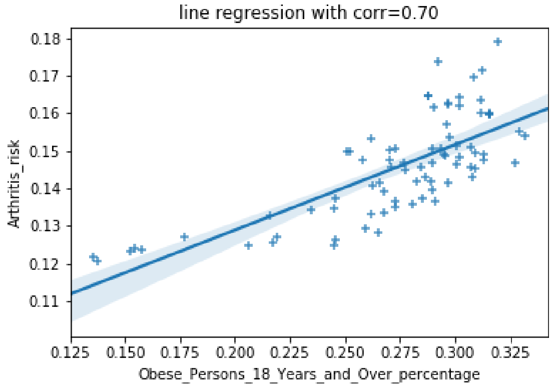
From the observation from figure(1) we pick some interesting relations. The community with higher percentage of adults drinking alcohol has a higher risk of having the circulatory system disease(r ≈ 0.72). The community with higher percentage of obese adults has a higher risk of having musculoskeletal system diseases(r ≈ 0.72) and arthritis(r ≈ 0.70). The community with higher percentage of adults who smoke results in a higher risk in asthma(r ≈ 0.71) and chronic obstructive pulmonary disease(r ≈ 0.88).



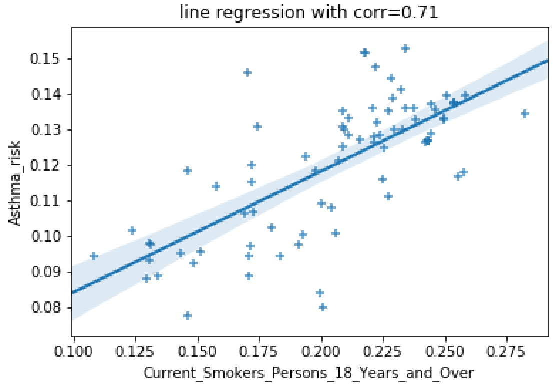
*figure(2) linear regression model for adult alcohol v.s. circulatory system disease*



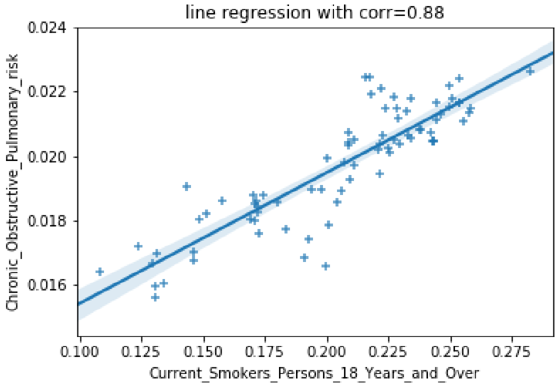
*figure(3) linear regression model for obese adults v.s. musculoskeletal system diseases risk*



*figure(4) linear regression model for obese adults v.s. arthritis risk*



*figure(5) linear regression model for adult smokers percentage v.s. asthma risk*



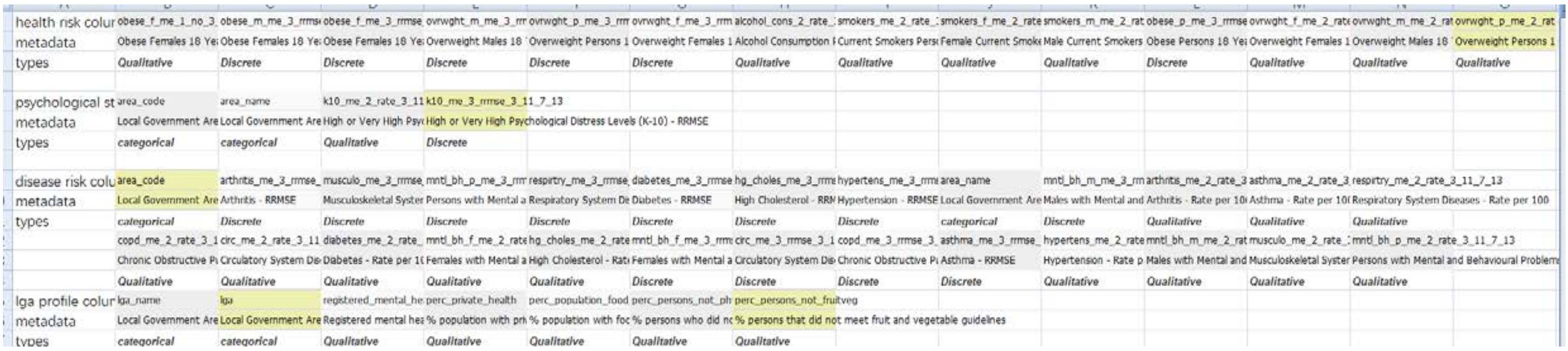
*figure(6) linear regression model for adult smoker percentage v.s. chronic obstructive pulmonary risk*

***To sum it up, several chronic diseases are related citizens' health risk as presented.***

## 8 Further Investigation & feasibility

Figure(2)–(6) is drawn by regplot in seabon library. As demonstrated by figure(2)–(6), there exists the correlation between chronic disease risk and community residents' health risk. Moreover, there are some interesting outliers that are not lied near the regression model nor the major group of lga. And this is what is going to be further investigated. Additionally, more community features like economical disadvantage index will also be introduced into analysis in Q(2). The technique like parallel coordinates will be used to discuss how multiple features influence health risks. The principle coordinates analysis will be used to find the major influence that determines health risks.

Grouping techniques will be used to find similar communities which will provide useful information of how to make policies for various communities. And more data columns will be covered in further discussion.



| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| health risk colur | obese_f_me_1_no_3 | obese_m_me_3_rrms | obese_f_me_3_rrmse | ovrwght_m_me_3_rr | ovrwght_p_me_3_rr | ovrwght_f_me_3_rrm | alcohol_cons_2_rate_ | smokers_me_2_rate_ | smokers_f_me_2_rate | smokers_m_me_2_rat | obese_p_me_3_rrmse | ovrwght_f_me_2_rat | ovrwght_m_me_2_rat | ovrwght_p_me_2_rat |
| metadata | Obese Females 18 Ye | Obese Females 18 Ye | Obese Females 18 Ye | Overweight Males 18 | Overweight Persons 1 | Overweight Females 1 | Alcohol Consumption | Current Smokers Pers | Female Current Smok | Male Current Smokers | Obese Persons 18 Ye | Overweight Females 1 | Overweight Males 18 | Overweight Persons 1 |
| types | *Qualitative* | *Discrete* | *Discrete* | *Discrete* | *Discrete* | *Discrete* | *Qualitative* | *Qualitative* | *Qualitative* | *Qualitative* | *Discrete* | *Qualitative* | *Qualitative* | *Qualitative* |
| | | | | | | | | | | | | | | |
| psychological st | area_code | area_name | k10_me_2_rate_3_11 | k10_me_3_rrmse_3_11_7_13 | | | | | | | | | | |
| metadata | Local Government Are | Local Government Are | High or Very High Psy | High or Very High Psychological Distress Levels (K-10) - RRMSE | | | | | | | | | | |
| types | *categorical* | *categorical* | *Qualitative* | *Discrete* | | | | | | | | | | |
| | | | | | | | | | | | | | | |
| disease risk colu | area_code | arthritis_me_3_rrmse_ | musculo_me_3_rrmse_ | mntl_bh_p_me_3_rr | respirtry_me_3_rrmse_ | diabetes_me_3_rrmse | hg_choles_me_3_rrm | hypertens_me_3_rrm | area_name | mntl_bh_m_me_3_rm | arthritis_me_2_rate_3 | asthma_me_2_rate_3 | respirtry_me_2_rate_3_11_7_13 | |
| metadata | Local Government Are | Arthritis - RRMSE | Musculoskeletal Syster | Persons with Mental a | Respiratory System De | Diabetes - RRMSE | High Cholesterol - RRM | Hypertension - RRMSE | Local Government Are | Males with Mental and | Arthritis - Rate per 10 | Asthma - Rate per 10( | Respiratory System Diseases - Rate per 100 | |
| types | *categorical* | *Discrete* | *Discrete* | *Discrete* | *Discrete* | *Discrete* | *Discrete* | *Discrete* | *categorical* | *Discrete* | *Qualitative* | *Qualitative* | *Qualitative* | |
| | copd_me_2_rate_3_1 | circ_me_2_rate_3_11 | diabetes_me_2_rate_ | mntl_bh_f_me_2_rate | hg_choles_me_2_rate | mntl_bh_f_me_3_rrm | circ_me_3_rrmse_3_1 | copd_me_3_rrmse_3_ | asthma_me_3_rrmse_ | hypertens_me_2_rate | mntl_bh_m_me_2_rat | musculo_me_2_rate_ | mntl_bh_p_me_2_rate_3_11_7_13 | |
| | Chronic Obstructive P | Circulatory System Di | Diabetes - Rate per 1( | Females with Mental a | High Cholesterol - Rat | Females with Mental a | Circulatory System Di | Chronic Obstructive P | Asthma - RRMSE | Hypertension - Rate p | Males with Mental and | Musculoskeletal Syster | Persons with Mental and Behavioural Problem | |
| | *Qualitative* | *Qualitative* | *Qualitative* | *Qualitative* | *Qualitative* | *Discrete* | *Discrete* | *Discrete* | *Discrete* | *Qualitative* | *Qualitative* | *Qualitative* | *Qualitative* | |
| lga profile colur | lga_name | lga | registered_mental_he | perc_private_health | perc_population_food | perc_persons_not_ph | perc_persons_not_fruitveg | | | | | | | |
| metadata | Local Government Are | Local Government Are | Registered mental hea | % population with pri | % population with foo | % persons who did nc | % persons that did not meet fruit and vegetable guidelines | | | | | | | |
| types | *categorical* | *categorical* | *Qualitative* | *Qualitative* | *Qualitative* | *Qualitative* | *Qualitative* | | | | | | | |