Sunnydale Health Characteristics Research

Data Dictionary and Cleaning Documentation

Murray Keogh

HDAT 9400

Table of contents

1. Data Dictionary

1.1. GP Data Dictionary

The table below summarizes the information for the variables collected and created for the GP dataset. The GP data contains the information about patients visiting a general practice in Sunnydale. The GP data contains information on a person-level for the clientele that visited the practice.

Variable name	Description	Format	Allowable Entries
ID	Unique person ID	Number	
GP_last	Date of most recent GP visit	Date (DDMMYY10.)	Dates in the range 01/01/2014 - 31/12/2014
age	Age of patient at the most recent GP visit	Number	
agecat	Age of patient at the most recent GP visit, bucketed into category	Number	1 = 0 - 29 2 = 30 - 39 3 = 40 - 49 4 = 50 - 59 5 = 60 and above
sex	Gender of the patient	Character	1 = male 2 = female
sex_clean	Sex (cleaned)	Character	1 = male 2 = female
cob	In what country were you born?	Number Cobf.	1 = Born in Australia 2 = Born overseas
cob_clean	Cob (cleaned)	Number Cobf.	1 = Born in Australia 2 = Born overseas
healthcare_card	Do you have a healthcare card?	Number Ynf.	1 = Yes 0 = No
healthcare_card_clean	Healthcare_card (cleaned)	Number Ynf.	1 = Yes 0 = No
ever_smoked	Have you ever been a regular smoker?	Number Ynf.	1 = Yes 0 = No
ever_smoked_clean	Ever_smoked (cleaned)	Number Ynf.	1 = Yes 0 = No
Smoke_now	Are you a regular smoker now?	Number Ynf.	1 = Yes 0 = No
Current_smoker_GP	Is the patient currently a smoker?	Number Ynf.	1 = Yes 0 = No
Curr_smoker1	age_stop = . and (smoke_now = 0 or smoke_now = .) and ever_smoked_clean = 1	Number Ynf.	1 = Yes 0 = No
Curr_smoker2	(age_start <= age < age_stop) and ever_smoked_clean = 1 and (smoke_now = 0 or smoke_now = .)	Number Ynf.	1 = Yes 0 = No
Age_start	How old were you when you started smoking regularly?	Number	Invalid if <10 or >105
Age_Stop	How old were you when you stopped smoking?	Number	Invalid if <10 or >105
Drinks_day	About how many alcoholic drinks do you drink per day?	Number	Invalid if >20
Drinks_day_clean	Drinks_day (cleaned)	Number	Invalid if >20
Risky_alchohol_GP	Risky alcohol defined as having more than 2 drinks per day	Number Ynf.	Drinks_day > 2 = Yes

Variable name	Description	Format	Allowable Entries
			Drinks_day <= 2 = No
height	How tall are you without shoes? (meters)	Number	Invalid if < 0.55m or > 2.40m
weight	About how much do you weigh? (kilograms)	Number	Invalid if < 0.4kg or >270kg
weight_clean	Weight (cleaned)	Number	Invalid if < 0.4kg or >270kg
BMI_GP	Calculated BMI from height and weight : Weight(kg) / Height(m)^2	Number	
Obese_GP	Obese defined as BMI greater or equal to 30	Number Ynf.	BMI >= 30 = Yes BMI < 30 = No
Adverse_reaction	Have you had any adverse reaction to any medication?	Number Ynf.	1 = Yes 0 = No
Syst_bp	Systolic blood pressure (mm/Hg)	Number	
Diast_bp	Diastolic blood pressure (mm/Hg)	Number	
reason	Reason for most recent GP bvisit	Character	HEADACHE NAUSEA TINNITUS VOMITING ITCHING ABDOMINAL PAIN DIZZINESS SKIN RASH PALPITATIONS HALLUCINATIONS

1.2. ED Data Dictionary

The table below summarizes the information for the variables collected and created for the ED dataset. The ED dataset contains the information from a single emergency department in Sunnydale – the same neighbourhood where Medical Plus GP is located. The ED data contains information on a record-level, with each record represents an ED presentation by the Sunnydale resident population. This can be joined with the GP dataset using patient ID variable as a key.

Variable name	Description	Format	Allowable entries
ID	Unique person ID		
Ed_admission	Date of ED presentation	Date (DDMMYY10.)	Dates in the range 01/01/2014 - 31/12/2014
Ed_seperation	Date of ED separation	Date (DDMMYY10.)	Dates in the range 01/01/2014 - 31/12/2014
Age_ed	Age of patient at ED presentation	Number	
Sex_ed	Gender of the patient	Number Sexf.	1 = male 2 = female
Cob_ed	In what country were you born?	Number Cobf.	1 = Born in Australia 2 = Born overseas
Cob_ed_clean	Cob_ed (cleaned)	Number Cobf.	1 = Born in Australia 2 = Born overseas
interpreter	An interpreter is needed?	Number Ynf.	1 = Yes 0 = No
Health_insurance	Do you have private health insurance?	Number Ynf.	1 = Yes 0 = No
Triage_category	Urgency of presentation	Number Triagef.	1 = Resuscitation 2 = Emergency 3 = Urgent 4 = Semi urgent 5 = Non urgent
dx1	Principal presenting diagnosis (ICD-10-AM codes)	Character	International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification 8th edition
Dx2-dx5	Up to 4 additional diagnosis (ICD-10-AM codes)	Character	International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification 8th edition
Separation_mode	Status of the person at separation from emergency department	Number Sepmodef.	1 = Admitted to hospital 2 = Departed ED 3 = Died in ED 4 = Dead on arrival

2. Data Cleaning Process

2.1. GP Data Cleaning Notes

1. Check the Data – I examined the contents of the GP dataset. There were 5,837 observations and 17 variables in the original dataset. The variables matched the original data dictionary, and there were no additional variables.
2. Check for Duplicates – I observed 11 exact duplicates and decided to remove all 11 exact duplicates. I observed 31 partial duplicates using the ID variable. When I observed the 31 duplicates (62 records), I noticed many inconsistencies and differing information between the duplicates. As a result, I decided to remove all 31 duplicates (62 records) from the dataset because I questioned the data. Once removed, the dataset now contained 5,764 records.
3. Data Quality Checks – I checked the GP_last field to ensure no GP visits fell outside the allowable range. There were no records observed. I checked the frequency tables for categorical variables and the summary statistics for continuous variables. I observed that Sex included both 1,2 and M,F and would have to be recoded as a new variable. I also observed that cob, healthcare_card, drinks_day, weight, and ever_smoked included values of 99, 998, or 999. I made the decision to recode each of these variables as a new variable. For the purpose of this analysis, I was not sure what a value of 99, 998, or 999 might represent so I decided to recode these values as missing. The number of these values for each variable was low, so there will be minimal impact on further analysis. Also, the original variable remained unchanged, so any information about these values can be updated in the future. I observed that age_start and age_stop both had approximately 80% missing values. This is expected as 65% of patients responded No to ever_smoked. All other variables appeared correct, having only values listed as allowable in the data dictionary. As a result of the new variables needed for recoding, the dataset now contained 23 variables.
4. Variable Creation - I created five new variables for the purpose of this analysis. I performed data exploration to ascertain the number of current smokers to create the current_smoker_GP variable. I first made sure the smoke_now variable had no inconsistencies, mainly that a patient did not say they were a current smoker but had entered an age_stop. There were no records observed. From this, I felt confident using the smoke_now = 1 as a reliable indicator for a current smoker. Additionally, I explored a few methods to increase the ascertainment of current smokers. I checked to ensure no patient had been listed as not currently smoking and had entered an age_start but no age_stop. There were no records observed. I then checked patients who had ever smoked, were not current smokers, but did not enter an age_stop. I observed 49 records that fell into this category. I decided to count these 49 patients as current smokers, since they were smokers and did not specify stopping smoking. I also checked patients who had ever smoked, were not current smokers, and had an age between age_start and age_stop. I observed 87 records that fell into this category. I decided to count these 87 patients as current smokers, since they were smokers and indicated a future age as their stopping age. This implied that they were still currently smoking. To define these records, I created two new variables, currsmoker1 and currsmoker2 for each of the respective situations. Overall, current_smoker_GP ended up with 933 current smokers compared to the original 797 who answered yes to smoke_now.

   I created the risky_alchohol_GP variable, defining any record with drinks_day_clean > 2 as 1 or yes. I created the BMI_GP variable, calculated using weight_clean / height^2. Using the

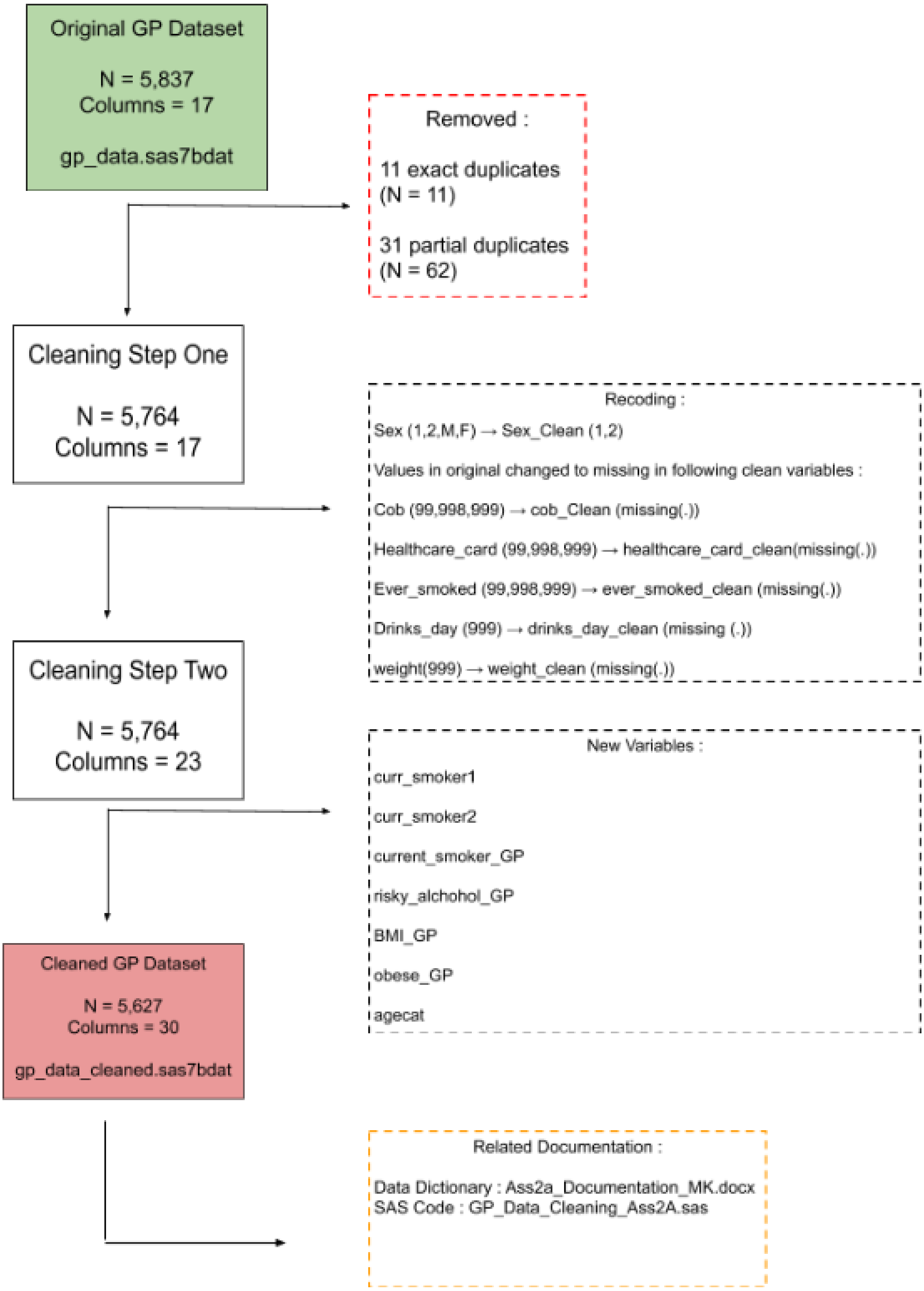
BMI_GP variable, I created the obese_GP variable, defining any record with BMI_GP >= 30 as 1 or yes. Finally, I created the agecat variable to bucket the ages into five categories to utilize in later demographic analysis.

All the recoded and created variables were checked after creation to ensure the logic and code was working correctly.

As a result of the created variables, the dataset now contained 30 variables.

5. Tidy and Save - The final dataset included 5,764 observations and 30 variables. The labels for the new columns were updated. The final cleaned dataset was saved permanently on a local drive to be used in future analysis.

2.2. GP Data Cleaning Flowchart



Original GP Dataset

N = 5,837
Columns = 17

gp_data.sas7bdat

Removed :

11 exact duplicates
(N = 11)

31 partial duplicates
(N = 62)

Cleaning Step One

N = 5,764
Columns = 17

Recoding :

Sex (1,2,M,F) → Sex_Clean (1,2)

Values in original changed to missing in following clean variables :

Cob (99,998,999) → cob_Clean (missing(.))

Healthcare_card (99,998,999) → healthcare_card_clean(missing(.))

Ever_smoked (99,998,999) → ever_smoked_clean (missing(.))

Drinks_day (999) → drinks_day_clean (missing (.))

weight(999) → weight_clean (missing(.))

Cleaning Step Two

N = 5,764
Columns = 23

New Variables :

curr_smoker1

curr_smoker2

current_smoker_GP

risky_alchohol_GP

BMI_GP

obese_GP

agecat

Cleaned GP Dataset

N = 5,627
Columns = 30

gp_data_cleaned.sas7bdat

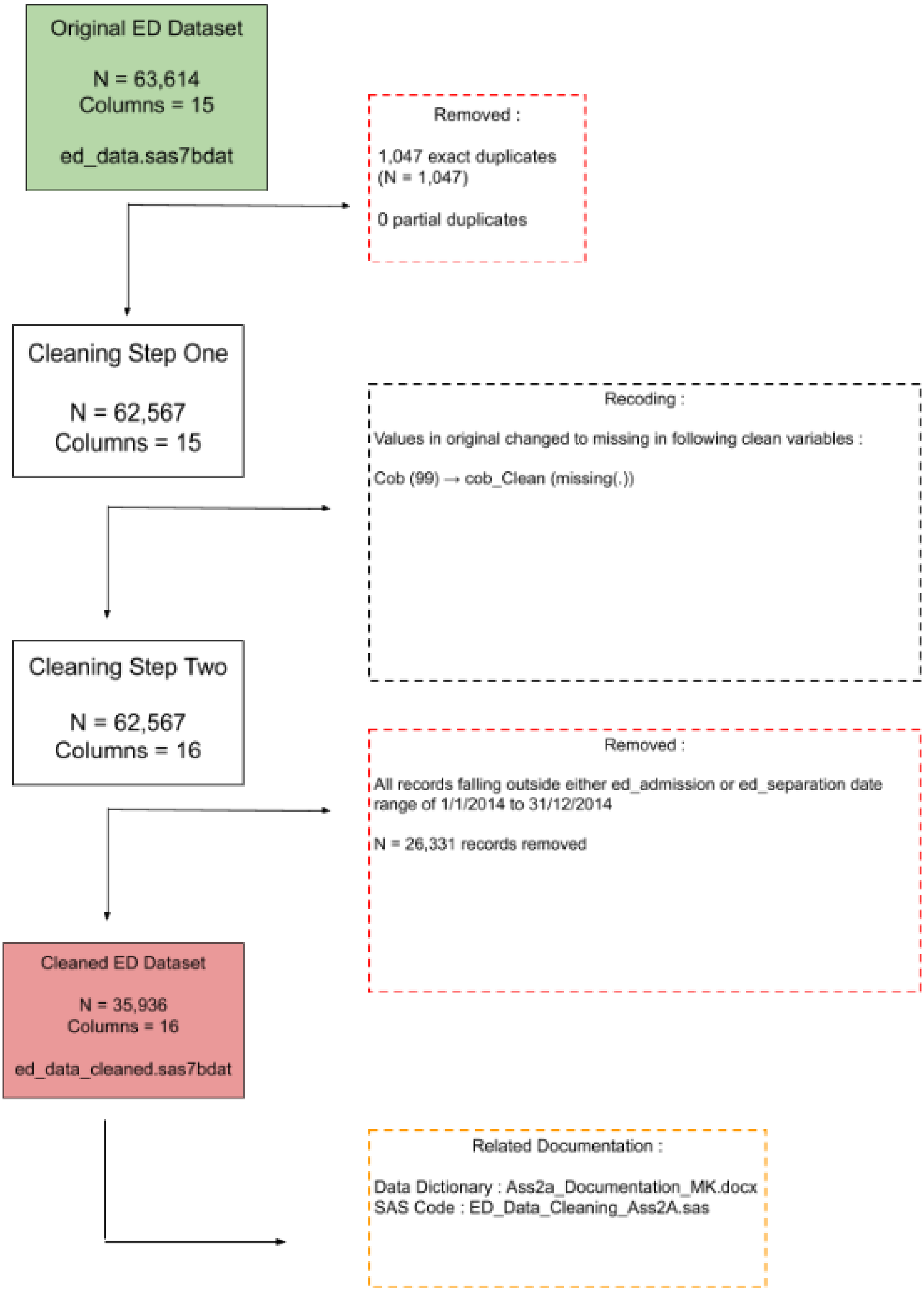
Related Documentation :

Data Dictionary : Ass2a_Documentation_MK.docx
SAS Code : GP_Data_Cleaning_Ass2A.sas

2.3. ED Data Cleaning Notes

1. Check the Data – I examined the contents of the ED dataset. There were 63,614 observations and 15 variables in the original dataset. The variables matched the original data dictionary, and there were no additional variables.
2. Check for Duplicates – I observed 1,047 exact duplicates and decided to remove all 1,047 exact duplicates. I checked for partial duplicates using the ID and Admission Date variables (since it is unlikely for multiple ED admissions in one day) but did not observe any duplicates. Once removed, the dataset now contained 62,567 records.
3. Data Quality Checks – I checked the ed_admission and ed_separation fields to ensure no admissions or separations fell outside the allowable range. I observed that 26,186 records had an ed_admission outside of 2014 and 26,266 observations had an ed_separation outside of 2014. I made the decision to remove all records with either an ed_admission or ed_separation outside the allowable range for the purpose of future analysis. The original file is unchanged and saved, in case future analysis require the use of data outside the allowable date ranges. I checked the frequency tables for categorical variables and the summary statistics for continuous variables. I observed that cob_ed included values of 99 (1%). I made the decision to recode cob_ed as a new variable. For the purpose of this analysis, I was not sure what a value of 99 might represent so I decided to recode these values as missing. The number of these values for each variable was low, so there will be minimal impact on further analysis. Also, the original variable remained unchanged, so any information about these values can be updated in the future. I observed that sex_ed, cob_ed, interpreter, and health_insurance all had missing values, though none had higher than 5%. This is important to note but should not have a large impact on future analysis. Additionally, I checked to ensure that all the Dx fields only included alphabetic and numerical characters. All records followed the $3. format and included only allowable characters. All other variables appeared correct, having only values listed as allowable in the data dictionary. As a result of the new variables needed for recoding, the dataset now contained 16 variables. As a result of excluding ed_admission and ed_separation dates outside of the allowable range, the dataset now contains 35,936 observations.
4. Variable Creation - No new variables were created.
5. Tidy and Save - The final dataset included 35,936 observations and 16 variables. The labels for the new columns were updated. The final cleaned dataset was saved permanently on a local drive to be used in future analysis.

2.4. ED Data Cleaning Flowchart


Original ED Dataset
N = 63,614
Columns = 15
ed_data.sas7bdat
Removed :
1,047 exact duplicates
(N = 1,047)
0 partial duplicates
Cleaning Step One
N = 62,567
Columns = 15
Recoding :
Values in original changed to missing in following clean variables :
Cob (99) → cob_Clean (missing(.))
Cleaning Step Two
N = 62,567
Columns = 16
Removed :
All records falling outside either ed_admission or ed_separation date range of 1/1/2014 to 31/12/2014
N = 26,331 records removed
Cleaned ED Dataset
N = 35,936
Columns = 16
ed_data_cleaned.sas7bdat
Related Documentation :
Data Dictionary : Ass2a_Documentation_MK.docx
SAS Code : ED_Data_Cleaning_Ass2A.sas

3. Research Question

Using the cleaned GP dataset, I wish to describe the demographic, lifestyle, and other characteristics of patients who visited the Medical Plus GP in Sunnyvale. The tables and findings are presented below.

Patients visiting the Medical Plus GP range from 2 to 96 years old, with the average age falling around 46 years old. 50% of the patients are aged between 38 and 55 years old. From the Age Category breakdown in Table 1, only 9.46% of patients visiting the Medical Plus GP are below 30 years old. Similarly, only 14.68% of patients are aged 60 or above. This indicates that the majority (75.86%) of the patients visiting Medical Plus GP fall into the middle-aged categories of 30-59 years old.

2,585 (44.93%) of the patients are male versus 3,169 (55.07%) female. 2,390 (41.93%) of the patients were born in Australia versus 3,310 (58.07%) who were born overseas. From the ABS, in June 2018, there were 98.5 males for every 100 females in NSW[1] . From NSW HealthStats, in 2016, 30.1% of residents had been born overseas[2]. This indicates a patient population that differs from the NSW population, since the patients visiting the Medical Plus GP in Sunnyvale have higher proportions of female patients and patients born overseas when compared to the overall proportions in NSW.

The descriptive statistics for weight and height are in Table 2. Also, in Table 2, the average patient BMI is 26.15 with the median being lower at 24.70. This indicates that about 50% of patients are either underweight or healthy, while about 50% of patients are either overweight or obese. 75% of patients fall under 29.31 BMI, indicating a relatively healthy patient population.

Demographic Variables		
Variables	Observations	Percentage
Age Category		
0 - 29	545	9.46%
30 - 39	1139	19.76%
40 - 49	1813	31.45%
50 - 59	1421	24.65%
60+	846	14.68%
Sex		
Male	2585	44.93%
Female	3169	55.07%
Country of Birth		
Australia	2390	41.93%
Overseas	3310	58.07%

Table 1.

Demographic Variables							
Variables	N	Minimum	25th Percentile	Mean	50th Percentile	75th Percentile	Maximum
Age	5,764.00	2.00	38.00	46.27	46.00	55.00	96.00
Weight (kg)	5,661.00	44.65	62.95	77.73	74.08	90.03	140.54
Height (m)	5,710.00	1.62	1.62	1.72	1.67	1.83	1.83
BMI	5,611.00	15.02	21.45	26.15	24.70	29.31	41.99

Table 2.

Patients visiting the Medical Plus GP in Sunnyvale exhibit signs of unhealthy lifestyles. Observing Table 3, 1,618 (30.23%) of patients have been smokers, while 933 (16.19%) are current smokers. The percentage of current smokers is 3% higher than the NSW average of 13%[3]. 679 (12.46%) of patients are also defined as risky alcohol users, which is defined as drinking more than 2 drinks per day. Compared to the 32.8% of adults who exhibit risky drinking behaviour in NSW[4], the patient population here appears to be exhibit less risky drinking behaviours. Finally, 1,203 (21.44%) of patients are defined as being obese, which is defined as having a BMI 30 or greater. This is very similar to the NSW average of 22.3%[5].

Lifestyle Variables		
Variables	Observations	Percentage
Has Ever Smoked		
No	3734	69.77%
Yes	1618	30.23%%
Current Smoker GP		
No	4831	83.81%
Yes	933	16.19%
Risky Alcohol User		
No	4771	87.54%
Yes	679	12.46%
Obese		
No	4408	78.56%
Yes	1203	21.44%

Table 3.

Finally, other characteristics are listed in Table 4 that provide further information about the patient population visiting Medical Plus GP. Only 1,557 (27.26%) of patients visiting the GP practice had a healthcare card. For billing and administrative purposes, patients should be advised to sign up for a healthcare card.

222 (3.85%) of patients had an adverse reaction to medicine. While low, this still represents an opportunity for improvement in prescription practices. Proper medicine could have prevented both

the negative health outcome for the patient and the resulting utilization of health resources needed to treat the adverse reaction.

2,390 (50%) of patients visiting the Medical Plus GP had headache as the reason for their visit. The next largest reason, nausea, only accounted for 638 (11.07%) of patient visits with all other reasons falling around 5% each respectively. With most reasons for GP visits falling under headache, there might be an opportunity to provide patients the information they need to treat headaches without a GP visit. Various suggestions include an online portal or a nurse hotline that patients can access to seek advice.

Other Characteristics		
Variables	Observations	Percentage
Has Healthcare Card		
No	4155	72.74%%
Yes	1557	27.26%%
Has had adverse reaction to medicine		
No	5542	96.15%
Yes	222	3.85%
Reason for Last GP Visit		
HEADACHE	2390	50.09%
NAUSEA	638	11.07%
DIZZINESS	299	5.19%
SKIN RASH	297	5.15%
TINNITUS	296	5.14%
ITCHING	293	5.08%
VOMITING	267	4.63%
ABDOMINAL PAIN	264	4.58%
HALLUCINATIONS	263	4.56%
PALPITATIONS	260	4.51%

Table 4.

4. References

1. Australia Bureau of Statistics, Australia, accessed 22 July 2020, <https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/3235.0Main%20Features22018?opendocument&tabname=Summary&prodno=3235.0&issue=2018&num=&view=>
2. Healthstats NSW, Australia, accessed 22 July 2020. <http://www.healthstats.nsw.gov.au/Indicator/dem_pop_age/cob_pop_cob_year#:~:text=The%20percentage%20of%20the%20NSW,%2C%20the%20Philippines%2C%20and%20Vietnam.>
3. Healthstats NSW, Australia, accessed 22 July 2020. <http://www.healthstats.nsw.gov.au/Indicator/beh_smocat/beh_smocat?&topic=Health-related%20behaviours&topic1=topic_beh&code=beh>
4. Healthstats NSW, Australia, accessed 22 July 2020. <http://www.healthstats.nsw.gov.au/Indicator/beh_alc_age/beh_alc_age?&topic=Health-related%20behaviours&topic1=topic_beh&code=beh>
5. Healthstats NSW, Australia, accessed 22 July 2020. <http://www.healthstats.nsw.gov.au/Indicator/beh_bmicat/beh_bmicat?&topic=Health-related%20behaviours&topic1=topic_beh&code=beh>