Research Question 1 (Primary care perspective) 30%

The Medical Plus GP is interested in knowing how many adult GP patients attended the local ED in 2014 and their characteristics.

1.1. Create and label a new variable**:**

New variable name	Value and Label	
Agegroup_GP	1	= Under 60 years old
	2	= 60 years old and older

See workdir.gp_data

1.2. Calculate and report the proportion of GP patients who attended the ED in 2014. Comment on the findings.

Total_GP = 5300

GP_attended_ED = 1283

$$\text{Proportion of GP patients who attended the ED in 2014} = \frac{\text{GP}_{\text{attended ED}}}{\text{Total}_{\text{GP}}} \times 100\%$$

$$= \frac{1283}{5300} \times 100\%$$

$$= 24.21\%$$

A rate of 24.21% indicates that nearly a quarter of GP patients also attended the ED in 2014. According to the Australian Bureau of Statistics, in 2021-22 compared to 2020-21, 14.8% compared to 13.4% of Australians aged 15 and over visited a hospital emergency department (ED) and 83.6% compared to 82.4% saw a GP (1). Thus, this is a significant proportion, and it suggests that a large number of patients required emergency care in addition to their GP care. Also, there might be a potential overlap in the services provided by GPs and the ED. Which indicate a need for improved primary care via investigate whether it's essential to visits both GP and ED, and further investigate the reasons behind such visits.

1.3. Calculate total number of monthly ED admissions for all GP patients. Create a figure to show monthly trends of ED admissions and interpret the findings.

Month–Year	total number of monthly ED admissions	Percentage
JAN 2014	482	7.91%
FEB 2014	475	7.80%
MAR 2014	542	8.90%
APR 2014	547	8.98%
MAY 2014	551	9.04%
JUN 2014	550	9.03%
JUL 2014	558	9.16%
AUG 2014	534	8.77%
SEP 2014	485	7.96%
OCT 2014	483	7.93%
NOV 2014	445	7.30%
DEC 2014	440	7.22%
Total	6092	100%

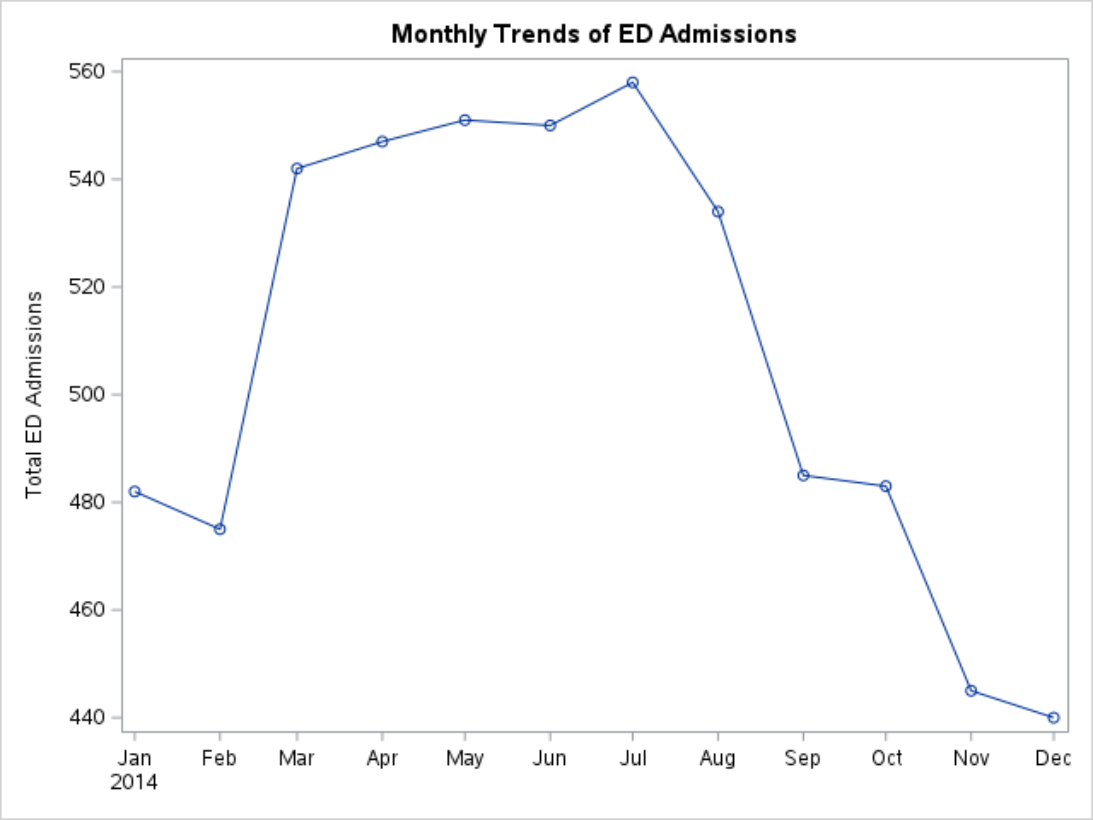


Table and line plot indicates the trend of ED admissions in 2014. The year starts with a relatively stable number of admissions in January and February with 7.91% and 7.80% of the yearly total, respectively. There's a noticeable increase in March (8.90%), which continues to rise slightly in April (8.98%) and peaks in May (9.04%) and June (9.03%). July reaches the highest percentage at 9.16%. A slight decline is observed in August (8.77%), which then drops more significantly in September (7.96%). October's admissions (7.93%) are almost at the same level as September's. November and December show the lowest admissions with 7.30% and 7.22% respectively.

Admissions to ED peak exist in the mid-year, particularly from March to July. This could be associated with several factors, such as flu season, which generally has a spike during colder months. There's a decline in the number of admissions toward the end of the year. This could be due to the holidays and festivities where people tend to avoid hospital visits unless necessary. Although there are spikes and drops, the number of admissions doesn't have drastic fluctuations, indicating a relatively consistent need of emergency services throughout the whole 2014 year.
It would be valuable to review previous years data to identify if these are recurring patterns or unique to 2014.

1.4. Create a table showing differences between patients who did and did not attend the ED in 2014 in terms of socio-demographic characteristics [sex, age group, country of birth, and health care card] and health-related factors [smoking, risky alcohol consumption, obesity, and high blood pressure]. Comment on the findings.

The table below provides a comparison of patients who did and didn't attend the ED in 2014 based on various socio-demographic and health-related factors. Here's an interpretation and commentary on the findings:
Sex:
Significance: The chi-square p-value is < 0.0001, indicating that the relationship between sex and ED attendance is statistically significant.
Findings: Males attending the ED is higher (27.24%) than females (21.80%).

Characteristic	Patient did attend the ED in 2014		Chi-Square Value	Chi-square p-value
	Yes (N=1283)	No (N=4017)		
socio-demographic characteristics				
sex				
Male	644 (27.24%)	1720 (72.76%)	21.0549	< 0.0001
Female	638 (21.80%)	2288 (78.20%)		
Age Group				
Under 60 years old	1091 (24.11%)	3435 (75.89%)	0.177	0.6739
60 years old and older	192 (24.81%)	582 (75.19%)		
Country of Birth				
Australia	586 (26.24%)	1647 (73.76%)	6.4010	0.0114
Overseas	697 (23.20%)	2307 (76.80%)		
Have Healthcare Card				
Yes	579 (34.3%)	1109 (65.70%)	138.2544	< 0.0001
No	701 (19.45%)	2903 (80.55%)		
health-related factors				
Current Smoker				
Yes	245 (32.36%)	512 (67.64%)	34.1843	< 0.0001
No	1014 (22.56%)	3481 (77.44%)		
Risky Alcohol Drinks				
Yes	439 (47.56%)	484 (52.44%)	331.7354	< 0.0001
No	779 (19.08%)	3304 (80.92%)		
Being Obese				
Yes	559 (34.89%)	1043 (65.11%)	145.1726	< 0.0001
No	700 (19.41%)	2907 (80.59%)		
Have high blood pressure				
Yes	548 (33.21%)	1102 (66.79%)	105.8806	< 0.0001
No	735 (20.14%)	2915 (79.86%)		

The chi-square value is a measure of how much the observed frequencies deviate from the expected frequencies, and the p-value is the probability of obtaining such a deviation or more extreme by chance. A common rule of thumb is to reject the null hypothesis of no association if the p-value is less than 0.05, which means that the deviation is unlikely to be due to chance. When the p-value (Prob) is <.0001, which is less than the usual significance level 0.05, so the null hypothesis is rejected. This means that there is a statistically significant association (e.g. between ED_attended and sex).

Age Group:
Significance: The chi-square p-value is 0.6739, suggesting no statistically significant relationship between age group and ED attendance.
Findings: Both age groups have roughly the same proportions for attending and not attending the ED

Country of Birth:
Significance: With a p-value of 0.0114, the difference is statistically significant.
Findings: Those born in Australia are slightly more likely to attend the ED (26.24%) compared to those born overseas (23.20%).

Have Healthcare Card:
Significance: The relationship between having a healthcare card and attending the ED is statistically significant with a p-value < 0.0001.
Findings: A larger proportion of individuals with a healthcare card attend the ED (34.3%) compared to those without (19.45%).

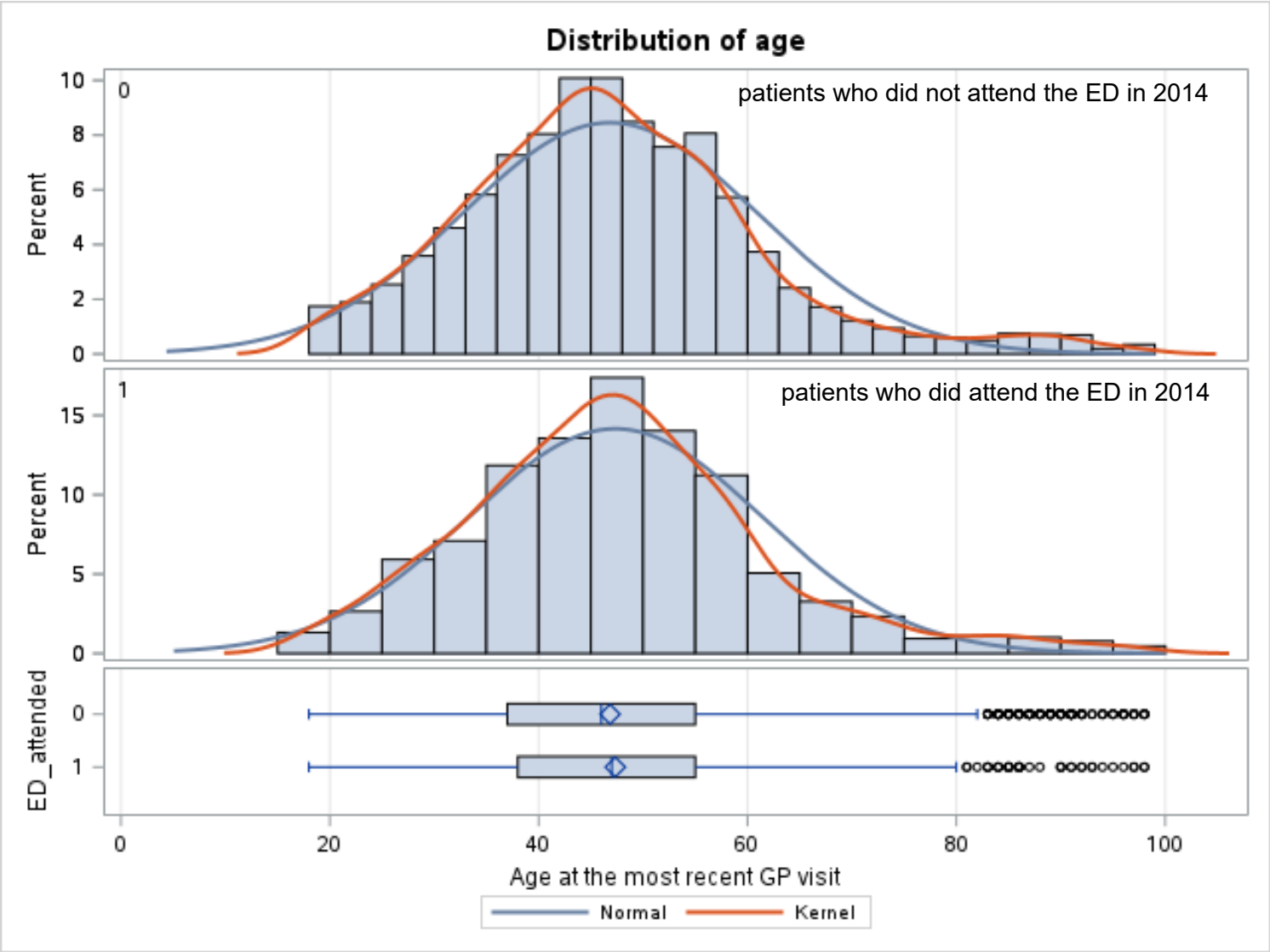
Current Smoker:
Significance: The relationship between smoking status and ED attendance is statistically significant.
Findings: Current smokers are more likely to attend the ED (32.36%) compared to non-smokers (22.56%).

Risky Alcohol Drinks:
Significance: The relationship between risky alcohol consumption and ED attendance is highly significant.
Findings: A much higher proportion of individuals with risky alcohol consumption habits attend the ED (47.56%) compared to those without such habits (19.08%).

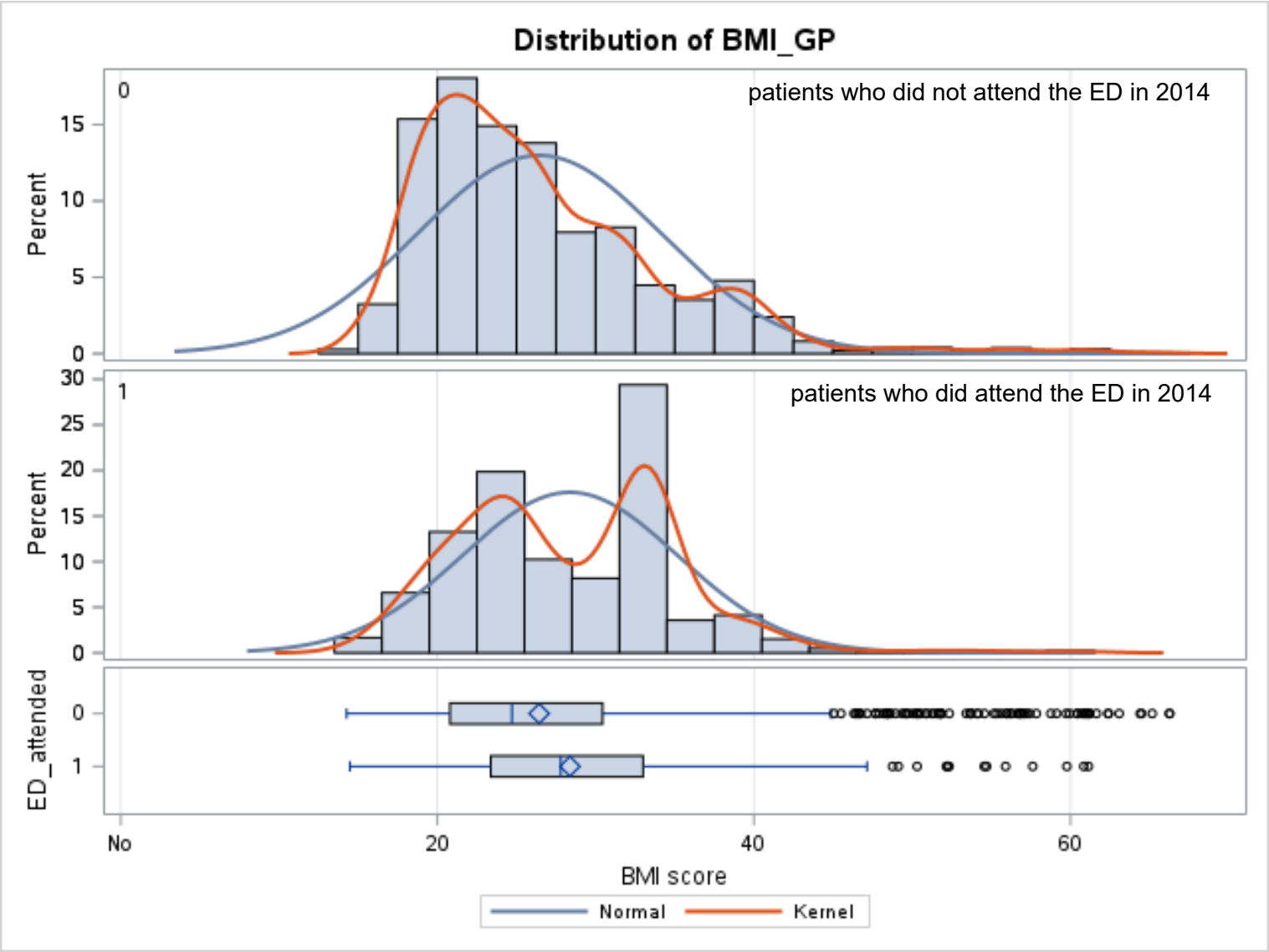
Being Obese:
Significance: Obesity's relationship with ED attendance is statistically significant.
Findings: Obese individuals are more likely to attend the ED (34.89%) compared to non-obese individuals (19.41%).

Have High Blood Pressure:
Significance: The relationship between having high blood pressure and attending the ED is statistically significant with a p-value < 0.0001.
Findings: Individuals with high blood pressure have a higher rate of attending the ED (33.21%) compared to those without high blood pressure (20.14%).

In summary, while sex shows a significant relationship with ED attendance, the practical significance is relatively minor. Healthcare cardholders have a notably higher likelihood of attending the ED. This could be due to the financial subsidy and convenience for the cardholders. All the health-related factors like smoking, risky alcohol consumption, obesity, and high blood pressure show a clear association with ED attendance. This underscores potential public health concerns.

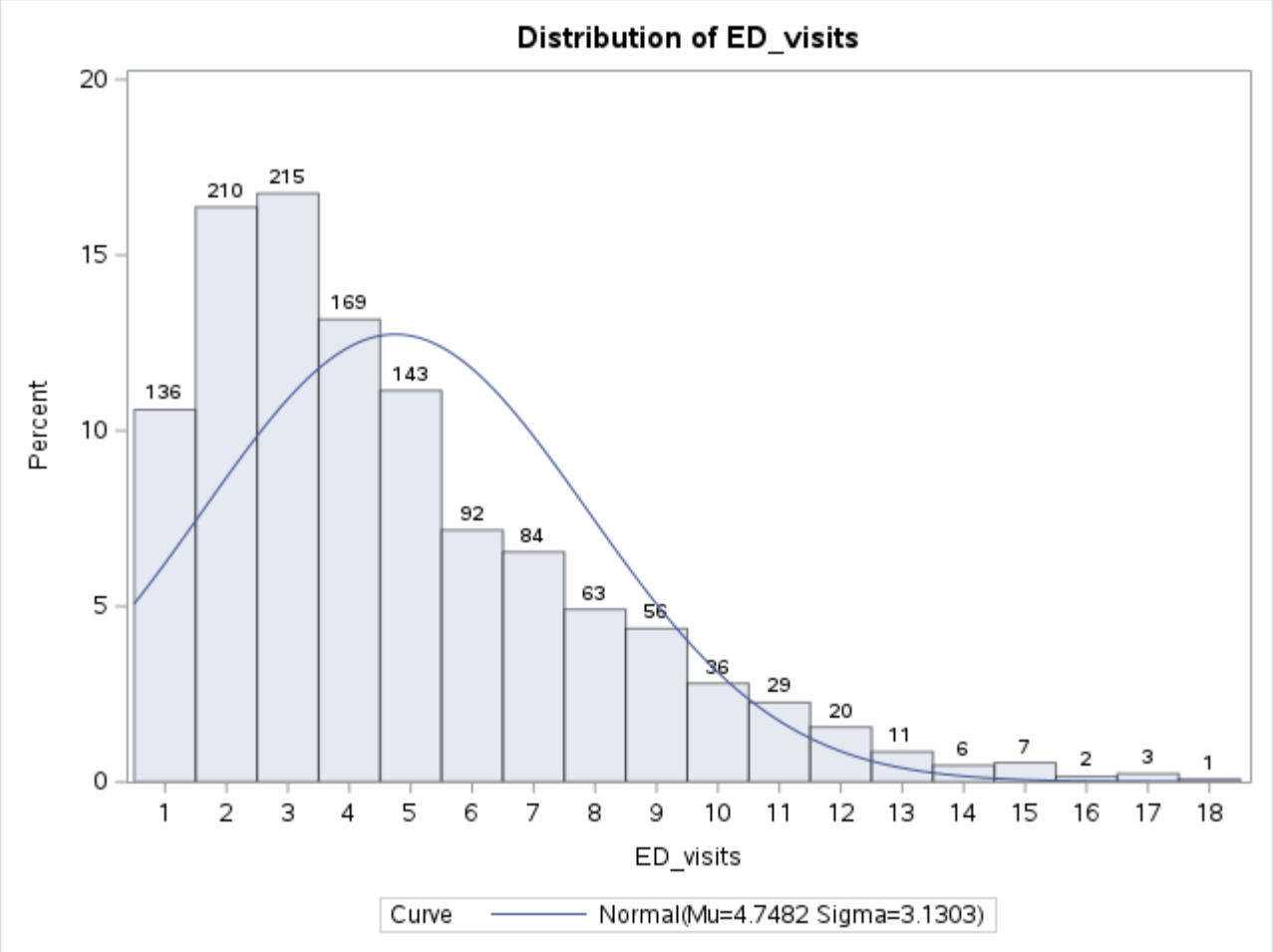


For those who did not attend the ED (`ED_attended=0`), the mean age is approximately 46.83 with a standard deviation of approximately 14.16. For those who attended the ED (`ED_attended=1`), the mean age is approximately 47.34 with a standard deviation of approximately 14.09. The mean value of the variable is slightly higher in the group that attended the ED in 2014 compared to those who didn't, but the difference is relatively small (0.5059 units). 95% Confidence Limits for this difference (Pooled):is -1.3949 to 0.3831. The p-value associated with the t-test is above the common significance level (0.05). Therefore, the observed difference in means between the two groups is not statistically significant, regardless of whether equal or unequal variances are assumed.



For those who did not attend the ED (`ED_attended=0`), the mean BMI is approximately 26.43 with a standard deviation of approximately 7.68. For those who attended the ED (`ED_attended=1`), the mean BMI is approximately 28.38 with a standard deviation of approximately 6.80. The difference in means (ED_attended 1 - ED_attended 0) is approximately -1.9452. This suggests that, on average, the BMI of those who attended the ED is almost 1.95 units higher than those who didn't. Both t-tests are statistically significant ($p < 0.0001$). There is a statistically significant difference in the average BMI between patients who attended the ED and those who did not.Patients who attended the ED had, on average, a higher BMI compared to those who did not.

1.5. Among GP patients who visited the ED, calculate the total number of ED admission for each person in 2014. Describe the distribution of numbers of ED admission using a histogram and descriptive statistics. Comment on the findings**.**



Based on histogram and the analysis results, here is the summary of the ED admission data:
Sample Size (N): 1283 patients
Total Number of ED Visits (Sum Observations): 6092 visits
Mean ED Visits: 4.748 visits per patient. However, the distribution is right-skewed, meaning there are some patients who visit the ED a lot more than the average, pushing the mean up.
Median ED Visits: 4 visits. Most GP patients (around 50%) have been to the ED 4 times or fewer, as indicated by the median. But there are outliers who visit much more frequently, with the maximum number of visits being 18.
Most Common Number of Visits (Mode): 3 visits
Standard Deviation: 3.130
Variance: 9.7985
Coefficient of Variation: 65.92%, which indicates a high relative variability in the number of visits.
Skewness: 1.1199, which indicates a right-skewed distribution (i.e., a longer tail on the right side of the distribution)
Kurtosis: 1.061, which indicates a relatively flat peak (platykurtic) as compared to a normal distribution. Given the skewness and the fact that the mode is 3 visits, many patients visit the ED only a few times, but a smaller number of patients have a high number of visits, which can be considered as frequent visitors.
Kurtosis: 1.061, which indicates a relatively flat peak (platykurtic) as compared to a normal distribution.
The Student's t-test result indicates that the mean number of ED visits is significantly different from 0. With a t-statistic of 54.33 and a p-value of < 0.0001, we can confidently reject the null

hypothesis. However, consider the 0 time of ED visit is not included in it doesn't make practical sense.

1.6. Continue with the results of step 1.5, select GP patients who had many ED visits (i.e. top 25% percentile). Examine and report socio-demographic and health-related characteristics of these patients.

Characteristic	GP Patient did attend the ED in 2014			
	All (N=1283)		High ED visits patients (top 25%) (N=410)	
socio-demographic characteristics	Number	Col Pct	Number	Col Pct
sex				
Male	644	50.23%	202	49.39%
Female	638	49.77%	207	50.61%
Age Group				
Under 60 years old	1091	85.04%	339	82.68%
60 years old and older	192	14.96%	71	17.32%
Country of Birth				
Australia	586	45.67%	234	57.07%
Overseas	697	54.33%	176	42.93%
Have Healthcare Card				
Yes	579	45.23%	326	79.71%
No	701	54.77%	83	20.29%
health-related factors				
Current Smoker				
Yes	245	19.46%	72	17.96%
No	1014	80.54%	329	82.04%
Risky Alcohol Drinks				
Yes	439	36.04%	342	88.60%
No	779	63.96%	44	11.40%
Being Obese				
Yes	559	44.40%	300	73.53%
No	700	55.60%	108	26.47%
Have high blood pressure				
Yes	548	42.71%	291	70.98%
No	735	57.29%	119	29.02%

The table in 1.6 presents the difference in characteristics between all GP Patient who did attend the ED in 2014 and the High ED visits patients who attend the ED 6 or more than 6 times (N=410). When examining the column percent (Col Pct), High ED visits patients are more likely:

to be 60 years old and older than all GP patients who attended the ED (17.32% vs 14.96%).
to be born in Australia than all GP patients who attended the ED (57.07% vs 45.67%).
to have a healthcare card than all GP patients who attended the ED (79.71% vs 45.23%).
to have risky alcohol drinks than all GP patients who attended the ED (88.60% vs 36.04%).
to be obese than all GP patients who attended the ED (73.53% vs 44.4%).
to have high blood pressure than all GP patients who attended the ED (71.0% vs 42.71%).

Research Question 2 (ED care perspective) 50%

The Sunnydale ED examines the quality of recording of adult patient smoking, risky alcohol consumption and obesity in the ED data, using the following ICD-10-AM codes:

- Smoking: '**F17**', '**Z72**'
- Risky alcohol consumption: '**F10**'
- Obesity: '**E66**'

2.1. Create three variables to flag ED records with these behaviours being recorded in any diagnosis field.
Please name these new variables as below

New variable name	Value and Label
smoker_flag	0=No 1=Yes, smoker
risky_alcohol_flag	0=No 1=Yes, drinker
obesity_flag	0=No 1=Yes, obese

See workdir.flaged_ed_data

2.2. Classify whether the patient smokes, drinks alcohol at risky level or is obese, if these risk factors are recorded in any ED records for a patient. Calculate and report the prevalence of smoking, risky alcohol consumption and obesity among ED patients.
Please name these new variables as below

New variable name	Value and Label
smoker_ED	0=No 1=Yes, smoker
risky_alcohol_ED	0=No 1=Yes, drinker
obesity_ED	0=No 1=Yes, obese

See workdir.classified_ed_data

Characteristic		Patient did recorded in 2014 ED data (N=5,637)	
health-related factors		Number	Col Pct
Current Smoker			
	Yes	879	15.59%
	No	4758	84.41%
Risky Alcohol Drinks			
	Yes	1130	20.05%
	No	4507	79.95%
Being Obese			
	Yes	1520	26.96%
	No	4117	73.04%

Out of all the ED patients, 15.59% are current smokers, 84.41% are not current smokers. This indicates that a majority of the ED patients do not currently smoke. 20.05% of the ED

patients are identified as having risky alcohol consumption habits. Correspondently, 79.95% of the ED patients do not consume alcohol at risky levels. Over a quarter of the ED patients (26.96%) are obese, indicating that obesity might be a common health-related factor in this patient group.
It can be inferred that while the majority of these patients are not current smokers, do not engage in risky alcohol consumption, and are not obese, a significant proportion still presents with these health concerns.

2.3. Examine whether there are any differences between ED patients who did and did not visit a GP in terms of sex, age, country of birth, private health insurance, smoking, risky alcohol consumption and obesity. You can categorise patient age into two groups (under 60 /60 and older). Interpret your findings.

Sex:
Of the patients who visited a GP, males represented 50.23% while females accounted for 49.77%. In contrast, of those who did not visit a GP, males comprised 51.10% and females 48.90%. The difference in the proportion of males and females between the two groups (those who did and did not visit a GP) is very small. The chi-square test value is 0.2968 with a p-value of 0.5859. This denotes no statistically significant sex-based difference between ED patients relative to their GP visits.

Age Group:
For those who saw a GP, 84.33% were below 60 years old and 15.67% were 60 years or older. In contrast, among non-visitors, 71.96% were under 60 and 28.04% were 60 or older. There's a notable difference in age group distribution between the two groups. This suggests a difference in age distribution between the groups. A chi-square value of 80.4787 and a p-value less than 0.0001 confirm this statistically significant distinction.

Country of Birth:
Of those attending a GP, 45.67% were Australian-born and 54.33% were born overseas. Among the patients who did not visit a GP, a vast majority (77.14%) were born in Australia and only 22.86% were born overseas. There's a clear distinction between the two groups based on the country of birth. The chi-square test yielded a value of 459.7318 with a p-value of <0.0001 suggest a marked difference in birthplace distribution between the two groups.

Health Insurance:
A slight majority of both groups had health insurance, with 53.99% of GP visitors and 54.58% of non-visitors insured. With a chi-square value of 0.1338 and a p-value of 0.7145, the data suggests no significant distinction based on insurance status.

Current Smoker:
Within the GP-visited cohort, 27.67% reported risky alcohol consumption, while this figure was 17.80% for non-visitors. With a chi-square value of 60.2301 and a p-value less than 0.0001, there exists a significant divergence in this behavior between the groups.

Risky Alcohol Drinks:
Of the GP visitors, 27.67% reported risky alcohol consumption, while this figure was 17.80% for non-visitors. A higher proportion of patients who visited a GP consume alcohol at risky levels compared to those who did not visit a GP. The chi-square value is 60.2301 with a p-value of <0.0001, there exists a statistically significant difference in terms of risky alcohol consumption between the groups.

Being Obese:
Obesity rates between the two groups were comparable, with 25.49% of GP visitors and 27.40% of non-visitors categorized as obese. The chi-square value is 1.8414 with a p-value of 0.1748, indicating that there is no statistically significant difference between the groups concerning obesity status.

In summary, there are statistically significant differences emerged between ED patients who did and did not visit a GP in terms of age, country of birth, smoking, and alcohol consumption, as their p-values are all less than 0.0001. This means that these characteristics are associated with GP visits among ED patients. Conversely, sex distribution, health insurance status and obesity status do not significantly differ between ED patients based on GP visits, as indicated by p-values greater than 0.05.

Characteristic	ED patients who did or did not visit a GP in 2014					
	Did (N=1283)		Did Not (N=4354)			
socio-demographic characteristics	Number	Col Pct	Number	Col Pct	Chi-Square Value	Chi-square p-value
sex						
Male	644	50.23%	2206	51.10%	0.2968	0.5859
Female	638	49.77%	2111	48.90%		
Age Group						
Under 60 years old	1082	84.33%	3133	71.96%	80.4787	< 0.0001
60 years old and older	201	15.67%	1221	28.04%		
Country of Birth						
Australia	586	45.67%	3203	77.14%	459.7318	< 0.0001
Overseas	697	54.33%	949	22.86%		
Have Health Insurance						
Yes	676	53.99%	2300	54.58%	0.1338	0.7145
No	576	46.01%	1914	45.42%		
health-related factors						
Current Smoker						
Yes	319	24.86%	560	12.86%	108.4544	< 0.0001
No	964	75.14%	3794	87.14%		
Risky Alcohol Drinks						
Yes	355	27.67%	775	17.80%	60.2301	< 0.0001
No	928	72.33%	3579	82.20%		
Being Obese						
Yes	327	25.49%	1193	27.40%	1.8414	0.1748
No	956	74.51%	3161	72.60%		

2.4. Calculate overall sensitivity (Sn) and specificity (Sp) of the **recording of patient smoking in the ED data**, using patient smoking information in the GP data as the gold standard. Comment on overall quality of ED data on patient smoking.

Sensitivity (Sn) and specificity (Sp) are calculated when comparing a diagnostic test (in this case, recording of patient smoking in the ED data) against a gold standard (patient smoking information in the GP data).
After merge the gp_data (with smoking information as the gold standard) with ed_data_classified (with smoker_flag), create a 2x2 contingency table for smoking status.

		smoker in ED data	
		NO	YES
smoker in GP data	NO	893	121
	YES	49	196

True Positive (TP): Patients who are smokers in both gp_data and ed_data – 196.
True Negative (TN): Patients who are non-smokers in both datasets – 893.
False Positive (FP): Patients who are non-smokers in gp_data but smokers in ed_data – 121.
False Negative (FN): Patients who are smokers in gp_data but non-smokers in ed_data - 49.

$$\text{Sensitivity (Sn)} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{196}{196 + 49} = 77.17\%$$

$$\text{specificity (Sp)} = \frac{\text{TN}}{\text{FP} + \text{TN}} = \frac{893}{121 + 893} = 88.07\%$$

Sensitivity (Sn): The proportion of true positives out of the number of actual positives. It measures how often the diagnostic test correctly identifies the condition (smoking, in this case) when it's actually present according to the gold standard. A sensitivity of 77.17% suggests that the ED data correctly identified 77.17% of the smokers as identified in the GP data (gold standard). Accordingly, approximately 23% of smokers were not identified as such in the ED data.

Specificity (Sp): The proportion of true negatives out of the number of actual negatives. It measures how often the diagnostic test correctly identifies the absence of the condition when it's actually absent according to the gold standard. A specificity of 88.07% means that the ED data correctly identified 88.07% of the non-smokers as identified in the GP data. This suggests that ED data may incorrectly classify approximately 12% of non-smokers as smokers.

Overall, these results suggest that the ED data on patient smoking is fairly good. Both sensitivity and specificity are reasonably high. However, there is a very high rate of missing data. In this analysis 78% of the data were missing due to sample size is 1259 and frequency missing is 4378. With such a high level of missing data, the analysis might be biased. This is because the data included in the analysis may not be representative of the overall population, which could lead to biased estimates of sensitivity and specificity. Meanwhile, the missing data limit the generalizability of your results. The results might only apply to the specific group of patients whose data was not missing.

2.5. Repeat calculation of Sn and Sp of the recording of smoking in ED data, **separately for each** patient's sex, age group, country of birth and private health insurance (i.e. stratified by sociodemographic factors). Comment on whether recording of smoking information in ED data differs by patient sociodemographic characteristics.

patient smoking in the ED data							
socio-demographic characteristics	TP	TN	FP	FN	missing	Sensitivity (Sn)	specificity (Sp)
sex							
Male	116	410	70	31	78%	78.91%	85.42%
Female	80	482	51	18	77%	81.63%	90.43%
Age Group							
Under 60 years old	174	746	97	45	75%	79.45%	88.49%
60 years old and older	22	147	24	4	86%	84.62%	85.96%
Country of Birth							
Australia	48	453	68	9	85%	84.21%	86.95%
Overseas	148	440	53	40	59%	78.72%	89.25%
Have Health Insurance							
Yes	82	407	53	25	77%	76.64%	88.48%
No	106	457	64	24	78%	81.54%	87.72%

This table examines the sensitivity and specificity of smoking information recorded in Emergency Department (ED) data, considering various patient socio-demographic characteristics. The analysis explores differences based on patient sex, age group, country of birth, and health insurance status. The sample size was 1259 patients. Despite notable sensitivity and specificity variations, the presence of missing data is the major limitation. It might cause potential bias and make these differences not carry clinical significance. In addition, cautious interpretation and further investigation into underlying factors is necessary.

Sex-based Comparison: Sensitivity for females was slightly higher (81.63%) compared to males (78.91%), while specificity for females was also higher (90.43%) than for males (85.42%). Considering the missing data, these modest sex-based differences may not carry clinical significance. Possible reasons for variations include healthcare-seeking behaviors and provider-patient interactions influenced by gender norms.

Age Group Comparison: Sensitivity was slightly higher for patients aged 60 and older (84.62%) compared to those under 60 (79.45%), while specificity was slightly higher for patients under 60 (88.49%) compared to those aged 60 and older (85.96%). Age-related health risks might contribute to these differences, but elucidation of the underlying mechanisms needs further research.

Country of Birth Comparison: Sensitivity was higher for patients born in Australia (84.21%) compared to those born overseas (78.72%), while specificity was slightly higher for patients born overseas (89.25%) compared to those born in Australia (86.95%). Familiarity and acceptability of cultural as well as healthcare-seeking behaviors formed since childhood might be the reasons of this pattern.

Health Insurance Comparison: Sensitivity was slightly higher for patients without health insurance (81.54%) compared to those with health insurance (76.64%), while specificity was consistently for patients with health insurance (88.48%) and those without health insurance (87.72%).

2.6. What suggestions do you have for the ED manager for improving their ED screening and recording of patient smoking status based on your findings from 2.4 and 2.5?

Smoking status in ED data is determined by ICD-10-AM codes such as 'F17' (Mental and behavioral disorders due to use of tobacco) and 'Z72' (Problems related to lifestyle). These codes are not necessarily routinely or universally applied for every patient who is a smoker. It's very possible that this coding system is under-utilized, and therefore, many smokers aren't identified as such in the ED data. In emergency departments, healthcare providers often prioritize the most acute and urgent medical conditions. Less acute issues, like smoking status, might not be fully addressed or recorded during the visit, especially if they aren't directly related to the presenting issue. In addition, using ICD-10-AM codes to define smoking status may not capture all patients who are smokers. For instance, if a patient is not currently suffering from a tobacco-related mental or behavioral disorder or if their smoking is not considered to be a problem related to lifestyle, then they might not be assigned the relevant ICD-10-AM codes even if they are indeed smokers. Therefore, while ICD-10-AM codes can provide useful information, they should not be the sole source of data for determining patients' smoking status.

To improve the completeness and accuracy of the ED data on patient smoking status, other data sources or methods for identifying smokers might need to be incorporated. In addition, further education and training of data collection and recording may be required to ensure accurate reporting of smoking status. Regular audits and feedback on data quality can also be used to improve the accuracy and reliability of the data.

While variations in sensitivity and specificity across patient characteristics were observed, the presence of high missing data underscores the need for cautious interpretation. Possible influencing factors, such as healthcare-seeking behaviors, cultural biases, and provider-patient dynamics, should be investigated further to enhance the accuracy of smoking data recording in ED settings.

Future research should delve into specific healthcare provider behaviors, communication styles, and cultural biases that could impact smoking data recording. Additionally, exploring interventions to minimize missing data and improve data quality is essential.

Research Question 3 (Tobacco control perspective) 20%

The Sunnydale Population Health Department is developing a trial for quitting smoking which includes the use of smoking cessation medicines in addition to behavioural therapies. Smoking cessation medicines include varenicline, bupropion and nicotine replacement therapy (NRT). To inform the design of the trial, the Department investigates smoking prevalence using three data sources and assesses the baseline uptake of smoking cessation medicines using PBS data linked to GP and ED data.

Information about Sunnydale residents who had a dispensing of a smoking cessation medicine is contained in the PBS data, with medicines coded using Anatomical Therapeutic Chemical (ATC) classification. In 2014 varenicline, bupropion and NRT patches were subsided by PBS, only for smoking cessation indication and not for other means. The ATC codes to identify these therapies in PBS data include:

- N07BA01 – NRT patches
- N06AX12– Bupropion
- N07BA03 – Varenicline

Analyses for Question 3 leverage on your prior analyses for Questions 1 and 2.

3.1. Create a cohort of Sunnydale residents who smoke using information from the GP and ED data sources. How many smokers could you identify in the GP data alone, ED data alone, and using a combination of both GP/ED data sources.
I combined both GP/ED data as well as PBS data (See workdir.sunnydale_smoker).

Cohort of Sunnydale smoker (N=1,457)				
smoker in ED data	smoker in GP data	smoker in PBS data	number	percent
.	.	YES	17	1.17%
.	YES	.	437	29.99%
.	YES	YES	75	5.15%
NO	YES	.	37	2.54%
NO	YES	YES	12	0.82%
YES	.	.	514	35.28%
YES	.	YES	48	3.29%
YES	NO	.	101	6.93%
YES	NO	YES	20	1.37%
YES	YES	.	152	10.43%
YES	YES	YES	44	3.02%
		total	1457	100.00%

After examining a cohort of 1,457 individuals smoker from Sunnydale.

Data identified in single data set alone:

437 individuals (29.99% of the cohort) were identified as smokers through GP records without corresponding evidence in ED or PBS data. This considerable number indicates that a significant fraction has their smoking habits recorded at GP clinics but not elsewhere.

514 individuals (35.28%) were tagged as smokers based solely on ED data, without corresponding records in the GP or PBS databases. This could imply that a significant portion of the cohort has had episodes related to smoking in emergency situations

without consistent documentation in primary care or pharmacological intervention.

17 individuals (1.17% of the cohort) were identified as smokers exclusively through the PBS data. This suggests a segment of the population depends on smoking cessation medications without smoking records in the ED or GP databases.

Overlap of two or more data sets:

Overlap of GP and PBS Data: 75 patients (5.15% of the cohort) were identified in both the GP and PBS datasets, highlighting those who, having a smoking record in the GP data, also resorted to pharmacological aid for smoking cessation.

Overlap of ED and PBS Data: 48 patients (3.29%) have smoking records in the ED and PBS databases but not in GP records. This subset likely represents smokers who experienced emergencies, possibly linked to smoking, and also pursued pharmacological cessation aids, all while bypassing or not being recorded in regular GP visits.

Overlap of ED and GP Data: 152 patients (10.43%) were recorded as smokers in both the ED and GP databases but not in the PBS data. This sizable number underscores individuals who have their smoking habits well-documented in clinical settings but haven't necessarily sought or been prescribed pharmacological cessation interventions.

Overlap of ED and PBS, but Not in GP: 20 individuals (1.37%) had records in the ED and PBS datasets but not in the GP database. These individuals likely have acute episodes related to smoking and have also sought pharmaceutical interventions, yet have no corresponding records in GP datasets.

Overlap Across All Three Datasets: 44 individuals (3.02%) were identified as smokers across all three datasets. This subset, having consistent records across various health touchpoints, can be considered with high confidence as smokers who actively sought cessation aids.

Conflict results:

Confirmed Non-smokers in ED with Smoking Record in GP Data: 37 individuals (2.54%) were found to be non-smokers in ED data but had a smoking record in the GP database. This discrepancy might arise due to time lags between quitting and recording, or inaccuracies in data collection.

Confirmed Non-smokers in ED with Records in Both GP and PBS: 12 patients (0.82%) were identified as non-smokers in ED data but had records in both the GP and PBS datasets. The reasons could be similar to the previous point, with the added dimension of these individuals seeking smoking cessation aids.

Confirmed Smokers in ED but Not in GP Data: 101 individuals (6.93%) were found to be smokers in ED data but did not have a corresponding record in the GP database. The reasons could vary, from these individuals not visiting GPs often, to lapses in recording at GP clinics.

Summary:

The utilization of multiple datasets provides a more comprehensive understanding of the smoking habits of the Sunnydale cohort. While the GP data alone identifies a significant percentage of the cohort as smokers, ED data and PBS data also identify the smokers. The overlaps offer valuable insights into the health-seeking behaviors of smokers and the documentation in different healthcare settings.

Moreover, it is essential to acknowledge the limitations of a single dataset. Notably, there are some smokers identified by one dataset and not the others, suggesting potential gaps in data recording (as I mentioned before about ED data is determined by ICD-10-AM codes) or the presence of exclusive smokers' categories, such as those only seeking pharmacological help. This underlines the importance of holistic data analysis, leveraging multiple sources to arrive at a more nuanced understanding of the smoking patterns of a population.

While some conflict results can be attributed to data collection lapses or the dynamic nature of smoking habits, they also point towards the diverse health-seeking behaviors and pathways of smokers in Sunnydale. For example, those identified only through PBS data might be self-initiating their cessation using OTC without seeking help from doctor. Thus, there is no consistent documentation in other clinical settings.

In conclusion, considering information from GP, ED, and PBS datasets, presents a a more comprehensive, nuanced understanding on smoking habits within the Sunnydale cohort. It will enabling more effective health interventions, policy recommendations, and patient management strategies.

3.2. Examine PBS data against the cohort defined in Step 3.1 and comment on the value of PBS data as an additional data source (i.e. on top of GP and ED data) to identify people who smoke and who were not identified in GP or ED data.

Unique Identifications from PBS Data:

Out of the cohort, 17 individuals (1.17% of the total) were exclusively identified as smokers through the PBS data. These individuals had no corresponding smoking records in the GP or ED datasets.

Overlap between PBS and Other Datasets:

75 individuals were identified as smokers in both the GP and PBS datasets.

48 individuals were found in both the ED and PBS datasets.

44 individuals were consistently identified across all three datasets (GP, ED, and PBS).

The PBS dataset managed to identify a subset of the population that wasn't captured in the GP or ED data. The individuals identified through the PBS data are those who sought out smoking cessation medications. This reflecting active cessation efforts of these individuals. For those identified across multiple datasets, the PBS data acts as an additional validation point. For instance, those identified in both the GP and PBS datasets or across all three datasets can be regarded with a higher degree of confidence as current or former smokers. The presence of smokers in the PBS data but not in GP or ED data can highlight potential gaps or lapses in recording smoking habits in regular clinical settings. This makes the PBS data invaluable as it capturing those seeking pharmacological interventions for smoking cessation, provides insights into the health-seeking behaviors of the cohort.

3.3. For the cohort created in Step 3.1, calculate the proportion of smokers who used any of the smoking cessation therapies, as well as each of the three individual medicines in 2014. Comment on your findings.

Smoking Cessation Therapy (N=216)			
	number	percent in PBS	percent in Sunnydale smoker Cohort
NRT patches	60	27.78%	4.12%
Bupropion	1	0.46%	0.07%
Varenicline	158	73.15%	10.84%

The table outlines the distribution of different smoking cessation therapies used by a segment of three individual medicines of the Sunnydale smoker cohort.

Varenicline was the most commonly used therapy, with 158 individuals (or 73.15% of those using cessation therapies) using it. In the broader context, about 10.84% of all smokers in Sunnydale are using Varenicline as a cessation method. This could be due to its efficacy, lesser side effects, affordability

NRT patches were the second most popular choice, with 60 individuals (or 27.78% of those using cessation therapies) utilizing them. When comparing to the entire Sunnydale smoker cohort, 4.12% of all smokers in Sunnydale used NRT patches for smoking cessation.

Bupropion had minimal usage among the cohort, with only 1 person opting for this therapy.

It is worth noting that although the percentages within the PBS are high, it's essential to contextualize these numbers. For example, although 73.15% of people in the PBS use Varenicline, this constitutes only 10.84% of the entire Sunnydale smoker cohort.

In addition, it's crucial to understand the reasons behind such preferences. For example if Varenicline is indeed more effective, it worth to be promoted more widely. However, if the preference is driven by other non-efficacy factors (like cost or promotion), then efforts should be made to ensure that all smokers are well-informed about all available options and can choose what's best for their individual needs.

GP data dictionary

Variable	Description	Variable type	Format name	Allowable entries
ID	Unique person ID	Number		
GP_last	Date of most recent GP visit	Number	DDMMYY10.	Dates in the range 01/01/2014 – 31/12/2014
age	Age at the most recent GP visit in 2014	Number		
sex	Sex of patient	Number	SEXF.	1=male 2=female
cob	Country of birth	Number	COBF.	1= Born in Australia 2= Born overseas
healthcare_card	Have a healthcare card[1]	Number	YNF.	1= Yes 0= No
drinks_day	Number of alcohol drinks per day	Number		Invalid if >20
height	Body height (metres)	Number		Invalid if <0.55m or >2.40m
weight	Body weight (kilograms)	Number		Invalid if <5.0kg or >270kg
adverse_reaction	Had any reaction to any medication	Number	YNF.	1= Yes 0= No
syst_bp	Systolic blood pressure (mmHg)	Number		
diast_bp	Diastolic blood pressure (mmHg)	Number		
reason	Reason for the most recent GP visit	Character		HEADACHE NAUSEA TINNITUS VOMITING ITCHING ABDOMINAL PAIN DIZZINESS SKIN RASH PALPITATIONS HALLUCINATIONS
Smoke_current_GP	Being a current smoker	Number	YNF.	1= Yes 0= No
Risky_alcohol_GP	Have two or more alcohol drinks per day	Number	YNF.	1= Yes 0= No
BMI_GP	BMI score (weight kg / height squared)	Number	YNF.	1= Yes 0= No
Obese_GP	Being obese (BMI>=30)	Number	YNF.	1= Yes 0= No
HighBP_GP	Have high blood pressure (>=135/85mmHg)	Number	YNF.	1= Yes 0= No

Agegroup_GP		Number		1 = Under 60 years old 2 = 60 years old and older
ED_attended		Number		1 = patients who did attend the ED in 2014 0 = patients who did not attend the ED in 2014

ED data dictionary

Variable	Description	Variable type	Format name	Allowable entries
ID	Unique person ID	Number		
ed_admission	Date of ED attendance	Number	DDMMYY10.	Dates in the range: 01/01/2014 – 31/12/2014
ed_separation	Date of ED separation	Number	DDMMYY10.	Dates in the range: 01/01/2014 – 31/12/2014
age_ed	Age of patient at ED attendance	Number		
sex_ed	Sex of patient	Number	SEXF.	1=male 2=female
cob_ed	Country of birth	Number	COBF.	1= Born in Australia 2= Born overseas
interpreter	An interpreter is required	Number	YNF.	1= Yes 0= No
health_insurance	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 diagnoses (ICD-10-AM codes)	Character		International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification 8th edition
separation_mode	Status at separation from emergency department	Number	SEPMODEF.	1 = Admitted to hospital 2 = Departed ED 3 = Died in ED 4 = Dead on arrival
smoker_ED		Number		1 = Yes, smoker 0 = No

risky_alcohol_ED		Number		1 = Yes, drinker 0 = No
obesity_ED		Number		1 = Yes, obese 0 = No
Agegroup_ED		Number		1 = Under 60 years old 2 = 60 years old and older

PBS data dictionary

Variable	Description	Variable type	Format name	Allowable entries
ID	Unique person ID	Number		
supply_date	Date of medication dispensed	Number	DDMMYY10.	
ATC	Anatomical Therapeutic Chemical (ATC)	Character		Code as per ATC Classification code allocated by the WHO Collaborating Centre for Drug Statistics Methodology https://www.whocc.no/atc_ddd_index/
drug_name	Generic name of medicine	Character		As per PBS medicine listing by drug: https://www.pbs.gov.au/browse/medicine-listing
item_code	PBS item code	Character		As per PBS medicine listing by drug: https://www.pbs.gov.au/browse/medicine-listing
form_strength	Form and strength of medicine	Character		As per PBS medicine listing by drug: https://www.pbs.gov.au/browse/medicine-listing
Smoking_Cessation_Therapy		Number		0 = Not take 1 = Take
NRT_patches		Number		0 = Not take 1 = Take
Bupropion		Number		0 = Not take 1 = Take
Varenicline		Number		0 = Not take 1 = Take

Reference:
1. Statistics ABo. Patient Experiences. 2022. Patient Experiences, 2021-22 financial year | Australian Bureau of Statistics (abs.gov.au)