Assignment 2A: Assessing Data Quality and Producing Reproducible Notes

Group 2

PART 1: Data cleaning, documentation and data dictionary update

A. Cleaning of Medical Plus General Practice data

1. Data exploration

Prior to analysis, data exploration and cleaning were conducted to familiarise ourselves with the content and quality of the data. The data received from the data custodian was first saved. There are no additional publicly available data dictionaries associated with this data.

SAS studio for academics was utilised for data exploration and cleaning purposes. The library "/home/u63486273/HDAT9400/Assignment2" was assigned and the supplied SAS format file applied. Throughout the process, actions were documented as seen below and the code was annotated for reproducibility. A flow chart was simultaneously created to reflect the approach (Figure 1).

The data dictionary and data received were checked for format, completeness and the total number of records received was noted. The contents of the dataset were examined using PROC CONTENTS. Then, summary statistics were run using PROC MEANS which confirmed the expected data length was present. All observations were within the study period (GP_last). Frequency, mean, minimum, maximum values, range, median and standard deviation analysis was run on all variables to identify illogical results or inconsistencies in trends.

2. Data cleaning

Data was cleaned and standardised to match formats in the data dictionary. For example, for sex, responses included 1, 2, M, F. These were standardised to 1 and 2 in accordance with the data dictionary. The same was done for all other categorical variables.

For variables where there were invalid or missing results, these were changed to "." to enable analysis. Common invalid responses included "99", "999", "998" which are outside expected ranges per the data dictionary. It is difficult to infer from the context what these may represent so they are treated as missing.

Data outside of data dictionary ranges were also cleaned from the results, for example, values of age of stopping smoking less than 10 or more than 105 years of age were excluded. Noting year values (YYYY e.g., 2014) in response to this question, the age of the participant at the time of data collection (2021) was calculated by age-(2021- age_stop). Additionally, the data dictionary was updated to exclude the description "or when did you stop smoking" due to potential confusion about the requested information format. Year values were converted to patient age of ceasing smoking. The reason variable was also mapped in line with the conventions for the other categorical variables.

After removing invalid responses, cleaned variables were cross-tabulated with the respective original variable. Analytics including norow, nocol, nopercent missing were conducted to confirm the cleaning did not lose any responses. Old variables were dropped and the cleaned variables renamed back to the original for ease of use.

Data was then checked for logic. This included checking for any cases where patients reported their smoking status as currently smoking (smoke_now=1) but had a smoking stop date (age_stop >0). 12 patients were found to fall within this group. Given it is not possible to establish which data point is

correct, the decision was made to treat both of these variables as missing for these patients. Next, it was checked if there were any illogical cases where patients had a start or stop age but responded that they had never smoked. Any patients who responded they were still smoking (smoke_now=1) but also responded no to ever smoking were also investigated. No such responses were found.

Further checks were conducted to examine low height values to determine if they belonged to children. By running a frequency check for height <1.40m and then tabulating the corresponding ages, it was found that the ages of people with height below 1.4m were between 1 and 13 years. This is reasonable. Similarly, weight was examined to determine if recorded low weights were logical. It was found that people below 30 kg in weight were between 1 and 13 years of age, also reasonable.

The presence of duplicates was then checked. For exact duplicates (same data for all variables), with PROC SORT, one out of the set of duplicates was removed (n=11). The data was then checked for duplicates of patient ID only. 31 duplicated IDs were found. Examples were then reviewed to guide the treatment of partial duplicates. On reviewing each partial duplicate, there seemed to be one duplicate with more complete information than the other in general. However, there were also contradictions to this pattern: ID 2897. As no clear patterns to explain the duplication was found, both duplicates were dropped for all partial ID duplicates (n=31).

Data was further checked for temporal relationship between month of collection and completeness. Missing values for variables were evenly distributed so this was deemed unlikely to be the cause.

Observing occurrence for missing values across all variables, smoking start and stop age were two variables with a larger proportion of missing data (>90% missingness). There remained a high proportion of missingness when considering just the proportion of missing data for patients who had ever smoked. Despite this, it was decided to retain this data given it could potentially provide useful signals for future analysis.

3. Review and summary statistics

Summary statistics including range, median, and missingness were then reviewed post-cleaning. Descriptive statistics and summary information for the final cleaned dataset were generated and exported to a CSV file. There were 5764 remaining unique IDs. No variables were completely removed from the dataset in the cleaning process.

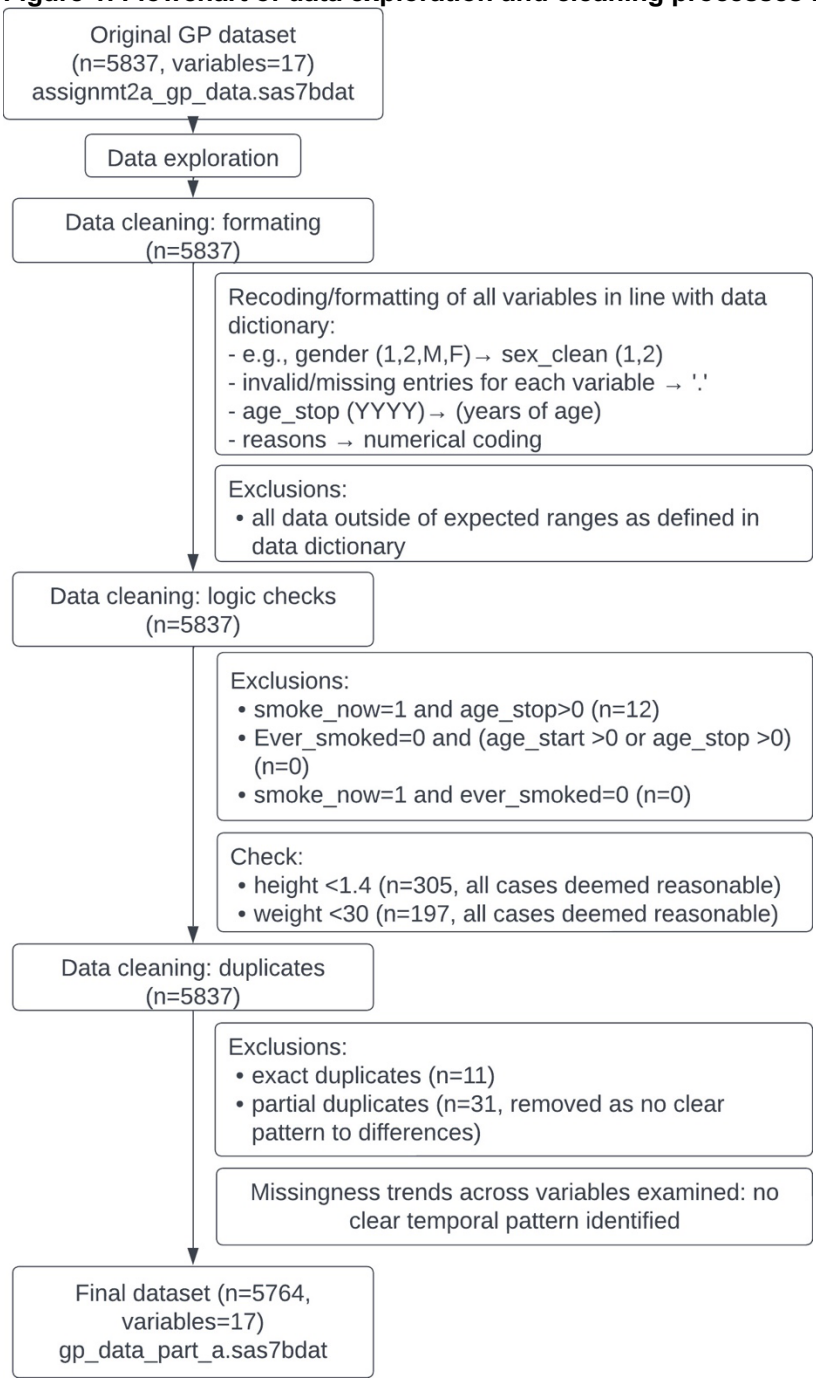
4. Independent review

Finally, data exploration and cleaning were independently reviewed by a second group member. An updated data library was created.

5. Saving cleaned data

The cleaned data was saved as a new dataset (gp_data_part_a.sas7bdat).

Figure 1. Flowchart of data exploration and cleaning processes for GP data



B. Cleaning of Emergency Department (ED) data

1. Data Exploration

As with the GP dataset, a library was assigned to save a dataset from the session. Then the format file and ED dataset was loaded from the path "/home/u63486273/HDAT9400/ Assignment2". The data was examined for format, completeness, and the total number of records using PROC CONTENTS and PROC MEANS. Summary statistics were calculated to confirm the expected data length, missingness and assess the distribution of numeric variables. There were 15 variables and 63614 observations.

2. Data Cleaning

Any data from visits outside the dates stipulated in the provided data dictionary were removed (n=27101 removed). The remaining observations were then formatted and recoded in accordance with the data dictionary. Similar to the procedure utilised for cleaning the GP dataset, it was decided to code invalid or missing results as "." to enable analysis. A new data file “ed_data_cleaned” was created. Original variables were dropped, and the new cleaned variables renamed for ease. PROC CONTENTS was run post-recoding to ensure no data was lost (n=36513, 15 variables).

The presence of duplicates after standardisation of formatting and recoding was then addressed. While it is possible a patient represents to ED on the same day for the same reason, this is considered unlikely thus only one copy of the duplicated row was kept in the dataset (n= 577 removed, n=35936 remaining)

For partial duplicates, it was considered possible for a patient to represent to ED multiple times throughout the study period, including multiple times on the same day. Therefore, handling of partial duplicates was more complex than for the GP dataset which looked at the most recent presentation date (GP_last) only. Where there was discrepancy in variables that should not change between visits such as country of birth and sex, all entries for the given patient ID were excluded as it is not possible to distinguish which is the correct result. Age should not have more than 2 distinct values during the study period (allowing for 2 possible ages given patients may have an interval birthday between presentations). It is allowable to have duplication of reason for presentation (Dx) etc. Where there were more than 2 ages reported, these rows were also excluded. Once these rows were removed, there was n=32391 remaining.

3. Review and Summary Statistics

Summary statistics were reviewed on the cleaned data, "ed_data_cleaned," to ensure that the cleaning process did not introduce any unexpected changes.

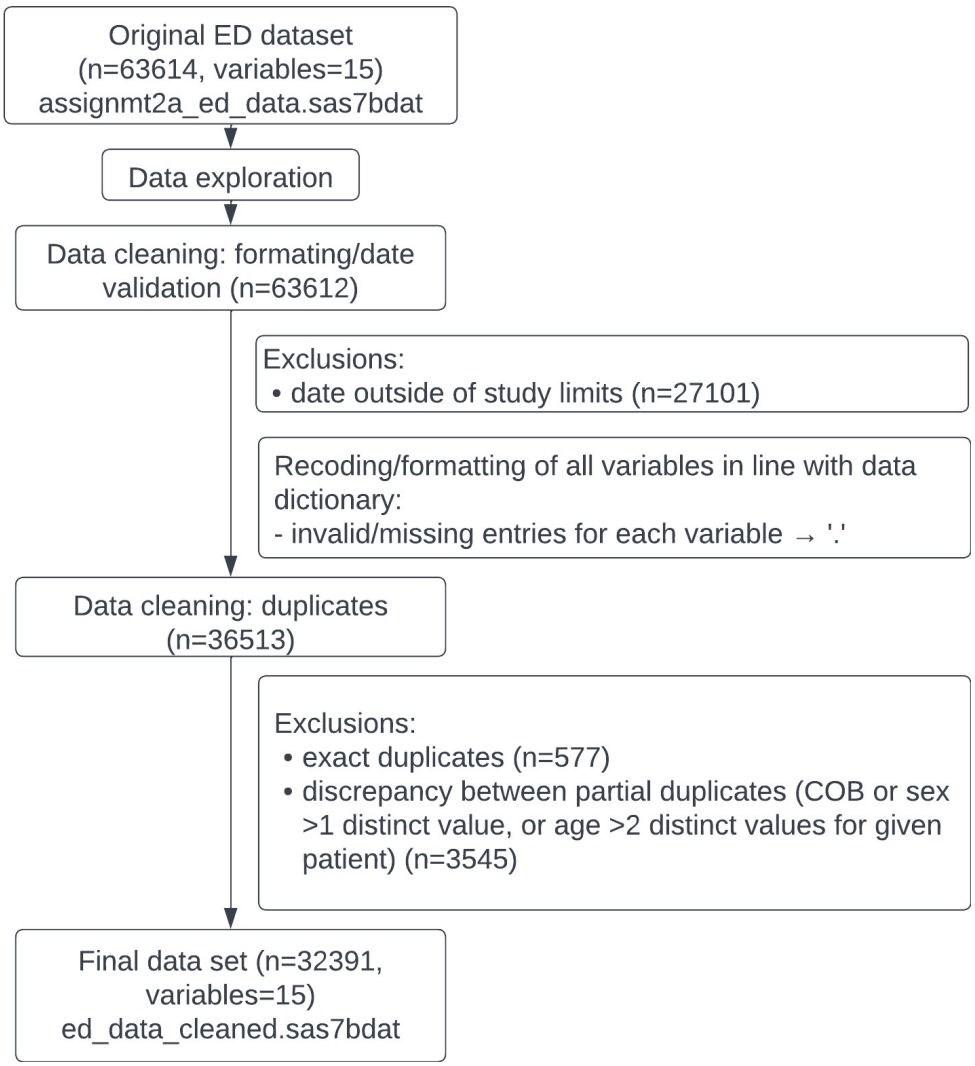
4. Independent Review

Data exploration and cleaning were independently reviewed by another team member to validate the process and ensure accuracy of results.

5. Saving Cleaned Data

The cleaned data, named "ed_data_cleaned," was saved to a new dataset, "Ass2.ed_data_cleaned," in the specified dataset path ("/home/u63486273/HDAT9400/Assignmt_2").

Figure 2. Flowchart of data exploration and cleaning of ED data



C. New GP data variables

New variables were created in line with the provided definitions (SAS file appended). As per the instructions, only participants over the sage of 18 were considered for this component. Variables were formatted as per the updated data dictionary.

- Smoke status

This variable was created to provide further insight into a patient's current and past smoking status. The data for smoking was handled as explained in Part 1A above with illogical results removed.

If ever_smoked=0 then smoke_status could be assumed to 0 (never smoked) given the previous checks which ensured that there was no conflicting age_start or age_stop for these patients.

If ever_smoked=1 then smoke_status was taken as 1. It is not able to be inferred from having an age_start but not a age_stop that these participants are still smoking as the data could be missing given the large proportion of missingness for these variables. As such, these cases were not included as 1 but rather treated as missing.

Finally, if age_stop >0 or ever_smoked=1 and smoke_now=0 then smoke_status was taken as 2.

D. Updated data dictionaries

The data dictionary provides a comprehensive overview of the variables and their characteristics in two datasets: assignmt2a_gp_data and assignmt2a_ed_data. These datasets contain information about the people and services in a fictitious neighbourhood called Sunnydale, which has a high proportion of culturally and linguistically diverse populations.

The assignmt2a_gp_data dataset represents the client information recorded in the most recent general practice (GP) visits in 2014. It includes variables such as unique person ID, date of the most recent GP visit, age, gender, country of birth, healthcare card status, smoking history, alcohol consumption, blood pressure readings, and reason for the GP visit. The data dictionary was modified to include the new variables created in Part C. Other modifications included the phrasing of the description for age_stop for clarity for the expected data format.

Table 1. Updated GP data dictionary. Data dictionary updated after data exploration and cleaning including new Variables smoke_status_GP, risky_alcohol_GP, BMI_GP, obese_GP and highBP_GP.

Variable	Description	Variable type	Format name	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 in 2014	Number		
sex	Gender of the patient	Character		1=male 2=female .=missing/invalid
cob	In what country were you born?	Number	cobf.	1= Born in Australia 2= Born overseas .=missing/invalid
healthcare_card	Do you have a healthcare card[1]?	Number	ynf.	1= Yes 0= No .=missing/invalid
ever_smoked	Have you ever been a regular smoker?	Number	ynf.	1= Yes 0= No .=missing/invalid
smoke_now	Are you a regular smoker now?	Number	ynf.	1= Yes 0= No .=missing/invalid
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
smoke_status_GP	Patient smoking status	Number		0=Never smoked 1=Current smoker 2=Ex-smoker .=missing/invalid
drinks_day	About how many alcoholic drinks do you drink per day?	Number		Invalid if >20
risky_alcohol_GP	Classification of patient alcohol	Number		0=No (≤2 drinks per day) 1=Yes (>2 drinks per day) .=missing/invalid

height	How tall are you without shoes? (metres)	Number		Invalid if <0.55m or >2.40m
weight	About how much do you weigh? (kilograms)	Number		Invalid if <5.0kg or >270kg
BMI_GP	Patient BMI (Weight/Height^2)	Number		Invalid if <10 or >60
obese_GP	Patient BMI meets criteria for obesity (BMI ≥30)	Number		0=No (BMI <30) 1=Yes (BMI ≥30) .=missing/invalid
adverse_reaction	Have you had any adverse reaction to any medication?	Number	ynf.	1= Yes 0= No
syst_bp	Systolic blood pressure (mmHg)	Number		
diast_bp	Diastolic blood pressure (mmHg)	Number		
highBP_GP	Patient hypertensive status	Number		0=Normal blood pressure (systolic<135mmHg & diastolic<85mmHg) 1=High blood pressure (systolic≥135mmHg or diastolic≥85mmHg) .=missing/invalid
reason	Reason for the most recent GP visit	Number		1= headache 2= nausea 3= tinnitus 4 = vomiting 5= itching 6= abdominal pain 7= dizziness 8= skin rash 9= palpitations 10= hallucinations

The assignmt2a_ed_data dataset contains information about emergency department (ED) attendances by Sunnydale residents. It includes variables such as unique person ID, dates of ED presentation and separation, age at ED presentation, gender, country of birth, interpreter requirement, private health insurance status, triage category, principal presenting diagnosis, additional diagnoses, and separation mode. This dataset consists of 63614 rows and 15 columns.

Table 2. Updated ED data dictionary

Variable	Description	Variable type	Format name	Allowable entries
ID	Unique person ID	Number		
ed_admission	Date of ED presentation	Date	DDMMYY10.	Dates in the range 01/01/2014 – 31/12/2014
ed_separation	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 .= missing/invalid
cob_ed	In what country were you born?	Number	cobf.	1= Born in Australia 2= Born overseas .= missing/invalid
interpreter	An interpreter is needed?	Number	ynf.	1= Yes 0= No .= missing/invalid
health_insurance	Do you have private health insurance?	Number	ynf.	1= Yes 0= No .= missing/invalid
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 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 .= missing/invalid

PART 2: Research Question

A. Results of analysis

Table 3. Summary of demographic, health and lifestyle data

Variable	Mean	Minimum	Maximum	Range	Median	Standard deviation
Age (years) (n=5764)	43.7	1.0	98.0	97.0	45.0	17.3
Height (m) (n=5676)	1.68	0.62	1.89	1.27	1.73	0.2
Weight (kg) (n=5663)	74.6	7.0	140.5	133.5	72.4	22.9
BMI (kg/m^2) (n=5187)	26.6	14.2	66.3	52.0	24.9	7.6
Blood pressure						
Systolic blood pressure (mmHg) (n=5690)	118.5	81.0	181.0	100.0	119.0	15.8
Diastolic blood pressure (mmHg) (n=5690)	76.1	51.0	132.0	81.0	76.0	13.5
Smoking						
Age start (years) (n=494)	27.7	13.0	58.0	45.0	26.0	8.6
Age stop (years) (n=556)	40.7	21.0	79.0	58.0	39.0	7.8
Alcohol						
Drinks per day (standards) (n=5455)	0.9	0	18.0	18.0	0	1.4

Table 4. Demographic information by age

		Sex		**Country of birth**		**Healthcare card**	
		Male	*Female*	*Australia*	*Overseas*	*Yes*	*No*
Age	*≤20*	277 (47.9%)	301 (52.1%)	184 (32.0%)	391 (68%)	392 (74.0%)	138 (26.0%)
	21-40	755 (46.9%)	856 (53.1%)	413 (25.9%)	1182 (74.1%)	1117 (69.3%)	494 (30.7%)
	41-60	1295 (44.8%)	1599 (55.3%)	1331 (46.5%)	1534 (53.5%)	2128 (73.4%)	772 (26.6%)
	61-80	196 (38.9%)	308 (61.15%)	346 (69.2%)	154 (30.8%)	383 (76.0%)	121 (24.0%)
	>80	62 (37.1%)	105 (62.9%)	116 (70.3%)	49 (29.7%)	135 (80.8%)	32 (19.2%)
Total		2585 (44.9%)	3169 (55.1%)	2390 (41.9%)	3310 (58.1%)	4155 (72.7%)	1557 (27.3%)

Table 5. Analysis of relationship between variables. Obesity status, smoking status, risky alcohol status and hypertensive status include only individuals over the age of 18 at the time of their last GP visit.

		Obesity status		**Smoking status**			**Risky alcohol status**		**Hypertensive status**	
		No	*Yes*	*Current*	*Ex*	*Never*	*No*	*Yes*	*High*	*Normal*
Age	*≤20*	66 (78.6%)	18 (21.4%)	14 (20.0%)	6 (8.6%)	50 (71.4%)	72 (90.2%)	8 (9.8%)	41 (41.4%)	58 (58.6%)
	21-40	1154 (73.0%)	428 (27.1%)	259 (16.6%)	253 (16.2%)	1052 (67.3%)	1358 (89.2%)	164 (10.78%)	435 (26.9%)	1180 (73.1%)
	41-60	2082 (72.9%)	775 (27.1%)	406 (14.0%)	457 (15.8%)	2029 (70.2%)	2393 (86.9%)	361 (13.11%)	847 (29.2%)	2053 (70.8%)
	61-80	353 (70.7%)	146 (29.3%)	60 (11.9%)	77 (15.3%)	366 (72.8%)	410 (86.6%)	64 (13.5%)	126 (25.0%)	378 (75.0%)
	>80	129 (78.2%)	36 (21.8%)	23 (13.8%)	22 (13.2%)	122 (73.1%)	131 (85.6%)	22 (14.4%)	31 (18.6%)	136 (81.4%)
Sex	*Male*	1431 (61.8%)	885 (38.2%)	427 (18.5%)	426 (18.5%)	1456 (62.1%)	1956 (87.6%)	276 (12.37%)	952 (40.3%)	1408 (59.7%)
	Female	2346 (82.0%)	516 (18.0%)	318 (11.1%)	400 (13.9%)	2158 (75.0%)	2402 (87.6%)	341 (12.4%)	525 (18.0%)	2390 (82.0%)
Country of birth (COB)	*Australia*	1617 (74.1%)	564 (25.9%)	207 (9.5%)	315 (14.4%)	1669 (76.2%)	1752 (82.2%)	380 (17.8%)	570 (26.0%)	1647 (74.0%)
	Overseas	2118 (71.9%)	826 (28.1%)	536 (18.2%)	504 (17.1%)	1903 (64.7%)	2613 (91.6%)	239 (8.4%)	885 (29.5%)	2111 (70.5%)
Smoking status	*Current*	516 (70.2%)	219 (29.8%)	-	-	-	624 (88.5%)	81 (11.5%)	229 (30.6%)	520 (69.4%)
	Ex	574 (70.7%)	238 (29.3%)	-	-	-	692 (88.0%)	94 (11.96%)	242 (29.3%)	585 (70.7%)
	Never	2633 (74.0%)	926 (26.0%)	-	-	-	2983 (87.3%)	434 (12.7%)	947 (26.2%)	2672 (73.8%)
Obesity status	*No*	-	-	516 (13.9%)	574 (15.4%)	2633 (70.7%)	3142 (88.0%)	428 (12.0%)	1006 (26.6%)	2778 (73.4%)
	Yes	-	-	219 (15.8%)	238 (17.2%)	926 (67.0%)	1149 (86.5%)	180 (13.5%)	433 (30.9%)	970 (69.1%)
Total		3784 (73.0%)	1403 (27.1%)	3619 (69.7%)	762 (14.7%)	815 (15.7%)	4366 (87.6%)	619 (12.4%)	3805 (72.0%)	1480 (28.0%)

B. Interpretation of results

The analysis of the GP dataset provides valuable insights into the characteristics of the sampled patients from the Medical Plus GP practice, including their health status, demographic information and lifestyle characteristics. There are 5764 unique IDs in the GP dataset.

1. Socio-demographic characteristics

Table 3 and 4 present the socio-demographic characteristics of the clients. The majority of the clients in the dataset are adults, with a median age of 45. Ages range from 1-98, capturing a wide variety of life stages. 50.3% of patients fall within the 41-60 age range, with the fewest amount of people falling into the >80 age group. The gender distribution shows a slightly high proportion of females to males in the dataset with 44.9% males and 55.1% females. Particularly in the older age groups, there was a slightly higher proportion of females than males (Table 4).

Overall, the number of patients born overseas was greater than those born in Australia as shown in Table 4. Older patients were more likely to be born in Australia than their younger counterparts. 72.7% of patients of the clinic held a healthcare card. These statistics demonstrate that the population of Sunnydale who attend the Medical Plus GP practice is diverse, involving a range of ages, and cultural backgrounds, with a subsequent diversity of potential healthcare needs.

2. Lifestyle factors

30% of study participants have been or are currently regular smokers; this encompasses 14% who are current smokers and 16% who reported being ex-smokers. Males were more likely to have a smoking history (be they current or ex-smokers) than women. The age at which clients started smoking ranges from 13-58 with an average age of 28, and the age at which they stopped smoking ranges from 21-79 (Table 3). This suggests that starting to smoke regularly may occur as early as adolescence, and there is a wide age range for smoking cessation. There was no clear correlation between smoking status and demographic e.g., age, gender, or other lifestyle variables, e.g., obesity status, and alcohol intake thus there may be other factors, not explored in this study, that influence smoking behaviours (Table 5). Further research or analysis could explore other variables or factors that may contribute to differences in smoking status among patients in Sunnydale. There was significant missing data for variables around smoking initiation and cessation. GPs at the practice may benefit from taking thorough smoking histories from patients, including these details, to further develop their understanding of patients' risk factors and psychosocial histories. Optimising health outcomes for the current smoking group may represent a potential target for future health interventions. Such interventions may benefit from being tailored towards population groups with high rates of smoking, e.g., gender-specific support groups, targeting younger rather than older people.

The average number of drinks per day in the patients studied was 0.9, with a minimum value of 0 and maximum of 18 drinks per day. Risky drinking rates were similar across genders (12.4% in males, 12.4% females) and increased slightly with age (up to 14.4% in >80s) (Table 5). Notably, there is less data available in the >80 years age group decreasing the external validity of these results. Risky drinkers are at increased risk of developing alcohol dependence, and associated health issues. This is another subset of patients who may warrant further investigation and intervention.

3. Health status

Considering reasons for presenting to the GP practice, the most commonly reported was headaches (representing almost 50% of all reported reasons). Information is not available on the severity of presentations, management or follow-ups i.e., whether this is a first presentation for the reason.

The Body Mass Index (BMI) of the clients ranges from 14.2-66.3, with 27.1% meeting the criteria for obesity (Table 3). The median BMI was 26.6 which falls within the overweight range. There was a

significant difference between the obesity rates of males and females studied, with 38.2% of males meeting the criteria for obesity compared to 18.0% of women (Table 5). There was no clear trend in BMI with age. Further research into gender-specific factors that may be impacting body weight may be valuable in the future to improve health outcomes in Sunnydale. Furthermore, A slightly higher proportion of people within the obese BMI category had hypertension (30.9%) compared to those who are not obese (26.26%).

In general, the systolic and diastolic blood pressure measurements indicate that 28.0% of patients in the dataset have high blood pressure (Table 5). Amongst these patients, hypertension rates were higher amongst men than women (40.3% and 18.0% respectively). Rates were similar among patients born in Australia versus overseas (26.0% and 29.5% respectively). There was no clear relationship between smoking status and hypertension. Hypertension rates appeared higher in the <18 years age group, however, there is notably a smaller sample size for this age range which may impact accuracy of results. Furthermore, whether patients are on antihypertensive therapies is not included in the available data which may skew results. Understanding the correlation between hypertensive status and other variables studied can help medical practitioners screen for and manage hypertension.

These findings suggest the need for healthcare coordination within the Medical Plus GP practice to address the diverse needs of their patients. Strategies can be implemented to promote smoking cessation programs, alcohol awareness and moderation, and interventions to manage obesity and high blood pressure. Additionally, further analysis and follow-up assessments can be conducted to gain a deeper understanding of patients; health status and identify potential areas for improvement in their overall well-being.

Overall, the analysis of the GP dataset provides valuable insights into the characteristics of the clients at Medical Plus GP practice, enabling informed decision-making and targeted interventions to enhance healthcare coordination and improve patient outcomes. The current analysis also reveals avenues for further research to better understand these interactions.

C. SAS code

The relevant SAS code is appended in the attached SAS file.

Part 3: Data Linkage

A. Data linkage strategies

After the preparation of the dataset by each data custodian for linkage by checking data quality, standardising variables and anonymising sensitive information, as well as establishing a formal agreement between each entity, data would be linked. The CHeReL data linkage unit will be employed as they hold jurisdiction over New South Wales.

CHeReL's services are supported by their Master Linkage Key (MLK); a system of continuously updated links within and between core health-related datasets in NSW and ACT[1]. Currently, the MLK consists of 32 datasets containing pointers to over 328 million records relating to over 18 million individuals. Requests for linked data of datasets contained within the MLK can be completed relatively quickly. Linkage of datasets not contained within the MLK can be completed via the use of deterministic or probabilistic linkage strategies.

The datasets requested for linkage for this project include the Medical Plus GP data, Registry of Births, Deaths and Marriages (RBDM) deaths, and PBS data. RBDM death data is contained within the CHeReL MLK, while the Medical Plus GP data and PBS data require additional strategies for linkage. The available identifiers for each dataset are listed in Table 6..

Table 6. Datasets and shared variables. Blue: data available from RBDM and Medical Plus GP. Red: data available from Medical Plus GP and PBS.

Dataset:	**RBDM deaths**	**Medical Plus GP**	**PBS**
Identifiers available:	• Names • Address • Date of birth	• Names • Address • Date of birth • Medicare number	• Medicare number

Deterministic (exact) matching can be performed by linking unique identifiers such as a Medicare number, or via a proxy linkage key such as SLK-581. The SLK-581 uses variables such as first name, last name, date of birth, and sex to create a client identifier for linkage purposes. Probabilistic linking is used where deterministic linking is not possible or is of insufficient quality. Probabilistic linking involves a process in which common identifiers between datasets are linked and evaluated. Probabilities (m probability and u probability) are assigned for each field and used to determine 'links', 'possible links', and 'non-links'. Possible links go through a manual clerical review process to determine which links are matches and should stay together.

Based on the available identifiers for each dataset listed in Table 6, the CHeReL strategy for linking datasets requested by the Medical Plus GP would be:

1. **Link the Medical Plus GP data and PBS data deterministically based on Medicare number.**

   The data custodian for the PBS data only has access to the Medicare number for identification purposes. The data custodian for the Medical Plus GP data also has access to Medicare number, allowing for near deterministic matching between the two datasets. It is important to note that although Medicare numbers can be used as a reasonable unique identifier, there are potential issues that may arise via this method. These issues stem from the fact that multiple individuals can be listed on the same card, and data collection may fail to include or incorrectly record the Individual Reference Number. Ideally, it would be best to perform probabilistic linking using Medicare number and other identifiers. However, Medicare number is the only common identifying variable available which can be used for linkage in this scenario.

2. Link the Medical Plus GP data and RBDM deaths data probabilistically via the shared identifiers (names, address and date of birth).

Both the data custodian for RBDM deaths data and the data custodian for Medical Plus GP data have names, address, and date of birth variables available for identification. While these variables closely align with those required to form the SLK-581 (consisting of names, date of birth, and sex), the custodians lack access to the sex identifier needed to perform deterministic linkage. Instead, probabilistic linkage will be conducted by matching shared variables and assigning probabilities to establish links, possible links, and non-links. To further determine links and non-links, a clerical review will take place to sort through possible matches. This step is crucial due to the possibility of address changes over time and the potential for errors in data collection affecting other variables.

References

1. Centre for Health Record Linkage. CHeReL Master Linkage Key. Datasets. 2023. Accessed July 12, 2023. https://www.cherel.org.au/master-linkage-key

B. Graphical representation of information interchange

The data exchange process involves a series of steps and different stakeholders to ensure privacy, data security, and compliance with legal and ethical requirements. These steps are outlined below and visualised in Figure 3 below:

1. Analyst prepares a research question and determines which data are needed

The analyst will determine which data will be needed to address their research question. In this scenario, the Medical Plus GP manager would like to examine medication compliance among their patients. They have decided that the Medical Plus GP data, PBS data, and RBDM deaths data is needed to adequately address the research question.

2. Apply for data and approvals

The data analyst will initiate a request to obtain access to the relevant datasets from each respective data custodian. To proceed, approval must be obtained from each data custodian, as well as CHeReL (the appropriate data linkage unit), and the Human Research Ethics Committee.

3. Data Custodians split the data and send identifiers to CHeReL

To ensure privacy, each data custodian will split the data into two separate files. One file will contain identifying information such as name, address, date of birth, and Medicare number etc., while the other file will contain de-identified content data. These files are stored and managed separately to ensure the security of the data. The file containing identifying information is sent to CHeReL by each data custodian for the purpose of data linkage.

4. Data Linkage

CHeReL receives the identifying information from each dataset and uses a combination of deterministic and probabilistic matching methods to establish links for each individual. Clerical review is conducted to sort through possible links. Once matched, identifier data for each individual is assigned a Project-specific Person Number (PPN). The identifying information, now associated with a PPN, is returned to each respective data custodian.

5. Data Integration

Each data custodian receives the identifying information and PPN for their dataset. The identifying variables are removed, and the PPNs are attached to the de-identified content data. The de-identified

content data with PPNs are sent by each data custodian back to the analyst via a secure transfer system.

6. Analyst creates a research dataset

The analyst is able to use the PPN to combine records for each individual without requiring access to any identifying information. This process is used to maintain privacy.

Figure 3. Proposed data matching strategy