ASSIGNMENT 2B: CREATING ANALYSIS-READY DATA FOR REPRODUCIBLE RESEARCH

Individual assignment (each person submits their own work)

Due: **Monday 7th August 2023 by 9am AEDT**

Weight: 40% of the total grade

Submission: Open Learning

Submission

This is an **individual** assignment. Please submit the following two documents:

1) Assignment answers (word/pdf document) with clearly labeled parts/questions
2) SAS code (saved as .sas) or a .zip if you have more than one SAS code

The assignment consists of three parts, please address all the questions:

1. Research question 1
2. Research question 2
3. Research question 3

Learning outcomes assessed

LO3: Evaluate data quality
LO5: Develop and implement data merging and cleaning rules
LO6: Generate syntax (code) required to produce analysis-ready datasets

Penalty for late submission

A **penalty** will apply for **late submissions** of assessment tasks (5% per day) if special consideration has not been granted. Assessments will not be marked if submitted more than 5 days after the assessment due date (in line with UNSW policy), and will receive a value of 0. For example, if you submit your assessment 2 days late, then 10% (5% x 2 days) will be deducted from the assessment mark. Thus, if your assessment was marked as 75% but was submitted 2 days late, then your final mark will be 65%.

In case of illness or misadventure, you may apply for an extension, only if requested *before* the assignment's due date. Special consideration requests are handled by central student administration and should be submitted via https://student.unsw.edu.au/special-consideration.

Assignment instructions

Data for the assignment

For assignment 2B, you will be using three simulated (made-up) datasets – GP, ED and PBS data – all of which contain information about people and services in a fictitious neighbourhood called Sunnydale, which has a high proportion of culturally and linguistically diverse population. The context of the GP and ED datasets for this assignment is the same as for assignment 2A, however, there are modifications to variables and data values in the GP and ED datasets.

Medical Plus general practice data (GP)
The Medical Plus GP serves 60% of Sunnydale's population and has multilingual staff. The GP data contains information about the clientele that visited the practice. It was developed by the Practice Manager, with one record for one client, based on the client's information recorded in the most recent GP visit in 2014. The information about variables in the GP data is outlined in the GP data dictionary (page 7).

Emergency Department data (ED)
In Sunnydale neighbourhood, there is a single Emergency Department (ED). The ED data was extracted from the database of this Emergency Department. It contains information about each ED attendance by Sunnydale residents. The information about variables in the ED data is outlined in the ED data dictionary (page 8).

Pharmaceutical Benefit Schemes data (PBS)
This data contains information about claims for PBS-subsidised medicines that were dispensed at local pharmacies to Sunnydale residents. Each record in the PBS data represents a medicine dispensed at a time. The data custodian agreed to provide PBS data of individuals who had a dispensing of medications for smoking cessation in 2014. More information can be found in the PBS data dictionary (page 9).

The three datasets and a format program are saved in zipped file on Open Learning as below:

GP data:	assignmt2b_gp_data.sas7bdat
ED data:	assignmt2b_ed_data.sas7bdat
PBS data:	assignmt2b_pbs_data.sas7bdat
SAS format program:	assignmt2b_formats.sas.

Examine data dictionaries, SAS datasets and format program to familiarise yourself with the data.

For this assignment, data cleaning is not required, but it is a good practice to do exploratory data analysis (EDA) before analyses.

Provide written answers and interpretation to each assignment question in a Word/PDF document, fully supported by SAS code and updated data dictionaries, as required

Where applicable, summarise your results in a table format and provide written interpretation of the findings. Table(s) should be presented in an academic format similar to what would be found in the results section of a published journal article. You can present more than one table.

Think about principles in creating the analysis ready dataset(s) and the merges that will be required to combine the information from multiple data sources.

Assignment questions

In your assignment you'll analyse data to address three research areas from these perspectives:

- **Primary care perspective:** Primary care manages patients in the community and aims to keep patients out of hospitals. Through linkage to ED data, the Medical Plus GP quantifies their patients attending the ED in 2014 and characteristics of these patients.
- **ED care perspective:** Prior research has reported that people might be less likely to reveal their lifestyle behaviours in unfamiliar settings such as hospitals. The Sunnydale ED links ED data to GP data to investigate how well lifestyle and health behaviours are documented in the ED data and whether data quality varies according to patient populations. This would inform a targeted intervention for patient assessment and screening for risky behaviours.
- **Tobacco control perspective:** The Sunnydale Population Health Department is developing a clinical trial for smoking cessation which includes the use of medicines in addition to behavioural therapies. To inform the design of the trial, the Department assesses the baseline uptake of medicines for smoking cessation using PBS data linked to GP and ED data.

Population of interest are adults (i.e. 18 years old and older).

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 **(1 mark):**

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

1.2. Calculate and report the proportion of GP patients who attended the ED in 2014. Comment on the findings (**4 marks**).

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 (**4 marks**).

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 **(6 marks**).

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 **(6 marks).**

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 (**6 marks**).

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 (**3 marks**).
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

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 **(6 marks)**.
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

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 **(16 marks)**.

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 (**6 marks**).

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 **(10 marks)**.

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 **(5 marks)**?

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 **(6 marks)**.

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 **(4 marks)**.

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 medicine in 2014. Comment on your findings **(7 marks)**.

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

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

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

Grading rubric

The grading rubric is outlined below.

Criteria	HD 85 – 100	D 75 – 84	Credit 65 – 74	Pass 50 – 64	Fail Less than 50
Question answers	Correct answers to all questions, thorough interpretation, recommendations well-supported by findings	Correct answers to all questions, moderate interpretation, reasonably supported recommendations	Mainly correct answers, minor misinterpretation, sound recommendations	Incorrect answers, misinterpretation in parts, sound recommendations	Mostly incorrect answers, lack of interpretation, No answer provided.
Data assembly (SAS)	Data merged correctly with level of detail exceeding examples in class	Data merged correctly with adequate level of detail	Data merged with mostly correct detail	Data merged with some incorrect detail	Data not merged
Documentation (SAS)	Fully functioning and reproducible code with extensive and consistent annotation	Fully functioning and reproducible code with moderate but consistent annotation	Functioning code with some annotation	Code with some errors, lacking annotation.	Code with errors in most parts, or no code provided
Presentation and organisation of information	Outstanding use of academic and linguistic conventions, coherent and succinct discussion	Coherent and clear explanation. Minor editing required.	Clear expression. Flow of discussion mostly smooth. Minor editing required.	Inconsistent and ambiguous expression in parts. Significant editing is required	Lack of coherence, difficult to follow. Significant editing is required.

The scoring rubric is outlined below.

Assessment of written questions answers* and SAS code**	Max score	Your mark
Research question 1 (30 marks)		
1.1. Create age group variable	1	
1.2. Proportion of GP patients attended ED in 2014	4	
1.3. Trends of monthly ED admissions	4	
1.4. Differences between patients who attended and did not attend the ED	6	
1.5. Distribution of numbers of ED attendance	6	
1.6. Characteristics of top 25% ED attending patients	6	
Presentation and organisation of information	3	
Research question 2 (50 marks)		
2.1. Create variables to flag ED records	3	
2.2. Create person-level variables and Report prevalence	6	
2.3. Differences between patients who did and did not visit a GP	12	
2.4. Sensitivity (Sn) and Specificity (Sp) of smoking in ED data	6	
2.5. Sn and Sp of smoking by sex, age group, country of birth, private health insurance	10	
2.6 Suggestions for improving ED screening and recording	5	
Presentation and organisation of information	4	
Research question 3 (20 marks)		
3.1. Create a cohort of smokers using GP and ED data, with numbers	6	
3.2. PBS data as an additional source to identify people who smoke	4	
3.3. Proportion of smokers using any, and each smoking cessation medicine	7	
Presentation and organisation of information	3	
TOTAL MARK	**100**	

*: Assessment of written answers includes the use of academic and linguistic conventions, coherent and succinct discussion with supporting tabular and figure information presented.
**: Assessment of SAS code include i) correct use of common SAS commands introduced in the course, ii) effective creation of variables that takes into account variable type and values, and iii) clear annotation that enables reproducibility.