Sunnydale Health Characteristics Research

Data Linkage and Analysis

Murray Keogh

HDAT 9400

Table of contents

1. Research Question 1

1.1. Creating New Variables

To begin the analysis needed for the GP Practice Manager, I first examined both the ED and GP datasets. The GP dataset included 5,730 observations and 10 variables. Each observation corresponded to a unique GP patient visit, with no duplicates among IDs, meaning there are 5,730 unique patients in the GP population. The ED dataset included 30,392 observations and 15 variables, and all the ED visits occurred in 2014. I also checked for full duplicates in the ED dataset and there were none. Both datasets included only fields present in the provided data dictionary.

I created four new variables that would be needed in the analysis. The four variables created are risky_alcohol_GP, BMI_GP, obese_GP, and agecat. All the recoded and created variables were checked after creation to ensure the logic and code was working correctly.

1.2. ED Visits

In order to perform the analysis for GP patients who visited the ED, I had to merge the GP and ED datasets. SAS was used to merge the dataset and create a final dataset for unique patients with a flag for whether the patient visited the ED or not in 2014. Table 1 below shows the proportion of GP patients who visited the ED in 2014. Out of the 5,730 unique GP patients, 1,378 (24.05%) had a visit to the ED in 2014.

GP Patients With Visit to ED		
	Observations	Percentage
Visit to ED	1378	24.05%
Did not visit the ED	4352	75.95%
Total	5730	

Table 1

1.3. Differences in Demographic Variables

Further analysis conducted for the GP Practice Manager aimed to analyse whether there were differences in demographic and health characteristics for GP patients who did and did not visit the ED. Raw numbers, proportions, and chi-square tests were calculated and conducted to assess the differences. Table 2 below highlights the overall findings for this analysis.

Sex, Country of Birth, Current Smoker, and Obese variables all had statistically significant differences between GP patients who did and did not visit the ED. Healthcare Card, Age Group, and Risky Alcohol User did not have statistical differences between the two groups.

Within the GP population, females appear to visit the ED less than males. GP patients born in Australia visit the ED more than patients born overseas. GP patients who are current smokers also visit the ED more than non-smokers. Lastly, GP patients who are obese visit the ED more than

patients who are not obese. These findings can be used by the GP Practice Manager to support programs and support for patients to hopefully decrease negative health outcomes associated with higher ED utilization.

Demographic and Health Characteristics			
Characteristic	Visit to ED?		Chi-Square P-Value
Sex	Yes (N=1378)	No (N=4352)	
Male	692 (50.22%)	1871 (43.15%)	**<.0001**
Female	686 (49.78%)	2474 (56.85%)	
Country of Birth			
Australia	628 (45.57%)	1836 (42.19%)	**0.0269**
Overseas	750 (54.43%)	2516 (57.81%)	
Healthcare Card			
No	1016 (73.73%)	3166 (72.75%)	0.4744
Yes	362 (26.27%)	1186 (27.27%(	
Age Group			
< 60	1173 (85.12%)	3710 (85.25%)	0.9094
>= 60	205 (14.88%)	642 (14.75%)	
Current Smoker			
No	1101 (79.9%)	3721 (85.5%)	
Yes	253 (18.36%)	540 (12.41%)	**<.0001**
Missing/Unknown	24 (1.74%)	91 (2.09%)	
Risky Alcohol User			
No	1150 (83.45%)	3585 (82.38%)	
Yes	158 (11.47%)	523 (12.02%)	0.6234
Missing/Unknown	70 (5.08%)	244 (5.61%)	
Obese			
No	1020 (74.02%)	3358 (77.16%)	
Yes	321 (23.29%)	878 (20.17%)	**0.0446**
Missing/Unknown	37 (2.69%)	116 (2.67%)	

Table 2

1.4. ED Visits Data Analysis

Analysis was also conducted on the GP patients who had visited the ED to assess how many times these patients made a visit to the ED in 2014. The table below highlights key summary statistics for total ED visits by each patient and the histogram visualizes the distribution of this metric.

Out of the 1,378 GP patients who visited the ED in 2014, the mean number of ED visits was 4.72 and the median was 4. The minimum was 1 and the maximum was 18. 50% of the patients had between 2 and 6 ED visits. The standard deviation was 3.12. The histogram shows that the total ED visits has a right-skewed distribution and the majority of patients had 5 or less ED visits. This analysis can help the GP Practice Manager identify high utilizers within the patient population to provide extra support and resources in order to decrease ED utilization.

Number of ED Visits Per Person (2014)								
N	Minimum	25th Percentile	Mean	50th Percentile	75th Percentile	Maximum	St Dev	Variance
1378	1.00	2.00	4.72	4.00	6.00	18.00	3.12	9.73

Table 3

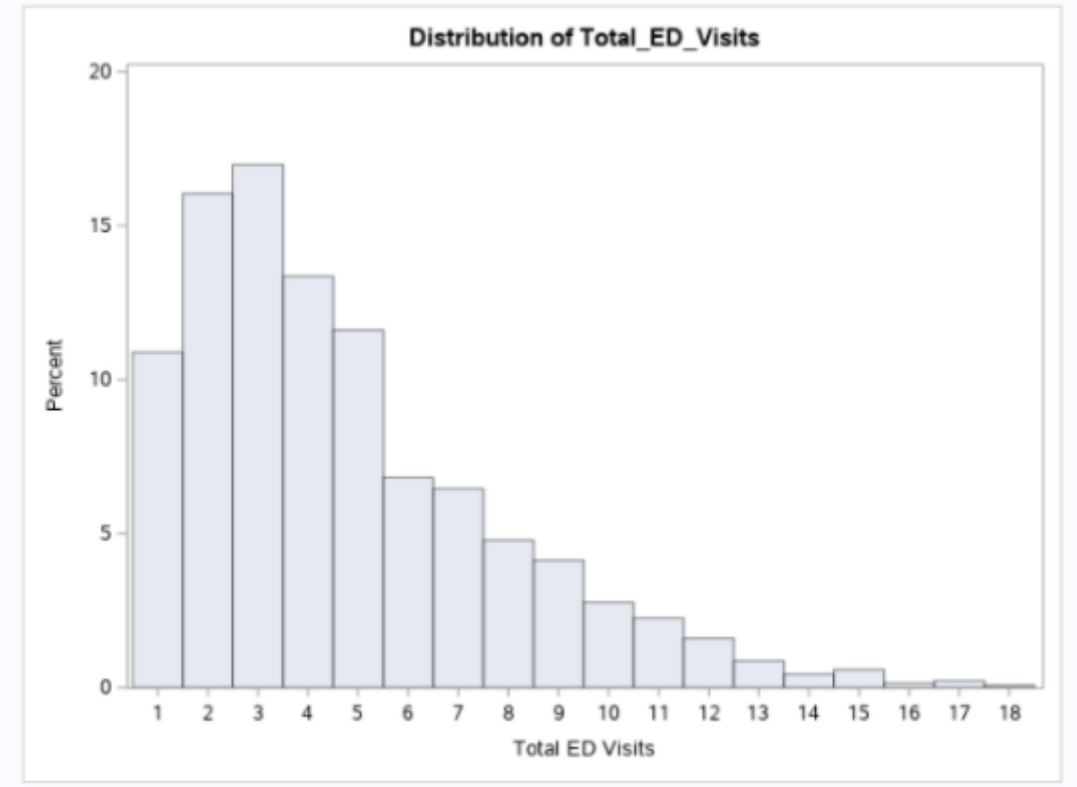


Figure 1

2. Research Question 2

2.1. Creating New Variables

Since both the GP and ED dataset were examined in Research Question 1, I did not need to repeat the preliminary examination steps. Since this analysis is for the ED Manager, I did calculate the unique IDs present within the ED dataset. There were 5,604 unique IDs in the ED dataset, which will be important to remember as I continue my analysis.

I created three new variables that would be needed in the analysis. The three variables created are smoker_flag, risky_alcohol_flag, and obesity_flag. These flags were based on the five possible diagnosis codes that were coded for each ED visit. All the recoded and created variables were checked after creation to ensure the logic and code was working correctly.

2.2. ED Risk Factors

Using the newly created risk factor flags, I am able to create a person level dataset to ascertain the prevalence of these risk factors among the ED patient population.

In order to calculate this, I created three new variables. The three variables created are smoker_ED, risky_alcohol_ED, and obesity_ED. If a patient had any of the risk factor flags for an ED visit, then the three variables above were set to 1 for all records for the patient ID.

Utilizing these variables, I then created a new dataset that only included one unique row for each patient ID. With this person level dataset, I was able to calculate the risk factor prevalence shown below in Table 4. In the ED patient population, 15.67% of patients are smokers, 20.09% of patients are risky alcohol users, and 26.96% are obese.

Risk Factors			
Characteristic		Observations	Percentage
Smoker			
	No	4726	84.33%
	Yes	878	15.67%
Risky Alcohol User			
	No	4478	79.91%
	Yes	1126	20.09%
Obese			
	No	4093	73.04%
	Yes	1511	26.96%

Table 4

2.3. GP Visits

Similar to Research Question 1, SAS was used to merge the GP and ED datasets and create a final dataset for unique patients with a flag for whether the ED patient also visited the GP. Table 5 below shows the proportion of ED patients who visited the GP. Out of the 5,604 unique ED patients, 1,378 (24.59%) had a visit to the GP.

ED Patients with GP Visit		
	Observations	Percentage
Visit to GP	1378	24.59%%
Did not visit the GP	4226	75.41%
Total	5604	

Table 5

2.4. Differences in Risk Factors

Further analysis conducted for the ED Manager aimed to analyse the differences in the risk factors for ED patients who did and did not visit the GP. Raw numbers, proportions, and chi-square tests were calculated and conducted to assess the differences. Table 6 below highlights the overall findings for this analysis.

The Smoker and Risky Alcohol User variables had statistically significant differences between ED patients who did and did not visit the GP. Obesity did not have a statistical difference between the two groups.

Within the ED population, both smokers and risky alcohol users visited the GP more than their respective counterparts. I assume that both of these patient populations suffer more health issues overall that would lead to them visiting their GP more often. There is also a possibility that many patients visiting the ED only do so because of an acute injury or illness. Besides this acute ailment, these patients are healthier and thus do not need to visit the GP on a regular basis.

Health Characteristics				
Characteristic		Visit to GP?		Chi-Square P-Value
Smoker		Yes (N=1378)	No (N=4226)	
	No	1027 (74.53%)	3699 (87.53%)	**<.0001**
	Yes	351 (25.47%)	527 (12.47%)	
Risky Alcohol User				
	No	996 (72.28%)	3482 (82.39%)	**<.0001**
	Yes	382 (27.72%)	744 (17.61%)	
Obese				
	No	1028 (74.60%)	3065 (72.53%)	0.132
	Yes	350 (25.40%)	1161 (27.47%)	

Table 6

2.5. Smoker Sensitivity and Specificity

The ED Manager expressed concern that the ED process for recording smoking status might have issues. I carried out the following analysis to ascertain the sensitivity and specificity of the ED smoking status when compared to the GP smoking status, the gold standard. Please note that missing values for smoking status in the GP dataset were not included in the below numbers. The sensitivity and specificity table and calculation are shown below in Figure 2.

The ED process for recording smoking status does a better job at ascertaining non-smokers than smokers, as evidenced by the specificity value of 88.19% and the sensitivity value of 80.63%.

While both values are high, there is room for improvement in the process to move these numbers closer to 100%. The next section will stratify these metrics by key demographic variables to get a better idea of where improvements can be made.

ED Dataset : Smoking Status	GP Dataset : Current Smoker		Total
	Yes	No	
Yes	204	130	334
No	49	971	1020
Total	253	1101	1354

Note : 24 missing current smoker values not included

Sensitivity :	80.63%
Specificity :	88.19%

Figure 2

2.6. Stratified Smoker Sensitivity and Specificity

In order to dig deeper into the ED process for recording smoking status, the sensitivity and specificity were calculated for stratified key demographic variables including sex, country of birth, healthcare card, and age category.

The stratified sensitivity and specificity for sex is shown below in Figure 3. The first table and calculations are for males, and the second table and calculations are for females. Females had a higher sensitivity and specificity than males, indicating an opportunity in the process when recording male smoking status especially.

ED Dataset : Smoking Status	GP Dataset : Current Smoker Sex = Male		Total
	Yes	No	
Yes	118	72	190
No	32	452	484
Total	150	524	674

Note : 18 missing current smoker values not included

Sensitivity :	78.67%
Specificity :	86.26%

ED Dataset : Smoking Status	GP Dataset : Current Smoker Sex = Female		Total
	Yes	No	
Yes	86	58	144
No	17	519	536
Total	103	577	680

Note : 6 missing current smoker values not included

Sensitivity :	83.50%
Specificity :	89.95%

Figure 3

The stratified sensitivity and specificity for country of birth is shown below in Figure 4. The first table and calculations are for patients born in Australia, and the second table and calculations are for patients born overseas. Patients born in Australia had a higher sensitivity than patients born overseas, but a lower specificity. This indicates an opportunity to work on recording overseas patients who are smokers.

ED Dataset : Smoking Status	GP Dataset : Current Smoker Country of Birth = Australia		Total
	Yes	No	
Yes	55	71	126
No	8	488	496
Total	63	559	622

Note : 6 missing current smoker values not included

Sensitivity :	87.30%
Specificity :	87.30%

ED Dataset : Smoking Status	GP Dataset : Current Smoker Country of Birth = Overseas		Total
	Yes	No	
Yes	149	59	208
No	41	483	524
Total	190	542	732

Note : 18 missing current smoker values not included

Sensitivity :	78.42%
Specificity :	89.11%

Figure 4

The stratified sensitivity and specificity for healthcare card is shown below in Figure 5. The first table and calculations are for patients with no healthcare card and the second table and calculations are for patients with a healthcare card. Patients with a healthcare card had a higher sensitivity and specificity than patients with no card, indicating an opportunity in the process when recording the smoking status for patients with no healthcare card.

ED Dataset : Smoking Status	GP Dataset : Current Smoker Healthcare Card = No		Total
	Yes	No	
Yes	150	99	249
No	39	712	751
Total	189	811	1000

Note : 16 missing current smoker values not included

Sensitivity :	79.37%
Specificity :	87.79%

ED Dataset : Smoking Status	GP Dataset : Current Smoker Healthcare Card = Yes		Total
	Yes	No	
Yes	54	31	85
No	10	259	269
Total	64	290	354

Note : 8 missing current smoker values not included

Sensitivity :	84.38%
Specificity :	89.31%

Figure 5

The stratified sensitivity and specificity for age category is shown below in Figure 6. The first table and calculations are for patients aged less than 60 and the second table and calculations are for patients aged 60 or above. Patients 60 or above had a higher sensitivity than patients less than 60, but a lower specificity. This indicates an opportunity to work on recording younger patients who are smokers.

ED Dataset : Smoking Status	GP Dataset : Current Smoker Age < 60		Total
	Yes	No	
Yes	180	106	286
No	45	822	867
Total	225	928	1153

Note : 20 missing current smoker values not included

Sensitivity :	80.00%
Specificity :	88.58%

ED Dataset : Smoking Status	GP Dataset : Current Smoker Age >= 60		Total
	Yes	No	
Yes	24	24	48
No	4	149	153
Total	28	173	201

Note : 4 missing current smoker values not included

Sensitivity :	85.71%
Specificity :	86.13%

Figure 6

2.7. Recommendations

From above, I note several opportunities to improve the ED screening and recording of smoking status. First, there is a significant overall opportunity to increase both the sensitivity and specificity closer to 100%. By analysing the stratified sensitivity and specificity, I have the following recommendations to the ED manager:

- To increase the sensitivity of the process, focus primarily on accurate recording for males, patients born overseas, patients with no healthcare card, and patients less than 60.
- To increase the specificity of the process, focus primarily on accurate recording for males, patients born in Australia, patients with no healthcare card, and patients 60 or older.
- Overall, I would recommend primarily focusing efforts on male patients born overseas to increase both the sensitivity and specificity.

3. Research Question 3

3.1. Create Smoking Cohort

In order to assess the baseline uptake of medicines for smoking cessation using PBS data linked to GP and ED data, I first must create a cohort of smokers. To create the smoking cohort, I first identify all smokers in each of the GP and ED datasets and the first date when they were identified as a smoker. In the GP dataset, I choose the patient ID and visit date as start date since each patient record is unique. I find 793 total smokers in the GP dataset. In the ED dataset, I choose the first ED visit where a patient had a smoking diagnosis and use the ED admission date as the start date. I find 878 total smokers in the ED dataset. To create the final smoking cohort, I append these two datasets, retaining the patient ID, the respective start date, and the dataset identifier. I then sort the records by start date and keep only the first record for each ID (earliest identification of smoker status) to create the final smoking cohort. Table 7 shows the breakdown of unique patients in the smoking cohort, and the dataset in which they were first identified as a smoker.

Smoking Cohort		
Dataset	Observations	Percentage
ED	755	51.47%
GP	712	48.53%
Total	1467	

Table 7

3.2. Therapy Usage

Using my smoking cohort, I set out to ascertain the therapy usage of these patients by merging this dataset with the PBS data. Before I do so, I conduct an examination of the PBS dataset. The PBS dataset contains 3,216 observations and 4 variables, and all supply dates are from 2014. I did observe 21 full duplicates in the PBS dataset, but I decided to keep the duplicates as similar prescriptions can be filled by the same person on the same day. I also observe that there are 79 NRT patch, 1 Bupropion, and 235 Varenicline records in the PBS dataset, which can have multiple records for each ID.

First, I merged the smoking cohort dataset with the PBS dataset, creating a flag to identify patients in the cohort that had a matching PBS record. Then, I created four new variables that would be needed in the analysis. The four variables created are NRT_use, Bupropion_use , Varenicline_use, and medicine_flag. These flags were based on the act codes in the PBS dataset. The medicine_flag identified a patient that used at least one of the therapies. Using these flags, I created a person level dataset that included each patient in my smoking cohort and which therapy (or therapies) they used. Table 8 below shows the proportion of smokers who used any of the therapies and the overall proportion of smokers who used a therapy in 2014.

Cohort Therapy Usage			
Therapy		Observations	Percentage
Varenicline			
	No	1316	89.71%
	Yes	151	10.29%
NRT patches			
	No	1410	96.11%
	Yes	57	3.89%
Bupropion			
	No	1466	99.93%
	Yes	1	0.07%
Any Therapy			
	No	1261	85.96%
	Yes	206	14.04%

Table 8

3.3. Update Smoking Cohort

I am also interested in utilizing the PBS data to update and enhance the smoking cohort already created. Since the PBS data includes the supply date for the various smoking related therapies, I can use this data combined with my smoking cohort to assess the impact on cohort size.

Similar to the process conducted to create the initial smoking cohort, I find all records in the PBS dataset that were supplied a smoking related therapy. I then sort by supply date and keep only the first time an ID received a medication. From this, I observe that there are 216 unique patients who received a smoking related medication. 216 is higher than the 206 I originally found linked to my smoking cohort, so I am assuming there will be additions to the smoking cohort.

I append this dataset to the already created smoking cohort using the supply date as the start date. I then sort this dataset and keep the first identification of smoking using the new data from PBS. Table 8 shows the breakdown of unique patients in the smoking cohort, and the dataset in which they were first identified as a smoker. I observe that the overall smoking cohort increased from 1467 to 1477 patients. I also see that 96 of the cohort comes from the PBS data, and the cohort numbers from both the ED and GP dataset decreased. I would absolutely use the PBS as an additional source to identify people who smoke for two reasons. One, the overall cohort number increases. Two, since the PBS identified patients earlier, I have better information to conduct longitudinal studies and aid the Sunnydale Population Health Department in their clinical trial.

Smoking Cohort with PBS Data		
Therapy	Observations	Percentage
ED	720	48.75%
GP	661	44.75%
PBS	96	6.50%
Total	1477	

Table 9

4. Research Question 4

4.1. Data Linkage Strategy

A sound data linkage strategy will need to be employed by CHeReL in order to link the GP, RBDM, and PBS datasets for the GP Practice Manager. A sound data linkage strategy will result in a high-quality dataset that can be used to examine medication compliance among the GP patients.

The RBDM dataset contains names, addresses, and dates of birth as patient identifiers. The GP dataset contains names, addresses, dates of birth, and the Medicare number as patient identifiers. Finally, the PBS dataset contains only the Medicare number as a patient identifier. In order to create a high-quality dataset, the CHeReL will need to employ both deterministic and probabilistic linking algorithms to link these three datasets.

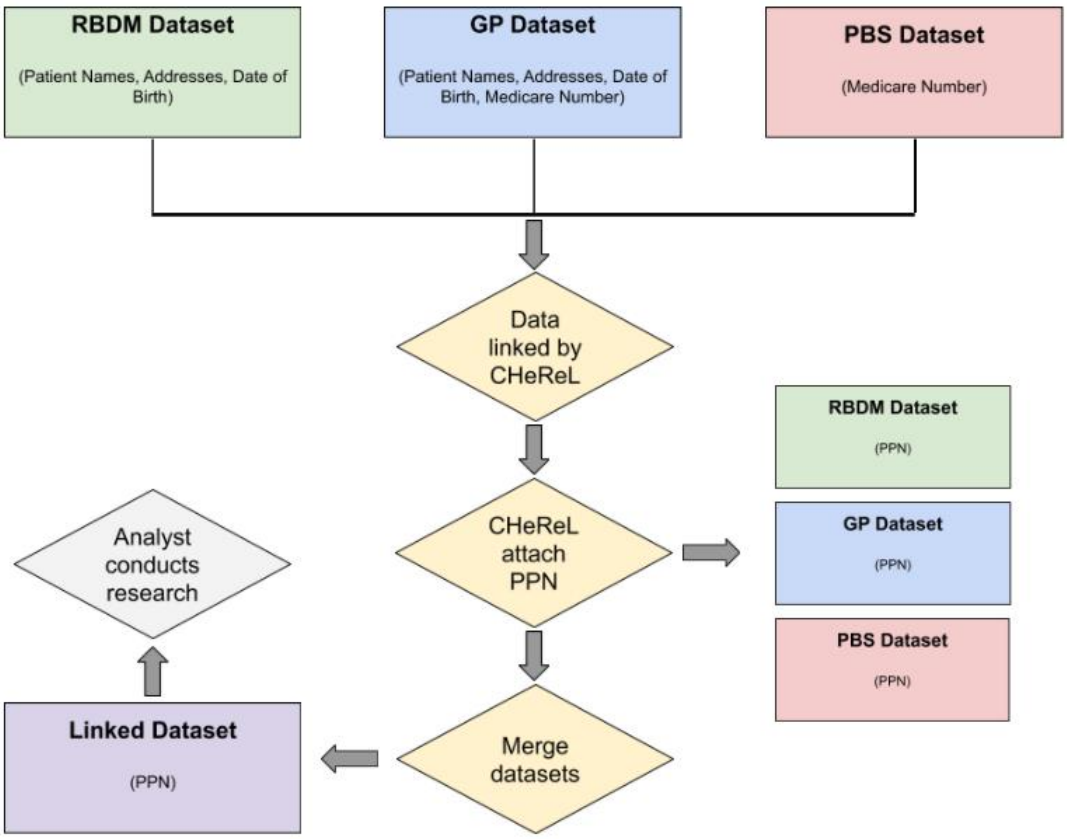
Linking between the RBDM and GP dataset will need to employ a probabilistic linking algorithm based on the name, address, and date of birth fields. This process involves calculating a m (likelihood of agreeing given records are the same) and u (likelihood of agreeing given records are different) probability for each record[1]. From these values, a field weight is calculated and if the field weight exceeds a defined cut-off threshold, the record pair is deemed a link[1]. The CHeReL aims to achieve a 5/1000 false positive and negative rate and will update the cut-off threshold to meet this goal.

To link to the PBS dataset, a deterministic linking algorithm will be used on the Medicare number field between the GP and PBS dataset. Records will be linked if the Medicare number unique identifier is a match. This method does have one drawback, as the Medicare number is not a perfect unique identifier. Multiple people can be listed under one Medicare card, and one person can have multiple Medicare cards. In this case, CHeReL will have to assume that the person attending the GP with a given Medicare number will also be the same person receiving medication under the Medicare number.

Using these two data linkage strategies, CHeReL will be able to create a high-quality dataset that can aid the Medical Plus GP Manager in examining medication compliance among the GP patient population.

4.2. Variable Exchange Diagram

The diagram below depicts the flow of variable exchange information between the data custodians, the CHeReL, and the analyst conducting the research. All information in parentheses highlights the personal identification present in that specific dataset.



4.3. Journal Submission

Prior to submitting to the Medical Journal of Australia, the data analyst should ensure the research is reproducible and the privacy of patients is protected. The analyst needs to include proper documentation detailing the research and should conduct a thorough statistical disclosure control process.

Proper documentation detailing the research includes but is not limited to describing the study setting, design, analysis methods, and results. If possible, the analysis code used in the research study should be shared. The documentation should include the data cleaning steps, especially any record exclusions that occurred during the process. The documentation should also

include updated data dictionaries, including newly created variables. If possible, the analyst should also have a data retention strategy in the event that future research requires the use of the data. The goal of this documentation is that an independent research team can accurately replicate the findings.

Depending on the granularity and type of research conducted, there might be sensitive topics covered and certain reported numbers might increase the risk of disclosing personal information. The analyst will want to address any risks present in the reported tabular data. Risks of disclosure includes small n cells, cells with a high contribution from only a few respondents, and cells where external information is available that could be used to disclose confidential information. If any of these risks are apparent, then techniques such as perturbation, cell suppression, or controlled tabular adjustment need to be used in order to protect the information.

5. References

1. National Statistical Service, Australia Government, accessed 11 August 2020, <https://media.openlearning.com/9E7L7zqmxrwgQGsQxYGJnptG4TQFpJefVjD5B5RkvdfDaeZ4NLuLMjJsbBBJQ98S.1538704733/NSS_probabilistic_datalink.pdf>