MSCR 534: Analysis Exercise 1

Anish Shah

February 14, 2020

Assignment Description

- Present epidemiological data in table format for communicating findings
- Logistic regressions are used for both crude and adjusted associations
- Demonstrate software skills to create logistic models, interprate model results, and build multivariable models

General models for analysis:

- Model A
  - Exposure = HIV status
  - Outcome = site of extra-pulmonary TB (EPTB)
- Model B
  - Exposure = country of birth
  - Outcome = prison diagnosis of TB

Table 1. Bivariate associations and crude OR for outcome site of EPTB

Requirements

- *Table 1* style figure (similar to prior Analysis Exercise 1 from fall semester)
- Instead of Total column, present crude odds ratios (95% CI) between covariates and site of EPTB
- Dichotomous outcome variable from XPSITE should be used as well
- Include 5 participant characteristics in Table 1 (including 1 continuous variable), along with primary exposure of HIV
- Each characteristic should have crude odds ratios (95% CI), indicate referent category for each

Data intake and tidying

The SAS dataset *EPTB* was read in and processed/cleaned, to make the following simple data set with covariates chosen by clinical importance.

Covariates chosen for Table 1

ID	XPSITE_NOM	HIV	GEN	AGE	CD4ADM	ARV	PreviousTB
7334	Other	Positive	Male	40	73	No	No
7259	CNS	Positive	Male	50	99	No	No
3307	Other	Positive	Female	42	6	Yes	No
3300	CNS	Positive	Male	56	84	No	Yes
3291	Other	Positive	Male	46	240	No	No
3250	Other	Positive	Female	38	883	No	Yes

The previous data was presented in the following "Table 1" style format, with crude odds ratios for each parameter.

Table 1. Characteristics by EPTB Site

	Other N=228	CNS N=67	OR	p.ratio
HIV Status:				
Negative	117 (51.3%)	25 (37.3%)	Ref.	Ref.
Positive	111 (48.7%)	42 (62.7%)	1.76 [1.01;3.12]	0.045
Sex:				
Female	76 (33.3%)	16 (23.9%)	Ref.	Ref.
Male	152 (66.7%)	51 (76.1%)	1.58 [0.86;3.05]	0.143
Age (years)	39.7 (11.6)	41.5 (12.9)	1.01 [0.99;1.04]	0.270
Anti-retroviral Use:				
No	131 (57.5%)	44 (65.7%)	Ref.	Ref.
Yes	12 (5.26%)	7 (10.4%)	1.75 [0.61;4.68]	0.289
'Missing'	85 (37.3%)	16 (23.9%)	0.56 [0.29;1.05]	0.071
CD4 Count	156 (183)	156 (188)	1.00 [1.00;1.00]	0.995
H/o Active TB:				
No	162 (71.1%)	47 (70.1%)	Ref.	Ref.
Yes	18 (7.89%)	9 (13.4%)	1.73 [0.69;4.05]	0.229
'Missing'	48 (21.1%)	11 (16.4%)	0.80 [0.37;1.62]	0.541

Table 2. Bivariate associations and crude OR for outcome TB diagnosed in prison

- Similar to Table 1 above, however outcome is prison diagnosis, and exposure is country of birth
- Create new prison dx variable - assume *missing* prison dx had actually been diagnosed in prison (use this as outcome variable)
- Instead of Total column, present crude odds ratios (95% CI) between covariates and prison diagnosis. Columns should consist of those who had prison dx and those who did not.
- Include 5 participant characteristics in Table 1 (including 1 continuous variable), along with primary exposure of country of birth
- Each characteristic should have crude odds ratios (95% CI), indicate referent category for each

Date intake and tidying

The following covariates were chosen, and presented as a sample data set below.

Covariates chosen for Table 2

ID	PRISON_DX	COUNTRY	PreviousTB	COPU	DrugUse	AlcoholAbuse
7334	No	USA	No	Yes	Yes	Yes
7259	No	USA	No	No	Yes	Yes
3307	Yes	USA	No	Yes	Yes	Yes
3300	Yes	USA	Yes	Yes	Yes	Yes
3291	No	USA	No	Yes	Yes	Yes
3250	No	USA	Yes	No	Yes	Yes

A "table 1" style figure was created to show the association between covariates and the outcome of TB diagnosis while in prison.

Table 2. Characteristics by Prison Diagnosis of TB

	No N=233	Yes N=63	OR	p.ratio
Birthplace:				
USA	175 (75.1%)	57 (90.5%)	Ref.	Ref.
Foreign	58 (24.9%)	6 (9.52%)	0.33 [0.12;0.74]	0.006
Sex:				
Female	85 (36.5%)	7 (11.1%)	Ref.	Ref.
Male	148 (63.5%)	56 (88.9%)	4.49 [2.08;11.3]	<0.001
Age (years)	40.5 (12.4)	38.3 (9.53)	0.98 [0.96;1.01]	0.180
H/o Active TB:				
No	171 (73.4%)	39 (61.9%)	Ref.	Ref.
Yes	21 (9.01%)	6 (9.52%)	1.27 [0.44;3.22]	0.638
'Missing'	41 (17.6%)	18 (28.6%)	1.93 [0.98;3.69]	0.056
Concurrent Pulmonary Dz:				
No	140 (60.1%)	34 (54.0%)	Ref.	Ref.
Yes	93 (39.9%)	29 (46.0%)	1.28 [0.73;2.25]	0.386
Illegal Drug Use:				
No	169 (72.5%)	12 (19.0%)	Ref.	Ref.
Yes	59 (25.3%)	35 (55.6%)	8.21 [4.09;17.6]	<0.001
'Missing'	5 (2.15%)	16 (25.4%)	42.2 [13.9;153]	<0.001
Alcohol Abuse:				
No	139 (59.7%)	14 (22.2%)	Ref.	Ref.
Yes	90 (38.6%)	34 (54.0%)	3.71 [1.92;7.54]	<0.001
'Missing'	4 (1.72%)	15 (23.8%)	34.8 [10.9;140]	<0.001

Table 3. Multivariable model for association between HIV and site of EPTB

- Model A purpose is to estimate unbiased association between HIV and EPTB
- Create DAG to demonstrate hypothesized causal relationship of variables in Table 1 (including covariates)
- Based on DAG, build model A with crude and adjusted OR in Table 3. Adjusted model (regardless of important covariates or DAG) should include HIV and previous TB as independent variables

Table 4. Multivariable model for association between country of birth and TB diagnosis in prison

- Similar to table 3 above. Model B purpose is to estimate unbiased association between country of birth and prison diagnosis
- Create DAG to demonstrate hypothesized causal relationship of variables in Table 2 (including covariates). Relationship between country of birth and covariates should be thoroughly explored.
- Based on DAG and observed bivariate associations, build model B and report crude and adjusted OR in Table 4. Adjusted model should at minimum contain country of birth and age as independent variables.

Questions

1. Titles

1. What are the titles for Tables 1-4?

2. Table 1

2A. How was the outcome variable of XPSITE dichotomized? Why was the decision made to dichotomize the variable using the categories chosen?

2B. Which covariate had the strongest measure of association with the site of EPTB variable you created?

2C. Interpret the measure of association reported in part 2B using one sentence.

2D. Which covariate had the lowest p-value in its association with the site of EPTB variable you created? What was the statistical test used and what was the p-value?

2E. Interpret the p-value reported in part 2D using one sentence.

3. Table 2

3A. What was the prevalence ratio for a TB prison diagnosis, comparing those born in the US to those born outside the US?

3B. What was the odds ratio for prison diagnosis, comparing those born in the US to those born outside the US?

3C. Does the odds ratio estimate the prevalence ratio? Why or why not? Explain your answer in one sentence.

3D. Did the assumption that those with missing prison information had actually been diagnosed in prison change the estimated measure of association toward the null, away from the null, or had no effect?

4. Table 3

4A. Consider your DAG for Model A. Are there any unblocked backdoor paths from HIV to the EPTB variable? If yes, list the pathway.

4B. Consider your DAG for Model A. Are there any colliders? If yes, list the colliders.

4C. Based on bivariate analyses, are any covariates strongly associated with the outcome (EPTB site) or with the exposure of interest (HIV)? Are there any significantly associated with both?

4D. State the expression for the odds ratio of EPTB site comparing someone with HIV to someone without HIV in the adjusted model.

5. Table 4

5A. In a short paragraph (<= 4 sentences) describe the model building strategy used for Model B.

5B. As if you were writing a sentence for a Results section of a scientific article, report the main findings of Model B. Use only one sentence.

5C. In Model B, did the assumption that those with missing prison information had actually been diagnosed in prison change the adjusted estimated measure of association toward the null, away from the null, or had no effect? How does this compare to part 3D?