MSCR 534: Analysis Exercise 1

Anish Shah

February 14, 2020

```
# Knitr options
knitr::opts_chunk$set(
  cache = TRUE,
  warning = FALSE,
  eval = TRUE,
  echo = TRUE,
  include = TRUE,
  message = FALSE,
  dpi = 600,
  dev = "png",
  options("scipen" = 999, "digits" = 3),
  tinytex.verbose = TRUE,
  tidy = FALSE
)

options(xtable.comment = FALSE)
```

```
library(tidyverse)
library(knitr)
library(rmarkdown)
library(magrittr)
library(compareGroups)
library(haven)
library(kableExtra)
library(DiagrammeR)
library(stargazer)
```

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.

```
# Data
eptb <- read_sas("eptb.sas7bdat")
df1 <- eptb

# Create dichotomous variable for XPSITE
# Original labels were ... XPSITE 2 = CNS, 5 = Lymph
# For us, will do dichotomous for CNS = 1, other = 0
df1$XPSITE_NOM[df1$XPSITE == 2] <- 1
df1$XPSITE_NOM[df1$XPSITE %in% c(1,3:9)] <- 0 # There is 1 NA value here

# Outcome = xpsite, exposure = HIV
# Select 5 covariates: GEN, AGE, RACECORR, CD4ADM, ARV, PreviousTB,

### Clean up data

# Outcome
df1$XPSITE_NOM %<>%
    factor(., levels = c(0, 1), labels = c("Other", "CNS"))
attr(df1$XPSITE_NOM, "label") <- "EPTB Site"

# Exposure
df1$HIV %<>%
    factor(., levels = c(0, 1), labels = c("Negative", "Positive"))
attr(df1$HIV, "label") <- "HIV Status"

# Sex
df1$GEN %<>% factor(., levels = c(0, 1), labels = c("Female", "Male"))
attr(df1$GEN, "label") <- "Sex"

# Age
attr(df1$AGE, "label") <- "Age (years)"

# Previous TB
df1$PreviousTB %<>%
    factor(., levels = c(0,1), labels = c("No",  "Yes"))
attr(df1$PreviousTB, "label") <- "H/o Active TB"

# CD4ADM
```

```r
attr(df1$CD4ADM, "label") <- "CD4 Count"

# ARV
df1$ARV %<>%
    factor(., levels = c(0,1), labels = c("No", "Yes"))
attr(df1$ARV, "label") <- "Anti-retroviral Use"

### Present this data

# Head
subset(df1, select = c(ID, XPSITE_NOM, HIV, GEN, AGE,
                       CD4ADM, ARV, PreviousTB)) %>%
    head(.) %>%
    kable(., "latex", caption = "Covariates chosen for Table 1",
          booktabs = TRUE) %>%
    kable_styling(latex_options = c("HOLD_position"))
```

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

Table 1 creation

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

```
# Data from above, will lose attributes if I subset it

# CompareGroups style table
compareGroups(
    XPSITE_NOM ~ HIV + GEN + AGE + ARV + CD4ADM + PreviousTB,
    data = df1, include.label = TRUE, simplify = FALSE,
    include.miss = TRUE
    ) %>%
    createTable(
        show.n = FALSE, show.p.overall = FALSE,
        show.ratio = TRUE
    ) %>%
    export2md(format = "latex", caption = "Table 1. Characteristics by EPTB Site")
```

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.

```
# Data to use
df2 <- eptb

### Create new variable about prison status
# If NA, then actually dx in prison = 1
df2$PRISON_DX <- df2$PRISON
df2$PRISON_DX[is.na(df2$PRISON)] <- 1

### Clean the data

# Outcome
df2$PRISON_DX %<>%
    factor(., levels = c(0,1), labels = c("No", "Yes"))
attr(df2$PRISON_DX, "label") <- "TB Dx in Prison"

# Exposure
df2$COUNTRY %<>%
    factor(., levels = c(1,0), labels = c("USA", "Foreign"))
attr(df2$COUNTRY, "label") <- "Birthplace"

### Select / clean covariates: AGE, GEN, PreviousTB, Homeless, DrugUse, AlcoholAbuse, COPU

# Sex
df2$GEN %<>% factor(., levels = c(0, 1), labels = c("Female", "Male"))
attr(df2$GEN, "label") <- "Sex"

# Age
attr(df2$AGE, "label") <- "Age (years)"

# Previous TB
df2$PreviousTB %<>%
    factor(., levels = c(0,1), labels = c("No", "Yes"))
attr(df2$PreviousTB, "label") <- "H/o Active TB"

# Pulmonary disease
df2$COPU %<>%
    factor(., levels = c(0,1), labels = c("No", "Yes"))
attr(df2$COPU, "label") <- "Concurrent Pulmonary Dz"
```

```r
# Drug Use
df2$DrugUse %<>%
    factor(., levels = c(0,1), labels = c("No", "Yes"))
attr(df2$DrugUse, "label") <- "Illegal Drug Use"

# Alcohol Abuse
df2$AlcoholAbuse %<>%
    factor(., levels = c(0,1), labels = c("No", "Yes"))
attr(df2$AlcoholAbuse, "label") <- "Alcohol Abuse"

### Present this data

# Head
subset(df2, select = c(
    ID, PRISON_DX, COUNTRY, PreviousTB, COPU, DrugUse, AlcoholAbuse
    )) %>%
    head(.) %>%
    kable(., "latex", caption = "Covariates chosen for Table 2",
          booktabs = TRUE) %>%
    kable_styling(latex_options = c("HOLD_position"))
```

Covariates chosen for Table 2

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

Table 2 creation

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

```
# Data from above

# CompareGroups style table
compareGroups(
    PRISON_DX ~
        COUNTRY + GEN + AGE + PreviousTB + COPU + DrugUse + AlcoholAbuse,
    data = df2, include.label = TRUE, simplify = FALSE,
    include.miss = TRUE
) %>%
    createTable(
        show.n = FALSE, show.p.overall = FALSE,
        show.ratio = TRUE
    ) %>%
    export2md(format = "latex", caption = "Table 2. Characteristics by Prison Diagnosis of TB")
```

Table 2. Characteristics by Prison Diagnosis of TB

	No N=233	Yes N=63	OR	p.ratio
Birthplace:				
USA	58 (24.9%)	6 (9.52%)	Ref.	Ref.
Foreign	175 (75.1%)	57 (90.5%)	3.07 [1.35;8.40]	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

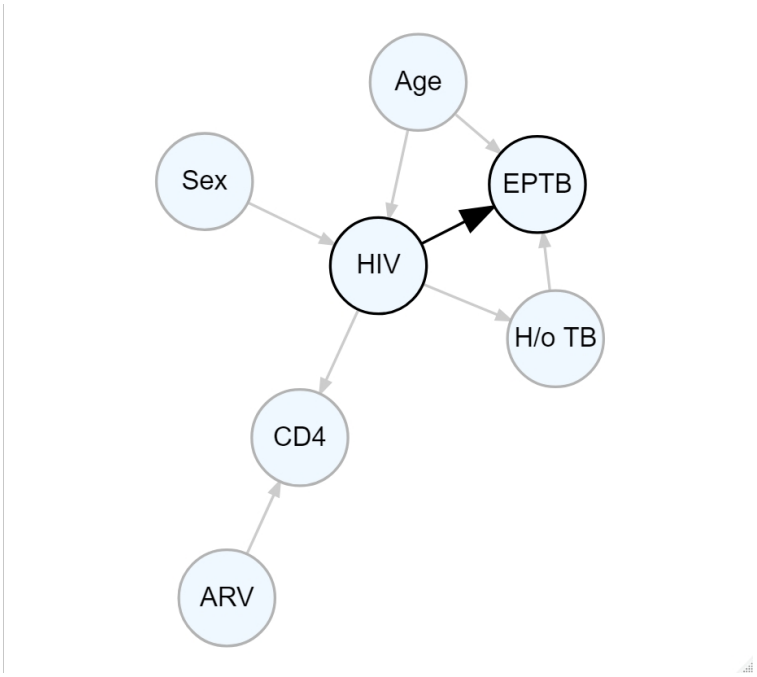
- Model A purpose is to estimate unbiased association between HIV and EPTB
- Create DAG to demonstrate hypothesized causal relationship of variables in Table 1 (including covariates)
- Based on DAG, build model A with crude and adjusted OR in Table 3. Adjusted model (regardless of important covariates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

Directed acyclic graph for HIV and EPTB

Figure 1. DAG for Table 1 Variables. Exposure is HIV, and outcome is EPTB. Hypothetical relationship of other covariates shown.

```
# Need to make DAG, using diagrammer likely

### Include all nodes:

# Diagrammer approach

# Nodes
# XPSITE_NOM, HIV, GEN, AGE, ARV, CD4, PreviousTB
nodes <-
    create_node_df(
    n = 7,
    type = c("outcome", "exposure", rep("covariate", 5)),
    label = c("EPTB", "HIV", "Age", "Sex", "ARV", "CD4", "H/o TB")
    )

# Relationships
edges <-
    create_edge_df(
        from = c(2,7,3,4,5,2,2,3),
        to =   c(1,1,2,2,6,6,7,1)
    )

# Modifications
nodes$color[nodes$id == 1 | nodes$id == 2] <- "black"
edges$color[edges$id == 1] <- "black"
edges$arrowsize[edges$id == 1] <- 1.2

# Make graph
# Save this as it won't insert into a PDF otherwise
create_graph(
    nodes_df = nodes, edges_df = edges,
    graph_name = "DAG of Table 1 Variables"
) %>%
    render_graph(layout = "kk")
```

```
include_graphics("./dag1.jpeg")
```



Model A

The following, Table 3, shows the crude and adjusted odds ratios as developed from the Table 1 analyses and from the DAG in Figure 1.

```
# Data will be as above

# Crude analysis
m <- glm(XPSITE_NOM ~ HIV, family = binomial, data = df1)

# Adjusted analysis
n <- glm(XPSITE_NOM ~ HIV + AGE + PreviousTB, family = binomial, data = df1)

# Present as table
stargazer(m, n,
          type = "latex", header = FALSE,
          title = "Table 3. Effect of HIV status on EPTB Site",
          no.space = FALSE, single.row = FALSE,
          apply.coef = exp,
          ci = TRUE, p.auto = FALSE, report = "vc*s",
          ci.custom = list(exp(confint(m)), exp(confint(n))),
          dep.var.labels = c("CNS (versus other) EPTB Site"),
          covariate.labels = c("HIV Status", "Age", "Prior Active TB"),
          column.labels = c("Unadjusted", "Adjusted")
)
```

Table 3. Effect of HIV status on EPTB Site

	Dependent variable:	
	CNS (versus other) EPTB Site	
	Unadjusted	Adjusted
	(1)	(2)
HIV Status	1.770**	1.840*
	(1.020, 3.130)	(0.981, 3.520)
Age		1.010
		(0.981, 1.040)
Prior Active TB		1.510
		(0.600, 3.580)
Constant	0.214***	0.146***
	(0.136, 0.323)	(0.041, 0.490)
Observations	295	235
Log Likelihood	−156.000	−126.000
Akaike Inf. Crit.	316.000	261.000
Note:	*p<0.1; **p<0.05; ***p<0.01	

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.

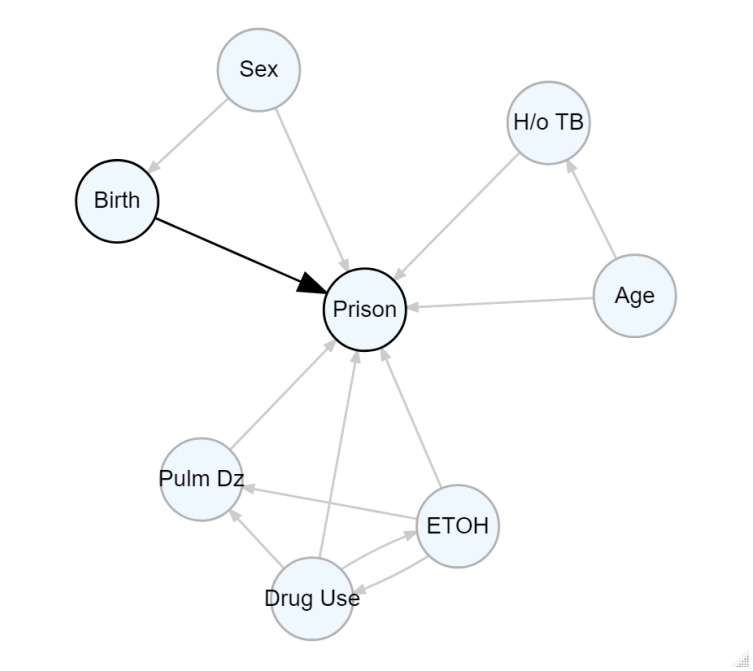
Directed acyclic graph

Below is the DAG for the covariates from table 2.

Figure 2. DAG for Table 2 Variables. Exposure is Birthplace, and outcome is prison diagnosis of TB. Hypothetical relationship of other covariates shown.

```
# Need to make DAG, using diagrammer likely

### Include all nodes:

# Diagrammer approach

# Nodes
# XPSITE_NOM, HIV, GEN, AGE, ARV, CD4, PreviousTB
nodes <-
    create_node_df(
    n = 8,
    type = c("outcome", "exposure", rep("covariate", 6)),
    label = c("Prison", "Birth", "Sex", "Age", "H/o TB", "Pulm Dz", "Drug Use", "ETOH")
    )

# Relationships
edges <-
    create_edge_df(
        from = c(2,3,4,4,5,7,8,7,8,6,3,8,7),
        to =   c(1,2,1,5,1,1,1,8,7,1,1,6,6)
    )

# Modifications
nodes$color[nodes$id == 1 | nodes$id == 2] <- "black"
edges$color[edges$id == 1] <- "black"
edges$arrowsize[edges$id == 1] <- 1.2

# Make graph
# Save this as it won't insert into a PDF otherwise
create_graph(
    nodes_df = nodes, edges_df = edges,
    graph_name = "DAG of Table 1 Variables"
) %>%
    render_graph(layout = "kk")
```

```
include_graphics("./dag2.jpeg")
```



Model B

Based on the DAG from Figure 2, and the covariates/significant predictors from Table 2, unadjusted and adjusted models were created (Cox proportional hazard models).

```
# Data will be as above

# Crude analysis
m <- glm(PRISON_DX ~ COUNTRY, family = binomial, data = df2)

# Adjusted analysis
n <- glm(PRISON_DX ~ COUNTRY + AGE + GEN + PreviousTB + COPU + DrugUse + AlcoholAbuse, family = binomial

# Present as table
stargazer(m, n,
          type = "latex", header = FALSE,
          title = "Table 4. Effect of Birthplace on Prison Dx of TB",
          no.space = FALSE, single.row = FALSE,
          apply.coef = exp,
          ci = TRUE, p.auto = FALSE, report = "vc*s",
          ci.custom = list(exp(confint(m)), exp(confint(n))),
          dep.var.labels = c("In-Prison Diagnosis of TB"),
          covariate.labels = c("Birth Country", "Age", "Sex", "Prior Active TB", "Pulmonary Dz", "Drug U
          column.labels = c("Unadjusted", "Adjusted")
)
```

Table 4. Effect of Birthplace on Prison Dx of TB

	Dependent variable:	
	In-Prison Diagnosis of TB	
	Unadjusted	Adjusted
	(1)	(2)
Birth Country	3.150**	2.050
	(1.380, 8.500)	(0.469, 14.400)
Age		0.967
		(0.925, 1.010)
Sex		5.280***
		(1.780, 19.900)
Prior Active TB		1.110
		(0.308, 3.550)
Pulmonary Dz		1.550
		(0.704, 3.490)
Drug Use		6.120***
		(2.730, 14.600)
Alcohol Abuse		1.830
		(0.792, 4.360)
Constant	0.103***	0.028***
	(0.040, 0.221)	(0.002, 0.226)
Observations	296	231
Log Likelihood	−149.000	−82.800
Akaike Inf. Crit.	303.000	182.000
Note:	*p<0.1; **p<0.05; ***p<0.01	

Questions

1. Titles

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

- Table 1: Characteristics by EPTB Site
- Table 2: Characteristics by Prison Diagnosis of TB
- Table 3: Effect of HIV status on EPTB Site
- Table 4: Effect of Birthplace on Prison Dx of TB

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?

The variable XPSITE was dichotomized into CNS-location versus other location. CNS penetration has an exceptionally high mortality, and was of particular interest as it is in an immuno-privileged location, particularly as it relates to HIV status. Thus, understanding if HIV affected CNS spread of TB was of additional interest.

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

Of the covariates, positive HIV status had the highest OR of 1.76 (1.01 - 3.12) out of the statistically significant covariates.

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

Having HIV increases the OR of developing CNS-site EPTB by 1.76 (1.01 - 3.12) compared to other EPTB sites.

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?

The covariate with the lowest p-value was HIV status ($p = 0.045$). The measure of association was an odds ratio. The test statistic was χ^2, and the confidence intervals were constructed using a median-unbiased estimation method.

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

With an $\alpha = 0.05$, and a $p = 0.045$, there is enough evidence to reject the null hypothesis and accept the alternative that $OR \neq 1$.

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?

The prevalence ratio (PR) is 2.621.

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

The odds ratio is 3.07 of diagnosing TB in prison if born in the US compared to foreign-born. In the table 2, it shows an OR of 0.33 (as being born outside of the US reduces the OR of having a prison diagnosis of TB) - this can be inverted to ~3.0.

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

In this case, the OR and PR are in the same direction, but are dissimilar in size (~15% difference between estimates). Because the prevalence of prison diagnosis is high in the exposed (USA birth) and unexposed (foreign birth), the $OR \neq PR$.

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?

```
# Temp data to check original prison dx with missing pts
x <- glm(PRISON ~ COUNTRY, family = binomial, data = eptb)

# Relabed missing
y <- glm(PRISON_DX ~ COUNTRY, family = binomial, data = df2)
```

If the missing were excluded, the OR would have been approximately 7.0. Our calculated OR by converting the missing values into prison diagnosis was approximately 3.0. Thus, our assumption change the measure of association towards the null (null assuming OR would be 1).

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.

Yes, there are unblocked, backdoor paths from HIV to EPTB. Per the DAG in Figure 1...

- Age = Age has a backdoor path to HIV, and goes forward to EPTB, creating a backdoor unblocked pathway from HIV to EPTB through Age

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

Yes, CD4 is a collider as it has an arrow going into it from HIV, as well as from ARV.

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?.

Per Table 1, HIV status was the only covariate significantly associated with EPTB. The relationship of the exposure was not previously tested with the other covariates. Before creating a model, this would need to be done (usually would perform a correlation table). However, for this question, will recreate a table of the covariates with HIV status.

```
# CompareGroups style table
compareGroups(
    HIV ~ XPSITE_NOM + GEN + AGE + ARV + CD4ADM + PreviousTB,
    data = df1, include.label = TRUE, simplify = FALSE,
    include.miss = TRUE
    ) %>%
    createTable(
        show.n = FALSE, show.p.overall = FALSE,
        show.ratio = TRUE
    ) %>%
    export2md(format = "latex", caption = "Characteristics by HIV Status")
```

Characteristics by HIV Status

	Negative N=142	Positive N=154	OR	p.ratio
EPTB Site:				
Other	117 (82.4%)	111 (72.1%)	Ref.	Ref.
CNS	25 (17.6%)	42 (27.3%)	. [.;.]	.
'Missing'	0 (0.00%)	1 (0.65%)	. [.;.]	.
Sex:				
Female	58 (40.8%)	34 (22.1%)	Ref.	Ref.
Male	84 (59.2%)	120 (77.9%)	2.43 [1.47;4.06]	0.001

Characteristics by HIV Status *(continued)*

	Negative	Positive	OR	p.ratio
Age (years)	40.3 (14.6)	39.8 (8.71)	1.00 [0.98;1.02]	0.737
Anti-retroviral Use:				
No	44 (31.0%)	132 (85.7%)	Ref.	Ref.
Yes	0 (0.00%)	19 (12.3%)	. [.;.]	.
'Missing'	98 (69.0%)	3 (1.95%)	. [.;.]	.
CD4 Count	430 (257)	137 (162)	1.00 [0.99;1.00]	<0.001
H/o Active TB:				
No	103 (72.5%)	107 (69.5%)	Ref.	Ref.
Yes	6 (4.23%)	21 (13.6%)	3.30 [1.34;9.42]	0.008
'Missing'	33 (23.2%)	26 (16.9%)	0.76 [0.42;1.36]	0.355

The measures that were significantly associated with HIV status were sex, CD4, and history of active TB. There were not enough values available to test the relationship between ARV use, but every individual on ARV had HIV positive, while none of the HIV negative patients were on ARVs.

There were no covariates 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.

The expression is below. Both sides can be exponentiated to retrieve the OR from the coefficient.

$$\begin{aligned} log\left(\frac{P(EPTB = CNS|HIV+)}{P(EPTB = CNS|HIV-)}\right) &= \beta_0 + \beta_1 HIV + \beta_2 AGE + \beta_3 PriorTB \\ &= \frac{\beta_1 HIV(1)}{\beta_1 HIV(0)} \\ &= \beta_1 HIV(1) \end{aligned}$$

5. Table 4

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

To build Model B, all clinically relevant covariates were selected taht could affect the primary outcome, Prison Diagnosis. The significant associations by bivariate analysis were noted (birthplace, sex, drug use, alcohol abuse) to be included in the final model. Then, a DAG was generated to show possible relationships of exposure to outcome. All variables that had a backdoor, open path from exposure to outcome were included. In addition, all competing exposures were included in the regression model.

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.

The unadjusted odds for being diagnosed with TB while in prison was 3.15 (1.38, 8.50) higher in those that were born in the USA compared to those born in foreign countries.

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?

In part 3D, we found that assuming missing data were those that had prison diagnosis of TB baised the estimated association towards the null. I created the model again and excluded missing (versus changing their coding).

```
# Data will be EPTB
df <- eptb
```

```r
# Exposure
df$COUNTRY %<>%
    factor(., levels = c(1,0), labels = c("USA", "Foreign"))
attr(df2$COUNTRY, "label") <- "Birthplace"


x <- glm(PRISON ~ COUNTRY, family = binomial, data = df)
y <- glm(PRISON_DX ~ COUNTRY, family = binomial, data = df2)

# Present as table
stargazer(x, y,
          type = "latex", header = FALSE,
          title = "Difference between reassignment v. exclusion of missing data",
          no.space = FALSE, single.row = FALSE,
          apply.coef = exp,
          ci = TRUE, p.auto = FALSE, report = "vc*s",
          ci.custom = list(exp(confint(x)), exp(confint(y))),
          dep.var.labels = c("In-Prison Diagnosis of TB"),
          covariate.labels = c("Birth Country"),
          column.labels = c("Excluded", "Reassigned")
)
```

Difference between reassignment v. exclusion of missing data

	Dependent variable:	
	In-Prison Diagnosis of TB	PRISON_DX
	Excluded	Reassigned
	(1)	(2)
Birth Country	7.620***	3.150**
	(2.260, 47.600)	(1.380, 8.500)
Constant	0.034***	0.103***
	(0.006, 0.110)	(0.040, 0.221)
Observations	281	296
Log Likelihood	−122.000	−149.000
Akaike Inf. Crit.	248.000	303.000
Note:		*p<0.1; **p<0.05; ***p<0.01

As per this new table, we can see that exclusion of patients with missing data had an OR of 7.6, while the missing reassigment had an OR of 3.2. This suggests that the reassignment of missing led to a bias of the estimate towards the null. This is similar to 3D, when we compared odds ratios. As this was done in logistic regressions that were unadjusted, we would expect the findings to be very similar to a crude odds ratio (which we did find as seen in 3D).