Week ON13

Elmer V Villanueva

18 May 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON14")
```

Announcements

- The remaining coursework assessment and the final paper have deadlines as follows

Assessment	Due Date	Days Till Deadline
Coursework 2	30 May	5
Final Paper	17 June	23

- All the assessments are to be completed INDIVIDUALLY. No collusion is allowed. This means that you cannot discuss this assignment with other students, nor can you share your approach or code.
- The running count for students forwarding errors is as follows:

Student	Items Identified
Yijia Jiang	6
Jing Wang	3
Yuxuan Wu	2
Xinwen Hu	1
Yuxin Zhang	1

- This is second last week of sessions. Next week will be our last week.

Reading

Read and understand Vittinghoff et al., Chapter 5.

Examples of Reporting of Logistic Regression Results in the Primary Literature

Similar to our examination of examples of the reporting or results from linear regression models, we will spend this sessions evaluating some examples showing logistic regression results. This will be helpful for your second coursework assessment.

Example 1: Oppong et al. [1]

Objectives

The authors wanted to study the association between race/ethnicity and breast density. By now, you should be able to find the objectives of the study.

Methods

Note how the dependent variable, originally classified using four categories (blue highlight in Figure 1), is "dichotomised" or reduced into two categories (yellow highlight in Figure 1) for the purposes of the analysis. This technique is easy to accomplish, but any reduction in data will result in a loss of information. Thus, you need to balance the advantages of dichotomisation against its downsides.

2.3.2 | Breast density classification—We used the American College of Radiology BI-RADS to classify density. BI-RADS classification consists of 4 categories: (1) almost entirely fat, (2) scattered fibroglandular densities, (3) heterogeneously dense, and (4) extremely dense. The most recent mammogram available in the electronic medical record maintained by MedStar Health, the health system to which the Ourisman Center and GUMC belong, was used. Two board-certified radiologists (E.M. and E.P.) independently classified each participant. For the CBCC subset, the recorded density on the mammography report was used (no independent review of imaging).

For this analysis, breast density was categorized into low (categories 1 and 2) and high (categories 3 and 4). Figure 1 depicts mammographic images of a fatty or low density breast (A) compared to a dense breast (B).[22]

Figure 1: Section 2.3.2 in Oppong et al.

The statistical analysis section is quite brief. Note that the authors only mention multivariable (or multiple) logistic regression as their modelling technique. There is also no information about the types of models diagnostics that were conducted. This is clearly inappropriate.

Results

Similar to the examples we evaluated previously, Oppong et al. use Table 1 to describe the characteristics of their study population. In this case, note how the group is divided according to levels of the dependent variable. That is, there is a column for low density and another for high density mammograms.

In Figure 2, I give a portion of the paper's Table 1. The p-value presented in the last column is derived from t-tests or chi-square tests, not from a logistsic regression. This was described in the statistical analysis section.

TABLE 1

Sample characteristics by mammographic density

Characteristics	Low density (n = 935)	High density (n = 1,211)	*P**
Age at screening (years), mean ± SD	53.33 ± 8.51	48.44 ± 7.89	<.01
Race, n (%)			
Hispanic	296 (33.1)	597 (66.9)	<.01
White	122 (38.9)	192 (61.1)	
Black	517 (55.1)	422 (44.9)	
Education, n (%)			
High school or less	449 (45.3)	543 (54.7)	.12
Some college	180 (45.6)	215 (54.4)	
College or more	224 (39.4)	344 (60.6)	
Missing	82 (42.9)	109 (57.1)	

Figure 2: Portion of Table 1 in Oppong et al.

The main results arising from the multiple logistic regression model are presented in Table 2 of the paper. The table is recreated in Figure 3.

TABLE 2

Odd ratios (OR) and 95% confidence intervals (95% CI) for the association between selected variables and mammographic density based on logistic regression

Variables	Adjusted OR* (95% CI)
Age	0.94 (0.93, 0.96)
Race	
Hispanic	1.00
Black	0.47 (0.38, 0.59)
White	0.38 (0.28, 0.53)
Reproductive parity	
Nulliparous	1.00
1 or 2	0.83 (0.54, 1.28)
3+	0.62 (0.40, 0.96)
Menopausal status	
Premenopausal	1.00
Postmenopausal	0.72 (0.56, 0.93)
Age of menarche	1.05 (1.00, 1.11)
Family history of breast cancer	
No	1.00
Yes	1.02 (0.80, 1.30)
BMI (kg/m^2)	
Nonobese (BMI <30 kg/m^2)	1.00
Obese (BMI ≥30 kg/m^2)	0.29 (0.23, 0.37)

OR, odds ratio; CI, confidence intervals; BMI, body mass index.

*OR = odds ratio for each variable is adjusted for the other variables.

Figure 3: Portion of Table 2 in Oppong et al.

Note that the authors do not present the regression equation. This is typical. As I mentioned in a previous lecture, presenting information in logit units is very difficult for most people to understand. Thus, the typical presentation of results is in odds ratios (ORs).

The second column of the table states that the odds ratios are already adjusted for all other variables in the table. This might be a little difficult to understand, so we will interpret some results. Let us begin with the main independent variable: race. This variable is itnerpreted in this manner:

- Compared to Hispanics, Blacks have an odds ratio of 0.45 (95% CI 0.38, 0.59) for high density compared to low density, after controlling for age, reproductive parity, menopausal status, age of menarche, family history of breast cancer and body mass index.
- Compared to Hispanics, Whites have an odds ratio of 0.62 (95% CI 0.40, 0.96) for high density compared to low density, after controlling for age, reproductive parity, menopausal status, age of menarche, family history of breast cancer and body mass index.

Would you be able to derive the regression equation from the data provided by the authors? Partially. The authors do not provide information about one regression estimate: the intercept. The others, however, may be estimated from the odds ratios that are presented. The regression equation for the Oppong et al. study is

$$logit(Low\ Density) = \beta_0 + ln(0.94)AGE + ln(0.47)BLACK + ln(0.38)WHITE + ln(0.83)PARITY12 + ln(0.62)PARITY3MORE + ln(0.72)POSTMENOPAUSAL + ln(1.05)MENARCHE + ln(1.02)FAMILYHISTORY + ln(0.29)OBESE.$$

This reduces to

$$logit(Low\ Density) = \beta_0 - 0.0619AGE - 0.7550BLACK - 0.9676WHITE - 0.1863PARITY12 - 0.4780PARITY3MORE - 0.3285POSTMENOPAUSAL + 0.0488MENARCHE + 0.0198FAMILYHISTORY - 1.238OBESE.$$

Finally, we have no information about the fitness of the model because no diagnostic tests were reported.

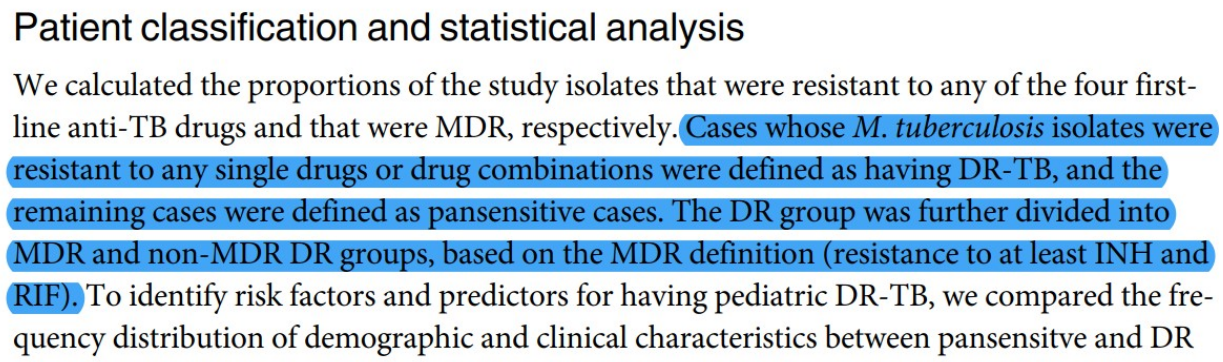
Example 2: Guo et al. [2]

Objectives

The authors wanted to evaluate the relationship between drug resistance to tuberculosis drugs and the characteristics of a paediatric patient or his or her clincial care.

Methods

The dependent variable is easy to understand conceptually, but notice how the authors make it very complicated (Figure 4). There are three groups, pansensitive TB, DR-TB and MDR-TB. Given that there are three groups, will a logistic regression work? Since a logistic regression only deals with binary outcomes, the authors seem to have run three separate logistic regression models. The first compares pansensitive with DR TB; the second compares pansensitive with MDR TB; the last compares DR-TB with MDR-TB. This is quite inefficient. There are models designed for dependent variables with more than two groups. The authors should have used those models, instead of three separate logistic regressions.



Patient classification and statistical analysis

We calculated the proportions of the study isolates that were resistant to any of the four first-line anti-TB drugs and that were MDR, respectively. Cases whose *M. tuberculosis* isolates were resistant to any single drugs or drug combinations were defined as having DR-TB, and the remaining cases were defined as pansensitive cases. The DR group was further divided into MDR and non-MDR DR groups, based on the MDR definition (resistance to at least INH and RIF). To identify risk factors and predictors for having pediatric DR-TB, we compared the frequency distribution of demographic and clinical characteristics between pansensitve and DR

Figure 4: Definition of Outcome Variable in Guo et al.

The rest of the statistical analysis section identifies univariate and multivariate logistic regression analysis as the main techniques used. Once again, there is no mention of diagnostic tests.

Results

The authors do not have a single table describing their study population. This is a major oversight. Instead, the ready is forced to examine Tables 2 to 4 to get a sense of the characteristics of the study population. Since these are the same tables that present the main results, it is quite difficult to read and understand the true study population.

Let us look at one of the logistic regression models reported by the authors. In Table 2 of the paper, the authors report the comparison between DR-TB and pansensitive TB. The table is shown in Figure 5.

Table 2. Comparison of demographic and clinical characteristics between pansensitive (n = 140) and DR-TB[a] (n = 56) diagnosed in the Children's Hospital of Chongqing Medical University during 2008–2013 using logistic regression models.

Characteristic	Pansensitive TB	DR-TB	p value[b]	DR-TB vs. Pansensitive TB	
	N (%)	N (%)		Crude OR (95% CI)	Adjust OR (95% CI)
Sex			0.36		
Male	85 (60.7)	30 (53.6)		Ref	Ref
Female	55 (39.3)	26 (46.4)		1.34 (0.72–2.50)	1.39 (0.70–2.77)
Age			0.80		
<5 year	67 (47.9)	24 (42.9)		Ref	Ref
5–9 years	29 (20.7)	12 (21.4)		1.16 (0.51–2.62)	1.25 (0.50–3.11)
≥ 10 years	44 (31.4)	20 (35.7)		1.27 (0.63–2.57)	1.74 (0.79–3.81)
Residence			0.67		
Rural	88 (62.9)	37 (66.1)		Ref	Ref
Urban	52 (37.1)	19 (33.9)		0.87 (0.45–1.67)	0.84 (0.41–1.73)
BCG history[c]			0.56		
Yes	87 (70.2)	38 (74.5)		Ref	Ref
No	37 (29.8)	13 (25.5)		0.80 (0.39–1.68)	0.68 (0.31–1.49)
Disease site			0.42		
Thoracic	62 (44.3)	25 (44.6)		Ref	Ref
Extrathoracic	22 (15.7)	5 (8.9)		0.56 (0.19–1.65)	0.61 (0.20–1.92)
Concurrent thoracic-extrathoracic	56 (40.0)	26 (46.4)		1.15 (0.60–2.22)	1.16 (0.55–2.42)
AFB[d] smear[e]			0.03		
Positive	69 (50.4)	18 (32.7)		Ref	Ref
Negative	68 (49.6)	37 (67.3)		2.09 (1.08–4.02)	2.33 (1.13–4.80)

[a] DR-TB represents drug-resistant tuberculosis.
[b] Based on Chi-square test.
[c] Based on 175 cases for which information on BCG vaccination history is available.
[d] AFB represents acid-fast bacilli.
[e] Based on 192 cases for which AFB smear was conducted.

Figure 5: Logistic regression results in Table 2 of Guo et al.

The results are presented as ORs and 95% CIs in the last two columns. The second last column presents "crude" or unadjusted ORs. That is, these only compare one variable at a time. The last column presents adjusted ORs. There is no notation, but the adjustment seems to have considered all other variables in the table.

Let us interpret some results.

Crude OR for sex. Female children have a 34% increase in the odds of having DR-TB versus pansensitive TB compared to male children.

Adjsuted OR for sex. Female children have a 39% icnrease in the odds of having DR-TB versus pansensitive TB compared to male children after controlling for age, residence, BCG history, disease site and AFB smear result.

Adjusted OR for age. Children aged between 5-9 years have an OR of 1.25 for DR-TB versus pansensitive TB compared to those aged less than 5 years, after controlling for sex, residence, BCG history, disease site and AFB smear result. Children at least 10 years of age have an OR of 1.74 for DR-TB versus pansensitive TB

compared to those aged less than 5 years, after controlling for sex, residence, BCG history, disease site and AFB smear result.

I leave to you the interpretation of results in Tables 3 and 4, as well as the derivation of the three partial regression equations.

References

1. Oppong BA, Dash C, O'Neill S, Li Y, Makambi K, Pien E, Makariou E, Coleman T, Adams-Campbell LL. Breast density in multiethnic women presenting for screening mammography. *Breast J* 2018 ;24:334-338.

2. Guo Q, PanY, Yang ZH, Liu RX, Xing LL, Peng Z, Zhu CM. Epidemiology and clinical characteristics of pediatric drug-resistant tuberculosis in Chongqing, China. *PLoS ONE* 2016;11:e0151303

THE END