Summative Quiz 3 (Multiple Cox Regression) Solutions:

1. What guides a researcher when deciding between using either linear, logistic, or Cox proportional hazards regression as an analysis tool?

Answer: The outcome variable type (continuous, binary, or time –to-event) for the particular analyses utilizing the regression model(s).

2. The generic formulation of a multiple regression model including age (x_1=age in years), sex (x_2= 1 if female, 0 if male), and smoking status (x_3 = 1 if smoker, and 0 if non-smoker) as predictors is as follows:

$$[LHS] = intercept + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

Where LHS = "left hand side", and intercept =$\hat{\beta}_o$ or $\ln(\hat{\lambda}_o(t))$.

What comparison is being made by the slope $\hat{\beta}_3$?

Answer: Smokers to non-smokers of the same sex and age.

Reasoning: Recall, the slope for a predictor from a multiple regression compares groups who differ by one-unit in the predictor value (here 1 – smoker, 0 = non-smoker), adjusted for all other predictors in the model

3. Suppose an interaction term x_4 is added to the model from question 2, where x_4=x_2*x_3 ("sex times smoking status") The resulting model is:

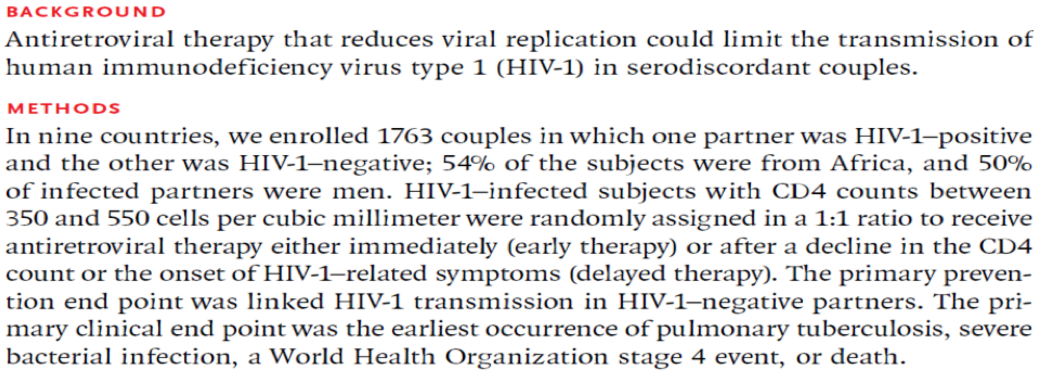
$$[LHS] = intercept + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$$

Where LHS = "left hand side", and intercept =$\hat{\beta}_o$ or $\ln(\hat{\lambda}_o(t))$.

Which of the following does the interaction term allow for?

Answer: Separate estimates of the age-adjusted relationship between the LHS and smoking status, for males and females.

4. Recall that article presenting the results from a randomized trial that enrolled HIV sero-discordant couples is several countries. As per the abstract:



BACKGROUND

Antiretroviral therapy that reduces viral replication could limit the transmission of human immunodeficiency virus type 1 (HIV-1) in serodiscordant couples.

METHODS

In nine countries, we enrolled 1763 couples in which one partner was HIV-1–positive and the other was HIV-1–negative; 54% of the subjects were from Africa, and 50% of infected partners were men. HIV-1–infected subjects with CD4 counts between 350 and 550 cells per cubic millimeter were randomly assigned in a 1:1 ratio to receive antiretroviral therapy either immediately (early therapy) or after a decline in the CD4 count or the onset of HIV-1–related symptoms (delayed therapy). The primary prevention end point was linked HIV-1 transmission in HIV-1–negative partners. The primary clinical end point was the earliest occurrence of pulmonary tuberculosis, severe bacterial infection, a World Health Organization stage 4 event, or death.

Additionally, the results from both simple and multiple Cox regression models were summarized in the following table:

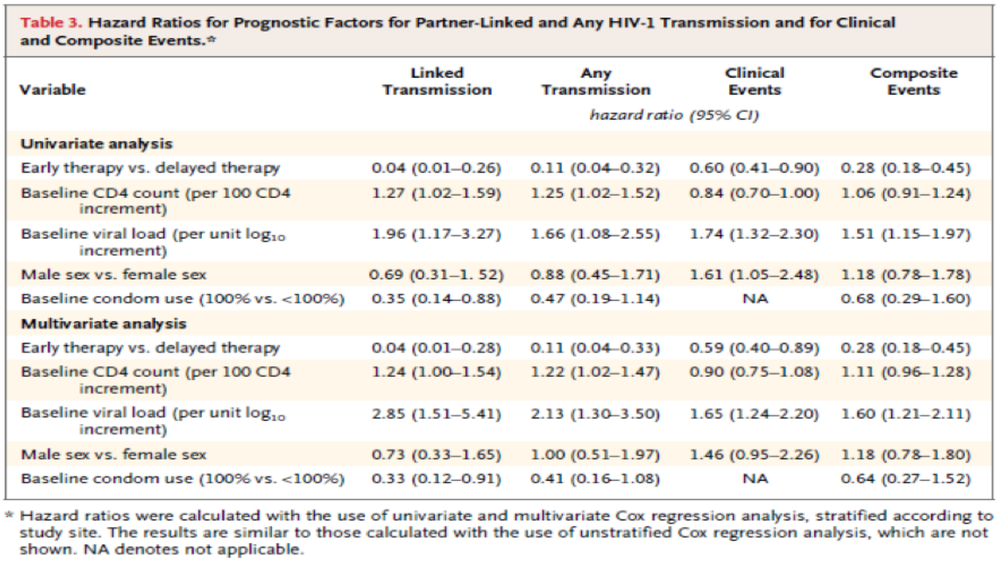


Table 3. Hazard Ratios for Prognostic Factors for Partner-Linked and Any HIV-1 Transmission and for Clinical and Composite Events.*

Variable	Linked Transmission	Any Transmission	Clinical Events	Composite Events
	hazard ratio (95% CI)			
Univariate analysis				
Early therapy vs. delayed therapy	0.04 (0.01–0.26)	0.11 (0.04–0.32)	0.60 (0.41–0.90)	0.28 (0.18–0.45)
Baseline CD4 count (per 100 CD4 increment)	1.27 (1.02–1.59)	1.25 (1.02–1.52)	0.84 (0.70–1.00)	1.06 (0.91–1.24)
Baseline viral load (per unit $\log_{10}$ increment)	1.96 (1.17–3.27)	1.66 (1.08–2.55)	1.74 (1.32–2.30)	1.51 (1.15–1.97)
Male sex vs. female sex	0.69 (0.31–1. 52)	0.88 (0.45–1.71)	1.61 (1.05–2.48)	1.18 (0.78–1.78)
Baseline condom use (100% vs. <100%)	0.35 (0.14–0.88)	0.47 (0.19–1.14)	NA	0.68 (0.29–1.60)
Multivariate analysis				
Early therapy vs. delayed therapy	0.04 (0.01–0.28)	0.11 (0.04–0.33)	0.59 (0.40–0.89)	0.28 (0.18–0.45)
Baseline CD4 count (per 100 CD4 increment)	1.24 (1.00–1.54)	1.22 (1.02–1.47)	0.90 (0.75–1.08)	1.11 (0.96–1.28)
Baseline viral load (per unit $\log_{10}$ increment)	2.85 (1.51–5.41)	2.13 (1.30–3.50)	1.65 (1.24–2.20)	1.60 (1.21–2.11)
Male sex vs. female sex	0.73 (0.33–1.65)	1.00 (0.51–1.97)	1.46 (0.95–2.26)	1.18 (0.78–1.80)
Baseline condom use (100% vs. <100%)	0.33 (0.12–0.91)	0.41 (0.16–1.08)	NA	0.64 (0.27–1.52)

* Hazard ratios were calculated with the use of univariate and multivariate Cox regression analysis, stratified according to study site. The results are similar to those calculated with the use of unstratified Cox regression analysis, which are not shown. NA denotes not applicable.

The "univariate analysis" results are from simple Cox regressions and are unadjusted, whereas the "multivariate analysis" results are from a multiple regression, and are adjusted. The predictors listed are characteristics of the HIV positive partner in each of the serodiscordant couples in the study.

The following question is related to the outcome of "Linked Transmission".
What two groups are being compared by the adjusted hazard ratio for "baseline CD4 count"?

Answer: Two groups of couples where the HIV positive partners differ by CD4 cell counts of 100, but are the same in terms of the other factors used in the multiple Cox regression model.

5. (this item references the same logistic regression results as item #4)

The following question is related to the outcome of "Linked Transmission".

Based on the adjusted results, estimate the hazard ratio (and 95% CI) of linked transmission in serodiscordant couples where the HIV partners had baseline CD4 counts of 500 compared to serodiscordant couples where the HIV partners had baseline CD4 counts of 200.

Answer: 1.9 (1.0, 3.65)

Reasoning: So these groups differ by 3 units on the CD4 count measure. On the ratio scale, the ratio, and its confidence interval endpoints can be raised to the 3rd power:

1.24^3 $(1.00^3, 1.54^3) \rightarrow$ 1.91 (1.00, 3.65)

Just a reminder of the mathematics as to why the above works: on the regression scale, the results look like this:

$$\ln(\text{hazard of death: t, xs}) = \ln(\hat{\lambda}_o(t)) + \hat{\beta}_1 x_1 + (\text{other } \beta\text{s and xs})$$

, where x_1= CD4 count measure (as defined in the table), and $\hat{\beta}_1 = \ln(1.24)$
So, on the regression scale, if were to compute the adjusted difference in the ln(hazard), equivalent to the ln(hazard ratio) for two groups who differ be 3 units, this difference is $3\hat{\beta}_1$ $(\hat{\beta}_1 + \hat{\beta}_1 + \hat{\beta}_1)$. Exponentiating this to get the hazard ratio yields $e^{3\hat{\beta}_1} = (e^{\hat{\beta}_1})^3 = 1.24^3$.

6. (this item references the same logistic regression results as item #4)

Likely, why are the unadjusted and adjusted hazard ratios for Early therapy (versus delayed) therapy identical, and the resulting unadjusted and adjusted 95% CIs so similar?

Answer: As this was a randomized study, the relationship between linked transmission and treatment is unlikely to be confounded by other factors.

7. (this item references the same logistic regression results as item #4)

   The following question is related to the outcome of "Linked Transmission".
   What is the estimated adjusted hazard ratio of linked transmission for couples where the HIV positive partner had a baseline viral load of 1,000,000 copies/ml versus couples where the HIV positive partner had a baseline viral load of 10,000 copies/ml?

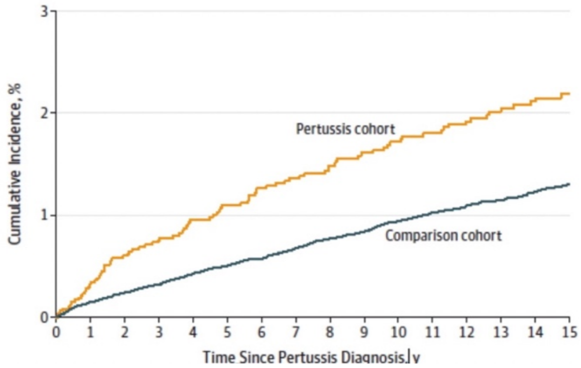
Answer: 8.10

Reasoning: 2.85^2= 8.12 which rounds to 8.1

8. (An observational cohort study appearing in *JAMA* examines the relationship between pertussis and the development of epilepsy in matched cohort of Danish children. The primary endpoint is a diagnosis of epilepsy. This will be referred to as the "epilepsy" in the related exercises. Subjects with pertussis were followed from the date of pertussis diagnosis until epilepsy (or censoring). Subjects without pertussis (the comparison cohort) were followed from the date of pertussis diagnosis for their matched subject with pertussis until epilepsy (or censoring) (reference: Olson M, et al.The Risk and Prevention Study Collaborative Group. Hospital-Diagnosed Pertussis Infection in Children and Long-term Risk of Epilepsy. *JAMA* (2015).)

   The unadjusted association between epilepsy and pertussis is visually displayed in the following Kaplan-Meier curve:

Figure. Cumulative Incidence of Epilepsy in 4700 Patients With a Hospital Diagnosis of Pertussis in Denmark During 1978-2011 and a Comparison Cohort Matched (1:10) on Birth Year and Sex



The Cox proportional hazard model used to quantify this (unadjusted) association is as follows:

$$\ln(hazard\ of\ the\ pimary\ endpoint; t, x_1) = \ln\left(\hat{\lambda}_o(t)\right) + 0.6x_1$$

where $x_1 = 1$ for subjects with pertussis and 0 for subjects without pertussis (comparison cohort). *The standard error of the slope for x_1 is 0.10*

What does the function ln(lambda_hat[t]) characterize?

Answer: The ln(hazard) of epilepsy for the comparison cohort as a function of time across the follow-up period.

Reasoning: Recall, the "intercept" for a Cox proportional hazards model is a function of time: it tracks the ln(hazard) of the outcome over time, for the group whose predictor values are all 0. In this case, there is only one dichotomous predictor, x_1: $x_1 = 0$ for those in the comparison group.

9. (this item makes a reference to the result presented with item #8)

   Report the unadjusted hazard ratio (and 95% CI) for the hazard of epilepsy for the pertussis cohort compared to the comparison across the 15 years follow-up period.

   Answer: 1.8 (1.5, 2.2)

   Reasoning: The slope estimate, $\hat{\beta}_1$=0.60, and the 95% CI for the population level slope is

   $\hat{\beta}_1 \pm 2\widehat{SE}\left(\hat{\beta}_1\right) \rightarrow 0.60 \pm 2(0.10) \rightarrow (0.40,0.80)$. To get to the odds ratio scale, exponentiate the results. :$e^{\widehat{\beta}_1} = e^{0.6} \approx 1.8$, and $(e^{0.4}, e^{0.8}) \rightarrow (1.5, 2.2)$.

10. (this item makes a reference to the result presented with item #8)

What is the (unadjusted) difference in the 15 year (cumulative) risk of epilepsy for the pertussis cohort compared to the comparison cohort?

Answer: 1%

Reasoning: While it is true that this cannot be estimated from the simple Cox regression model without additional information (and even then it would not be a straightforward computation, but would involve a computer), the Kaplan-Meier curve shows the unadjusted cumulative risk of epilepsy over time for both the pertussis and comparison cohorts. At 15 years, thus cumulative risk for the pertussis group is (slightly greater than) 2% as compared to (slightly greater than) 1% in the comparison group, for a difference of approximately 1%. (This also serves as a continued reminder that relative comparisons, such as hazard ratio, only quantify part of the "story" with regard to binary and time-to-event outcomes.)

11. (this item makes a reference to the result presented with item #8)

Lower gestational age is associated with increased risk for epilepsy in this cohort of Danish children. However, the gestational age adjusted hazard ratio (and 95% CI) of epilepsy for children the pertussis cohort compared to the comparison is almost identical to the unadjusted hazard ratio (and 95% CI). What is the best explanation for this among the following?

Answer: In this observational cohort study, gestational age is not related to pertussis.

Reasoning: Because the unadjusted and age-adjusted results are nearly identical, this indicates little to no confounding by gestational age. In order to confound the relationship, gestational age must be related to both the outcome (epilepsy) and the predictor (pertussis/no pertussis). It is given that gestational age is related to the outcome: as such, gestational age must not be related to the predictor (pertussis/no pertussis) for there to be no confounding.

12. (this item makes a reference to the result presented with item #8)

Suppose the researchers are interested in whether the relationship between epilepsy and pertussis is modified by gestational age (4 categories). What should they do to investigate this possibility of effect modification?

Answer: Estimate separate hazard ratios (and 95% CIs) of epilepsy for the pertussis groups compared to the controls for each of the four gestational age groups. Compare these four gestational age specific hazard ratios (and 95% CIs).

Note: This could be done by running a multiple Cox regression model with the following predictors: pertussis (yes/no), age b(four categories requiring 3 indicator x's), and interactions terms between the pertussis x and the 3 age category indicator xs

:

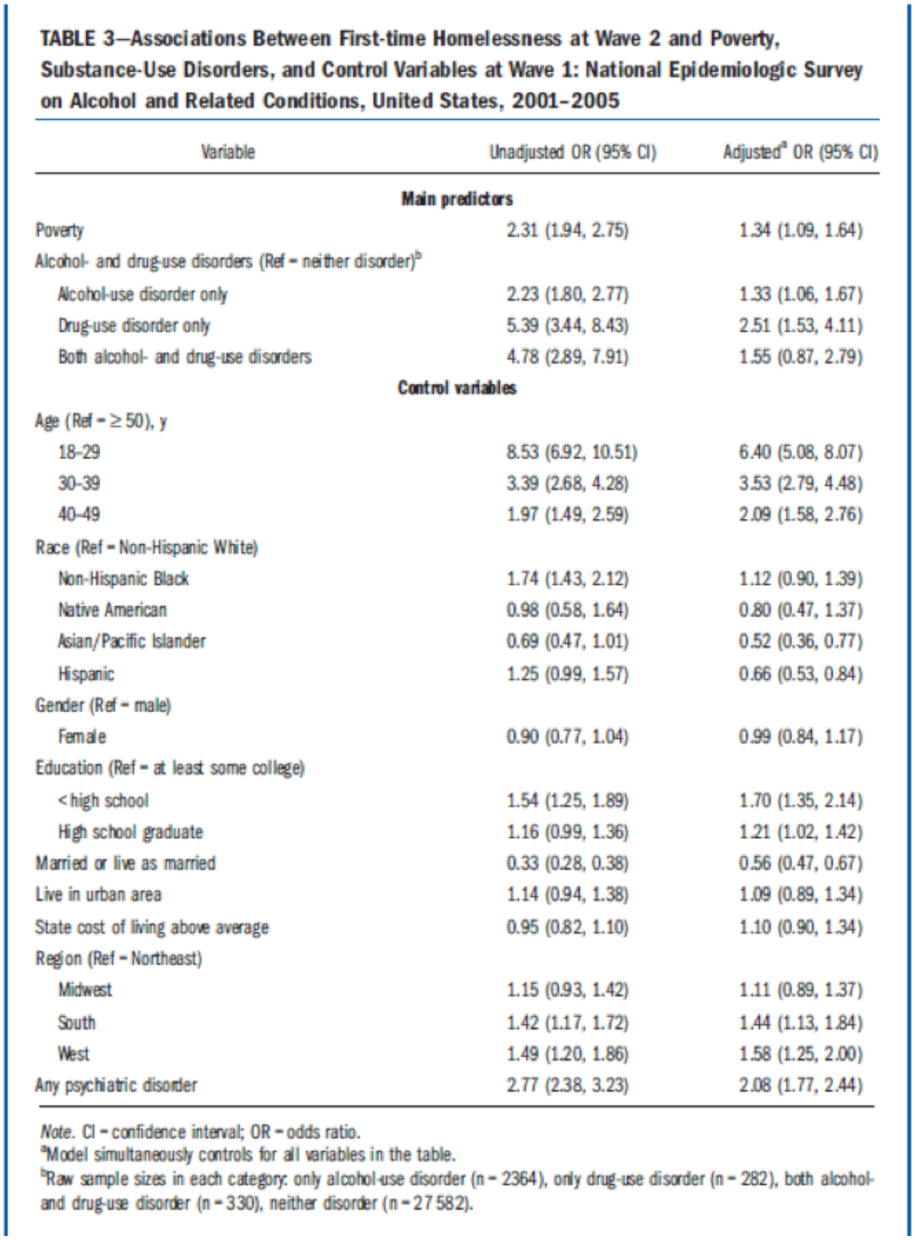


TABLE 3—Associations Between First-time Homelessness at Wave 2 and Poverty, Substance-Use Disorders, and Control Variables at Wave 1: National Epidemiologic Survey on Alcohol and Related Conditions, United States, 2001–2005

Variable	Unadjusted OR (95% CI)	Adjusted[a] OR (95% CI)
Main predictors		
Poverty	2.31 (1.94, 2.75)	1.34 (1.09, 1.64)
Alcohol- and drug-use disorders (Ref = neither disorder)[b]		
Alcohol-use disorder only	2.23 (1.80, 2.77)	1.33 (1.06, 1.67)
Drug-use disorder only	5.39 (3.44, 8.43)	2.51 (1.53, 4.11)
Both alcohol- and drug-use disorders	4.78 (2.89, 7.91)	1.55 (0.87, 2.79)
Control variables		
Age (Ref = ≥ 50), y		
18-29	8.53 (6.92, 10.51)	6.40 (5.08, 8.07)
30-39	3.39 (2.68, 4.28)	3.53 (2.79, 4.48)
40-49	1.97 (1.49, 2.59)	2.09 (1.58, 2.76)
Race (Ref = Non-Hispanic White)		
Non-Hispanic Black	1.74 (1.43, 2.12)	1.12 (0.90, 1.39)
Native American	0.98 (0.58, 1.64)	0.80 (0.47, 1.37)
Asian/Pacific Islander	0.69 (0.47, 1.01)	0.52 (0.36, 0.77)
Hispanic	1.25 (0.99, 1.57)	0.66 (0.53, 0.84)
Gender (Ref = male)		
Female	0.90 (0.77, 1.04)	0.99 (0.84, 1.17)
Education (Ref = at least some college)		
< high school	1.54 (1.25, 1.89)	1.70 (1.35, 2.14)
High school graduate	1.16 (0.99, 1.36)	1.21 (1.02, 1.42)
Married or live as married	0.33 (0.28, 0.38)	0.56 (0.47, 0.67)
Live in urban area	1.14 (0.94, 1.38)	1.09 (0.89, 1.34)
State cost of living above average	0.95 (0.82, 1.10)	1.10 (0.90, 1.34)
Region (Ref = Northeast)		
Midwest	1.15 (0.93, 1.42)	1.11 (0.89, 1.37)
South	1.42 (1.17, 1.72)	1.44 (1.13, 1.84)
West	1.49 (1.20, 1.86)	1.58 (1.25, 2.00)
Any psychiatric disorder	2.77 (2.38, 3.23)	2.08 (1.77, 2.44)

Note. CI = confidence interval; OR = odds ratio.

[a] Model simultaneously controls for all variables in the table.

[b] Raw sample sizes in each category: only alcohol-use disorder (n = 2364), only drug-use disorder (n = 282), both alcohol- and drug-use disorder (n = 330), neither disorder (n = 27 582).

Generally speaking, how do the adjusted results for substance abuse disorders compare to the unadjusted results?

Answer: While the three substance abuse disorders are still positively and statistically significantly associated (with the exception of alcohol and drug use disorder) with the risk of homelessness, the magnitudes of these associations attenuated (decreased) after adjustment

13. (this item refers to the same results presented in item #11)

Based on the multiple logistic regression results (the “adjusted column”) what is the estimated odds ratio of homelessness for :20 year olds living in poverty who have a drug use disorder compared to 55 year olds not in poverty with no substance abuse disorder, where both groups are otherwise comparable on the other adjustment variables?

Answer: 21.5.

Reasoning: 1.34*2.51*6.4=21.5

These differences are multiplicative on the odds ratio scale. (If you were to go back to the logistic regression scale, the difference would be additive:

The ln(odds ratio) would be $\hat{\beta}_{age\ 18-19} + \hat{\beta}_{poverty} + \hat{\beta}_{drug\ use\ disorder}$. Exponentiating this yields $e^{\hat{\beta}_{age\ 18-19}+\hat{\beta}_{poverty}+\hat{\beta}_{drug\ use\ disorder}} =$ $\left(e^{\hat{\beta}_{age\ 18-19}}\right)\left(e^{\hat{\beta}_{poverty}}\right)\left(e^{\hat{\beta}_{drug\ use\ disorder}}\right) =$ $(\widehat{AOR}_{age\ 18-19})(\widehat{AOR}_{poverty})(\widehat{AOR}_{drug\ use\ disorder})$, ie: the product of the 3 adjusted odds ratios.