1. **PRACTICE QUESTIONS** (please attempt to answer these practice questions 1a-e, but it is not required that you type your answers into this document).

   Use the data set “Epi203survival” in Stata and consider CD4 count (a measure of immune function) at baseline as the outcome variable and HIV viral load (the amount of HIV in the blood) at baseline as the exposure variable. The variable for the dichotomous classification of CD4 count is `cd4_c`, where 1=CD4 <350 and 0=CD4 ≥ 350. The variable “hivload_c” dichotomizes HIV viral load: 0 if <10,000 copies/mL and 1 if ≥ 10,000.

   (a) Construct the 2 x 2 table looking at the association between HIV viral load, classified as ≥10,000 and <10,000, and the presence of a CD4 count of less than 350. The Stata syntax for this can be found in the `epitab` command and in the lecture slides.

       **Practice question; click to see answer**

   (b) What is the absolute difference in the prevalence of a CD4 count below 350 in the two HIV viral load groups?

       **Practice question; click to see answer**

   (c) What is the prevalence ratio comparing the two HIV viral load groups for having a CD4 count less than 350?

       **Practice question; click to see answer**

   (d) What is the odds ratio comparing the two HIV viral load groups for having a CD4 count less than 350?

       **Practice question; click to see answer**

   (e) If the prevalence of a CD4 count less than 350 in this population was increased by a factor of 2 (i.e. doubled) in both the exposed and unexposed, how would this affect the prevalence ratio? How would this affect the odds ratio?

       **Practice question; click to see answer**
2. Use the Stata dataset “Epi203survival.dta” to evaluate the association between baseline CD4 count and subsequent AIDS diagnosis among HIV-infected men. The variable for the dichotomous classification of CD4 count is cd4_c, where 1=CD4 < 350 cells/mm³ (“exposed”) and 0=CD4 ≥ 350 cells/mm³ (“unexposed”). Assume no competing events. Clinical note: CD4 count is a measure of the status of the human immune system. Lower CD4 counts indicate more severe damage to the immune system.

(a) Construct a single graph showing the Kaplan-Meier estimate of the cumulative incidence of AIDS in the two groups (i.e., among men with CD4 < 350 and among men with CD4 ≥ 350). Show the point-wise confidence bands for each group. Make sure to label all axes. (1 pt)

(b) What is the log rank test statistic and p value for the comparison between the two curves in part (a)? Show your Stata output. (0.5 pts)

(c) What is the null hypothesis for this test? What does the log rank test tell you about the strength of association between CD4 count and subsequent AIDS? (0.5 pts)

(d) Calculate the risk ratio comparing a baseline CD4 lymphocyte count < 350 cells/mm³ to a baseline CD4 lymphocyte ≥ 350 cells/mm³ for the development of an AIDS diagnosis by 10 years. The Stata command for obtaining separate Kaplan-Meier estimates by levels of an exposure variable (assuming your data set has already been declared survival data: “stset”) is “sts list, by(exposure_variable)”. Describe your finding about the risk ratio in a clear, informative, and comprehensible sentence. (1 pt)

(e) Calculate the risk difference comparing a baseline CD4 lymphocyte count < 350 cells/mm³ to a baseline CD4 lymphocyte ≥ 350 cells/mm³ for the development of an AIDS diagnosis by 10 years. Describe your finding about the risk difference in a clear, informative, and comprehensible sentence that incorporates actual numbers of patients — in order to give the measure a flavor of public health or clinical impact. Assume that the role of baseline CD4 lymphocyte count in the occurrence of AIDS is causal and can be modified. (1 pt)

(f) Using the respective average incidence rates, calculate the rate ratio comparing a baseline CD4 lymphocyte count < 350 cells/mm³ to a baseline CD4 lymphocyte ≥ 350 cells/mm³ for the development of an AIDS diagnosis. Use all available observation time. Describe your finding about the rate ratio in a clear, informative, and comprehensible sentence. (1 pt)

(g) Using Stata, construct a single graph showing the hazard of AIDS in the two groups (i.e., among men with CD4 < 350 and among men with CD4 ≥ 350). Show the point-wise confidence bands for each group. Make sure to label all axes. (1 pt)
(h) FOR DISCUSSION IN SECTION ONLY:

Proportional hazards regression will estimate a hazard ratio between two or more exposure groups. It assumes that the hazard ratio between groups is the same across time (this is the “proportional hazards assumption”), and it will estimate this assumed constant hazard ratio by averaging the different hazard ratios at each time point over the entire range of the observation period. Another way to think about this is that the regression formula assumes that the hazard ratio is indeed constant over time in truth if you could study everyone but that any actual differences in the hazard ratio we see in our data are just due to the random variation we often see when just studying a sample of people. To get the best estimate of the constant hazard ratio, the machinery calculates an average of the hazard ratios at many different time points. When applied to these CD4 and AIDS data, the following output for a proportional hazards regression model is obtained. The model shown investigates the role of baseline CD4 count as a dichotomous exposure variable in influencing the subsequent occurrence of AIDS.

```
. stcox cd4_c
    failure _d: aids
    analysis time _t: obstime

Iteration 0:   log likelihood = -1207.6701
Iteration 1:   log likelihood = -1199.9626
Iteration 2:   log likelihood = -1195.0411
Iteration 3:   log likelihood = -1194.9652
Iteration 4:   log likelihood = -1194.9651
Refining estimates:
Iteration 0:   log likelihood = -1194.9651

Cox regression -- Breslow method for ties
No. of subjects =          375                Number of obs   =       375
No. of failures =          220
Time at risk    =  2664.071234
LR chi2(1)      =     25.41                Log likelihood  =   -1194.9651
Prob > chi2     =    0.0000
------------------------------------------------------------------------------
    _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
    cd4_c |   2.742937   .4910432     5.64   0.000     1.931225    3.895818
------------------------------------------------------------------------------
```

Why do you think that the incidence rate ratio calculated in part (e) is different from the hazard ratio of 2.74 calculated by proportional hazards regression?

(i) Numerically compare the rate ratio to the 10-year risk ratio. Give a conceptual explanation as to why the rate ratio differs from the 10-year risk ratio. In a context like the disease in this study, where the incidence of outcome is high, what happens to our ability to interpret the risk ratio as a measure of causation as observation time gets longer and longer? (Clinical note: In this dataset, the definition of AIDS is based on the original combination of various clinical manifestations and it does not specifically include any CD4 count criterion.) (2 pts)
3. There is considerable interest in getting eligible San Francisco residents registered to vote. Several years ago, two different civic organizations in San Francisco each developed registration drives which aimed to increase the number of new voter registrations. Six months prior to the voter registration deadline, each organization was randomly assigned 1000 unregistered San Francisco residents (who were eligible to vote) to target, and each organization used a different short-term (i.e., lasting just one week) technique to encourage voter registration. Of note, please assume that once a person is registered to vote, various political campaigns or public officials are not able to use this information in real-time in any way until the voter registration deadline has passed. After the election cycle was over, the civic organization leaders sought to determine which approach was more effective. They obtained the following data from the San Francisco voter registration office on the 2000 individuals (i.e., the 1000 targeted unregistered individuals by each of the two different civic organizations): did the individual register to vote (yes/no), and if yes, the date of registration. There were no deaths or movements out of San Francisco among the 2000 individuals studied. What measure of association should be used to judge which campaign was more effective? Explain your answer. (Subject matter note: In San Francisco during this period of time, one needed to be registered in order to be able to vote in public elections. The registration deadline was one month prior to the election.) (1 pt)
4. In a cross-sectional study of blood pressure in 642 persons, 148 were identified with a diastolic blood pressure greater than 90 mm Hg, which was defined as hypertension (high blood pressure) for the analysis. When participant age was dichotomized at < 50 years and ≥ 50 years, there were 381 participants 50 years or older, of whom 118 had hypertension.

(a) Construct the 2 x 2 table looking at the association between age, classified as ≥ 50 years old and < 50 years old, and the presence of hypertension. (1 pt)

(b) Calculate the absolute difference in the prevalence of hypertension in the two age groups and describe your result in a clear, informative, and comprehensible sentence or two. (1 pt)

(c) Calculate the prevalence ratio comparing the older age group to the younger age group for the presence of hypertension and describe your result in a clear, informative, and easily comprehensible sentence. (1 pt)

(d) Calculate a 95% confidence interval for the prevalence ratio. (Use the “cs1” immediate command in Stata or see Appendix A.3 in the S & N textbook; use the relative risk formula. In this formula, “log” means natural logarithm, often designated as “ln.”). (1 pt)

(e) Calculate the odds ratio comparing the two age groups for having hypertension and describe your result in a clear sentence. (1 pt)

(f) Calculate the odds ratio for not having hypertension (i.e., being normotensive) in the two age groups, comparing those aged ≥ 50 years with those < 50. (1 pt)

(g) Calculate the prevalence ratio for not having hypertension (i.e., being normotensive) in the two age groups, comparing those aged ≥ 50 years with those < 50. (1 pt)

(h) For the same set of data, odds ratios are typically more extreme in magnitude (i.e., farther away from 1) than prevalence ratios. Name the situation where this is not the case. (Note: do not consider situations where 0s are present in any cells). (1 pt)
5. A cross-sectional study was performed evaluating exposure to second-hand smoke as a cause of self-reported chronic fatigue. It showed the following (Diseased = fatigue; Non-Diseased = absence of fatigue):

<table>
<thead>
<tr>
<th></th>
<th>Diseased</th>
<th>Non-Diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>450</td>
<td>300</td>
</tr>
<tr>
<td>Unexposed</td>
<td>300</td>
<td>450</td>
</tr>
</tbody>
</table>

prevalence ratio = (450/750)/(300/750) = 1.5
odds ratio = (450/300)/(300/450) = 2.25 (formally, this is a prevalence odds ratio)

The investigators were asked by the NIH to show their results separately by gender:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease</td>
<td>Non-Disease</td>
</tr>
<tr>
<td>Men</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed</td>
<td>300</td>
<td>75</td>
</tr>
<tr>
<td>Unexposed</td>
<td>225</td>
<td>150</td>
</tr>
</tbody>
</table>

prevalence ratio = (300/375)/(225/375) = 1.33
odds ratio = (300/75)/(225/150) = 2.67

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease</td>
<td>Non-Disease</td>
</tr>
<tr>
<td>Women</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed</td>
<td>150</td>
<td>225</td>
</tr>
<tr>
<td>Unexposed</td>
<td>75</td>
<td>300</td>
</tr>
</tbody>
</table>

prevalence ratio = (150/375)/(75/375) = 2.0
odds ratio = (150/225)/(75/300) = 2.67

When the investigators calculated a weighted average of the prevalence ratios for men and women, it was equal to 1.5, which was equal to the prevalence ratio in the entire population. The weighted average for the odds ratios for men and women, however, was (obviously) 2.67. The investigators were puzzled by this odds ratio of 2.67 because gender was not associated with exposure to second-hand smoke, as shown in the data table below:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exposed</td>
<td>Unexposed</td>
</tr>
<tr>
<td>Men</td>
<td>375</td>
<td>375</td>
</tr>
<tr>
<td>Women</td>
<td>375</td>
<td>375</td>
</tr>
</tbody>
</table>

Notice the measure of association in the entire population and compare it with the weighted average of the measure of associations in the sub-groups. Look at both the prevalence ratio and the odds ratio. What do we call what is happening here, i.e., what is the name of this phenomenon (in epidemiologic terminology)? (2 pts)
6. Consider the following abstract from the article titled “Why Review Articles on the Health Effects of Passive Smoking Reach Different Conclusions” published in *JAMA*.

**Objective.** To determine whether the conclusions of review articles on the health effects of passive smoking are associated with article quality, the affiliations of their authors, or other article characteristics. **Data sources.** Review articles published were identified through electronic searches of MEDLINE and EMBASE and from a database of symposium proceedings on passive smoking. **Article selection.** An article was included if its stated or implied purpose was to review the scientific evidence that passive smoking is associated with 1 or more health outcomes. Articles were excluded if they did not focus specifically on the health effects of passive smoking or if they were not written in English. **Data synthesis.** A total of 106 reviews were identified; 31 were from authors affiliated with the tobacco industry. Overall, 37% (39/106) of reviews concluded that passive smoking is not harmful to health; 74% (29/39) of these were written by authors with tobacco industry affiliations. In multiple logistic regression analyses controlling for article quality, peer review status, article topic, and year of publication, the only factor associated with concluding that passive smoking is not harmful was whether an author was affiliated with the tobacco industry (odds ratio, 88.4; 95% confidence interval, 16.4-476.5; \( P < .001 \)). **Conclusions.** The conclusions of review articles are strongly associated with the affiliations of their authors. Authors of review articles should disclose potential financial conflicts of interest, and readers of review articles should consider authors' affiliations when deciding how to judge an article's conclusions.

(a) The authors report an odds ratio that was obtained via multiple logistic regression (i.e., logistic regression adjusting for multiple exposure variables). Can you calculate the crude or unadjusted odds ratio (i.e., the odds ratio prior to adjustment for anything)? If so, describe in a single sentence your interpretation of the crude odds ratio. (1 pt)

(b) Describe whether you think the use of the odds ratio in this scenario is accurate and appropriate for all readers. (1 pt)

(c) Calculate a different unadjusted ratio measure of association that the authors could have used instead of the odds ratio. Show your calculation. (0.5 pt)

(d) Using your unadjusted ratio measure of association in part (c), describe in a clear, informative, and comprehensible sentence what can be concluded. (0.5 pt)
7. Remdesivir is the name of a drug synthesized by Gilead Sciences as part of their antiviral drug discovery program for agents with activity against RNA viruses. Despite promising in vitro activity against Ebola virus, remdesivir did not demonstrate clinical effectiveness in human studies. Nonetheless, all of the research performed with remdesivir in the Ebola epidemic helped to establish a dosage for humans and a relatively safe side effect profile. When COVID-19 appeared, it was quickly established that remdesivir had in vitro activity against SARS-CoV-2, the viral causative agent. One of the initial randomized trials of remdesivir for the treatment of COVID-19 studied adults hospitalized with COVID-19 who had evidence of lower respiratory tract involvement (Beigel et al. New England Journal of Medicine 2020 DOI: 10.1056/NEJMoa2007764). Participants were randomly assigned by the investigators to receive either remdesivir or placebo. The authors designated discharge from the hospital as their primary outcome, but in the Supplementary Appendix they provided data on all-cause mortality. One such appendix figure, shown below, examined death among a sub-group of participants who required supplemental oxygenation (albeit not of high intensity) at time of randomization. The remdesivir group is shown in the blue and placebo group in the pink.

Supplementary Appendix Figure: Death in participants who required supplemental oxygenation at time of randomization to either remdesivir or placebo

(a) Name at least two measure(s) of associations that can be readily reported using data directly available from the figure? (1 pt)

(b) What do you think is the preferred, or best, measure of association to emphasize or report to colleagues to describe the public health/clinical impact of the findings? Why? (0.5 pt)

(c) How many patients would you have to treat with remdesivir rather than placebo in order to prevent one additional death by 28 days after the start of therapy? (1 pt)

(d) What area of the figure can be interpreted as the restricted mean survival time difference? (0.5 pt)
8. Read the Schulman et al. article (NEJM 340:618, 1999) and the accompanying Sounding Board by Schwartz et al. (NEJM 314: 279, 1999) posted on the website. In answering these questions, we ask you to focus on the findings comparing black and white patients and ignore the findings regarding a possible interaction between race and gender. Interaction will be a topic at the end of this course.

(a) In table 4 in the Shulman article, describe what exactly are the numbers presented in the second column (labeled “Mean referral rate”). (1 pt)

(b) Do you think that the column heading “Mean Referral Rate” is an accurate description of the data in this column? (1 pt)

(c) In the Sounding Board, the authors suggest “three headlines that would have better characterized the findings.” The first headline they suggest reads: “Study Shows Blacks Referred 7 Percent Less Often Than Whites.” They compare this to actual headlines that did appear saying that Blacks were referred 40% less often than Whites. The actual headlines were based on an odds ratio reported by the study and featured in the press releases. What does their proposed revised headline of “7 percent less often” refer to? In other words, how do you think they arrived upon this number? (1 pt)

(d) What other measure might readers think this proposed headline (“7 percent less often”) refers to? (1 pt)

(e) FOR DISCUSSION IN SECTION ONLY:
   Can you suggest better wording (i.e., less ambiguous) for a headline stating the amount of difference found between blacks and whites in the study?
9. For the following questions, please refer to the article by Grosso et al. to be discussed in the upcoming Journal Club. (Grosso et al. *Oral bisphosphonates and risk of atrial fibrillation and flutter in women: a self-controlled case-series safety analysis.* *PLoS ONE.* 2009; 4(3):e4720)

(a) In the Introduction, the authors state that “Although evidence from randomized trials is less likely to be biased, the trials were not designed, or powered, to detect differences in the risk of AF. Case-control studies have better power…”. Do you agree, in general, with this statement about case-control studies? Explain your answer. (1 pt)

(b) Succinctly state the primary research question in this article. (1 pt)

(c) Accurate dating of the outcome is critical in getting the correct answer in this study design. What did the authors do to ensure accurate dating? (1 pt)
10. **FOR DISCUSSION IN SECTION ONLY:**
Think of a longitudinal study in your research field, or in a field in which you are interested, in which it only matters what percentage of participants get the event in question and not how fast this happens. Briefly describe the study population, exposure and outcome.
ANSWERS TO PRACTICE QUESTIONS 1a-e:

(a) Answer:
> . cs cd4_c hivload_c

<table>
<thead>
<tr>
<th>HIV Viral Load; 0 is</th>
<th>&lt;10000</th>
<th>1 is &gt;=10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>Unexposed</td>
<td>Total</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Cases</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>Noncases</td>
<td>221</td>
<td>93</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Total</td>
<td>263</td>
<td>94</td>
</tr>
<tr>
<td>Risk</td>
<td>.1596958</td>
<td>.0106383</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Point estimate</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk difference</td>
<td>.1490575</td>
</tr>
<tr>
<td>Risk ratio</td>
<td>15.01141</td>
</tr>
<tr>
<td>Attr. frac. ex.</td>
<td>.933384</td>
</tr>
<tr>
<td>Attr. frac. pop</td>
<td>.9116774</td>
</tr>
</tbody>
</table>

+-------------------------------------------------
| chi2(1) = 14.52  Pr>chi2 = 0.0001 |

(b) Answer: 0.15
(Prevalence of CD4 < 350 in exposed – prevalence of CD4 < 350 in unexposed = 0.1597-0.0106)

(c) Answer: 15.0
(Prevalence of CD4 < 350 in exposed/prevalence of CD4 < 350 in unexposed = 0.1597/0.0106)

(d) Answer: This can achieved by adding the “, or” option to your Stata command. The OR is 17.7.
> . cs cd4_c hivload_c, or

<table>
<thead>
<tr>
<th>HIV Viral Load; 0 is</th>
<th>&lt;10000</th>
<th>1 is &gt;=10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>Unexposed</td>
<td>Total</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Cases</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>Noncases</td>
<td>221</td>
<td>93</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Total</td>
<td>263</td>
<td>94</td>
</tr>
<tr>
<td>Risk</td>
<td>.1596958</td>
<td>.0106383</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Point estimate</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk difference</td>
<td>.1490575</td>
</tr>
<tr>
<td>Risk ratio</td>
<td>15.01141</td>
</tr>
<tr>
<td>Attr. frac. ex.</td>
<td>.933384</td>
</tr>
<tr>
<td>Attr. frac. pop</td>
<td>.9116774</td>
</tr>
</tbody>
</table>

| Odds ratio             | 17.67421             | 3.023509    . (Cornfield) |

+-------------------------------------------------
| chi2(1) = 14.52  Pr>chi2 = 0.0001 |

Epi Methods Problem Set 4  12  Disease Association I
(e) **Answer:** The PR $(0.16*2)/(0.01*2)$ would remain the same $(15.07)$, but the OR $(0.3/0.7)/(0.02/0.98) = 21$ would increase. **Comment:** Using the original data, one can see that because the prevalence of the outcome was not uncommon (e.g., above 0.1) in at least one of the groups that the odds ratio $(17.4)$ did not closely approximate the prevalence ratio. The odds ratio will only closely approximate the prevalence ratio when the prevalence of outcome is well below 0.1 in both the exposed and unexposed groups. This is known as the “rare disease assumption” and refers to what needs to be assumed if one wants an odds ratio to approximate a prevalence ratio. When the prevalence of outcome was doubled in this example, the odds ratio became even more disparate from the prevalence ratio. In other words, when one strays farther away from the rare disease assumption there will be greater discrepancy between the odds ratio and prevalence ratio. In cross-sectional and cohort studies, the rarity of the outcome (prevalence or incidence respectively) need not be assumed; it can be directly measured. It is only in case-control studies, in which absolute measures of disease occurrence are typically not available, that one needs to assume the rarity of the outcome.