1. A randomized, two-arm trial is conducted comparing a control treatment (A) to an experimental treatment (B). The primary outcome is all-cause mortality.

(a) After completion of the trial (between 1.5 and 3 years follow-up), we observe the following table:

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Deaths</th>
<th>Person-years follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>500</td>
<td>241</td>
</tr>
<tr>
<td>B</td>
<td>500</td>
<td>219</td>
</tr>
</tbody>
</table>

Assume exponential survival and compute the hazard ratio, the Wald test statistic for the log-hazard ratio and a 95% confidence interval for the hazard ratio. (See Section 5.8.3 of the notes).

The hazard ratio is

\[ \frac{\hat{\lambda}_B}{\hat{\lambda}_A} = \frac{219/876}{241/835} = 0.866 \]

From page 15 of the notes from lecture 8, the variance of \( \hat{\beta} = \log \frac{\hat{\lambda}_B}{\hat{\lambda}_A} \) is

\[ \text{Var}(\hat{\beta}) = \frac{1}{241} + \frac{1}{219} = 0.00875 \]

and the Wald statistic is

\[ \frac{(\log 0.866)^2}{0.00875} = 2.37 \sim \chi^2_1 \text{ under } H_0 \]

or

\[ \frac{\hat{\beta}}{\sqrt{\text{Var}(\hat{\beta})}} = -1.54 \sim N(0, 1) \text{ under } H_0 \]

(This corresponds to a two-sided p-value of 0.12, so it does not reach conventional levels of statistical significance.)

You could also calculate a Wald test on the hazard ratio scale, rather than the log-hazard ratio scale. By the delta-method

\[ \text{Var}(e^{\hat{\beta}}) = e^{2\hat{\beta}} \text{Var}(\hat{\beta}) = 0.866^2 \times 0.00875 = 0.00656 \]

and the test is

\[ \frac{(0.866 - 1)^2}{0.00656} = 2.73 \]

or

\[ \frac{0.866 - 1}{\sqrt{0.00656}} = -1.65 \]
The 95% CI for $\beta$ is $\log 0.866 \pm 1.961\sqrt{0.008715} = (-0.327, 0.039)$, and the 95% CI for $\lambda_B/\lambda_A = (0.721, 1.040)$

(b) It is noted that many subjects do not adhere to their assigned treatment and when subjects are classified by their level of adherence, we observe the following table:

<table>
<thead>
<tr>
<th>Adherence</th>
<th>Subjects</th>
<th>Deaths</th>
<th>Person-years follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>A &gt; 80%</td>
<td>110</td>
<td>52</td>
<td>180</td>
</tr>
<tr>
<td>50%-80%</td>
<td>240</td>
<td>153</td>
<td>352</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>150</td>
<td>36</td>
<td>303</td>
</tr>
<tr>
<td>B &gt; 80%</td>
<td>145</td>
<td>51</td>
<td>262</td>
</tr>
<tr>
<td>50%-80%</td>
<td>280</td>
<td>148</td>
<td>463</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>75</td>
<td>20</td>
<td>151</td>
</tr>
</tbody>
</table>

Compute hazard ratios for subjects within each stratum based on adherence (> 80%, 50%-80%, ≤ 50%), and note the differences with the overall comparison in part (a). Which result is more credible as an assessment of the effect of treatment and why?

The within-stratum hazard ratios are

- > 80%: $\frac{51}{262} / \frac{52}{180} = 0.674$
- 50%-80%: $\frac{148}{463} / \frac{153}{352} = 0.735$
- ≤ 50%: $\frac{20}{151} / \frac{36}{303} = 1.115$

The hazard ratios for the “better compliers” (> 80% and 50%-80% strata) are much smaller than the overall comparison (the intention-to-treat (ITT) analysis). It might be tempting to conclude that this analysis provides evidence of treatment benefit that is not evident in the intention-to-treat analysis, however, the within-stratum analysis is not credible for several reasons.

- The randomization ensures that the treatment assignments are independent of outcomes, and therefore, the ITT analysis is a valid test of the null hypothesis that there is no net causal effect of treatment assignment on outcomes.
- Adherence to assigned treatment is a post-randomization characteristic, and analysis conditional on post-randomization characteristics do not in general have valid causal interpretations.

Thus, the apparent benefit of B relative to A in the stratified analysis does not represent a valid causal effect of treatment.

The table below shows subjects (hypothetically) assigned to one of four unobservable strata as a function of what their adherence would have been had they been assigned either A or B. For simplicity, the 50%-80% and ≤ 50% categories have been collapsed to a single ≤ 80% category. The four strata are

I. adherence > 80% on both A and B
II. adherence > 80% were they to be assigned to A but ≤ 80% were they to be assigned to B
III. adherence ≤ 80% were they to be assigned to A but > 80% were they to be assigned to B
IV. Adherence ≤ 80% on both A and B.

Note that these are inherent subject characteristics that are present at baseline, but we cannot fully observe them. We cannot observe which of the B-adherence categories subjects assigned A are in because we cannot observe their B-adherence status, and vice versa for subjects assigned to B.

The first column of this table corresponds to the good “A-adherers” and we observe the overall result at the bottom of the column, but we cannot know how these are divided between good “B-adherers” and poor “B-adherers”. Similarly, the first row corresponds to the good “B-adherers” and we observe the overall result on the right, but we cannot know how these are divided between good “A-adherers” and poor “A-adherers”. Each entry is the color-coded sample size and observed hazard ratio for subjects Assigned A, Assigned B. We can see that the good B-adherers, composed of strata I and III which have low risk (about .2), have overall low risk, but the good A-adherers, composed of strata I (low risk) and II (high risk) has higher risk.

<table>
<thead>
<tr>
<th>Assigned B</th>
<th>Assigned A</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 80%</td>
<td>≤ 80%</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>83, (\hat{\lambda} = 26/127 = 0.205)</td>
<td>63, (\hat{\lambda} = 26/123 = 0.211)</td>
</tr>
<tr>
<td></td>
<td>81, (\hat{\lambda} = 27/134 = 0.201)</td>
<td>64, (\hat{\lambda} = 24/128 = 0.188)</td>
</tr>
<tr>
<td>II</td>
<td>27, (\hat{\lambda} = 26/53 = 0.491)</td>
<td>327, (\hat{\lambda} = 163/532 = 0.306)</td>
</tr>
<tr>
<td></td>
<td>39, (\hat{\lambda} = 22/55 = 0.400)</td>
<td>316, (\hat{\lambda} = 146/559 = 0.261)</td>
</tr>
<tr>
<td>IV</td>
<td>145, (\hat{\lambda} = 51/262 = 0.195)</td>
<td>355, (\hat{\lambda} = 168/614 = 0.274)</td>
</tr>
<tr>
<td>All</td>
<td>110, (\hat{\lambda} = 52/180 = 0.289)</td>
<td>390, (\hat{\lambda} = 189/555 = 0.289)</td>
</tr>
</tbody>
</table>

Note that within each cell, the observed event rates are similar by treatment. This is guaranteed by randomization if treatment has no effect in any stratum. The differences are simply due to sampling variability.

From a potential outcomes point of view, the potential outcomes given assignment to A for subjects in stratum II are included in the per-protocol analysis, but would not be had they been assigned to treatment B. Conversely, the potential outcomes given assignment to B for subjects in stratum III are included in the per-protocol analysis, but would not be had they been assigned to treatment A. Hence the difference in rates for the adherers does not represent a causal effect.

2. Suppose that a hypothetical trial is conducted whose primary outcome is time to a non-fatal event (there are no deaths during the study, so all subjects can continue to be followed beyond the occurrence of the event). The treatment is intended to be given during the entire study, however, some subjects discontinue their assigned treatment prior to the end of follow-up. Suppose further that we are able to follow all subjects until their planned administrative censoring time so that we can observe events occurring after treatment discontinuation. We also observe two baseline variables that are associated with outcomes.

The data file “data4.csv” (in csv format, comma delimited) contains the following variables:

- \(z\): Treatment variable (0=control, 1=experimental)
- risk1, risk2: Baseline covariates associated with risk
- time: Time to the event of interest
- event: Indicator of the event of interest (0=censored, 1=event)
- time.ot: Time to the event of interest or end of treatment
- event.ot: Indicator that the event of interest happened on treatment (0=censored, 1=event)
- eot.time: Time of treatment termination or end of follow-up
- off.trt: Indicator of treatment termination (0=censored, 1=treatment terminated)

Note that the variable time.ot is the earliest of time and eot.time and that event.ot = event unless time>eot.time.

(a) Conduct the Intention-to-treat analysis of the primary outcome:

i. Plot Kaplan-Meier survival curves,

```r
> km <- survfit(Surv(time, event)~z, data=surv)
> plot(km, lty=1:2) ## experimental treatment is dashed line
```

![Kaplan-Meier survival curves](image)

ii. Conduct log-rank test,

```r
> survdiff(Surv(time, event)~z, data=surv)
...
N Observed Expected (O-E)^2/E (O-E)^2/V
z=0 313 144 126 2.64 5.18
z=1 319 113 131 2.53 5.18
Chisq= 5.2 on 1 degrees of freedom, p= 0.0228
```
iii. Compute hazard ratio for the experimental versus the control adjusted for the baseline variable `risk1`.

```r
> coxph(Surv(time, event) ~ z + risk1, data=surv)
```

<table>
<thead>
<tr>
<th>coef</th>
<th>exp(coef)</th>
<th>se(coef)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>-0.349</td>
<td>0.705</td>
<td>0.126</td>
<td>-2.77</td>
</tr>
<tr>
<td>risk1</td>
<td>0.367</td>
<td>1.444</td>
<td>0.046</td>
<td>7.97</td>
</tr>
</tbody>
</table>

Likelihood ratio test=74.2 on 2 df, p=1.11e-16  n= 632

What do you conclude regarding the effect of:

i. treatment

The experimental treatment reduces the risk of the event by approximately 30% (the adjusted and unadjusted analyses agree closely).

ii. `risk1`

Each unit increase in `risk1` appears to increase risk of the event by about 44% (after accounting for treatment).

(b) Conduct the “on-treatment” analysis of the primary outcome (i.e., censor failure times at the time subjects discontinue treatment).

i. Plot Kaplan-Meier survival curves,

```r
> km.ot <- survfit(Surv(time.ot, event.ot) ~ z, data=surv)
> plot(km.ot, lty=1:2)  ## experimental treatment is dashed line
```
ii. Conduct log-rank test,

> survdiff(Surv(time.ot, event.ot)~z, data=surv)
... 
N Observed Expected (O-E)^2/E (O-E)^2/V
z=0 313 98 110.5 1.42 3.91
z=1 319 80 67.5 2.33 3.91
Chisq= 3.9 on 1 degrees of freedom, p= 0.0481

iii. Compute hazard ratio for the experimental versus the control.

> coxph(Surv(time.ot, event.ot)~z, data=surv)
... 
coef exp(coef) se(coef)  z  p
z 0.303  1.35  0.154  1.97  0.049
Likelihood ratio test=3.83 on 1 df, p=0.0503
n= 632, number of events= 178

iv. Compute hazard ratio for the experimental versus the control adjusted for the baseline variable risk1.

> coxph(Surv(time.ot, event.ot)~z+risk1, data=surv)
... 
coef exp(coef) se(coef)  z  p
z 0.0175  1.02  0.1572  0.111  9.1e-01
risk1 0.3545  1.43  0.0573  6.191  6.0e-10
v. Compute hazard ratio for the experimental versus the control adjusted for the baseline variable risk2.

\[
\text{> coxph(Surv(time.ot, event.ot)~z+risk2, data=surv)}
\]

\[
\begin{array}{cccccc}
\text{coef} & \text{exp(coef)} & \text{se(coef)} & \text{z} & \text{p} \\
\hline
z & -0.231 & 0.793 & 0.1563 & -1.48 & 0.14 \\
risk2 & 0.991 & 2.693 & 0.0977 & 10.14 & 0.00 \\
\end{array}
\]

Likelihood ratio test=135 on 2 df, p=0
n= 632, number of events= 178

Recall that covariate adjustment helps in two ways: reducing variability in the estimates of the parameter of interest when the covariate explains much of the variability between subjects, and to account for potential confounding between the treatment and the response. How do the results of (b)iv and (b)v help explain the differences you see in the models in part (a) and part (b)iii.

The ITT analysis in part (a) indicates that assignment to the experimental treatment reduces the risk of the event. On the other hand, the “on-treatment” analysis seems to suggest that treatment increases risk. While the covariate adjustment in (a) does not alter in any meaningful way the apparent effect of treatment, the covariate adjustment in (b) has no effect on the standard error of the coefficient, it significantly changes the estimates, actually reversing the sign of the log-HR from positive to negative. This suggests that there is confounding between the treatment, outcome and covariate. (Because treatment is randomly assigned, there cannot be confounding with the assignment itself, however, since we’re only counting events prior to treatment discontinuation, we’re implicitly considering “treatment” to be both assignment and duration and the latter is not randomly assigned and therefore subject to confounding.) Hence, the unadjusted OT analysis may be driven as much by confounding as by an effect of treatment.

In addition, adjustment for risk1 induces a smaller change in the coefficient for treatment, suggesting that it is less strongly associated with treatment and/or outcome.

(c) Consider the association between treatment discontinuation and outcomes by creating 2 × 2 tables of outcome (event=0/1) versus treatment discontinuation (off.trt=0/1) for each treatment group separately. What does this suggest regarding the assumptions underlying the models in part (b)?
> tab0 <- table(surv$event[surv$z==0], surv$off.trt[surv$z==0])
> tab0
   0 1
  0 125 44
  1  56  88
> chisq.test(tab0, correct=F)
Pearson’s Chi-squared test
data: tab0
X-squared = 39.224, df = 1, p-value = 3.779e-10
> tab1 <- table(surv$event[surv$z==1], surv$off.trt[surv$z==1])
> tab1
   0 1
  0 33 173
  1  56  57
> chisq.test(tab1, correct=F)
Pearson’s Chi-squared test
data: tab1
X-squared = 40.8033, df = 1, p-value = 1.683e-10

In both treatment groups treatment discontinuation is strongly associated with events, however, the relationship is reversed between groups. For control (z=0), the OR is $125 \times 88 / 44 \times 56 = 4.46$ whereas for experimental (z=1), the OR is $33 \times 57 / 56 \times 173 = 0.194$.

The key assumption for any time-to-event analysis is independence between censoring and the event. These tables clearly show an association between censoring (going off treatment) and events, violating the key assumption.

(d) Suppose that we had discontinued follow-up at the time of treatment discontinuation (so that we observed only `time.ot` and `event.ot` and not `time` and `event`) and that we measured baseline variable `risk1` but not `risk2`. What would we have concluded regarding the effect of treatment and why?

The unadjusted OT analysis suggests that the experimental treatment has an adverse effect on events, while the adjusted (for `risk1`) analysis suggests no difference between treatments. The variable `risk2` accounts for most of the confounding observed induced by the dependent censoring, however, we are now assuming that we do not observe `risk2` so it will be impossible to adjust for `risk2`. It would be impossible to reach the correct conclusion that the experimental treatment is superior. (A “valid” on-treatment analysis requires an assumption of “no unmeasured confounders” which could be clearly violated in this case.)