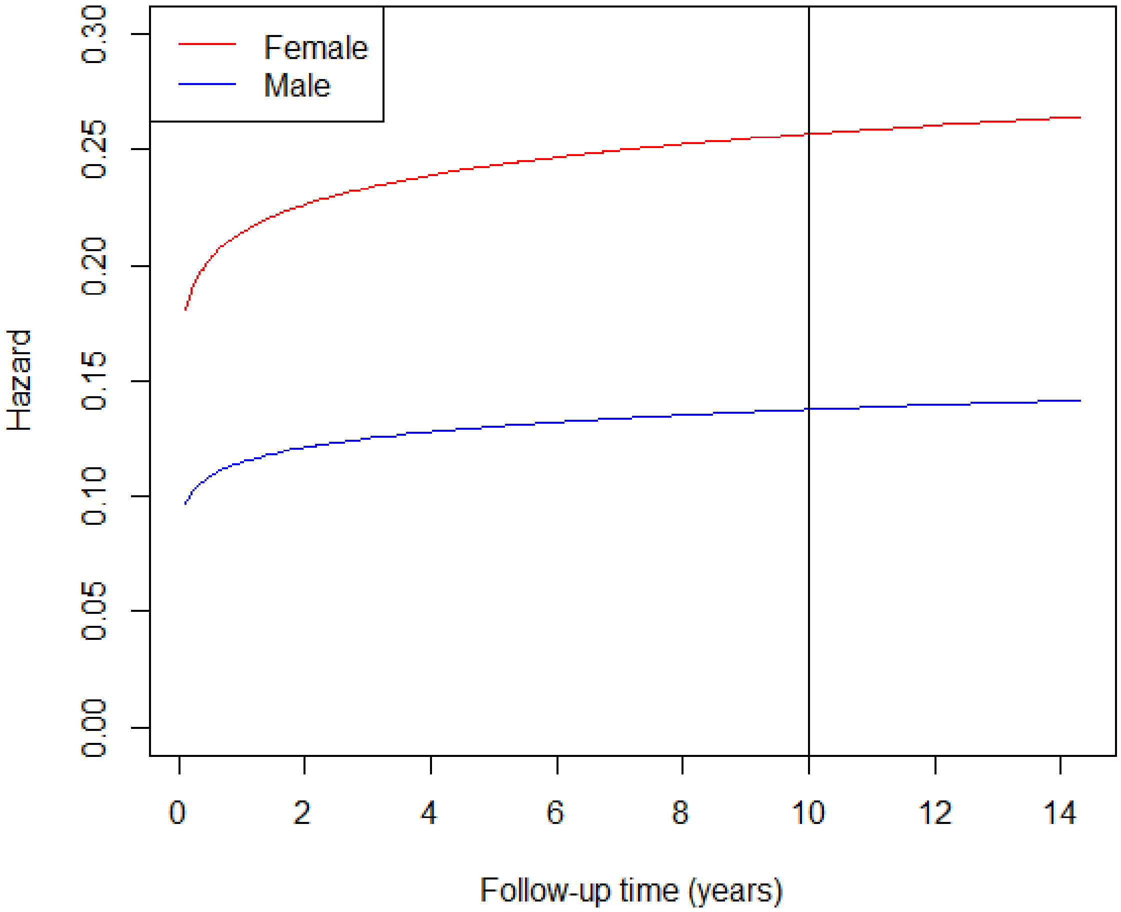
Recap: Survival Analysis

- **Survival**
- **Hazard**
- **Kaplan-Meier estimator**
- **Proportional hazards model**
- **Cox proportional hazards model**
- **Parametric proportional hazards models**

PBC Study: Weibull hazard function



HR: 1.87

Time-varying covariates & Two-stage Models

Day 4

- Endogenous vs. exogenous covariates
- Extended Cox PH for time-dependent variables
- Two-stage model for longitudinal and survival data

Breakout Session #4

1. What would be the consequence of including a time-dependent variable (e.g., transplant status) that occurs during follow-up as a time-independent fixed effect?
2. What are two reasons we cannot predict survival using an Extended Cox model? someone dead before the
3. What is the benefit of using the two-stage model versus the extended Cox model?
4. What is the assumption made by the extended Cox model and the two-stage model that is likely unrealistic for biomarkers?
5. What's your #1 productivity tip?

selection for the future information
for prediction

prediction based on the future

Review

- We looked at how to model a continuous longitudinal outcome, using a linear mixed effects model

$$y_i(t) = X_i'(t)\beta + Z_i'b_i + \epsilon_i(t), \qquad \epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$$

- We also learned how to model a time-to-event outcome, using a proportional hazards model

$$h_i(t) = h_0(t)\exp(\gamma' w_i)$$

Review

- Now, what can we do if the longitudinal and survival outcomes are related
- Leads to questions such as:
  - Survival outcome: What if the trajectory of bilirubin (i.e., how it changes over time), impacts the risk of death?
  - Longitudinal outcome: If patients with higher bilirubin values are more likely to die, will that affect our estimates of the trajectory of bilirubin over time?
- **Goal:**
  - How does serum bilirubin change over time, and are those changes associated with survival?

Time Dependent Covariates

- There is often interest in the association between a time-dependent covariate and the risk of an event

$$h(t) = h_0(t) \exp[\gamma_1 \boxed{w_1(t)}]$$

- Examples:
    - Treatment changes with time (e.g., dose)
    - Time-dependent exposure (e.g., smoking, diet)
    - Markers of disease or patient condition (e.g., blood pressure, PSA levels, serum bilirubin)

- Different from time-dependent covariate effects

$$h(t) = h_0(t) \exp[\boxed{\gamma_1(t)} w_1]$$

Time Dependent Covariates

- To answer the question of interest we need to postulate a model that relates the marker of interest with the time-to-event outcome
- The association between baseline marker levels and the risk of death can be estimated with standard statistical tools (e.g., Cox regression)
- By using only baseline observations we are throwing away a lot of useful information
- When we move to a time-dependent setting, a more careful consideration is required

Time Dependent Covariates

- There are two types of time-dependent covariates

1. **Exogenous** (external): the future path of the covariate up to any time $t > s$ is not affected by the occurrence of an event at time point s

$$\Pr(\mathcal{Y}_i(t)|\mathcal{Y}_i(s), T_i^* \geq s) = \Pr(\mathcal{Y}_i(t)|\mathcal{Y}_i(s), T_i^* = s)$$
$$\text{where } 0 < s \leq t \text{ and } \mathcal{Y}_i(t) = \{y_i(s), 0 \leq s < t\}$$

- Affects the failure process directly, but are not involved in the failure mechanism
- Variables that change in a known way
  - E.g. dose of drug
- Variables that exists totally independently of all individuals
  - E.g. air pollution

Time Dependent Covariates

- There are two types of time-dependent covariates

2. **Endogenous** (internal)
   - Affects the failure process, but can also be impacted by the failure mechanism

- Variables that relate to the individual and can only be measured when the individual is alive

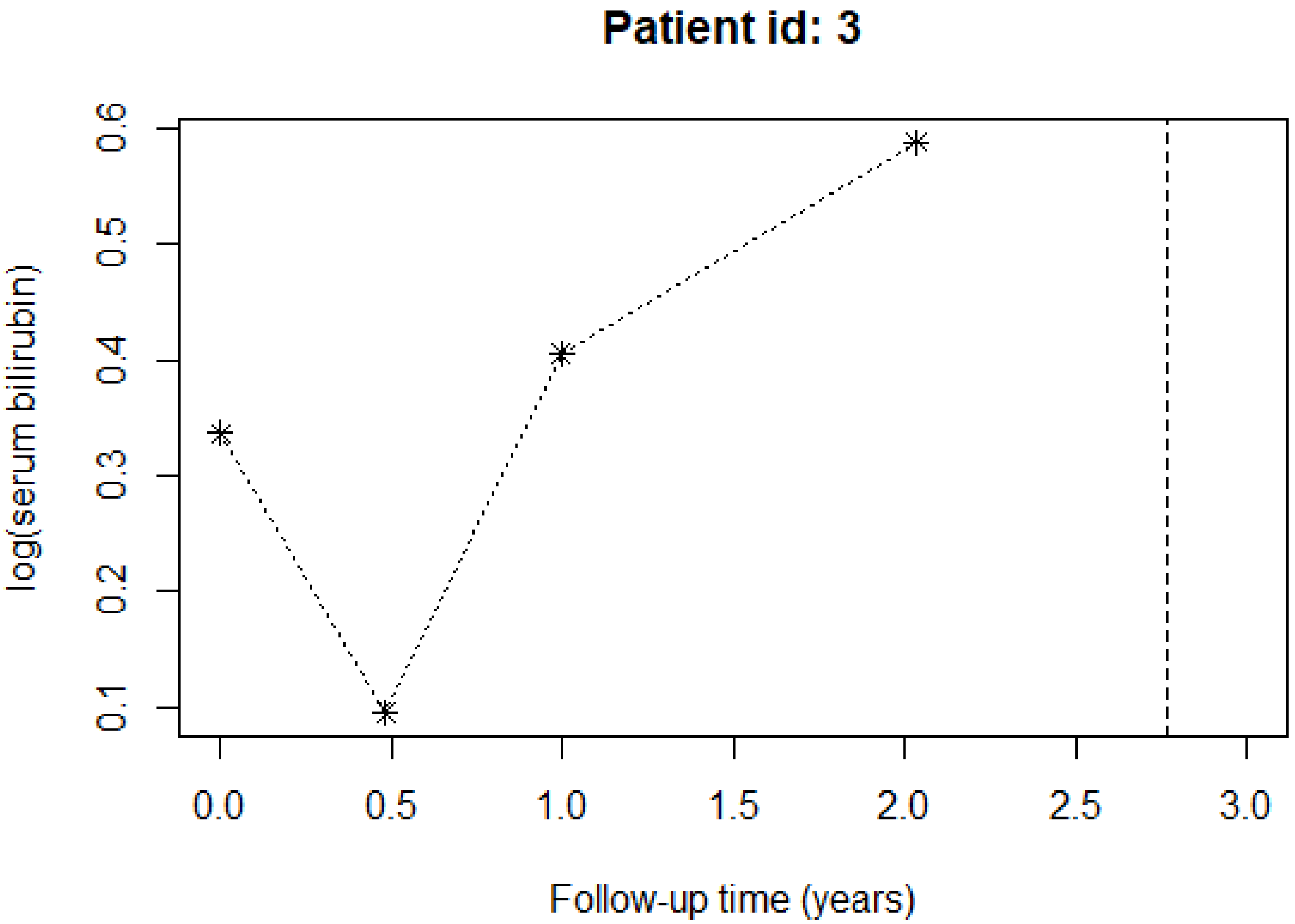
Endogenous Covariates

- Time-dependent measurements taken on the subject
- It is important to distinguish between these two types of time-dependent covariates because the type of covariate dictates the appropriate analysis
- Biomarkers are endogenous covariates
  - Their existence is directly related to failure status
  - Measured with error (i.e., biological variation)
  - The complete history is not available (observed at measurement times)

Endogenous vs. Exogenous covariates

- Time of day
- Blood pressure
- Age of the individual
- Weight of the individual

Time Dependent Covariates



Extended Cox Model

- The Cox model presented earlier can be extended to handle time-dependent covariates

$$h_i(t|\mathcal{Y}_i(t), w_i) = h_0(t) \exp\{\gamma' w_i + \alpha y_i(t)\}$$

- $y_i(t)$ denotes the observed value of the time-varying covariate at t
- $\exp(\alpha)$: relative increase in the risk of an event at time t that results from one unit increase in $y_i(t)$ at the same time point
- Hazard ratio is not constant in time (no longer making PH assumption)

Extended Cox Model

- Using the counting process formulation

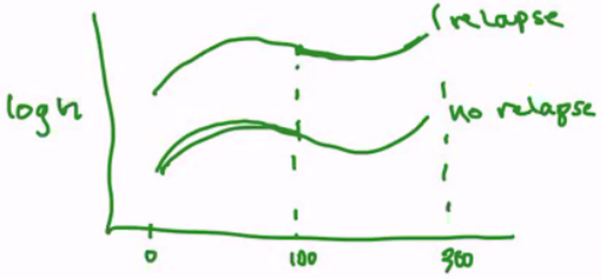
$$h_i(t|\mathcal{Y}_i(t), w_i) = h_0(t)R_i(t)\exp\{\gamma' w_i + \alpha y_i(t)\}$$

  - $N_i(t)$ is a counting process which counts the number of events for subject i by time t
  - $h_i(t)$ denotes the intensity process for $N_i(t)$
  - $R_i(t)$ denotes the at-risk process ('1' if subject i is still at risk at time t)

- Parameters are estimated based on the log-partial likelihood

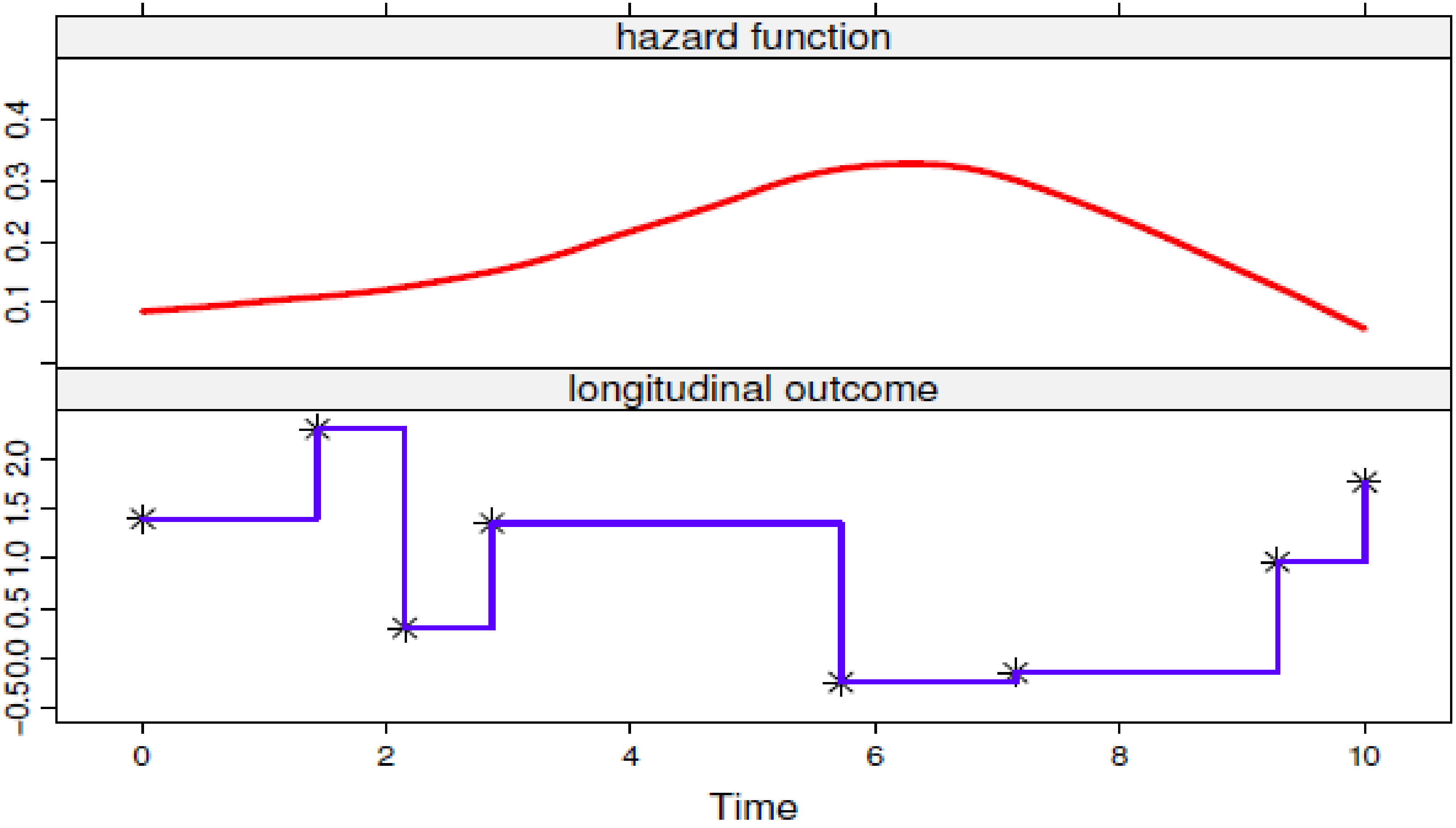
$$Y_i(t) = \begin{cases} 1 \text{ relapse} \\ 0 \text{ no "} \end{cases}$$

T_i	δ_i	Y_i
		0
(0 , 100]	0	0
(100 , 300)	1	1



Extended Cox Model

lag effects or
irreversible damage

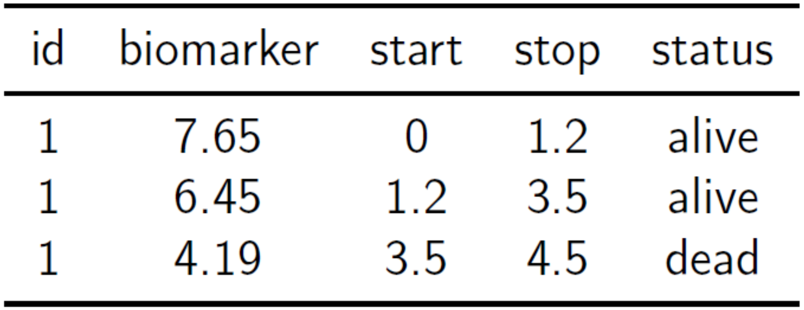


Example: PBC Study

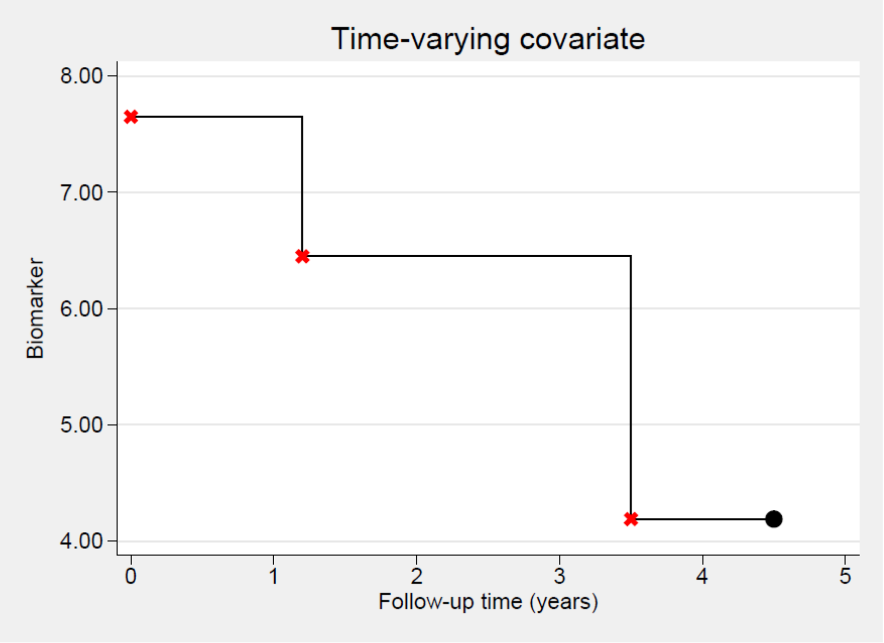
- We must first organize our data in a start/stop format
- Consider a hypothetical patient that had measurements taken at baseline, 1.2, and 3.5 years and died at 4.5 years.

we cannot use the future time to describe the situation for now.

id	year	biomarker	years	status
1	0	7.65	4.5	dead
1	1.2	6.45	4.5	dead
1	3.5	4.19	4.5	dead



id	biomarker	start	stop	status
1	7.65	0	1.2	alive
1	6.45	1.2	3.5	alive
1	4.19	3.5	4.5	dead

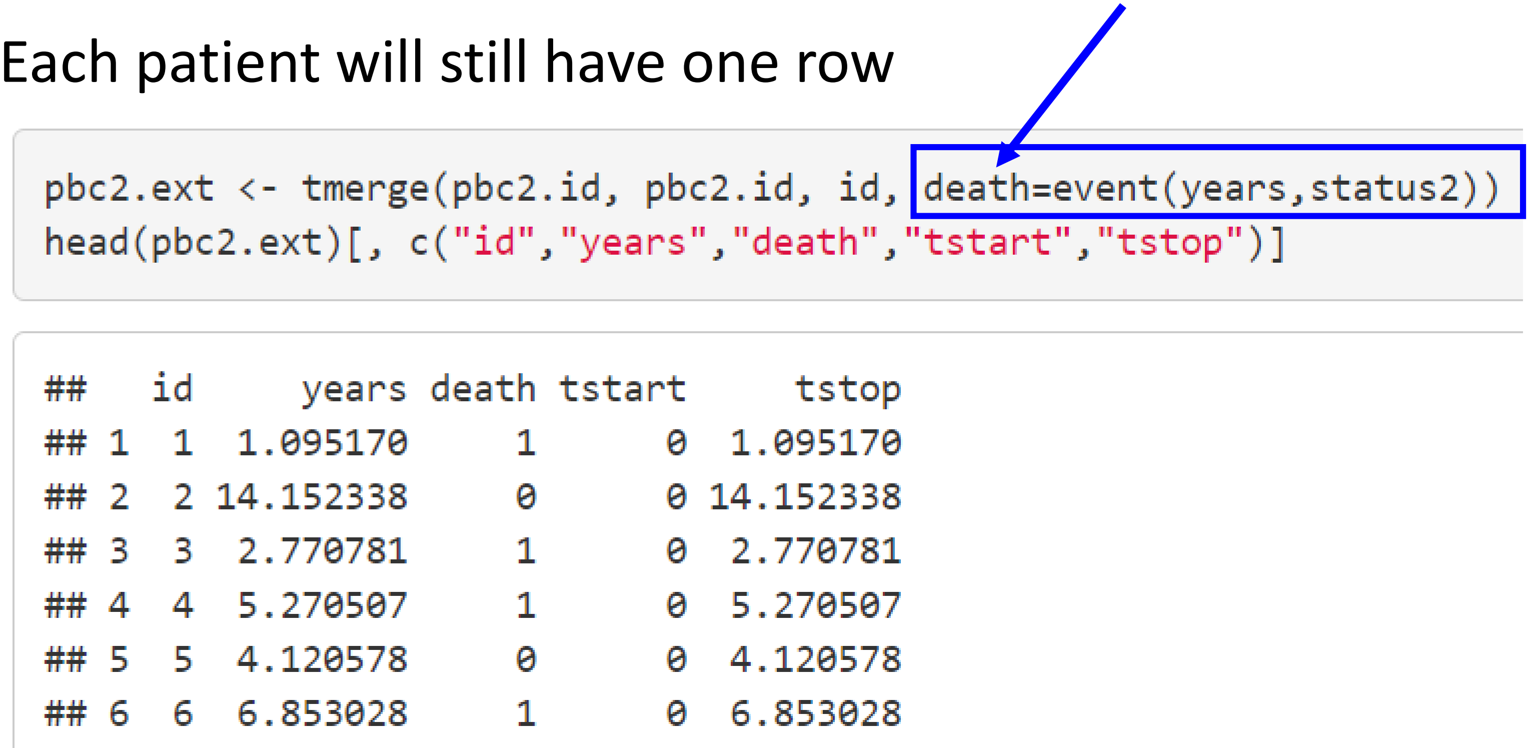


Example: PBC Study

- Use “tmerge” function from “survival” package

1. Apply to a unique data set with one row per subject

- death=event([observed follow-up time], [event indicator])
- Creates new columns: “tstart”, “tstop”, “death”
- Each patient will still have one row



```
pbc2.ext <- tmerge(pbc2.id, pbc2.id, id, death=event(years,status2))
head(pbc2.ext)[, c("id","years","death","tstart","tstop")]
```

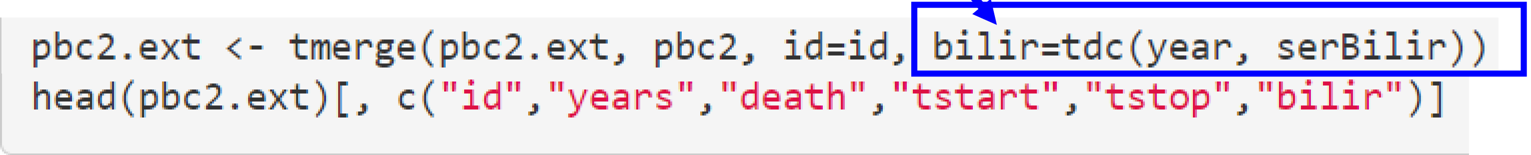
```
##   id     years death tstart     tstop
## 1  1  1.095170     1      0  1.095170
## 2  2 14.152338     0      0 14.152338
## 3  3  2.770781     1      0  2.770781
## 4  4  5.270507     1      0  5.270507
## 5  5  4.120578     0      0  4.120578
## 6  6  6.853028     1      0  6.853028
```

Example: PBC Study

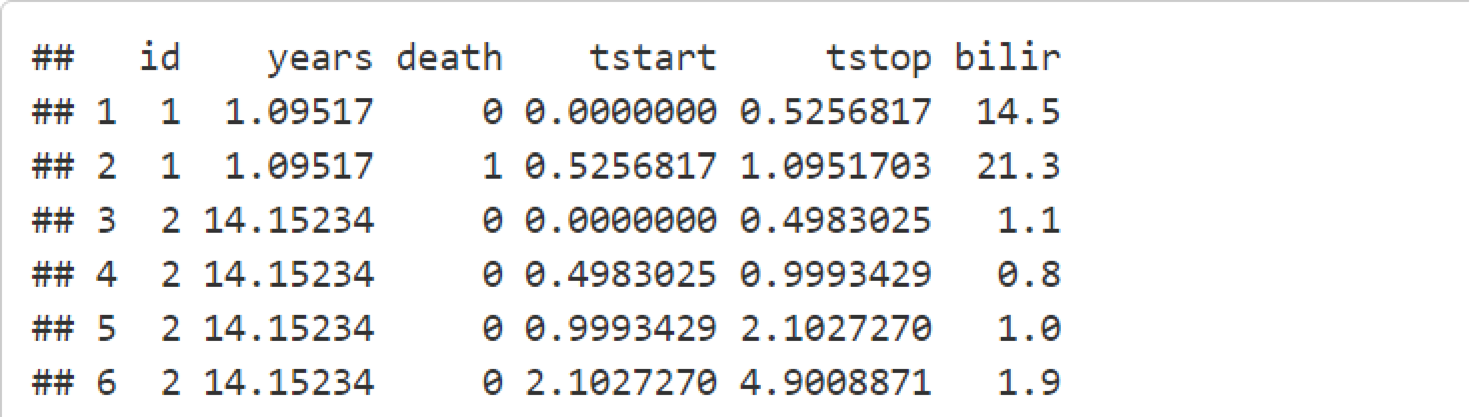
for the cox model each row is an individual

2. Apply “tmerge” to the new data set (on the previous slide) and your longitudinal data set (one row for each measurement)

- bilir=tdc([measurement time], [biomarker value])
- Creates new column: “bilir”
- Each patient will have multiple rows



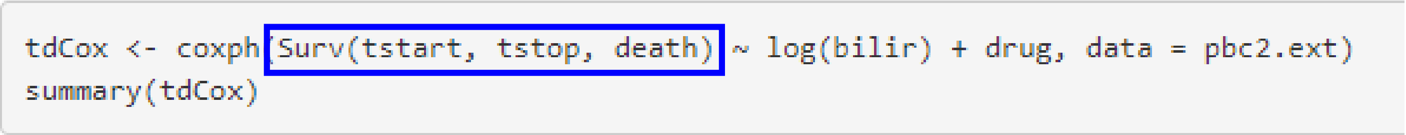
```
pbc2.ext <- tmerge(pbc2.ext, pbc2, id=id, bilir=tdc(year, serBilir))
head(pbc2.ext)[, c("id","years","death","tstart","tstop","bilir")]
```



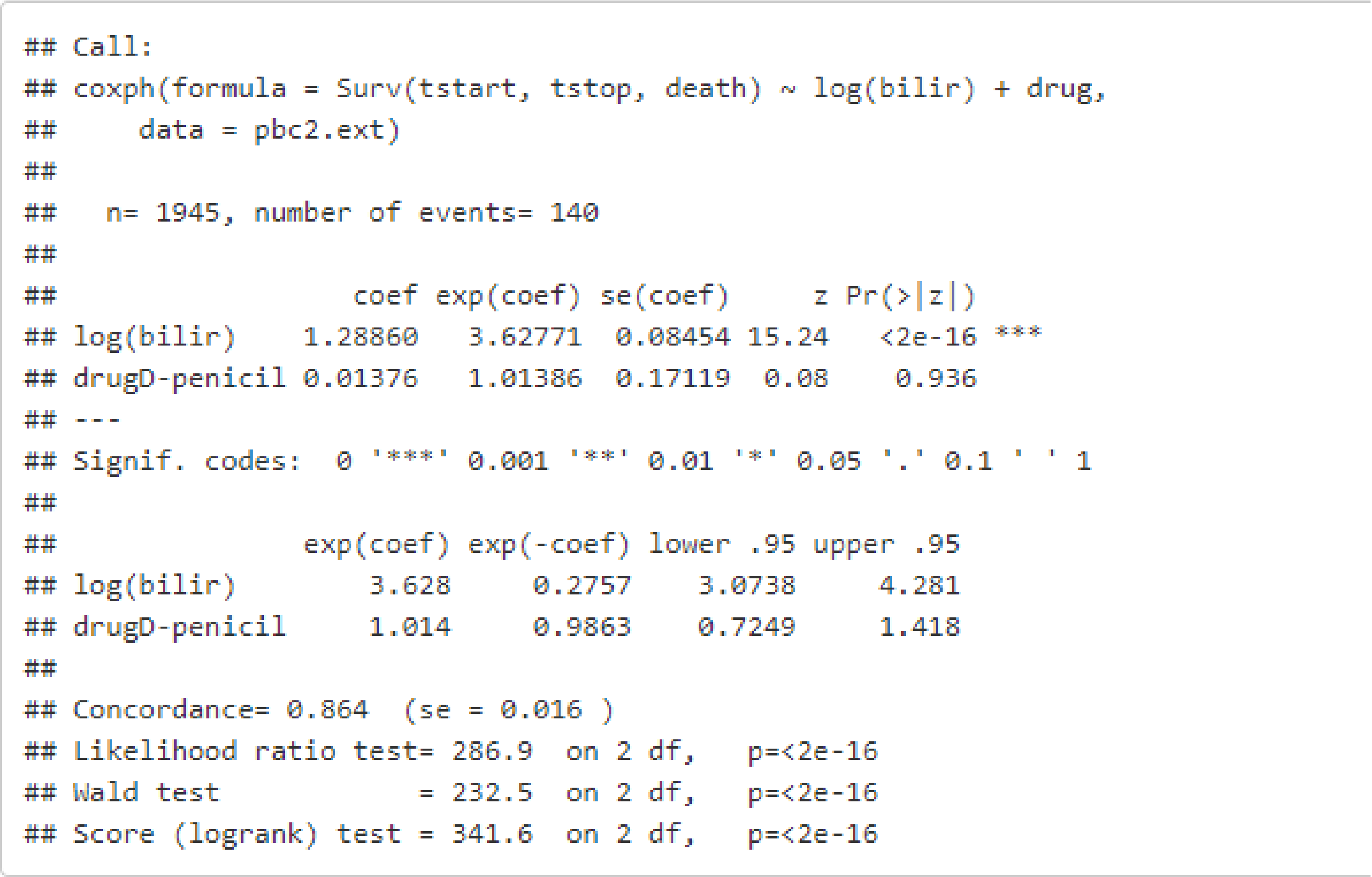
```
##    id    years death    tstart     tstop bilir
## 1   1  1.09517     0 0.0000000 0.5256817  14.5
## 2   1  1.09517     1 0.5256817 1.0951703  21.3
## 3   2 14.15234     0 0.0000000 0.4983025   1.1
## 4   2 14.15234     0 0.4983025 0.9993429   0.8
## 5   2 14.15234     0 0.9993429 2.1027270   1.0
## 6   2 14.15234     0 2.1027270 4.9008871   1.9
```

Example: PBC Study

$$h(t) = h_0(t)\exp\{\gamma_1 \log y_i(t) + \gamma_2 \text{D-penecillin}_i\}$$

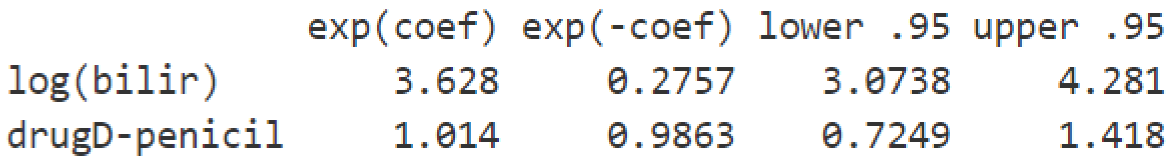


```
tdCox <- coxph(Surv(tstart, tstop, death) ~ log(bilir) + drug, data = pbc2.ext)
summary(tdCox)
```



```
## Call:
## coxph(formula = Surv(tstart, tstop, death) ~ log(bilir) + drug,
##     data = pbc2.ext)
##
##   n= 1945, number of events= 140
##
##                   coef exp(coef) se(coef)     z Pr(>|z|)
## log(bilir)     1.28860   3.62771  0.08454 15.24   <2e-16 ***
## drugD-penicil 0.01376   1.01386  0.17119  0.08    0.936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## log(bilir)         3.628     0.2757    3.0738     4.281
## drugD-penicil      1.014     0.9863    0.7249     1.418
##
## Concordance= 0.864  (se = 0.016 )
## Likelihood ratio test= 286.9  on 2 df,   p=<2e-16
## Wald test            = 232.5  on 2 df,   p=<2e-16
## Score (logrank) test = 341.6  on 2 df,   p=<2e-16
```

Example: PBC Study

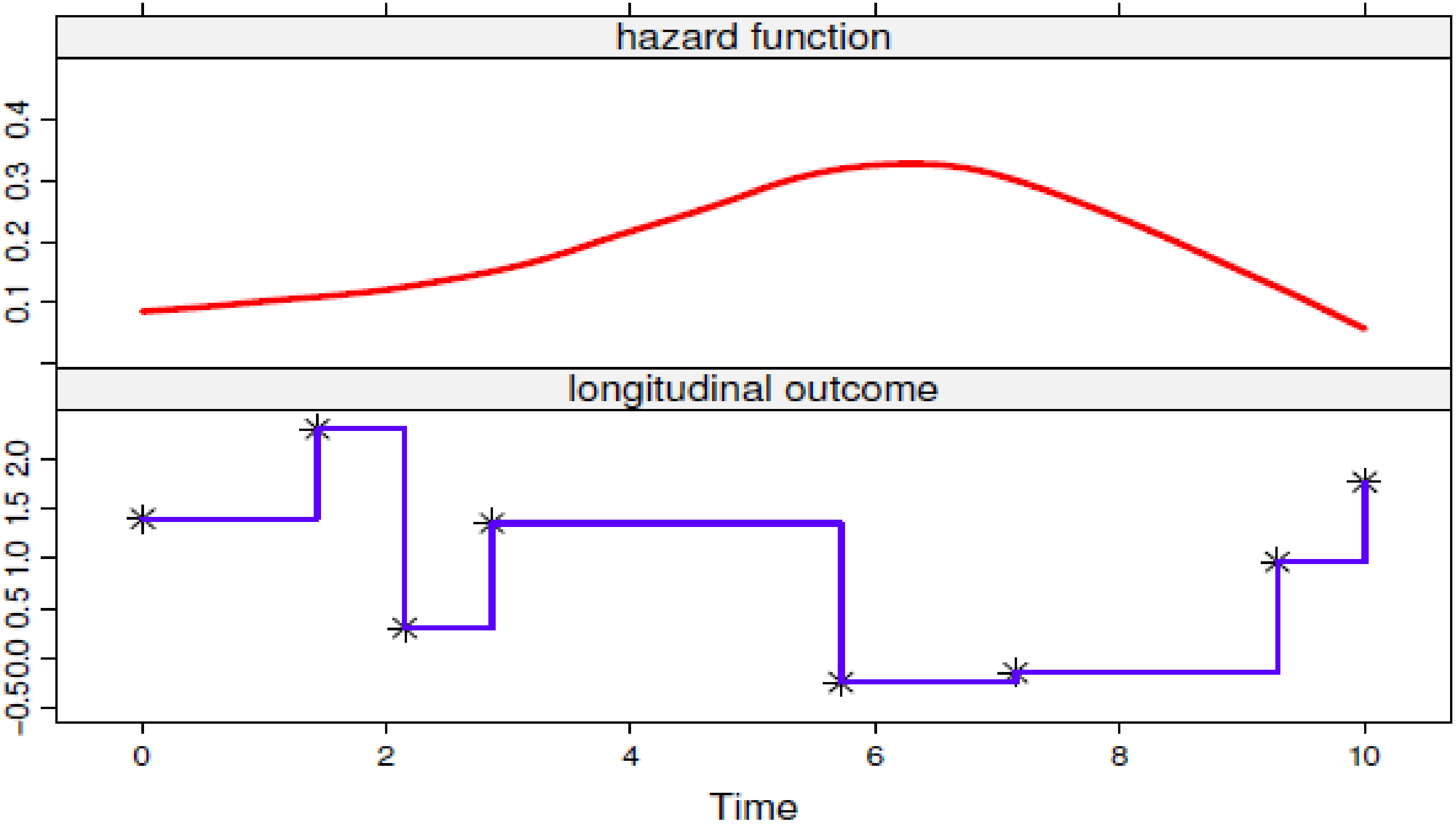


```
              exp(coef) exp(-coef) lower .95 upper .95
log(bilir)        3.628     0.2757    3.0738     4.281
drugD-penicil     1.014     0.9863    0.7249     1.418
```

the bilirubin level at current time point

- After adjusting for treatment, bilirubin is strongly associated with the risk of death, with one unit increase in log-bilirubin resulting in a 3.62-fold increase in the risk of death (95% CI: 3.07-4.28; p<0.001).

Extended Cox Model



Extended Cox Model

- How does the extended Cox model handle time-varying covariates?
  - Step-function path
  - Assumes no measurement error
  - Existence of the covariate is not related to failure status
- Thus, the extended Cox model is only valid for exogenous time-dependent covariates
- Treating endogenous covariates as exogenous may produce spurious results!

Two-stage models

- The extended Cox model assumes that the time-varying covariate is error-free
- Previously, we have modeled the biomarker using a linear mixed effects model to account for measurement error
- Instead of using the observed biomarker values, we can use subject-specific predictions of the true, unobserved biomarker values instead

Two-stage models

- **Stage 1**: Fit a Linear Mixed Effects Model and obtain subject-specific predictions of the true marker value, $\widehat{m}_i(t)$

$$y_i(t) = m_i(t) + \epsilon_i(t), \qquad \epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$$

where

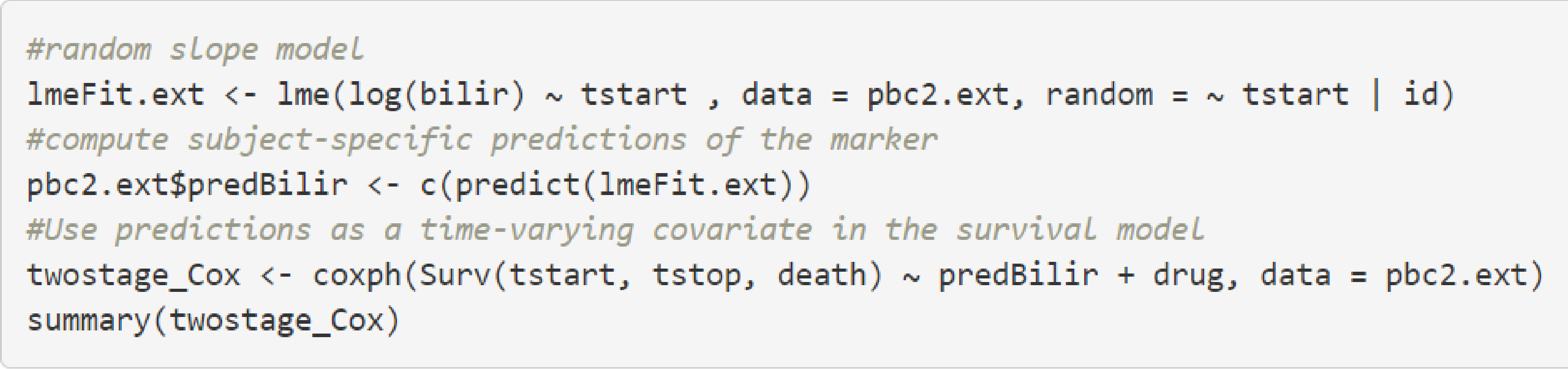
$$m_i(t) = X_i'(t)\beta + Z_i'(t)b_i \qquad b_i \sim N(0, D), b_i \perp\!\!\!\perp \epsilon_i$$

- **Stage 2:** Use the subject-specific predictions, $\widehat{m}_i(t)$, as our time-varying covariate in the Extended Cox model

$$h_i(t) = h_0(t)\exp[\gamma' w_i + \alpha \hat{m}_i(t)]$$

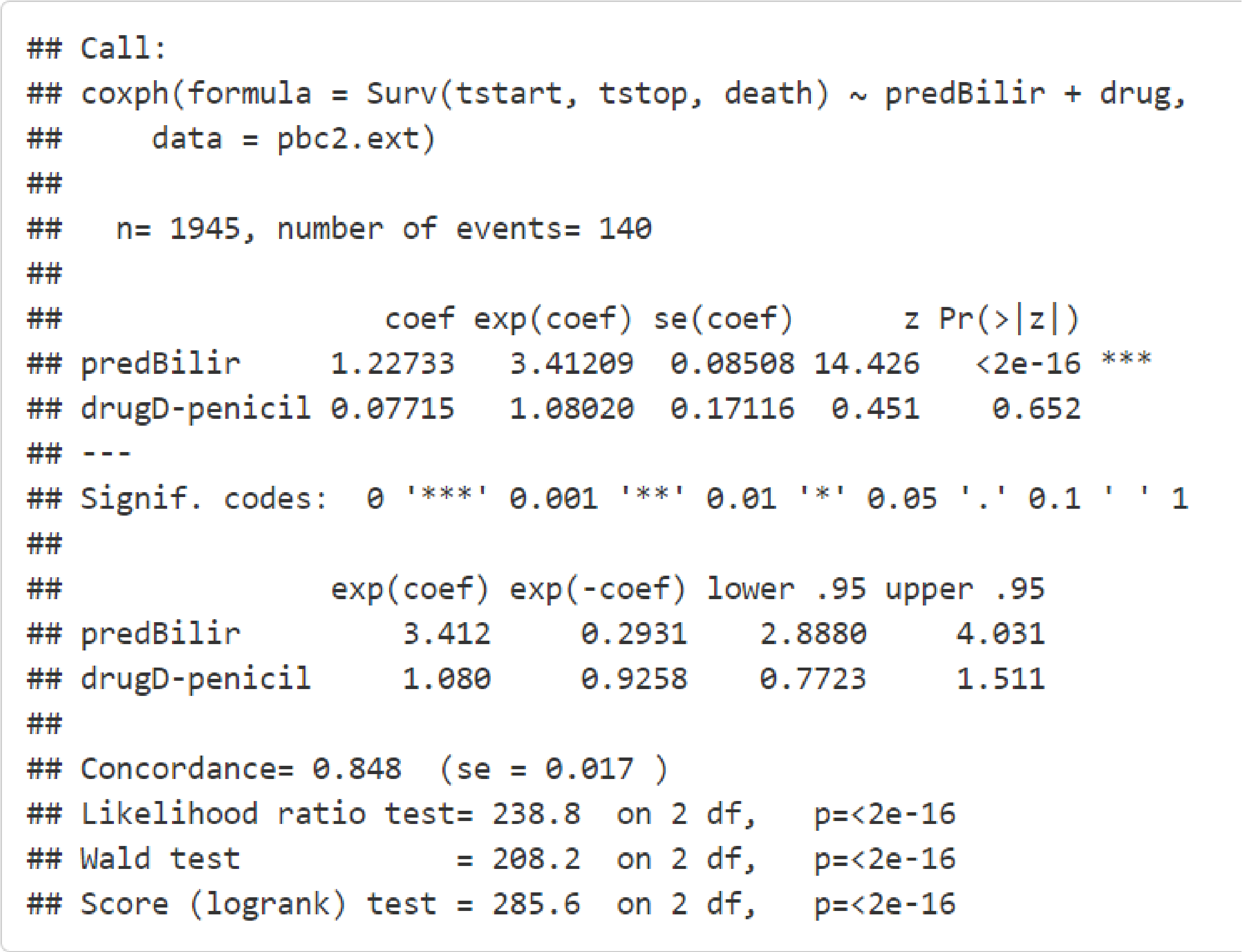
Example: PBC data

- We can fit this model using the start/stop data set that we previously created



```
#random slope model
lmeFit.ext <- lme(log(bilir) ~ tstart , data = pbc2.ext, random = ~ tstart | id)
#compute subject-specific predictions of the marker
pbc2.ext$predBilir <- c(predict(lmeFit.ext))
#Use predictions as a time-varying covariate in the survival model
twostage_Cox <- coxph(Surv(tstart, tstop, death) ~ predBilir + drug, data = pbc2.ext)
summary(twostage_Cox)
```

Example: PBC data



```
## Call:
## coxph(formula = Surv(tstart, tstop, death) ~ predBilir + drug,
##     data = pbc2.ext)
##
##   n= 1945, number of events= 140
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## predBilir      1.22733   3.41209  0.08508 14.426   <2e-16 ***
## drugD-penicil  0.07715   1.08020  0.17116  0.451    0.652
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## predBilir         3.412     0.2931    2.8880     4.031
## drugD-penicil     1.080     0.9258    0.7723     1.511
##
## Concordance= 0.848  (se = 0.017 )
## Likelihood ratio test= 238.8  on 2 df,   p=<2e-16
## Wald test            = 208.2  on 2 df,   p=<2e-16
## Score (logrank) test = 285.6  on 2 df,   p=<2e-16
```

Example: PBC data

- Comparing log hazard ratios for serum bilirubin

Model	Log HR	SE	95% CI
Time-varying covariate	1.29	0.0845	(1.12, 1.45)
Two-stage	1.23	0.0851	(1.06, 1.39)

standard error will be smaller than the real
it is part of assignment

- We are accounting for measurement error in the two-stage model
- Estimates differ (can be substantial)

Two-stage models

- There are still issues with the two-stage approach
  - The uncertainty in our estimates from the first stage are not carried through to the second stage
  - Thus, our estimates of association are too precise
  - We are still assuming that the values do not change between observation times
- However,
  - It has been shown to greatly reduce bias compared to the time-varying covariate approach
  - It allows us to fit complex models very quickly
  - Can handle multiple time-dependent covariates
- Next step: Joint models!

Breakout Session #4

1. What would be the consequence of including a time-dependent variable (e.g., transplant status) that occurs during follow-up as a time-independent fixed effect?
2. What are two reasons we cannot predict survival using an Extended Cox model?
3. What is the benefit of using the two-stage model versus the extended Cox model?
4. What is the assumption made by the extended Cox model and the two-stage model that is likely unrealistic for biomarkers?
5. What's your #1 productivity tip?

Immortal Time Bias

- Why should we not use future values of the time-dependent covariate as fixed baseline covariates?
- For example, our time-dependent variable is a binary indicator of transplant
- Patients that died early would not have the chance to have a transplant
- Patients that received a transplant have to live long enough to have a transplant (immortal until they receive the transplant)
- The two groups being compared are selectively biased favouring transplant patients

Prediction and Time-Dependent Covariates

- Why can we not predict survival with an Extended Cox model?

1. The model depends on the value of a changing quantity, for which we do not know the future values

- Recall our survival function

$$S(t) = \exp\{-\int_0^t h(s)ds\} = \exp\{-\int_0^t h_0(t)\exp(\gamma' w + \alpha y(s))ds\}$$

- To compute future survival we need to integrate over future values of the time-dependent covariate

2. If we did know the future covariate value, its existence would imply that the subject is still alive