UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

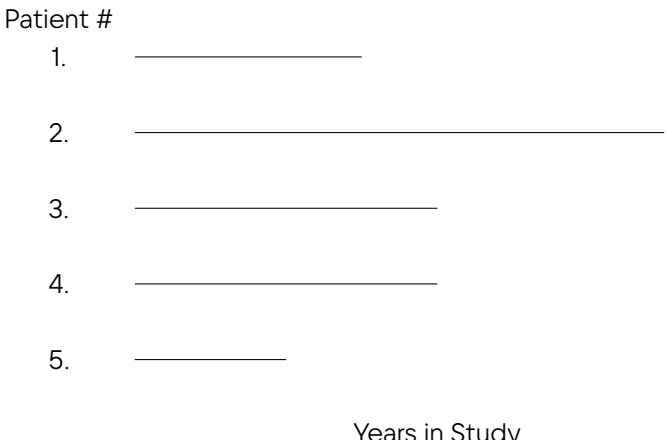
# An Introduction to Censored Covariates

Marissa Ashner

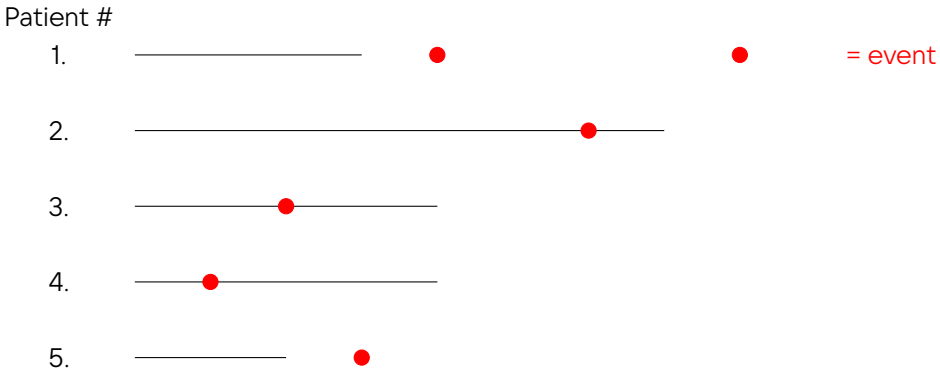
STA779: *Applied Survival Analysis (Spring 2023)*

February 28, 2023

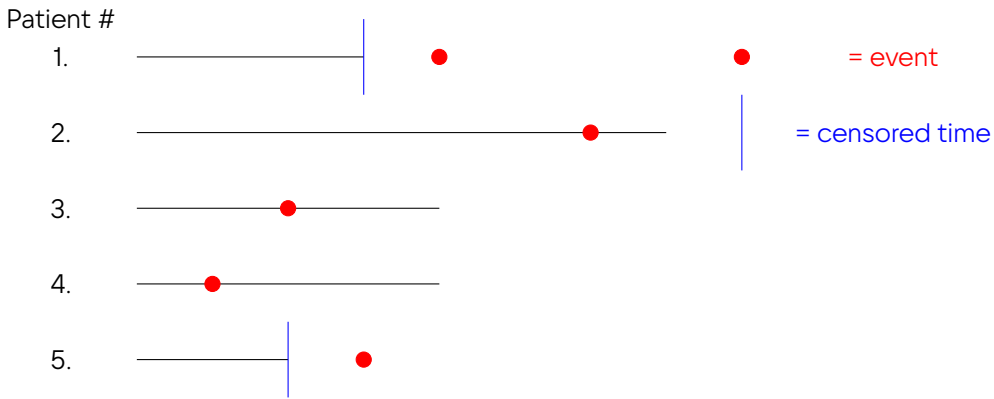
# REVIEW: Censoring



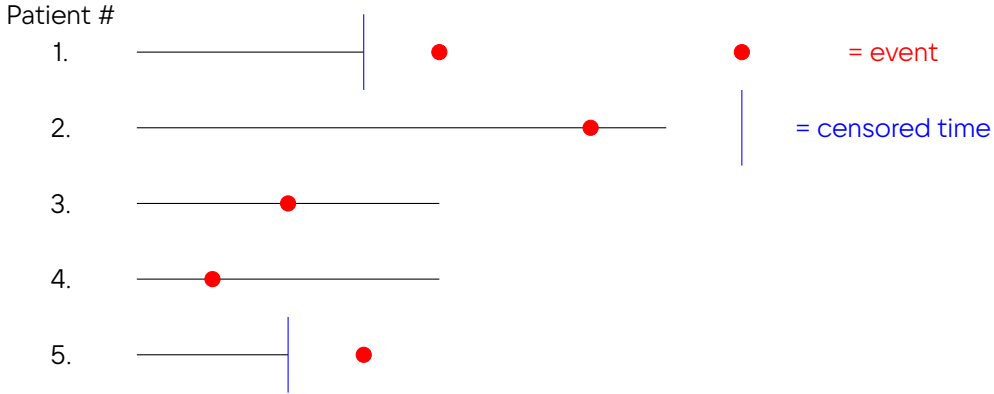
# REVIEW: Censoring



# REVIEW: Censoring

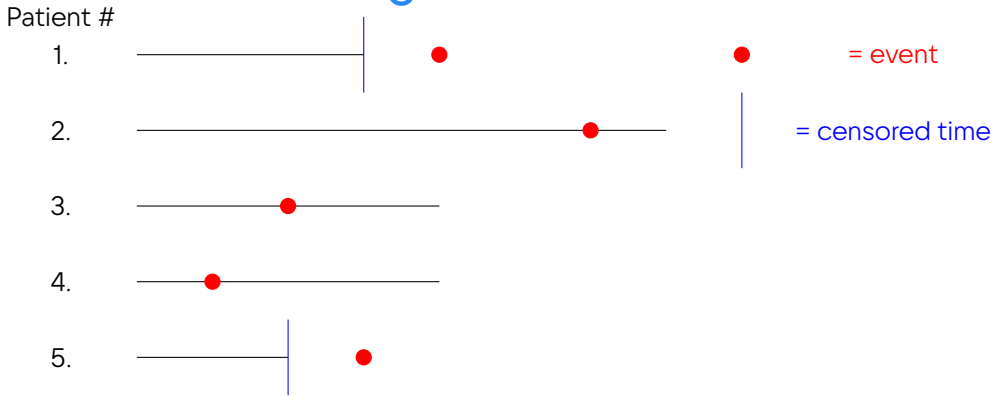


# REVIEW: Censoring




**What kind of censoring is this?:** *Right?, Left?, Interval?*

# REVIEW: Censoring




**What kind of Type I censoring could this be?:** *Fixed? Progressive? Generalized?*

# Censored Outcomes

- What you will *primarily* learn about in this class (except today)

# Censored Outcomes

- What you will *primarily* learn about in this class (except today)
- The whole idea around survival analysis, where the *outcome* you are interested in is some 'time until event' (or some function of that time)

# Censored Outcomes

- What you will *primarily* learn about in this class (except today)
- The whole idea around survival analysis, where the *outcome* you are interested in is some 'time until event' (or some function of that time)
- It is often the case that you will adjust for some other factors, or covariates, that could affect this outcome

# Censored Outcomes

- What you will *primarily* learn about in this class (except today)
- The whole idea around survival analysis, where the *outcome* you are interested in is some 'time until event' (or some function of that time)
- It is often the case that you will adjust for some other factors, or covariates, that could affect this outcome
- or maybe you want to see which factors determine survival of this event

# How do you determine where to put variables in a model?

- What if I want to measure the linear association between height and weight?

# How do you determine where to put variables in a model?

- What if I want to measure the linear association between height and weight?
- Is this my model?

$$H = \beta_0 + \beta_1 W + \epsilon$$

# How do you determine where to put variables in a model?

- What if I want to measure the linear association between height and weight?
- Is this my model?

$$H = \beta_0 + \beta_1 W + \epsilon$$

- Or is this my model?

$$W = \beta_0 + \beta_1 H + \epsilon$$

# How do you determine where to put variables in a model?

- What if I want to measure the linear association between height and weight?
- Is this my model?

$$H = \beta_0 + \beta_1 W + \epsilon$$

- Or is this my model?

$$W = \beta_0 + \beta_1 H + \epsilon$$

- What if I want to predict height based on someone's weight?

# How do you determine where to put variables in a model?

- What if I want to measure the linear association between height and weight?
- This depends heavily on your **research question and analysis goal**

# Censored Covariates

- What if I want to measure the progression of Huntington disease impairment as a function of a patient's diagnosis time, among other factors?

# Censored Covariates

- What if I want to measure the progression of Huntington disease impairment as a function of a patient's diagnosis time, among other factors?
- The patient's diagnosis time here, is a covariate, based on our research question

# Censored Covariates

- What if I want to measure the progression of Huntington disease impairment as a function of a patient's diagnosis time, among other factors?
- The patient's diagnosis time here, is a covariate, based on our research question
- AND it is right-censored, since not all patients will be diagnosed during a Huntington disease observational study

# Censored Covariates

- What if I want to measure the progression of Huntington disease impairment as a function of a patient's diagnosis time, among other factors?
- The patient's diagnosis time here, is a covariate, based on our research question
- AND it is right-censored, since not all patients will be diagnosed during a Huntington disease observational study
- Another example: What if I want to measure the impact of income on college matriculation but income is top-coded, meaning anyone with an income over a certain number, let's say $\$250K$, is simply labeled as $\geq \$250K$.

# Types of censored covariates

- **Limit of Detection Censoring**
  - Also called top-coded or bottom-coded, like the income example
  - A lot of biological markers often have a lower limit of detection, meaning the device to measure the factor cannot pick up levels below a certain threshold
  - Often, observations have the same limit of detection, but they can also have different limits
  - These limits are known a priori, and are considered constants
  - Similar to Type I censoring that you discussed last week

# Types of censored covariates

- **Limit of Detection Censoring**
  - Also called top-coded or bottom-coded, like the income example
  - A lot of biological markers often have a lower limit of detection, meaning the device to measure the factor cannot pick up levels below a certain threshold
  - Often, observations have the same limit of detection, but they can also have different limits
  - These limits are known a priori, and are considered constants
  - Similar to Type I censoring that you discussed last week
- **Random Censoring**
  - Like the Huntington disease example
  - Censoring values differ for each observation, and we don't consider these values known a priori
  - Instead, we consider the set of these values as a random variable, with it's own distribution

# Some Notation

- Let $X$ denote the censored covariate of interest
- Let $C$ denote the censoring value (considering random censoring)
- We observe $W = \min(X, C)$ and $\Delta = I(X \leq C)$
- Let $Y$ be the outcome of interest and $Z$ be the set of fully-observed covariates

# Missing Data vs. Censored Data

- Censored Data $\subset$ Missing Data

# Missingness and Censoring Mechanisms

**Why don't we know $X$?**: *This is the most important question to answer when considering missing data techniques*

# Missingness and Censoring Mechanisms

**Why don't we know $X$?:** *This is the most important question to answer when considering missing data techniques*

## Missing Data

- **Missing at Random (MAR):** The probability that $X$ is missing depends on **OBSERVED** variables (e.g., $Y$, $Z$)
-
-

# Missingness and Censoring Mechanisms

**Why don't we know $X$?:** *This is the most important question to answer when considering missing data techniques*

## Missing Data

- **Missing at Random (MAR):** The probability that $X$ is missing depends on **OBSERVED** variables (e.g., $Y$, $Z$)
- **Missing NOT at Random (MNAR):** The probability that $X$ is missing depends on **UNOBSERVED** variables (e.g., $X$ itself)
-

# Missingness and Censoring Mechanisms

**Why don't we know $X$?:** *This is the most important question to answer when considering missing data techniques*

## Missing Data

- **Missing at Random (MAR):** The probability that $X$ is missing depends on **OBSERVED** variables (e.g., $Y$, $Z$)
- **Missing NOT at Random (MNAR):** The probability that $X$ is missing depends on **UNOBSERVED** variables (e.g., $X$ itself)
- These missingness mechanisms rely on the censoring indicator $\Delta$

# Missingness and Censoring Mechanisms

**Why don't we know $X$?:** *This is the most important question to answer when considering missing data techniques*

## Missing Data

- **Missing at Random (MAR):** The probability that $X$ is missing depends on **OBSERVED** variables (e.g., $Y$, $Z$)
- **Missing NOT at Random (MNAR):** The probability that $X$ is missing depends on **UNOBSERVED** variables (e.g., $X$ itself)
- These missingness mechanisms rely on the censoring indicator $\Delta$

## Censored Data

- The **PROBABILITY** that $X$ is censored is **INHERENTLY** dependent on the value of $X$, by definition –– remember $\Delta = 1$ if $X \leq C$
-
-

# Missingness and Censoring Mechanisms

**Why don't we know $X$?:** *This is the most important question to answer when considering missing data techniques*

## Missing Data

- **Missing at Random (MAR):** The probability that $X$ is missing depends on **OBSERVED** variables (e.g., $Y$, $Z$)
- **Missing NOT at Random (MNAR):** The probability that $X$ is missing depends on **UNOBSERVED** variables (e.g., $X$ itself)
- These missingness mechanisms rely on the censoring indicator $\Delta$

## Censored Data

- The **PROBABILITY** that $X$ is censored is **INHERENTLY** dependent on the value of $X$, by definition –– remember $\Delta = 1$ if $X \leq C$
- Always **MNAR**
-

# Missingness and Censoring Mechanisms

**Why don't we know $X$?:** *This is the most important question to answer when considering missing data techniques*

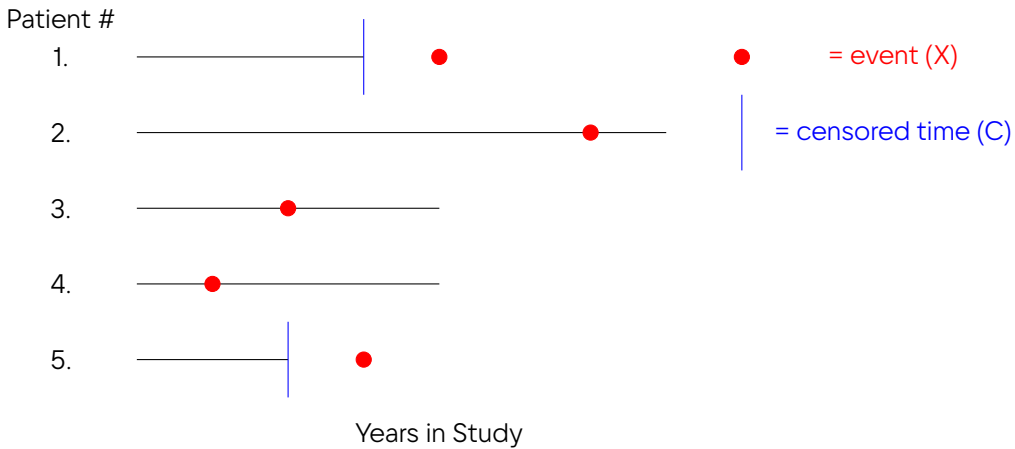
## Missing Data

- **Missing at Random (MAR):** The probability that $X$ is missing depends on **OBSERVED** variables (e.g., $Y$, $Z$)
- **Missing NOT at Random (MNAR):** The probability that $X$ is missing depends on **UNOBSERVED** variables (e.g., $X$ itself)
- These missingness mechanisms rely on the censoring indicator $\Delta$

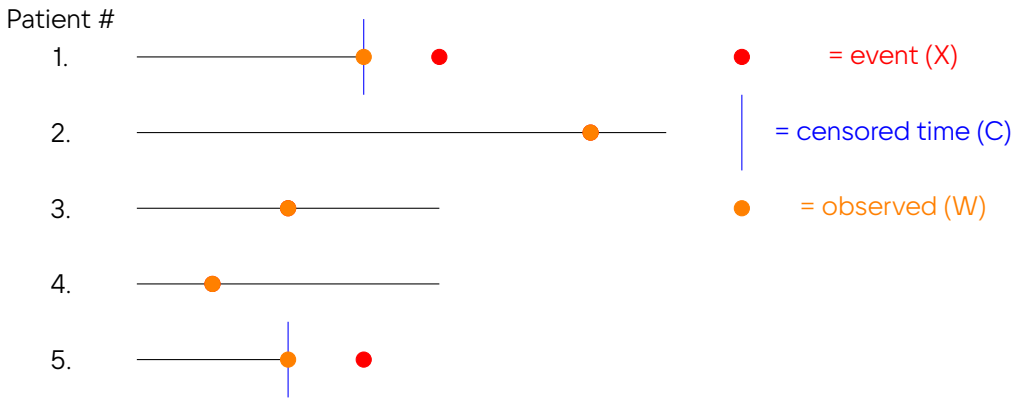
## Censored Data

- The **PROBABILITY** that $X$ is censored is **INHERENTLY** dependent on the value of $X$, by definition –– remember $\Delta = 1$ if $X \leq C$
- Always **MNAR**
- $\Delta \not\perp\!\!\!\perp X$

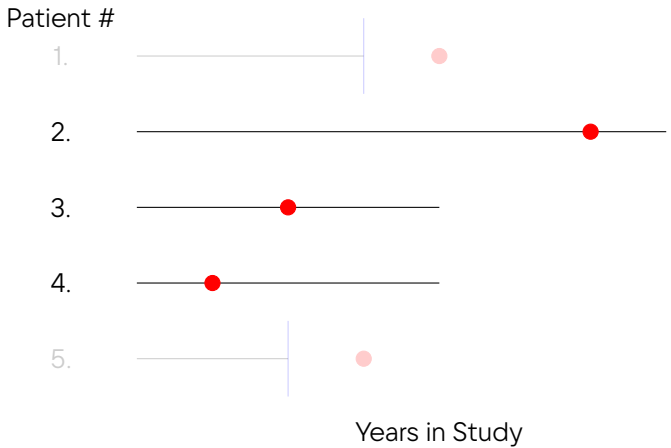
# What happens when we replace $X$ with $W$?



# What happens when we replace $X$ with $W$?



# What happens when we delete the censored observations? (i.e., a complete case analysis)



# What happens when we delete the censored observations? (i.e., a complete case analysis)

- Consider a general missing data setting (not necessarily censoring)

# What happens when we delete the censored observations? (i.e., a complete case analysis)

- Consider a general missing data setting (not necessarily censoring)
- valid if the probability of being a complete case depends on $X, Z$ but not $Y$ (Little and Rubin, 2002)

# What happens when we delete the censored observations? (i.e., a complete case analysis)

- Consider a general missing data setting (not necessarily censoring)
- valid if the probability of being a complete case depends on $X, Z$ but not $Y$ (Little and Rubin, 2002)
- i.e. $\Delta \perp\!\!\!\perp Y | X, Z$

# What happens when we delete the censored observations? (i.e., a complete case analysis)

- Consider a general missing data setting (not necessarily censoring)
- valid if the probability of being a complete case depends on $X, Z$ but not $Y$ (Little and Rubin, 2002)
- i.e. $\Delta \perp\!\!\!\perp Y | X, Z$
- What happens if our COVARIATE is censored?

# What happens when we delete the censored observations? (i.e., a complete case analysis)

- Consider a general missing data setting (not necessarily censoring)
- valid if the probability of being a complete case depends on $X, Z$ but not $Y$ (Little and Rubin, 2002)
- i.e. $\Delta \perp\!\!\!\perp Y | X, Z$
- What happens if our COVARIATE is censored?
- What happens if our OUTCOME is censored (i.e., switch the roles of X and Y?)

# Let’s Try it!

Click here for the code on my GitHub!

# Thank you! Any questions?

✉: mashner@live.unc.edu