Confounding

Prior to this lecture we will have already discussed confounding conceptually. This will be the first detailed example.

Example

- **Study design:** Prospective cohort
- **Population:** Diagnosed with head and neck cancer within the past 30 days at Duke Cancer Institute
- **Outcome variable:** Death at 5 months post-diagnosis (yes/no)
- **Exposure variable:** Tumor site (oral cavity/oropharynx)
- **Confounder:** Stage at diagnosis (I/II or III/IV)

Biological Basis for the Research
(mostly reminders from previous classes)

- Head and neck cancer refers to squamous cell cancers at the following sites:
  - Voicebox (larynx)
  - Throat (pharynx)
  - Airway behind the nasal cavity and including the tonsils and back of the tongue (oropharynx)
  - Mouth including the front of the tongue (oral cavity)
- Cancer at these sites has different etiology
  - Larynx, pharynx, and oral cavity cancers are caused primarily by tobacco and alcohol use
  - Oropharynx cancers are usually HPV-related and often occur in non-smokers
  - Therefore, tumor site is a proxy for “HPV-related” vs. “HPV-unrelated” head and neck cancer
- Cancer at these sites is **hypothesized** to have different prognosis
  - Differences in tumor biology (which arise due to etiologic differences) may lead to different probabilities of survival with current treatment approaches
- But stage at diagnosis also influences survival
  - Later stage indicates advanced disease, and advanced disease is less likely to be curable
  - For some reason (not entirely known) cancer at some sites tends to be diagnosed at later stage
  - This creates a problem called ‘confounding’ that can lead to problems with testing the study hypothesis

Statistical Background for the Example

- This study uses the relative risk (RR) to compare the probability of death at 5 months between patients who were diagnosed with oral cavity vs. oropharynx cancer
- This study estimates the relative risk to be 1.47
- This implies patients who have oral cavity cancer have a 47% higher risk of death at 5 months than patients with oropharynx cancer
- But this estimate is *biased* because it does not account for prognostic differences between patients with each type of cancer
  - Oral cavity cancers are more likely to be diagnosed at late stage (high risk of death)
  - Oropharynx cancers are more likely to be diagnosed at early stage (low risk of death)
- We will introduce the concept of adjustment for this kind of bias—called confounding—by using a statistical technique called conditioning.

Example of Confounding

	Dead at 5 Years	Alive at 5 Years	
Oral Cavity	38	51	89
Oropharynx	40	98	138
			227

Crude (unadjusted) relative risk:

RR= (38/89) / (40/138) = 1.47 (95% CI: 1.03, 2.10)

After adjusting for confounding by stage at diagnosis we obtain the following estimate, which is substantially attenuated compared to the unadjusted estimate:

RR = 1.22 (95% CI: 0.78, 1.92)

A simple calculation shows us the magnitude of the bias:

$(RR_{crude} – RR_{adjusted}) / Rr_{crude}$

(1.47-1.22/1.47)*100 = 17%

Based on an example from *Nephron Clin Pract* 2010;116:c143–c147

Intuition About Confounding: Stratifying on Stage

	Dead at 5 years	**Alive at 5 Years**	
Oral Cavity	38	51	89
Oropharynx	40	98	138
			227

	Stage I/II			**Stage III/IV**		
	Dead	**Alive**	**Total**	**Dead**	**Alive**	**Total**
Oral Cavity	7	25	32	31	26	57
Oropharynx	17	67	84	23	31	54
Total	24	92	**116**	54	57	**111**

- All we've done here is separate the full sample (in the top table) into subgroups according to stage at diagnosis (the confounder).
- Separating the sample into subgroups of stage at diagnosis is called ***conditioning***

Based on an example from *Nephron Clin Pract* 2010;116:c143–c147

Intuition About Confounding: Stratifying on Stage

	Stage I/II			Stage III/IV		
	Dead	**Alive**	**Total**	**Dead**	**Alive**	**Total**
Oral Cavity	7	25	32	31	26	57
Oropharynx	17	67	84	23	31	54
Total	24	92	**116**	54	57	**111**

A few interesting things are revealed by conditioning on stage at diagnosis (i.e., separating the sample into subgroups of stage at diagnosis).

One observation is that the group of patients who have oral cavity cancer is overburdened with tumors that have a poor prognosis. You can observe this through simple descriptive analysis:

- A greater proportion of oral cavity cancers are Stage III/IV (57/89=64%) than oropharynx cancers (54/138=39%)
- The risk of death for Stage III/IV cancer (54/111=49%) is much higher than in for Stage I/II cancer (24/116=21%)

Based on an example from *Nephron Clin Pract* 2010;116:c143–c147

Intuition About Confounding: Stratifying on Stage

	Stage I/II			Stage III/IV		
	Dead	**Alive**	**Total**	**Dead**	**Alive**	**Total**
Oral Cavity	7	25	32	31	26	57
Oropharynx	17	67	84	23	31	54
Total	24	92	**116**	54	57	**111**

RR = (7/32) / (17/84) = 1.08 RR = (31/57) / (23/54) = 1.27

Another observation: Conditioning allows us to conduct the analysis for each stratum separately.

But why would we want to do this?

By conditioning on stage and analyzing within each stratum we have "held stage constant" for all patients in that stratum.

In other words, conditioning removes the effect of stage at diagnosis within each subgroup (because everyone in each subgroup has the same stage).

Based on an example from *Nephron Clin Pract* 2010;116:c143–c147

Intuition About Confounding: Stratifying on Stage

	Stage I/II			Stage III/IV		
	Dead	**Alive**	**Total**	**Dead**	**Alive**	**Total**
Oral Cavity	7	25	32	31	26	57
Oropharynx	17	67	84	23	31	54
Total	24	92	**116**	54	57	**111**

RR = (7/32) / (17/84) = 1.08 RR = (31/57) / (23/54) = 1.27

But this conditional analysis leaves us with 2 separate comparisons of the risk of death in oral cavity vs. oropharynx cancer.

The “eyeball test” suggests something may be wrong: the two estimates look different from each other, and they are both different from the overall estimate of 1.47.

If the relative risks really are different in the population then this is **interaction**. Unlike confounding, interaction is not a form of bias and can be ignored if we choose. Later lectures will cover analysis of interaction.

For now, we will move on with the goal of this study: to find a single estimate of the relative risk that isn’t affected by stage at diagnosis.

Based on an example from *Nephron Clin Pract* 2010;116:c143–c147

Intuition About Confounding: Stratifying on Stage

	Stage I/II			Stage III/IV		
	Dead	**Alive**	**Total**	**Dead**	**Alive**	**Total**
Oral Cavity	7	25	32	31	26	57
Oropharynx	17	67	84	23	31	54
Total	24	92	**116**	54	57	**111**

RR = (7/32) / (17/84) = 1.08 RR = (31/57) / (23/54) = 1.27

So, how do we combine the stratum-specific estimates of RR to get 1 estimate that isn't influenced by stage at diagnosis?

An intuitive approach would be to simply average the stratum-specific estimates: (1.08+1.27)/2 = 1.18

But this is imperfect because the strata are different sizes.

Based on an example from *Nephron Clin Pract* 2010;116:c143–c147

Intuition About Confounding: Stratifying on Stage

	Stage I/II			Stage III/IV		
	Dead	**Alive**	**Total**	**Dead**	**Alive**	**Total**
Oral Cavity	7	25	32	31	26	57
Oropharynx	17	67	84	23	31	54
Total	24	92	**116**	54	57	**111**

RR = (7/32) / (17/84) = 1.08 RR = (31/57) / (23/54) = 1.27

A better approach is to weight the average such that the weight reflects how much each stratum contributes to the mean.

The Mantel-Haenszel method does this by using a specific form of *standardization.* Standardization is covered in 709.

Regression takes a different approach: rather than averaging stratum-specific RRs, it uses a model to estimate the effect of tumor site **within levels of stage** (a conditional effect), assuming a particular form of the relationship.

You will learn something about how this estimation process works in the 705 course. In the meantime, we will focus simply on interpretation rather than the mechanics of estimation.

Based on an example from *Nephron Clin Pract* 2010;116:c143–c147

Some Regression Basics

A simple regression model has only 1 predictor:

$Y = B0 + B1*X$

In this kind of model B1 is an “unconditional” estimate of the association between X and Y

We call it unconditional because there’s nothing else (aside from X) on the right-hand side of the model.

A Model for Our Example

A simple model for our example would have 1 predictor, which is the tumor site.

Y = number dead at 6 months

Site = {0=oropharynx, 1=oral cavity}

For reasons we don't need to get into here, a common regression model is for the log of Y. This is called a Poisson model.

$\ln(Y) = B0 + B1*Site$

In this model, B1 is the mean difference (on the log scale) in the number of deaths at 5 months comparing the oral cavity to oropharynx cancer.

This estimate does not condition on stage because stage is not included in the model.

Using Regression to Condition

We can ask the software to condition on stage by specifying a model with 2 predictors:

Y = number dead at 5 months

Site = {0=oropharynx, 1=oral cavity}

Stage = {0=I/II, 1=III/IV}

$\ln(Y) = B0 + B1*Site + B2*Stage$

The estimate for B1 will now be conditional on stage. But because B1 is a mean difference in the log-count we can exponentiate to get a ratio of the count (recall, a difference on the log scale is a ratio on the anti-log scale):

RR = exp(B1) = 1.22 in our example

This allows us to say "**If we hold stage constant**, then the risk of death at 5 months is 22% higher in oral cavity cancer."

Therefore, regression is a method for adjusting for confounding.

Probability Theory is Present Here!

- You learned about the following in the 701 course:
  - Joint probability
  - Conditional probability
  - The law of total probability
- All of these things appear in the contingency table analysis

What is the joint distribution?

$$P[T = t, S = s, D = d] = \frac{Number\ of\ patients\ with\ (t, s, d)}{Total\ number\ of\ patients}$$

Tumor Site (T)	Stage (S)	Dead (D=1)	Alive (D=0)	Total
Oral	I/II	7	25	32
Oral	III/IV	31	26	57
Oropharynx	I/II	17	67	84
Oropharynx	III/IV	23	31	54
Total		78	149	227

A couple examples:

$$P[T = Oral, S = I/II, D = 1] = \frac{7}{227} = 0.03$$

$$P[T = Oropharynx, S = I/II, D = 1] = \frac{17}{227} = 0.07$$

What is the joint distribution?

The Law of Total Probability says that all probabilities must sum to 1 in order for this distribution to be a proper probability distribution.

Tumor Site (T)	**Stage (S)**	**Dead (D=1)**	**Alive (D=0)**	**Total**
Oral	I/II	7 (0.03)	25 (0.11)	32
Oral	III/IV	31 (0.14)	26 (0.11)	57
Oropharynx	I/II	17 (0.07)	67 (0.30)	84
Oropharynx	III/IV	23 (0.10)	31 (0.14)	54
Total		78	149	227

0.03 + 0.14 + 0.07 + 0.10 + 0.11 + 0.11 + 0.30 + 0.14 = **1**

What is the conditional distribution?

$$P[D = 1|T = t, S = s] = \frac{Number\ of\ patients\ with\ D = 1\ and\ (t, s)}{Total\ number\ of\ patients\ with\ (t, s)}$$

Tumor Site (T)	**Stage (S)**	**Dead (D=1)**	**Alive (D=0)**	**Total**
Oral	I/II	7	25	32
Oral	III/IV	31	26	57
Oropharynx	I/II	17	67	84
Oropharynx	III/IV	23	31	54
Total		78	149	227

A couple examples:

$$P[D = 1|T = Oral, S = I/II] = \frac{7}{32} = 0.22$$

$$P[D = 1 \mid T = Oropharynx, S = I/II\ ] = \frac{17}{84} = 0.20$$

What is the conditional distribution?

The Law of Total Probability still applies, but within the subgroups that are created by conditioning.

Tumor Site (T)	**Stage (S)**	**Dead (D=1)**	**Alive (D=0)**	**Total**
Oral	I/II	7 (0.22)	25 **(0.78)**	32
Oral	III/IV	31 (0.54)	26 **(0.46)**	57
Oropharynx	I/II	17 (0.20)	67 **(0.80)**	84
Oropharynx	III/IV	23 (0.43)	31 **(0.57)**	54
Total		78	149	227

0.22 + **0.78** = **1**
0.54 + **0.46** = **1**
0.20 + **0.80** = **1**
0.43 + **0.57** = **1**

A Brief Preview of Causal Inference

Summary of the Problem

The data are telling us there is an association between tumor site and survival at 5 months

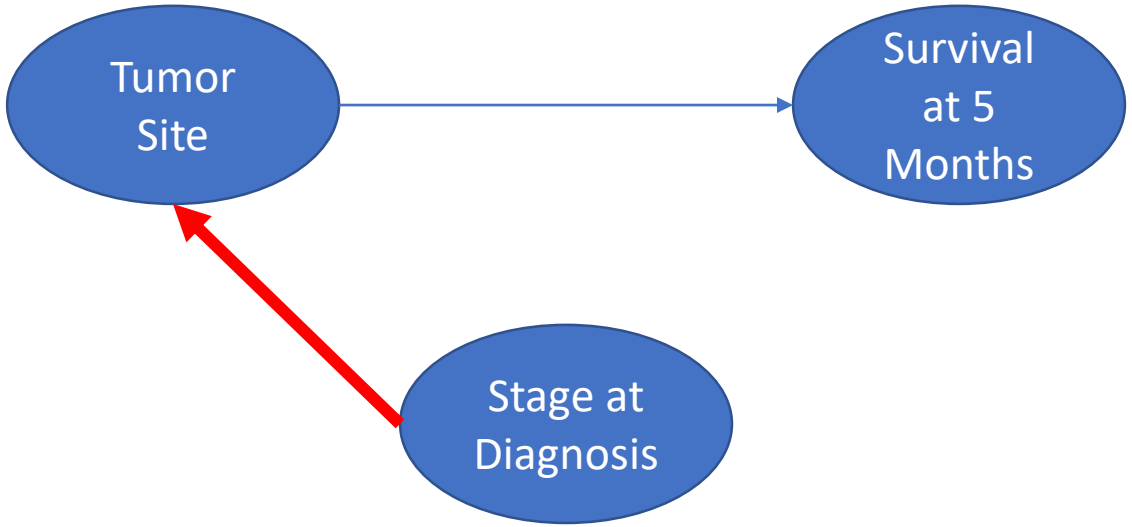
The estimate of the association is strong! (RR=1.47)

But it's not a terribly precise estimate (95% CI: 1.03, 2.10)

Since the CI includes shoulder-shrugging values as well as eye-popping values it's a bit hard to decide how excited we should be about this finding.

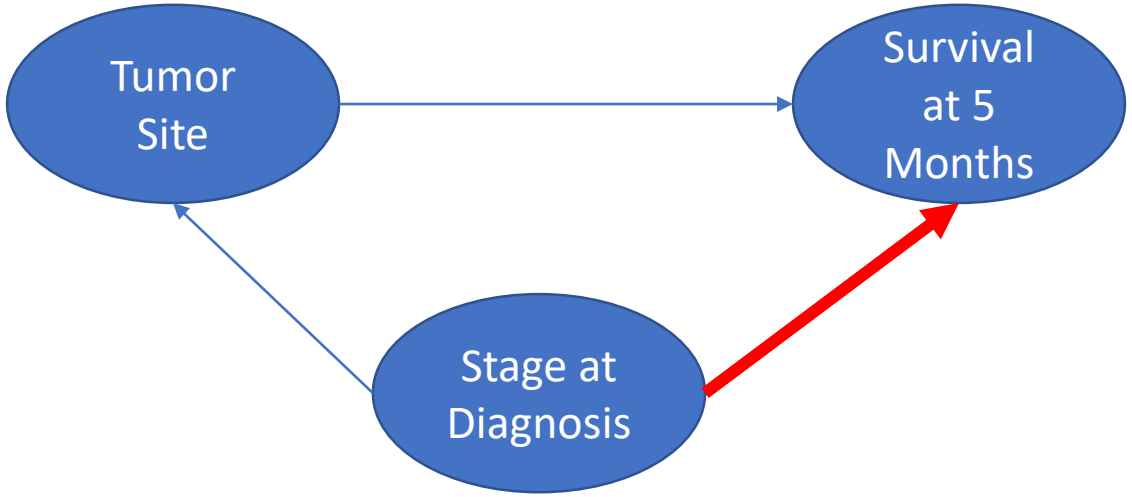
BUT…there is an even bigger problem here, which is…

Summary of the Problem



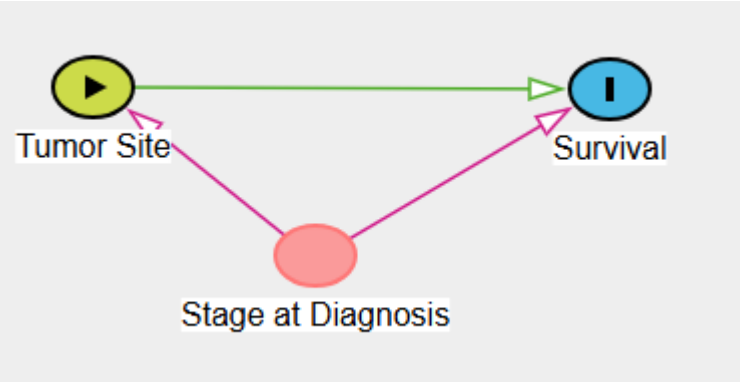
- We saw that later stage at diagnosis was more common in oral cancer than oropharynx cancer
- This could reflect a real phenomenon <u>in the population</u>
  - But whether it is “causal” is questionable (more on this in a moment)
- So, we can say that we’ve observed an association between stage at diagnosis and tumor site
  - Note that we don’t test for statistical significance of this association because it’s not the focus of our research
  - The mere presence of this imbalance in the sample suggests potential trouble ahead

Summary of the Problem



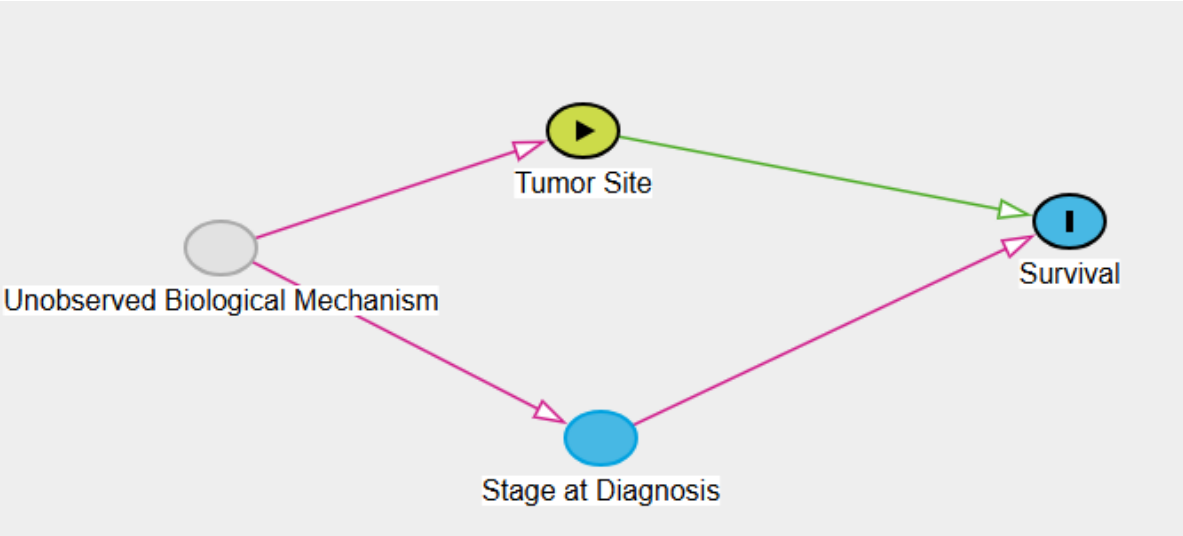
- We also noticed that patients in this sample who had later stage at diagnosis had a higher risk of death at 5 months than patients who had earlier stage at diagnosis
- So, it appears that stage at diagnosis has associations with both the exposure variable we're studying (tumor site) and the outcome (survival at 5 months)
- The pattern of this association is such that oral cavity cancers have a high burden of poor prognostic features relative to oropharynx; thus, our comparison of oral vs. oropharynx cancers might look worse than it really is.

Definition of Confounding
(using the language of causal inference)



- A variable is a **confounder** if it is a common cause of exposure and outcome
- Some authors will not draw the line that represents the hypothesized relationship between exposure and outcome, but we have drawn it here
  - Doing so enables us to use Daggity to identity Stage at Diagnosis as a confounder, indicated by the pink color of the node in the DAG
- But is stage at diagnosis really a “cause” of tumor site?

An Improved Causal Diagram



- Stage at diagnosis is plausibly related to tumor site in this population, but probably not through a direct mechanism
- Instead, tumor site and stage are probably related through a common cause (or multiple common causes)
- This common cause is not observable when we enroll patients at the time of diagnosis
- Nonetheless, it creates a backdoor path of association from stage at diagnosis to tumor site
- The question is, how do we incorporate stage at diagnosis into our analysis to block this back door path?

There are many ways to adjust for a confounder

- The adjustment method we select will provide an unbiased estimate of association for a specific target population
- This implies subtle differences in interpretation for different adjustment methods
- This leads to consideration of the estimand for the study, which you'll learn more about in other courses

A list of some adjustment methods

- Mantel-Haenszel for contingency tables
- Standardization (there are several approaches)
- Regression
- Inverse probability weighting

If you’re interested in this…

…then sign up for BIOSTAT 709 in the spring!