Biostatistics

Mithat Gonen

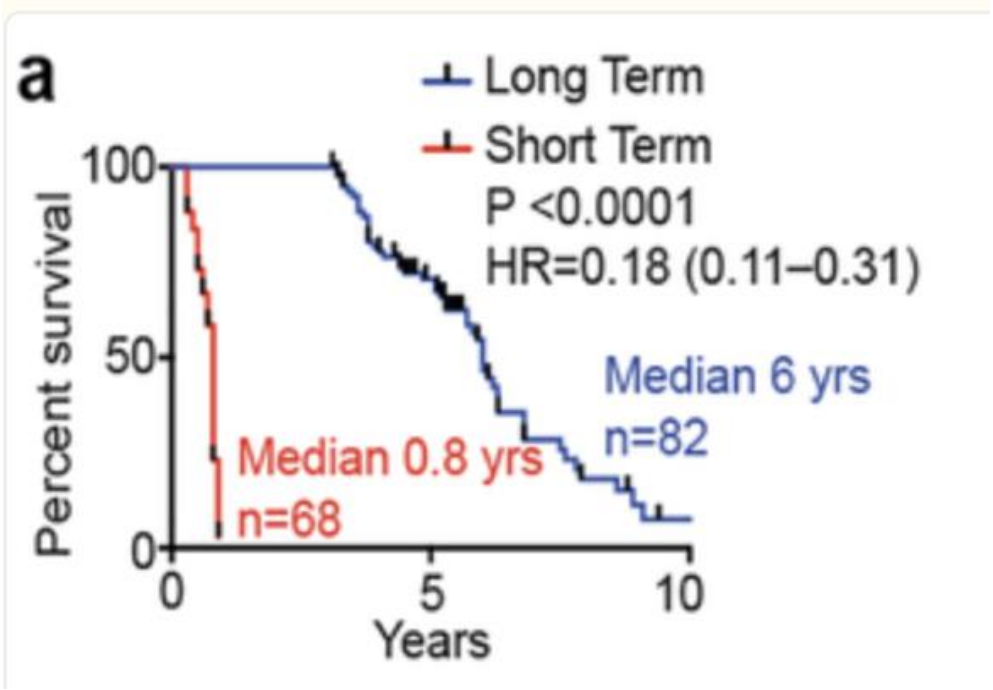
Charlie White

Today's Lecture

- Survival Analysis (cont'd)
  - Immortal Time Bias Revisited
  - Left Truncation
- Study Design
  - Confounding
  - Matching and Stratification
  - Regression

Letter | Published: 01 November 2017

Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer



What are the groups?

- Surviving at least 4 years (long term)
- Surviving at most 1 year (short term)

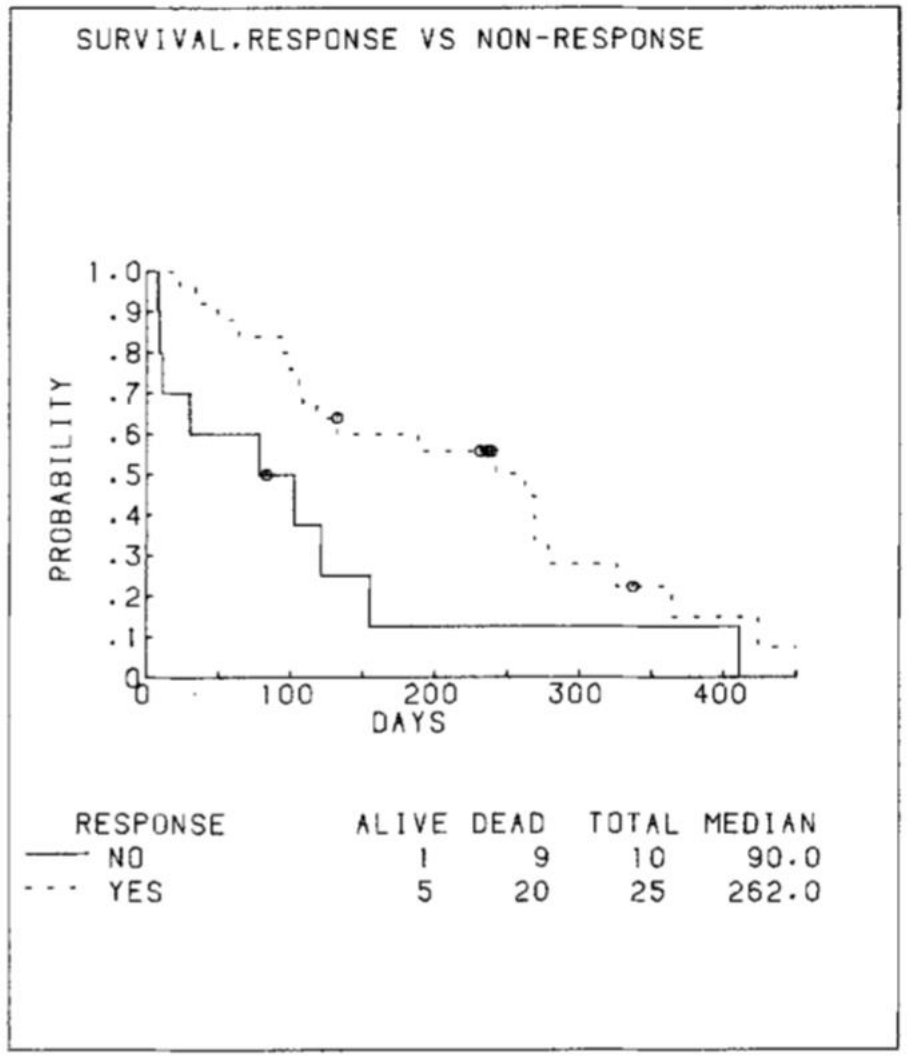


Fig. 2. Evaluation of survival for responders versus nonresponders by the usual method. Data from Eastern Cooperative Oncology Group Study EST 3477: Phase II master protocol for evaluation of agents in patients with multiple myeloma.

Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. J Clin Oncol. 1983 Nov;1(11):710-9. doi: 10.1200/JCO.1983.1.11.710. PMID: 6668489.

What are the groups?

- Responder vs non-responder?
- Or?
  - Responder and surviving at least 6 weeks
  - Non-responder or surviving less than 6 weeks

Landmarking

- Move the baseline to the landmark time
- Use the response status at landmark time (responders after landmark are considered non-responders) and exclude those who died before the landmark time
- How to choose the landmark time?
  - More art than science

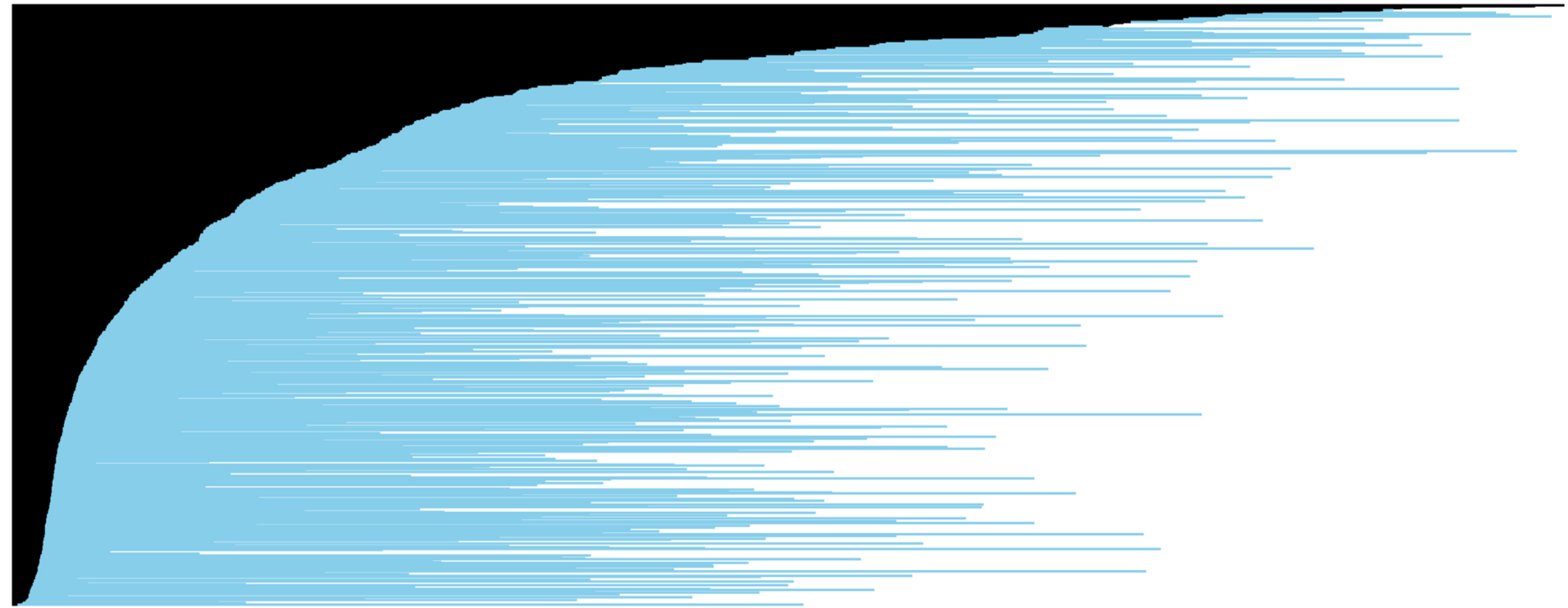
Left Truncation

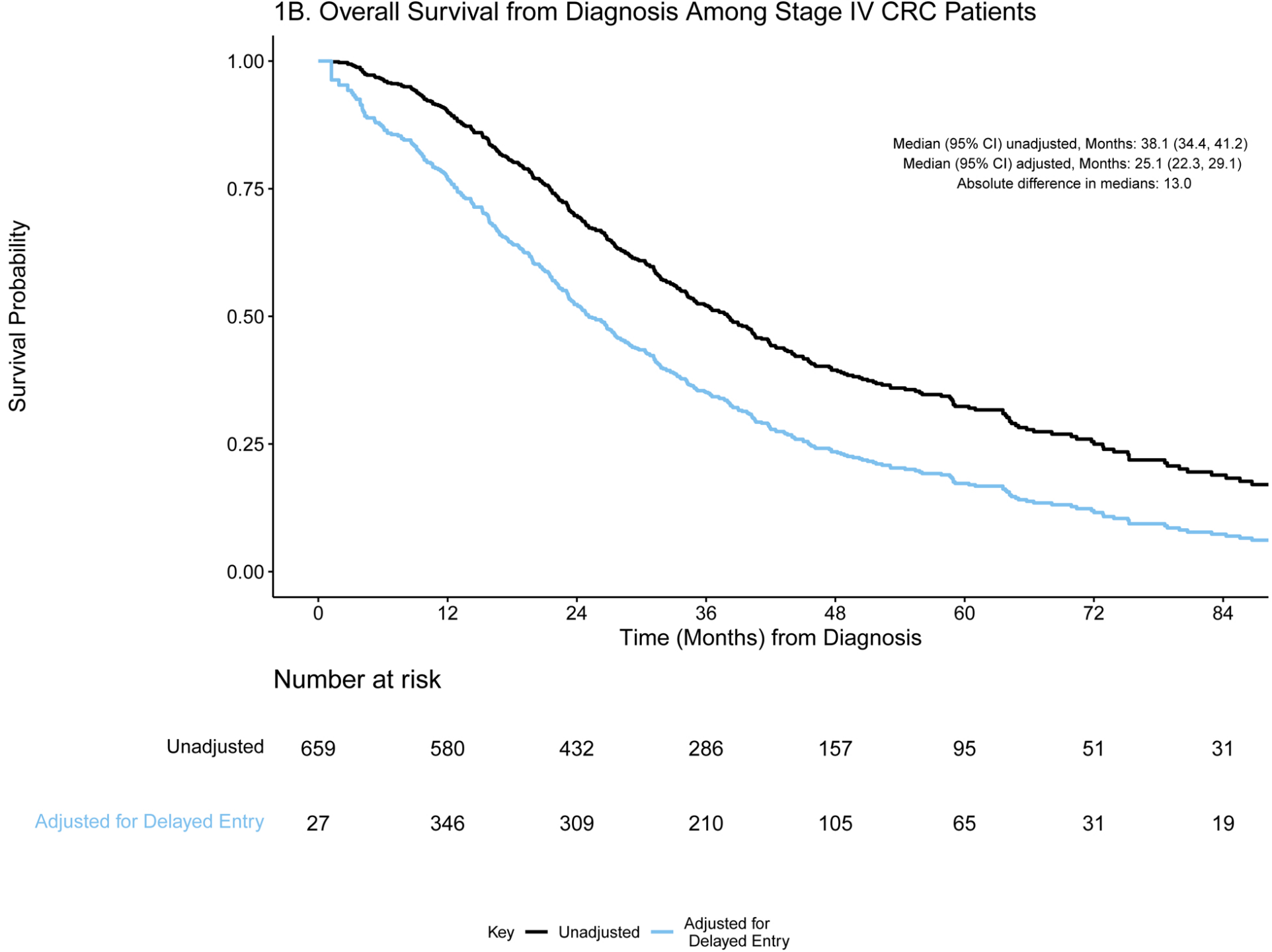
- Common in clinical genomic studies
- Most obvious when some patients' genomic material is harvested at progression, but we want to analyze time from diagnosis
- By definition, that patient was not at risk of death during the period between diagnosis and progression
- This is called left truncation (late entry into the risk set)
- We need to account for left truncation in these studies

Selection Bias

- When there is left truncation there is usually selection bias; some patients do not make it to sequencing at all
- This is mostly a problem of selective sequencing. If selection is for clinical reasons than it could add to bias as well.
- You can see if there is selection bias by plotting the KM curve of unsequenced patients to those sequenced (adjusted for left truncation)
  - Brown S, Lavery JA, Shen R, Martin AS, Kehl KL, Sweeney SM, Lepisto EM, Rizvi H, McCarthy CG, Schultz N, Warner JL, Park BH, Bedard PL, Riely GJ, Schrag D, Panageas KS; AACR Project GENIE Consortium. Implications of Selection Bias Due to Delayed Study Entry in Clinical Genomic Studies. JAMA Oncol. 2022 Feb 1;8(2):287-291. doi: 10.1001/jamaoncol.2021.5153. PMID: 34734967; PMCID: PMC9190030.

1A. Event History for Overall Survival from Diagnosis Among Stage IV CRC Patients (N=659)




1B. Overall Survival from Diagnosis Among Stage IV CRC Patients
Survival Probability
1.00
0.75
0.50
0.25
0.00
Median (95% CI) unadjusted, Months: 38.1 (34.4, 41.2)
Median (95% CI) adjusted, Months: 25.1 (22.3, 29.1)
Absolute difference in medians: 13.0
0
12
24
36
48
60
72
84
Time (Months) from Diagnosis
Number at risk
Unadjusted 659 580 432 286 157 95 51 31
Adjusted for Delayed Entry 27 346 309 210 105 65 31 19
Key Unadjusted Adjusted for Delayed Entry

STUDY DESIGN

Example

- Use of adjuvant hepatic arterial infusion in treating liver metastases from colorectal cancer
- There are randomized trials out there but there is no agreement in the field on whether HAI should be routinely used
- So you want to contribute to this by analyzing data from your own center
- First thing you need is a study design

Retrospective Study Design

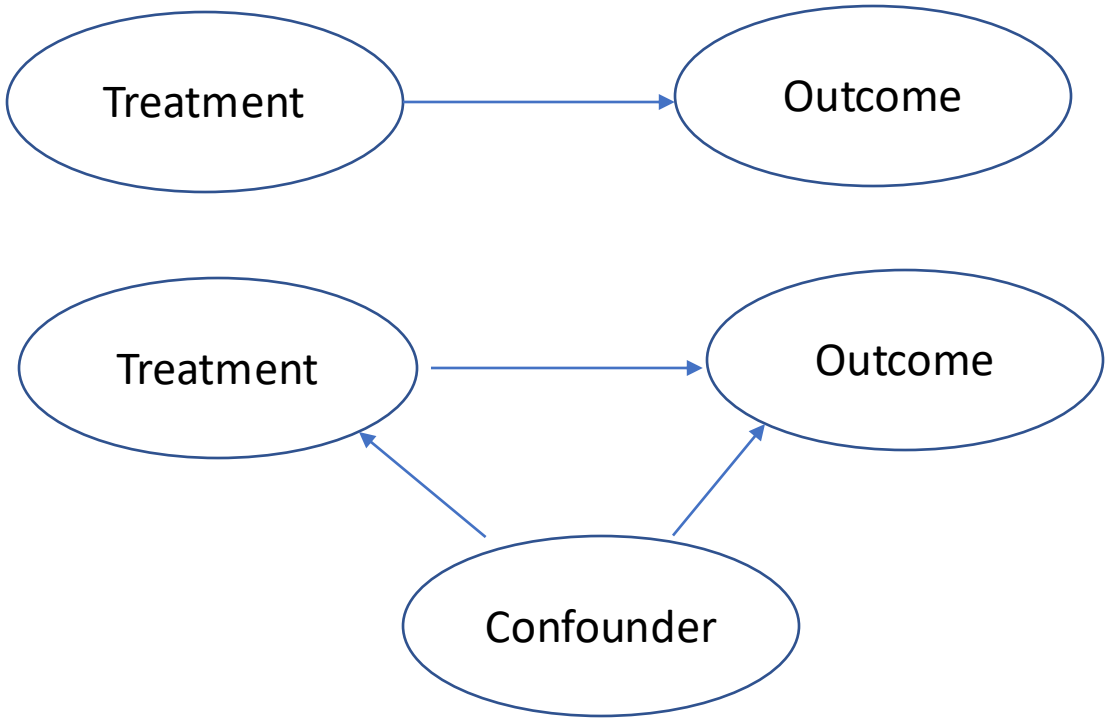
- Sounds like an oxymoron
- But hugely important
- Act as if this is a clinical trial
  - Primary objective/endpoint
  - Secondary objectives/endpoints
- Inclusion/Exclusion Criteria
- Define treatment arms
- …

Example (continued)

- Primary objective/endpoint: Compare overall survival between patients who were treated with adjuvant pump vs not
- Secondary objectives/endpoints: Time to Recurrence, Time to Hepatic Recurrence
- Inclusion/Exclusions: Completely resected patients, at least x doses of delivered by infusion ….
- What are the treatment arms:
  - Arm 1: Anyone who received HAI, additional treatment allowed
  - Arm 2: Anyone who did not receive HAI (?), or maybe anyone who received adjuvant treatment but not HAI

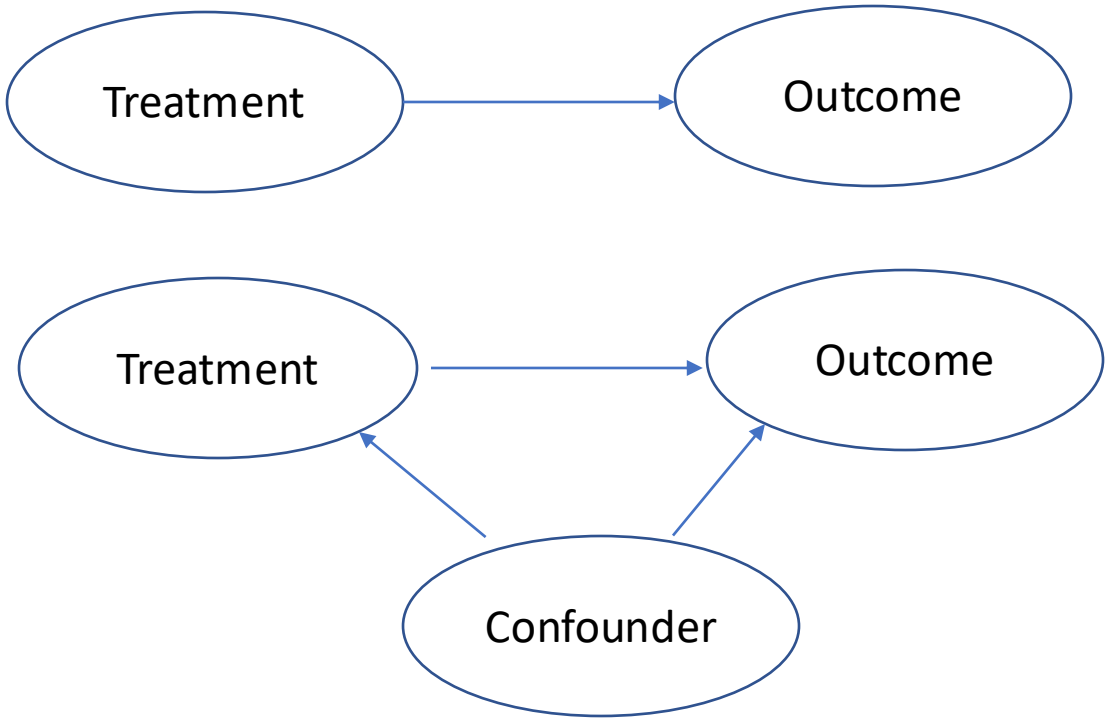
What is a confounder?

- Something that is associated with both the treatment (exposure) and the outcome



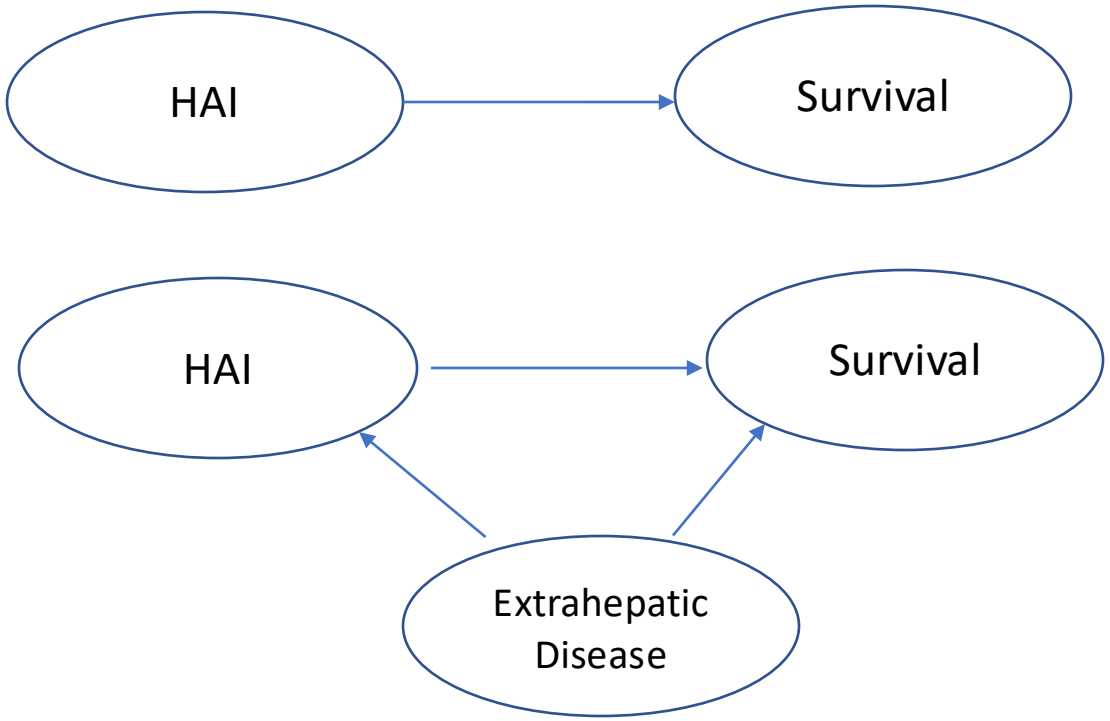
Aside: Causal Diagrams

- Direction of the arrows is important. It means treatment causes outcome.



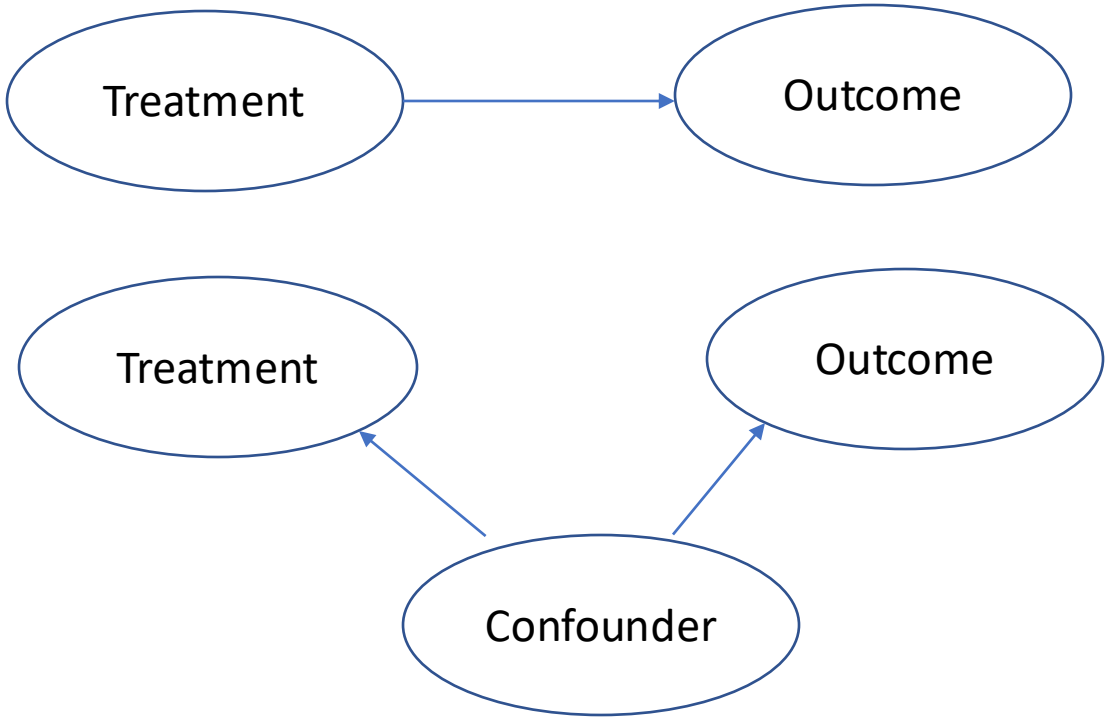
Back to the example

- Can low risk disease be a confounder? Low risk patients are more likely to receive the treatment and also more likely to survive?



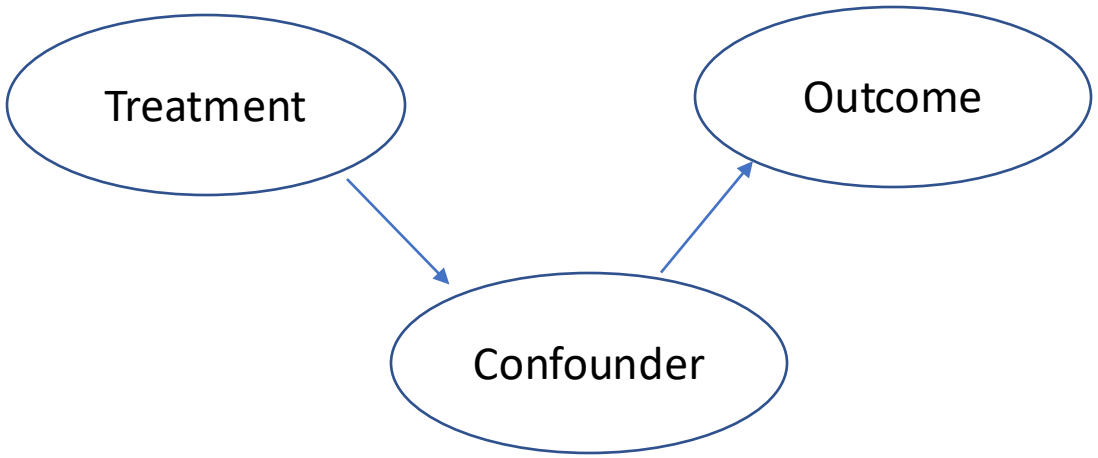
Why is this a problem?

- Even if there is no direct link between treatment and outcome, there is a still a link through confounding. Any analysis that do not account for the confounder will attribute that link to the treatment



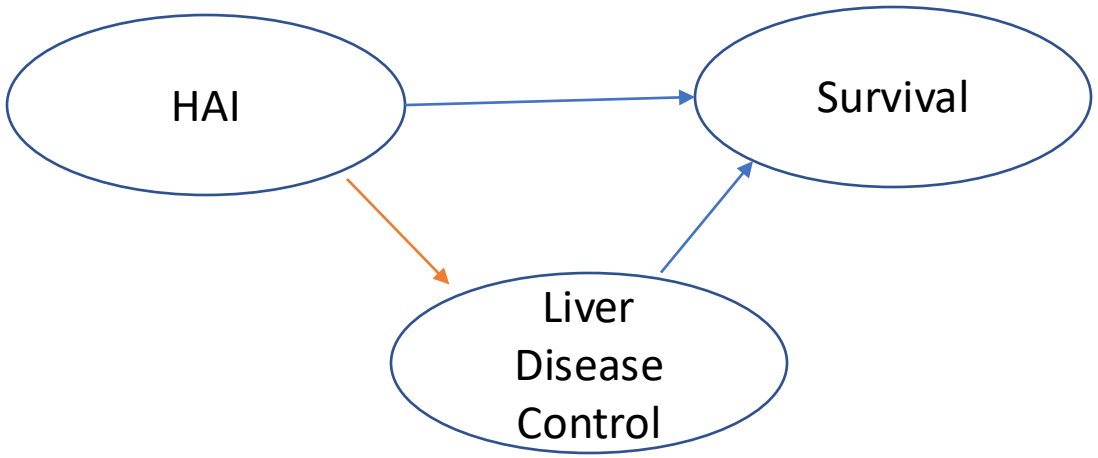
What if?

- The arrow between treatment and confounder is reversed. Is this still confounding?



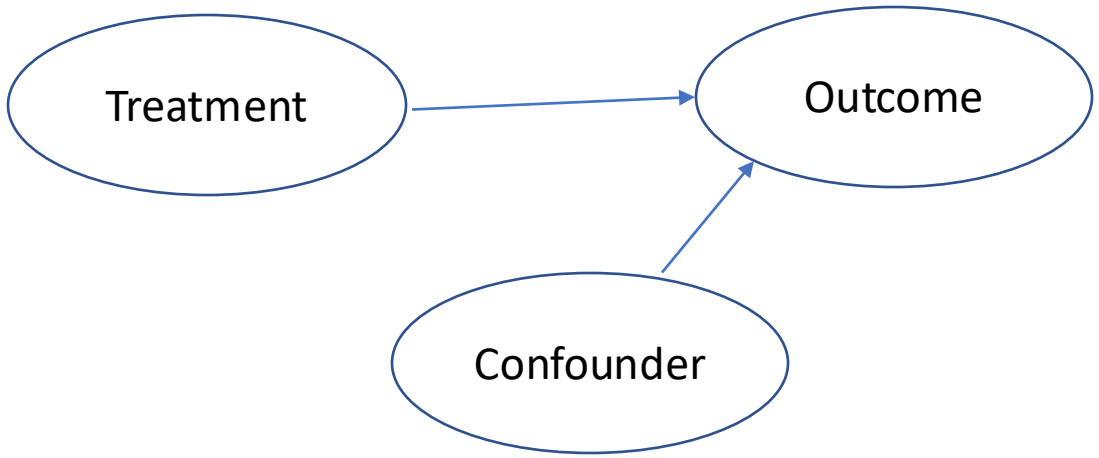
It is not confounding

- (Partial) Mechanism



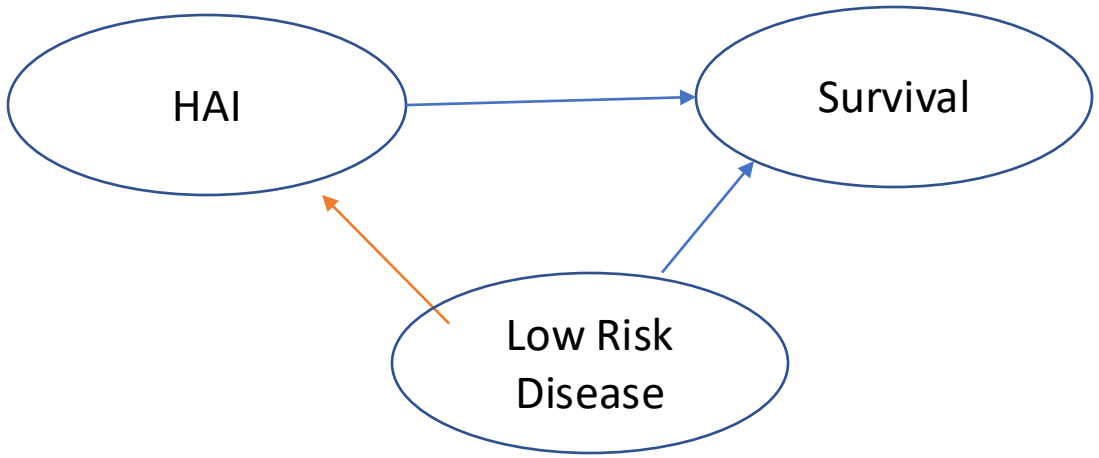
What if?

- There is no arrow between treatment and confounder. Is this still confounding?
- It is called a risk factor, might still need to be adjusted for



How do we deal with confounding

- We need to find ways to remove the link between Treatment and Confounder through data analytic methods
- Remember: That's what an RCT does (removes the link) by design. In an observational study we cannot use random assignment, hence we will need to achieve a similar effect by data analysis.



Matching

- If we compare all treated with all control, confounder raises its head
- What is we find a twin (a match) for each treated patient from the untreated (control) group.
  - Twin = Has the same value for the confounder
  - Suppose low risk is defined as number of liver tumors, time between primary and mets dx and CEA
  - For each patient who received HAI, I find someone who had exactly the same number of tumors, disease free interval and CEA but did not receive HAI
  - Now I have a data set where there is no link between Low Risk Disease and Treatment (I made it so, by “design”).
  - If there is still an association between treatment and outcome then it must be that the treatment causes the outcome

Matching

- If we compare all treated with all control, confounder raises its head
- What is we find a twin (a match) for each treated patient from the untreated (control) group.
  - Twin = Has the same value for the confounder
  - Suppose low risk is defined as number of liver tumors, time between primary and mets dx and CEA
  - For each patient who received HAI, I find someone who had exactly the same number of tumors, disease free interval and CEA but did not receive HAI
  - Now I have a data set where there is no link between Low Risk Disease and Treatment (I made it so, by “design”).
  - If there is still an association between treatment and outcome then it must be that the treatment causes the outcome

A hidden assumption

- “If there is still an association between treatment and outcome, then it must be that the treatment causes the outcome”
- I am making a very critical assumption here that I have not stated
- No other confounders !!!
- Hugely important.
- We are not making this assumption in a randomized study. Random assignment balances all confounders, observed and unobserved.
- Data analysis can at best balance observed and recognized confounders.

Back to Matching

- Find a twin (a match) for each treated patient from the untreated (control) group.
- Easier said than done
- How to choose which variables to match on?
- Node positive primary is also a risk factor in this disease. Should we use that as well?

More on Matching

- Find a twin (a match) for each treated patient from the untreated (control) group.
- Easier said than done
- CEA: do I need an exact match? If a treated patient has a CEA of 87 and there is no control patient with a CEA of 87, but there is one with 86 is it OK to match them? Are they still twins?
- The difference we allow in matching is called a caliper. If we have a caliper of 10 for CEA, then 77 to 97 match a CE of 87.
- How to choose a caliper? Usually not obvious but can be consequential

Categorized Risk Factors Are a Problem in Matching

- CEA >= 200 is used in this disease as a risk factor. Can we match on that?
- Sure, but it is actually worse than a caliper
  - 1 matches 199
  - But 199 does not match 200

Isn’t it possible to fix this?

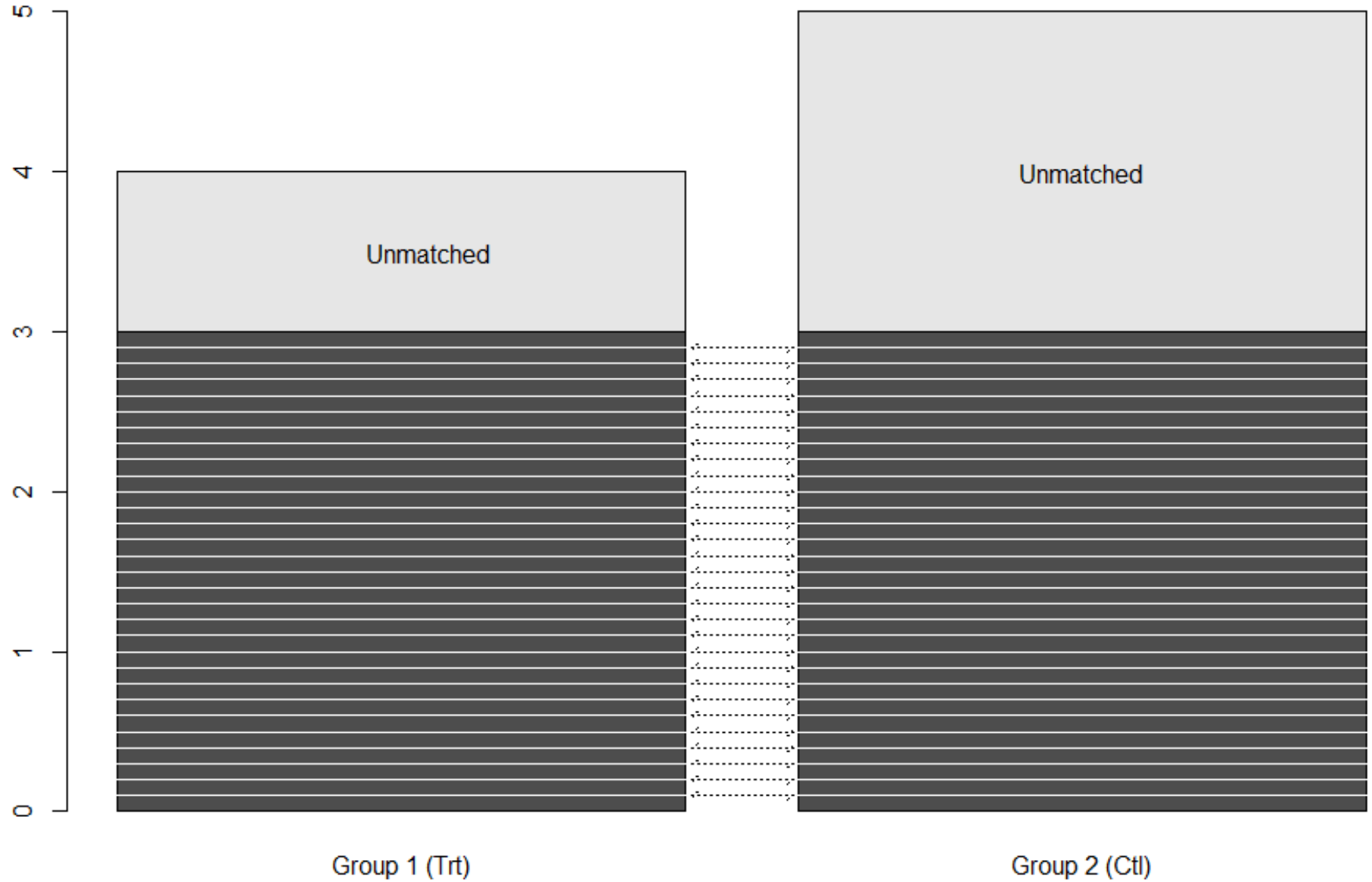
- Use as many confounders as you can think of to match
- Use a very small caliper (such as CEA) or insist on exact matching (number of tumors)
- Most of the time you cannot find a match for every treated patients if you insist on strict matching standards
- Exclude unmatched?

What happens if there are unmatched patients?

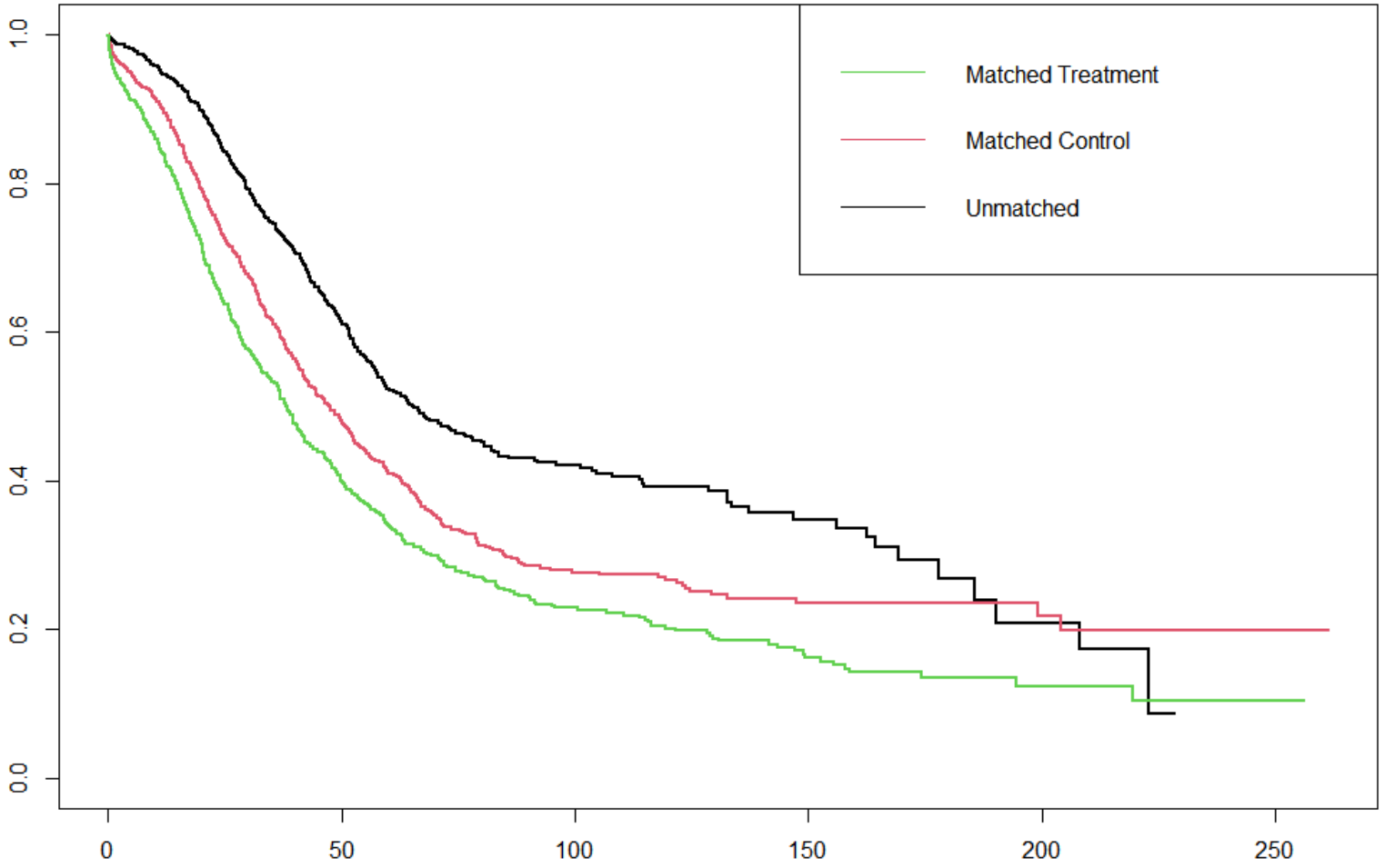
- Suppose we had 100 HAI patients to begin with
- We insisted on strict matching and we were able to match 60 of them
- And go ahead analyze this 60-60 matched cohort
- To what population does this generalize to?
- Only the population where the 60 matched HAI patients came from
- Can you define that population? The original 100 is (presumably) well-defined because you had inclusion/exclusion in your design (see the importance of design)

Leave no patients unmatched

- This means relax the matching criteria
  - Fewer confounders
  - Wider calipers
- But this means less twins more siblings → weaker control of confounding
- The entire field of dealing with confounders can be summarized with this struggle:
  - Bias/Validity tradeoff


5
4
3
2
1
0
Unmatched
Unmatched
Group 1 (Trt)
Group 2 (Ctl)

Example: Suppose everyone treated got matched



Bias/Validity Tradeoff

- Bias is the outcome difference between treated and untreated patients that is due to confounders
- We match to make the treatment and control groups comparable
- If we there are unmatched patients we lose on validity
- If we relax matching rules to improve validity then the groups are less comparable and bias creeps in
- No good solution to this, kind of a Heisenberg principle for empirical research

A note on “Group” Matching

- Some people would call what I described as 1:1 matching
- And they would call also matching if treated and untreated groups are matched on average (i.e. mean CEA is the same in both). “Groups are well-matched, groups are matched on means etc”
- This really is not matching. It is something clinical literature made up. Do not use it.

A note on 1:k matching

- Sometimes useful to capture the variability of the outcome in the control group
- Requires a large control group and most of the time impractical
- 1:2 match on 100 treated patients → 300 patient study
- Less powerful than 1:1 matched on 150 patients
- The rate limiting step is the number of treated patients.
- Do not 1:k match because it you think will increase your power, do it only to capture the variability in outcome

Statistical analysis of matched studies

- You cannot use typical two-sample (two group) tests
  - No two-sample t-test
  - No chi-square test
  - No log-rank test
- Instead
  - Paired t-test
  - McNemar test
  - Paired log-rank test

Fundamental idea of paired tests

- Remember the two-sample tests?
- Mean in group 1 minus mean in group 2
- Divided by the standard error of this mean difference
- Does not use at all the matching information
- Instead take the difference within each pair first
- Then calculate the mean of the differences and its standard error
- → Paired t-test

All paired tests work on this principle

- Do not compare group means (or medians etc)
- Compare the pairs and average over the pairs
- This way you are truly comparing like to like (to the extent your matching created likes to likes)

Summary of Matching

- Most matches are not twins. They are at best siblings with many differences between them
- Unmatched patients are a threat to the validity of conclusions
- Most of the time matching is not a great way to deal with confounding for these reasons
- But it has great face value: a lot of clinicians think “a matched cohort” is great even if they do not understand where we traded off between bias and validity

Brain Teaser

- You have a matched cohort and analyzed it two ways
  - Ignore matching, compare groups
  - Do an appropriate paired test
  - Which p-value will be smaller? Why?

Brain Teaser

- You have a matched cohort and analyzed it two ways
  - Ignore matching, compare groups
  - Do an appropriate paired test
  - Which p-value will be smaller? Why?
- Paired test p-value will be smaller
  - Group tests include variability between all pairs of patients (conceptually)
  - Paired tests only focus on the between pair variability

Stratification

- Can we match on a single categorical variables?
- Consider a different example: adjuvant treatment in localized colon cancer. All stage III's get it and so do some stage II's.
- If we are doing an observational treatment comparison can we simply match on stage II vs III?

Technically yes

- For each treated stage II patient, randomly choose an untreated stage II patient
- Can we call this a match?
- In the eye of the beholder
- If the matching group definitions are very broad and if there are only a few categories to match, then it may be better to use stratification instead of matching

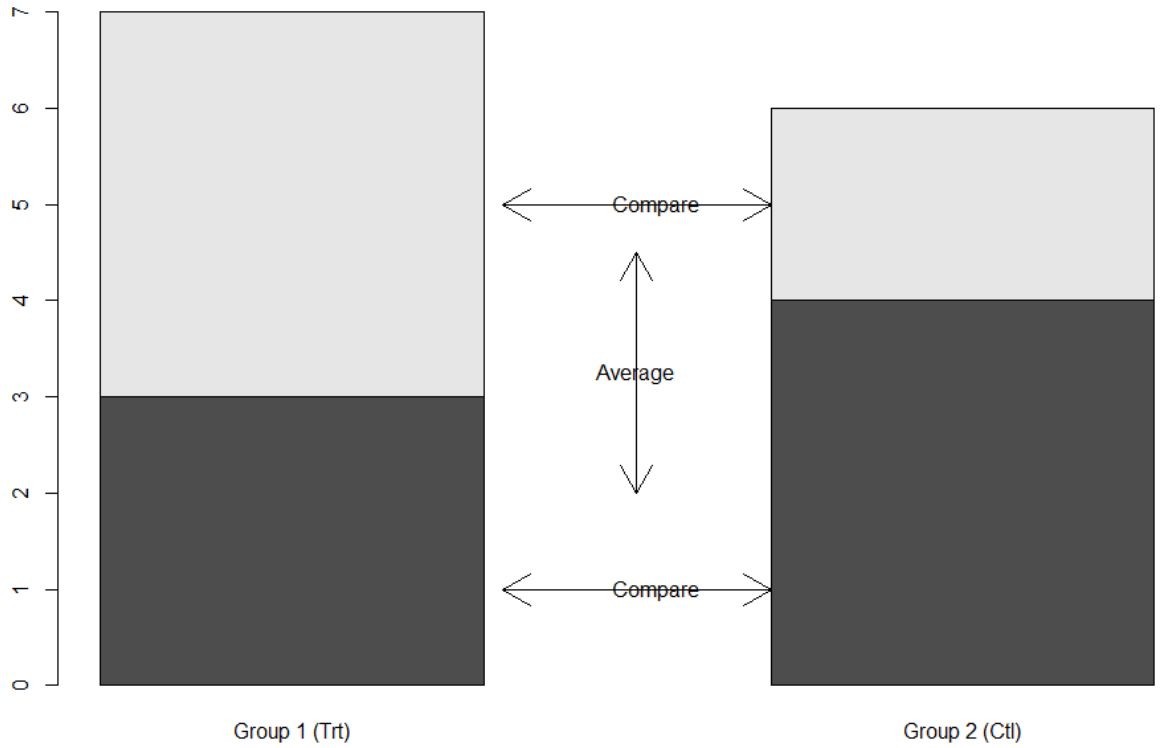
What is stratification?

- Form (a few) strata from the confounding variables
- Compare treatment and control within each strata (using everyone in that strata)
- Average these comparisons across strata
- No patients excluded, so validity is intact
- What about bias?

Bias in stratification

- Stratification also compares like to like, except that it defines “like” based on a very loosely defined criteria (like Stage II vs III)
- In that sense it is a little like matching on a few variables with large caliper
- → Bias is a concern
- Strengths
  - Transparent: no excluded patients, no arbitrary calipers

Stratification



Regression

- Start simple, outcome Y and one input variable (predictor, covariate) X, both continuous.
- I will write Y = f(X), where f(.) generically denotes a function. Our general aim is to figure this f(.) thing out
- If you have one X you can try many types of f's or even leave it unspecified. But for multivariable regression (many X's) we will limit ourselves in this class to a linear form.
- $E(Y) = \alpha + \beta X$, where E(.) means "expected value" or "mean of"
- α and β are parameters (remember populations vs sample; parameter vs estimate)

Linear Regression

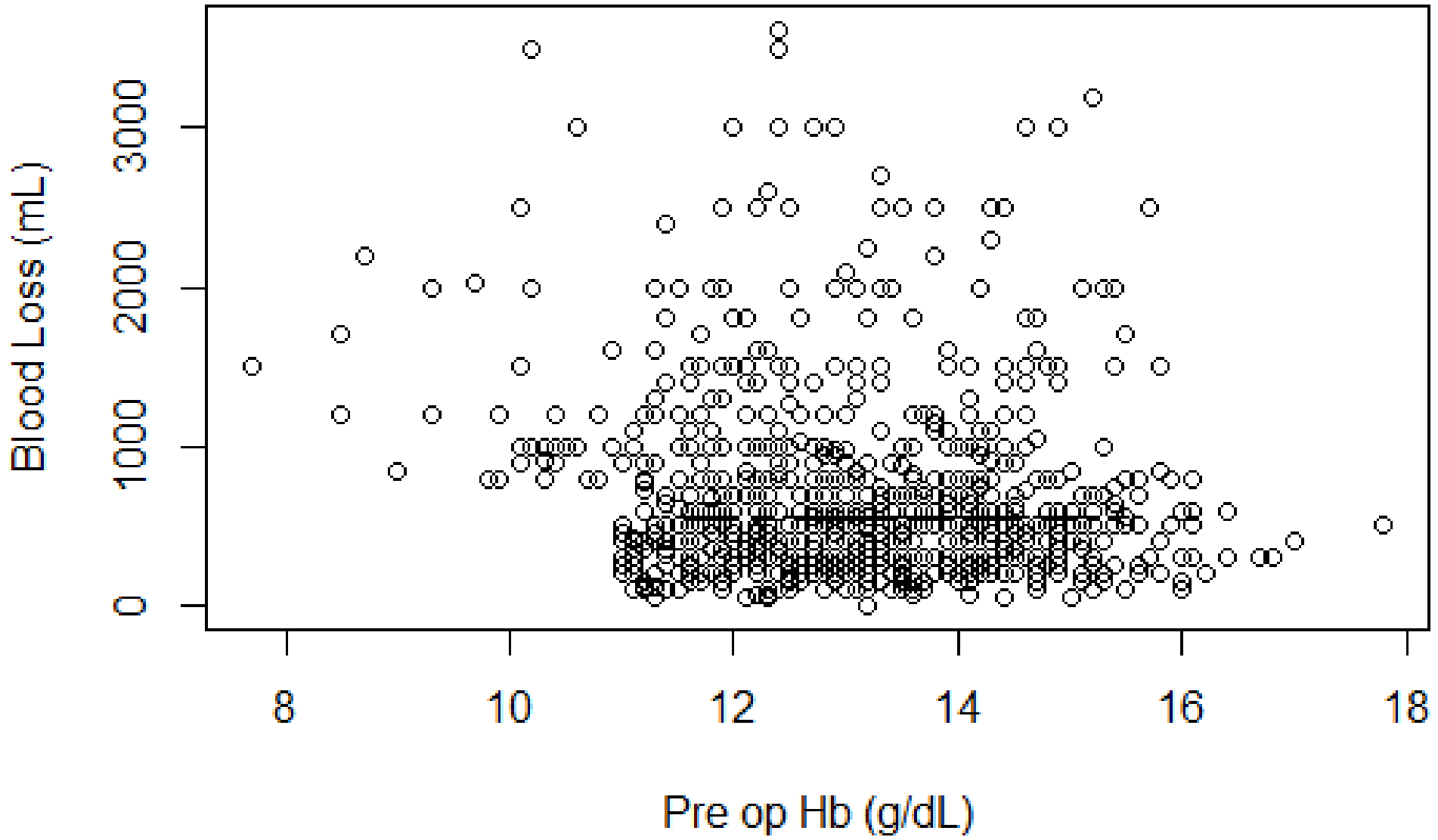
- $E(Y) = \alpha + \beta X$, where E(.) means “expected value” or “mean of”
- α and β are parameters (remember populations vs sample; parameter vs estimate)
- When have estimates instead of parameters the equation will look like
- $Pred(Y) = a + bX$
- a and b are estimates
- Pred(Y) means predicted value of Y

Estimation vs Prediction

- Finding the best-fit value of a parameter → Estimation
  - a and b are parameters
- Finding the best-fit value of an observation → Prediction
  - Pred(Y) is a prediction
- In regression and almost all other models
  - We first estimate the parameters (sometimes called fit the model)
  - We then generate predictions of the outcome

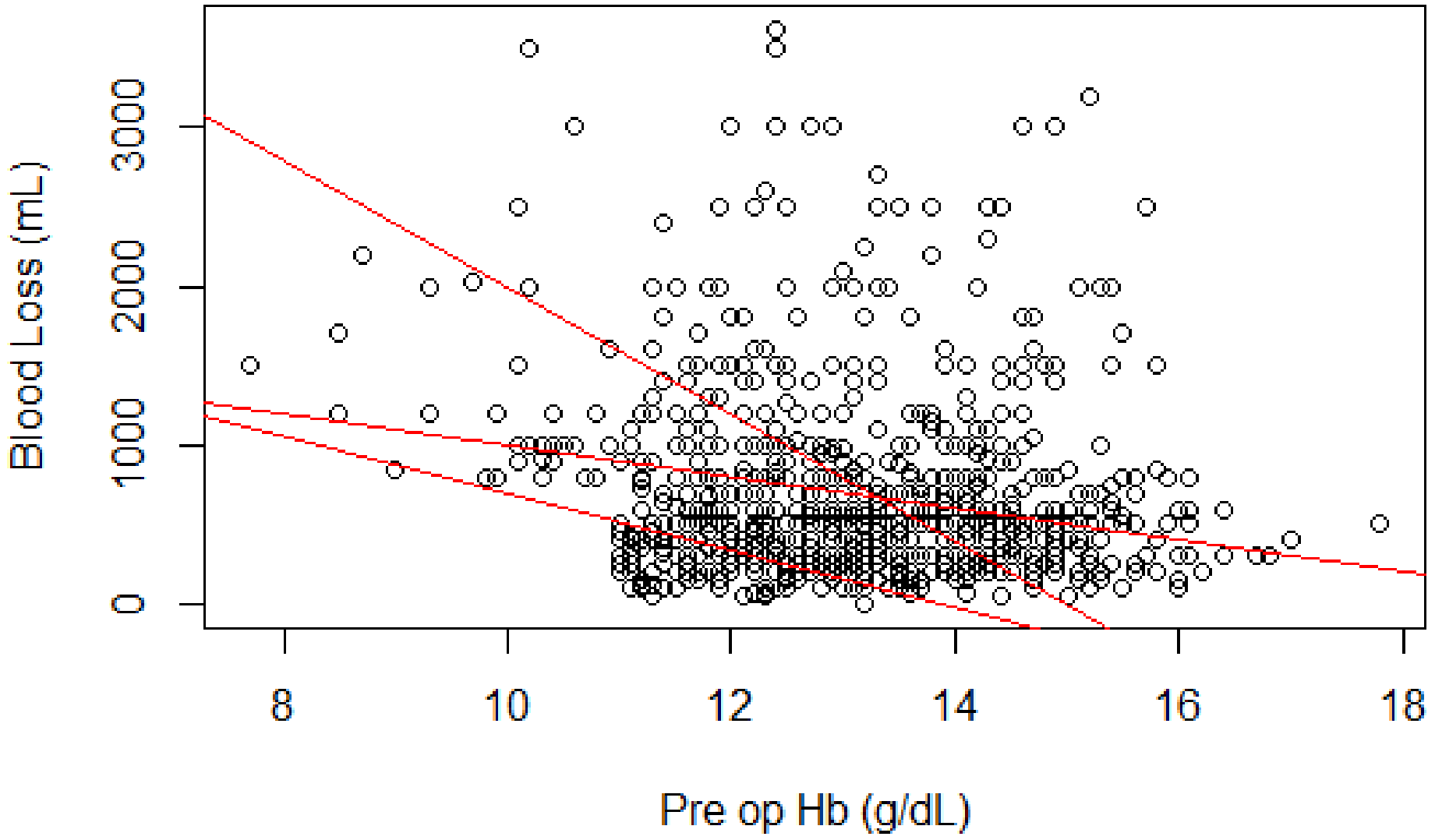
When Y is continuous

- No good examples in oncology
- Our interesting outcomes are either binary or censored
- But regression is best taught with a continuous outcome
- So we will spend this lecture on using a somewhat artificial example
- Pre-operative hemoglobin vs surgical blood loss
  - I doctored the data a little to make my points so do not conclude anything medical from this analysis


Blood Loss (mL)
0
1000
2000
3000
8
10
12
14
16
18
Pre op Hb (g/dL)

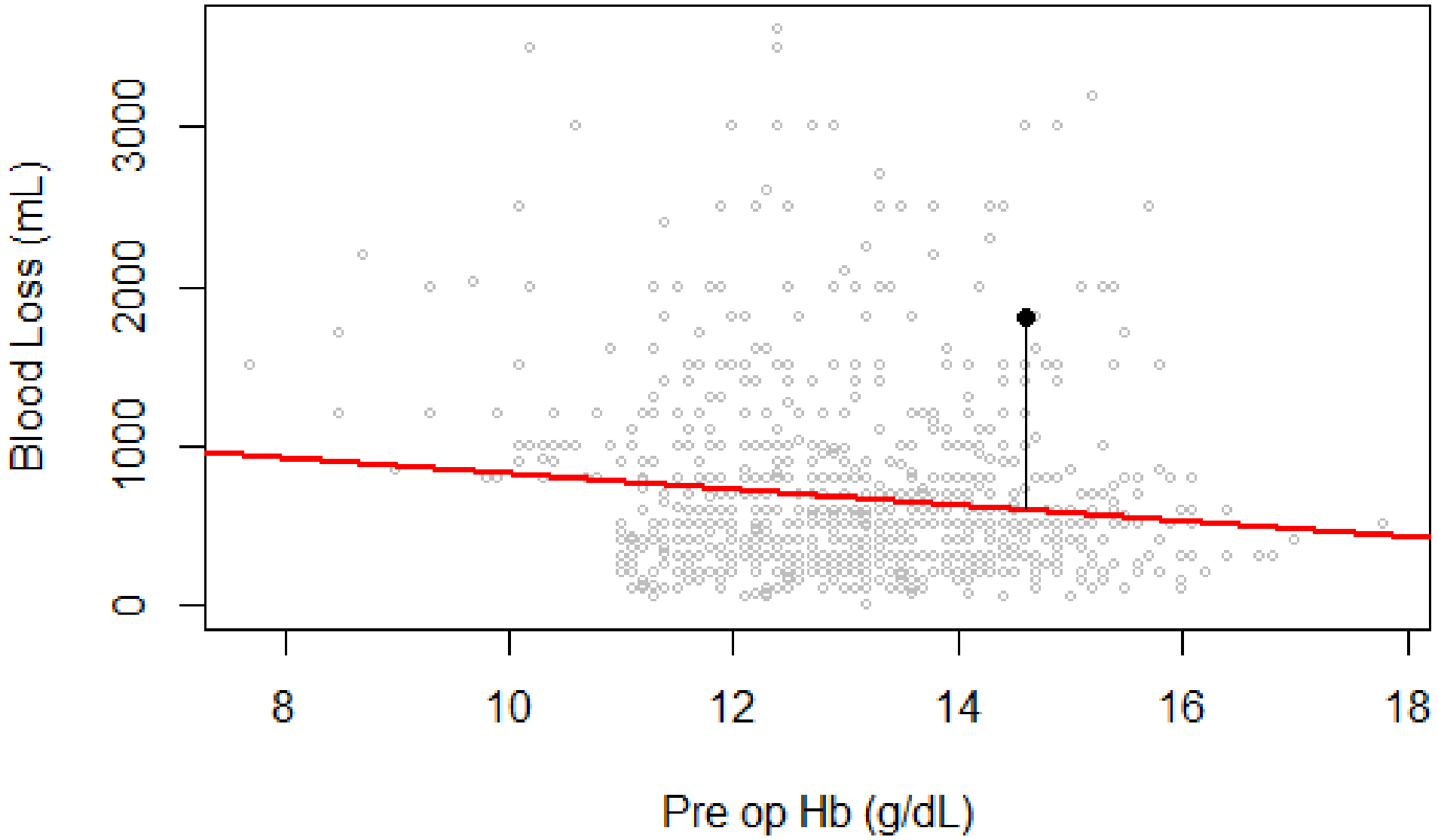
How to find the best fitting line

- Never mind (for the purposes of this class) that the general trend in this scatterplot does not look like a line or much of anything
- Imagine yourself (next set of slides) trying many different lines
- Which one fits best?
- What does “best fit” mean?


Blood Loss (mL)
0
1000
2000
3000
8
10
12
14
16
18
Pre op Hb (g/dL)

Best fit

- In this context of continuous Y, the most commonly used best fit criterion is “minimize the deviations from the fitted line”
- Deviations: the distance between a point and the fitted line
- Imagine going through this
  - For every possible line going through this data set
  - Calculate the deviation for each point
  - Add them up
  - Choose the line with the smallest sum of deviations
- Known as the least squares method


Blood Loss (mL)
0
1000
2000
3000
8
10
12
14
16
18
Pre op Hb (g/dL)

Least Squares

- We do not really try all the lines, there is a formula that gives a and b for a given set of points
- Least Squares is the oldest method for estimating regression coefficients
- Widely used
- DOES NOT generalize to other outcomes (binary, censored)
- We will not spend any appreciable time on it

Residuals

- Y – pred(Y) are called residuals
- Can you see on the previous graph that residual is the same as deviation?
- Residuals are very important in least squares regression
- Just like least squares does not generalize well to other outcomes
- But the idea of a deviation generalizes and we will continue to use that concept

Goodness of fit

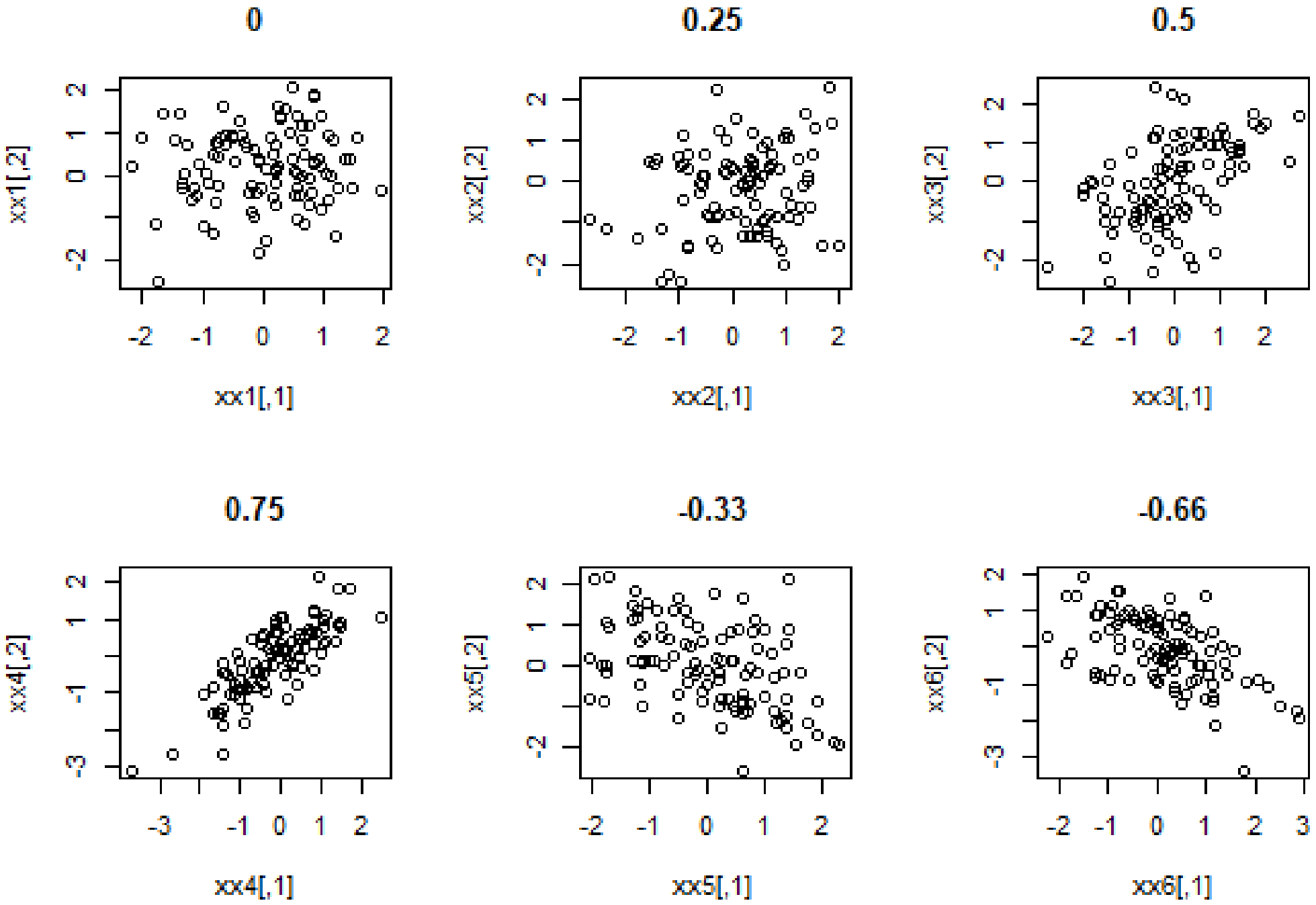
- Best fit does not mean good fit
- Can we quantify how well the best fit line fits the data?
- When Y is continuous we use R^2
- R^2 does not generalize well either but commonly reported used when someone uses least squares
- Between 0 and 1: higher values indicating better fit

Correlation

- An everyday word with a precise meaning in statistics
- Correlation is the (signed) square root of R2
- Sign comes from the sign of b (slope)
- Sleight of hand: parameter or estimate?

More on correlation

- Actually a parameter but its definition requires more math than we want here
- As most parameters it can be estimated
- Square root of R2 is one way to estimate: Pearson correlation
- Many other ways: Spearman (rank), Kendall’s tau, ….
- Does not generalize well either, but comes in handy as a concept and also in variable selection

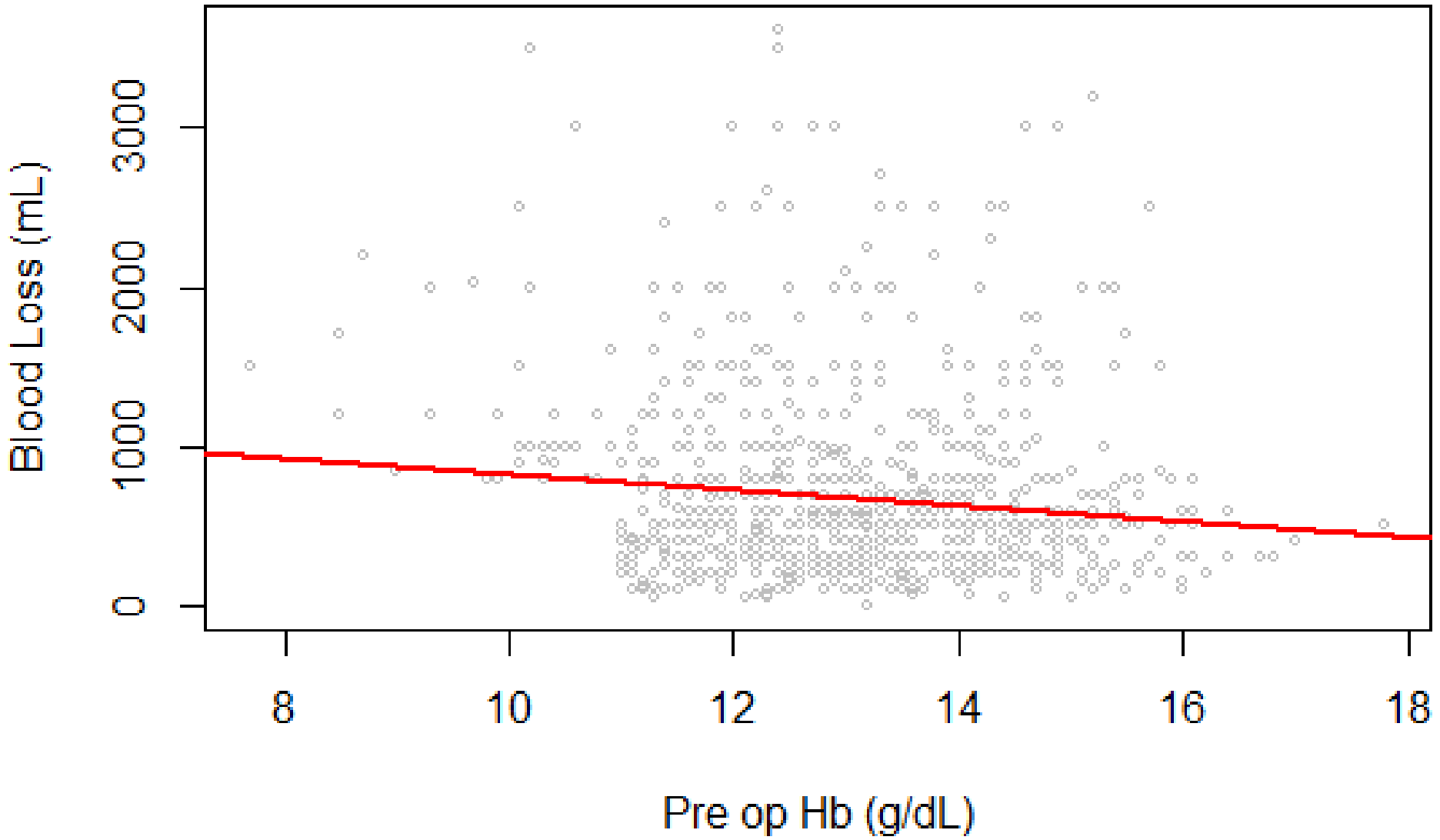

0
xx1[,2]
xx1[,1]
0.25
xx2[,2]
xx2[,1]
0.5
xx3[,2]
xx3[,1]
0.75
xx4[,2]
xx4[,1]
-0.33
xx5[,2]
xx5[,1]
-0.66
xx6[,2]
xx6[,1]

Back to the Example

- How to report a regression analysis?
- Report point estimates of a and b, along with confidence intervals and p-values for testing if the underlying coefficient is 0
- Report R2

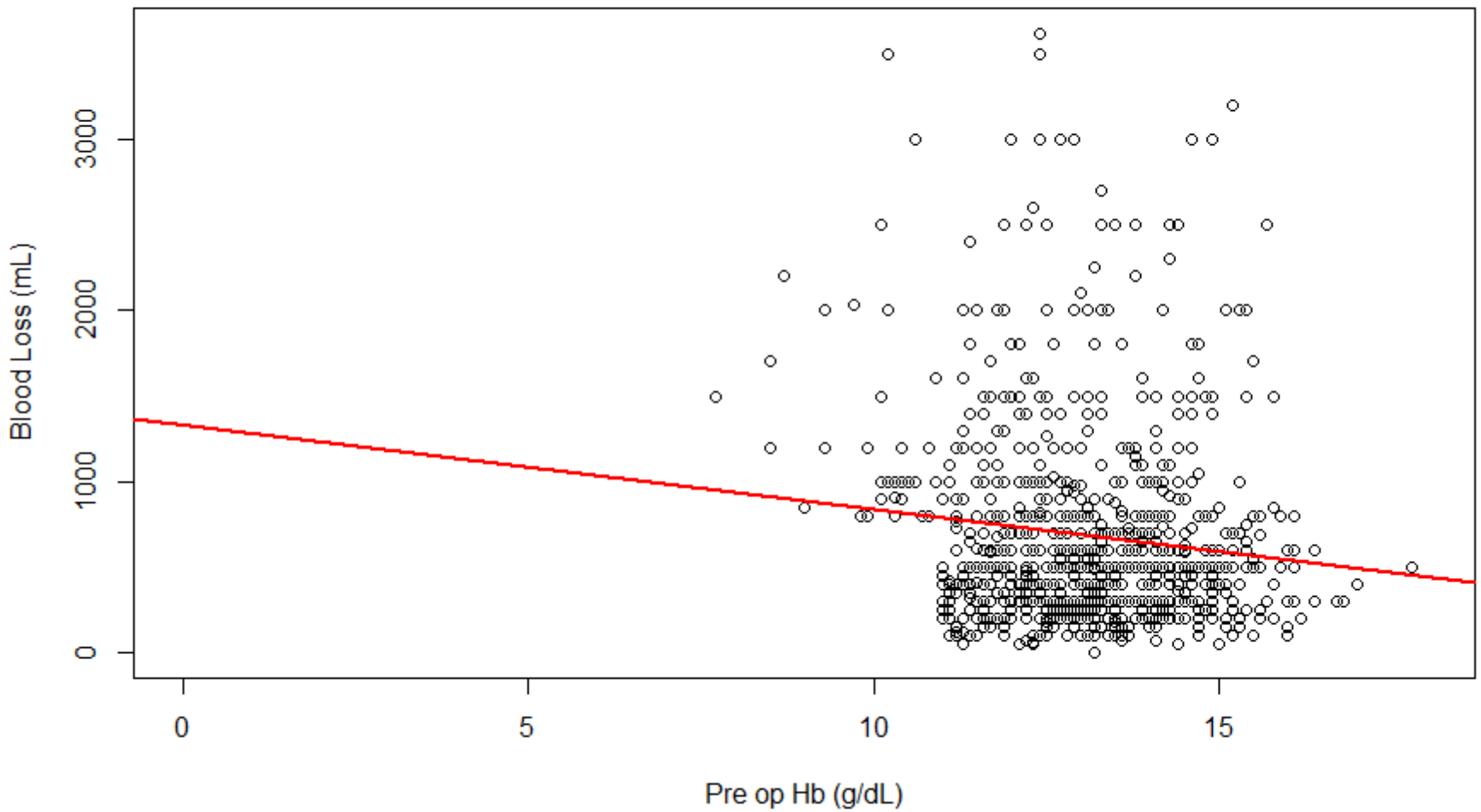
Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	1330.35	966.20	1694.51	?
Hb	-49.41	-77.10	-21.72	?
R2	?			

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	1330.35	966.20	1694.51	<0.0001
Hb	-49.41	-77.10	-21.72	0.0005
R2	?			


Blood Loss (mL)
0
1000
2000
3000
8
10
12
14
16
18
Pre op Hb (g/dL)

Intercept and Slope

- Pred(Y) = a + b*X
  - When X = 0 → Pred(Y) = a + b*0 = a
  - Intercept (a) is the point where the line crosses the vertical axis (Y value for X = 0)
- Slope
  - For X: Pred(Y) = a + b*X
  - For X+1: Pred(Y) = a + b*(X+1)
  - The difference in Pred(Y) when X goes up by one unit is: a + b*(X+1) – (a + b*X) = a + b*X + b – a – b*x = b
  - Slope is the change in Y when X changes one unit


Blood Loss (mL)
0
1000
2000
3000
0
5
10
15
Pre op Hb (g/dL)

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	1330.35	966.20	1694.51	<0.0001
Hb	-49.41	-77.10	-21.72	0.0005
R2	0.012 !!!!!!			

Multivariable (multivariate) regression

- We have more than one X
- And some of these X's can be continuous, some can be categorical
- I will first add a categorical variable to the mix
- It is very very very important to write the regression equation every time
- X1 is continuous, X2 is binary
- Pred(Y) = a + b1*X1 + b2*X2

Continuous and binary variables in regression

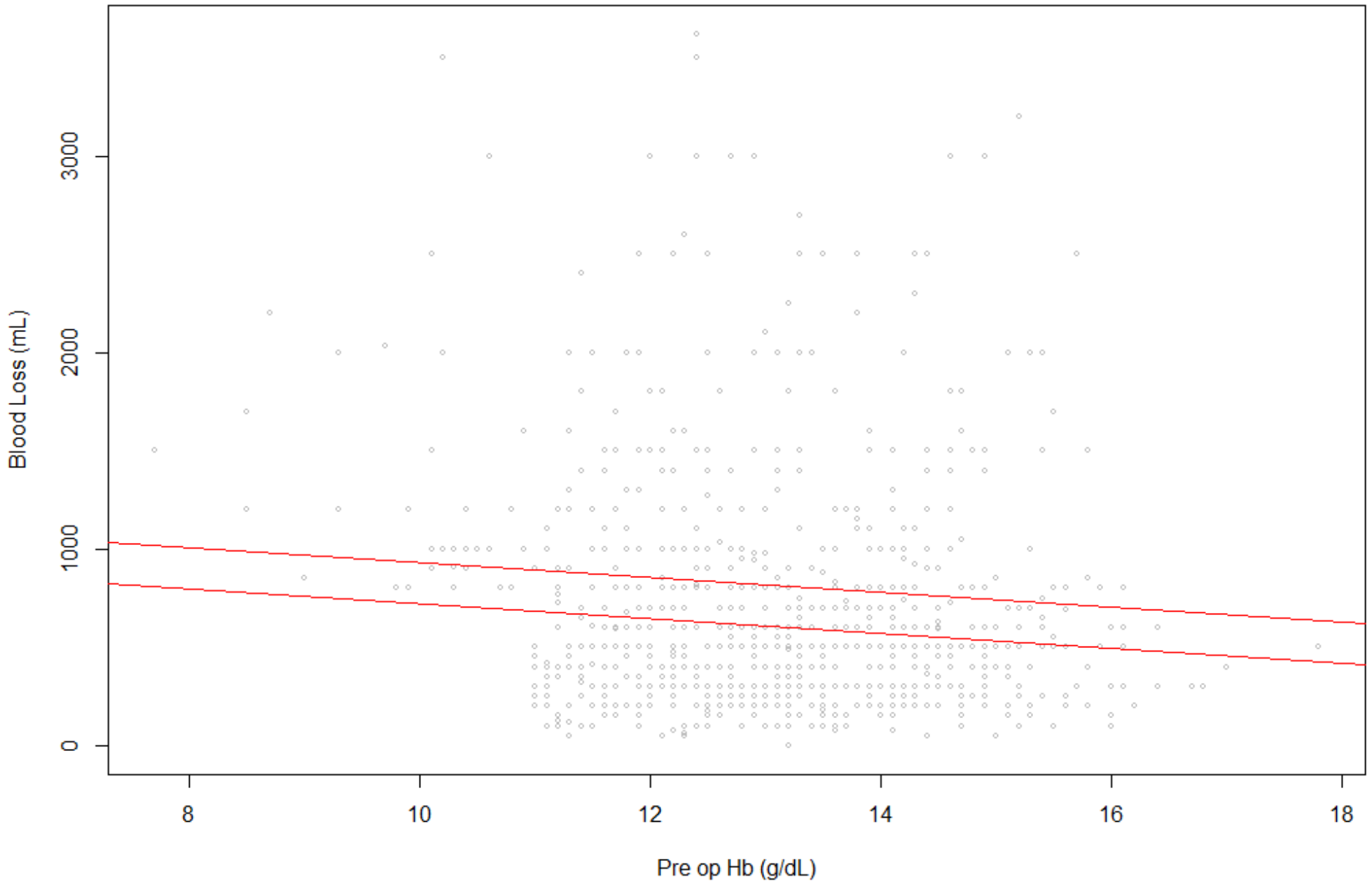
- Pred(Y) = a + b1*X1 + b2*X2
- Remember X2 is binary, so either 0 or 1
- When X2 = 0
  - Pred(Y) = a + b1*X1
- When X2 = 1
  - Pred(Y) = a + b1*X1 + b2
  - Pred(Y) = (a + b2) + b1*X1

Continuous and binary variables in regression

- Pred(Y) = a + b1*X1 + b2*X2
- Remember X2 is binary, so either 0 or 1
- When X2 = 0
  - Pred(Y) = a + b1*X1
- When X2 = 1
  - Pred(Y) = a + b1*X1 + b2
  - Pred(Y) = (a + b2) + b1*X1

Continuous and binary variables in regression

- Pred(Y) = a + b1*X1 + b2*X2
- X2 = 0 → Pred(Y) = a + b1*X1
- When X2 = 1 → Pred(Y) = (a + b2) + b1*X1
- Same slope, different intercept
- Parallel lines


Blood Loss (mL)
0
1000
2000
3000
8
10
12
14
16
18
Pre op Hb (g/dL)

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	1103.69	734.50	1472.89	<0.0001
Hb	-37.93	-65.56	-10.29	0.0071
Binary Variable	206.26	127.55	284.98	<0.0001
R2	0.012 !!!!!!			

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	1103.69	734.50	1472.89	<0.0001
Hb	-37.93	-65.56	-10.29	0.0071
Binary Variable	206.26	127.55	284.98	<0.0001
R2	0.041			

Pred(Y) for X2 = 0 → 1103.69 – 37.93*Hb

Pred(Y) for X2 = 1 → 1309.95 – 37.93*Hb

What if we want different slopes

- We need to use an interaction
- Pred(Y) = a + b1*X1 + b2*X2 + b12*X1*X2
- When X2 = 0
  - Pred(Y) = a + b1*X1
- When X2 = 1
  - Pred(Y) = (a + b2) + (b1 + b12)*X1
- Different intercepts and slopes

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	864.78	501.49	1428.06	<0.0001
Hb	-27.44	-62.23	7.35	0..122
Binary Variable	574.90	-171.88	1321.68	0.131
Hb*Binary Variable	-28.44	-85.73	28.86	0.330

Pred(Y) for X2 = 0 → 864.78 – 27.44*Hb

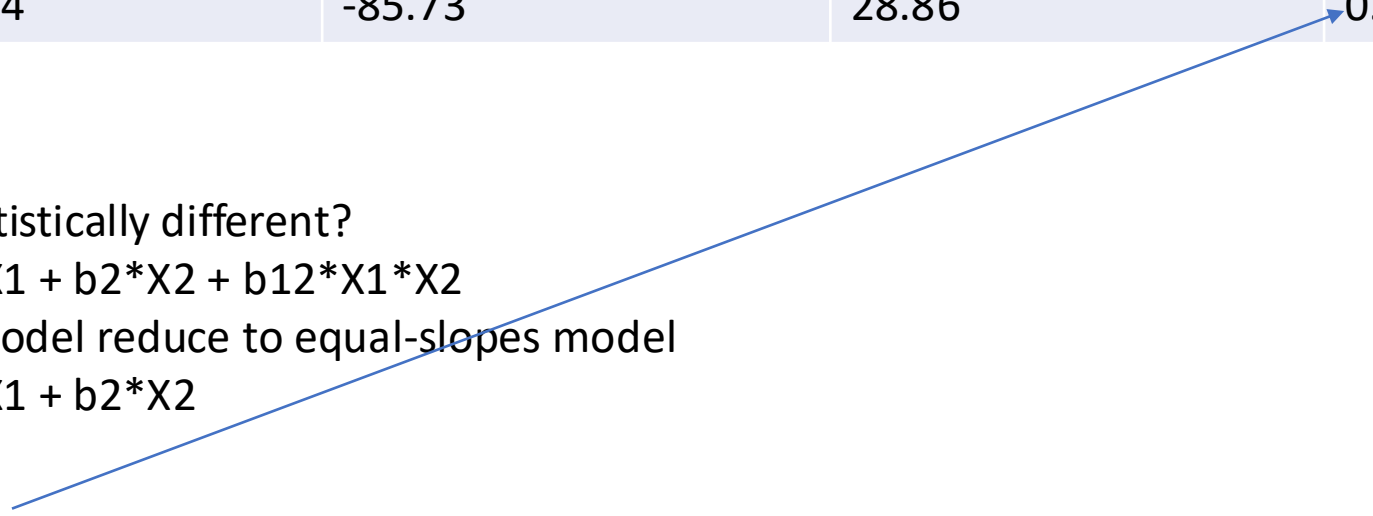
Pred(Y) for X2 = 1 → 1449.68 – 55.88*Hb

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	864.78	501.49	1428.06	<0.0001
Hb	-27.44	-62.23	7.35	0..122
Binary Variable	574.90	-171.88	1321.68	0.131
Hb*Binary Variable	-28.44	-85.73	28.86	0.330

Are the slopes statistically different?
Pred(Y) = a + b1*X1 + b2*X2 + b12*X1*X2
When does this model reduce to equal-slopes model
Pred(Y) = a + b1*X1 + b2*X2

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	864.78	501.49	1428.06	<0.0001
Hb	-27.44	-62.23	7.35	0..122
Binary Variable	574.90	-171.88	1321.68	0.131
Hb*Binary Variable	-28.44	-85.73	28.86	0.330

Are the slopes statistically different?
Pred(Y) = a + b1*X1 + b2*X2 + b12*X1*X2
When does this model reduce to equal-slopes model
Pred(Y) = a + b1*X1 + b2*X2
When b12 = 0!
So test for b12 = 0



Adding more variables

- Once you have two continuous predictors (X's) in the model it becomes difficult to visualize, more than 2 impossible
- So we need to "imagine" points in higher dimensional spaces that only exist in our minds (and in mathematical representations)
- But the idea of a "slope", "residual" etc applies

Parameter	Estimate	95% Lower	95% Upper	P
Intercept	1303.07	501.49	1428.06	<0.001
Hb	-37.52	-62.23	7.35	0.008
Binary Variable	207.473	-171.88	1321.68	<0.001
Age	-3.48	-85.73	28.85	0.018
R2	0.046			

Interpretation of coefficients

- Change in Pred(Y) corresponding to one unit change in X keeping others constant
- Binary: one unit means going from 0 to 1
- Continuous: depends on units of measurement
- Age: Take two patients with same Hb and X2, Pred(Y) goes down by 3.48 when age goes up by one year
- X2: Take two patients with same Hb and Age. The one with X2 = 1 has Pred(Y) higher by 207.47

Take a look at all the different models we fit

- Estimates jumped around quite a bit
- Model results are usually sensitive to which variables are included and excluded
- Variable selection is the Achilles heel of multivariate regression
- More on this in next lectures