Clinical Epidemiology

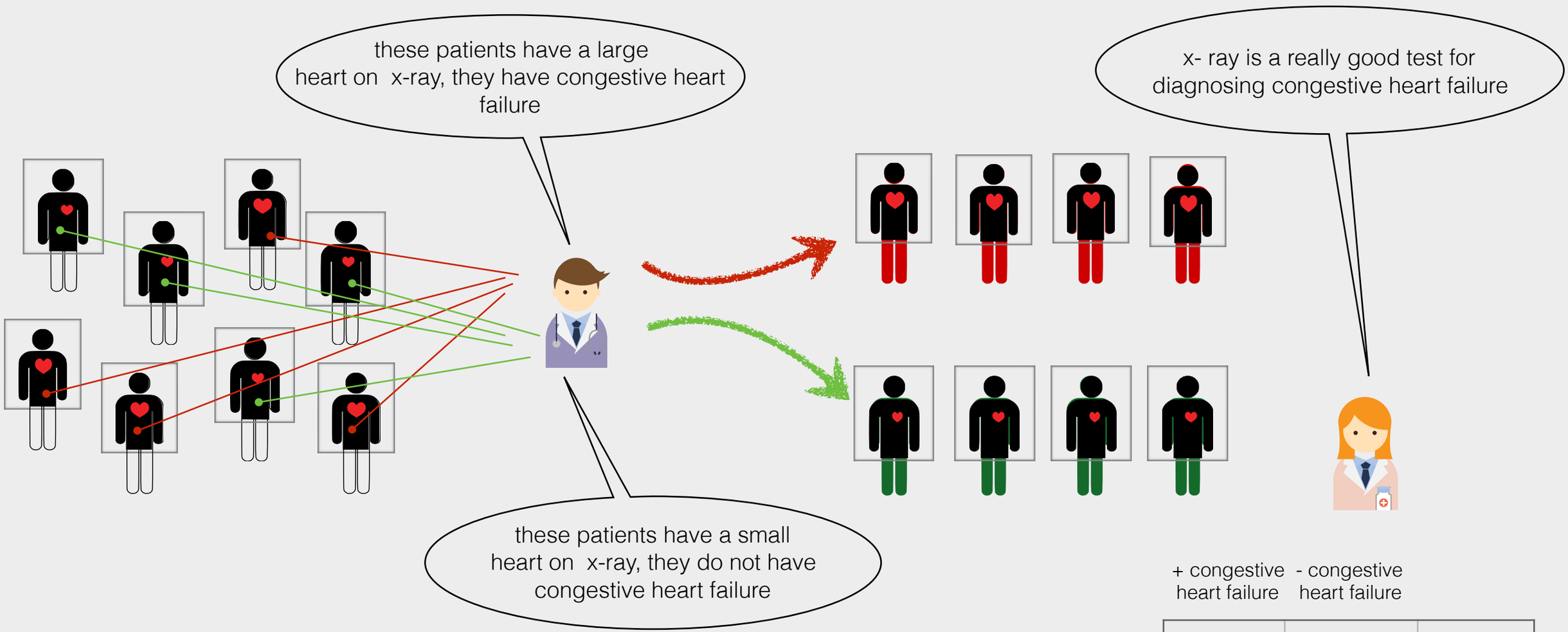
Lecture notes 2015 - Part II

Lecture 4
Studies of diagnostic tests

Bias in studies of diagnostic tests

Incorporation bias

the result of the index test is used to establish final diagnosis



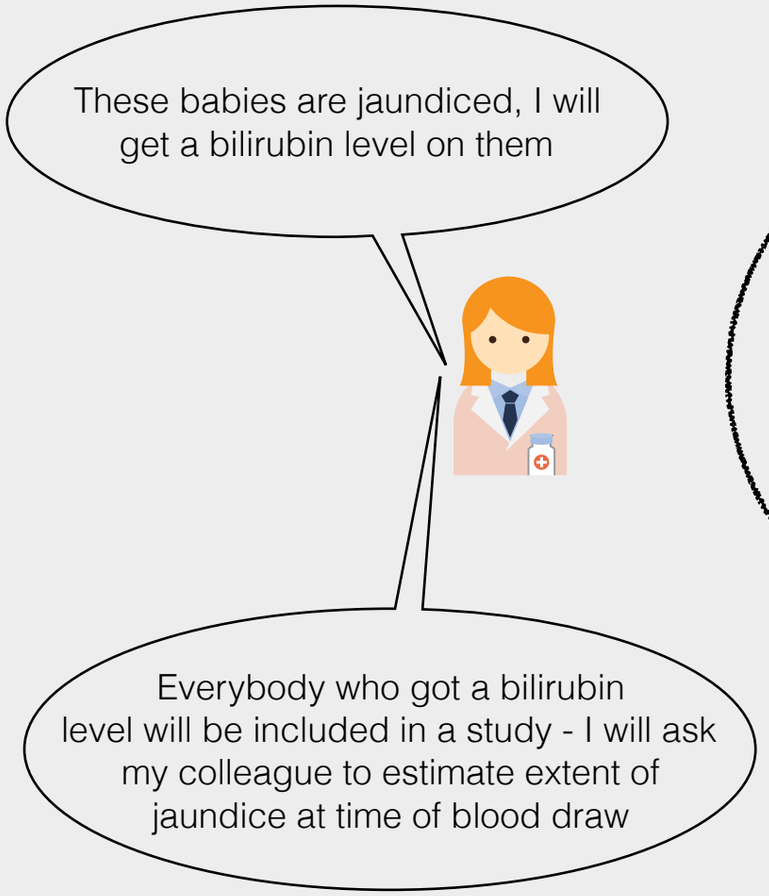
	+ congestive heart failure	- congestive heart failure	
big heart	**A**	B	A+B
small heart	C	**D**	C+D
	A+C	B+D	A+B+C+D

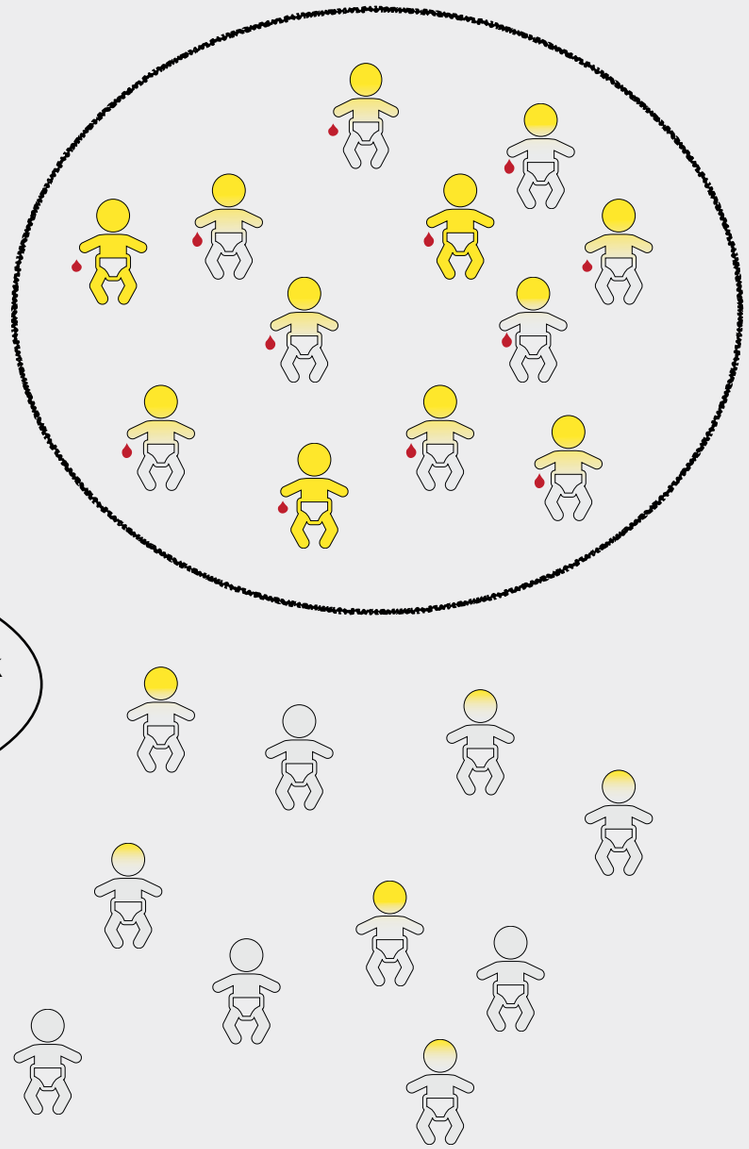
overrepresentation of cell A and D falsely elevates sensitivity (A/(A+C)) and specificity (D/(B+D))

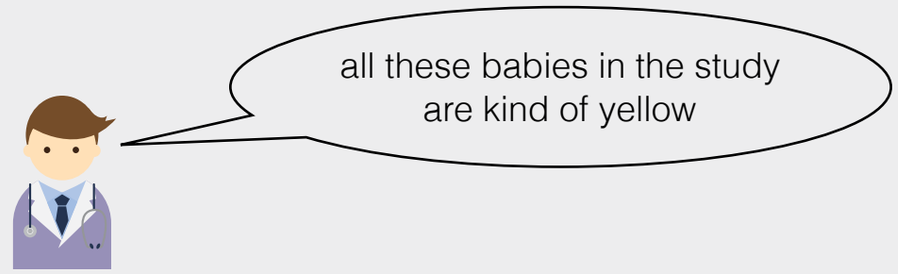
Bias in studies of diagnostic tests

Verification bias

subjects with positive index test are more likely to be included in the study AND to get gold standard







study population

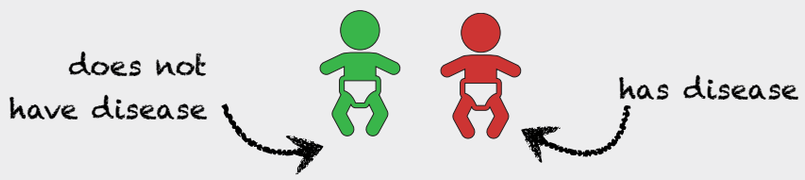
not a representative sample of the population

	high bilirubin level	low bilirubin level	
clinically yellow	**A**	**B**	A+B
clinically not yellow	C	D	C+D
	A+C	B+D	A+B+C+D

underrepresentation of cells C and D leads to falsely elevated sensitivity (A/(A+C)) and falsely low specificity (D/(B+D))

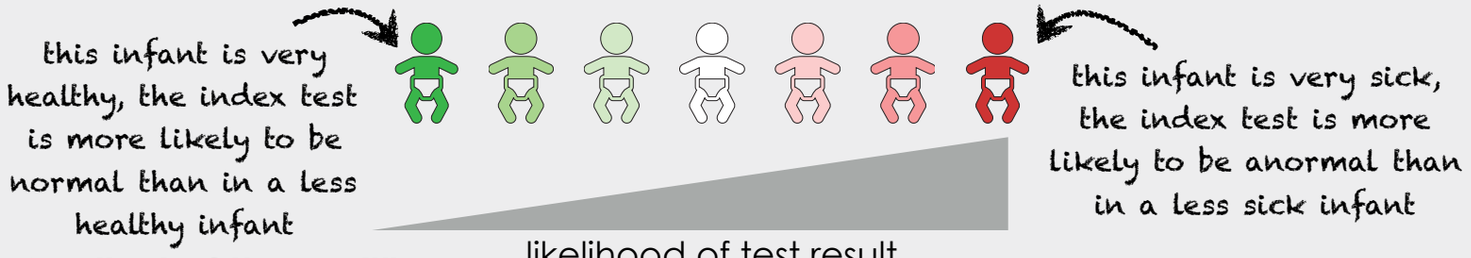
Bias in studies of diagnostic tests

Simplified assumptions



We think of a patient normally as either having the disease or not having the disease, however, in reality there is a **spectrum of health and disease**

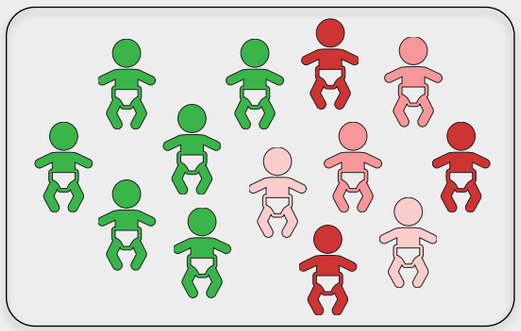
Reality



Spectrum of Health and Disease

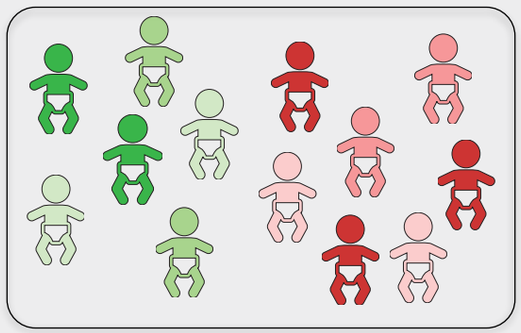
Spectrum bias

sensitivity and specificity depend on spectrum of health and disease in sample population



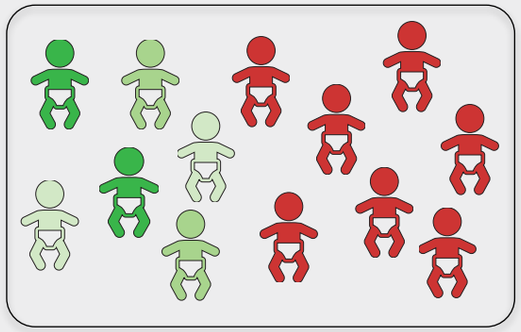
This population sample contains the entire spectrum of disease but only the "wellest of the well" without the disease

specificity of test is falsely high since "wellest of the well" without disease are more likely to have negative test result



This population sample contains the entire spectrum of disease and non-disease

sensitivity and specificity of test are accurate



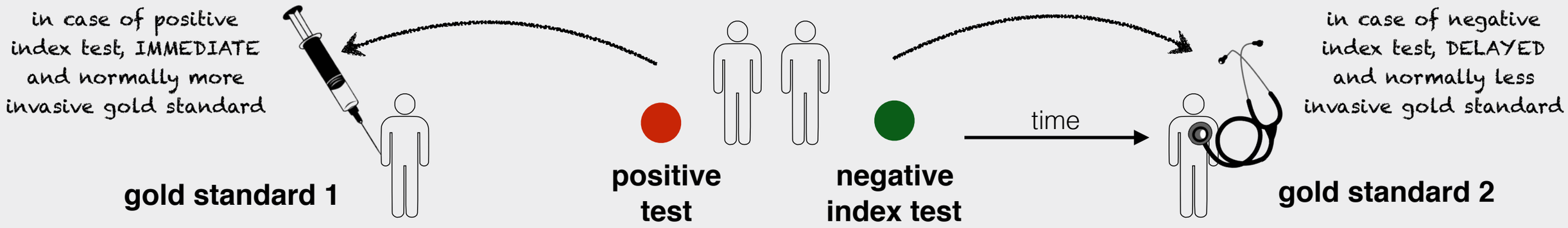
This population sample contains only the "sickest of the sick" with the disease and the entire spectrum of non-disease

sensitivity of test is falsely high since "sickest of the sick" with the disease have more likely to have positive test result

Bias in studies of diagnostic tests

Double Gold standard bias

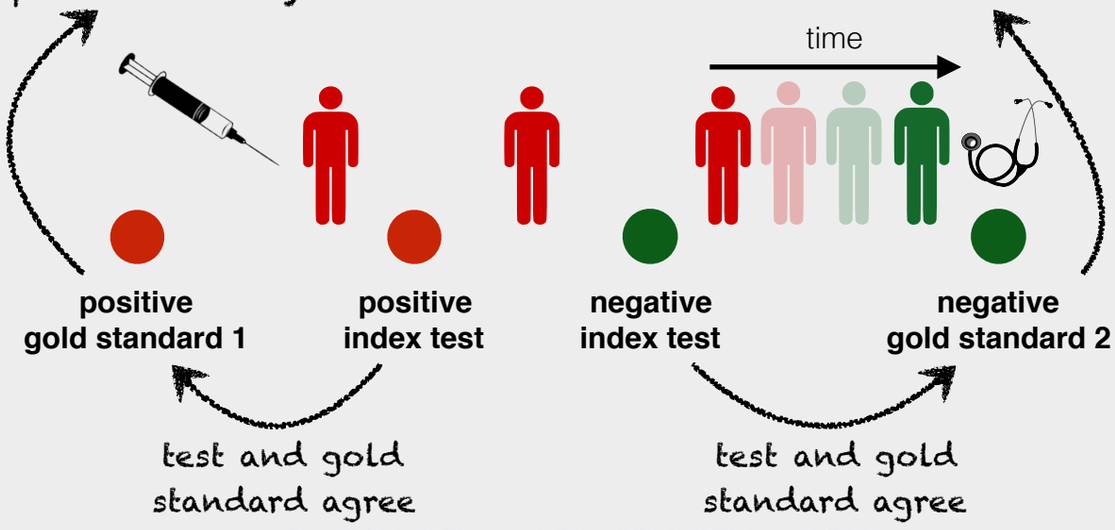
different gold standard is applied depending on result of index test



Spontaneously resolving disease

immediate invasive gold standard diagnoses the disease because patient currently has disease

delayed clinical follow up is negative because disease resolved



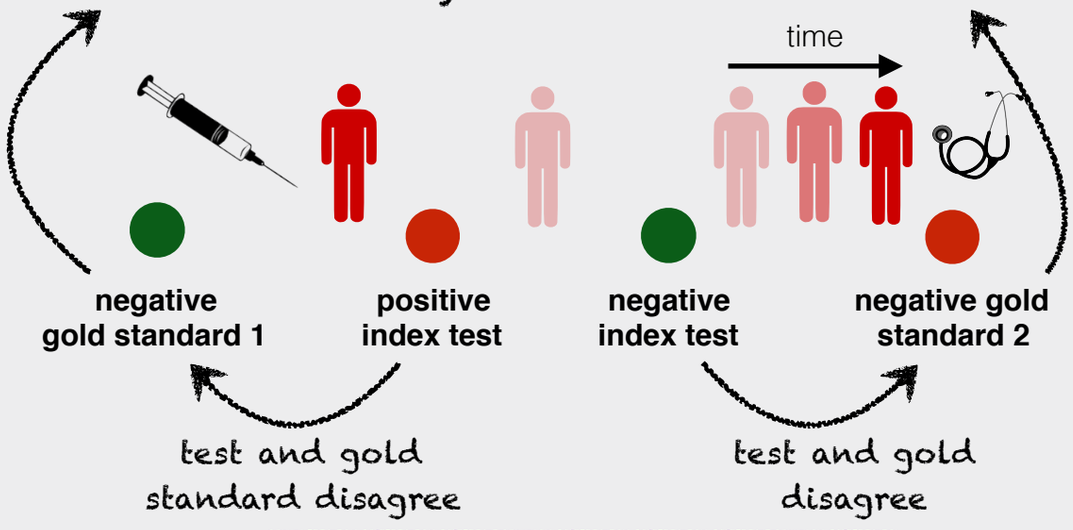
sensitivity and specificity are both falsely high

TEST CANNOT BE WRONG

Newly occurring or detectable disease

immediate invasive gold standard is negative because disease is not detectable yet

delayed clinical follow up is positive because disease becomes detectable



sensitivity and specificity are both falsely low

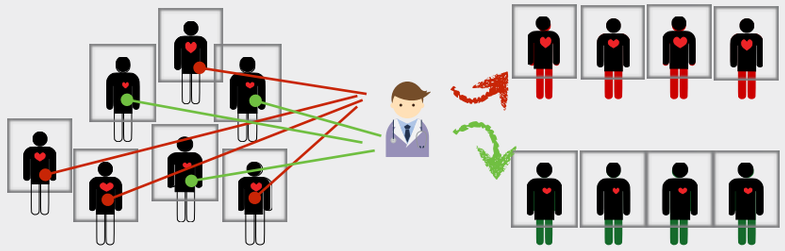
TEST CANNOT BE RIGHT

Bias in studies of diagnostic tests

Summary

Incorporation bias

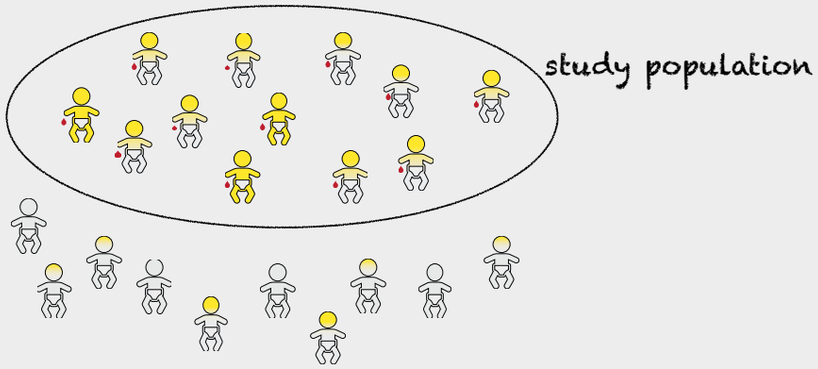
the result of the index test is used to establish final diagnosis



sensitivity ↑ specificity ↑

Verification bias

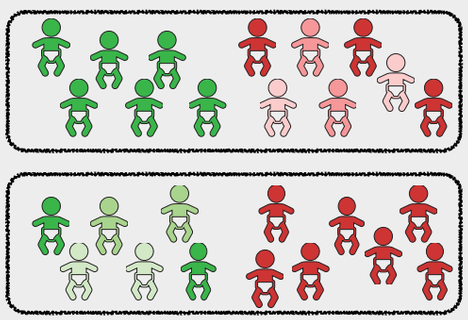
subjects with positive index test are more likely to be included in the study AND to get gold standard



sensitivity ↑ specificity ↓

Spectrum bias

sensitivity and specificity depend on spectrum of health and disease in sample population



"wellest of the well"

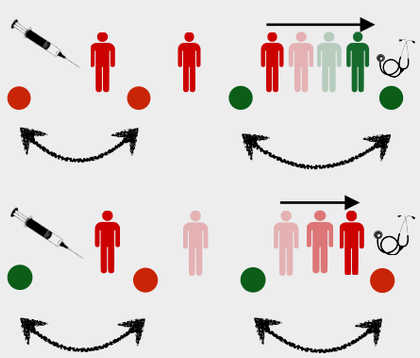
sensitivity → specificity ↑

"sickest of the sick"

sensitivity ↑ specificity →

Double Gold Standard bias

different gold standard is applied depending on result of index test



spontaneously resolving disease

sensitivity ↑ specificity ↑

newly occurring or detectable disease

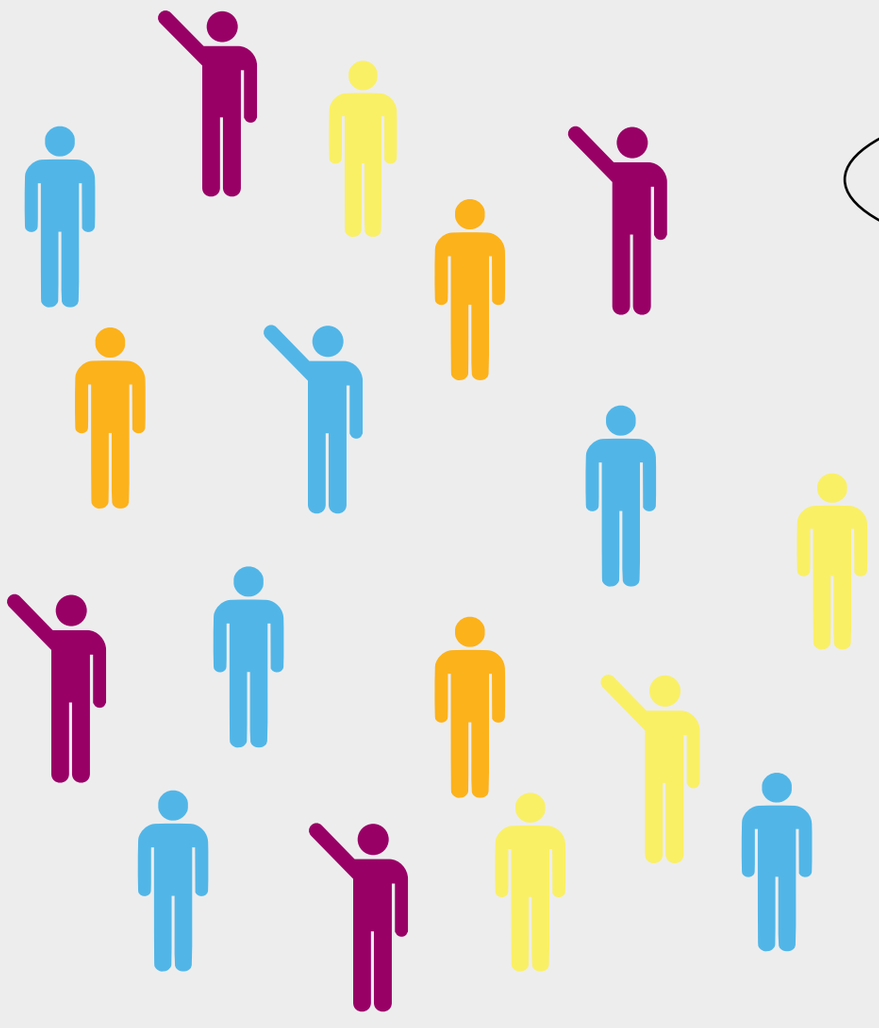
sensitivity ↓ specificity ↓

Lecture 5
Screening tests

Screening tests

Volunteer bias

people who volunteer for screening differ from those who not



people who volunteer might be more educated, more interested, healthier, etc., this group is not representative of the population

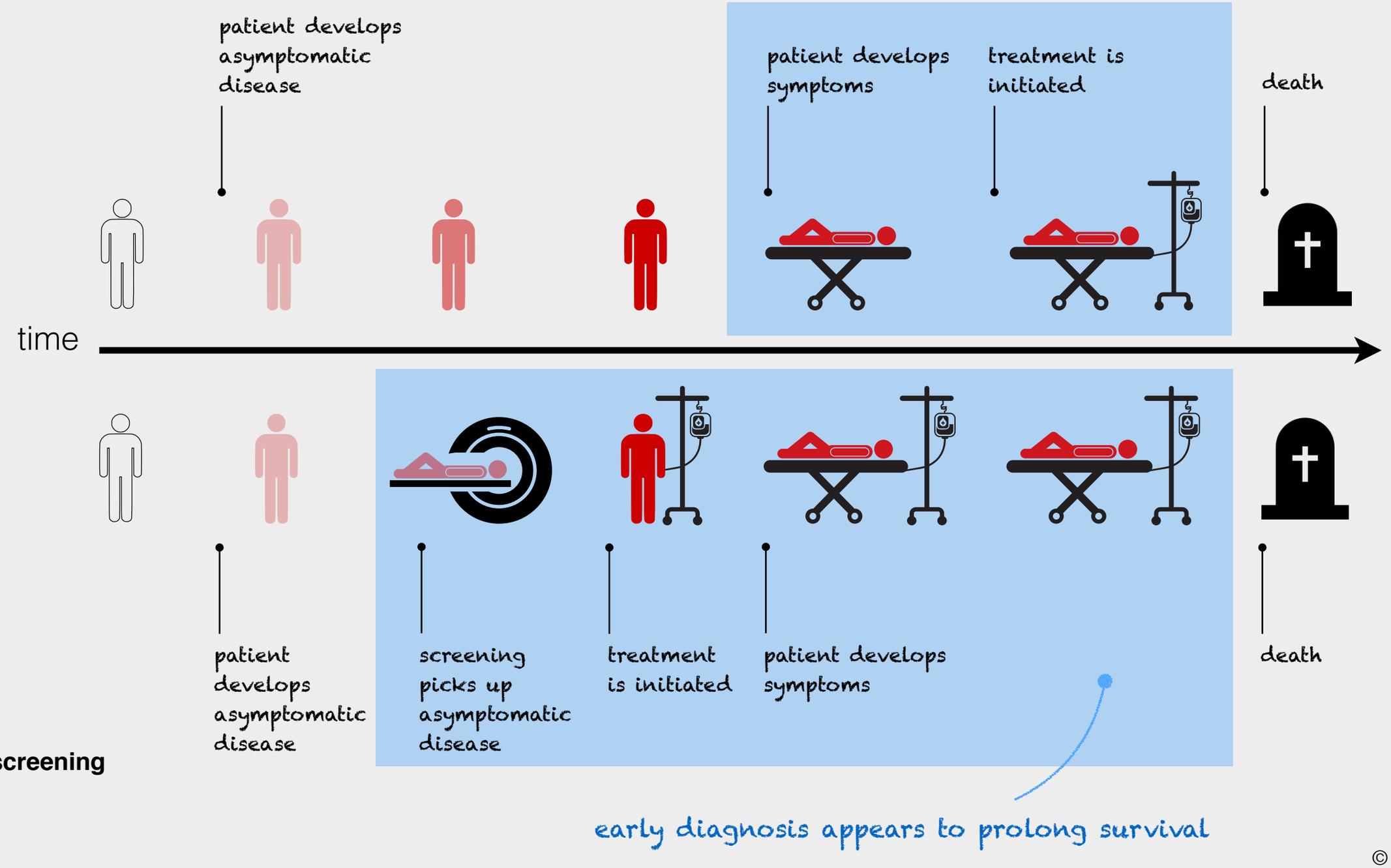
e.g. woman who volunteer for mammograms have lower heart disease rates

Screening tests

Lead time bias

Screening identifies disease during a latent period before it becomes symptomatic If survival is measured from time of diagnosis, screening will always improve survival even if there is no benefit of early diagnosis

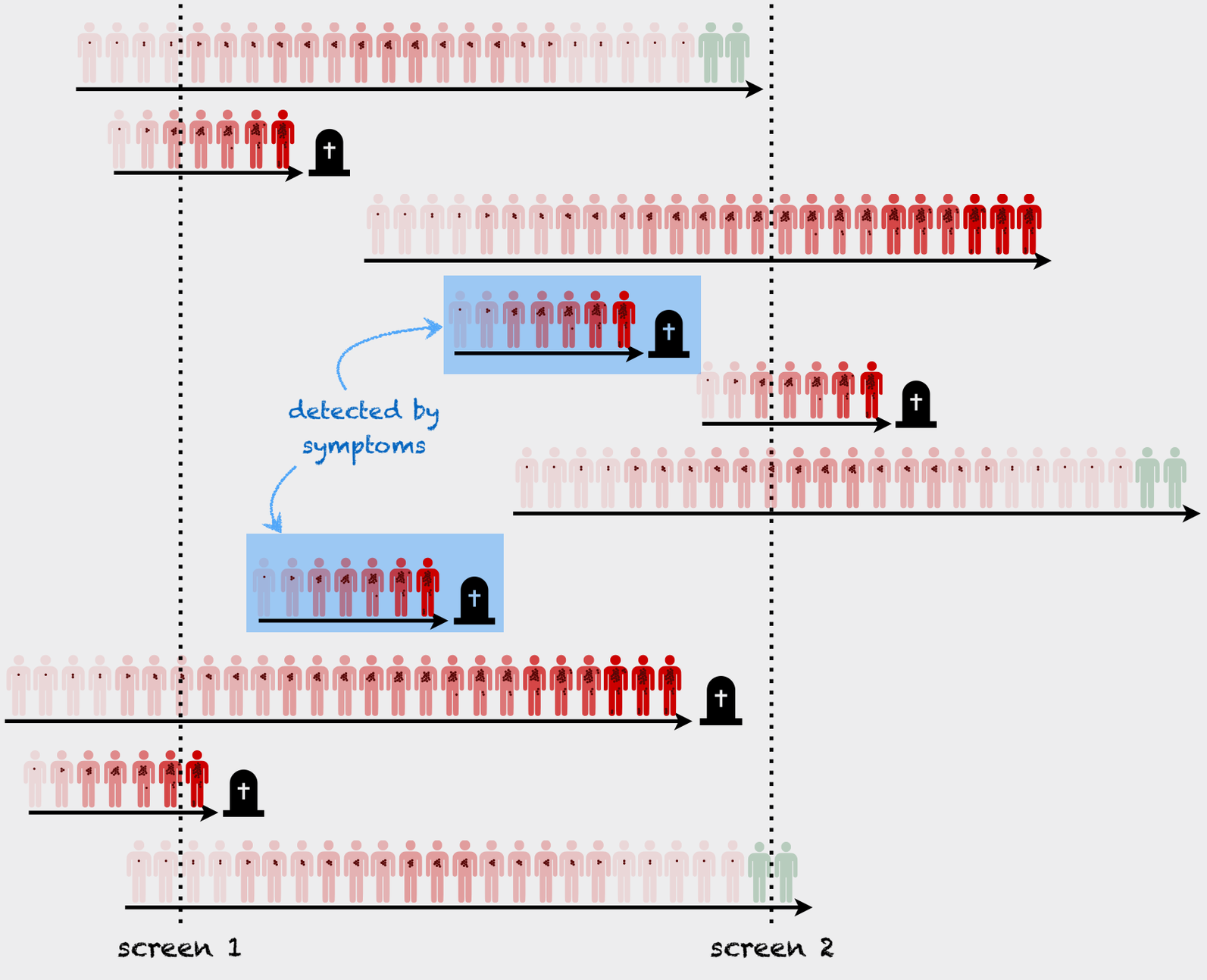
no screening



Screening tests

Length bias

Screening picks up prevalent disease, benign courses of disease have a greater prevalence



Prevalence = incidence x duration

Slower growing tumors have greater duration in the pre-symptomatic phase, therefore greater prevalence

Screening picks up *prevalent* cases, which therefore will be disproportionately slower growing

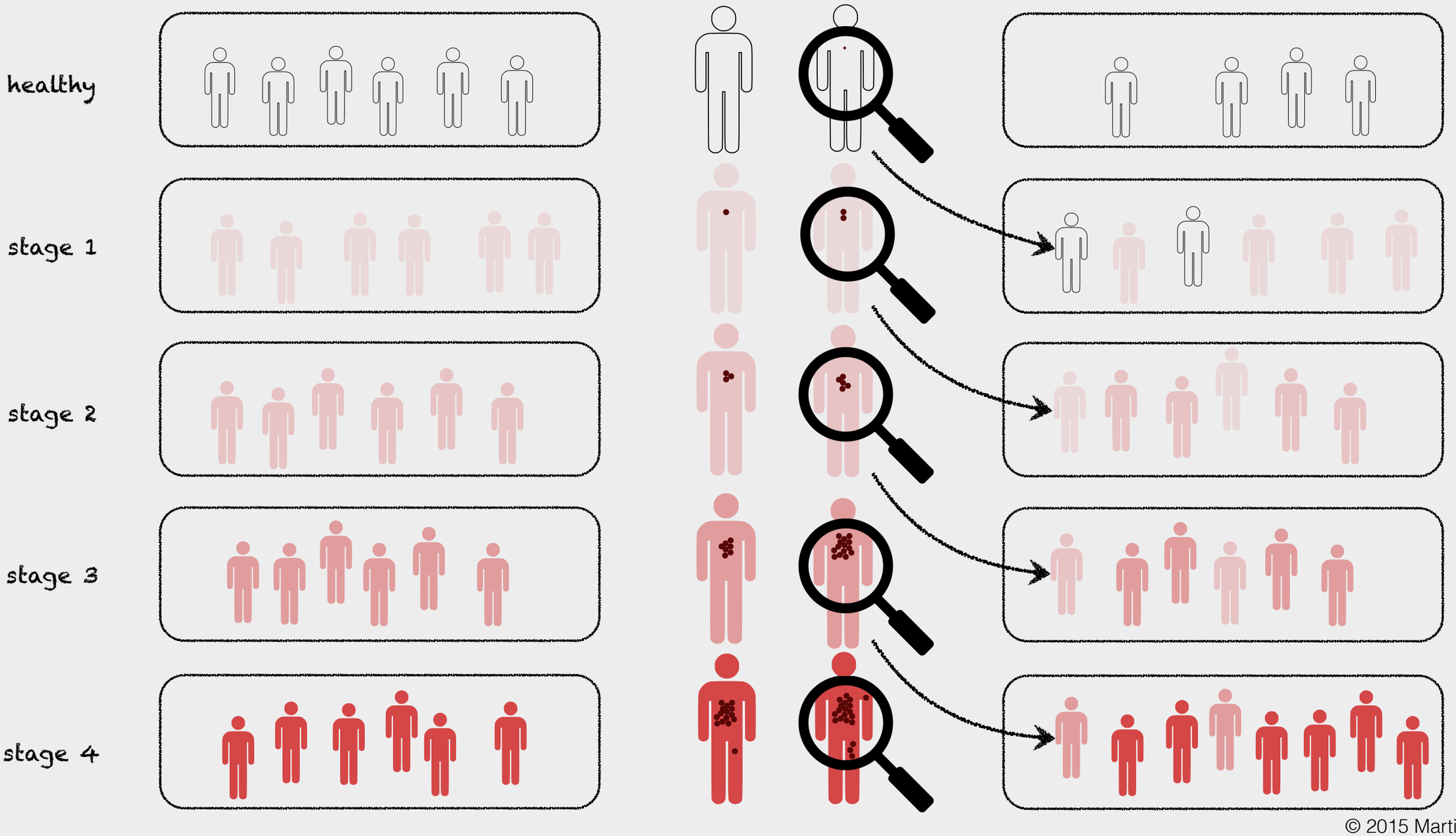
Screen 1 and **screen 2** detect a total of 8 tumors: 5/8 have a good prognosis, 3/8 have a bad prognosis.

Symptoms lead to detection of 2 tumors - both of them had a bad prognosis.

Screening tests

Stage migration bias

Newer tests classify people into more advanced stages. Stage specific survival improves.

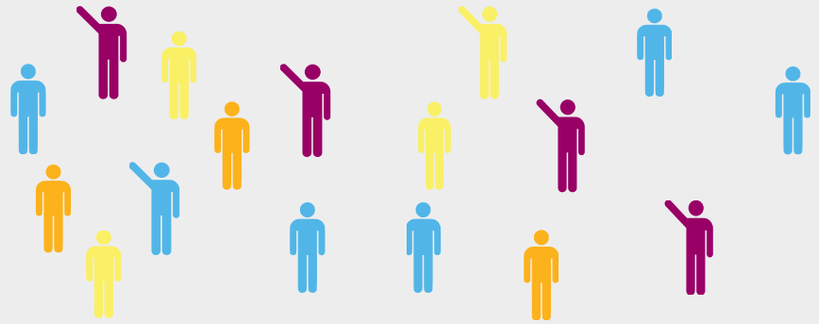


Screening tests

Summary

Volunteer bias

people who volunteer for screening differ from those who do not

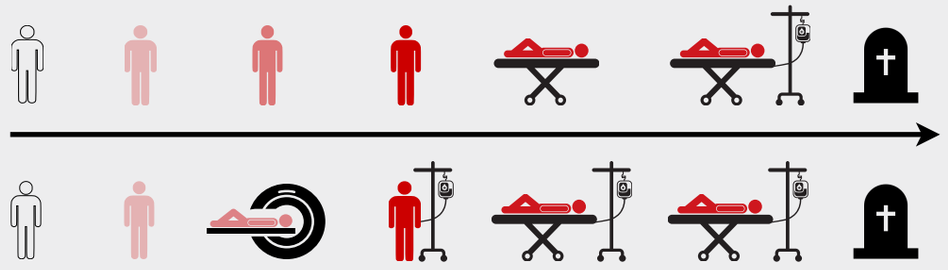


how to avoid

- ✳ randomize people to screened and unscreened
- ✳ otherwise, try to control for factors (confounders) associated with both screening and outcome

Lead time bias

screening identifies disease during a latent period before it becomes symptomatic - screening appears to improve survival

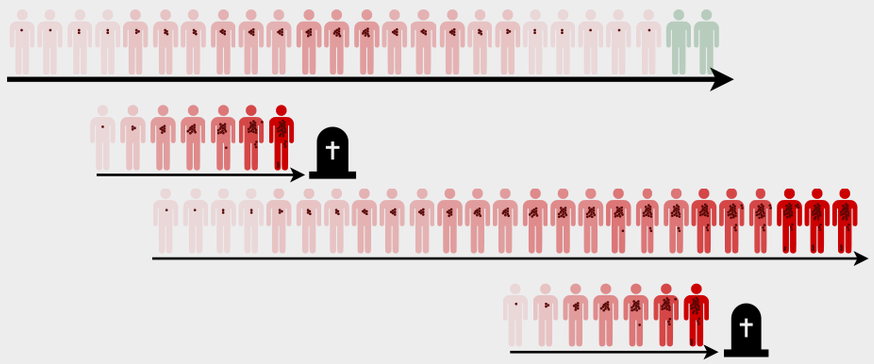


how to avoid

- ✳ measure mortality and not survival
- ✳ count from date of randomization

Length bias

screening picks up prevalent disease and benign courses of disease have a greater prevalence

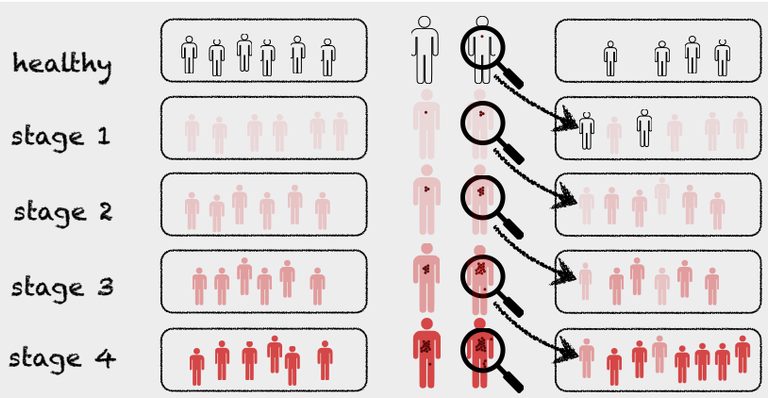


how to avoid

- ✳ compare mortality in the ENTIRE screened group to the ENTIRE unscreened group

Stage migration bias

newer tests classify people into more advanced stages - stage specific survival improves.



how to avoid

- ✳ beware of stage migration in any stratified analysis
- ✳ check OVERALL survival in screened vs. unscreened group

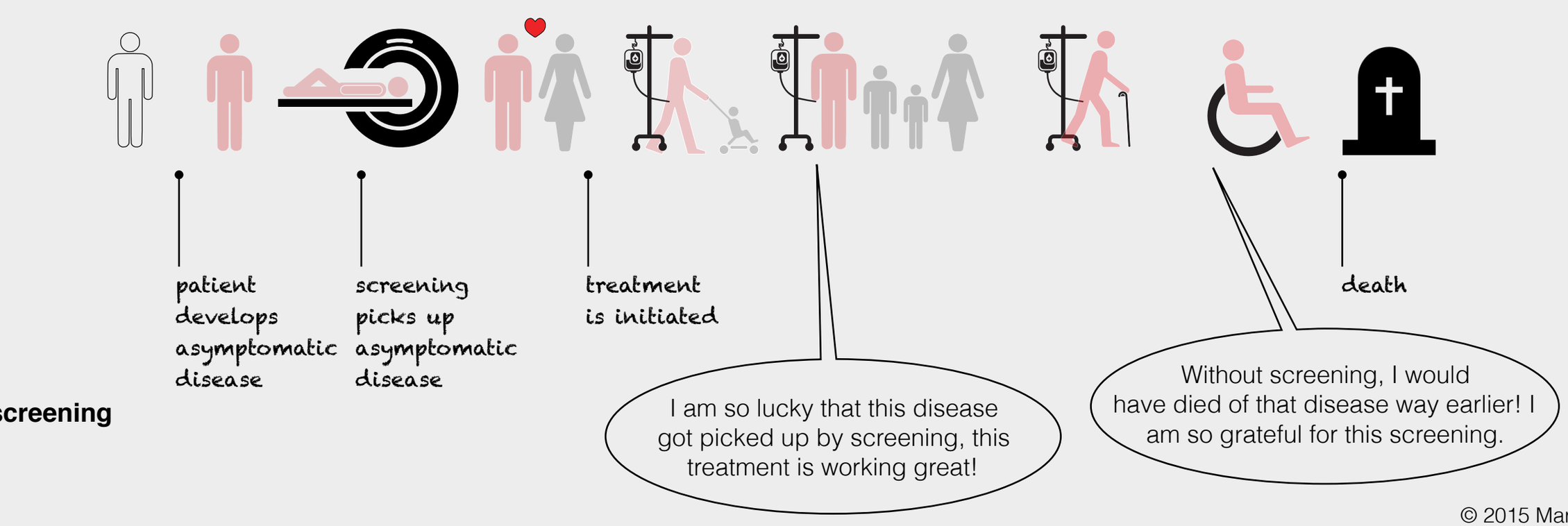
Screening tests

Pseudodisease or Overdiagnosis

Overdiagnosis is sort of like lead time bias with a ***latent phase equal to the patient's life expectancy***. That is, the disease would NEVER have become symptomatic if it had not been diagnosed by screening.

no screening

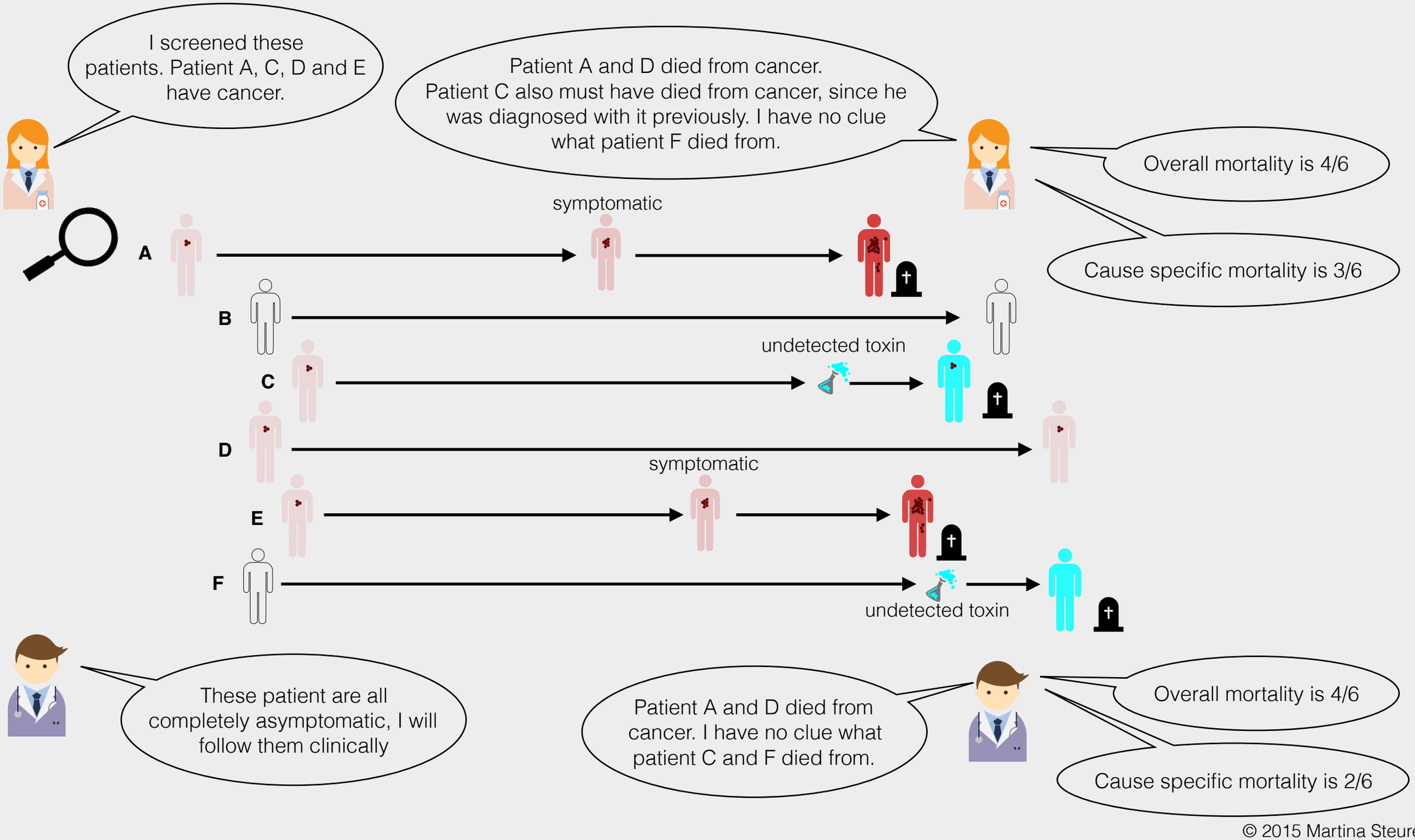
screening



Screening tests

Sticky diagnosis bias

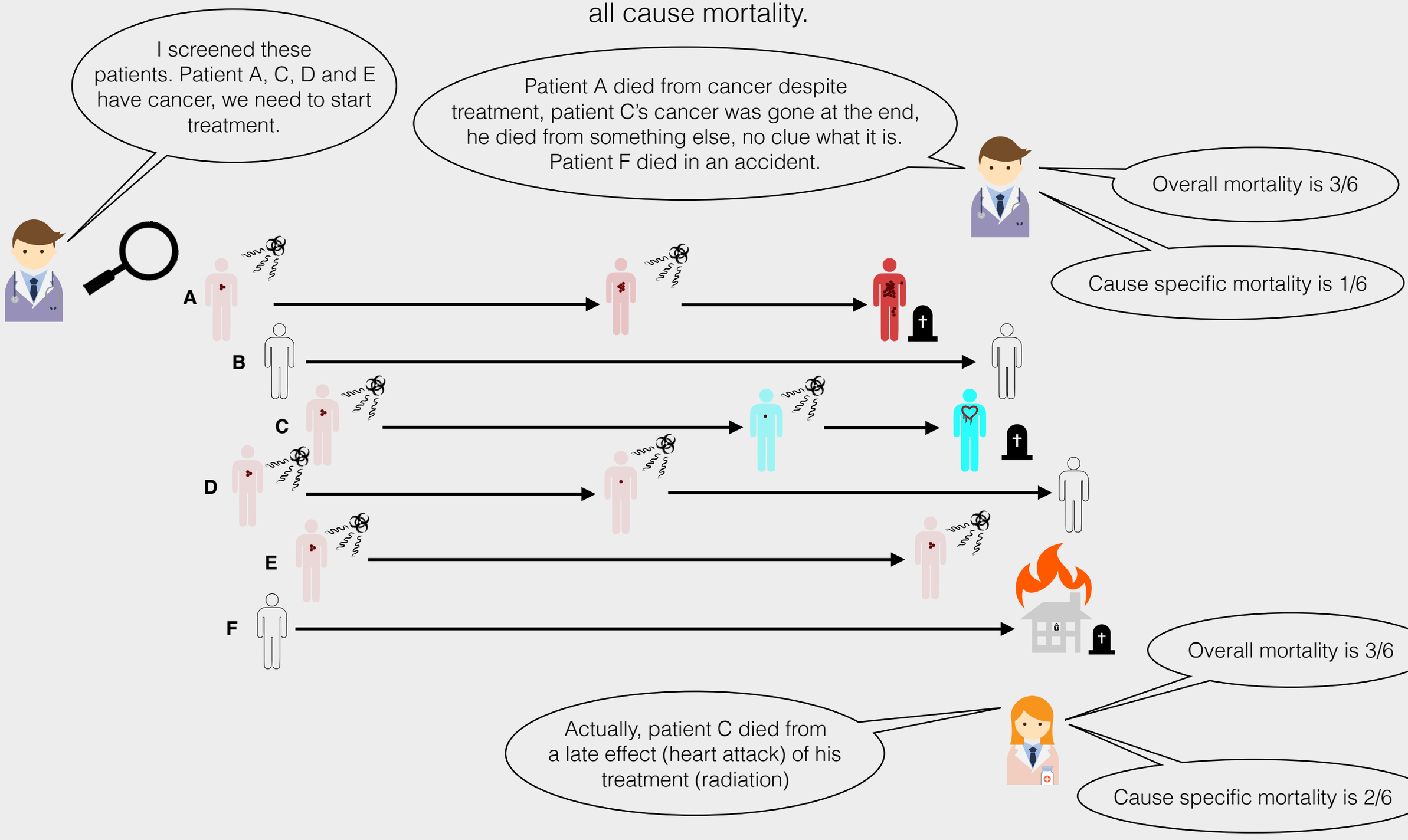
Overestimation of cause-specific mortality due to incorrect assignment of cause of death. Can be avoided by measuring all cause mortality.



Screening tests

Slippery linkage bias

Underestimation of cause-specific mortality due to incorrect assignment of cause of death. Late deaths due to complications of screening or treatment will not be counted in cause-specific mortality. Can be avoided by measuring all cause mortality.

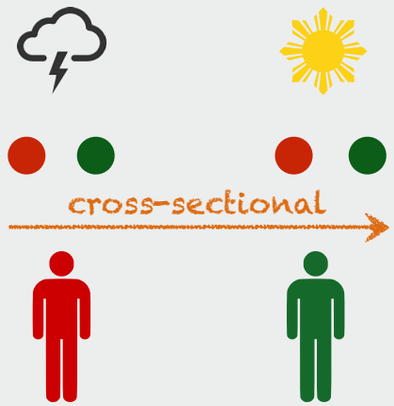


Lecture 6
Prognostic tests

Prognostic tests

Diagnostic test

identifies prevalent disease



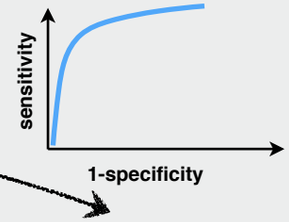
1. random events happen to patient that influence development of disease — happen BEFORE test
2. perform test for disease and obtain result r — result is +/-, ordinal or continues
3. gold standard determines true disease state
4. calculate P(r|D+), P(r|D-), LR(r) for each test result

	D+	D-	total
positive test	A	B	A+B
negative test	C	D	C+D
total	A+C	B+D	A+B+C+D

P(r|D+) = sens

P(r|D-) = 1-spec

LR(r) = sens/1-spec



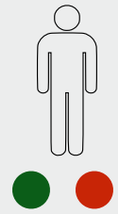
this is not always true remember spectrum bias

wellest of the well — sickest of the sick

sensitivity, specificity and LR are SEPARATE (independent of) from pretest probability

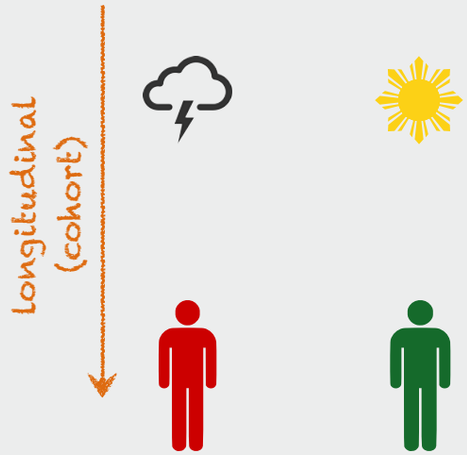
Prognostic test

predicts incident disease/outcome



result estimates risk directly

1. perform test to obtain result r to predict risk of disease/outcome
2. random events happen to patient that influence development of disease/outcome — happen AFTER test
3. follow patients over time and see if they develop disease/outcome



4. assess calibration and discrimination

discrimination

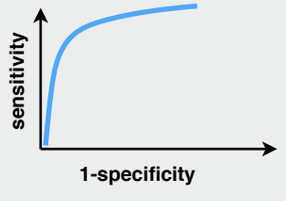
how well can the test separate subjects in the population from the mean probability to values closer to zero or 1?

- often measured with C-statistic (AUROC)

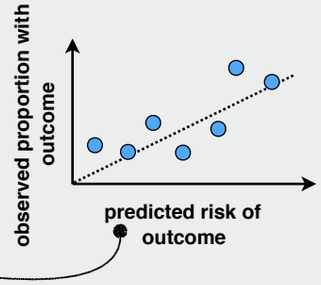
calibration

how accurate are the predicted probabilities?

- break the population into groups
- compare actual and predicted probabilities for each group



COMBINE discrimination with estimation of pretest probability, report PV P(D|r)



the ROC curve depends only on ranking, not calibration

Lecture 7
Combining tests and RCT

Combining tests

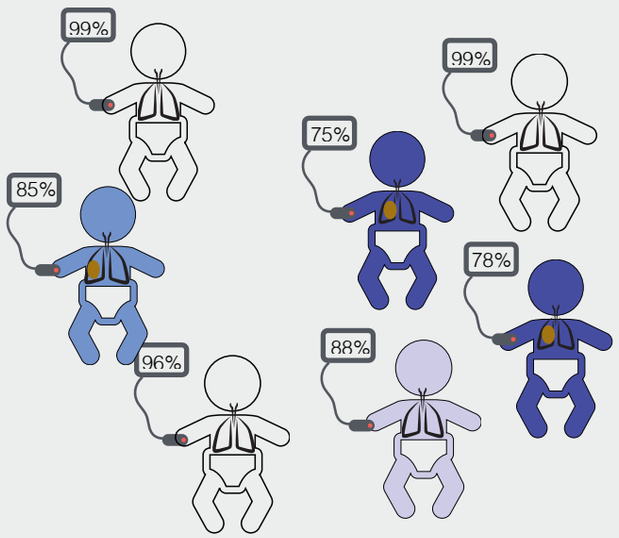
Test independence

Definition: Two tests are "independent" if the LR for any combination of results on the two tests is equal to the product of the LR for the result on the first test times the LR for the result of the second test

Explanation: among people who have the disease, knowing the result of test 1 tells you nothing about the probability of a certain result on test 2 and that the same is true among people who do not have the disease - this is "*conditional independence*", i.e. independence **conditional** on disease status.

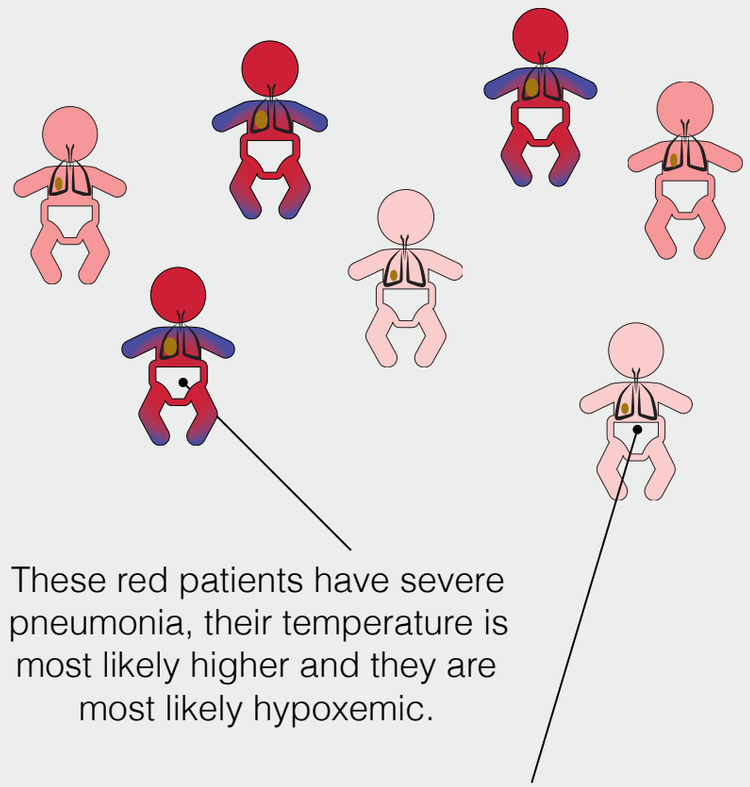
There are 3 reasons for test non-independence:

1. Tests measure similar things



Visual cyanosis and oxygen saturation measure the same aspect of disease: hypoxemia. Therefore, among both babies with and without pneumonia, those who have low oxygen saturation are more likely to be cyanotic.

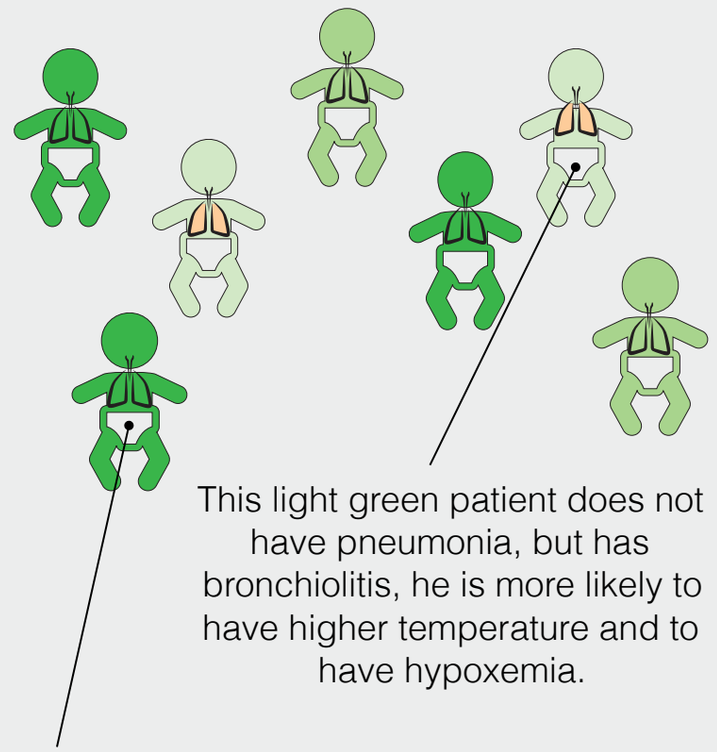
2. Heterogeneity of disease



These red patients have severe pneumonia, their temperature is most likely higher and they are most likely hypoxemic.

These light red patients have mild pneumonia, their temperature is probably lower and they are less likely hypoxemic.

3. Heterogeneity of non-disease



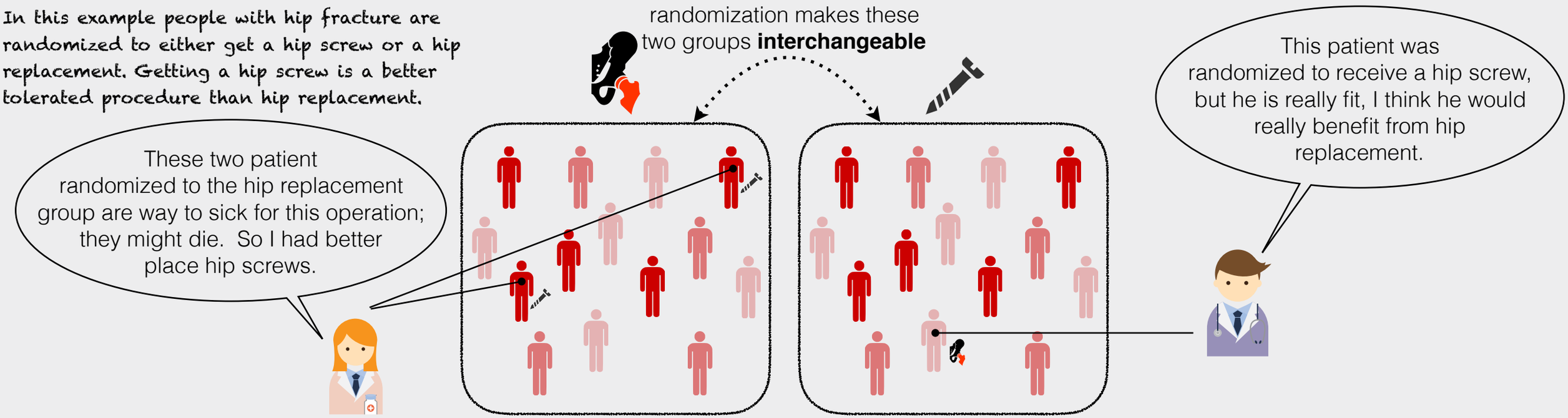
This light green patient does not have pneumonia, but has bronchiolitis, he is more likely to have higher temperature and to have hypoxemia.

These green patients is very healthy and has most likely normal temperature and is not hyperemic.

Randomized controlled trial

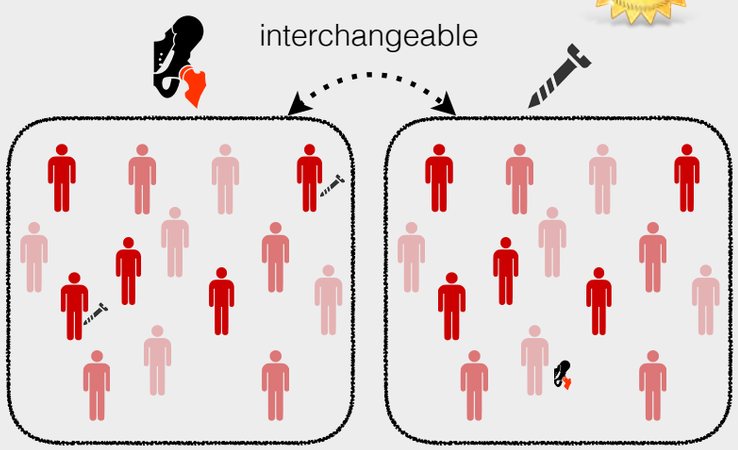
randomization

In this example people with hip fracture are randomized to either get a hip screw or a hip replacement. Getting a hip screw is a better tolerated procedure than hip replacement.



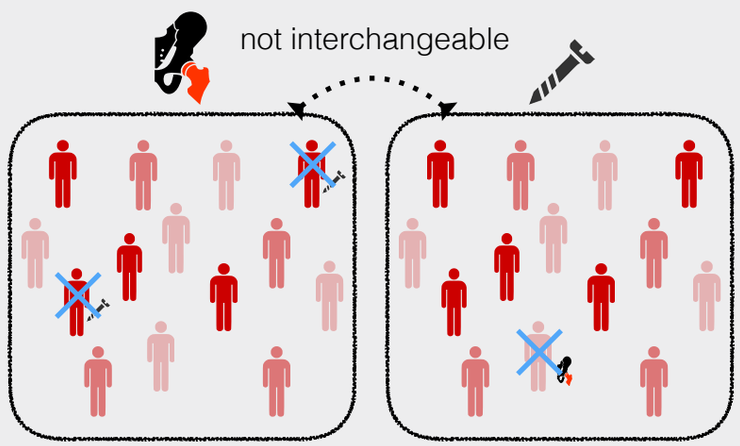
analysis

Intention to treat



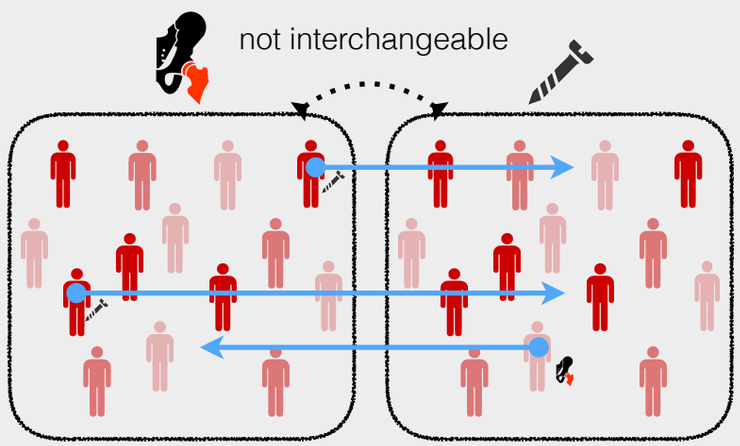
By keeping each study subject in the randomly assigned group regardless of the treatment they actually received results represent interchangeable groups without bias

Per protocol



In this scenario, the hip replacement group gets depleted of very sick patients while the hip screw group gets depleted of less sick patients, so the groups are no longer interchangeable. This leads to bias: the hip replacement treatment appears better than it is.

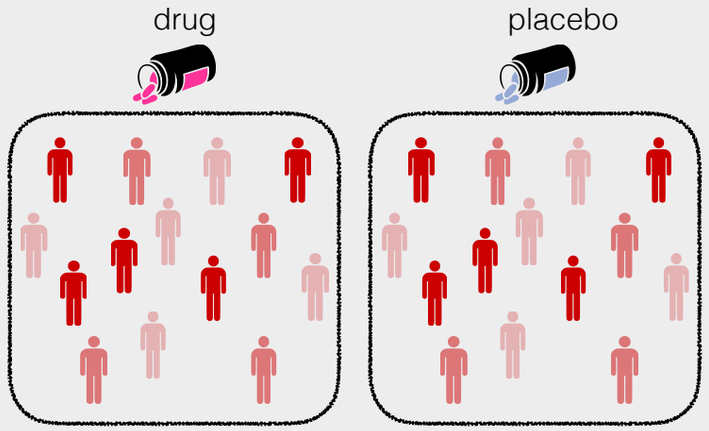
As treated



In this scenario, the hip replacement group gets depleted of very sick patients and receives less sick patients while the hip screw group gets depleted of less sick patients while receiving sick patient, the groups are no longer interchangeable. This leads to bias: the hip replacement treatment appears better than it is.

Randomized controlled trial

Blinding



1. Blinding of patients

I have no clue if this is placebo or the actual drug, my report of symptoms will be as objective as possible

I know I am taking this new drug. I can already feel it helping, my symptoms are much improved.

Differential co-interventions and biased patient report of signs and symptoms makes these groups not interchangeable anymore and introduces bias

2. Blinding of treatment team

I have no clue who is getting the drug and who is getting the placebo. I treat everybody the same.

I know who is getting the placebo. I will give some Vitamins to the placebo group so I at least do something good for them

Differential co-interventions makes the groups not interchangeable anymore and introduces bias

3. Blinding of outcome assessment

I have no clue who got the drug and who got placebo, my outcome assessment is unbiased

I know this patient was getting this awesome new drug. I am sure it helped and the patient's exam improved

Outcome assessors will not be objective, which introduces bias

Lecture 8
Alternatives to Randomized Trials for Estimating Treatment Efficacy (or Harm)

The holy grail: causation

In order to provide evidence that an effect is causal, the two populations (i.e. exposed and unexposed, or treated and untreated) need to be INTERCHANGEABLE (conditional on covariates)

It is possible to estimate causal effect in cohort studies by applying different statistical methods to get rid of confounding

multivariate analysis
propensity score
instrumental variable

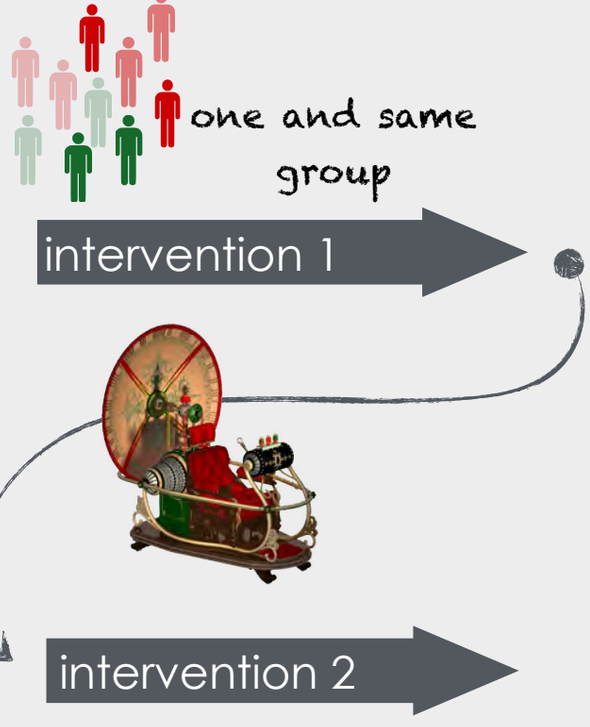
Causal effects

Correctly designed and analyzed RCT measure causal effects

Disease Association

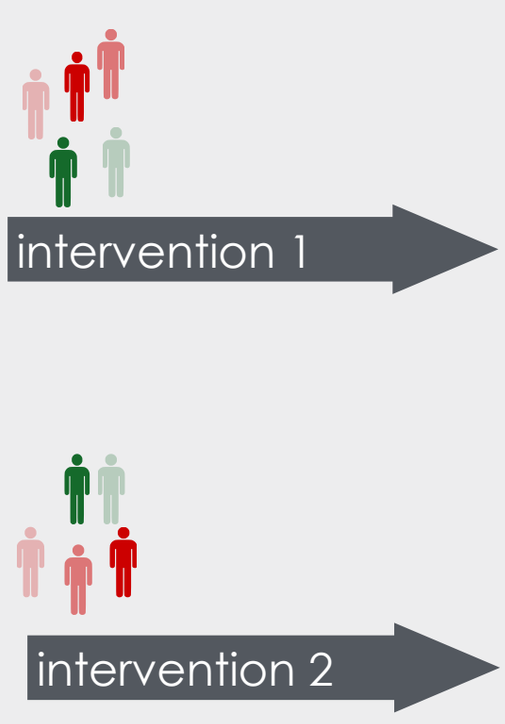
Cohort studies measure associations

Time travel



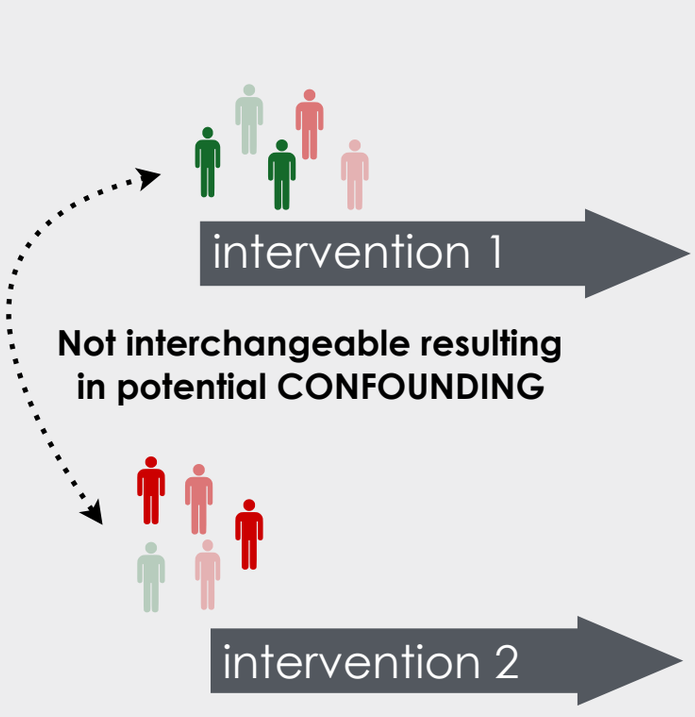
Time travel ensures that both groups are equal, sadly it is still not possible.

RCT



Randomization is a method to make groups interchangeable.

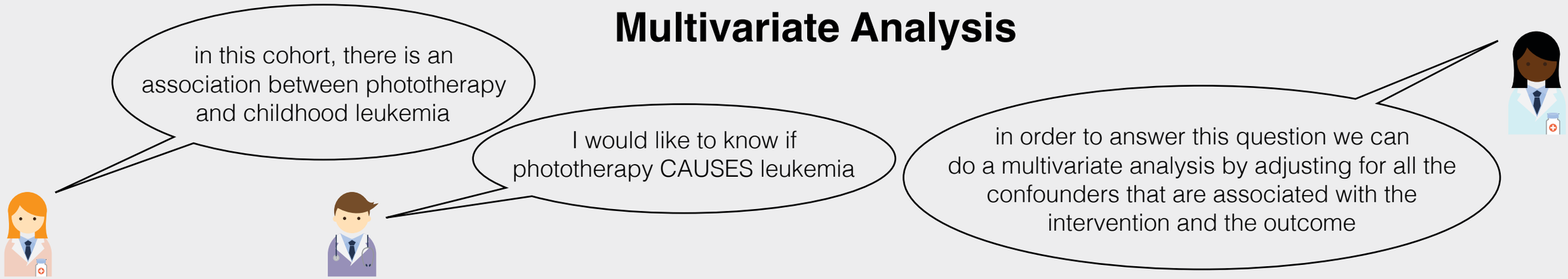
Cohort study

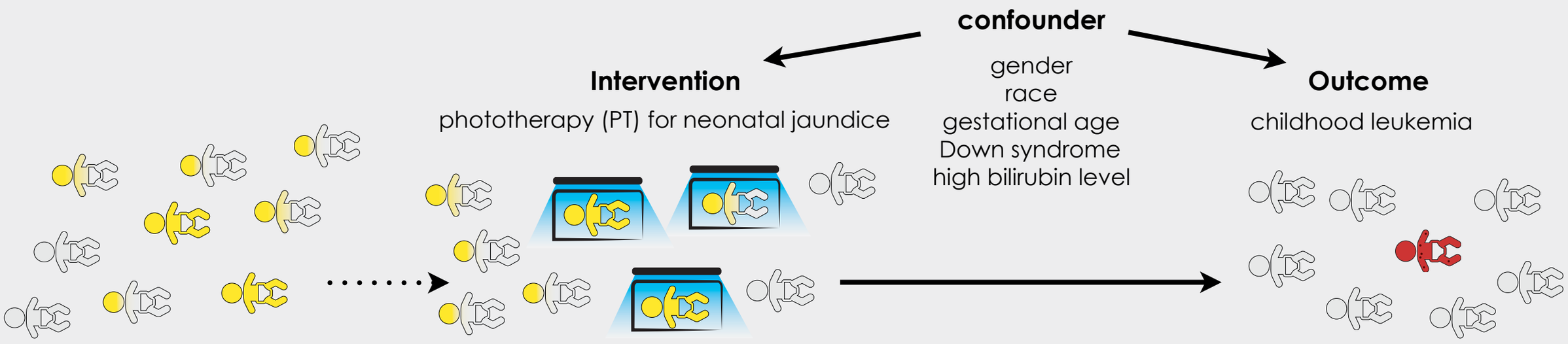


There is probably a reason why people got one or the other intervention: e.g. healthier people are more likely to receive hip replacement while sicker people are more likely to get a hip screw.

Alternatives to RCT

Multivariate Analysis





Propensity scores

method:

1. create new variable, ***propensity*** to be treated with intervention

   build a model to estimate the propensity of getting phototherapy, including all known confounders

2. match on, stratify for or include "propensity" in multivariate analysis

   e.g. match treated and not related infants based on propensity

advantages:

1. better power to control for co-variables: because receipt of the *intervention* may be much more common than occurrence of the outcome

   e.g. phototherapy is much more common than leukemia

2. you can more easily tell when treated and untreated groups are not comparable

   i.e. propensity scores don't overlap

caveats:

1. can only compare subjects whose propensity scores overlap: can only generalize to subjects who could have received either treatment

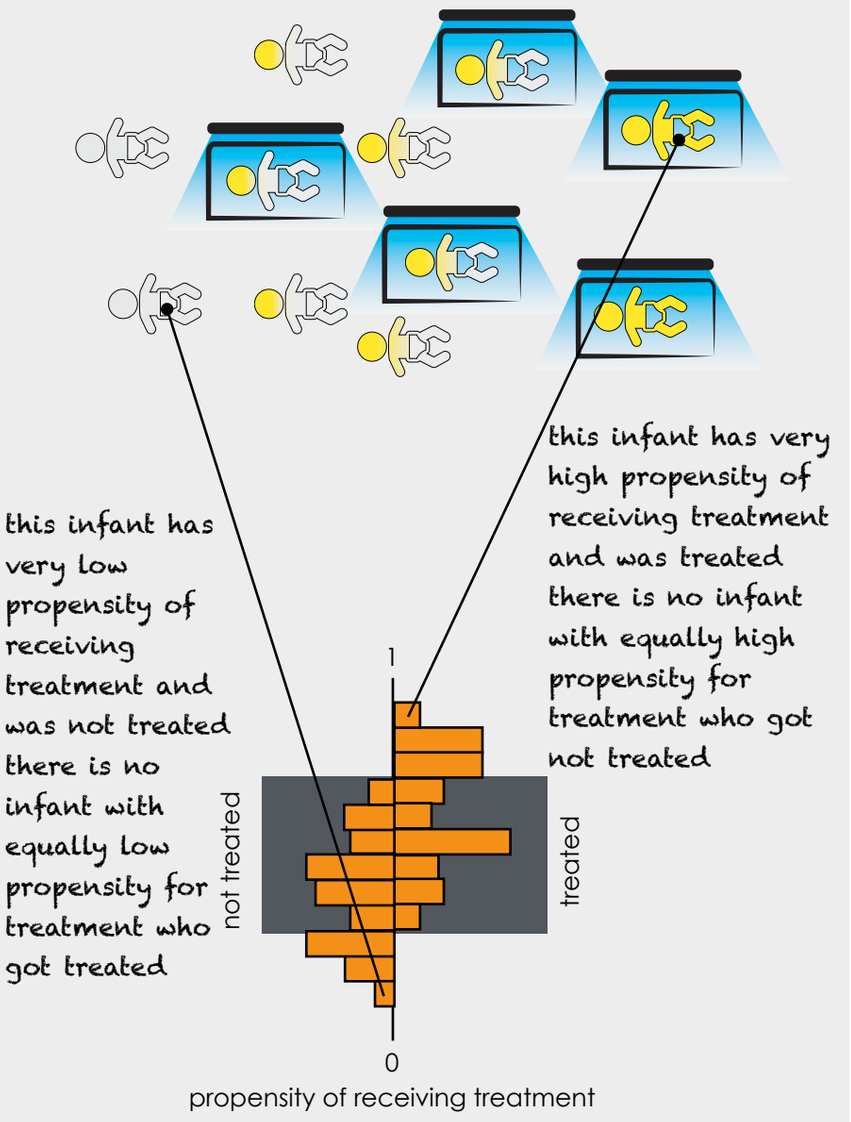
   e.g. infant with extremely high bilirubin has probably a very high propensity of getting PT and there is probably no infant with such a high propensity who did not get PT

2. important variable might be missing from model

Alternatives to RCT

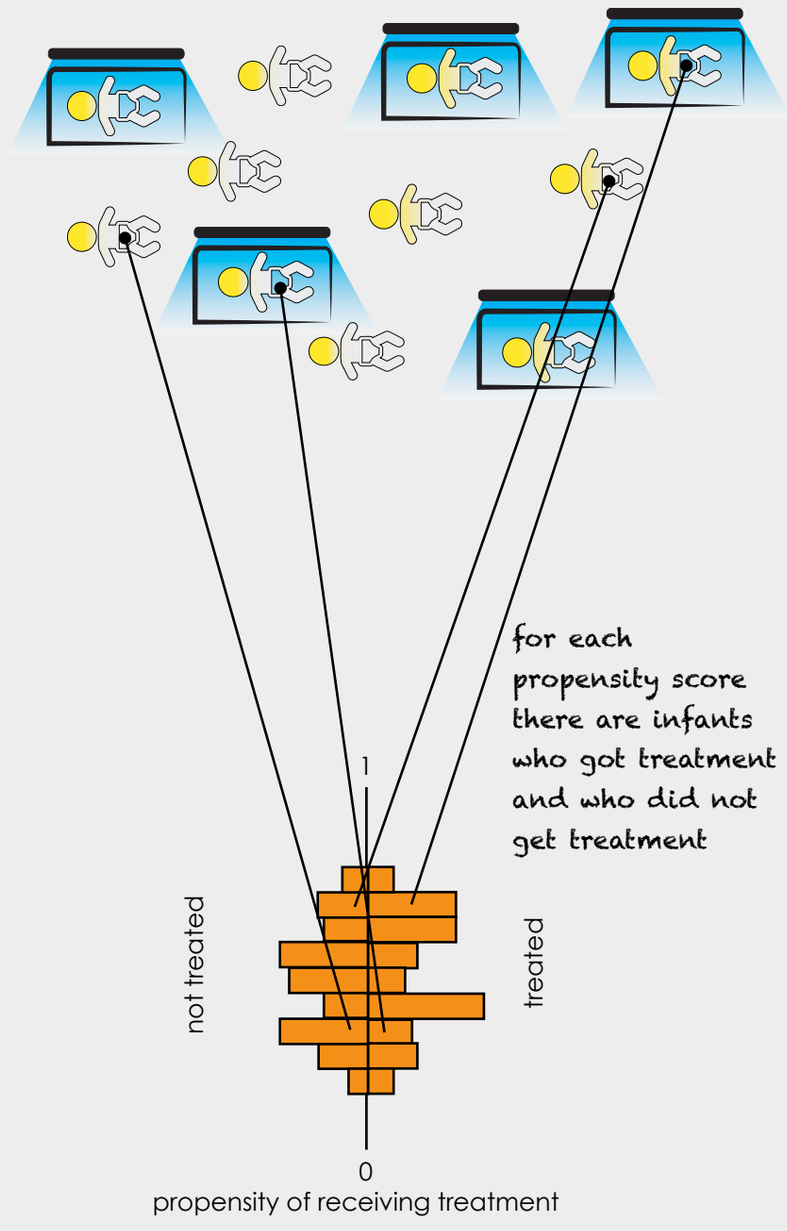
Propensity scores: distribution and overlap

1. good overlap



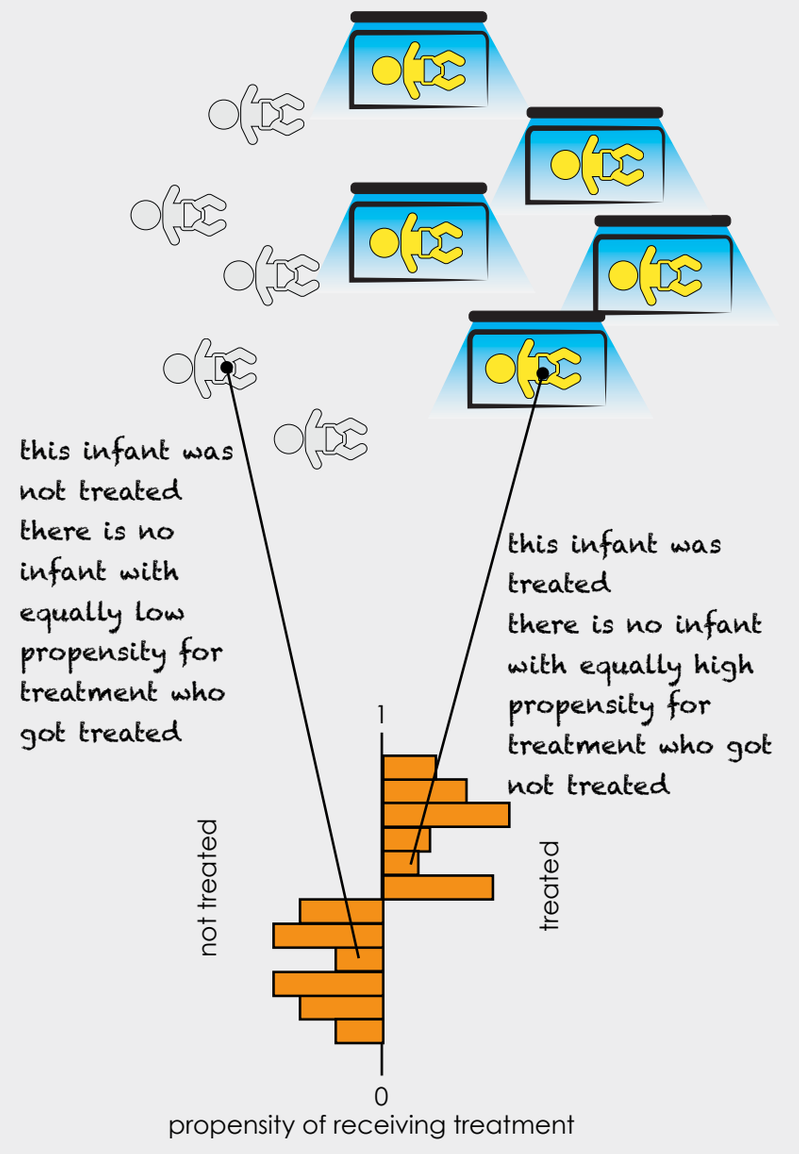
Good overlap in propensity scores. The subjects in the overlapping parts of the distribution can be studied.

2. perfect overlap



Propensity distributions are nearly identical. A propensity analysis is not necessary as groups are already matched or treatment was randomly assigned.

3. no overlap



Propensity scores do not overlap; treated and untreated groups are not comparable. A propensity analysis cannot be done and andy comparison between groups is hazardous.

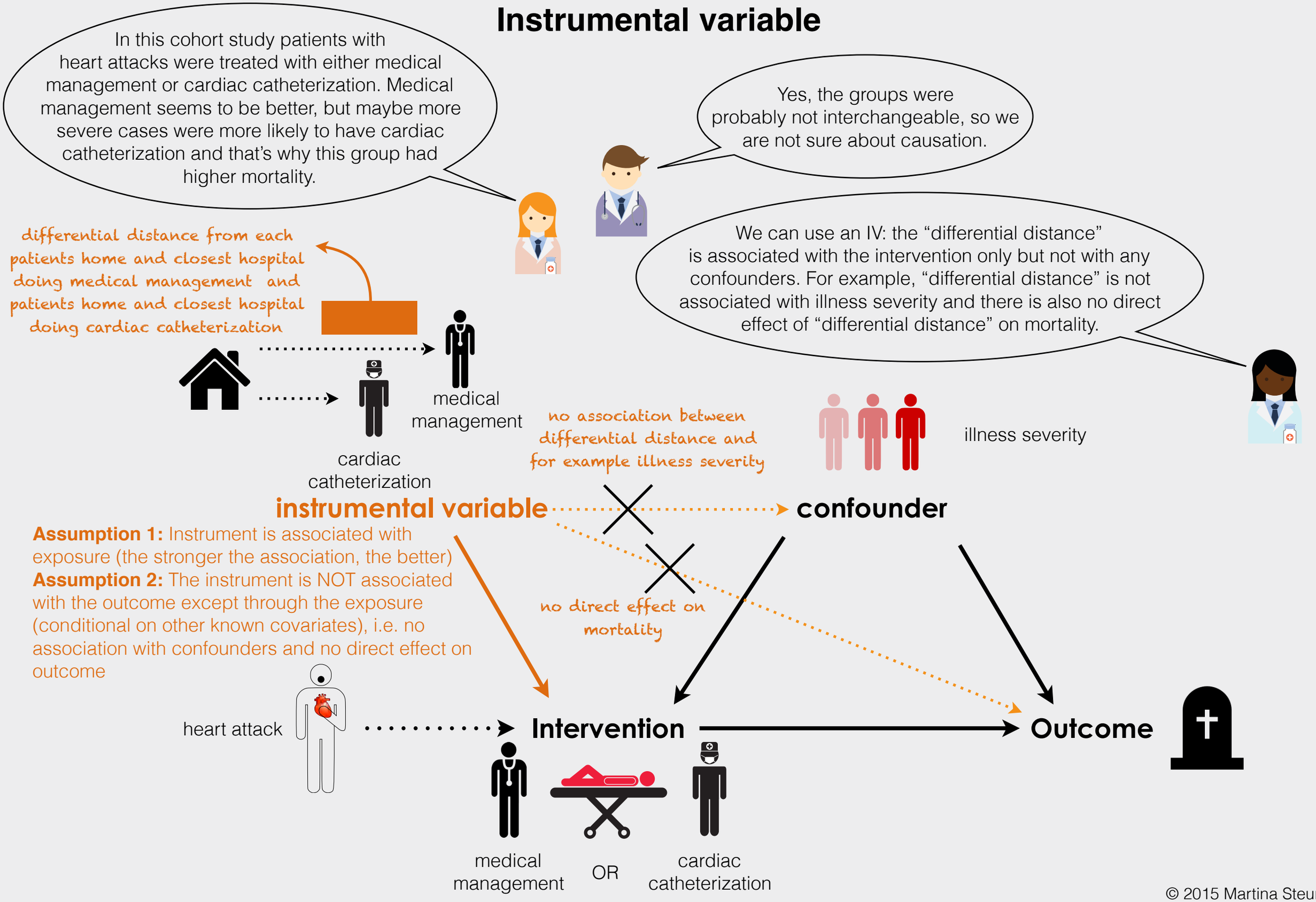
Alternatives to RCT

Instrumental variable

In this cohort study patients with heart attacks were treated with either medical management or cardiac catheterization. Medical management seems to be better, but maybe more severe cases were more likely to have cardiac catheterization and that's why this group had higher mortality.

Yes, the groups were probably not interchangeable, so we are not sure about causation.

We can use an IV: the "differential distance" is associated with the intervention only but not with any confounders. For example, "differential distance" is not associated with illness severity and there is also no direct effect of "differential distance" on mortality.



Assumption 1: Instrument is associated with exposure (the stronger the association, the better)
Assumption 2: The instrument is NOT associated with the outcome except through the exposure (conditional on other known covariates), i.e. no association with confounders and no direct effect on outcome