Selection Bias: Past, Present, & Future

Timothy L. Lash, DSc, MPH
Emory University

Outline

Present perspectives

Allusions to past perspectives

Challenges going forward

Present-ish perspectives

Definition of selection bias

"Selection bias arises when—in a study population—an estimate of disease occurrence, or an estimate of the effect of an exposure contrast on disease occurrence, differs from the estimate that would have been obtained in the study population's source population because of the way the study population was selected, either by design or analytic choice."

Fox, MacLehose, Lash (*Applying Quantitative Bias Analysis to Epidemiologic Data, 2nd edition*)

Definitions of populations

The source population is the population from which persons will be sampled and included in a measurement of disease frequency.

For example, the source population of the original Framingham Heart Study included men and women between the ages of 30 and 62 who were residents of the town of Framingham, Massachusetts in 1948.

From Greenland, Rothman, Lash, *Modern Epidemiology, 4th edition*

Definitions of populations

The study population is the subset, up to a complete census, of the source population whose experience (people over time) is included in a measurement of disease frequency.

Not all men and women who were eligible to join the Framingham Heart Study were invited to participate, and not all who were invited to participate agreed to participate (31% refusal rate, 13% initial loss to follow-up). Those who were invited, agreed, and were free of prevalent cardiovascular disease comprised the study population.

From Greenland, Rothman, Lash, *Modern Epidemiology, 4th edition*

Definitions of populations

The target population comprises the persons for whom information gleaned by the measurement of disease frequency will be relevant.

Information about risk factors for cardiac disease gleaned from the Framingham Heart Study has contributed to a nearly 75% decline in mortality related to cardiovascular disease in most industrialized societies.

From Greenland, Rothman, Lash, *Modern Epidemiology, 4th edition*

Definition of selection bias

"Selection bias arises when—in a study population—an estimate of disease occurrence, or an estimate of the effect of an exposure contrast on disease occurrence, differs from the estimate that would have been obtained in the study population's source [~~target~~] population because of the way the study population was selected, either by design or analytic choice."

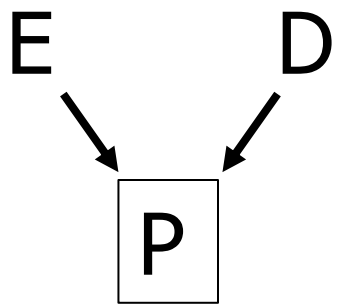
Fox, MacLehose, Lash (*Applying Quantitative Bias Analysis to Epidemiologic Data, 2nd edition*)

Two main structures for selection bias

Structure one of selection bias

Selection bias under the null arises from conditioning on a collider that is a descendant of both exposure (E) and outcome (D).

Conditioning is usually by restriction of the source population to members of the study population: analyses are restricted to persons who agree to participate or who agree to continue to participate.

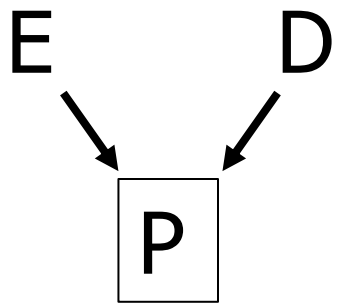


Hernán et al. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25.

Past perspective: Berkson's Bias

Selection bias under the null arises from conditioning on a collider that is a descendant of both exposure (E) and outcome (D).

Exposure health condition (E) associated with second outcome health condition (D) because source population restricted to study population of persons hospitalized (P)



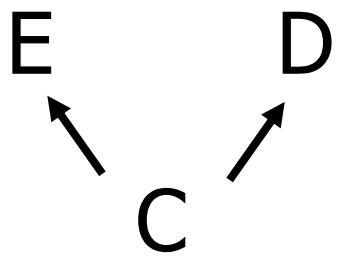
Hernán et al. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25.

Past perspective: Healthy Worker Effect

Workers have lower risk of many health outcomes than "general population" to which they were often compared.

Historically characterized as a selection bias (see Modern Epidemiology 1st and 2nd editions).

2004 paper clearly showed it is a confounding bias. (C)=fitness is associated with readiness to work (E) and outcomes (D)



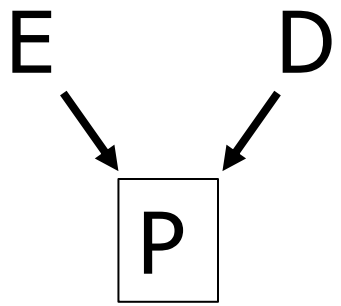
Hernán et al. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25.

Retrospective versus Prospective Design

Selection bias under the null arises from conditioning on a collider that is a descendant of both exposure (E) and outcome (D).

In order for (E) and (D) to affect P, they must be known at the time of study conduct.

Restricts selection bias of this structure to retrospective design



Hernán et al. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25.

Prospective

Retrospective

http://fantagear.files.wordpress.com/2011/06/good_and_bad_egg.jpg

"Retrospective" = "bad"

Pharmacological Management of Psychiatric Illness During Pregnancy

The extent to which these findings are of consequence to humans has yet to be established. Behavioral outcomes after prenatal exposure to psychotropics, including tricyclic antidepressants (TCAs), benzodiazepines, lithium, and, more recently, fluoxetine, have been reported (Laegreid et al. 1992; Misri and Sivertz 1991; Nulman et al. 1997; Schou 1976). Except for the fluoxetine data, most data are extremely limited by retrospective design and by small sample size. Furthermore, relevant control groups with psychiatric disorders but no psychotropic exposure have not been included in the analysis.

Viguera & Cohen. In *Psychiatric Illness in Women*. 2002; p. 579.

"Prospective" = "good"

Evidence-Based Treatment Decisions for Extremely Preterm Newborns

Results of the cohort study by Bader et al, published in this issue of *Pediatrics*,[2] are an important addition to the evidence base for these decisions. Theirs is one of only a few recent population-based studies that are free of the referral biases in virtually all center-based studies. Other important strengths include prospective data collection and entry into a computerized database with error checks. Prenatal and postnatal risk factors were related to predischarge mortality for >99% of the 3768 infants born alive at 23 to 26 weeks' gestational age (GA) in

Parikh *et al. Pediatrics* 2010;125:813–6.

The words “retrospective” and “prospective” have powerful connotations that imply study quality.

One would therefore like to think that they have widely accepted and well understood definitions.

In fact, there are three very different definitions, all routinely used, and with vastly different implications for inherent quality.

Let’s examine the three definitions.

http://thinkfuture.com/wp-content/uploads/2013/02/Thumbs-Up-Thumbs-Down.jpeg

Definition #1: Prospective=cohort design

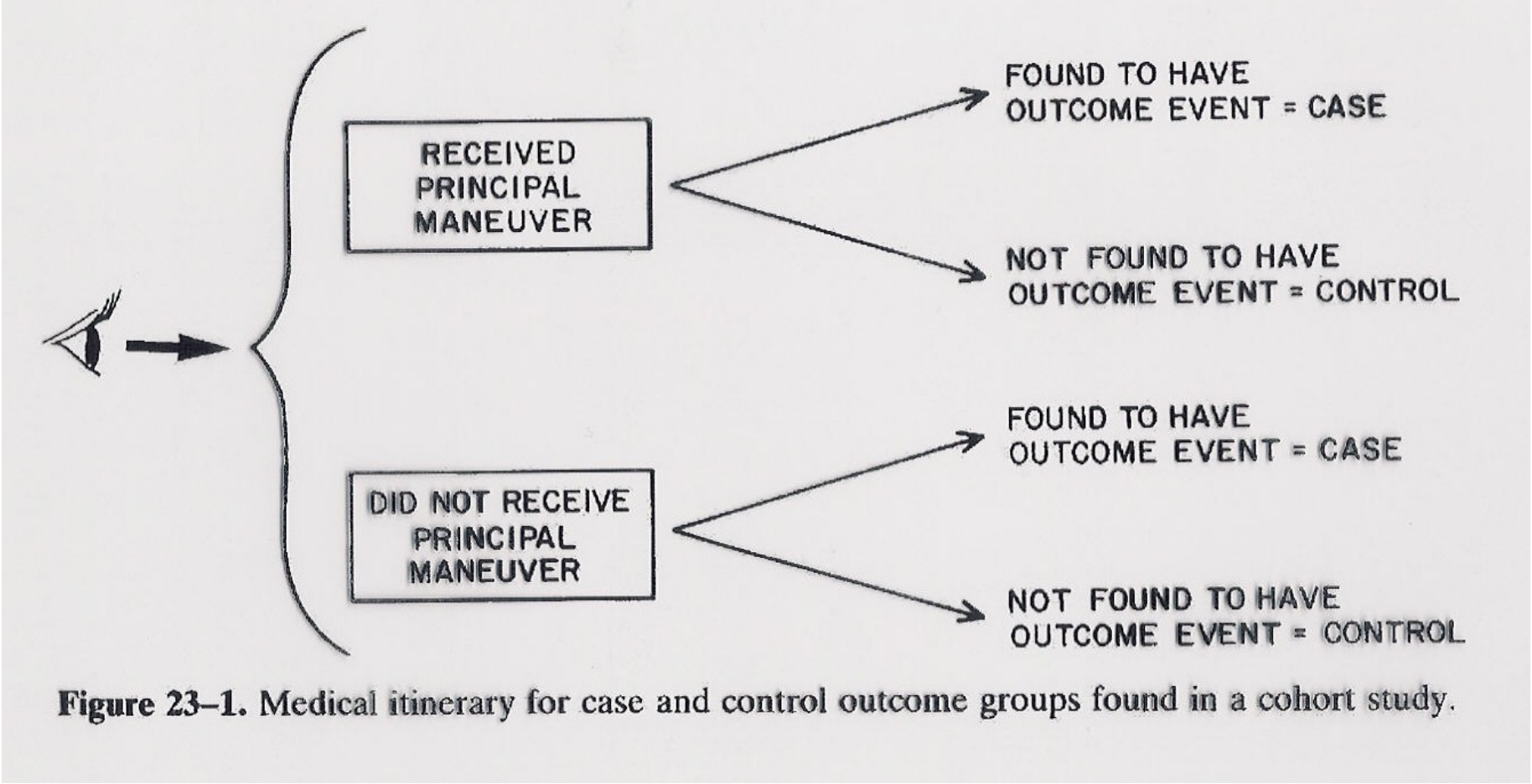


Figure 23–1. Medical itinerary for case and control outcome groups found in a cohort study.

Feinstein A. *Clinical Epidemiology* 1985.

Definition #1: Retrospective=case-control design

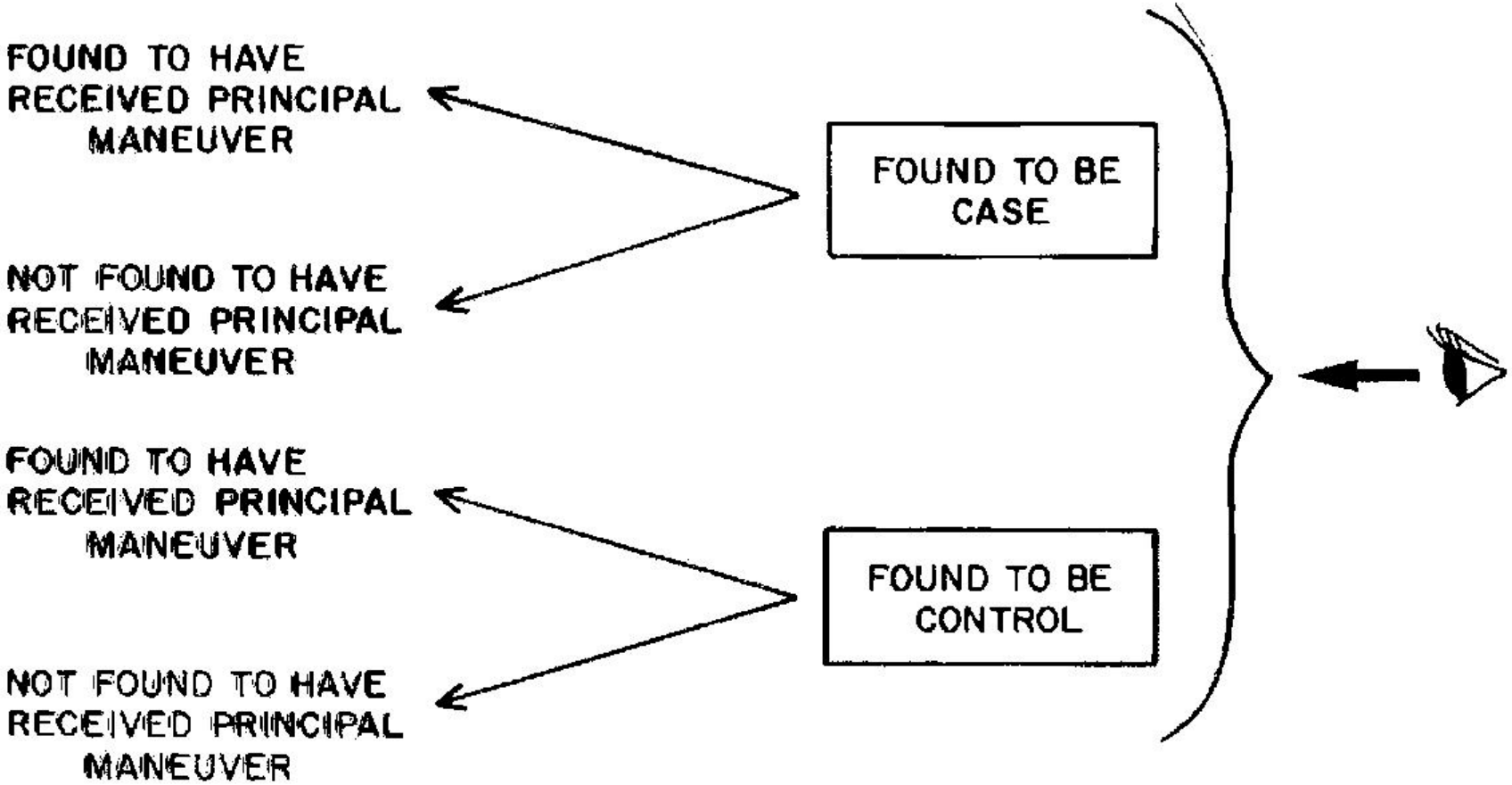


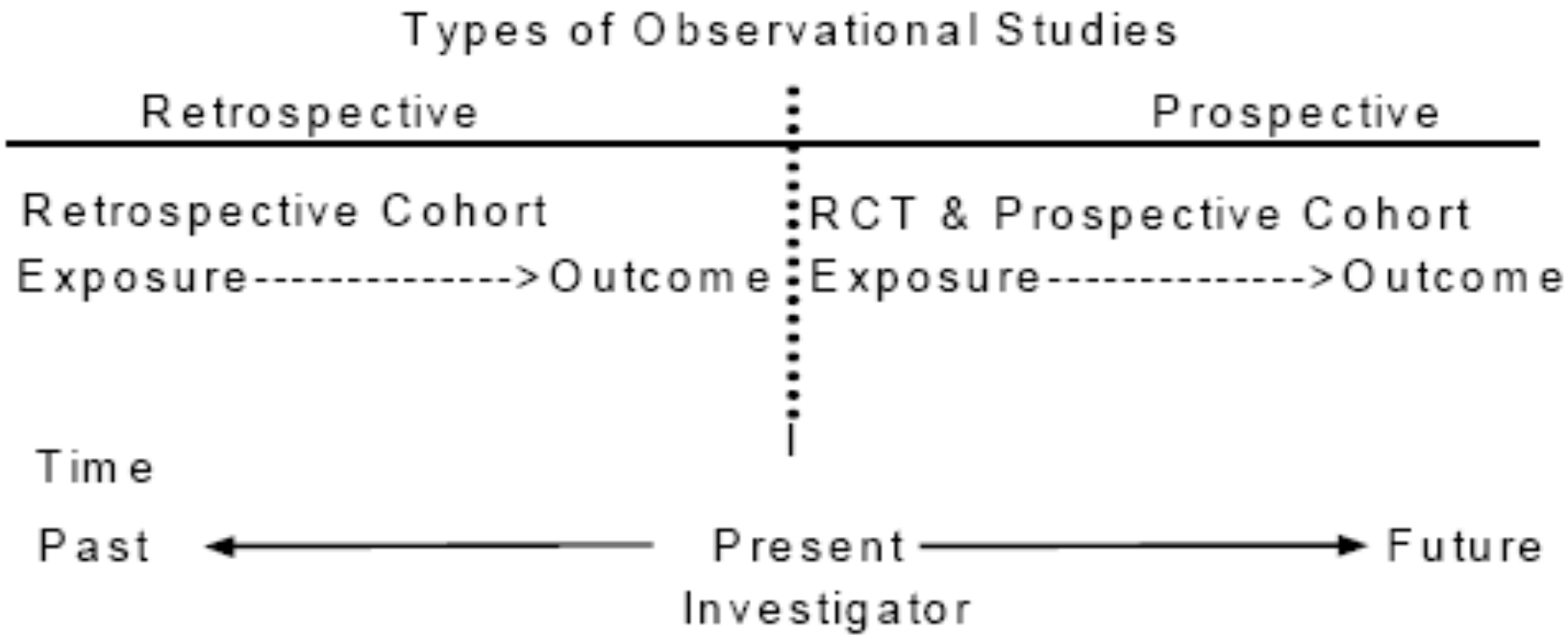
Figure 23–2. Investigative sequence in a retrospective case-control study.

Feinstein A. *Clinical Epidemiology* 1985.

Definition #2:
The investigator's perspective

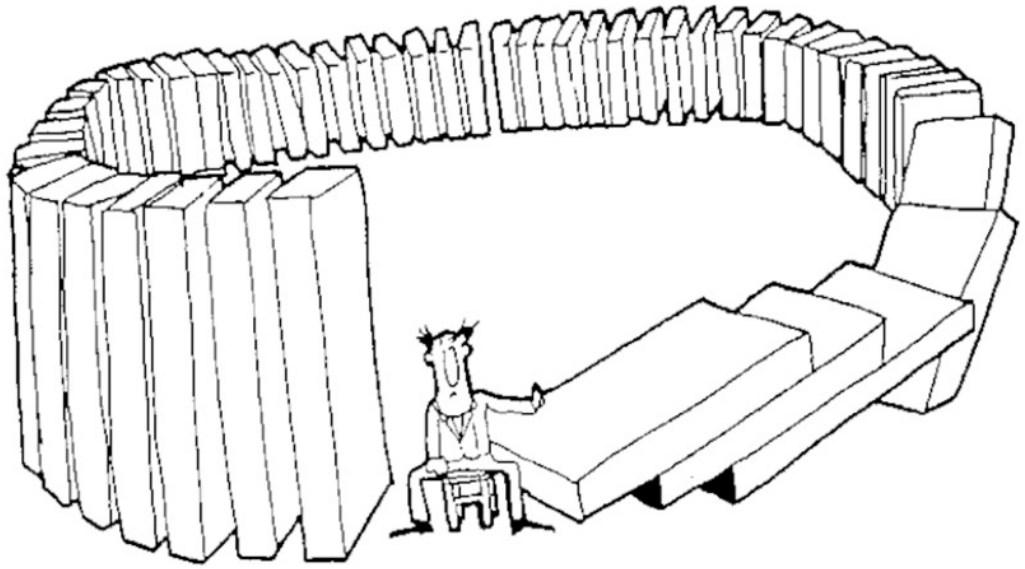
http://skinnymom.com/wp-content/uploads/2011/12/past-and-future.jpg

Definition #2:
The investigator's perspective



http://www.myoops.org/cocw/tufts/courses/1/content/D193106/C207560.jpg

Definition #3:
Exposure record in relation to outcome

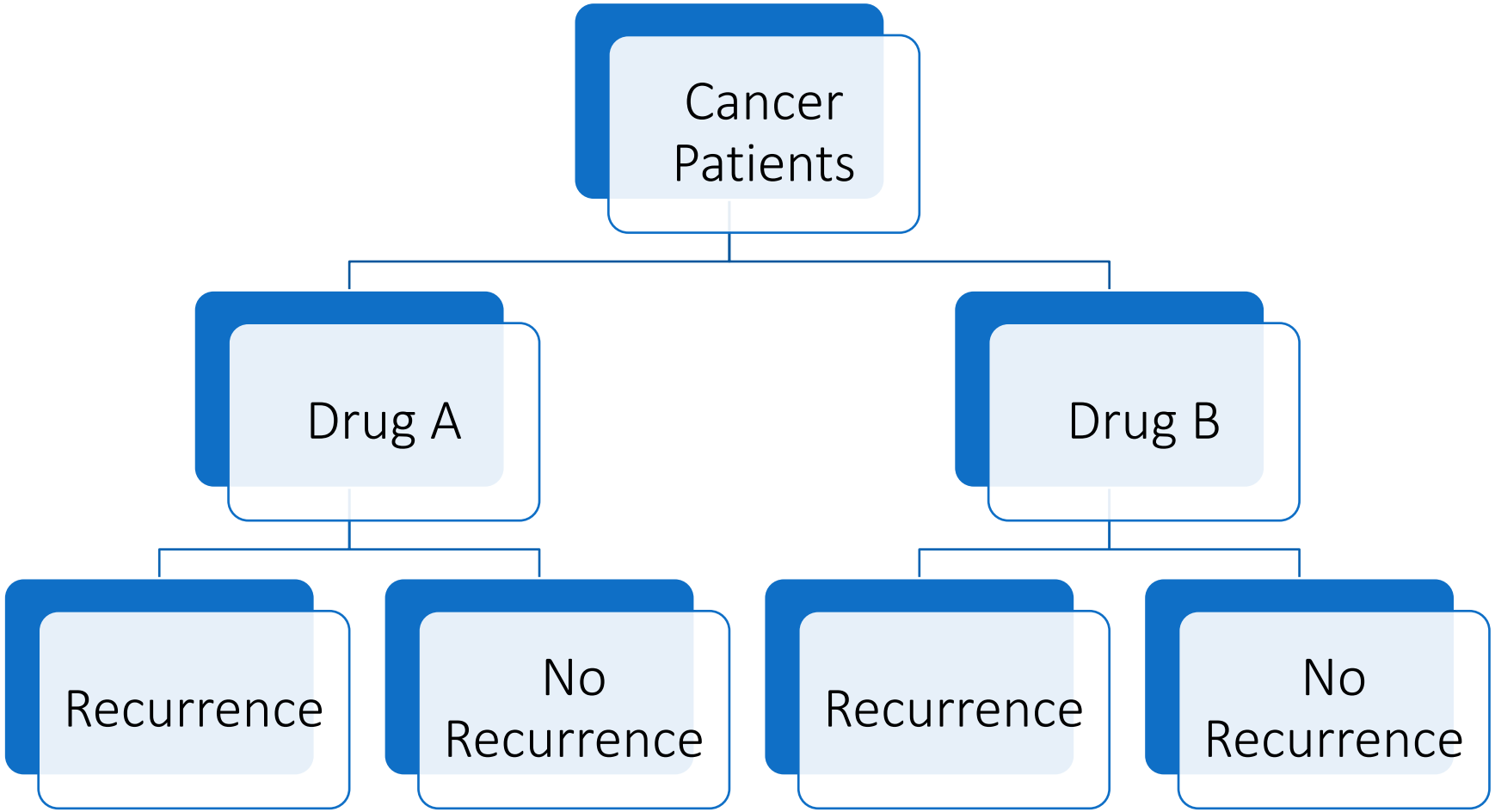


If the record of exposure was created before the outcome occurred, then prospective

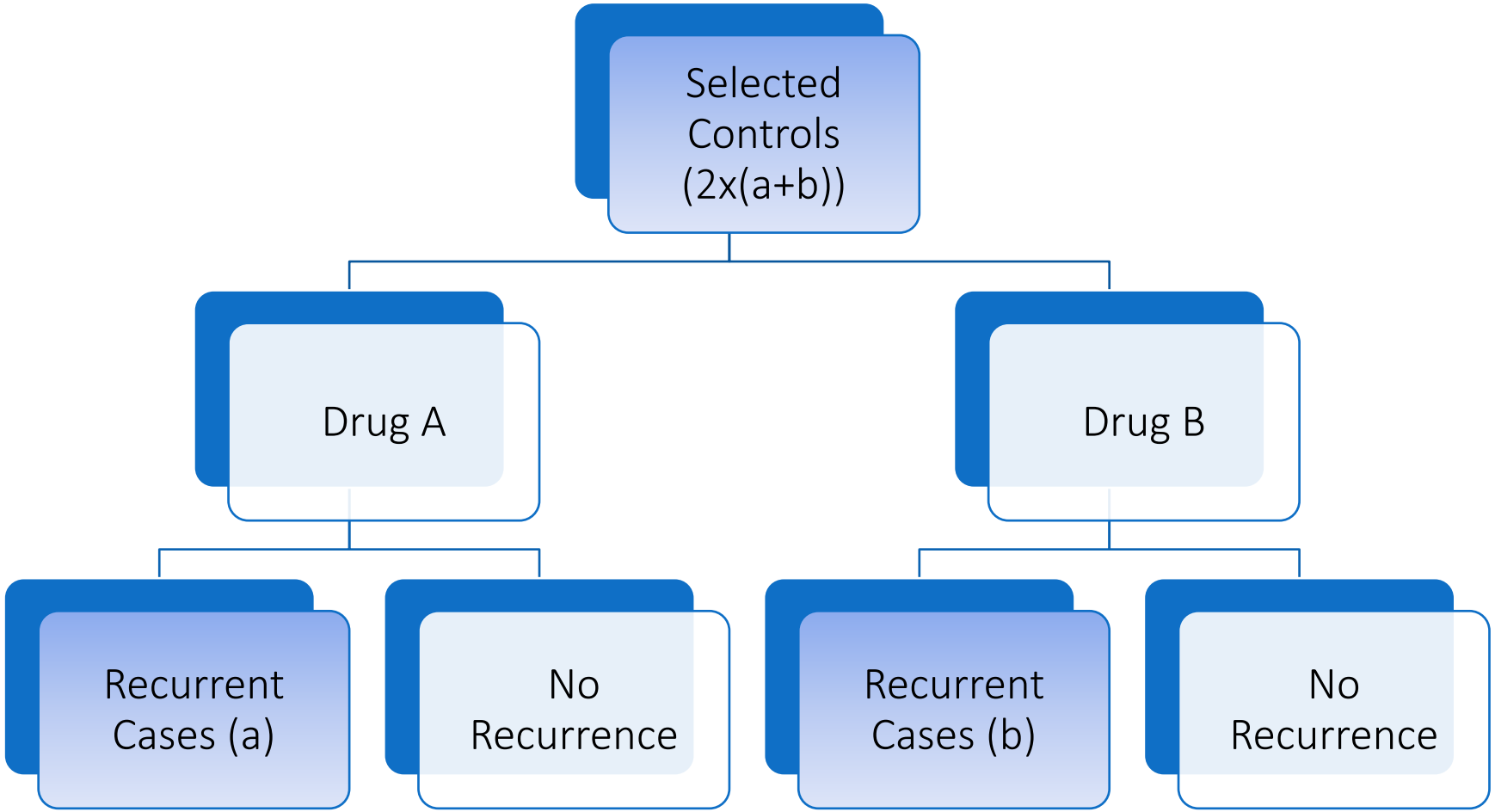
If the record of exposure was created after the outcome occurred, then retrospective

http://connie-gulick.com/images/cause-effectB.png

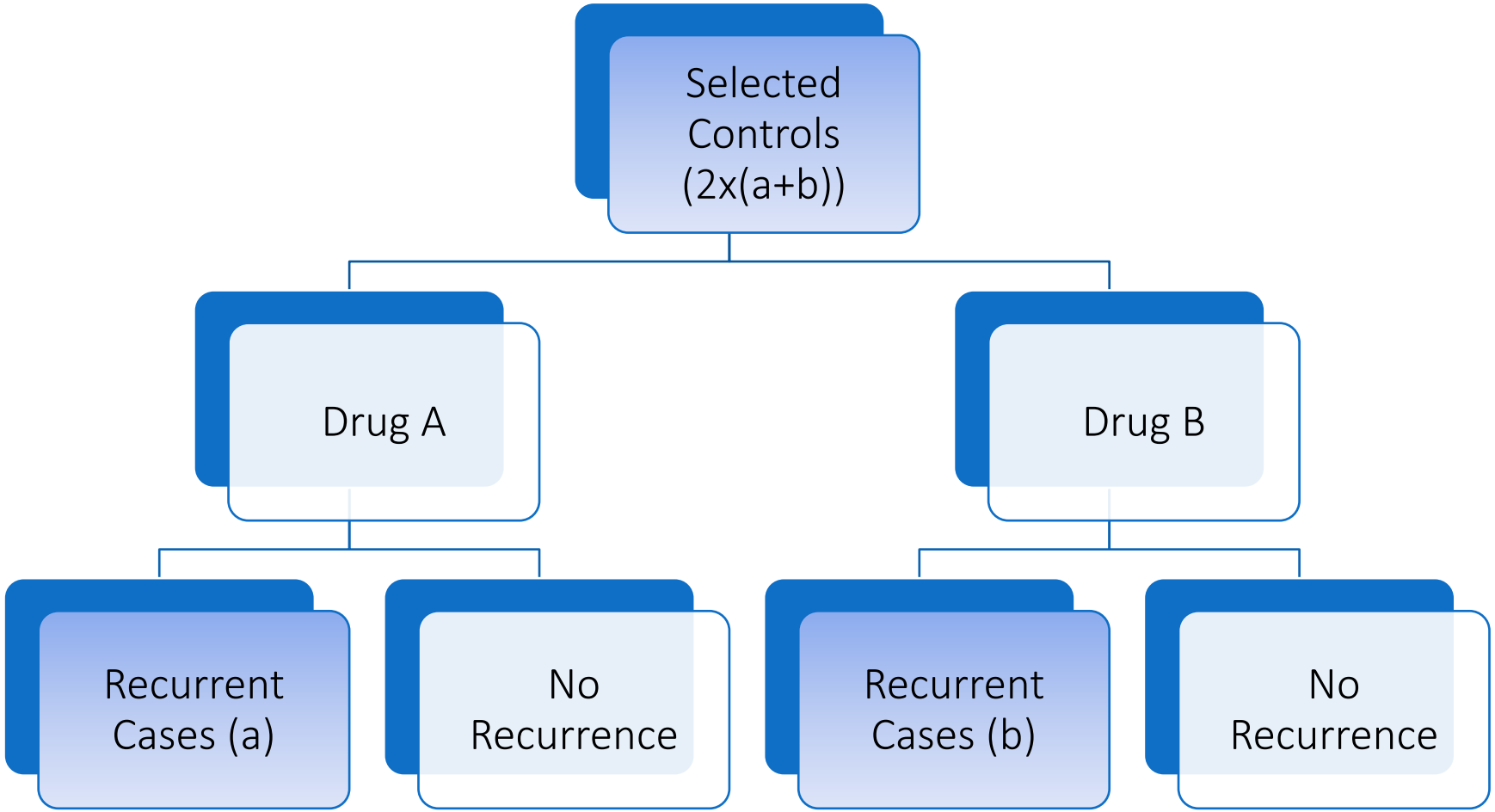
Randomized Trial Example
Study 1: The Intent to Treat Trial



Randomized Trial Example
Study 2: *a priori* case-cohort study of a biomarker



Randomized Trial Example
Study 3: *a posteriori* case-cohort study of genotype



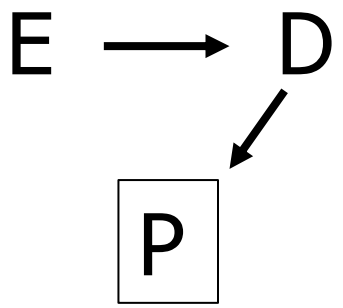
Randomized Trial Example Retrospective or Prospective?

Definition	ITT Trial	*A priori* case-cohort	*A posterior* case-cohort
1: Cohort or case-control	Prospective	Retrospective	Retrospective
2: Investigator perspective	Prospective	Prospective	Retrospective
3. Exposure before outcome	Prospective	Prospective	Prospective

Structure two of selection bias

In an example from Greenland 1977, the exposure does not directly affect participation (P), but it does affect the outcome (D). Now "off the null" (now an arrow from E to D), not "under the null."

While we must restrict the analysis to participants, this no longer induces a collider bias.

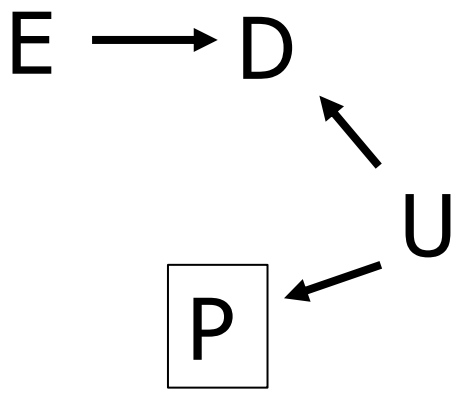


Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol. 1977 Sep;106(3):184-7.

The structure of selection bias (2)

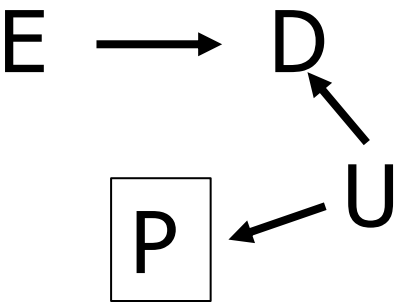
To be consistent with the causal graphs in a Hernán 2017 commentary, we will re-draw with a common ancestor of P (which is C in the commentary) and D.

U is a variable that predicts the outcome, and also predicts continued follow-up in the study. This could be some prognostic marker, such as hypertension is a prognostic marker for heart disease.



Hernán MA. Selection Bias Without Colliders. Am J Epidemiol 2017. 1048–1050,

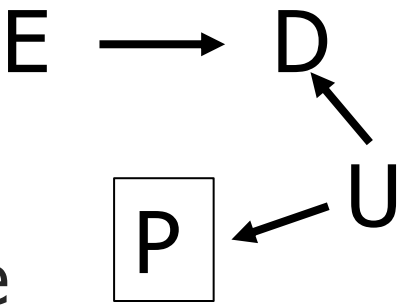
The structure of selection bias (2)



	Complete Cohort		Observed Cohort	
	E+	E-	E+	E-
D+	18	32	16	23
D-	232	718	167	517
N	250	750	183	540
RR	1.69		2.05	

Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol. 1977 Sep;106(3):184-7.

The structure of selection bias (2)



Selection was ~nondifferential because it was associated with the outcome (the odds ratio was 1.38) but not with the exposure (the odds ratio was 1.06)."

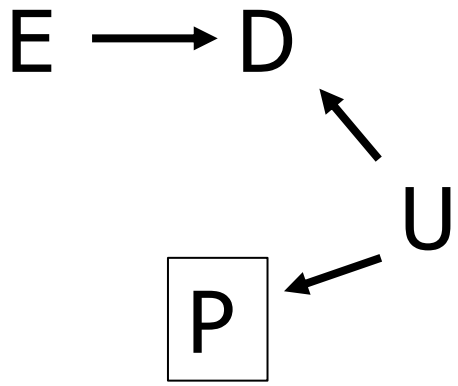
Participation		
E+	E-	Total
89%	72%	78%
72%	72%	72%
73%	72%	72%

Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol. 1977 Sep;106(3):184-7.

The structure of selection bias (2)

"Selection bias is a theoretical possibility whenever correlates of the outcome capable of influencing study participation are existent in some individuals at the beginning of the study."

"The inescapable conclusion is that collider stratification is not a necessary condition for selection bias."

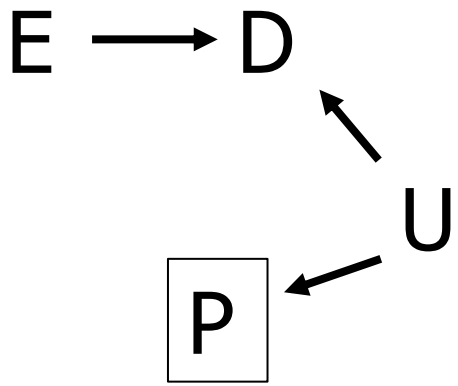


Hernán MA. Selection Bias Without Colliders. Am J Epidemiol 2017. 1048–1050,

Past perspective: Volunteer Bias

Cohorts created by allowing persons to volunteer to join (as opposed to invited to join) may be subject to this form of bias.

Volunteers are aware of their exposure history and may be aware of prognostic factors associated with the outcome.

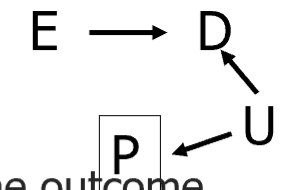


Future Challenges

Weighting to what?

Note that inverse probability of weighting of participation proportions for the study population recovers the unbiased estimate in the source population.

The structure of selection bias (2)



Selection was ~nondifferential because it was associated with the outcome (the odds ratio was 1.38) but not with the exposure (the odds ratio was 1.06)."

Participation		
E+	E-	Total
89%	72%	78%
72%	72%	72%
73%	72%	72%

Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol. 1977 Sep;106(3):184-7.

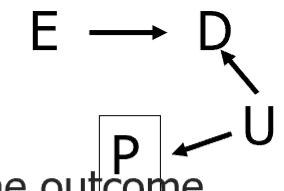
D+E+ weight = 1/0.89 = ~1.13
All others ~ 1/0.72 = ~1.89

Weighting to what?

But why weight study population to source population?

Why not weight to target population?

The structure of selection bias (2)



Selection was ~nondifferential because it was associated with the outcome (the odds ratio was 1.38) but not with the exposure (the odds ratio was 1.06)."

Participation		
E+	E-	Total
89%	72%	78%
72%	72%	72%
73%	72%	72%

Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol. 1977 Sep;106(3):184-7.

D+E+ weight = 1/0.89 = ~1.13
All others ~ 1/0.72 = ~1.89

Current / future selection bias topics

Selection bias as a continuum from target to study population and back again.

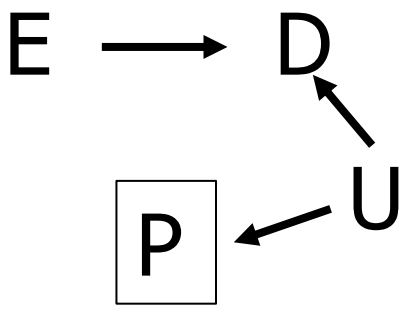
Implies change in thinking of internal and external validity; is it just one thing?

Implications for generalizability, transportability, representativeness, and generalization.

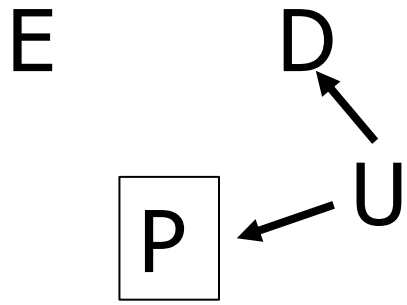

The End

The structure of selection bias (2)

"Yet, in the absence of collider stratification, selection bias is not guaranteed to arise."



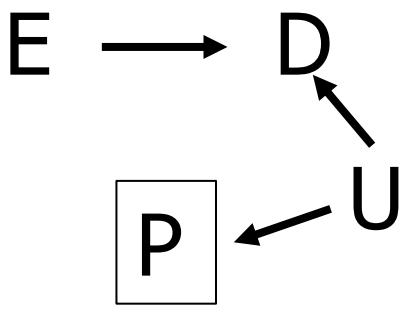
"E has a non-null causal effect on disease D for some individuals in the population (and therefore the causal risk ratio is different from 1)"

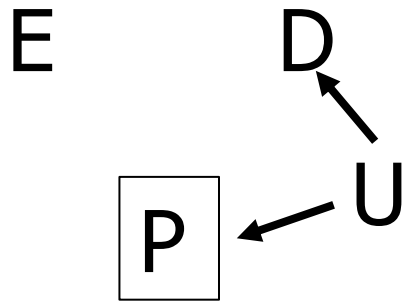


"exposure E has a null causal effect on the disease D of all individuals in the population (and therefore the causal risk ratio equals 1)"

The structure of selection bias (2)

"Yet, in the absence of collider stratification, selection bias is not guaranteed to arise."

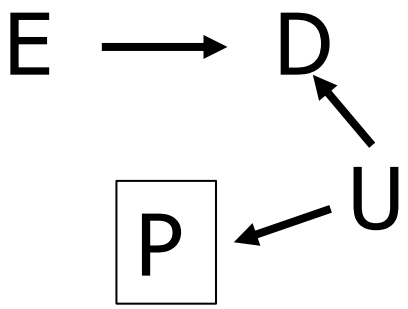


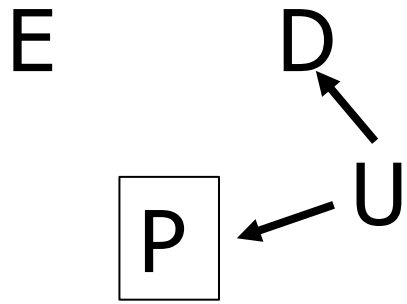


D-separation: "E and D are independent whether the analysis includes all individuals (no box around P) or is restricted to the uncensored individuals (box around P); that is, no selection bias arises when the sharp null hypothesis holds—the exposure has no effect on the outcome of any individuals—and censoring is independent of the exposure."

The structure of selection bias (2)

"Yet, in the absence of collider stratification, selection bias is not guaranteed to arise."

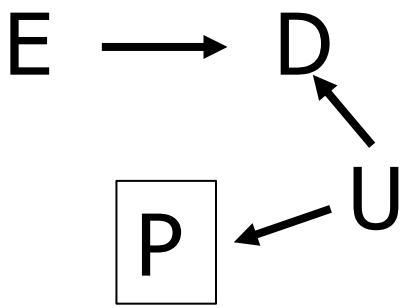




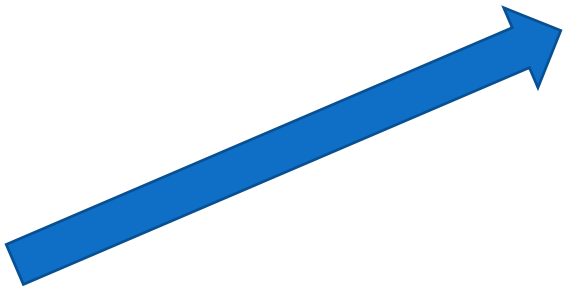
D-separation: "E and D are associated whether the analysis includes all individuals in the population (no box around P) or is restricted to the uncensored individuals (box around P). We will say that there is selection bias for the population parameter whenever the association in the uncensored individuals is different from the association in the entire population."

The structure of selection bias (2)

"Yet, in the absence of collider stratification, selection bias is not guaranteed to arise."



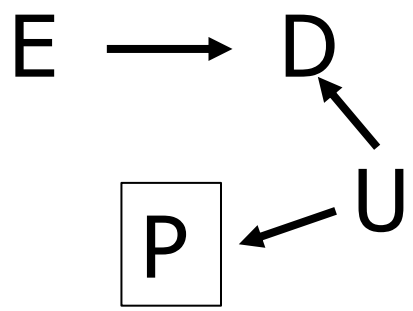
	Complete		Participants		Non-participants	
	E+	E-	E+	E-	E+	E-
D+	18	32	16	23	2	9
D-	232	718	167	517	65	201
Total	250	750	183	540	67	210
Risk	0.072	0.043	0.087	0.043	0.030	0.043
E+/E- RR	1.69		2.05		0.70	
P+/P- RR			2.93	0.99		



"conditioning on the noncollider P induces selection bias under the alternative hypothesis of a non-null effect of the exposure on the outcome, or selection bias off the null. Selection bias off the null further requires that the association between P and D varies across levels of E on the scale (e.g., risk ratio, risk difference) used to measure the population effect of E on D."

The structure of selection bias (2)

"Yet, in the absence of collider stratification, selection bias is not guaranteed to arise."



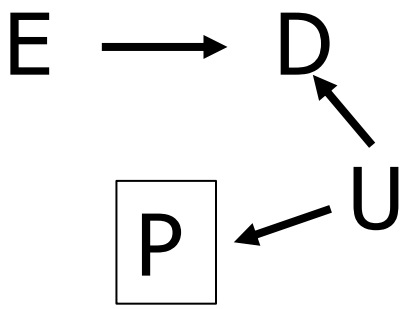
	Complete		Particpants		Non-Participants	
	E+	E-	E+	E-	E+	E-
D+	18	32	16	23	2	9
D-	232	718	167	517	65	201
Total	25	750	183	540	67	210
Risk	0.072	0.043	0.087	0.043	0.03	0.043
RR_{ED}	1.69		2.05		0.7	
P+/P- RR			2.93	0.99		

The bias is dependent on the measure used to estimate the effect: If instead, there had been no heterogeneity on the risk ratio scale, there would have been no selection bias for the population risk ratio, but there would have been selection bias for the population risk difference. This is because, off the null, no heterogeneity for the causal risk ratio implies heterogeneity for the causal risk difference. (recall this from interaction assessment in EPI540)

Causal directed acyclic graphs are nonparametric: Causal diagrams fail to depict selection bias off the null and thus cannot generally encode biases that depend on a particular parameterization of the effect.

The structure of selection bias (2)

"Yet, in the absence of collider stratification, selection bias is not guaranteed to arise."



	Complete		Participants		Non-participants	
	E+	E-	E+	E-	E+	E-
D+	18	32	16	23	2	9
D-	232	718	167	517	65	201
Total	250	750	183	540	67	210
Risk	0.072	0.043	0.087	0.043	0.030	0.043
E+/E- RR	1.69		2.05		0.70	
P+/P- RR			2.93	0.99		

Difference between selection bias and confounding: "This is also the reason why the distinction between bias under the null and bias off the null is important for selection bias but not for confounding. The presence of common causes of exposure and outcome is expected to induce an association (confounding bias) between treatment and outcome on all scales (risk ratio, risk difference, etc.), regardless of whether the exposure does or does not have an effect on the outcome."