Causal Inference

Center for Data Research and Analytics

Cause and Effect

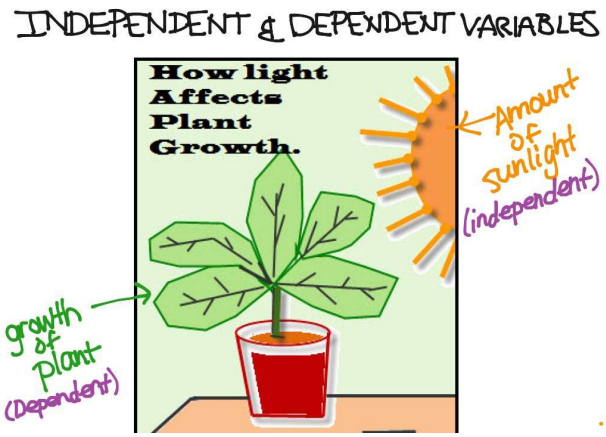
- Outcome: Any event may be defined as outcome. For example, a stroke, or developing a heart disease, or may be a GPA-5 in the SSC exam.
- Exposure: Any reason or factor that might affect the outcome is termed as exposure or exposure variable. For example, taking aspirin. Exposure varies among subjects.
- An exposure is said to be a cause of an outcome if the right-side conditions hold
- Causal association: In the analytic study, our intention is to identify a controllable exposure variable that causes an outcome. For example, intervention of smoking could reduce the prevalence of cancer in a country.

- *The level of exposure is determined before the occurrence of the disease (level of exposure and Chronology)*
- *Exposure X happened before the outcome Y*
- *Strong association between X and Y*
- *The other conditions for both exposed and unexposed will remain identical. No other things are responsible for the association*

We modify the exposure as an intervention to change the risk of an outcome.

Association/correlation

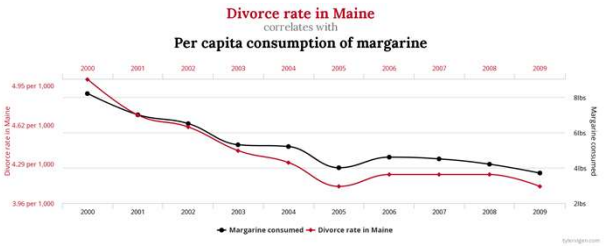
- Independent variable: A random variable who does not depend on the other variables. For example, time.
- Dependent Variable: Variable that depends on the other variables. For example, height.
- Correlation/Association: Association between variable A and B indicates whether increase/decrease in A can be reflected in B. You might be familiar with several method to estimate the association.

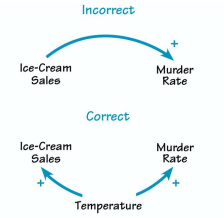


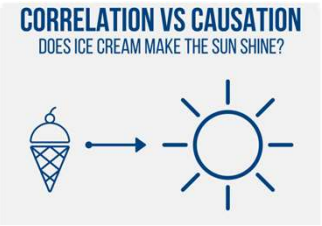
We use the measures of association to determine the causal effect as well

Correlation vs Causal Correlation

- Correlation does not necessarily mean causal association
- Sometimes correlations are spurious
- A correlation may be causal if the independent and independent variables follow the rules
  - Time condition
  - Strong association condition
  - Condition of no influence of the other variable







Additional care is needed to assess an association to be causal

Analytic study and measure of associations

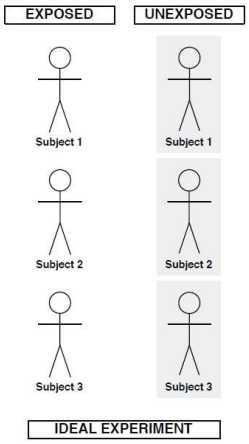
- Comparison group is involved in a study termed as analytical
- We have at least two groups of exposure
- We compare the risk of disease/any outcome between the exposed and unexposed group
  - Risk difference
    - a/(a+b) – c/(c+d)
  - Odds Ratio
    - ad/bc
  - Relative Risk/risk ratio
    - a/(a+b) ÷ c/(c+d)

	Disease	Not-disease	Total
Exposed	a	b	a + b
Not Exposed	c	d	c +d
Total	a + c	b + d	a + b + c + d

Sweetener	Bladder Cancer		
	Yes	No	Total
Yes	a	b	a + b
No	c	d	c +d
Total	a + c	b + d	a + b + c + d

Ideal causal experiment

- Suppose we want to understand the effect of coffee drinking on pancreatic cancer
- Take each subject, from a population, “expose” them to coffee drinking, and observe whether pancreatic cancer occurs over an appropriate follow-up period.
- Then, on the same individual, imagine turning back the clock to run the same experiment once more, but this time making the subject a coffee abstainer, measuring again if they develop pancreatic cancer or not.



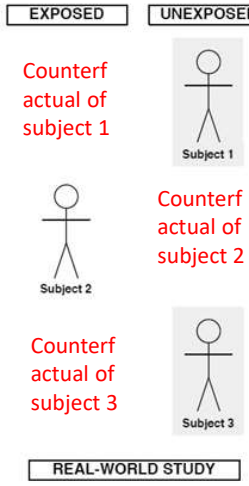
Group	Exposed	Not exposed	Number
1	D	D	Np1
2	D	D'	Np2
3	D'	D	Np3
4	D'	D'	Np4

$$RR_{Causal} = \frac{P(D/E)}{P(D/E')} = \frac{p_1 + p_2}{p_1 + p_3}$$

In the ideal world, causal association make sense easily.

Real world experiment

- But we can only observe one side of a subject.
- Unobserved response is called counterfactual
- Measures of causal effect depends on the distribution of counterfactuals
- Can we estimate the causal RR from this scenario?
- Theoretically, if the exposure is randomly distributed, then on average, the RR measured from this study design gives the causal association



Group	Exposed	Not exposed	Number
1	D		Np1
2		D'	Np2
3	D'		Np3
4		D'	Np4

Exposure allocation needed to be random to be a cause of an outcome

Study designs

We design our studies following the basic requirements of causality so that our observed association could translate into causal association.

The basic study designs in epidemiological research are as follows

- Observational
  - Cross Sectional
  - Case-Control study
  - Cohort Study
- Experimental
  - Randomized control trial

Cross-sectional study

At a time select a *random sample* from the study population and measure the exposure and outcome

		Disease Status		
		D	not D	
Coffee Drinking	E	7	52	59
	not E	7	134	141
		14	186	200

Cannot infer the temporal sequence of E-D

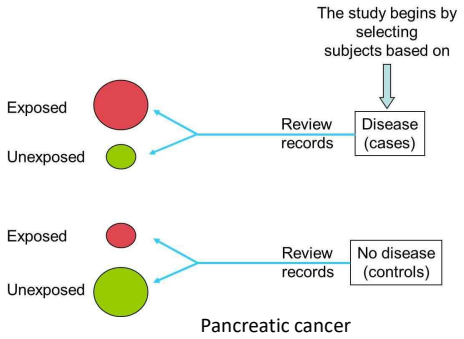
Proportion of disease among coffee drinker, $R_1 = \frac{7}{59} = 0.119$

Proportion of disease among coffee abstainer, $R_0 = \frac{7}{114} = 0.05$

$$RR = \frac{R_1}{R_0} = \frac{0.119}{0.05} = 2.39$$

Case-control Study

- Identify the two sub-group of population of diseased and not diseased
- Select a random samples from diseased and not diseased group separately
- Measure subsequently the presence and absence of E of everyone in the past in both samples
- We cannot estimate the RD or RR ass the total number of diseased person in each group is determined by a researcher



Pancreatic cancer

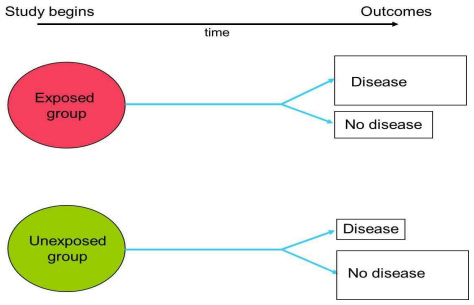
		Low	Normal	
Coffee Drinking	E	50	28	78
	not E	50	72	122
		100	100	200

Odds Ratio,

$$OR = \frac{ad}{bc} = \frac{50 \times 72}{50 \times 28} = 2.57$$

Prospective Cohort design

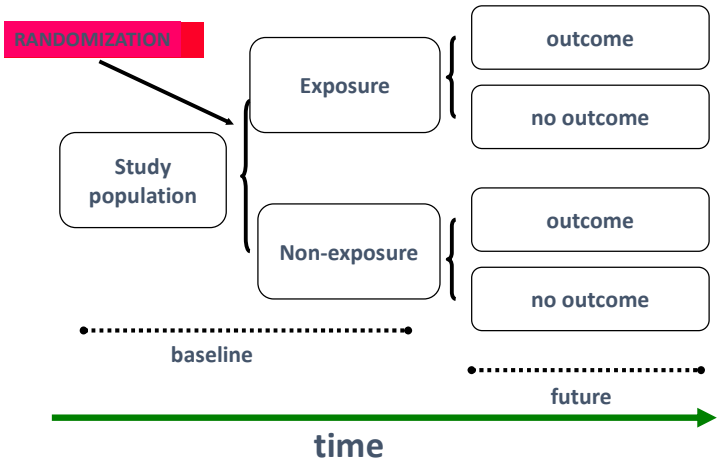
- Identify two sub-groups of the population based on their exposure level and follow-up
- Take a *random sample* from each of the two subgroups
- Measure subsequently the presence or absence of disease of the selected individuals of the both subgroups
- RD, RR and OR are valid estimate of causal association for cohort study design



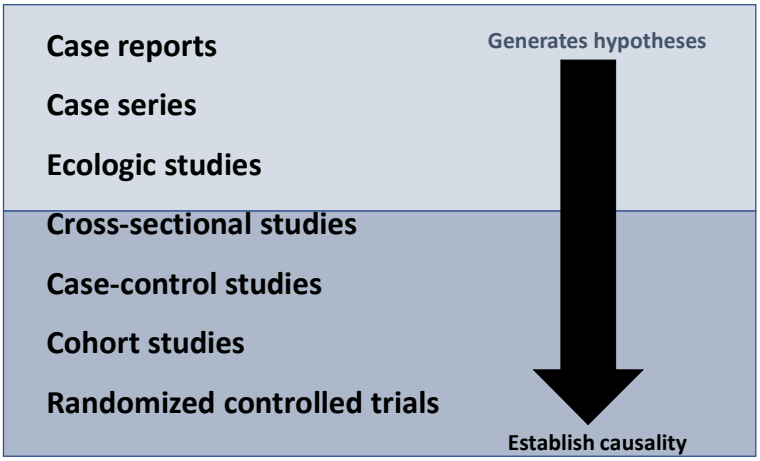
Pancreatic Cancer

		D	not D	
Coffee Drinking	E	12	88	100
	not E	5	95	100
		17	183	200

Randomized Control Trial



Hierarchy of Epidemiologic Study Designs



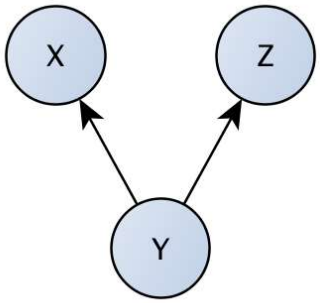
Tower & Spector, 2007

Evaluation of association

- Practically, we have data of exposure E (0, 1) and outcome D (0,1)
- The way we showed in the table is not computationally feasible under this data structure.
- We use regression, log-binomial/Poisson regression, and logistic regression to estimated RD, RR and RD respectively.
- In either model, E is considered as covariate, and D is the dependent variable.
- The estimated coefficient of E in regression model is the estimate of RD
- Similarly, for log-binomial and Poisson models.
- Logistic regression exp(coefficient) gives the odds ratio.
- You can include additional covariate to adjust for bias due to confounding.
- Adjust for outlier and robustness is also easier in the regression setup.
- Easily get the p-values and confidence intervals

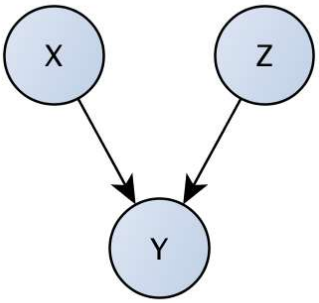
Control for extraneous factors

- In any study design and analysis, make sure you have included confounding variable in the model



- If not, despite no association, X and Z might become associated due to ignoring Z from the regression model (Confounding bias)

- Again, do not dump all variable as covariate that might have another bad consequence



- In this situation inclusion of Y in the regression covariate could hide the associate between X and Z.

Thanks

To all participants, Dr. Jaynal Abedin, and Dr. Rezaul Karim