

The Epidemiology Study Guide

Chia-Yen Chen

John Jackson

Chien-Chang Lee

Mitchell Machiela

Andreas Neophytou

Thank you L^AT_EX for making our dreams come true!

Special thanks to all the faculty, friends, and fellow epidemiologists (past and present) whose thoughts and ideas fill the pages of this guide.

*Draft version 21.10 - updated by Ellie M.
April 27, 2013*

Table of Contents

Table of Contents	3
Part 1. Measures & Designs	7
Chapter 1. Epidemiological Concepts	9
1. Sampling Variability & Estimation	9
2. Probability	10
3. Causation	11
Chapter 2. Measures of Frequency and Occurrence	13
1. Useful Concepts	13
2. Prevalence	14
3. Risk	16
4. Odds	19
5. Incidence Rate	20
Chapter 3. Measures of Association	23
Absolute Measures	23
1. Risk Difference	23
2. Rate Difference	24
3. Attributable Measures	24
4. Preventable Measures	28
Relative Measures	29
5. Risk Ratio	29
6. Incidence Rate Ratio	29
7. Odds Ratio	29
Chapter 4. Measures of Classification	33
1. Sensitivity & Specificity	33
2. Receiver Operating Characteristic Curves	35
Chapter 5. Study Designs	37
1. Study Population	37
2. Randomized Experiments	38
3. Observational Studies	39
4. Cohort Studies	39
5. Case-Control Sampling Schemes	41
6. Cross Sectional Studies	55
7. Ecological Studies	55
Part 2. Causal Inference	63

Chapter 6. Concepts & Definitions	65
1. Philosophical Viewpoints of Causality	65
2. Randomized Experiment Paradigm	66
3. Counterfactual Definition of Causal Effect	67
4. Conditions for Causal Inference	68
5. Effect Modification	73
6. Interaction	77
Chapter 7. Bias	87
1. Directed Acyclic Graphs: concepts	87
2. Confounding	100
3. Selection Bias	110
4. Information Bias	119
Chapter 8. Advanced Topics in Causal Inference	133
1. Directed Acyclic Graphs: theory	133
2. Direct and Indirect Effects	136
3. Static & Dynamic Observation Plans	139
4. Static & Dynamic Treatment Regimes	146
Part 3. Inference without Models	149
Chapter 9. Statistical Inference	151
1. Hypothesis Testing	151
2. Confidence Interval	151
3. Sources of Random Variability	152
Chapter 10. Traditional Methods	153
1. Road Map	157
2. Stratification	164
3. Matching	165
4. Matched Cohort study	167
5. Matched Case-Control study	167
6. Informative and non-informative risk sets	168
7. Analysis of Matched Case-control Studies	170
Chapter 11. G-Methods	173
1. Standardization	173
2. G-formula	179
3. Inverse Probability Weighting (IPW)	182
4. G-estimation	187
5. Instrumental Variables	188
Part 4. Inference with Models: Concepts & Methods	193
Chapter 12. Modeling Concepts	195
1. Modeling Concepts in Epidemiology	195
2. Why we need models	198
3. Stratification-based models	200

4. Model Specification	203
5. Hypothesis Testing & Estimation	203
6. Interaction in Regression Models	204
 Chapter 13. Time	209
1. Person-Time Analysis	209
2. Age-Period-Cohort Effects	210
3. Person-Time Data Structure	212
4. Time Scales & Cohorts	214
5. Time-Based Definitions of Exposure	216
 Chapter 14. Continuous Exposures & Confounders	221
1. Indicator Method	221
2. Score-Test	223
3. Fractional Polynomials	224
4. Splines	226
5. Comparisons	229
 Chapter 15. Missing Data	231
1. Types of Missing Data	231
2. Complete Case	233
3. Missing Indicator	234
4. Inverse Probability Weighting	235
5. Imputation	236
6. Maximum Likelihood	239
 Part 5. Inference With Models: Types	241
 Chapter 16. Models for Statistical Inference	243
1. Stratified Analysis using Regression Models	243
2. Linear Regression	244
3. Mantel-Hanszel as a Regression Function	244
4. Unconditional Logistic Regression	245
5. Poisson	246
6. Pooled Logistic Regression	250
7. Conditional Logistic Regression	251
8. Cox-Proportional Hazards Regression	253
9. Accelerated Failure Time Regression	255
 Chapter 17. Models for Statistical Inference: Special Topics	257
1. Model Selection for Case-control Data	257
2. Covariate Selection	258
3. Stratified Sampling Fractions	259
 Chapter 18. Models for Causal Inference	261
1. Parametric G-Formula	261
2. Marginal Structural Models	261
3. Structural Nested-Models	269
4. Propensity Scores	274

5. Instrumental Variables	276
6. G-Methods Analysis Example	276

Part 1

Measures & Designs

CHAPTER 1

Epidemiological Concepts

1. Sampling Variability & Estimation

1.1. Random Error.

- fluctuations in a parameter estimate
- we can never calculate the true average causal effect, we can only estimate it
 - sampling variability: random error due to the fact we are studying a subset of the population
 - nondeterministic (stochastic) counterfactuals: probability of a counterfactual event is not deterministic (0 or 1), but rather a probability between 0 and 1 (quantum mechanics)

1.2. Estimator.

- a function of observable sampled data used to approximate an unknown population parameter (the *estimand*)
 - **consistent estimator**: as the sample size approaches infinity, the estimated parameter approaches the truth
 - **biased estimator**: the expected value of the estimator does not equal the true value for the population parameter of interest

1.3. Efficiency.

- the precision of an estimation procedure
- the more efficient an estimator, the smaller its estimate of variance
 - *variance*: an estimation of uncertainty around a population parameter
 - *information*: inverse of the variance ($\frac{1}{variance}$)

1.4. Bias.

- the difference between the true value and the estimated value of a population parameter
- if the true value = expected value, there is no bias
- if the true value \neq expected value, there is bias

- bias can arise from:
 - (1) confounding
 - (2) selection bias
 - (3) information bias
- often there is a trade-off between bias and efficiency (ex: stratifying more finely will remove more bias, but result in less efficiency)
- for more information on bias, see Chapter 7

2. Probability

2.1. Variables.

- random variables: a variable whose values are realized from a probability distribution
 - It may have different values for different individuals.
 - For notation, an uppercase letter A (stands for action).
 - * continuous: may take any value
 - * discrete: can only be specified values (ex: integers)
 - polytomous: can take more than two values
 - dichotomous: can only take two values
- particular value: the realized value of a random variable. For notation, a lowercase letter a .

2.2. Probability.

- marginal (unconditional) probability
 - $Pr[Y = 1]$ “The probability that Y equals 1”
 - estimated for the entire sample
- joint probability
 - $Pr[Y = y, A = a] = Pr[Y = y \cap A = a]$
 - “The probability that $Y = y$ and $A = a$ ”
 - equals $Pr[Y = y|A = a] \times Pr[A = a] = Pr[A = a|Y = y] \times Pr[Y = y]$ (Bayes theorem)
 - estimated for the entire sample
- conditional probability
 - $Pr[Y = y|A = a] = Pr[Y = y, A = a]/Pr[A = a]$ “The probability that $Y = y$ given $A = a$ ”

- estimated in a subset of the sample (subjects with $A = a$)
 - * ex. $Pr[Y = y | A = a, L = l] = Pr[Y = y, A = a, L = l] / Pr[A = a, L = l]$

2.3. Independence.

- denoted $\perp\!\!\!\perp$
- $A \perp\!\!\!\perp Y = Y \perp\!\!\!\perp A, Pr[Y = 1 | A = 1] = Pr[Y = 1 | A = 0] = Pr[Y = 1]$
- information on one variable gives no information on another variable
- no correlation or association between variables
 - mean independence
 - * for a continuous outcome, $E[Y | A = 1] = E[Y | A = 0] = E[Y]$
 - * independence and mean independence are the same for a dichotomous outcome

2.4. Association.

- if there is an association $\rightarrow Pr[Y = 1 | A = 1] \neq Pr[Y = 1 | A = 0]$
- conditionally estimated (comparing the outcomes of two mutually exclusive subsets of the population under different scenarios)

3. Causation

- if there is a causal effect $\rightarrow Pr[Y^{a=1}] \neq Pr[Y^{a=0}]$
 - Y^a is defined as the outcome if everyone had received treatment $A = a$
- marginally estimated (comparing the outcomes of the whole population under different scenarios)

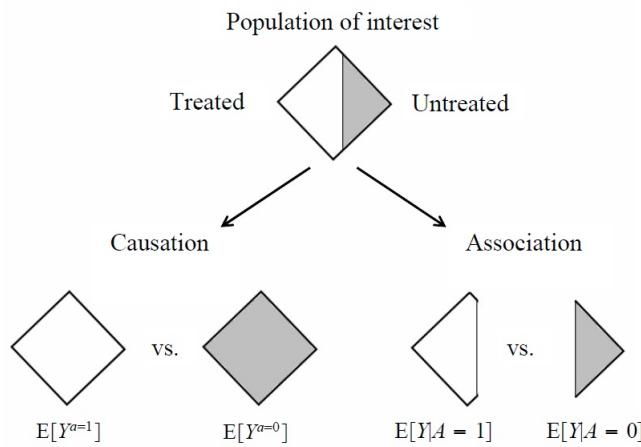


FIGURE 1.1. Causation is in the entire population, association is only in a subset of the population

CHAPTER 2

Measures of Frequency and Occurrence

1. Useful Concepts

1.1. Stationary Population (steady state assumption).

- the number entering the population is balanced by the number exiting population in *any* period of time within levels of all determinants of risk.
- no *net* migration
- allows calculation of the prevalence odds: $P/(1-P) = I \times D$, where P is prevalence, I is incidence, and D is disease duration
- any time period would be suitable for estimating the incidence rate because it is assumed to be *constant* over time

1.2. Person Time.

- is the amount of time an individual contributes to a group's observation, usually in a homogeneous state of risk
- has units of *person × time*
- 3 equivalent ways of calculating person-time
 - in a study of n subjects i , total person-time (PT) of observation =

$$(a) \sum_{i=1}^n [time]_i \quad (b) E[(time)_i] \times \sum_{i=1}^n i$$

$$(c) \sum time \times E \left[\sum_{i=1}^n i \right]$$

- (a) for each person, identify amount of person time contributed to group's observation, then sum the times of individual persons to get total person time for the group
- (b) multiply the number of persons under observation by the average duration of observation per person
- (c) multiply the length of the period of observation by the average number of persons under observation during the period

- classification of at-risk person-time by exposure status and other covariates assumes that the incidence rate of outcome is constant within these strata

1.3. Risk-Set.

- group of cohort members who would have been observed to have event, had they done so at the moment at which the event occurs.
 - unadjusted analysis: Kaplan-Meier
 - adjusted: Mantel-Haenszel matched on time
 - multivariate: Cox Proportional Hazards Regression
- when...
 - no event or censored $\Rightarrow S(t) = 1$
 - event occurs $\Rightarrow S(t) = (n - 1)/n$ and $F(t) = 1/n$
- assumption is that underlying hazard is unchanging over the interval of the risk set

2. Prevalence

Definition.

- the total number of individuals who have an attribute or disease at a particular time (or period) divided by the total population under study
- there are several types. Each corresponds to the proportion of individuals with disease or condition
 - **Point:** at a particular point in time
 - **Period:** during a specified period of time. Includes cases that (a) arise before but extend into the interval of study or (b) arise during the interval of study
 - **Lifetime:** for at least part of their lives at *any time* during their life
- time period and population must be specified

Characteristics.

General Properties.

- range $[0, 1]$
- a proportion (dimensionless)
- binomial distribution used for inference
- is a population parameter

Challenges for Causal Inference.

- is sensitive to causal and non-causal factors
- \uparrow occurrence $= \uparrow$ prevalence
- \uparrow duration (survival) $= \uparrow$ prevalence
- sampling prevalent cases can select on consequence of outcome and introduce bias (see *case-control* and *selection bias sections*)

Challenges for Public Health Planning.

- measures extent of disease in population
- short duration and survival rate may mask disease burden

Estimation.

- from cross-sectional or longitudinal data
 - prevalent case: existing case or new case arising during specified time interval (counting a person only once)
 - $\frac{\# \text{ prevalent cases}}{\text{total population}}$ during a specified interval of time
 - $Pr[Y = 1]$ for a dichotomous Y {1 = prevalent case, 0 = non-prevalent case}
- assuming steady state:
 - let N be population size, P be prevalence pool (# prevalent cases in population)
 - let I be incidence rate of disease, I' be incidence of exit from prevalence pool, \bar{D} be average duration of disease

inflow to prevalence pool = outflow from prevalence pool

$$I \times (N - P) \times \Delta \text{time} = I' \times P \times \Delta \text{time}$$

$$\begin{aligned} I \times (N - P) \times \Delta \text{time} &= \left(\frac{1}{\bar{D}}\right) \times P \times \Delta \text{time} \\ \frac{P}{N - P} &= I \times \bar{D} \end{aligned}$$

$$-\frac{\text{prevalence}}{1 - \text{prevalence}} = \text{Incidence Rate} \times \text{Average Duration} \implies \frac{P}{1 - P} = I \times \bar{D}$$

$$-\text{applying rare disease assumption } p/(1 - p) \approx p$$

$$* \text{prevalence} \cong \text{Incidence Rate} \times \text{Average Duration} \implies P \cong I \times \bar{D}$$

- note that “steady state assumption” refers to constant prevalence over time period
- this method is not too useful in practice due to unrealistic constraints

- does not apply to age-specific prevalence unless assumptions hold for that particular age group

3. Risk

Definition.

- has many synonyms and can refer individual or average risk:
 - **risk:** an individual's probability of developing disease or attribute of interest *during a specified time interval*. Can take values of either 0% or 100%.
 - **incidence proportion:** the average risk in a group of individuals during a specified time interval. Can take values between 0% and 100% (inclusive).
 - survival proportion is complement of incidence proportion $\implies S = 1 - R$. The average probability of not having event during specified time interval. Can take values between 0% and 100%.
 - **cumulative incidence:** synonym for incidence proportion
- average risk (incidence proportion, cumulative incidence) is estimated from population data and is used to *estimate* the risk experienced by individuals. It is usually just referred to as "risk".

Characteristics.

3.0.1. General Properties.

- range $[0, 1]$, proportion (dimensionless)
- binomial distribution used for inference
- makes no reference to when events occurred during follow up. Estimates of risk from different studies may relate to different patterns of disease accrual over time.
- risk accumulates over follow up time
 - towards 1 for inevitable outcome
- need to specify:
 - (1) time period of observation
 - (2) population at risk
- risks are only comparable across studies if reference time is the same (ex: 1% one-year risk vs. 1% lifetime risk)

Challenges for Causal Inference.

The incidence proportion is not well defined when there are:

- **competing risks:** death from another cause before occurrence of event.

- the resulting risk estimate will be *lower* than the true risk. (if competing risk unrelated to exposure will bias *risk ratio* towards null).
- the magnitude of this problem increases with follow up length, so it can be ignored if time of follow up is short
- precludes use Kaplan-Meier, exponential formula, and many survival methods because identifiability conditions are harder to justify for competing risks. Unlike loss-to-follow up, the treatment to remove a competing risk is not well defined, making it harder to assume exchangeability between the censored and uncensored
- **loss to follow up:** no ascertainment of outcome in individuals who drop out or otherwise cannot be contacted
- **open cohort:** open if new members are added to the cohort during follow up and/or some persons are right-censored (exit without occurrence of event).

Challenges for Public Health Planning.

- can mask pattern of disease accrual over time if follow up is over a long time
- since it's directly proportional to follow up time, it is interpretable only if time-frame specified

Estimation.

(1) Closed Cohort

- incident case: a case that arose during the period of follow-up
- $$\frac{\text{# of incident cases between } t_0 \text{ and } t_1}{\text{# of at risk subjects at } t_0 \text{ followed from } t_0 \text{ and } t_1}$$
- probability of disease occurrence during follow up
 - (a) marginal $\rightarrow (Pr[Y = 1])$
 - (b) conditional $\rightarrow (Pr[Y = 1|A = 1])$
- if constant incidence rate:

$$\text{incidence proportion} = \text{incidence rate} \times \text{length of follow up}$$

$$R = I \times \Delta T$$

- logistic regression $Pr[Y = 1|X] = \frac{\exp(\beta X)}{1 + \exp(\beta X)}$

(2) Open Cohort, Loss to Follow Up, & Variable Incidence Rate

- **Kaplan-Meier or Product-Limit Method:**

$$R = 1 - \left(\prod_{k=1}^K (1 - R_k) \right)$$

Necessary assumptions:

- (a) closed population (if open, censoring is random)
- (b) event is inevitable (i.e. no competing risks)
- (c) intervals k have constant incidence rates

Procedure:

- (a) divide follow-up time into K intervals $1, 2, \dots, k$
- (b) From incidence rate and interval length, calculate incidence proportion for each interval

$$\text{Incidence Proportion}_k = \text{Incidence Rate}_k \times \Delta t_k$$
- (c) Convert to survival proportions for each interval

$$\text{Survival Proportion}_k = 1 - \text{Incidence Proportion}_k$$
- (d) Calculate average survival proportion over entire follow up (i.e. joint probability of survival during each interval)

$$\text{Survival} = \prod_{k=1}^K \text{Survival Proportion}_k$$
- (e) Calculate average incidence proportion over entire follow up

$$\text{Incidence Proportion} = 1 - \text{Survival Proportion}$$

• **Exponential Formula:**

$$\text{Incidence Proportion} \approx 1 - \exp \left(- \left[\sum_{k=1}^K \text{Incidence Rate}_k \times \Delta t_k \right] \right)$$

Necessary assumptions:

- (a) closed population (if open, censoring is random)
- (b) event is inevitable (i.e. no competing risks)
- (c) $\text{Risk} = I \times \Delta t$ is small. Can be satisfied by stratifying follow up time into very fine intervals k such that each event occurs at its own unique time

Derivation and further forms:

$$\ln(1) = 0$$

Let us stratify time so finely that $R_k = I_k \times \Delta t_k$ is very small

$$\begin{aligned} \ln(1 - R_k) &\approx -R_k \\ (1 - R_k) &\approx \exp(-R_k) \end{aligned}$$

Substituting into product limit gives exponential formula...

$$\begin{aligned}
 R_{\Delta t} &= 1 - S_{\Delta t} \\
 &= 1 - \left(\prod_{k=1}^K (1 - R_k) \right) \\
 &\approx 1 - \exp \left(- \sum_{k=1}^K R_k \right) \\
 &\approx 1 - \exp \left(- \sum_{k=1}^K [I_k \times \Delta t_k] \right)
 \end{aligned}$$

(a) If $I_k = I$ for all k (constant incidence rate) then...

$$R_{\Delta t} \approx 1 - \exp(-I \times \Delta t)$$

(b) If $R_{\Delta t} < .10$ (i.e. I or Δt small) then... (Nelson-Aalan estimator)

$$R_{\Delta t} \approx \sum_{k=1}^K [I_k \times \Delta t_k]$$

(c) If $I_k = I$ for all k and $R_{\Delta t} < .10$ (i.e. I or Δt small) then... (cocktail party estimator)

$$R_{\Delta t} \approx I \times \Delta t$$

Save this formula for the Annual Epidemiology Department Winter Cocktail Party and stick to Kaplan-Meier approach (product-limit method)

(3) Competing Risks

- *independent competing risks* assumption: interval specific rates would not change if competing risks were removed
- necessary for product-limit method and exponential formula in the presence of competing risks

Relation to other Frequency/Occurrence measures. The hazard rate (or instantaneous incidence rate) is often referred to as *instantaneous risk* because it estimates the limit of the probability of event occurrence as $\Delta t \rightarrow 0$

4. Odds

- $[0, \infty]$, ratio (dimensionless)
- Odds = $P/(1-P)$
- Prob = Odds / (1+Odds)

- $\frac{Pr[Y=1|A=1]}{Pr[Y=0|A=1]} = \frac{Pr[Y=1|A=1]}{1-Pr[Y=1|A=1]}$
- ratio of probability of success to probability of failure
- estimates risk when disease is rare (i.e. $Pr[Y = 1] < 10\%$)
- logistic regression: $\frac{P}{(1-P)} = \exp(\beta X)$

5. Incidence Rate

Definition.

- as with risk, there are instantaneous and average forms

(1) Average Incidence Rate:

- the average rate of occurrence of disease in a population over a specified period of time
- number of incident cases per unit of person-time of follow up

(2) Incidence Density: The rate of disease occurrence over an infinitely small period of time

- synonym: Instantaneous Incidence Rate
- synonym: Hazard
- proper interpretation of its value requires selection of a time unit. Unit usually chosen so at least one digit to left of decimal place.
- is the limiting value of the incidence proportion as $\Delta t \rightarrow 0$. In this sense it is synonymous with *instantaneous risk* (i.e. risk per unit time, where units of time are infinitesimally small)

Characteristics.

5.0.2. General Properties.

- $[0, \infty]$, rate (units of inverse time)
- $1/\text{incidence rate} = \text{average waiting time}$ (time until disease)
- Poisson distribution usually used for inference

Causal Inference & Public Health Utility.

- is well defined even with
 - an open cohort
 - loss to follow up
 - competing risks
- not as intuitive as the concept of risk

Estimation.

- *average incidence rate* = $\frac{\# \text{ incident cases between } t_0 \text{ and } t_1}{\text{total person-time at risk from } t_0 \text{ to } t_1}$
 - only incident cases in which PT was contributed to denominator
- Poisson regression: $\lambda = \exp(\beta X)$
- *instantaneous incidence rate* = $\lim_{\Delta T \rightarrow 0} (Pr[Y_{t+1} = 1 | Y_t = 0]) / \Delta T$

Relation to other frequency/occurrence measures.

- see *prevalence*
- see *risk*

Measure	Measures of Disease Occurrence			
	Prevalence	Cumulative Incidence	Survival	Incidence Rate
Minimum Value	0	0	0	0
Maximum Value	1	1	1	∞
Incorporation of time	None	Specified follow-up	Specified follow-up	Incremental
Dimension	Cases per population	Cases per population after specified time	Non-diseased per population after specified time	Cases per population per unit time
Parameter	Probability	Probability	Hazard	
Probability Distribution	Binomial	Binomial	Poisson	
Variance	$\frac{Pr(1-Pr)}{N}$	$\frac{CI(1-CI)}{N}$	$\frac{S(1-S)}{N}$	$\frac{IR}{P}$

TABLE 2.1. Measures of disease occurrence, adapted from Walker, AM. 1991 Observation and Inference.

CHAPTER 3

Measures of Association

Measures	Association measures (conditional)	Causal effect measures (marginal)
Risk Difference	$Pr[Y = 1 A = 1] - Pr[Y = 1 A = 0]$	$Pr[Y^{a=1} = 1] - Pr[Y^{a=0} = 1]$
Risk Ratio	$Pr[Y = 1 A = 1]/Pr[Y = 1 A = 0]$	$Pr[Y^{a=1} = 1]/Pr[Y^{a=0} = 1]$
Odds Ratio	$\frac{Pr[Y=1 A=1]/(1-Pr[Y=1 A=1])}{Pr[Y=1 A=0]/(1-Pr[Y=1 A=0])}$	$\frac{Pr[Y^{a=1}=1]/(1-Pr[Y^{a=1}=1])}{Pr[Y^{a=0}=1]/(1-Pr[Y^{a=0}=1])}$

TABLE 3.1. Measures of association and causal effect

Absolute Measures

- absolute difference in risk, rate or odds
- effect is on the additive scale (e.g. exposure adds to the baseline risk/rate)

1. Risk Difference

- entire range $[-1, 1]$
 - $[-1, 0]$ for preventative effect
 - $[0, 1]$ for causative effect

1.1. Number Needed to Treat. The average number of individuals needed to receive treatment a to reduce the number of cases by one

$$\frac{1}{|Preventative\ Risk\ Difference|}$$

1.1.1. *Number Needed to Harm.* The average number of individuals needed to receive treatment a to increase the number of cases by one

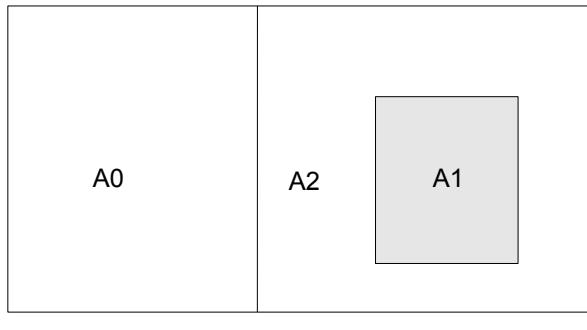
$$\frac{1}{|Causative\ Risk\ Difference|}$$

2. Rate Difference

- entire range $[-\infty, \infty]$
 - $[-\infty, 0]$ for preventative effect
 - $[0, \infty]$ for causative effect

3. Attributable Measures

3.1. Types of exposed cases.



A0: non-attributable cases
 A1: excess cases
 A2: etiologic but not excess cases
 A1+A2: all etiologic cases (including all excess cases)

FIGURE 3.1. Definitions of types of cases for attributable fraction. All individuals in the figure are exposed cases.

- A_0 : number of exposed cases that the exposure had no impact on the cases' incidence time or disease outcome, *non-attributable cases*
- A_1 : number of exposed cases where the subject will be a case if exposed but the subject will not become a case if unexposed, *excess cases*
- A_2 : number of exposed cases that the exposure made the cases' incidence time earlier than it would have been in the absence of exposure; the subject will eventually become a case even without exposure, *etiological cases* but not *excess cases*
- $A_1 + A_2$: all *etiological cases*
- $M = A_0 + A_1 + A_2$: total number of *exposed cases*
- $A_{exposed} = M = A_0 + A_1 + A_2 = \sum_i \{Y_i^{a=1} = 1 | A = 1\}$, cases that would have occurred if everyone exposed were actually exposed
- $A_{unexposed} = A_0 + A_2 = \sum_i \{Y_i^{a=0} = 1 | A = 1\}$, cases that would have occurred if everyone exposed had not been exposed

3.2. Attributable Risk.

- a measure of the contribution of an exposure A to the total incidence in the *exposed*
- defined as contrast between observed risk in the exposed vs. the counterfactual risk in the exposed had exposure been withheld
- alternate definitions
 - the difference in risk comparing exposed to unexposed
 - the risk associated with exposure among the exposed
 - if causal, the change in average risk that would occur in the exposed had they not been exposed
- there are analogous definitions for the entire population and rate differences (later)
- only includes excess cases A_1 . Does not include non-excess etiologic cases A_2 .

Formally speaking

- *regular notation:*

$$\frac{A_{exposed}}{N_{exposed}} - \frac{A_{unexposed}}{N_{unexposed}} = Risk_{exposed} - Risk_{unexposed}$$

- the two are only equal if the identifiability conditions for causal inference are met
- *counterfactual notation:*

$$\begin{aligned} & Pr[Y^{a=1} = 1 | A = 1] - Pr[Y^{a=0} = 1 | A = 1] \\ & \implies Pr[Y^{a=1} = 1 | A = 1] - Pr[Y^{a=0} = 1 | A = 0] \text{ (by exchangeability)} \\ & \implies Pr[Y = 1 | A = 1] - Pr[Y = 1 | A = 0] \text{ (by consistency)} \end{aligned}$$

3.3. Attributable Risk Fraction. Due to the differences between case types, the attributable fraction includes two distinct concepts:

(1) **etiologic fraction:**

- regular notation: $\frac{A_1 + A_2}{M}$
- interpreted as proportion of cases whose disease originated from a biologic process in which treatment had an effect (i.e. change in survival time or disease status)
- the etiologic fraction is non-identifiable because it requires extremely strong exchangeability assumptions

(2) **excess fraction:** the fraction of disease in the exposed that would be eliminated if the exposed were unexposed

- regular notation: $\frac{A_1}{M} = \frac{A_0 + A_1 + A_2}{M} - \frac{A_0 + A_2}{M} = \frac{A_{exposed} - A_{unexposed}}{A_{exposed}}$

- For a closed cohort, given the sample size of the entire cohort (N); the exposed risk ($Risk_{exposed}$); the unexposed risk ($Risk_{unexposed}$); the risk ratio ($RR = \frac{Risk_{exposed}}{Risk_{unexposed}}$), the excess fraction is given as:

$$\begin{aligned} \frac{A_{exposed} - A_{unexposed}}{A_{exposed}} &= \frac{A_{exposed}/N_{exposed} - A_{unexposed}/N_{unexposed}}{A_{exposed}/N_{exposed}} \\ &= \frac{A_{exposed}/N - A_{unexposed}/N}{A_{exposed}/N} = \frac{Risk_{exposed} - Risk_{unexposed}}{Risk_{exposed}} = \frac{RR - 1}{RR} \end{aligned}$$

- *counterfactual notation*: excess fraction in the treated

$$\begin{aligned} \frac{Pr[Y^{a=1} = 1|A = 1] - Pr[Y^{a=0} = 1|A = 1]}{Pr[Y^{a=1} = 1|A = 1]} &\implies \frac{Pr[Y = 1|A = 1] - Pr[Y^{a=0} = 1|A = 1]}{Pr[Y = 1|A = 1]} \\ &\implies \frac{Pr[Y = 1|A = 1] - Pr[Y = 1|A = 0]}{Pr[Y = 1|A = 1]} \\ &\implies \frac{Risk_{exposed} - Risk_{unexposed}}{Risk_{exposed}} \end{aligned}$$

- thus, the attributable fraction only has a causal interpretation if $Risk_{unexposed}$, $Risk_{exposed}$, and RR are causal
- can be estimated from case-control data if $OR \approx RR$

3.4. Extension to the entire population.

- a measure of the contribution of an exposure A to the total incidence in the *entire population*
- defined as contrast between observed risk in the entire population vs. the counterfactual risk in the entire population had exposure been withheld
- as before, these definitions only have causal interpretations if their components are causal; note that $(Risk_{exposed} - Risk_{unexposed})$ and $(\frac{Risk_{exposed} - Risk_{unexposed}}{Risk_{exposed}})$ are respectively, the *excess risk* and *excess fraction* as defined above

(1) Population Attributable Risk:

- *regular notation*:

$$\begin{aligned} Risk_{entire\ population} - Risk_{unexposed} \\ = (Risk_{exposed} - Risk_{unexposed}) \times \underline{Prev.\ of\ Exp\ in\ Population} \\ = excess\ risk \times \underline{Prev.\ of\ Exp\ in\ Population} \end{aligned}$$

- *counterfactual notation:*

$$\begin{aligned}
& Pr[Y = 1] - Pr[Y^{a=0} = 1] \\
&= Pr[Y = 1] - Pr[Y = 1|A = 0] \text{ (by consistency)} \\
&= (Pr[Y = 1|A = 1] \times Pr[A = 1] + Pr[Y = 1|A = 0] \times Pr[A = 0]) - Pr[Y = 1|A = 0] \\
&= (Pr[Y = 1|A = 1] \times Pr[A = 1] + Pr[Y = 1|A = 0] \times (1 - Pr[A = 1])) - Pr[Y = 1|A = 0] \\
&= Pr[Y = 1|A = 1] \times Pr[A = 1] + Pr[Y = 1|A = 0] - Pr[Y = 1|A = 0] \times Pr[A = 1] - Pr[Y = 1|A = 0] \\
&= Pr[Y = 1|A = 1] \times Pr[A = 1] - Pr[Y = 1|A = 0] \times Pr[A = 1] \\
&= (Pr[Y = 1|A = 1] - Pr[Y = 1|A = 0]) \times Pr[A = 1] \\
&= (Pr[Y^{a=1} = 1|A = 1] - Pr[Y^{a=0} = 1|A = 1]) \times Pr[A = 1] \text{ (by exchangeability)} \\
&= \text{excess risk} \times \text{Prev. of Exp in Population}
\end{aligned}$$

- excess risk of disease in the population that could be eliminated if exposure were eliminated from the population

(2) Population Attributable Risk Fraction:

- *regular notation:*

$$\begin{aligned}
& \frac{Risk_{entire\ population} - Risk_{unexposed}}{Risk_{entire\ population}} \\
&= \left(\frac{Risk_{exposed} - Risk_{unexposed}}{Risk_{exposed}} \right) \times \text{Prev. Exp in Cases} \\
&= \text{excess fraction} \times \text{Prev. Exp in Cases}
\end{aligned}$$

- *counterfactual notation:*

$$\begin{aligned}
\frac{Pr[Y = 1] - Pr[Y^{a=0} = 1]}{Pr[Y = 1]} &= \left(\frac{Pr[Y^{a=1} = 1|A = 1] - Pr[Y^{a=0} = 1|A = 1]}{Pr[Y^{a=1} = 1|A = 1]} \right) \times Pr[A = 1|Y = 1] \\
&\Rightarrow \left(\frac{Pr[Y = 1|A = 1] - Pr[Y^{a=0} = 1|A = 1]}{Pr[Y = 1|A = 1]} \right) \times Pr[A = 1|Y = 1]
\end{aligned}$$

- fraction of disease in the population that could be eliminated if exposure were eliminated from the population

3.5. Extension to Rates.

- Using risk to calculate excess fraction is fine because risk ratios can be translated into the ratio of numbers of cases
- using rate to calculate excess fraction is not correct
 - rates cannot be translated readily into the number of cases
 - use of rates results in the *incidence-density fraction*

Analogous Attributable Measures for Rates

- **Attributable Rate:** the excess rate of disease that would be removed from the exposed if the exposure had been withheld

$$Rate_{exposed} - Rate_{unexposed}$$

- **Attributable Rate Percent:** the fraction of disease associated with exposure among the exposed

$$\frac{Rate_{exposed} - Rate_{unexposed}}{Rate_{exposed}} = \frac{IRR - 1}{IRR}$$

- **Population Attributable Rate:** the excess rate of disease that would be removed from the entire population had exposure been withheld

$$\begin{aligned} & Rate_{entire\ pop.} - Rate_{unexposed} \\ &= (Rate_{exposed} - Rate_{unexposed}) \times \text{Prev. of Exposed PT in } \underline{\text{Population}} \end{aligned}$$

- **Population Attributable Rate Percent:** the fraction of disease associated with exposure among the entire population

$$\begin{aligned} & \frac{Rate_{entire\ pop.} - Rate_{unexposed}}{Rate_{entire\ population}} \\ &= \left(\frac{Rate_{exposed} - Rate_{unexposed}}{Rate_{exposed}} \right) \times \text{Prev. of Exposure PT in } \underline{\text{Cases}} \end{aligned}$$

4. Preventable Measures

- for when exposure is believed to protect against disease
- should be interpreted with caution as all or part of the apparent effect may be due to other factors associated with the apparent protective factor
- **Population Preventable Fraction:**

- the proportion of disease (in the population) that would be prevented if the whole population were exposed to the factor

$$= \frac{Rate_{population} - Rate_{exposed}}{Rate_{population}}$$

- **Population Prevented Fraction**

- the proportion of the hypothetical total load of disease in the population that has been prevented by exposure to the factor

$$= \frac{Rate_{unexposed} - Rate_{population}}{Rate_{unexposed}}$$

Relative Measures

- based on the ratio of risks, rates, or odds
 - effect is on the multiplicative scale (e.g. exposure multiplies the baseline risk/rate)
- “Relative Risk” is often used to refer to the following effect measures:

5. Risk Ratio

- $[0, \infty]$, multiplicative scale (*see figure 3.3*)
 - preventative $[0, 1]$
 - causative $[1, \infty]$
- ideal for causal inference, harder to estimate using models

6. Incidence Rate Ratio

- $[0, \infty]$, multiplicative log scale
 - preventative $[0, 1]$
 - causative $[1, \infty]$
- not ideal for causal inference, easier to estimate using models
- average hazard ratio can mask protective or hazardous period specific hazard ratios
- period-specific hazard ratios have built in selection bias. Depletion of susceptibles among exposed can make exposure look protective towards the end of the study
- the hazard ratio is sensitive to length of follow up and studies of different length can report different findings

7. Odds Ratio

- $[0, \infty]$, multiplicative log scale
 - preventative $[0, 1]$
 - causative $[1, \infty]$
- below risk of 0.1, odds approximates risk ($OR \cong RR$), rare disease assumption
- OR will always be further from null than RR
- not ideal for causal inference, easier to estimate using models
- less interpretable
- non-collapsible

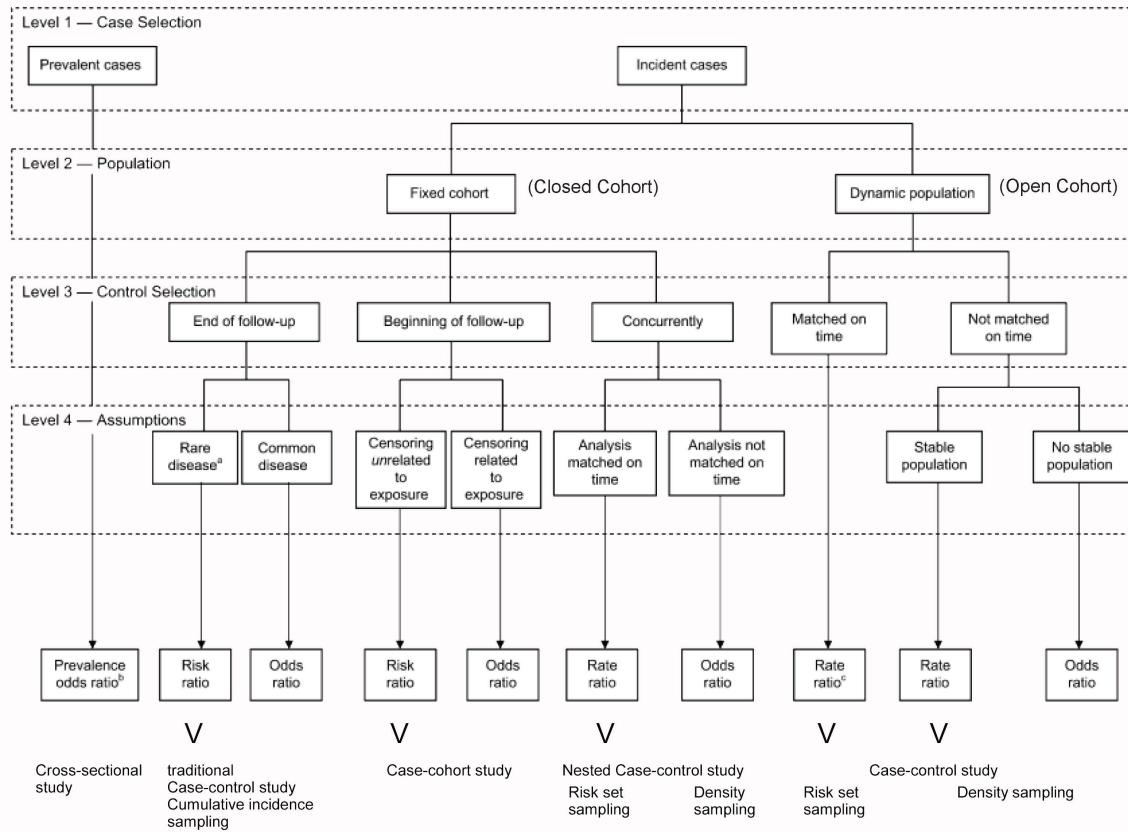


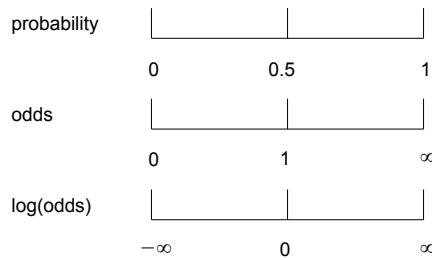
FIGURE 3.2. Classification of estimates from different study designs and sampling schemes

– is only a weighted average of stratum specific effects when there is NO heterogeneity

7.1. Interpretation of ORs from Case-Control Studies.

- in a case-control study, the OR can approximate various effect measures depending on the sampling scheme for controls and the underlying assumptions made (see Figure 3.2)

Odds ratio:
use logistic regression as canonical transformation



Rate ratio:
use Poisson regression as canonical transformation

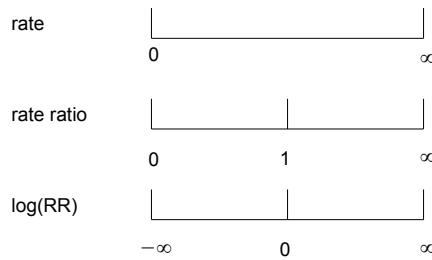


FIGURE 3.3. Transformation of odds ratio and risk ratio scales

CHAPTER 4

Measures of Classification

		Truth		
		$Y = 1$	$Y = 0$	
Test	$Y^* = 1$	T_p	F_p	
	$Y^* = 0$	F_n	T_n	
		$P[Y]$	$1 - P[Y]$	N

TABLE 4.1. Confusion Matrix

1. Sensitivity & Specificity

Measurement $Y^* : \{0 = \text{Negative}, 1 = \text{Positive}\}$

Truth $Y : \{0 = \text{Negative}, 1 = \text{Positive}\}$

- (1) **Sensitivity (Sn):** the probability of a positive test given that the truth is positive, also known as the *true positive rate* (TPR)

$$= Pr(Y^* = 1|Y = 1)$$

$$= \frac{T_p}{T_p + F_n}$$

- *False Negative Rate (FNR):* the probability of a negative test given that the truth is positive $= Pr(Y^* = 0|Y = 1) = 1 - Sn$

- $TPR + FNR = 1$

- (2) **Specificity (Sp):** the probability of a negative test given that the truth is negative, also known as the *true negative rate* (TNR)

$$= Pr(Y^* = 0|Y = 0)$$

$$= \frac{T_n}{T_n + F_p}$$

- *False Positive Rate (FPR):* the probability of a positive test given that the truth is negative $= Pr(Y^* = 1|Y = 0) = 1 - Sp$

- $TNR + FPR = 1$

- (3) **Positive Predictive Value (PPV):** The probability of an individual classified by the test as having the disease, actually having the disease. The probability of the truth being positive given a positive test = $Pr(Y = 1|Y^* = 1)$

$$\begin{aligned}
 &= \frac{Sn \times Pr[Y=1]}{(Sn \times Pr[Y=1]) + ((1-Sn) \times (1-Pr[Y=1]))} \implies \text{depends on the prevalence of } Y \\
 &= \frac{Tp}{(Tp+Fp)} \\
 \bullet \quad &Pr[Y = 1] = \frac{Tp+Fn}{Tp+Fn+Fp+Tn}
 \end{aligned}$$

- (4) **Negative Predictive Value (NPV):** The probability of an individual classified by the test as disease-free, actually being disease-free. The probability of the truth being negative given a negative test = $Pr(Y = 0|Y^* = 0)$

$$\begin{aligned}
 &= \frac{Sp \times (1-Pr[Y=1])}{(Sp \times (1-Pr[Y=1])) + ((1-Sp) \times Pr[Y=1])} \implies \text{depends on the prevalence of } Y \\
 &= \frac{Tn}{(Tn+Fn)}
 \end{aligned}$$

- (5) Likelihood Ratio

- in terms of PPV

$$\underbrace{\frac{Pr(Y = 1|Y^* = 1)}{Pr(Y = 0|Y^* = 1)}}_{\text{Posterior odds}} = \underbrace{\frac{Pr(Y^* = 1|Y = 1)}{Pr(Y^* = 1|Y = 0)}}_{\text{Likelihood Ratio}} \times \underbrace{\frac{Pr(Y = 1)}{Pr(Y = 0)}}_{\text{Prior odds}}$$

$$\frac{PPV}{1 - PPV} = \frac{Sn}{1 - Sp} \times \frac{Pr(Y = 1)}{1 - Pr(Y = 1)}$$

- LR(+) tells us how much more likely a condition is, given a positive test result

- in terms of NPV

$$\underbrace{\frac{Pr(Y = 1|Y^* = 0)}{Pr(Y = 0|Y^* = 0)}}_{\text{Posterior odds}} = \underbrace{\frac{Pr(Y^* = 0|Y = 1)}{Pr(Y^* = 0|Y = 0)}}_{\text{Likelihood Ratio}} \times \underbrace{\frac{Pr(Y = 1)}{Pr(Y = 0)}}_{\text{Prior odds}}$$

$$\frac{1 - NPV}{NPV} = \frac{1 - Sn}{Sp} \times \frac{Pr(Y = 1)}{1 - Pr(Y = 1)}$$

- LR(-) tells us how much more likely a condition is, given a negative test result

2. Receiver Operating Characteristic Curves

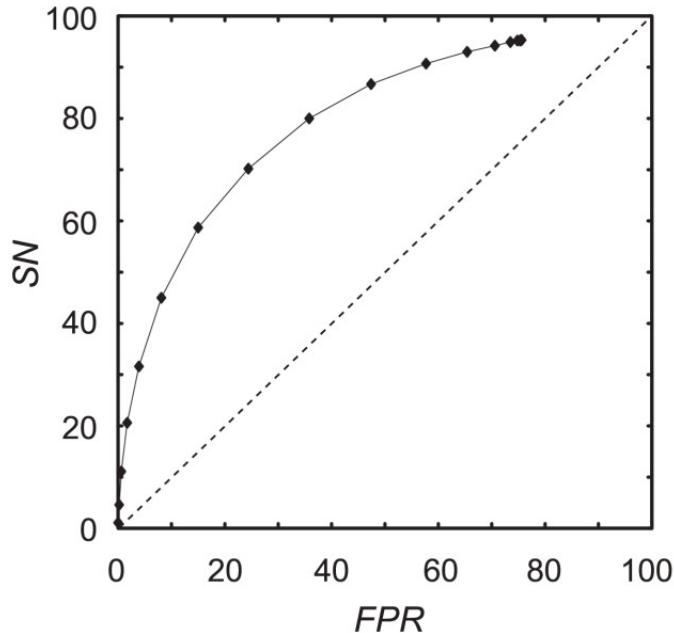


FIGURE 4.1. ROC curve (solid line) with lower limit (dotted line), points indicate values for different cutoff points used in the prediction algorithm

- A Receiver Operating Characteristic Curve (ROC) is drawn by graphing different cutoff values of a prediction algorithm on a graph of 1-Specificity (FPR) on the x-axis versus Sensitivity (TPR) on the y-axis (Figure 4.1)
- the area under the ROC curve (AUC) can be estimated by the area underneath this curve
 - AUC has the range $[0.5, 1]$, and can be interpreted as the discriminative ability of a model
 - simply put, the AUC is the probability that given a pair of individuals, one with disease and the other disease-free, the two individuals are correctly labeled by the prediction algorithm
 - an AUC of 0.5 means the prediction algorithm has no more predictive ability than simply tossing a coin
 - an AUC of 1.0 means the prediction algorithm has perfect discriminative ability
 - the closer a curve comes to the top left corner of the graph of 1-Sp vs Se, the higher its discriminative ability (AUC)

CHAPTER 5

Study Designs

1. Study Population

1.1. Types of populations.

populations: the focus on populations in epidemiology distinguishes it from other biomedical disciplines

There are four types of populations:

- (1) **study population:** a *sample* of subjects for whom we make observations and collect data on, a subset of the source population
- (2) **source population:** a group of individuals who are designated by personal, geographic, or temporal characteristics as eligible for the study
 - should be clearly defined by the investigator (ex: all 18-50 year old men living in Suffolk county between the years 2010-2011 at risk for prostate cancer)
 - it is not necessary to identify every member of the source population
 - also called the *sampling frame*
- (3) **base population:** group of individuals *at risk* of developing the outcome of interest, the population from which study cases arise
 - the group of individuals we would like to make valid causal inferences on
 - can be equivalent to the source population
 - is the sum of all eligible person-time
 - **primary study base:** base population is identified first, then cases are ascertained
 - clearly defined study population
 - roster often exists (ex: nested case-control study, study of current HMO members)
 - if the study base is known, baseline rates estimated for the population can be used in case-control studies
 - **secondary study base:** cases are ascertained first, then the base population is identified

- base often defined as the population that if they had developed the outcome of interest *would have been* a case in the study
 - a case-defined study (ex: hospital-based case-control study)
 - also known as the *study base*, *population at risk*, or *source population* (Rothman)
- (4) **target population:** population the investigator would like to generalize the study results to
- designated by the investigator according to their objectives
 - typically not observed by the investigator
 - investigator hopes the study results are externally valid to this population

1.2. Case-Definition & the Study Base.

- **Case:** is a person who developed the outcome and was *diagnosed* as a case
- must be part of the study base when outcome occurs
- if there are persons with no opportunity of diagnosis one must either
 - (1) exclude them from the study-base
 - (2) include them by redefining the study base
 - relaxing or removing diagnostic criteria for case-ness
- Case definition is implicitly intertwined with the availability of diagnosis
 - (1) if valid diagnostic procedure rarely performed or highly subjective
 - must be careful that exposure is not a determinant of diagnosis, otherwise could induce a selection bias
 - (2) if valid diagnostic procedure is widely available
 - benefits of rigorous case definition most often outweighs risk of bias from selective access / diagnosis
- Studies of benign conditions run a high risk for mistaking correlates of access to medical care for etiologic factors

2. Randomized Experiments

- randomized experiments: a study where the investigator uses a stochastic (random) mechanism to assign treatment
 - randomization is *expected* to balance the distribution of risk factors for the outcome between treatment groups (groups are comparable)
 - intention-to-treat (ITT) analysis: estimates the effect of the instrument (treatment assignment) and not the treatment itself

- the study you would want to conduct (or are trying to mimic) to estimate causal effects
- not always practical or ethical to do a randomized experiment
- ideal randomized experiment: groups are marginally exchangeable
 - * treatment groups are exchangeable: allows estimation of average level causal effects
 - * near infinite sample size (no sampling variability)
 - * no loss to follow-up
 - * full adherence
 - * double-blind
- crossover randomized experiment: individuals are observed during two or more periods
 - * individual is exchangeable: allows estimation of individual level causal effects
 - * assumptions:
 - exposure effect has short duration
 - exposure has no carry-over effect
 - outcome is an abrupt event that completely resolves by the next period (reversible event)
 - all prognostic factors other than treatment remain constant across periods
- unconditional (marginal) randomization: treatment probability is independent of any covariates, $Pr[A] \perp\!\!\!\perp L$
- conditional randomization: treatment probability depends on the level of a covariate or set of covariates, $Pr[A = a'|L = l'] \neq [A = a''|L = l'']$

3. Observational Studies

- a study set up by nature where the investigator observes exposures and outcome
- prone to biases, especially confounding
- include cohort studies and case-control studies

4. Cohort Studies

- Cohort: observational studies analogous to randomized experiments, but the investigator does not assign exposure

- groups defined by exposure status
- occurrence of disease compared in follow up time

4.1. Closed Cohorts.

- membership defining event
 - not necessarily at one point in time, depends on time scale (ex: birth cohort, military service)
- once a member always a member
- fixed population: no new members can enter (static cohort)
- measures of association
 - Cumulative Incidence Ratio/Difference: can be directly measured
 - Odds Ratio
 - Incidence Rate Ratio/Difference
- drawbacks
 - loss to follow up
 - competing risks
 - changes in exposure status over time (except fixed cohorts)
 - incidence rate chosen measure of association due to these drawbacks
 - in the case of an inevitable outcome (e.g. death) the cumulative incidence will approach 1 as follow-up increases

4.2. Open Cohorts.

- membership defined by state that can change over time (ex: MA residents, HMO members)
- PT from changing roster of participants (dynamic population)
 - individuals may contribute small or large amounts of person time
 - individuals may come and go
 - individuals may contribute person time to multiple exposure categories
- rates can be directly measured, but cumulative incidence cannot be directly measured
- to get cumulative incidence we need $C = 1 - \exp[-\sum(I * t)]$

4.3. Practical issues in Cohorts.

- (1) are expensive for rare diseases with long induction periods. Four solutions:
 - (a) use registry to reduce cost of monitoring
 - (b) use a historical or retrospective cohorts
 - #1 and #2 often use records, thus may suffer from
 - poor accuracy & misclassification
 - missing records that are related to exposure in some way (selection bias)
 - (c) use general population information (if exposure is rare)
 - beware of “healthy worker effect”:
 - the exposure group are fundamentally different from the population
 - exposure depends on meeting certain criteria that are not met in general population
 - (d) perform a nested case-control or case-cohort study
- (2) timing of outcome events hard to determine for chronic diseases with insidious onset
 - measurement is at discrete intervals
 - use of a minimum lag period is advisable when long latent periods are inevitable
 - examples: undiagnosed disease or prodromal periods
 - can be viewed as a source of measurement error and source of bias
- (3) tracing individuals must be feasible
 - loss to follow up can damage validity
 - if related to both exposure and disease will have bias
 - will add to cost

5. Case-Control Sampling Schemes

- Material for this section was drawn from:
 - (Mittleman 2009; Rothman 2008; Walker 2009)
 - (Miettinen 1976; Wacholder 1987; Langholz 1990; Wacholder 1991)
- The case-control paradigm is founded upon the odds ratio's symmetry

$$\frac{Pr[E|D]/Pr[\bar{E}|D]}{Pr[E|\bar{D}]/Pr[\bar{E}|\bar{D}]} = \frac{Pr[D|E]/Pr[\bar{D}|E]}{Pr[D|\bar{E}]/Pr[\bar{D}|\bar{E}]}$$

5.1. Useful Concepts.

Definitions.

- *Risk-Set*: the set of persons at risk when an event occurs
 - implicitly matched on time
 - may be constrained further to additional matching factors (e.g. age)
 - comparing a case to her risk set is very useful for dealing with time trends in exposure or other characteristics of an open cohort that are influenced by time
- *Incident Case*: outcome event occurred during follow up. These are usually sampled from the case-series.
- *Prevalent Case*: outcome event occurred before the start of follow up. These are usually excluded from the case-series.
- *Sampling Person-Time*: sampling a control with probability proportional to at risk-person-time experience.
- *Sampling Fraction*: or selection frequency, is the probability of being selected into the study.
 - if either the selection frequency for cases f or controls g are related to *exposure*, our sample estimate $\widehat{OR} = \frac{ad}{bc}$ is a biased estimator of the true population $OR = \frac{AD}{BC}$ because of selection bias.
 - by design cases are oversampled in a case-control study and thus $f > g$ (selection frequencies usually do differ by disease status)

Sampling Fractions.

- Define sampling fractions:

$$f' = \frac{a}{A} \quad f'' = \frac{b}{B} \quad g' = \frac{c}{C} \quad g'' = \frac{d}{D}$$

a/b is the distribution of exposure in sampled cases. It estimates A/B , the distribution of exposure in cases of the entire population

c/d is the distribution of exposure in controls. It estimates C/D , the distribution of exposure in the *person-time* of the entire population

- the incidence rate ratio estimate is

$$\widehat{IRR} = \widehat{OR} = \frac{(f' \cdot A) \times (g' \cdot D)}{(f'' \cdot B) \times (g'' \cdot C)} = \frac{a \times d}{b \times c}$$

- we have bias if sampling fractions for cases or controls are related to exposure ($f' \neq f''$ or $g' \neq g''$)
- note that due to sampling variability of a, b, c, d , \widehat{OR} provide an estimate of IRR with a margin of error. Thus we say it is a consistent estimator of the *IRR*

Case-Control Sampling.

$$\widehat{IRR} = \frac{I_1}{I_0} = \frac{\frac{A}{N_1}}{\frac{B}{N_0}} = \frac{\frac{A}{B}}{\frac{N_1}{N_0}} \times \frac{(f/f)}{(g/g)} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \times d}{b \times c} = \widehat{OR}$$

- assuming $f = f' = f''$ and $g = g' = g''$
- **cases:** give relative information about the numerator of the rates you would have calculated in a cohort study, $\frac{a}{b}$
 - a sample of the cases may be used, but the cases must be sampled independent of exposure
- **controls:** give relative information about the denominator of the rates you would have calculated in a cohort study, $\frac{N_1}{N_0}$
- cases and controls must represent the same study base *over the time period of the study*, conditional upon all measured risk factors
- non-random samples can be used with complete case ascertainment only if...
 - distribution of exposure is same in control series as in a random sample of the secondary base
- **stratified sampling fractions:** we can allow sampling fractions to vary by covariate level. thus within strata that define the sampling fraction, the rate ratio is consistent.
 - this allows us to ensure adequate numbers of controls in each confounder level (i.e. we can oversample some strata), for better control of confounding
 - it's *like* the idea of a matched case but it's not the same

Non-random selection of controls.

- **pseudo-random sampling:** choose controls through a process that is arguably random in nature (e.g. next live birth)
- **homology argument:** employed when a random sample not possible
 - (1) compare prevalence of risk factors for main outcome among exposed and unexposed, stratified by the diseases (or other factors) used to assemble control series
 - if see pattern probably reflects source population remove ones that are markedly different
 - (2) look at idiosyncratic features (person, time, and place) and ask if, hypothetically, these characteristics would result in a distribution of risk factors that is different than the unknown distribution in the actual source population.
 - if you can't think of any, probably represents source population

Case-Definition.

- as mentioned before, a *case* implies diagnosis
 - subjective or rare diagnostic procedures can spuriously relate case-definition to exposure
 - for mild disorders that do not require medical attention should either...
 - * ascertain all cases
 - * restrict base to those that would use medical services if develop disease
 - * ensure that medical treatment seeking is not associated with exposure
- feasibility can often be achieved by refining case definition (to better define a practical study base)
- if redefine case eligibility definition be sure to redefine control eligibility as well

Control Sources.

- (1) population controls: random sampling of a population list
- (2) hospital/clinic controls: sample control disease that have similar referral patterns for case disease, *but* admission is unrelated to exposure of interest
 - subjects with case disease would have been admitted for control disease (vice versa)
 - ideally want to use several diseases to lessen the chance of an association with exposure
 - for unrelated diseases, don't want their occurrence, detection, or likelihood of hospitalization to be related to outcome or exposure.
 - *case definition* should be handled with care. Cases that occur...
 - (a) after arrival to the hospital: base is persons who *while in the hospital*, if they developed the outcome, they would have been included a case in the study, *regardless of admission diagnosis*
 - (b) before arrival to the hospital: base is persons who if they developed the outcome would have been referred (or came on own accord) to hospital (i.e. may be a referring practice, HMO plan, or catchment area)
 - if both types of cases used you have a composite study base
- (3) random digit dialing: sample of households with telephones, several methodological issues (secondary study base)
- (4) dead controls: comparability issues, not members of the base population
- (5) death registry for mortality studies: deaths from causes whose occurrence is unrelated to exposure
- (6) deterministic (non-random)

- divide base into non-overlapping strata → 100% sample from case's strata
- ex. neighborhood controls: sample by residence
- ex. family/friend controls: case identifies a control, may lead to *overmatching*

Case-Control Paradigm.

- regardless of design (cohort or case-control), the comparison is always between the exposed and unexposed
- the design employs sampling to ascertain exposure distribution in cases and the *study base*
- is analogous to sampling from a full cohort, even if not possible to enumerate
 - assumption that the unsampled persons' data is *missing at random* (MAR)
 - (1) values of covariates for those who do (or do not) develop disease do not depend on whether the person is included as a case (or control) in the study
 - (2) met only if sampling fractions for cases and controls are independent of exposure
 - absolute rates can be estimated, with complete case ascertainment, if we have either
 - (1) known sampling fractions
 - (2) external information about the base (i.e. incidence rate for entire population)
- an advantage of case-control over cohort is by studying fewer people, can put more effort/resources into ascertainment and measurement of confounders for better accuracy
- fallacies
 - (1) “*trohoc*” fallacy: false belief that controls should be healthy and absent of disease
 - distracts from the goal of creating a control group that efficiently samples from the base population
 - exchangeability between cases and controls is *not* the same as comparability of cases and controls
 - * want comparability of cases and controls where they are both coming from the same base population
 - removing a causal intermediate from the control group is inappropriate
 - * to do so would change the distribution of exposure in the controls such that it is no longer representative of the study base that gave rise to the cases

- (2) *at risk for exposure*: restricting the control series to persons (or periods) that are at risk for exposure is akin to overmatching, even in a case-crossover study

5.2. Case-Cohort. Sampling at baseline

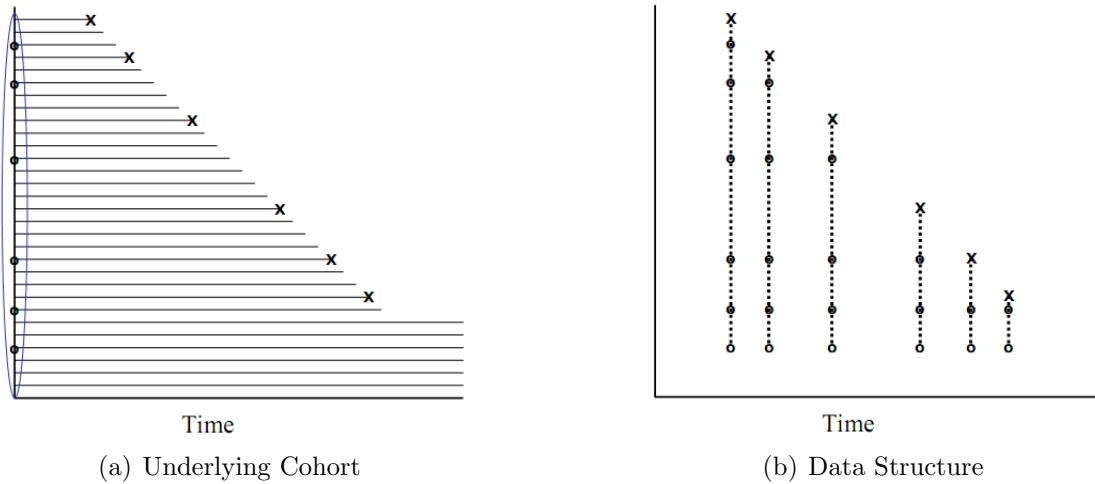


FIGURE 5.1. Case-Cohort Sampling

Sampling Frame.

- controls sampled
 - at baseline
 - independently of exposure and outcome
 - randomly from entire cohort
- forms sub-cohort from which risk-sets are formed during analysis

Measurement & Ascertainment.

- incident cases are observed for entire cohort
- date of entry, termination of follow up, and outcome ascertainment collected on entire cohort
- exposures, confounders, and effect modifiers measured only for sub-cohort and incident cases
 - identification and ascertainment of controls takes place *before* identification and ascertainment of cases

Strengths (compared to Nested-Case-Control).

- control ascertainment does not depend on case identification
 - faster data collection - can identify controls immediately, before cases occur
 - can collect expensive data “in bulk” and save money

- controls suitable for multiple outcomes
 - * however, 95% CI require adjustment because risk sets are correlated
 - * in regression the data has a correlated structure
- cases during extended follow up do not need new controls
- multiple time scales allowed → better control of confounding by time
- data can be re-analyzed using risk-set sampling (e.g. post-hoc studies)
- provides data for external comparisons (e.g. entire cohort/population)
- more efficient when little late entry and little right-censoring
- can estimate baseline risk and risk ratio in full cohort

Weaknesses (compared to Nested-Case-Control).

- power cannot be estimated until end of follow up because it depends on the size of the exposure overlap in the final risk-sets
- cases and controls are assessed at different times, need to ensure comparable measurement accuracy and compliance
- cases outside of sub-cohort have no baseline data, must be assessed at event
- not well suited for time-varying exposures or confounding because must be measured in every control at every outcome event (more \$\$\$)
- less efficient for staggered entry and substantial right-censoring
- risk-sets are not randomly sampled and are therefore correlated → complicated analysis
- forced to measure deteriorating lab samples at end of follow up, to avoid differential measurement error between sub-cohort cases, controls, and cases outside the sub-cohort

5.3. Density-Sampling. *Sampling PT over follow up*

Sampling Frame.

- the probability of sampling a control should be proportional to at risk person-time
- this is satisfied one of three ways:
 - (1) matched on time (see risk-set and nested-case-control sections)
 - (2) sampling controls at a constant rate over follow up
 - (3) STEP 1. randomly sample (with replacement) a date and a person from base
 STEP 2. if person under observation and at risk on that date, select as a control.
 - (#3) requires that

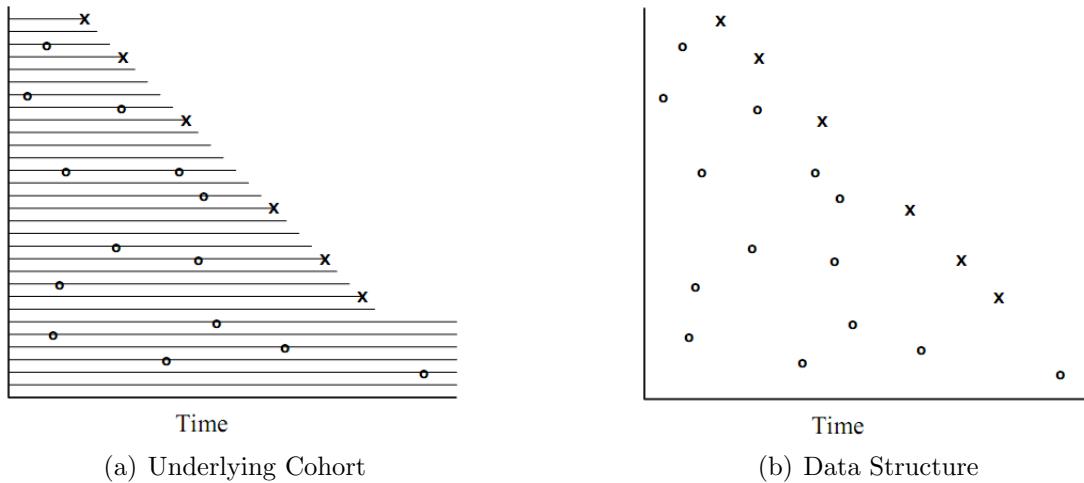


FIGURE 5.2. Density Sampling

- (a) controls are eligible for selection during the same period that they are eligible to become a case
 - (#2) and (#3) are unmatched on time
 - matching on time (#1) is subsumed within sampling with a steady rate (#2), which is subsumed with sampling in proportion to interval at risk (#3)
- steady state assumption is only required for #2 and #3 because risk-sets are not used in sampling or analysis. Specifically this means constant incidence and prevalence (see Chapter 2: Measures of Frequency and Occurrence for more detail)

Measurement & Ascertainment.

- incident cases are observed for entire cohort
- date of entry, termination of follow up, and outcome ascertainment collected on entire cohort
- exposures, confounders, and effect modifiers measured for all controls *and* incident cases

If not sampling at steady rate:

- control ascertainment cannot occur until end of follow up and the full period of risk can be specified

Strengths (as compared to Cumulative-Incidence).

- better for chronic disease epidemiology
 - NO need for rare disease assumption to estimate incidence rate ratio
 - long follow up makes prevalent controls (end of FU) more likely to induce selection bias
 - incidence density sampling avoids this by sampling over time

Weaknesses of Plain Density Sampling (as compared to Risk-Set).

- steady state assumption may not be reasonable
- identification of controls is much slower

Weaknesses of Steady Rate Density Sampling (as compared to Risk-Set).

- more resource intensive

5.4. Risk-Set Sampling. *Sampling PT matched on time*

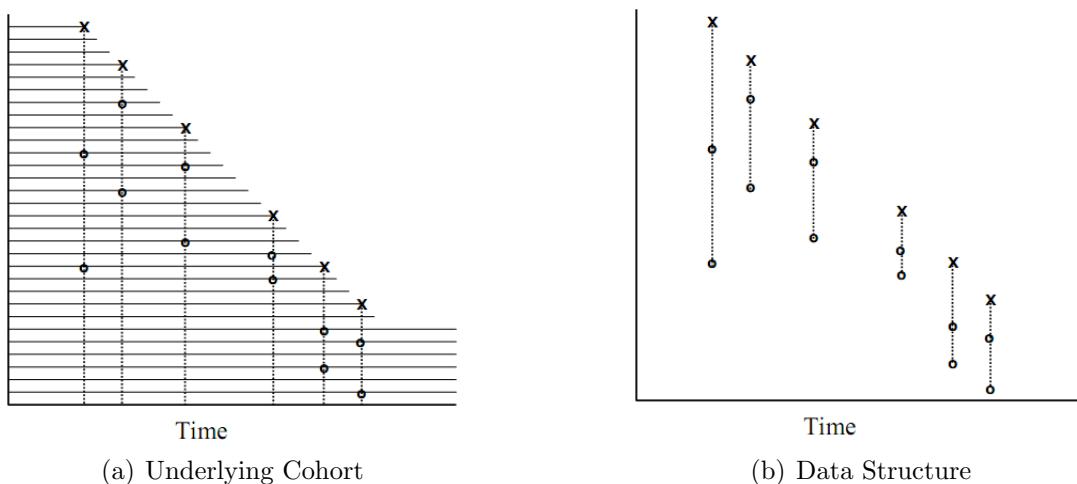


FIGURE 5.3. Risk-Set Sampling

Sampling Frame.

- controls derived from a random sample of case's risk-set *after* case ascertainment
 - cases and controls are implicitly *matched on time*
 - risk-set can be further specified in terms of other covariates (age, gender, etc.)
 - * matching factors should be a characteristic that could be easily assessed for the entire population
 - * and should be something where you're not interested its effect

Measurement & Ascertainment.

- incident cases are sampled from the study base
- date of entry, termination of follow up, and outcome ascertainment for cases and controls
- exposure, covariates, and any matching factors assessed on case index day

Strengths (compared to unmatched density sampling).

- no steady state assumption needed
 - * hazard ratios are estimated within risk-sets (e.g. at every failure time) and pooled using a weighted average
- matching on time provides
 - * comparable measurement between cases and controls
 - * allows assessment of exposures that vary over time (also confounding)
 - * lab samples are matched on “deterioration” → comparability

if there are time trends in exposure then a matched analysis is required (assuming no other matching factors)

Weaknesses.

- matching on time subjects the analysis to several constraints
 - * time scale is restricted to the one used to form risk-sets → potential residual confounding by time
- slow data collection - selection of controls depends on identification and *confirmation* of a case

5.5. Nested-Case Control. *Risk-set sampling in an enumerated cohort*

Sampling Frame.

- **technically** any sampling frame can be used
- most often risk-set sampling is used to tightly control for confounding by time
- here the study base is a fully enumerated cohort
- what follows applies to the nested-case-control under risk-set sampling

Measurement & Ascertainment.

- incident cases are observed for entire cohort
- date of entry, termination of follow up, and outcome ascertainment collected on entire cohort
- exposures, confounders, and effect modifiers measured only for the risk-sets and incident cases (*only for index dates*)
 - * identification and ascertainment of controls takes place *after* identification and ascertainment of cases

Strengths (compared to Case-Cohort).

- power is virtually independent of the size of the entire cohort
- matching on time provides

- * comparable measurement between cases and controls
 - * allows assessment of exposures that vary over time (also confounding)
 - * data on time-varying exposures in cases and controls does not need to be collected beyond follow up of case (less money)
 - * lab samples are matched on “deterioration” → comparability
 - simpler analysis because risk-sets are randomly sampled (e.g. independent)
 - note that a matched analysis is required if time is related to exposure (assuming no other matching factors)
- Weaknesses (compared to Case-Cohort).*
- matching on time subjects the analysis to several constraints
 - * time scale is restricted to the one used to form risk-sets → potential residual confounding by time
 - * not well suited for multiple outcomes
 - * difficult to make external comparisons because controls are not a random sample of entire cohort
 - * cannot be analyzed as case-cohort (i.e. post-hoc) because controls’ selection is conditional upon being free of outcome at time of selection. Could introduce selection bias.
 - slower data collection - selection of controls depends on identification and *confirmation* of a case

5.6. Cumulative-Incidence. Sampling at end of follow up

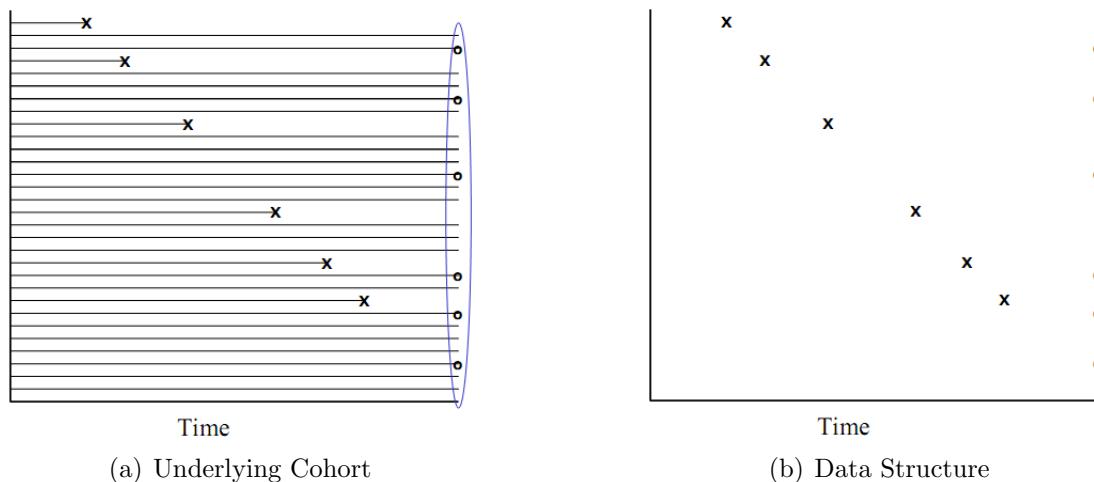


FIGURE 5.4. Cumulative Incidence Sampling

Sampling Frame.

- requires a closed cohort
- incident cases are selected over follow up
- non-cases are sampled as controls at the end of follow up

Measurement & Ascertainment.

- retrospective measurement of exposure

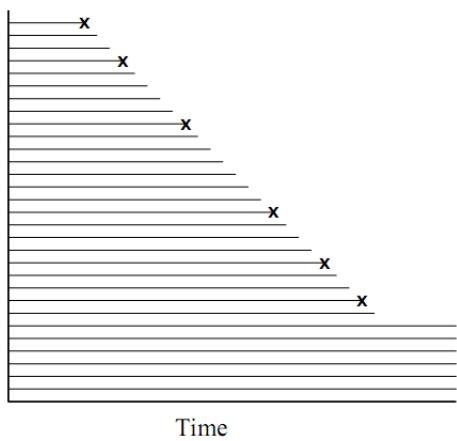
Strengths.

- FAST and resource efficient
- short follow-up makes them suitable for acute-epidemic outbreaks of short duration

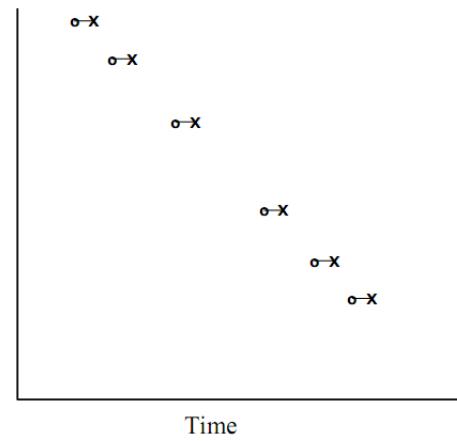
Weaknesses (as compared to all others).

- **this is the only design that requires rare disease assumption** to interpret the incidence odds ratio as the incidence rate ratio
- not ideal for chronic disease epidemiology when follow up periods are long
 - * use of prevalent controls likely to introduce bias because controls' selection is conditional upon survival to the end of follow up
 - * potential for recall bias since baseline exposure is measured retrospectively
- potential for selection bias if cases and controls selected from same source

5.7. Case-Crossover. *Case become their own control*



(a) Underlying Cohort



(b) Data Structure

FIGURE 5.5. Case-Crossover Sampling

Sampling Frame.

- requires transient exposure and acute effect on the outcome
- only cases are enrolled, use self as control
- a short period immediately before the outcome would be defined as the case window
- a matched length of period before the case period is defined as control window
- Control window can be randomly selected outside the case window, can be prospective, retrospective, or bidirectional
- the selection of control is based on the assumption of **conditional independence of the exposure and random distribution of the exposure over time**

Measurement & Ascertainment.

- retrospective measurement of exposure

Strengths.

- suitable for transient exposure where control is hard to find
- appropriate for environmental study where exposure (eg. air pollution) is common on all, no concurrent control is available
- do not require controls to be recruited
- immune to selection bias
- effectively control for both **measured and unmeasured inter-individual confounding**

Weaknesses (as compared to all others).

- It cannot control the **intra-individual confounding** (eg. someone always drink coffee after sex)
- potential for recall bias since exposure is measured retrospectively
- potential for exposure misclassification since induction period is usually unknown
- susceptible to exposure trend bias, if the baseline exposures are associated with a longitudinal trend
- cannot be used for exposures that do not vary over time (at the individual level)

Analysis of case crossover study.

- Each case-control period pair can be viewed as stratum
- unit of analysis is window of time: time just before acute event (case window) compared with other times (control windows)

- Mantel-Haenszel stratified analysis can be used to compute summary rate ratio (person-time data)
- The **rate ratio is identical to risk ratio**
 - * Because the follow-up periods for each control and case were fixed and identical and there was no loss to follow-up.
- The RGB (Robins-Greenland-Breslow) variance can be used to compute confidence interval and hypothesis test.

$$\hat{Var}_{RGB} = \frac{A}{BC} = \frac{\frac{M_{1i}N_{1i}N_{0i}}{T_i^2}}{\frac{a_iN_{0i}}{T_i} \frac{b_iN_{1i}}{T_i}}$$

- MH analysis in case-crossover study is algebraically identical to matched analysis
- The McNemar method or conditional logistic regression can also be used to obtain the effect estimate and confidence interval

Induction period analysis.

- The determination of the induction period is based on both biological plausibility and sensitivity analysis.

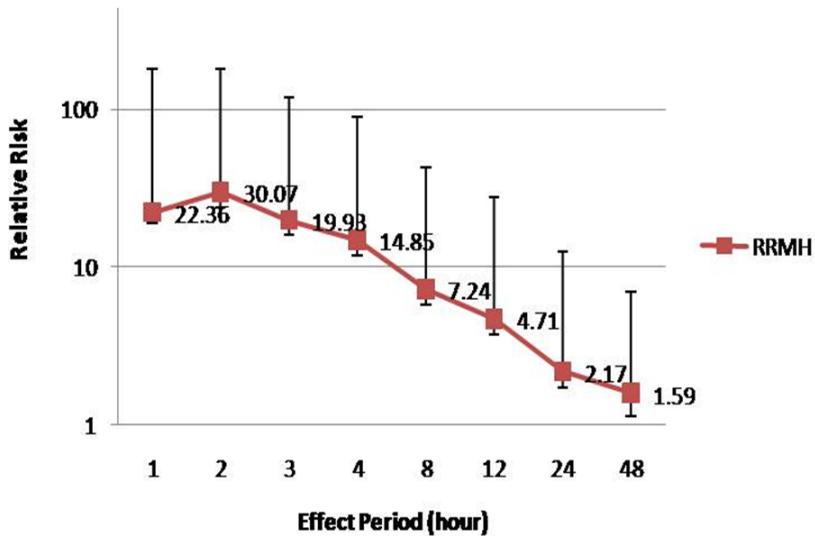


FIGURE 5.6. Induction period analysis

Case-time-control correction of exposure trend bias.

- require to recruit control
- control-crossover odds ratio would be computed

- control-crossover OR is used for adjustment of the case-crossover OR and obviate the potential exposure trend bias
- The validity of such adjustment is based on the assumption that baseline exposure trend of case and control is exchangeable, which is usually not the case



FIGURE 5.7. Case-time-control study

5.8. Case-Only.

6. Cross Sectional Studies

7. Ecological Studies

TABLE 5.1. Source population, Design, & Estimand

Observation Space	Design	Additional Sampling	Design	Estimand
Persons	Non-cases	Closed cohort with fixed follow up	Traditional Case-Control (cumulative incidence)	Odds Ratio
	All persons	Case base	Risk Ratio	
Person-Time	Cohort with changing baseline incidence	Closed cohort with changing exposure or open cohort	Pool of Person Time (incidence density)	Modern Case-control rate ratio
	Non-cases in risk-set	Matched on time	case-control on time	Rate ratio matched on time
	Sub-cohort reuse in risk sets	Case-Cohort		Rate ratio matched on time
Matched on person	Cases only	Case-Crossover	Rate ratio	
	Non-cases	Case-Time Control	Rate ratio	

TABLE 5.2. Comparison between study designs

Randomized Trial	
<i>Assumptions</i>	<ul style="list-style-type: none"> – random assignment of treatments guarantees that the distributions of measured and unmeasured confounders in treated group and in untreated group are identical (ideally) – often considered as the gold standard
<i>Strengths</i>	<ul style="list-style-type: none"> – can eliminate all potential confounding by design (ideally) – prospective study \Rightarrow no recall bias – double-blind \Rightarrow reduce measurement error of the outcome – placebo controlled \Rightarrow reduce behavior and life style change due to intervention
<i>Limitations</i>	<ul style="list-style-type: none"> – non-compliance \Rightarrow intention-to-treat analysis – limited follow-up often prevent observation of the outcome for chronic disease <ul style="list-style-type: none"> * surrogate outcome (e.g. blood pressure instead of CHD) – lack of generalizability <ul style="list-style-type: none"> * a trade-off between internal validity and external validity * non-compliance the pattern of non-compliance in the trial and in general population may be different * unrepresentative patients to make the treatment effect easier to detect \Rightarrow select patients with fewer comorbidities, fewer medications, not pregnant, and at earlier stage of disease * unrepresentative treatments the range of exposure in a trial is often narrower than the range of exposure in real life – expensive (especially when the treatment is expensive)
<i>Practical issues</i>	<ul style="list-style-type: none"> – informed consent (applies to all kinds of study mentioned below) – need to pre-specify sample size (statistical power) – need to pre-specify stopping rule – need to consider equipoise between treatment arms – cluster randomized trial: for vaccination, infectious disease, and community intervention, cluster randomized trial may be a good choice instead of individually randomized trial <ul style="list-style-type: none"> * the number of randomized units is usually small and balancing between confounders may not be achieved – randomized crossover study <ul style="list-style-type: none"> * control confounding by using an individual as his own control * need a washout period between two crossover periods – blocking in randomized trial: randomization within levels of factors (effect modifiers) <ul style="list-style-type: none"> * e.g. randomization within males and within female to ensure that exactly half of the male and half of the female participants received the treatment

TABLE 5.3. Comparison between study designs (cont.)

Closed cohort study	
<i>Assumptions</i>	<ul style="list-style-type: none"> – components of a closed cohort study <ul style="list-style-type: none"> (1) a heterogeneous group (2) a fixed follow-up (3) uniformly observed outcome – defined by “membership defining event”
<i>Strengths</i>	<ul style="list-style-type: none"> – establish temporality – can estimate risk if none of the limitations occur
<i>Limitations</i>	<ul style="list-style-type: none"> – limitations prevent estimation of risk <ul style="list-style-type: none"> (1) loss to follow up (2) competing risks (3) changes in exposure status over time – expensive for rare diseases
<i>Practical issues</i>	<ul style="list-style-type: none"> – risk eventually reaches 1 for inevitable outcomes (e.g. death) – should put equal efforts on data collection for each exposure group
Open cohort study	
<i>Assumptions</i>	<ul style="list-style-type: none"> – components of a open cohort study <ul style="list-style-type: none"> (1) person-time at risk for each exposure category (2) number of outcome events within each exposure category – need to assume constant rate during the entire follow-up if want to use a single summary rate estimate – defined by “membership defining state”
<i>Strengths</i>	<ul style="list-style-type: none"> – can handle non-differential loss to follow-up – can handle time-varying exposure <ul style="list-style-type: none"> * one individual can contribute person-time to more than one exposure group
<i>Limitations</i>	<ul style="list-style-type: none"> – suitable for multiple outcome event – differential loss to follow-up – competing risks – expensive for rare diseases – timing of outcome events hard to determine – prospectively need to wait for a long time for events to occur – hard to control pre-baseline confounding (e.g. need pre-treatment covariates to get propensity score for the treatment at baseline)
<i>Practical issues</i>	<ul style="list-style-type: none"> – can estimate rate ratio, but cannot estimate risk ratio directly <ul style="list-style-type: none"> (can calculate risk ratio by exponential formula $Risk = 1 - e^{-\sum_{i=1}^t \text{incidence rate} \times \Delta t}$) – use stratified analysis, Poisson regression, Cox PH model, pooled logistic regression to analyze person-time data

TABLE 5.4. Comparison between study designs (cont.)

Case-control study	cumulative incidence sampling
<i>Assumptions</i>	<ul style="list-style-type: none"> – requires a closed cohort as study base <ul style="list-style-type: none"> * incident cases are selected over follow-up * non-cases are sampled as controls at the end of follow up – requires rare disease assumption for risk ratio interpretation
<i>Strengths</i>	<ul style="list-style-type: none"> – fast – efficient – can estimate risk ratio directly
<i>Limitations</i>	<ul style="list-style-type: none"> – rely on rare disease assumption – inherited limitations of closed cohort study <ul style="list-style-type: none"> * subject to 1). loss to follow-up, 2). competing risk, and 3). time-varying exposure/confounder
<i>Practical issues</i>	<ul style="list-style-type: none"> – often suitable for outbreak of infectious disease – not suitable for chronic disease (may be a lot of competing risks and loss to follow-up in the base cohort)
Case-control study	incidence density sampling
<i>Assumptions</i>	<ul style="list-style-type: none"> – controls are used to estimate the relative distribution of person-time (not persons) across the exposure categories – assume constant rate over entire follow-up (steady state assumption)
<i>Strengths</i>	<ul style="list-style-type: none"> – can estimate rate ratio – high cost and time efficiency compared to open cohort study – do not rely on rare disease assumption
<i>Limitations</i>	<ul style="list-style-type: none"> – cannot estimate baseline rate unless know the sampling fractions, compared with open cohort study – cannot directly estimate risk ratio, compared with cumulative incidence sampling – not suitable for studies on exposure time trend in disease risk
<i>Practical issues</i>	<ul style="list-style-type: none"> – can allocate more resources on collecting detailed information from each participant, compared with open cohort study

TABLE 5.5. Comparison between study designs (cont.)

Nested Cs-Cn study	risk-set sampling
<i>Assumptions</i>	<ul style="list-style-type: none"> – components of nested case-control study <ul style="list-style-type: none"> * date of entry, termination of follow-up, and outcome ascertainment for cases and controls for the entire cohort * exposures, confounders, and effect modifiers measured only for cases and controls at the index day – matching on time (index day) and other matching factors which define the risk set – constant rate within a risk set (instantaneous incidence rate \Rightarrow hazard)
<i>Strengths</i>	<ul style="list-style-type: none"> – no steady state assumption for entire follow-up needed, compared with incidence density sampling – reduce confounding by matching – can pre-specify the sample size, compared with case-cohort study – less expensive to assess time-varying exposure and confounders, compared with case-cohort study – can assess the exposure of cases and controls at the same time (index day), compared with case-cohort study (no differential sample degradation)
<i>Limitations</i>	<ul style="list-style-type: none"> – cannot assess effect of matching factors (can assess the effect modification by the matching factor) – time scale is restricted to the one used to form risk-sets \Rightarrow potential residual confounding by time – not well suited for multiple outcomes – difficult to make external comparison, compared with case-cohort study

TABLE 5.6. Comparison between study designs (cont.)

Case-cohort study	sub-cohort sampling at baseline
<i>Assumptions</i>	<ul style="list-style-type: none"> – components of case-cohort study <ul style="list-style-type: none"> * date of entry, termination of follow-up, and incidence cases were observed for entire cohort * exposures, confounders, and effect modifiers measured only for sub-cohort and incident cases
<i>Strengths</i>	<ul style="list-style-type: none"> – control ascertainment does not depend on case identification, compared with nested case-control study – sub-cohort can serve as control groups for multiple outcomes – can estimate baseline risk and risk ratio in full cohort if sampling fraction is known
<i>Limitations</i>	<ul style="list-style-type: none"> – power cannot be estimated until end of follow-up – the exposure (and confounders) of cases and controls are assessed at different times ⇒ measurement error <ul style="list-style-type: none"> * at the end of follow-up otherwise ⇒ sample degradation – high cost for time-varying exposures or confounding ⇒ must assess every control at every time point with an outcome event – need to account for correlation between observations from sub-cohort in the analysis (GEE and robust variance) – more sensitive to censoring for the sub-cohort – the exposure/confounder distribution may be different at the end of follow-up and at baseline for the sub-cohort

TABLE 5.7. Comparison between study designs (cont.)

Case-crossover study	
<i>Assumptions</i>	<ul style="list-style-type: none"> – requires transient exposure and acute effect on the outcome – components pf case-crossover study <ul style="list-style-type: none"> * case window \Rightarrow a short period immediately before the event * a matched length of period other than the case window is defined as control window – the selection of control window is based on the assumption of conditional independence of the exposure and random distribution of the exposure over time
<i>Strengths</i>	<ul style="list-style-type: none"> – only cases are needed – suitable for transient exposure where control is hard to find – appropriate for environmental study where exposure is common and no concurrent control is available – immune to selection bias – effectively control for both measured and unmeasured inter-individual confounding – no loss to follow-up.
<i>Limitations</i>	<ul style="list-style-type: none"> – cannot control the intra-individual confounding – potential for recall bias since exposure is measured retrospectively – potential for exposure misclassification since induction period is usually unknown – susceptible to exposure trend bias, if the baseline exposures are associated with a longitudinal trend – cannot account for cumulative effect of exposure
<i>Practical issues</i>	<ul style="list-style-type: none"> – analyze as match case-control study – The rate ratio is identical to risk ratio – to adjust for time trend of exposure \Rightarrow use bidirectional control window or do case-time control study

Part 2

Causal Inference

CHAPTER 6

Concepts & Definitions

1. Philosophical Viewpoints of Causality

1.1. Popperian Philosophy. knowledge accumulates only by falsification

- scientific hypotheses have empirical content and are falsifiable
- testing of an hypothesis occurs by attempting to falsify them
- hypotheses that have been tested and not falsified remain reasonably good explanations of natural phenomenon

1.2. Hill Criteria. proposed set of criteria for determining causation in biomedical research

- (1) *Strength*: strong associations are particularly compelling
- (2) *Consistency*: consistent findings across studies
- (3) *Specificity*: a cause leads to a single effect, *or* an effect has one cause (not multiple causes)
- (4) *Temporality*: necessity that the cause precede the effect
- (5) *Biologic gradient*: presence of dose response or exposure related curve, not always true
- (6) *Plausibility*: scientific plausibility of an association
- (7) *Coherence*: cause-and-effect interpretation for an association does not conflict with what is known
- (8) *Experimental evidence*: deducing or eliminating an exposure and investigating if disease frequency declines
- (9) *Analogy*: hypotheses derived from insight
 - none of the criteria can bring indisputable evidence for or against causality
 - however, temporality is clearly required to establish causation, but not sufficient

1.3. Sufficient and Component Cause Model.

1.4. Rothman's Model. a general model of causation to conceptualize epidemiologic problems, (causal pies)

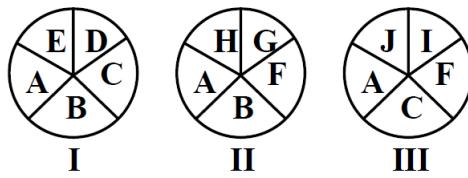


FIGURE 6.1. Rothman's Causal Pies where *I*, *II*, and *III* are sufficient causes, *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *I*, and *J* are component causes, and *A* is a necessary cause

- cause: an event or condition that plays an essential role in producing an outcome
 - * can be a constellation of components that act in concert
 - * **sufficient cause:** set of minimal conditions that *inevitably* produce disease
 - * **component cause:** an event or characteristic required by a given sufficient cause
 - * **necessary cause:** a component present in every sufficient cause
- strength of cause: predictive ability of a cause is determined by the relative prevalence of component causes
 - * ex: a rare factor is a strong cause if its complementary component causes are common (ex: PKU and Diet Coke[®])
- interactions among causes
 - * component causes in a single sufficient cause are considered to have mutual biologic interaction (ex: *D* and *E* in Figure 6.1)
 - * component causes present in multiple sufficient causes are considered to have super-additive effects (ex: *B* and *C* in Figure 6.1)
- **induction period:** period of time from causal action of a component cause to disease initiation, varies by component cause (ex: final component cause has an induction period of 0, Figure 6.2)
- **latent period:** amount of time between disease occurrence (sufficient cause) and clinical detection

2. Randomized Experiment Paradigm

- often, randomized trials are considered as gold standard of all epidemiological study designs

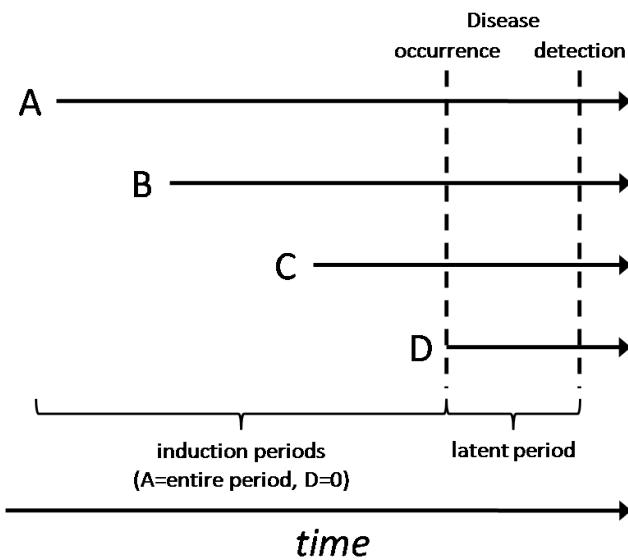


FIGURE 6.2. Timeline for a sufficient cause. A , B , C , and D are component causes, each with its own induction time.

- the analysis of observational studies should be mimicking the analysis of randomized trial

3. Counterfactual Definition of Causal Effect

3.1. Possible Worlds.

- actual world: the way things actually are, what we observe
- possible world: the way things might be, for example, a *counterfactual world*
 - * closest possible world: minimally different from actual world, only difference is the intervention of interest took place
 - * Y_i^a : the outcome of the subject in the possible world that is closest to the actual world and where the subject was treated with a
 - * the data generating model is the same in all worlds, only a different treatment or treatments have taken place

3.2. Null Hypothesis.

- null hypothesis of no average causal effect: $E[Y^{a=1}] = E[Y^{a=0}]$
- **sharp causal null hypothesis**: $Y_i^{a=1} = Y_i^{a=0}$ for all subjects i .
 - * absence of average causal effect does not imply absence of individual causal effect, *BUT* absence of individual causal effect does imply absence of average causal effect

- * the presence of individual causal effect implies the presence of average causal effect

3.3. Individual Causal Effect.

$$Y_i^{a=1} \neq Y_i^{a=0}$$

- with the exception of crossover studies, we can never calculate or estimate an individual causal effect because of the missing data problem (counterfactuals are not observed)
- causal contrast: a comparison between two counterfactual outcomes for two clearly defined treatment levels
- interference: when one individual's counterfactual is influenced by other individuals' treatments
 - * usually make the assumption of no interference for causal inference in most situations (exception: herd immunity from vaccination)

3.4. Average Causal Effect.

$$E[Y^{a=1}] \neq E[Y^{a=0}]$$

$$Pr[Y^{a=1} = 1] \neq Pr[Y^{a=0} = 1]$$

- we can never *calculate* the true average causal effect because of the missing data problem (counterfactuals are not observed)
- we can, however, *estimate* the true average causal effect when certain identifiability assumptions hold (more on this later)
- information needed: outcome, actions, counterfactual outcomes
 - * outcome: need an observed outcome (Y)
 - * actions (A)
 - treatment must be able to be manipulated
 - a well defined contrast is needed (ex: 150 mg aspirin a day versus no aspirin)
 - * a clearly defined study population whose counterfactual(s) (Y^a) are to be compared

4. Conditions for Causal Inference

observational study: a study in which the investigator does not assign treatment but simply observes treatments assigned by nature

- general paradigm
 - * make observational study look as much as possible like non-randomized experiment

- non-randomized experiment: treatment is deterministically assigned based on measured covariates L
- *non-blinded* conditionally randomized experiment
- * assume that conditional on measured covariates, treatment was randomly assigned
 - requires some sort of covariate adjustment
- * for each causal question need to carefully describe:
 - (1) the randomized experiment we would like to do
 - (2) the observational study that mimics the randomized experiment
 - (3) meet the identifiability conditions of *exchangeability*, *positivity* and *consistency*
 - (4) we need data and identifiability conditions to be met (or assumed) for an observational study
- information from observational study data is insufficient for causal inference, we need to make assumptions

4.1. Identifiability.

- an average causal effect is *non-parametrically* identifiable when the distribution of the observed data is consistent with a *single* value of the effect measure
- can get only one effect estimate from the observed data at hand (e.g. for a given observed dataset, the OR estimate can be 4 or 10 → non-identifiable; the OR estimate is 3 → identifiable.)
- from observational studies, it requires that the data is *supplemented* with three assumptions/conditions:
 - * exchangeability
 - * positivity
 - * consistency

4.2. Exchangeability.

- exchangeability (exogeneity): $Y^a \perp\!\!\!\perp A$, that is treatment A is independent of the counterfactual outcome Y^a
- allows estimation of all causal effect measures
- necessary causal assumption
- for observational studies
 - * the question is whether *all* joint predictors of A and Y were measured and included in L

- * this is an assumption that cannot be empirically verified.
- * it will require us to know both $Y^{a=1} \amalg A$ and $Y^{a=0} \amalg A$ to verify this assumption, which is unverifiable
- * we can only rely on expert knowledge
 - no unmeasured confounding
 - d-separation: no unblocked backdoor pathway
 - *caution*: $Y^{A=a} \amalg A$ is different from $Y \amalg A$
 - * full exchangeability: $Y^{A=a} \amalg A$ for all a
 - stronger version of exchangeability
 - all possible counterfactual outcomes (continuous, ie: two or more) are independent of treatment
 - usually only a result of successful randomization
 - assumes that all the individual counterfactual outcomes Y^{a_i} that are not observed are missing completely at random (MCAR)
 - * partial exchangeability: $Y^{A=a'} \amalg A$ only for $A = a'$ (not for all $A = a$)
 - only assumes counterfactual outcome for treatment level a' is independent of treatment
 - only allows for estimation of SMR (indirect standardization)
 - assumes that only the individual counterfactual outcomes $Y^{a'_i}$ that are not observed are missing completely at random (MCAR)
 - * Some notes on *full* and *partial* exchangeability.
Again, say that A and Y are both binary variables.

We may calculate the causal risk ratio for the entire population with *full exchangeability*. ($Y^A \amalg A$, i.e. $Y^{a=1} \amalg A$ and $Y^{a=0} \amalg A$)

Causal risk ratio

$$\begin{aligned}
 &= \frac{Pr[Y^{a=1} = 1]}{Pr[Y^{a=0} = 1]} \\
 &= \frac{Pr[Y^{a=1} = 1 | A = 1]}{Pr[Y^{a=0} = 1 | A = 0]} \text{ (full exchangeability)} \\
 &= \frac{Pr[Y = 1 | A = 1]}{Pr[Y = 1 | A = 0]} \text{ (consistency)}
 \end{aligned}$$

However, with *partial exchangeability* (say, only $Y^{a=0} \amalg A$ but not $Y^{a=1} \amalg A$),

we can only estimate the causal risk ratio in a subset of the population (exposed group, $A=1$).

The causal risk ratio among the exposed is also known as SMR.

SMR

$$\begin{aligned}
 &= \frac{Pr[Y^{a=1} = 1 | A = 1]}{Pr[Y^{a=0} = 1 | A = 1]} \\
 &= \frac{Pr[Y^{a=1} = 1 | A = 1]}{\sum Pr[Y^{a=0} = 1 | A = 1, L = l] Pr[L = l | A = 1]} \text{ (introduce L)} \\
 &= \frac{Pr[Y^{a=1} = 1 | A = 1]}{\sum Pr[Y^{a=0} = 1 | A = 0, L = l] Pr[L = l | A = 1]} \text{ (partial exchangeability)} \\
 &= \frac{Pr[Y = 1 | A = 1]}{\sum Pr[Y = 1 | A = 0, L = l] Pr[L = l | A = 1]} \text{ (consistency)}
 \end{aligned}$$

With $Y^{a=1} \perp\!\!\!\perp A$ alone, we can only estimate causal effect in the unexposed group.

- * mean exchangeability: average counterfactual outcome (not each individual's) is exchangeable
 - $E[Y^a] \perp\!\!\!\perp A$
 - does not imply $Y^{a_i} \perp\!\!\!\perp A$ for all i
 - $Y^{a_i} \perp\!\!\!\perp A$ for all i does imply $E[Y^a] \perp\!\!\!\perp A$
- * **marginal exchangeability:** $Y^a \perp\!\!\!\perp A$
- * **conditional exchangeability:** $Y^a \perp\!\!\!\perp A | L$
 - assumes treatment groups are exchangeable within subsets of L
 - assumes that all the individual counterfactual outcomes Y^{a_i} that are not observed are missing at random (MAR)
- * if we cannot achieve conditional exchangeability with our measured covariates, then the unobserved Y^{a_i} are missing not at random (MNAR) and depend on unmeasured covariates

4.3. Positivity.

$$\begin{aligned}
 &Pr[A = a] > 0 \text{ for all } a \text{ involved in the causal contrast} \\
 &Pr[A = a | L = l] > 0 \text{ for all } l \text{ with } Pr[L = l] \neq 0
 \end{aligned}$$

- positivity: probability of being assigned to each treatment level is greater than zero (i.e. experimental treatment assumption)
- it is guaranteed in randomized experiment
- can be violated in an observational study

- is violated when there are either no treated or no untreated subjects in all levels of L you are interested in (i.e. the levels of L in your data)

4.3.1. *structural violation.*

- the probability of some combination of L and A is guaranteed to be zero
- example: A study investigating the impact of combat exposure in the past year (A) on post-traumatic stress disorder (Y) that is stratified by age (L) will not have any 5 year-olds exposed to combat ($Pr[A = 1|L = 5] = 0$) because structurally in the US there are laws prohibiting children from military service
- solution is to restrict study to all subsets with $Pr[A|L] > 0$ (i.e. only include participants > 18 years old)
- can usually determined by subject matter knowledge

4.3.2. *random violation.*

- strata are empty by chance (i.e. MCAR) due to sampling variability
- is more likely as stratification becomes finer
- must be assumed
- difficult to distinguish between structural vs. random violations of positivity
- positivity is only necessary for
 - the set of variables L required for exchangeability
 - the levels of L where $Pr[L = l] > 0$
- if L is composite (more than 1 variable)
 - consider the probability of A in the joint set of covariate values
 - for example, persons may be able to (1) change treatment (2) be treated if not observed, but (3) not be able to change treatment if not observed
 - for IPW, positivity depends on how the weights are specified (i.e. what is in the conditioning event)
- positivity can sometimes be verified
- more likely to be violated when estimating the effect of extreme interventions
- required for standardization and IPW
 - workaround: restrict to strata of L where $Pr[A = a|L = l] > 0$ for all a
 - potential loss of generalizability
 - workaround: combine nearby strata, but could have
 - residual confounding

4.4. Consistency.

- for every subject with $A_i = a_i$, $Y_i^a = Y^{A_i} = Y_i$
- requires that each level of A is clearly defined (could be assigned in an RCT)
- requires that the investigator be able to describe the actual mechanism that is operating in the population to assign treatment
- $Y^{a=1}$ and $Y^{a=0}$ are precisely defined
- taken for granted in RCT, but not so obviously true in observational studies
 - ex: obesity can be a result of exercise, diet, genes
 - each of these can have a different effect on the outcome (e.g. mortality)
- for an ill-defined treatment
 - Solution 1:
 - (1) the investigator can describe the assignment mechanism as changing the *distribution of determinants* of the ill-defined “treatment” (e.g. obesity)
 - (2) but the proposed assignment mechanism does not match any hypothetical intervention of interest and thus has little public health utility
 - Solution 2:
 - (1) change causal question to a specific determinant of the ill-defined treatment (e.g. exercise vs. “obesity”)
 - (2) but the new question is further away from the original research question
 - regardless of how well refined a causal question is, it still has an inherent amount of vagueness
- well-defined interventions
 - uncertainty of conditional exchangeability is exacerbated by ill-defined interventions
 - L for well-defined interventions may be easier to identify than for ill-defined interventions
 - positivity may not hold (ex: strong obesity gene resulting in no non-obese carriers)
 - ill-defined interventions may be unreasonable or extreme
 - effect estimates are only as well-defined as the interventions that are being compared

5. Effect Modification

5.1. Effect Modification: The Big Picture.

- refers to the change in the magnitude of an effect measure according to the value of a third variable, a *modifier*
- a modifier of the average causal effect may only be *associated* with the outcome
 - may not have a causal effect or else would be an *interaction*
 - effect modifier is associated with a third variable that has a causal effect on outcome and *interacts* with treatment
- an intrinsic, biological phenomenon and cannot be eliminated from a study by clever design
 - differs from confounding because confounding is a bias the investigator tries to avoid, whereas effect modification is an elaborated description of the effect itself
 - * confounder: must be an independent risk factor for disease and associated with exposure
 - * modifier: may be an independent risk factor for disease, but may also be an independent risk factor for disease in the presence of another risk factor
 - that is, a factor may have properties of a modifier, a confounder, both, or neither
 - the focus is on the identification of effect modification and not its particular mechanism (we are not able to identify such mechanisms)
- the choice to study effect modification depends on inferential goals and the research question of interest
 - example 1: fluorination of public drinking water → not interested in effect modification
 - example 2: effect of pharmacological treatment → interested in effect modification (especially for those for whom treatment is contraindicated)
- there is no such thing as *the* average causal effect because of heterogeneity
 - the average causal effect in a population depends on the characteristics of the population under study
 - each population may have a different frequency of effect modifiers, limiting the external validity of a study which is restricted to a particular sub-population
 - the estimated effect should be reported for each level m of modifier M
- causal hypotheses and effect modification
 - the null hypothesis of no average causal effect in the entire population does *not* imply that there is no average causal effect for particular subgroups of the population

- the sharp causal null hypothesis *does* imply that there is no average causal effect for any subgroup of the population
- effect modification and adjustment methods
 - discrepant effect measures result from different causal questions asked by each investigator rather than their choice of analytic approach
 - effect modification can be a major source of heterogeneity in effect estimates produced by IPW, standardization, stratification, and matching methods because they ask different questions
- it is not always obvious whether effect measure heterogeneity across levels of L is due to effect modification, may also be due to:
 - selection bias
 - information bias
 - confounding
 - chance

5.2. Effect Modification: The Specifics.

- M is a modifier of the effect of A on Y when the average causal effect of A on Y varies across levels of M
- modifier: pre-treatment variables M , in the absence of other biases (above)
 - mediator: cannot be considered as an effect modifier
- the presence and extent of effect modification depends on the scale on which the effect is assessed
 - **additive effect modification:** effect modification on the additive (absolute) scale

$$E[Y^{a=1}|M=1] - E[Y^{a=0}|M=1] \neq E[Y^{a=1}|M=0] - E[Y^{a=0}|M=0]$$
 - **multiplicative effect modification:** effect modification on the multiplicative scale

$$\frac{E[Y^{a=1}|M=1]}{E[Y^{a=0}|M=1]} \neq \frac{E[Y^{a=1}|M=0]}{E[Y^{a=0}|M=0]}$$
 - **qualitative effect modification:** causal effects across M are in opposite directions
 - * only for qualitative modification will we see effect modification on both the additive and multiplicative scales
 - effect modification on one scale does not imply effect modification on the other scale
 - *effect-measure modification* : synonym used to stress dependence of effect modification on a particular scale (measurement)

- **heterogeneity** or “heterogeneity of causal effects across strata of M ”

- test for homogeneity of difference measures

$$\chi^2_{m-1} = \sum_{i=1}^M \frac{(\hat{RD}_i - \hat{RD})^2}{\hat{Var}_i(\hat{RD}_i)}$$

* where \hat{RD} is the summary estimate of the RD assuming homogeneity and \hat{RD}_i is the stratum specific estimate for level m

* $\hat{Var}_i(\hat{RD}_i) = \frac{a_i(N_{1i}-a_i)}{N_{1i}^3} + \frac{b_i(N_{0i}-b_i)}{N_{0i}^3}$ for risk differences

* $\hat{Var}_i(\hat{RD}_i) = \frac{a_i}{N_{1i}^2} + \frac{b_i}{N_{0i}^2}$ for rate differences

- test for homogeneity of ratio measures

$$\chi^2_{m-1} = \sum_{i=1}^M \frac{(\log \hat{RR}_i - \log \hat{RR})^2}{\hat{Var}_i(\log \hat{RR}_i)}$$

* where \hat{RR} is the summary estimate of the odds/risk/rate ratio assuming homogeneity (RR_{MH}) and \hat{RR}_i is the stratum specific estimate for level m

* $\hat{Var}_i(\log \hat{RR}_i) = \frac{c}{a_i N_{1i}} + \frac{d}{b_i N_{0i}}$ for risk ratios

* $\hat{Var}_i(\log \hat{RR}_i) = \frac{1}{a_i} + \frac{1}{b_i}$ for rate ratios

* $\hat{Var}_i(\log \hat{OR}_i) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$ for odds ratios

- considerably more data is needed to assess heterogeneity (effect modification), thus a failure to reject the null hypothesis of no heterogeneity may be explained by a lack of power

- **stratification:** the causal effect of A on Y is calculated in each stratum of M

- for conditional risks

1): stratify by M

2): standardization or IPW by L

- How to calculate causal effect in $M=0$ group

- (1) Standardization
for risk difference:

$$\begin{aligned} \hat{RD} = & \sum_l E[Y^{a=1}|A=a, L=l, M=0] \times Pr[L=l|M=0] - \\ & \sum_l E[Y^{a=0}|A=a, L=l, M=0] \times Pr[L=l|M=0] \end{aligned}$$

for risk ratio:

using weights specific to M only make sense if there are no other effect modifiers in L

$$\hat{RR} = \frac{\sum_l E[Y^{a=1}|A=a, L=l, M=0] \times Pr[L=l|M=0]}{\sum_l E[Y^{a=0}|A=a, L=l, M=0] \times Pr[L=l|M=0]}$$

(2) IP weight

$$\frac{1}{Pr[A=a|L=l, M=0]}$$

- **generalizability:** external validity or transportability, only applicable when:
 - the distribution of effect modifier M is the same in both populations
 - there are no effect modifiers
- reasons investigators are interested in effect modification
 - (1) if M modifies the effect of A on Y , then the average causal effect in the population may not be generalizable to the other populations in which M is more or less frequent
 - (2) evaluating the presence of effect modification is helpful to identify the groups of subjects that would benefit most from an intervention
 - additive effect modification important for public health interventions
 - (3) identification of effect modification may help understand the biological, social, or other mechanisms leading to the outcome

5.3. Assessing Effect Modification.

- we only want to assess effect modification after we have controlled for any confounding or selection bias
 - that is, the stratum-specific effect estimates should be unbiased before we proceed to consider effect modification
- the use of stratification as an adjustment method to evaluate effect modification *always gets* $E[Y^{A=a}|M=m]$
- the use of IPW and Standardization as adjustment methods does not require the evaluation of effect modification because exchangeability and effect modification are handled separately *can get* $E[Y^{A=a}]$ and $E[Y^{A=a}|M=m]$

6. Interaction

$$Pr[Y^{A=a, E=1} = 1] \neq Pr[Y^{A=a, E=0} = 1]$$

Definition.

- the causal effect of A on Y after a joint intervention that set E to 1 differs from the causal effect of A on Y after a joint intervention that set E to 0
- interaction involves at least two treatments (joint interventions)
 - identifying conditions (exchangeability, positivity, and consistency) required for *all* treatments
 - must be able to block all non-causal paths for all treatments
 - interaction is assessed marginally
 - rather than viewing A and E as two distinct treatments with levels (0 or 1), they can be viewed as a combined treatment AE with levels (00, 01, 10, or 11)
- if no conditional exchangeability for E , one can still assess effect modification, however, cannot assess interaction without *identifiability* assumptions
- defining interaction by heterogeneity or a departure from expected joint effects are mathematically equivalent
- *synergism*: the presence of A increases the effect of E (vice versa)
- *antagonism*: the presence of A decreases the effect of E (vice versa)

6.0.1. *Interaction on the Additive Scale.*

- Defined by terms of Heterogeneity of Effects

$$\underbrace{Pr[Y^{a=1,e=1} = 1] - Pr[Y^{a=1,e=0} = 1]}_{\text{effect of } E \text{ when } A \text{ set to 1}} \neq \underbrace{Pr[Y^{a=0,e=1} = 1] - Pr[Y^{a=0,e=0} = 1]}_{\text{effect of } E \text{ when } A \text{ set to 0}}$$

and when identifiability conditions hold...

$$\begin{aligned} & \underbrace{Pr[Y = 1|A = 1, E = 1] - Pr[Y = 1|A = 1, E = 0]}_{\text{effect of } E \text{ in the subset } A = 1} \\ & \neq \\ & \underbrace{Pr[Y = 1|A = 0, E = 1] - Pr[Y = 1|A = 0, E = 0]}_{\text{effect of } E \text{ in the subset } A = 0} \end{aligned}$$

- Defined by Observed vs. Expected Joint Effects (i.e. the compact way)

$$\underbrace{\Pr[Y^{a=1,e=1} = 1] - \Pr[Y^{a=0,e=0} = 1]}_{\text{joint effect when } A \text{ and } E \text{ set to 1}} \neq \\
 \underbrace{(\Pr[Y^{a=1,e=0} = 1] - \Pr[Y^{a=0,e=0} = 1])}_{\text{main effect of } A, \text{ when } E \text{ set to 0}} + \underbrace{(\Pr[Y^{a=0,e=1} = 1] - \Pr[Y^{a=0,e=0} = 1])}_{\text{main effect of } E, \text{ when } A \text{ set to 0}}$$

and when identifiability conditions hold...

$$\underbrace{\Pr[Y = 1|A = 1, E = 1] - \Pr[Y = 1|A = 0, E = 0]}_{\text{joint effect } A \text{ and } E \text{ in total}} \neq \\
 \underbrace{(\Pr[Y = 1|A = 1, E = 0] - \Pr[Y = 1|A = 0, E = 0])}_{\text{main effect of } A \text{ in subset } E = 0} + \\
 \underbrace{(\Pr[Y = 1|A = 0, E = 1] - \Pr[Y = 1|A = 0, E = 0])}_{\text{main effect of } E \text{ in subset } A = 0}$$

- Definition of sub/super additive
 - subadditive: \neq can be replaced by $<$
 - superadditive: \neq can be replaced by $>$

6.0.2. Interaction on the multiplicative scale.

- Defined by Heterogeneity of effects

$$\frac{\Pr[Y^{a=1,e=1} = 1]}{\underbrace{\Pr[Y^{a=1,e=0} = 1]}_{\text{effect of } E \text{ when } A \text{ set to 1}}} \neq \frac{\Pr[Y^{a=0,e=1} = 1]}{\underbrace{\Pr[Y^{a=0,e=0} = 1]}_{\text{effect of } E \text{ when } A \text{ set to 0}}}$$

when identifiability conditions hold...

$$\underbrace{\frac{\Pr[Y = 1|A = 1, E = 1]}{\Pr[Y = 1|A = 1, E = 0]}}_{\text{effect of } E \text{ in subset } A = 1} \neq \underbrace{\frac{\Pr[Y = 1|A = 0, E = 1]}{\Pr[Y = 1|A = 0, E = 0]}}_{\text{effect of } E \text{ in subset } A = 0}$$

- Defined by Observed vs. Expected Joint Effects

$$\begin{aligned}
 & \left(\begin{array}{c} \Pr[Y^{a=1,e=1} = 1] \\ \underbrace{\Pr[Y^{a=0,e=0} = 1]}_{\text{joint effect when } A \text{ and } E \text{ set to 1}} \end{array} \right) \\
 & \neq \left(\begin{array}{c} \Pr[Y^{a=1,e=0} = 1] \\ \underbrace{\Pr[Y^{a=0,e=0} = 1]}_{\text{main effect of } A, \text{ when } E \text{ set to 0}} \end{array} \right) + \left(\begin{array}{c} \Pr[Y^{a=0,e=1} = 1] \\ \underbrace{\Pr[Y^{a=0,e=0} = 1]}_{\text{main effect of } E, \text{ when } A \text{ set to 0}} \end{array} \right)
 \end{aligned}$$

when identifiability conditions hold...

$$\begin{aligned}
 & \left(\begin{array}{c} \Pr[Y = 1|A = 1, E = 1] \\ \underbrace{\Pr[Y = 1|A = 0, E = 0]}_{\text{joint effect when } A \text{ and } E \text{ in total}} \end{array} \right) \\
 & \neq \left(\begin{array}{c} \Pr[Y = 1|A = 1, E = 0] \\ \underbrace{\Pr[Y = 1|A = 0, E = 0]}_{\text{main effect of } A \text{ in subset } E = 0} \end{array} \right) + \left(\begin{array}{c} \Pr[Y = 1|A = 0, E = 1] \\ \underbrace{\Pr[Y = 1|A = 0, E = 0]}_{\text{main effect of } E \text{ in subset } A = 0} \end{array} \right)
 \end{aligned}$$

- Definition of sub/super multiplicative
 - supermultiplicative: \neq can be replaced by $>$
 - submultiplicative: \neq can be replaced by $<$

6.0.3. *interaction vs. effect modification.*

- effect modification: effect of A on Y varies across levels of M , interested in Y^a
- interaction: joint effect of A and E on Y , interested in $Y^{a,m}$
- when treatment E is randomly assigned, methods to identify effect modification can be used to identify interaction
- interaction between A and E without effect modification is possible, but rare (requires exact cancellations)

6.1. Counterfactual Response Types & Sufficient Cause Model.

6.1.1. *Definitions.*

- **Counterfactual response type**
 - a particular counterfactual response under a specified treatment regime
 - may be related to one or more sufficient causes
- **Sufficient Cause**
 - the minimal set of factors (called U_i) that cause the outcome

- * a component is a necessary factor within the sufficient cause, together with the background factors it brings about the outcome
- * background factors U are not affected by treatment
- treatment not necessarily a component in all sufficient causes for outcome
 - * for simplicity those that do not contain treatment are lumped into U_0
- for a given outcome there are many sufficient causes

6.1.2. One binary treatment & Dichotomous Outcome.

- (1) there are four counterfactual response types for A

TABLE 6.1

Type	$Y^{a=0}$	$Y^{a=1}$	Component Causes
Doomed	1	1	U_0 or $\{U_1 \& U_2\}$
Preventative	1	0	U_2
Causative	0	1	U_1
Immune	0	0	none

- (2) there are three sufficient causes for Y

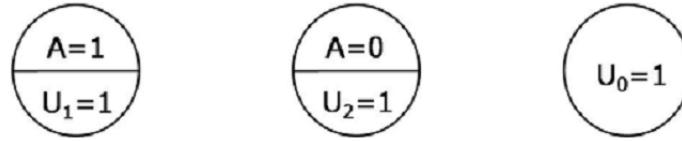


FIGURE 6.3. causative, preventative, and doomed

6.1.3. Two binary treatments & Dichotomous Outcome.

- (1) there are sixteen counterfactual response types for A and E

the presence of an *additive* interaction implies . . .

- individuals of type 1, 2, or 3 must exist
- i.e. counterfactual outcome for A requires taking E into account (vice versa)

the absence of an observed *additive* interaction implies . . .

- all individuals are of type 4
- or type 1, 2, and 3 exist but their departures from additivity cancel each other out

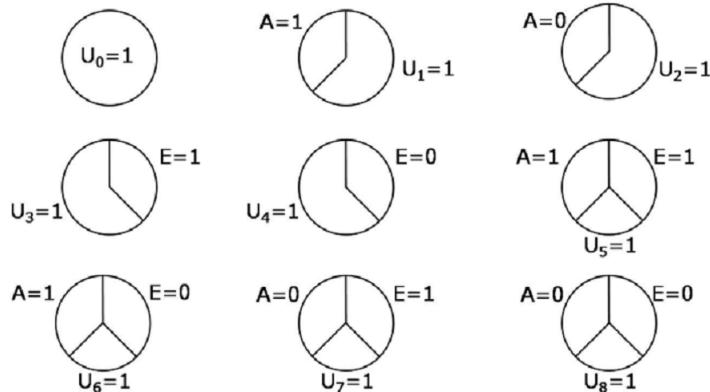
- (2) there are nine sufficient causes for A and E

the last four correspond to an *additive* interaction

TABLE 6.2

Type	Interaction	$Y^{a=1,e=1}$	$Y^{a=0,e=1}$	$Y^{a=1,e=0}$	$Y^{a=0,e=0}$	Component Causes
Type 1	Yes	1	0	0	0	U_5
		0	1	0	0	U_7
		0	0	1	0	U_6
		0	0	0	1	U_8
Type 2	Yes	1	0	0	1	U_5, U_8
		0	1	1	0	U_6, U_7
Type 3	Yes	1	1	1	0	U_5, U_6, U_7
		1	0	1	1	U_5, U_6, U_8
		1	1	0	1	U_5, U_7, U_8
		0	1	1	1	U_6, U_7, U_8
Type 4	No	1	1	1	1	U_0 or $\{U_1 \text{ to } U_8\}$
		0	0	0	0	none
		1	1	0	0	U_3, U_5, U_7
		0	0	1	1	U_4, U_6, U_8
		1	0	1	0	U_1, U_5, U_6
		0	1	0	1	U_2, U_7, U_8

FIGURE 6.4



6.2. Key points about Sufficient-Component Cause (SCC) model.

- (1) it is deterministic (recent extensions are making it stochastic)
- (2) it is limited to dichotomous treatments and outcomes (recent extensions for categorical/ordinal data)
- (3) the magnitude of the causal effect of treatment A depends on the distribution (prevalence) of effect modifiers
 - explains why an average causal effect depends on the characteristics of the population under study and may not be transportable
- (4) explains why it does not make sense to add attributable fractions for two separate

treatments A and E for the joint treatment $\{A, E\}$

- not all 9 sufficient-component causes for a dichotomous outcome and two dichotomous treatments exist in all settings
- persons can have more than one sufficient cause
 - some persons may have sufficient causes involving only A , only E , or both A and E
 - such persons would be counted more than once if two independent attributable fractions are added.
 - should calculate an ARF for a joint treatment by
 - (a) define exposure of interest, say $(A = 1, E = 1)$
 - (b) define non-exposure as all other levels, i.e. $(0-0, 1-0, 0-1)$
 - (c) excess risk (or rate) accordingly

6.3. Sufficient Causes & Exchangeability.

Definitions.

- (1) a binary outcome Y and binary treatment $A = 1$ if treated, $A = 0$ if untreated
- (2) three sufficient causes as follows U_0 (doomed), U_1 (causative), and U_2 (preventative)
- (3) see figure 6.1

6.3.1. Partial exchangeability.

- Only allows us to estimate effect in the subgroup of the treated or the untreated (depending on the type of partial exchangeability we have)
- $Y^{a=0} \amalg A$
 - to simulate the counterfactual outcome under no treatment for those who were *actually treated*, we need to be able to assume that the risk of death *under no treatment* is the same in both groups.
 - that is, the frequency of the combination of U_0 and U_2 (doomed and preventative persons) who would develop the outcome *in the absence* of treatment, is the same in the persons who were treated and untreated
- $Y^{a=1} \amalg A$
 - to simulate the counterfactual outcome under receiving treatment for those who were *actually untreated*, we would need to be able to assume that the risk of death *under receiving treatment* is the same in both groups.
 - that is, the frequency of the combination of U_0 and U_1 (doomed and causative persons) who would develop the outcome *in the presence* of treatment, is the same in persons who were treated and untreated

6.3.2. Full exchangeability.

- Allows us to estimate effect in entire population and in both subgroups of treatment levels
- $Y^a \perp\!\!\!\perp A$ for all a (continuous treatment)
- $Y^{a=0,a=1} \perp\!\!\!\perp A \rightarrow Y^a \perp\!\!\!\perp A$ for all a (dichotomous treatment)
- for a dichotomous treatment, we see that both cases of partial exchangeability must be assumed in order to assume full exchangeability.

6.4. Sufficient Cause Interaction.

6.4.1. Definition.

- sufficient cause interaction: interaction within the sufficient-component-cause framework
- defined as joint presence of treatments A and E in the same *sufficient cause*
- sufficient cause interaction can be *synergism* or *antagonism*

synergism: $A = 1$ and $E = 1$ present in the same sufficient cause

antagonism: $(A = 1 \text{ and } E = 0) \text{ or } (A = 0 \text{ and } E = 1)$ present in the same sufficient cause

- makes explicit reference to the causal mechanisms involving the treatments A and E
- however, sometimes it does not require knowledge of mechanism to identify sufficient cause interaction
- sufficient condition for synergism

$$Pr[Y^{a=1,e=1} = 1] - (Pr[Y^{a=0,e=1} = 1] + Pr[Y^{a=1,e=0} = 1]) > 0$$

– this is a sufficient but not necessary condition of synergistic interaction

* implies observed joint effect is greater than expected joint effect

* to see this divide all terms by $Pr[Y^{a=0,e=0}]$

– fulfillment of above condition do **NOT** always imply synergism

– synergism **always** imply the fulfillment of above condition

– this is a very strong condition that misses most cases of synergism

– conceptually, this condition implies

$$\{U_5 = 1\} - \{U_0 = 1\} - \{U_2 = 1\} - \{U_4 = 1\} - \{U_6 = 1\} - \{U_7 = 1\} > 0$$

6.4.2. Monotonicity.

- is a weaker sufficient condition of synergism
- refers to monotonic effect of A and E
 - no preventive type persons for the *joint* treatment $\{A, E\}$ in the population
 - the absence of one component A does not increase $E[Y]$ compared to the absence of A , holding the other component constant E
- sufficient condition for synergism with monotonicity assumption

$$Pr[Y^{a=1,e=1} = 1] - (Pr[Y^{a=0,e=1} = 1] > Pr[Y^{a=1,e=0} = 1] - Pr[Y^{a=0,e=0} = 1])$$
- conceptually, this condition implies

$$\{U_5 = 1\} > 0$$

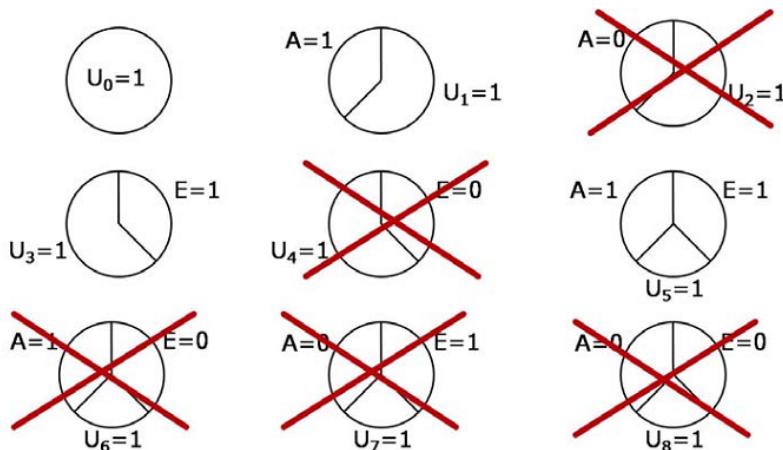


FIGURE 6.5. With monotonicity assumption, some sufficient causes are guaranteed not to exist

6.4.3. Interpretation.

- counterfactual framework and sufficient cause framework answer different questions
 - (1) counterfactual: the outcome that would have occurred under a particular action
 - what happens?
 - a way to study causes and effects
 - (2) sufficient cause: the mechanism that caused the outcome
 - how does it happen?

- a way to study the effects of causes
- biologic interaction
 - although sufficient cause interaction is often referred to as biologic interaction, sufficient cause interaction does not imply that A and E physically interact with each other

CHAPTER 7

Bias

1. Directed Acyclic Graphs: concepts

1.1. Definition.

- a DAG is a tool to...
 - represent our qualitative *expert knowledge* and a priori *assumptions* about the causal structure of interest
 - classify bias
 - identify potential problems in study design and analysis
- communicate our assumptions (parsimony is desirable)
- definitions of **directed acyclic graph** (DAG)
 - **node**: A, L, Y, U , etc.
 - **edge**: \rightarrow
 - properties
 - * *Directed*: No bi-directional arrows
 - * *Acyclic*: No loops
 - * *Graph*: it's a visual representation of statistical (or causal) associations
 - * temporally ordered DAG: time flows from left to right
 - in a causal DAG or its statistical equivalent (later):
 - * no arrow from X to D means:
 - (1) X has NO causal effect on D for *any individual* in the population
 - (2) we *assume* that X has NO causal effect on D for *any individual* in the population
 - * arrow from A to Y means:
 - (1) A has causal effect on Y for *at least one* individual in the population
 - (2) We do *NOT* want to assume that A does not have a causal effect on any individual in the population
 - * $A \rightarrow L \leftarrow Y$: L is a **collider**

- formal definition
 - (1) nodes: random variables $V = (V_1, \dots, V_M)$
 - (2) PA_m are the **parents** of V_m : the set of nodes with arrows into V_m
 - (3) V_j is a **descendant** of V_m if there is a sequence of nodes connected by edges between V_j and V_m
 - V_j need not be a direct descendant (child)

Conditioning.

- conditioning is calculating the effect measure *within* levels of the confounder (i.e. using only the subsets' data) to give stratum-specific effect measures (i.e. stratification or restriction)
- **conditional independence** implies that we can remove an association between two variables by stratifying on a third variable
 - we generally stratify on a variable L that lies along a path between treatment A and outcome Y to remove sources of association that are non-causal
 - how the stratification affects A and Y 's association depends on what kind variable L is
- let's assume treatment A , outcome Y , and covariate B , L , and/or U (unmeasured)
- stratifying on a third variable:
 - (1) Path contains only **intermediate variable** (see figure 7.1)
 - (a) prior to stratification A and Y are marginally *associated*
 - (b) stratifying on B removes the association between A and Y
 - (c) thus, A and Y are conditionally independent given B
 - (d) usually we prefer to not stratify on intermediates because:
 - (i) it removes the indirect effect of A on Y
 - (ii) there may be an unmeasured variable U that is a common cause of B and Y

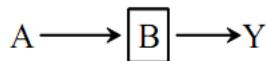


FIGURE 7.1. Conditioning on an Intermediate

- (2) Path contains a **common cause** (see figure 7.2)
 - (a) prior to stratification A and Y are marginally *associated*
 - (b) stratifying on L removes the association between A and Y
 - (c) thus, A and Y are conditionally independent given L

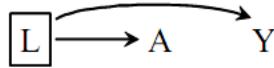


FIGURE 7.2. Conditioning on a Common Cause

- (3) Path contains a **common effect** (i.e. collider, *see figure 7.3*)
- prior to stratification A and Y are marginally *independent*
 - stratifying on L creates an association between A and Y
 - thus, A and Y are *not* conditionally independent given L when it is a common effect

when a variable is the common effect of two other variables, those two variables are guaranteed to provide information about each other within levels of the common effect (Sprinkler-Rain-Grass-is-wet analogy)

- the sprinkler is on an automatic timer
- both the sprinkler and rain cause the grass to be wet
- if the grass is wet and I tell you the sprinkler was off, do you know that it rained?

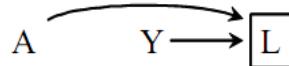


FIGURE 7.3. Conditioning on a Common Effect

- (4) Conditioning on a variable's **descendant**

- Let's say that A has a direct cause on B so that $A \rightarrow B$
- to the degree that A and B are associated, stratifying on A 's descendant B will cause stratification on A also. Two implications:

descendant of a common cause: (*see figure 7.4*)

- * cannot measure U directly but can measure its descendant L
- * adjusting L will control for confounding to the degree that L is associated with U
- * residual confounding occurs when the L is not highly correlated with U
- * the same concept applies to *measurement error* (L = truth, L^* = measurement)

descendant of a common effect: (*see figure 7.5*)

- * same effect as stratifying on the parent, which is a collider
- * the path that the parent formerly blocked is now open

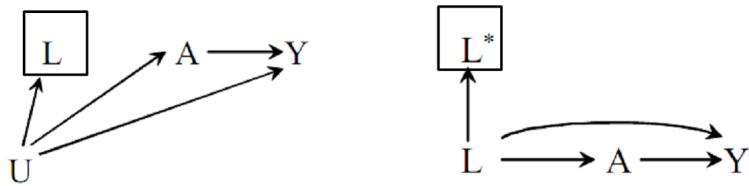


FIGURE 7.4. Conditioning on the Descendant of a Common Cause (Left: Surrogate Confounder; Right: Measurement Error)

- * in figure 7.5, A and Y are associated because C is stratified upon
- * this situation happens often in the study of direct effects, joint effects, and the effects of time-varying treatments.

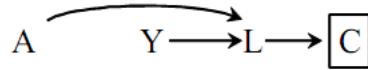


FIGURE 7.5. Conditioning on the Descendant of a Common Effect

1.2. Statistical assumptions in a DAG.

Markov Factorization.

- a DAG describes conditional associations between the included variables
- for a DAG, we can write the joint distribution as a Markov factorization

$$f(v) = \prod_{j=1}^M f(v_j | pa_j)$$

- the joint distribution captures the presence or lack of arrows (associations/effects)
- A complete DAG: each variable is a descendant of all variables preceding it

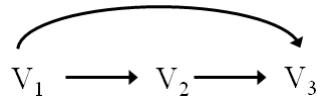


FIGURE 7.6. a Complete DAG

$$f(V_3, V_2, V_1) = \underbrace{f(V_3 | V_2, V_1) \times f(V_2 | V_1) \times f(V_1)}_{\text{Markov factorization}}$$

- with assumptions we remove arrows (the independence is reflected in the joint distribution)

$$f(V_3, V_2, V_1) = f(V_3 | V_1) \times f(V_2 | V_1) \times f(V_1)$$

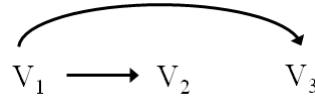
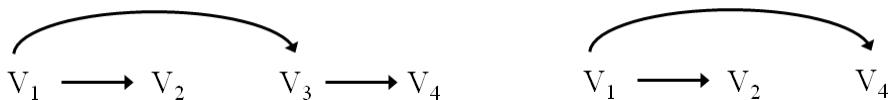


FIGURE 7.7. Incomplete DAG

- comparing the complete and incomplete DAGs we find that
 - $f(V_3|V_2, V_1) = f(V_3|V_1) \rightarrow V_3 \perp\!\!\!\perp V_2|V_1$
 - $f(V_2|V_1) = f(V_2|V_1) \rightarrow V_2 \perp\!\!\!\perp V_1$
- we can marginalize over a variable if the conditional associations of other variables are preserved
 - consider the two DAGs:

FIGURE 7.8. marginalize over V_3

- these two dags are statistically equivalent

1.3. Causal DAGs.

Intuitive Examples.

- figure 7.9(a):
 - (1) conditionally randomized experiment
 - (2) observational study with the assumption that A depends on L but there are *no other causes of Y*
- figure 7.9(b):
 - (1) marginally randomized experiment
 - (2) $\underbrace{Pr[Y^{a=1} = 1]}_{causation} \neq \underbrace{Pr[Y^{a=0} = 1]}_{causation}$ and $\underbrace{Pr[Y = 1|A = 1]}_{association} \neq \underbrace{Pr[Y = 1|A = 0]}_{association}$
- figure 7.9(c):
 - (1) confounding by L
 - (2) $\underbrace{Pr[Y^{a=1} = 1]}_{no causation} = \underbrace{Pr[Y^{a=0} = 1]}_{no causation}$ BUT $\underbrace{Pr[Y = 1|A = 1]}_{association} \neq \underbrace{Pr[Y = 1|A = 0]}_{association}$
- figure 7.9(d):
 - (1) L is a collider of A and Y

$$(2) \underbrace{Pr[Y^{a=1} = 1] = Pr[Y^{a=0} = 1]}_{\text{no causation}} \text{ and } \underbrace{Pr[Y = 1|A = 1] = Pr[Y = 1|A = 0]}_{\text{no association}}$$

(3) remember that **being a collider is path-specific!**

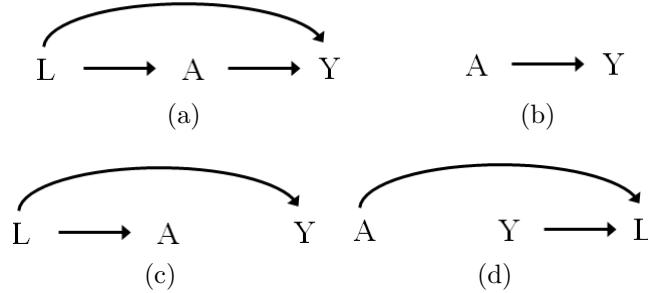


FIGURE 7.9. The simplest cases

Formal Definition.

- A DAG is a *causal* DAG if...
 - (1) lack of arrow means “no direct causal effect” with respect to other variables in the DAG
 - (2) *all* common causes are on the DAG (even unmeasured)

Statistical vs. Causal DAG.

- a statistical DAG represents associations because it does not have all common causes of treatment and outcome on the graph
- a causal DAG can be translated to a statistical DAG if all the implied conditional associations are preserved
- in a statistical DAG the following all represent association:

$$V_1 \rightarrow V_2 \rightarrow V_3 \quad V_1 \leftarrow V_2 \rightarrow V_3 \quad V_3 \rightarrow V_2 \rightarrow V_1$$

- we use a statistical DAG when we have unmeasured confounders in the causal DAG

Examples

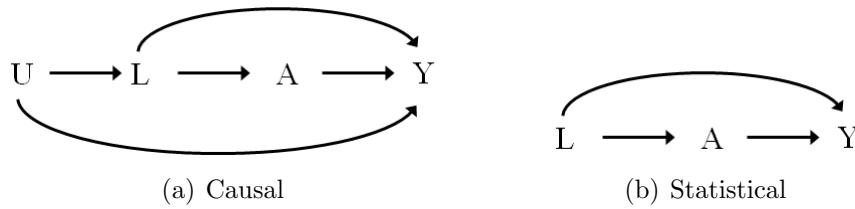
(1) For both DAGs $Y \perp\!\!\!\perp A$ and $Y \perp\!\!\!\perp A|L$ see figure 7.10

- for 7.11(a):
- for 7.10(b):

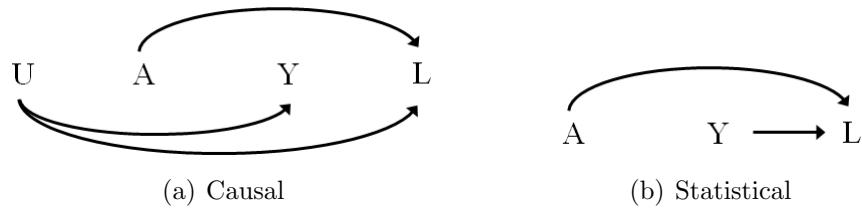
(2) For both DAGs $Y \perp\!\!\!\perp A$ and $Y \perp\!\!\!\perp A|L$ see figure 7.11

(3) Cannot get rid of U see figure 7.12

- because U is a direct cause of A

FIGURE 7.10. A randomized but L observed

- (a) $f(Y|A, L, U) \times f(A|L, U) \times f(L|U) \times f(U)$
 (b) $f(Y|A, L) \times f(A|L) \times f(L)$

FIGURE 7.11. A randomized but L observed

- (a) $f(L|A, U) \times f(Y|U) \times f(A)$
 (b) $f(L|Y, A) \times f(Y) \times f(A)$

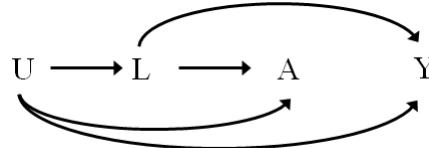


FIGURE 7.12. Observational Study

Causal Markov Assumption.

- need to make *assumptions* to **link the causal DAG to the observed data in a study**
- causal Markov assumption: $f(v) = \prod_{j=1}^M f(v_j|pa_j)$
- we assume this holds true for our causal DAG

Non parametric structural equation model (NPSEM).

- (1) a variable is a function of its parents and a random error term
 - deterministic unknown function $f_m = (pa_m, \epsilon_m)$
 - independent random errors ϵ_{v_m} equivalent to “no unmeasured confounders”
- (2) a DAG representing a NPSEM is a *causal* DAG, which satisfies the causal Markov assumption (implied by mutually independent ϵ)

- (3) a NPSEM is a causal DAG, *but* not all causal DAGs are NPSEMs
- (4) a NPSEM is a *fully randomized causally interpreted structured tree graph* (the IPW tree)

1.4. Faithfulness.

Definition.

- Faithfulness occurs when the associations implied by the *causal* DAG's structure exist in the data
- the distribution of the data is *not* faithful to the causal DAG when the joint distribution of the data does not reflect associations implied by the causal DAG.

Sources of Faithfulness Violations.

- the average causal null holds but the sharp causal null does not
 - ex: equal and opposite cancellation
- rarely occurs by chance, but can be guaranteed for some study designs (matching)

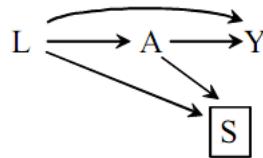


FIGURE 7.13. Matched Prospective Cohort Study

Example: Matched Prospective Cohort Study.

- S represents a matching factor that is determined by treatment and some confounding variable you wish to control for
 - S is a common cause because being selected depends on where the subject lies in the joint distribution of treatment and confounder
 - there may not be suitable matches for someone who has that particular level of confounder *and* has the opposite treatment history
- counter to the *causal* DAG, the data shows no association between L and A in the matched study population
- $L \rightarrow A$ association is equal in magnitude and opposite in sign to the $L \rightarrow_L S \leftarrow A$ association
- matching has the inverse magnitude of the propensity score
- there is no need to adjust for L to eliminate bias (in this particular example), however one might adjust for L for gains in statistical efficiency

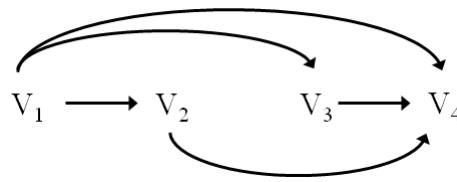
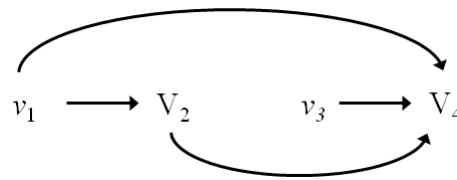


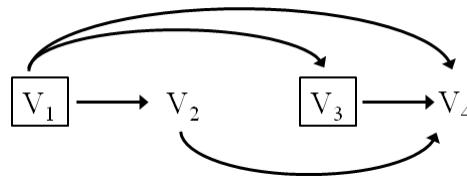
FIGURE 7.14. original DAG

1.5. Intervention vs. Stratification.

- g-methods remove arrows by simulating counterfactual worlds where we “intervene” and set treatment to a particular value
 - in each world, everybody receives the same level of treatment (thus treatment is constant)
 - in a DAG we express this by removing all arrows flowing into the “set” variable and making it lowercase
 - the causal effect contrasts the entire population under different treatment assignments

FIGURE 7.15. V_1 and V_3 set to v_1 and v_3

- stratification constricts the analysis to occur within strata
 - in a DAG we express this by placing boxes around all stratification factors
 - the causal effect contrasts two subgroups that happen to
 - have the same values for stratification factors
 - differ only in treatment status

FIGURE 7.16. V_1 and V_3 stratified upon

- when all stratification covariates are set, the causal interpretation of g-methods and stratification are equivalent

1.6. Exchangeability & D-separation.

Exchangeability.

- for basics on exchangeability, see Chapter 6. 4
- what we want to know is if the counterfactual outcome Y^a is independent of treatment assignment $A = a$ with respect to other covariates, that is are the treatment groups comparable
 - exchangeability can be a result of a random natural distribution or the result of randomization ($Y^a \perp\!\!\!\perp A$)
 - exchangeability can be conditional on a variable L ($Y^a \perp\!\!\!\perp A|L$)
- DAGS can be useful tools to assess exchangeability:

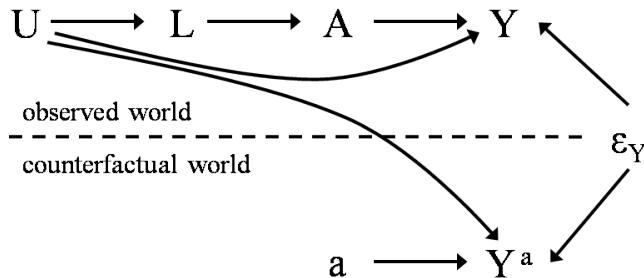


FIGURE 7.17. DAG integrating the counterfactual world $A = a$

- the DAG in figure 8.2 gives an example of how an unmeasured variable U can cause a lack of exchangeability (association) between treatment A and counterfactual outcome Y^a
- a simple rule of thumb is to think of Y^a as a pre-treatment variable (part of U) and assess if there is an unblocked backdoor pathway between A and U
- exchangeability can always be created by conditioning on the observed past (Markov factorization) of A (figure 7.18)

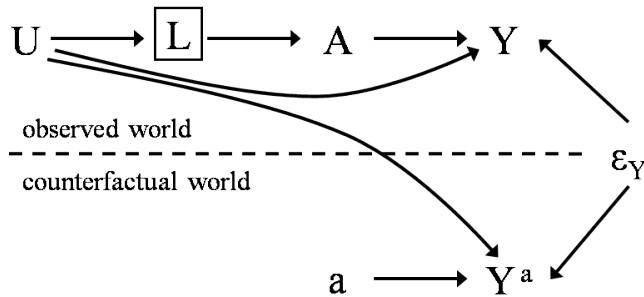


FIGURE 7.18. DAG showing how conditioning on the observed past (L) can create exchangeability ($Y^a \perp\!\!\!\perp A|L$)

D-separation. (Π_{d-sep})

- a set of graphical rules to decide if two variables in a DAG are independent (d-separated) or not independent (d-connected)
- “d-” stands for directional
- a path is any set of nodes (variables) connected by edges
- a path is blocked or unblocked depending on the following rules:
 - (1) A path is blocked if two arrowheads on a path collide
 - $L \rightarrow A \rightarrow Y$ is unblocked
 - $A \rightarrow Y \leftarrow L$ is blocked (collider)
 - (2) A path that contains a non-collider that is conditioned on is blocked
 - (3) A collider that has been conditioned on does not block a path
 - (4) A collider that has descendants that have been conditioned on does not block a path
 - (5) **being a collider is path-specific!**
- *unconditional d-separation*: A is d-separated from B in DAG G if and only if all paths between them are blocked, $(A \Pi_{d-sep} B)_G$
- *conditional d-separation*: A is d-separated from B in DAG G given L if and only if all paths between them are blocked when conditioning on L , $(A \Pi_{d-sep} B|L)_G$
- Conclusions:
 - a path is blocked if it contains a non-collider that has been conditioned on, or contains a collider that has not been conditioned on and that collider has no descendants that have been conditioned on
 - causes (ancestors) are not independent of their effects (descendants) and vice versa
 - generally, two variables are associated if they share a common cause, but sharing a common effect does not always imply the two causes are associated

1.6.1. *Back-door Criterion & Sequential Back-Door Criterion.*

- *Point Exposures*
 - $(Y \Pi_{d-sep} A|L)_G \implies Y^a \Pi A|L$
 - achieving d-separation with measured confounders L implies that, given L , we have conditional exchangeability for treatment A
- *Joint Exposures*
 - a stronger version, the *Sequential Back-Door Criterion*, is required for time-varying exposures, treatment regimes, etc.

- $(Y \amalg_{d-sep} A_k | \bar{L}_k, \bar{A}_{k-1})_G \implies Y^a \amalg A_k | \bar{L}_k, \bar{A}_{k-1}$
- implication: to estimate causal effect of a joint treatment
 - (1) need to fulfill backdoor criterion (i.e. conditional exchangeability) for *each component* of treatment A (i.e. A_i)
 - (2) required for all functions of a joint exposure (e.g. cumulative, average, etc.)
- recall that we fulfill the back-door criterion by testing for conditional independence between Y and A given a set of variables S
 - for point exposures, S is the set of variables in L
 - for joint exposures, S is the set of variables in \bar{A}_{k-1} and \bar{L}_k
 - where \bar{A}_k is history of treatment up to time $k \implies \{A_0, A_1, \dots, A_k\}$
 - where \bar{L}_k is covariate history up to time $k \implies \{L_0, L_1, \dots, L_k\}$
- fulfilling the back-door criterion can be achieved by removing or blocking all back-door pathways...
 - in the study design (restriction, matching)
 - in the analysis (stratification, g-methods)

1.7. Graphs, Counterfactuals, and Interventions.

- causal diagrams encode expert knowledge and assumptions about the causal structure of the problem
- emphasizes not all variables are created equal
 - requirements on treatment variables A that do not apply to other variables like L (exchangeability, positivity, and consistency)
 - decision nodes (representing the potential interventions) are not included for simplicity and to avoid redundancy

Identifying Assumptions Represented in Causal DAGs.

- *Nodes*
 - (1) *treatment node*: a node A for which we want $Y^{A=a'}$ vs. $Y^{A=a''}$
 - must fulfill positivity, consistency, and exchangeability
 - enclosed in circles in causal tree graphs (i.e. see IPW figure)
 - (2) *non-treatment node*: does *not* have to fulfill positivity and consistency
- *Positivity*
 - arrows from the nodes L to the treatment node A (i.e. $L \rightarrow A$) are *not* deterministic. When this is fulfilled there is variation in A for each level of L

- only concerns arrows arriving *into* treatment nodes ($\rightarrow A$)
- *Consistency*
 - the effect of treatment, represented by arrows from treatment node A to outcome Y (i.e. $A \rightarrow Y$) is well defined
 - only concerns arrows departing *from* treatment nodes ($A \rightarrow$)
- *Exchangeability*
 - lack of paths between A and Y nodes, other than those originating from A , that would result in an association between A and Y
 - i.e. back-door criterion fulfilled

1.8. Structure of Effect Modification.

- causal DAGs are helpful in illustrating effect modification
- $M \rightarrow Y \leftarrow A$
 - M is the effect modifier
 - Y is the outcome
 - A is the treatment
 - example: how quality of care M can modify the effect of heart transplant A on survival Y
- two important caveats:
 - (1) causal DAG would still be valid if it did not include M
 - because the causal question makes reference to M , M is included
 - (2) the causal diagram does not necessarily indicate the presence of effect modification by M and cannot distinguish between:
 - (a) the causal effect in stratum $M = 1$ is in the same direction as $M = 0$
 - (b) the causal effect in stratum $M = 1$ is in the opposite direction as $M = 0$
 - (c) treatment A has a causal effect in one stratum of M but not in another
- many effect modifiers do not have an effect on the outcome, but must be associated with the outcome (different from $M \rightarrow Y \leftarrow A$)
 - true effect modifier: M has a direct causal effect on Y
 - surrogate effect modifier: a variable associated with M that is associated with Y through M
 - * simply a variable associated with the true effect modifier
 - * can be associated with a true effect modifier M by structures including common causes, conditioning on common effects, or cause and effect

- causal diagrams are agnostic about interactions between treatments A and E
 - *however*, can encode information about interactions when using nodes that represent sufficient-component causes

1.9. Structural Classification of Bias.

- An association between two variables may exist because of:
 - (1) cause and effect
 - (2) bias
 - (3) chance (because of sampling variability)
- common causes and stratification on a common effect are structural sources of bias (inappropriate stratification)
- chance is not bias because it is not structural
- *bias* can only be defined in terms of data's structure, thus bias is the result of a structural source of non-causal association
- there are a finite number of causal structures that can give rise to bias:
 - (1) when treatment and outcome share a common cause
 - (2) when treatment and outcome share a common effect that it, or its descendant, is stratified upon
 - (3) cause and effect: reverse causation and information bias
 - reverse causation:** outcome precedes (and has a causal effect on) measurement of treatment
 - information bias:** measurement error of treatment, outcome, or confounders

2. Confounding

2.1. Association is not Causation.

- due to common cause or due to unblocked backdoor path
- 2 sources of association between A and Y
 - (1) $A \rightarrow Y$ (**A has causal effect on Y**)
 - (2) $A \leftarrow L \rightarrow Y$ (**backdoor path**)

2.2. Examples of Confounding.

- (1) healthy worker bias (figure 7.19 part (a))
- (2) confounding by indication or channeling (figure 7.19 part (b))
- (3) population stratification (figure 7.19 part (c))

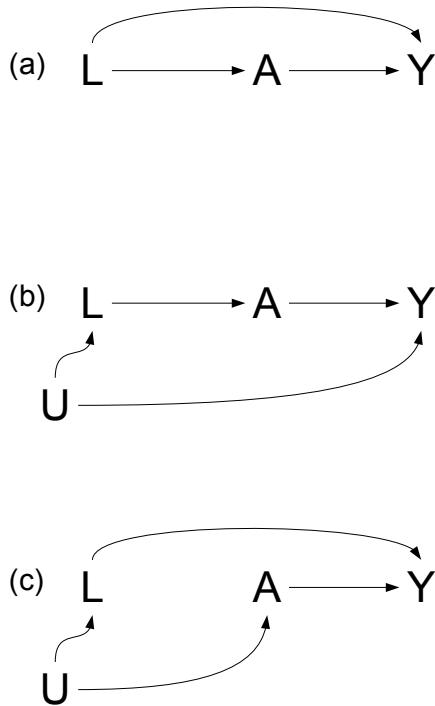


FIGURE 7.19. Example DAGs for confounding

2.3. Backdoor Criteria. The causal effect of A on Y is identifiable if all backdoor paths between them can be blocked.

- No common causes: marginally randomized experiment
- Common causes but enough measured variables to block all backdoor paths: conditionally randomized experiment
- 3 questions about confounding:
 - (1) Does confounding exist? Yes, if there is any unblocked path between A and Y
 - (2) Can confounding be eliminated? Yes, if all backdoor paths can be blocked using measured variables
 - (3) What variables are necessary to eliminate the confounding? The *minimal* set of variables that block all backdoor paths
- backdoor criteria does **NOT** show the magnitude or direction of confounding

2.4. Definition of a Confounder.

- any variable that can be used to reduce the bias caused by common cause between A and Y

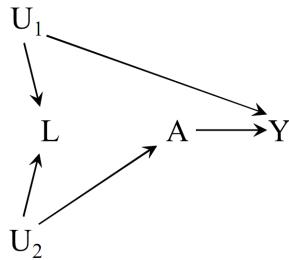


FIGURE 7.20. A counter example of traditional definition of confounder

- any variable that can be used to block backdoor path between A and Y
- traditional definition
 - (1) confounder is associated with A
 - (2) confounder is associated with Y *conditional on A* (or among the *untreated*)
 - (3) confounder is *not* on the causal pathway between A and Y
- according to traditional definition, in figure 7.19, all L s are confounders
- in figure 7.20, L matches the traditional definition of confounder **BUT** conditioning on L does *not* eliminate confounding (there was no confounding at all) but creates selection bias

2.5. Structural vs. Statistical Definition of Confounding.

- structural definition of confounding shows that **statistical criteria is NOT sufficient to define confounding**
- the structural definition first characterizes confounding as the bias resulting from the presence of common causes and then characterizes confounder as any variable that is necessary to eliminate the bias in the analysis
- confounding is an absolute concept; confounder is a relative concept
- confounding defined by counterfactuals
 - (1) $Y_a \perp\!\!\!\perp A$
 - no marginal exchangeability
 - no “marginally randomized experiment”
 - there is common cause of A and Y
 - there is confounding between A and Y
 - (2) $Y_a \perp\!\!\!\perp A|L$
 - no conditional exchangeability

- no exchangeability conditional on L
- no “conditionally randomized experiment”
- there is unblocked backdoor paths between A and Y given L
- there is *unmeasured* confounding between A and Y (residual and unmeasured confounding)
- our hope for observational study: even though there are common causes, there will be no *unblocked* backdoor path given measured L s
- additional assumptions (no selection bias and no measurement error) are needed for causal inference in observational study

2.6. Surrogate & Time-Varying Confounders.

Surrogate confounder: see figure 7.21

- (1) is not on any path between treatment A to Y
- (2) is a direct descendant of a variable that does lie on a path from treatment A to Y which, if measured, would be used to block all non-causal paths between A and Y
- (3) because it is a descendant of a confounder, stratifying on this variable causes stratification of its parent, the confounder
- (4) the path is blocked only to the degree that the surrogate confounder L and the unmeasured confounder U are correlated

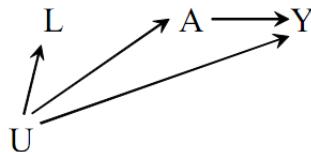


FIGURE 7.21. A surrogate Confounder

Time-varying confounder: a time-varying variable that is needed to block the backdoor paths of a time-varying treatment

2.7. Methods to adjust for confounding.

- can occur in the design or analysis phase
- ideal randomized experiment: do not need subject-matter knowledge; randomization does everything
- observational study: need subject-matter knowledge to identify L
 - (1) G-methods (“generalized”)
 - (a) standardization, IPW, G-estimation

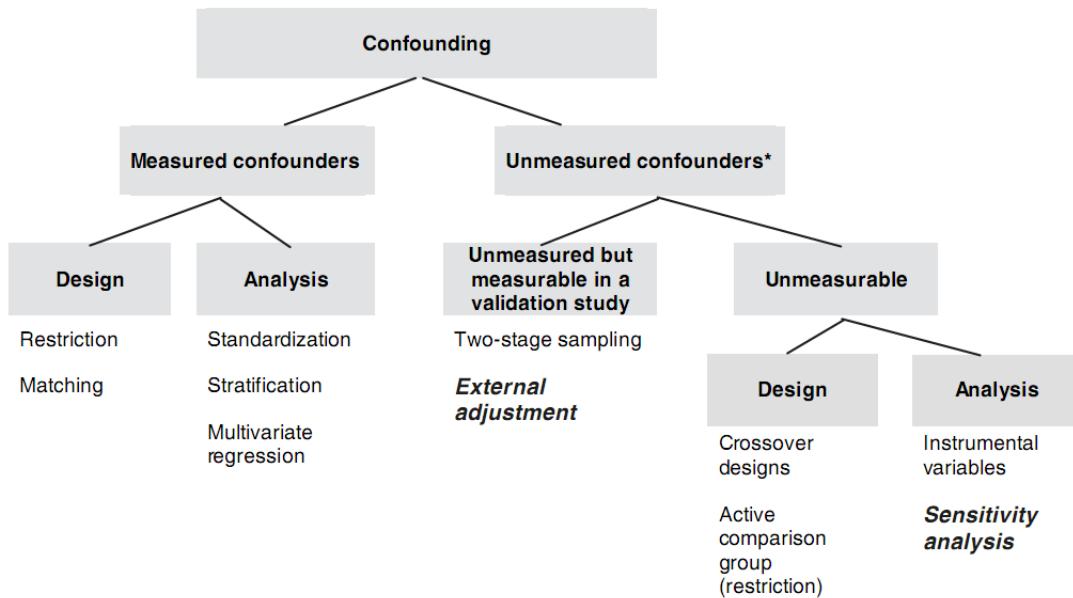


FIGURE 7.22. Strategies for handling confounding

- (b) **estimate causal effect**
 - (c) use conditional exchangeability in subsets defined by L to estimate the causal effect of A on Y in the entire population or in any subset of it
 - (d) simulate the distribution of counterfactual outcomes in the population if backdoor paths involving the measured variables L did not exist
 - (e) **in DAG:** delete the arrow from L to A
- (2) stratification-based methods
 - (a) stratification, restriction, matching
 - (b) **estimate association**
 - (c) use conditional exchangeability in subsets defined by L to estimate the association between A and Y in those subsets only
 - (d) estimate the association between A and Y in one or more subsets of the population in which the treated and the untreated are assumed to be exchangeable
 - (e) **in DAG:** stratification/restriction: box around L
 - (f) **in DAG:** matching: add a “selection” (S) node that is conditioned on

2.8. Identification vs Control of Confounding.

- both stratification-based and g-methods require conditional exchangeability given L to identify the causal effect of A on Y
 - (1) causal effect in *entire* population: conditional exchangeability in all levels of L
 - (2) causal effect in *one level of L* : conditional exchangeability in that level of L
- expert knowledge can be used to select confounders and avoid adjusting for nonconfounders
- the definition of confounding does not depend on the adjustment method

2.9. Strength and Direction of Confounding Bias.

- relationship between the confounder and the exposure of interest is a feature of the particular study base and the individuals sampled into study
 - (1) matching in *cohort* study
 - (2) restriction
 - (3) selection of a population base that is not characterized by the confounder-exposure association
 - (4) randomization
- confounding affects both *point estimates* and *validity of statistical inference* (e.g. p-value, confidence interval, etc.)
- **Direction**
 - apparent effect estimate is higher or lower than the true estimate
 - figure 7.23
 - *Definitions*

Upward Bias: apparent effect measure $>$ true effect measure

Downward Bias: apparent effect measure $<$ true effect measure

– *Signed Causal DAGs*

- (1) assess confounder-treatment and confounder-outcome associations
 - * add a positive sign \oplus over the arrow $L \implies A$ if the L has a positive causal effect on A (similarly for $L \implies Y$)
 - * add a negative sign \ominus over the arrow $L \rightarrow A$ if the L has a negative causal effect on A (similarly for $L \rightarrow Y$)
- (2) if $L \rightarrow A$ and $L \rightarrow Y$ are either (a) both positive or (b) both negative
→ upward bias
- (3) if $L \rightarrow A$ and $L \rightarrow Y$ are of opposite sign → downward bias
- the above procedure only works in simple settings and when the variables are dichotomous

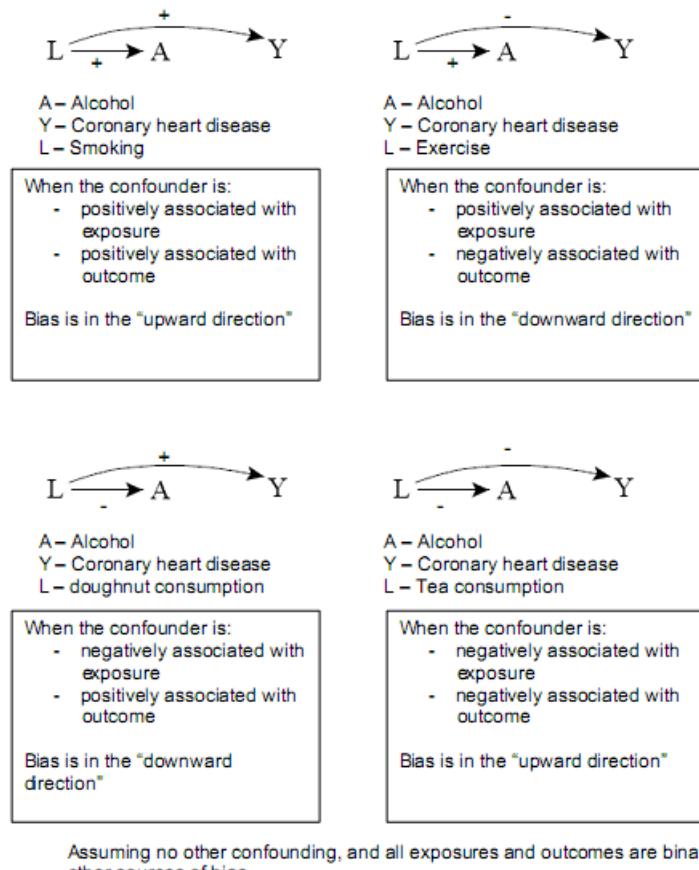


FIGURE 7.23. the direction of confounding

- **Magnitude**

- How far will the bias move my estimate?
- A **sensitivity analysis** can repeat the analysis under several assumptions
 - * for each analysis assume (simulate) the magnitude of the bias
 - * see how much your effect estimate changes under the different scenarios
 - * quantify the bias by setting bounds as the amount of confounding that would have to occur to meaningfully change your estimate
- For dichotomous outcome, treatment, and confounder the bias is limited
 - * Definitions
 - (1) RR_E : the RR for disease associated with the exposure of interest

$$RR_E = \frac{Risk_{exposed}}{Risk_{unexposed}}$$

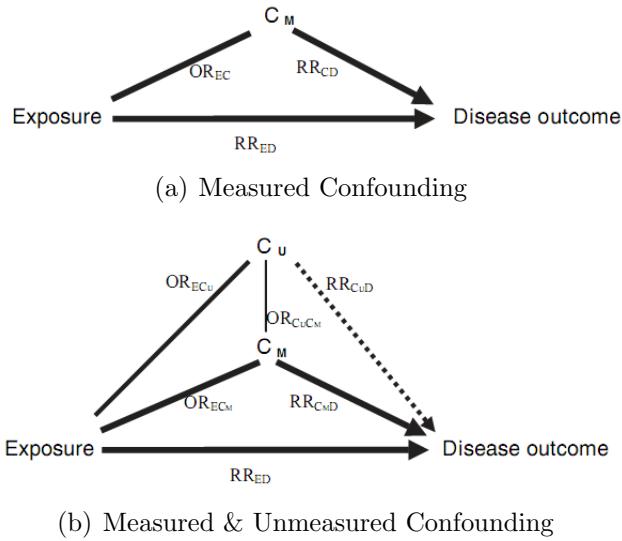


FIGURE 7.24. Strategies for controlling confounding

(2) RR_{CD} : the RR for disease associated with the confounder

$$RR_{CD} = \frac{Risk_{L=1}}{Risk_{L=0}}$$

(3) RR_{EC} : the RR for confounder associated with exposure

$$RR_{EC} = \frac{Pr[L=1|E=1]}{Pr[L=1|E=0]}$$

(4) $Pr[L=1]$: prevalence of confounder in the study population

* **assuming** $RR_E > 1$, $RR_{CD} > 1$, $RR_{EC} > 1$

* for preventive effect, recode the exposure

(1) $RR_{CD} > 1 \rightarrow RR_E < RR_{CD}$

(2) $RR_{CD} > 1$ and $RR_{EC} > 1 \rightarrow RR_E < RR_{EC}$

(3) thus, $1 < RR_E < \min(RR_{CD}, RR_{EC})$

(4) further limited by the prevalence of the confounder $\Rightarrow Pr[L=1]$

* lowest when $Pr(L=1) = 1$ or 0

(5) also, can use $OR_{EC} = \frac{Pr[L=1|A=1]/Pr[L=0|A=1]}{Pr[L=1|A=0]/Pr[L=0|A=0]} = \frac{Pr[L=1|A=1] \times Pr[L=0|A=0]}{Pr[L=1|A=0] \times Pr[L=0|A=1]}$ instead of RR_{EC}

	<p>Upper bound of confounding: Determined by $\log(\min(\text{RR}(C), \text{RR}(EC)))$</p> <p>The most extreme bias a preventive effect can have (+,+) or (-,-)</p>
	<p>Null value: $\log(\text{RR})=0$</p> <p>Lower bound of confounding: Determined by $\log(1/ \min(\text{RR}(C), \text{RR}(EC)))$</p> <p>The most extreme bias a causative effect can have (-,+) or (+,-)</p>

FIGURE 7.25. Bounds of the magnitude of confounding

Adjustment stage	Method	Strengths	Limitations
Design	Restriction	Completely removes confounding in crude analysis, may also preserve positivity	limits possible study questions and generalizability
	Matching	removes confounding in crude analysis for cohort studies, ensures positivity	for ca-c-: bias if matching on factor affected by both outcome and exposure, reduced efficiency if matching on factor only associated with exposure, bias without stratified analysis in case-control studies, cannot assess effect of matching factors only effect modification by them, inappropriate for time-varying situations and intermediates
Analysis	Standardization	can estimate effect in subsets of population, can be used in time-varying situations (g-formula) and for intermediate variables	not valid with non-positivity, G-null paradox
	IPW	effect in subsets of population, can be used in time-varying situations, for intermediates and for censoring, can also be used with random nonpositivity	non-positivity, treatment model assumptions that can be misspecified, cannot assess time varying EM
	G-estimation	causal effect and CIs, allows for interaction between exposure and confounders	computationally complex, additional model with assumptions, cannot use logistic models
	Stratification	can be used even if it results in non-positive cells, but with loss of power, simple	not appropriate for time-varying situations or intermediates, bias efficiency trade-off
	Regression	same as stratification	same as stratification

TABLE 7.1. Comparison between methods for adjusting for confounding with measured confounders

Adjustment stage	Method	Strengths	Limitations
Design	Two stage Analysis	Reduced cost in data collection	complex ML analysis, additional assumptions about subsample and external data
	Crossover designs	no need to measure or control time-invariant confounders	does not address time-varying confounders, assumptions about hazard periods and carry over effects
Analysis	Instrumental Variables Sensitivity Analysis	no unmeasured confounding no additional data needed	additional unverifiable assumptions assumptions about different model specification, external data

TABLE 7.2. Comparison between methods for adjusting for confounding with unmeasured confounders

3. Selection Bias

3.1. Definition of Selection Bias.

- *Selection bias* is formally defined as the bias resulting from conditioning on the common effect of two variables:
 - (1) treatment or a variable associated with treatment
 - (2) outcome or a variable associated with outcome
- the process by which individuals are selected into the analysis creates a joint distribution of treatment and outcome where the two variables bear an association that is *not causal*. In general this can happen in three ways:
 - (1) Selection into study (matching, sampling frame in case-controls, institutional study, survivors etc.)
 - (2) Effect measures (e.g. hazard ratios) that condition on survival when the distribution of confounders in the exposed and unexposed changes over time (i.e. differential loss to follow up or survival, etc.)
 - (3) Inappropriate stratification during data analysis
- In DAG language, selection bias *always* involves a collider that has been conditioned upon
- Selection bias lacks both internal and external validity. Thus it is *not* the same as a mere lack of generalizability

internal validity: the effect measure for the population under study is causal

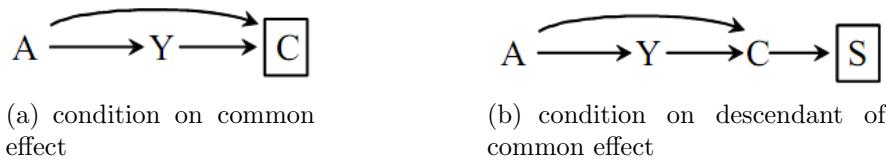


FIGURE 7.26. Selection Bias in a Randomized Trial

external validity: the effect measure for the population under study is both causal and transportable

3.2. Selection Bias vs Confounding.

- key differences
 - (1) What randomization **does** do
 - (a) protect against confounding
 - (b) protect against selection bias when treatment assignment occurs *AFTER* selection
 - (c) thus figures 7.27 and 7.28 could not represent *volunteer bias* in a randomized trial
 - figure 7.27 because in RCT treatment cannot determine selection
 - figure 7.28 because in RCT treatment is determined solely by the investigator
 - (2) What randomization **does not** do
 - (a) does not protect against selection bias when treatment occurs *BEFORE* selection
 - (b) does not protect against differential loss to follow up
- a confounder is therefore any variable L on which one has to adjust for to remove bias due to a common cause *or* common effect
- understanding the structure of the problem can help guide:
 - (1) the choice of study design
 - (2) the choice of analytical method
- selection on pre-treatment variables can introduce bias (i.e. *see figure 7.28*). This explains why some variables act as confounders in one study and not in another

3.3. The Structure of Selection Bias. The square around C indicates that the *analysis is restricted to subjects with $C = 0$* by study design or choice of analysis

- (1) **Conditioning on a common effect of treatment A and outcome Y**
 - *see figure 7.26*

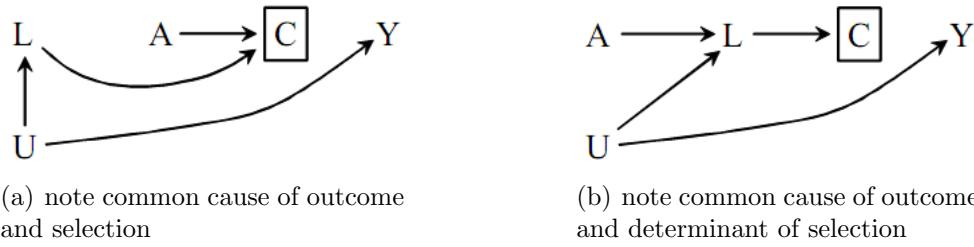


FIGURE 7.27. Selection Bias in a Randomized Trial

- association between A and Y is *induced* by conditioning on C

(a) conditioning on C opens the path $A \rightarrow C \leftarrow Y$

$$(b) \underbrace{\frac{Pr[Y=1|A=1, C=0]}{Pr[Y=1|A=0, C=0]}}_{\text{associational}} \neq \underbrace{\frac{Pr[Y^{a=1}]}{Pr[Y^{a=0}]}}_{\text{causal}}$$

(c) Example

- analysis of only survivors *see figure 7.26(a)*
 - selection into study or analysis depends on a variable affected by survival S *see figure 7.26(b)*

(2) *Conditioning on a common effect of treatment A and a variable L that is also a cause of the outcome*

- see figure 7.27
 - association between A and Y is *induced* by conditioning on C

(a) conditioning on C opens the path $A \rightarrow C \leftarrow U \rightarrow Y$

$$(b) \underbrace{\frac{Pr[Y=1|A=1, C=0]}{Pr[Y=1|A=0, C=0]}}_{\text{associational}} \neq \underbrace{\frac{Pr[Y^{a=1}]}{Pr[Y^{a=0}]}}_{\text{causal}}$$

(c) Example

- loss to follow up due to disease severity U and side-effects A . If L is measured could be used to block opened path *see figure 7.27(a)*
 - *prior* treatment A and disease severity have direct effects on symptoms L *see figure 7.27(b)*

(3) *Conditioning on a common cause of some trait that is associated with treatment and a variable L that is also a cause of the outcome*

- see figure 7.28
 - association between A and Y is *induced* by conditioning on C
 - (a) conditioning on C opens the path $A \leftarrow W \rightarrow C \leftarrow L \leftarrow$

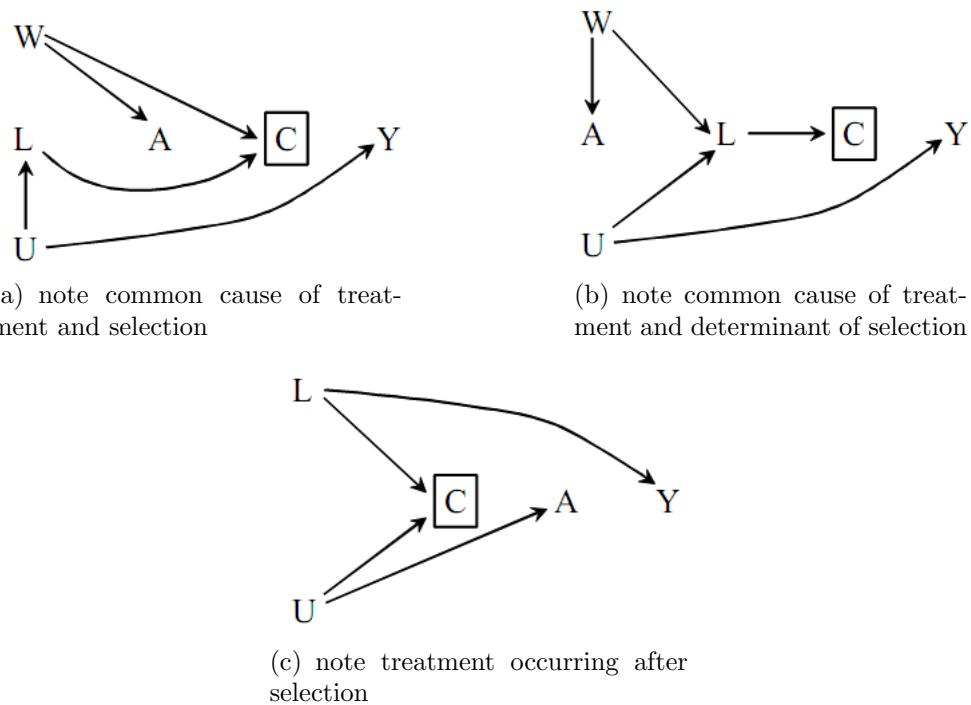


FIGURE 7.28. Selection bias in a *Non-Randomized* Trial

$$(b) \underbrace{\frac{Pr[Y=1|A=1, C=0]}{Pr[Y=1|A=0, C=0]}}_{\text{associational}} \neq \underbrace{\frac{Pr[Y^{a=1}]}{Pr[Y^{a=0}]}}_{\text{causal}}$$

(c) Example

- unmeasured variables that determine both treatment and attitudes towards participation or keeping study visit appointments *see figure 7.28(a)*
 - unmeasured variables that determine both treatment and threshold for reporting symptoms *see figure 7.28(b)*

3.4. Examples of Selection Bias.

3.4.1. *Differential loss to follow up.* Also known as informative censoring

- loss to follow up differs between the treated and untreated
 - diseased have a different rate of loss to follow up

3.4.2. Missing data bias & Non-response bias.

- missingness differs between the treated and untreated (perhaps more data collected on treated persons in observational study, or subjects receiving unblinded intervention more willing to disclose data)

- missingness also related to outcome (perhaps a greater willingness to disclose data among diseased subjects).

3.4.3. *Healthy worker bias.*

- underlying true health status is determinant of outcome and being at work
- study restricted to individuals at work at time of outcome ascertainment
- treatment reduces probability of being at work in the near future (directly or through common cause)

3.4.4. *Self-selection bias & Volunteer bias.*

- study restricted to those who agreed to participate
- can only occur when selection precedes treatment
- thus cannot occur in randomized trials

3.4.5. *Selection affected by treatment received before study entry.*

- treatment occurs before start of study
- or*
- there is a pre-study component

3.4.6. *Berksonian Bias.*

- treatment and disease are simultaneously associated with selection into an institution
- the joint distribution of treatment and disease in the institution differs from the source population
 - persons in institution more likely to have multiple diseases (or determinants) for selection into institution
 - sampling a control from a reference group (such as an “unrelated disease”) is likely to induce an association between treatment and outcome
- can happen even when selection frequencies for treatment and disease are equal

3.4.7. *Survivor Bias.*

- can be thought of as selecting on a consequence of the outcome (e.g. survival)
- for example, if outcome is cancer and only study those who are 60 years old
 - the treatment may have had such a strong effect that most of those who were exposed to it died
 - treatment would look beneficial after this point (e.g. crossing hazards)

3.4.8. Incidence-Prevalence Bias.

- occurs when sampling prevalent cases in a case control study, either at the end or during the course of follow up
- occurs when exposure affects disease duration
- when this happens exposure is associated with the outcome, resulting in selection bias

3.5. Selection Bias in Case-Control Studies.

- the arrow $Y \rightarrow C$ is present by design (cases are oversampled)
- the concern is if the arrow $A \rightarrow C$ or the structure $A \leftarrow L \rightarrow C$ is present
 - (1) ask yourself...has the selection process resulted in a different treatment distribution among cases and controls than exists in the source population?
 - (2) if yes, then selection is related to treatment and \rightarrow selection bias is present
 - (3) if there is selection into the institution where your study is based, the following are possible solutions:
 - (a) if the structure $A \leftarrow L \rightarrow C$ is present could collect data on L
 - (b) sample controls from a disease that is unrelated to your treatment
 - (4) want controls to be sampled independently of treatment

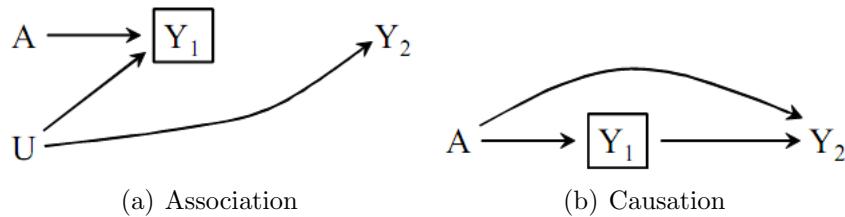


FIGURE 7.29. The selection bias from hazard ratios can render the causal effect unidentifiable - Which one is the true DAG?

3.6. The Built-in Selection Bias of Hazard Ratios.

- see figure 7.29
- in *discrete time*, the **hazard ratio** is defined as the probability of dying at time $T = t$ given that you have survived up to time t

$$IRR_{A,Y_t|Y_{t-1}} = \frac{Pr[Y_t = 1|A = 1, Y_{t-1} = 0]}{Pr[Y_t = 1|A = 0, Y_{t-1} = 0]}$$

- $IRR_{A,Y_t|Y_{t-1}}$ is the apparent hazard ratio and may not be causal
- by definition, the hazard ratio conditions upon survivors

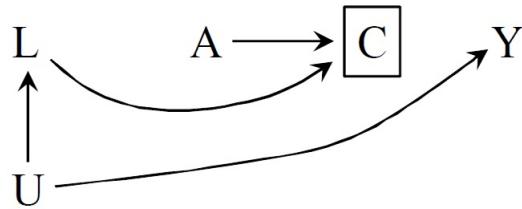


FIGURE 7.30. Identifiability and analytical control of censoring by treating it as an intervention

- the bias arises from conditioning on the common effect Y_{t-1} of treatment A and an unmeasured confounder U , where U is also a *cause* of Y_t
- this opens the path $A \rightarrow Y_{t-1} \leftarrow U \rightarrow Y_t$
- an unmeasured cause of death that is marginally unassociated with treatment such as U is often referred to as **frailty**
- this bias can occur in both randomized and observational studies

3.7. Identifiability of Causal Effects.

- censoring C can result in selection bias when there is differential censoring among treatment groups (Figure 7.30)
- censoring can be viewed as another “treatment” where measurement error and confounding need to be eliminated
 - causal question of interest: $Pr[Y^{a=1} = 1]$ vs $Pr[Y^{a=0} = 1]$
 - * what we are interested in is the results *had no one been censored*
 - $Pr[Y^{a=1,C=c}]$ vs $Pr[Y^{a=0,C=c}]$
 - * with censoring, information is only available for the subgroup $C = 0$
 - * only investigating $C = 0$ subset results in biased results due to selection bias
 - * A and C can be seen as joint interventions, making it important that exchangeability, positivity, and consistency hold for both A and C
 - * selection bias can be eliminated by analytic adjustment for confounding of “treatment” C
 - * selection bias does not hold for competing risks

3.8. Adjusting for Selection Bias.

- selection bias can sometimes be avoided by adequate design, however it is often unavoidable

- loss to follow-up, self-selection, and missing data bias can occur no matter how careful the investigator
 - selection bias needs to be corrected in the analysis
 - IPW (or standardization) can be used to correct selection bias
 - procedure
 - * formally, we modify IPW to weight by the inverse probability of receiving the *joint treatment* of A and $C = 0$ conditional upon variables L needed to block the back door path between
 - (1) A and Y
 - (2) C and Y
 - * we estimate the censoring and treatment parts of the probability of the joint treatment $\{C = 0, A\}$ separately and use them to construct a single inverse probability weight
- $$\underbrace{\frac{1}{Pr[C = 0, A = a|L]}}_{\text{joint treatment weight}} = \underbrace{\frac{1}{Pr[C = 0|A, L]}}_{\text{censoring weight}} \times \underbrace{\frac{1}{Pr[A|L]}}_{\text{treatment weight}}$$
- * assign the weight to each selected subject ($C = 0$) to account for themselves and other non-selected ($C = 1$) subjects with the same distribution of L and A
 - * the goal is to make a pseudo-population the same size as the original population where censoring C is independent of Y within levels of L : $Y \perp\!\!\!\perp C|L$
 - the association measure in the pseudo-population equals the effect measure in the original population if:

- (1) exchangeability: $E[Y|C = 0, L = l, A = a] = \overbrace{E[Y|C = 1, L = l, A = a]}^{\text{not observed}}$: the average outcome in $C = 0$ must equal the unobserved average outcome in $C = 1$ with the same values of A and L
 - * the probability of selection used in weighting is conditional upon
 - (a) treatment
 - (b) all factors that independently predict both selection and outcome
 - * if this holds the variables included in $Pr[C = 0|A, L]$ are sufficient to block all backdoor paths *between* C and Y
- (2) $Pr[C = 0|L = l] > 0$ for all l : all conditional probabilities of being uncensored given L must be greater than zero
 - * positivity assumption required for $C = 0$, but not for $C = 1$

- * not interested in inferring what would have happened if at least one subjects had been censored in each strata
- (3) consistency: the effect of treatment A is the same in the pseudo-population as in the original population if nobody had been censored
 - * interpretation is valid when censoring the result of loss to follow-up or nonresponse
 - * interpretation is questionable in the presence of competing events
 - could argue that stratification by level of L (rather than IPW or standardization) would be sufficient to remove selection bias
 - * care should be taken to ensure colliders are not opened

3.9. Selection without Bias.

- conditioning on a collider *always* induces an association between its common causes
 - the association may be restricted to only *certain* levels of the common effect
 - it is possible that selection on a common effect does not result in selection bias when the analysis is restricted to a single level of the common effect
 - * example: 3 cause-specific mortality variables (1) death from tumor Y_A , (2) death from heart attach Y_E , and (3) death from other causes Y_O (Figure 7.31)
 - * the path between surgery A and haplotype E though opened by conditioning on collider Y ($Y = 0$) is blocked by conditioning on the noncolliders Y_A , Y_E , and Y_O ($Y_A = Y_E = Y_O = 0$)
 - * A and E affect survival through independent mechanisms
 - * multiplicative survival model: when the conditional probability of survival given A and E ($Pr[Y = 0|A = a, E = e]$) is the product of functions a and e ($g(a) \times h(e)$)

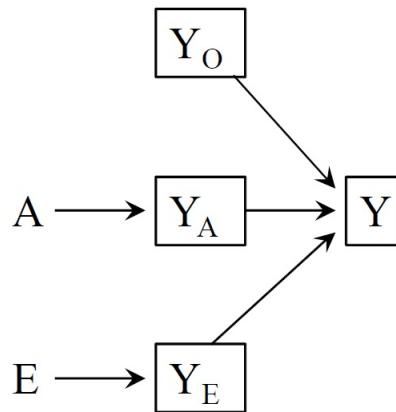


FIGURE 7.31. Selection without bias. When $Y = 0$, A and E are conditionally independent

- in practice it is important to consider the expected direction and magnitude of selection bias
 - direction
 - * depends on how the two causes A and E interact to cause Y
 - factor U is associated with A or E , A and E are negatively associated
 - factor U is associated with A and E , A and E are positively associated
 - * DAGs fail to distinguish between the *and* and *or* mechanisms
 - magnitude
 - * a large selection bias requires strong associations between the collider and both treatment and outcome

4. Information Bias

4.1. Measurement Error.

- synonymous for “misclassification” for discrete variable
- graphical representation:

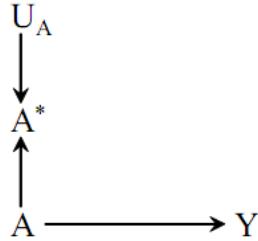


FIGURE 7.32. A : true A ; A^* : measured A ; U_A : a measurement error (All variables associated with A^* other than A)

- in the presence of measurement bias, the identifiability conditions exchangeability, positivity, and consistency are insufficient to compute the causal effect of treatment A and outcome Y . On risk ratio scale:

$$Pr[Y^* = 1|A^* = 1]/Pr[Y^* = 1|A^* = 0] \neq Pr[Y^{a=1} = 1]/Pr[Y^{a=0} = 1]$$

- * in general, methods for measurement error rely on a combination of modeling assumptions and validation samples
 - **validation samples** are subsets of the study sample in which key variables are measured with little or no error.

4.2. The Structure of Measurement Error.

- The structure of measurement error is classified according to two properties:

(1) Independence:

- the measurement error of variable A (i.e. U_A) is not associated with the *measurement error* of another variable Y (i.e. U_Y)
- Independence of U_A and $U_Y \implies f(U_A, U_Y) = f(U_A) \times f(U_Y)$

(2) Non-differentiability:

- the measurement error of variable A (i.e. U_A) is not associated with the *value* of another variable Y (i.e. $Y = y$)
- $f(U_A|Y) = f(U_A) \text{ or } f(U_Y|A) = f(U_Y)$

- Independence and non-differentiability produce distinct structures of measurement error:

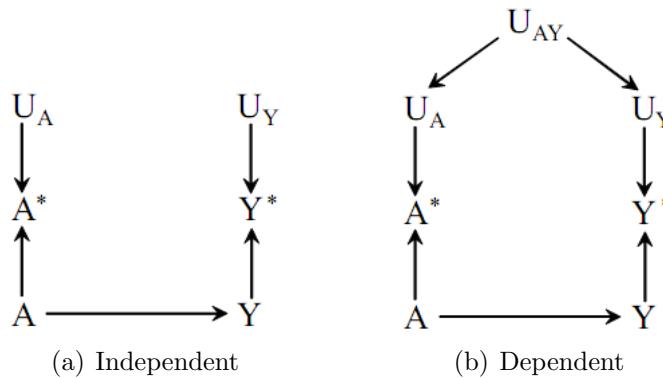


FIGURE 7.33. Non-differential measurement error

(1) Independent nondifferential

see figure 7.33(a) for DAG

- measurement errors of exposure and outcome are *not* associated and do *not* depend on the value of the other variable
- Y^* : measured Y ; U_Y : all variables associated with Y^* other than Y
- e.g. Data entry error for A^* and Y^*

(2) Dependent nondifferential

see figure 7.33(b) for DAG

- the measurement errors of exposure and outcome *are* associated but do *not* depend on the value of the other variable
- e.g. Recall error for A^* and Y^* , U_{AY} : ability to recall

(3) Independent differential

see figure 7.34(b) for DAG where outcome affects measurement of exposure

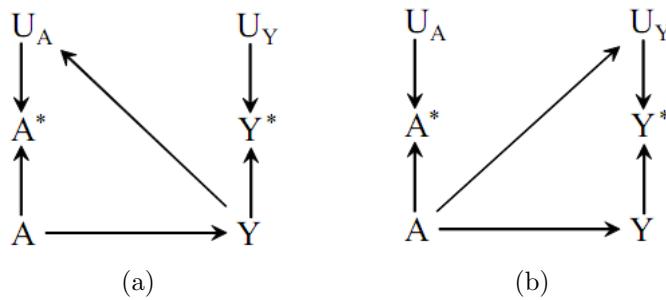


FIGURE 7.34. Independent differential measurement error

- measurement errors of exposure and outcome are *not* associated but *do* depend on the value of the other variable
- e.g. Recall bias: Study of the drug use in gestation period (A) on congenital malformation (Y), drug use ascertained by recall (U_A). Arrow: Outcome Y affects U_A
- e.g. Reverse causation bias: Study of liver toxicity (Y) related to drug (A), drug level (A^*) measured retrospectively, which is affected by the liver metabolism (U_A), liver metabolism was impaired due to liver toxicity ($Y \rightarrow U_A$)

see figure 7.34(b) for DAG

- e.g. Detection bias: the physician prescribed a new drug to patients and ordered more monitoring tests ($A \rightarrow U_Y$), resulting in higher frequencies of liver toxicity in the treatment group ($A \rightarrow Y$)
- e.g. Detection bias: estrogen and endometrial cancer
- e.g. Detection bias: mammography screening and breast cancer

(4) Dependent differential:

see figure 7.35

- measurement errors of exposure and outcome *are* associated and *do* depend on the value of the other variable

4.3. Mismeasured confounders.

see figure 7.36(a)

- e.g. Study of the drug therapy and liver toxicity, past history of hepatitis was a potential confounder (L) and measured by recall (L^*)
 - Mismeasured confounder can be viewed as surrogate confounder
 - Controlling for mismeasured confounder does not completely block the backdoor path (residual confounding), measurement bias or information bias

see figure 7.36(b)

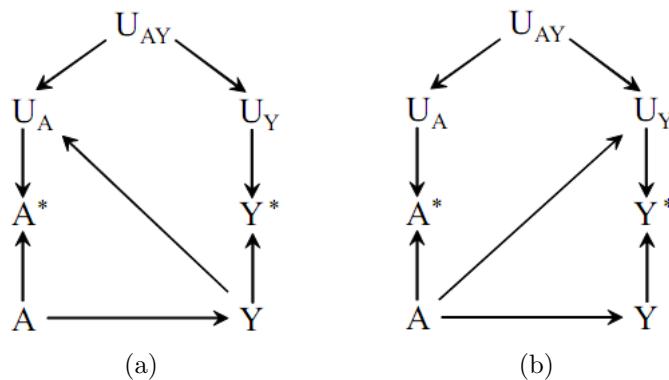


FIGURE 7.35. Dependent differential measurement errors

- e.g. L has an indirect relationship between A and Y through U , controlling for L^* does not block the backdoor path
 - Mismeasurement of confounders can even lead to what looks like effect modification
see figure 7.36(c)
- conditioning on a mismeasured common effect (C^*) still opens the backdoor path

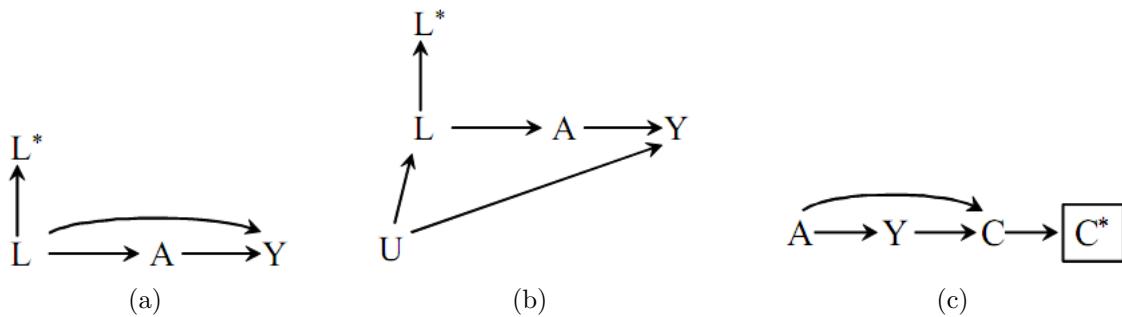


FIGURE 7.36. Mismeasured adjustment variables

4.4. Strength and direction of measurement bias.

- If A - Y association is null
 - only *independent and nondifferential* measurement error will not affect the results ($A^* - Y^*$ association is null)
 - Other than this circumstance, the measurement error may lead to bias either away from or close to the null
- For non-dichotomous treatment, measurement error may result in $A - Y$ and $A^* - Y^*$ association in opposite direction, even under independent and nondifferential measurement error, as long as $A^* - A$ association is non-monotonic.

- The magnitude of measurement bias depends on the measurement error. Measurement bias generally increase as the strength of $U_A \rightarrow A^*$ or $U_Y \rightarrow Y^*$ increase

4.5. Types of Misclassification and Expected Bias.

- Examples
 - *cumulative exposure* that include time outside of the induction period
 - *grouped exposures*, some of which do not contain the etiological agent
- Definitions
 - *Non-differential*: both Sn and Sp are the same in group A=1 and group A=0
 - *Differential*: either Sn or Sp varies by group status
 - *Dependent*: the error is associated with the *measurement error* of another variable

4.5.1. Non-Differential Misclassification.

- is “random” misclassification
- is expected to bias the *average value* of the effect measure (i.e. from repeated studies) to the null under the following conditions:
 - (1) the variable is dichotomous
 - (2) the misclassification is exactly non-differential
 - (3) misclassification errors are independent of measurement errors in other variables in the analysis
 - (4) misclassification errors are not structurally related to other biases (confounding, collider-bias, etc.)
 - even when these hold, in a given study with misclassification the effect measure may be *further away* from the null due to random variation
- **can also affect statistical inference:**
 - (1) reduce power of hypothesis tests
 - (2) incorrectly centered and narrower confidence intervals
 - (3) sample size calculations that overestimate power
 - (4) however the test of the null hypothesis is still valid (ie.e correct Type I error rate)

4.5.2. Differential misclassification.

- Examples
 - better (or more frequent) measurement applied to one group vs. the other

- differences in specimen handling, storage, assay, or time until analysis (e.g. deterioration)
- recall bias
- categorization of continuous variables can change nondifferential to differential error
- may bias towards or away from the null

4.5.3. Recall Bias.

- A form of differential measurement error/misclassification
- knowledge or feelings about the outcome affects how exposure is *recalled* (i.e. retrospectively measured)
- is expected to result in bias either away or towards the null
- Some things that may affect recall:
 - time since exposure can affect memory
 - parental guilt can affect recall of child or fetus exposure
- a possibility in any case-control study that relies on subject memory

4.6. Correcting for Misclassification of Dichotomous Variable.

(1) obtain sensitivity(θ) and specificity(ϕ)

- usually by performing a validation study in a randomly chosen subset of your sample
- in this sample exposure is measured with little or no misclassification

(2) obtain expected cell counts with misclassification:

		True Exposure		Misclassified Exposure	
		$A = 1$	$A = 0$	$A^* = 1$	$A^* = 0$
Disease	$Y = 1$	a	b	a^*	b^*
	$Y = 0$	c	d	c^*	d^*

$m_1 \xrightarrow{\theta, \phi} m_1$

TABLE 7.3. True and Misclassified Exposure

$$\begin{aligned}
 a^* &= \theta a + (1 - \phi)b \\
 b^* &= (1 - \theta)a + \phi b \\
 c^* &= \theta c + (1 - \phi)d \\
 d^* &= (1 - \theta)c + \phi d
 \end{aligned}$$

(3) calculate expected true cell counts without misclassification (see table 7.3)

$$a = \frac{\phi m_1 - b^*}{\theta + \phi - 1} \quad b = \frac{\theta m_1 - a^*}{\theta + \phi - 1}$$

$$c = \frac{\phi m_0 - d^*}{\theta + \phi - 1} \quad d = \frac{\theta m_0 - c^*}{\theta + \phi - 1}$$

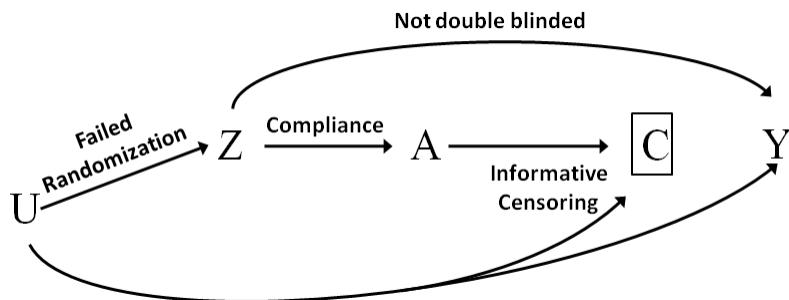


FIGURE 7.37. Characteristics of a simple randomized trial. U: Unmeasured patient characteristics, Z: Treatment assignment, A: Treatment received, C: Censoring Indicator, Y: Outcome

4.7. Noncompliance.

- in real randomized experiments, participants may not comply with the assigned treatment
- distinguish *assigned* treatment (Z) and *received* treatment (A)
- Z is a misclassified version of A
- noncompliance (of Z) is a special case of *treatment* misclassification
 - in general, A^* has NO causal effect on Y (refers to figure 7.32)
 - however, Z has causal effect on Y (see figure 7.38)
 - (1) $Z \rightarrow A \rightarrow Y$
 - (2) $Z \rightarrow Y$ (not through A , e.g. change life style)
 - causal effect of Z on Y combined both effects
 - how to eliminate $Z \rightarrow Y$, i.e. only effects of Z on Y are $Z \rightarrow A \rightarrow Y$
 - * *double-blind placebo-controlled randomized experiment*
 - no $Z \rightarrow Y$: *exclusion restriction* holds
 - * $Y^{z=0,a} = Y^{z=1,a}$ for all subjects and all values a

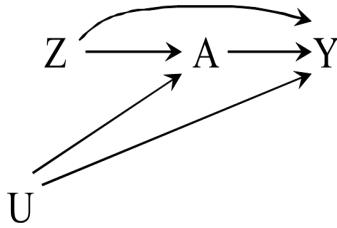


FIGURE 7.38. DAG showing noncompliance. Z is assigned treatment, A is received treatment.

4.8. Intention-to-treat Effect.

- as-treated effect (“efficacy”) **used for adverse effect**

see figure 7.38

- *nonrandom noncompliance* results in $Y^a \perp\!\!\!\perp A$
- the reason why participants received treatment ($A = 1$) is not random but rather associated with participants’ prognosis U
- A is associated with U ($U \rightarrow A$)
- $A \leftarrow U \rightarrow Y$
- confounding for the effect of A on Y
- does not have a causal interpretation unless adjust for confounding (by g-methods or instrumental variable estimation)
- intention-to-treat effect (“effectiveness”) **used for treatment effect**
 - $Y^z \perp\!\!\!\perp Z$ even when there is nonrandom noncompliance
 - “the effect of assigning participants to being treated with A ”
 - $\frac{Pr[Y=1|Z=1]}{Pr[Y=1|Z=0]} = \frac{Pr[Y^z=1]}{Pr[Y^z=0]}$
 - the magnitude of ITT effect depends on:
 - (1) the effect of A on Y ($A \rightarrow Y$)
 - (2) the effect of Z on A ($Z \rightarrow A$)
 - (3) the direct effect of Z on Y that is not mediated by A ($Z \rightarrow Y$) (null in double-blind randomized experiment)
 - ITT effect can only be computed when there is no loss to follow-up or any censoring
 - * (alternative) *pseudo-ITT effect*: $\frac{Pr[Y=1|Z=1,C=0]}{Pr[Y=1|Z=0,C=0]}$
- reasons to use ITT analysis

- (1) ITT effect is closer to the null than the effect of A on Y in a double-blind placebo-controlled randomized experiment
 - ITT effect is the lower bound of the effect of A on Y (conservative)
- (2) **null preservation:** if the sharp causal null hypothesis holds for A , the ITT average effect would also be null
- reporting only the ITT effect implies preference for misclassification bias over confounding, a preference needs to be justified in each application
- reporting ITT effect as primary findings from a randomized experiment is hard to justify for
 - (1) experiments that are not double-blind placebo-controlled
 - (2) study for the effect of a treatment *safety*
- caveats for ITT analysis
 - (1) not good for drug safety study (too conservative)
 - (2) not conservative when both treatments are “active” (no placebo was used in the study)
 - (3) the ITT effect measures the effect of assigned treatment under the adherence conditions observed in a particular experiment
 - (4) the above argument implies that we should refrain from conducting double-blind randomized clinical trial because, in real life, both doctors and patients are aware of the received treatment

	Definition	Example	Remedy
Confounding bias	Spurious association between exposure and outcome caused by common cause	Spurious association between match carrying and lung cancer	Stratification, restriction, matching, and regression
Confounding by indication	Spurious association between treatment and outcome caused by physician's judgment on patient's indication	Spurious causative effect of beta-agonist use and asthma death	Stratification, restriction, matching, and regression on potential confounders, propensity score analysis, instrumental variable analysis
Confounding by adherence/non-adherence	Adherence of certain drug may decrease/increase in symptomatic stage before diagnosis and cause spurious association between drug and disease	Spurious preventive effect of statin on cancer	Stratification, restriction, matching, and regression on adherence
Healthy user (worker) bias	The exposure is associated with individual's health status and cause a spurious association between exposure and disease	HRT is protective for CHD in the nurse health study, they are healthier	Stratification, restriction, matching, and regression on indicators of healthy status SMR

FIGURE 7.39. Summary of bias definition, example, and remedy

	Definition	Example	Remedy
Selection bias (collider stratification bias)	The selection of control or case cause the distortion of exposure- outcome relationship	See following examples	Prospective follow-up study design, G-methods to correct differential censoring
Healthy worker bias	Distortion of exposure-outcome relationship due to restricting study subjects on those remained on work common effect	Workers with occupational exposure paradoxically exhibit lower overall death rates than general population, because the ill are excluded from employment	Prospective follow-up study, collect information on censoring, correct informative censoring by IPW [[[Healthy worker effect can be viewed as confounding or selection bias]]]
Volunteer bias/ Self-selection bias	Distortion of exposure-outcome relationship due to restricting study subjects on those who volunteered to participate [[selection after exposure already began]]	Individuals with family history of CHD are more likely to volunteer to participate in study that examine the smoking-CHD relationship, and cause bias in either direction	Random sampling of exposed and nonexposed population, collect information to model self-selection process and corrected with IPW [[[Volunteer does not causes bias if all selection process before exposure/treatment start, it's only cause problem of lack of generalizability]]]
Prevalence-incidence bias, Survivor bias	Distortion of exposure-outcome association due to use prevalent cases rather than incident cases	Better survival increase disease prevalence	Exclude prevalent cases for analysis
Trohoc bias	Spurious association between exposure and disease due to selection of control that mimic case characteristics by excluding	Excluding controls with polyps in the smoking-colon ca study may cause downward bias	Selection control with comparability of source population in which case arose [[[Comparison: matched COHORT study seek comparability between EXPOSED and

FIGURE 7.40. Summary of bias definition, example, and remedy

	INTERMEDIATES	UNEXPOSED]]]
Berkson bias	Spurious association between exposure and disease due to different admission rate of case and control in the hospital based study	Spurious association between diabetes and cholecystitis use "refractive error" patients as control Sample control disease that have similar referral patterns for case disease, but admission is unrelated to exposure of interest Use several diseases to lessen the chance of an association with exposure
Length time bias	Bias due to selection of cases with longer survival	Spurious benefit of screening on cancer survival as screening tend to select indolent cases of cancer, patients with short survival period are less likely to be screened Multiple longitudinal screening
Left truncation bias	Bias due to differential left censoring in the exposed and unexposed, censoring distort the exposure disease association	Spurious association between use of spermicide and decreased fertility because hyperfertile couples were left truncated New user cohort design, exclude prevent user, follow up new user and count every incident outcome, use inception cohort rather than survivor cohort
Right truncation bias (informative censoring), non-respondent bias	Bias due to differential right censoring in the exposed and unexposed, censoring distort the exposure disease association	In randomized control trial, treatment groups have more loss of follow-up because of side effect, recovery, or death Collect information on the censoring mechanism and corrected with IPW

FIGURE 7.41. Summary of bias definition, example, and remedy

	Definition	Example	Remedy
Information bias	Measurement error of treatment, outcome, or confounders	See following examples	Subsample validation study, regression calibration, sensitivity analysis
Non-differential misclassification	Random misclassification of exposure or outcome, in the case of dichotomous variable, random misclassification will bias toward the null on average	Random misclassification of actual drug intake based on dispensing record	Subsample validation study and do sensitivity analysis to assess the magnitude and direction of possible bias
Detection bias	Differential outcome misclassification dependent on the exposure status	New drug users have higher incidence of impaired liver function due to more frequent testing	Stratifying on the frequency of tests
Recall bias / Reporting bias	Differential exposure misclassification dependent on the outcome	Mother of infant with congenital malformation recall/report more drug exposure	Study in a prospectively cohort with collection of exposure and confounder information independent of outcome
Reverse causation bias	Outcome precedes and has a causal effect on measurement of treatment	Estrogen used for treatment of uterine bleeding and cause a spurious association between estrogen and endometrial cancer	Lag analysis
Protopathic bias	Bias caused by prescribing pharmaceutical agent for an early manifestation of a disease that has not yet been diagnostically detected	Spurious association between use of antacid and risk of gastric cancer, because early symptoms of gastric cancer is dyspepsia	Lag analysis
Induction period bias	Nondifferential exposure misclassification due to unknown induction period, bias toward the null on average if the exposure is dichotomous	In the study of coffee drinking and MI, the induction period is unknown and misclassification of exposure person-time will biased the results toward the null	Induction period analysis

FIGURE 7.42. Summary of bias definition, example, and remedy

CHAPTER 8

Advanced Topics in Causal Inference

1. Directed Acyclic Graphs: theory

1.1. definition of DAG.

- use DAG to conceptualize problem
- use counterfactual-based methods to analyze data
 - (1) graphs have nodes with directed edges and no directed cycles
 - (2) nodes: random variables $V = (V_1, \dots, V_M)$
 - (3) PA_m are the *parents* of V_m : the *set* of nodes with arrows into V_m
 - (4) V_j is a *descendant* of V_m if there is a sequence of nodes connected by edges between V_j and V_m

1.2. definition of *causal* DAG.

- causal DAG is a specific type of DAG
- a causal DAG is a DAG with following features, which give DAG a causal interpretation:
 - (1) lack of arrow means “no direct causal effect” with respect to other variables in the DAG
 - (2) *all* common causes are on the DAG (even unmeasured)
- causal DAG is agnostic to counterfactuals

1.3. assumption to link DAGs to data: causal Markov assumption.

- we need an assumption to link **causal DAG** to **statistical data**; otherwise, a causal DAG is just a graph of no practical use
- to do so, we first need to define **DAGs** (all DAGs, not only causal DAGs) in terms of mathematical form
- **definition: a DAG represents the joint probability of all variables in the DAG ($f(\nu)$) if and only if $f(\nu)$ satisfies the Markov factorization**

$$f(v) = \prod_{j=1}^M f(v_j | pa_j)$$

$f(\nu)$ can be factorized as the product of the probability of each variable given its parents

= the non-descendants of a given variable is independent of that variable conditional on the variable's parents

- Markov factorization implies additional statistical independence:

conditional independence

$Y \perp\!\!\!\perp A | L$ if Y and A are **d-separated** given L in the DAG

marginal independence

$Y \perp\!\!\!\perp A$ if Y and A are **d-separated** without conditioning on any other variables in the DAG

- d-separation and d-connection

– d-separation

(1) d-separation means no open path between A and Y in the DAG

(2) d-separation between A and Y implies *independence* between A and Y

marginal independence: d-separated *without* conditioning on any L

conditional independence: d-separated *after* conditioning on L

(3) “**graphical rules to block a path**”

(a) a *collider* on the path \rightarrow block

(b) a *non-collider* on the path is *conditioned on* \rightarrow block

– d-connection

(1) there is open path between A and Y

(2) d-connection between A and Y does **NOT** always imply *dependence* between A and Y

(3) need assumption of **faithfulness** to say that two variables are dependent if they are d-connected

(4) “**graphical rules to open a path**”

(a) a *collider* on the path is *conditioned on* \rightarrow open

(b) a *non-collider* on the path \rightarrow open

- Now, if I want to link my causal DAG represents to my data, what should I do...

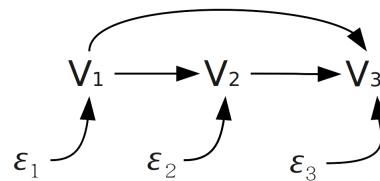
causal Markov assumption

The joint distribution of variables in my causal DAG satisfies the Markov factorization

1.4. non parametric structural equation model (NPSEM).

- NPSEM is a *causal DAG model* that includes counterfactuals random variables (e.g. $Y^{a=1}, Y^{a=0}, \text{etc}$)
- NPSEM combined both counterfactual theory and DAG
- notations of NPSEM
 - W : random variable
 - ω : the set of possible values of W
 - $\varpi = (\omega_1, \dots, \omega_m)$
 - R is any subset of V ; r is a value of R
 - V_m^r is a counterfactual value of V_m
- for a NPSEM presented by a DAG G with V
 - **model assumption**
 - (1) ϵ_m is the unobserved **mutually independent** random error
 - (2) $f_m = (pa_m, \epsilon_m)$ (deterministic unknown function)
 - * $V_1 = f_{\epsilon_1}$
 - * $V_m^{\bar{v}_{m-1}} \equiv V_m^{pa_m}$
 - * (both given by the function)
 - **model**
 - * V_m and V_m^r , for $m > 1$ can both be obtained recursively
 - * e.g. $V_3^{v_1} = V_3^{v_1, V_2^{v_1}}$ (counterfactual)
 - * e.g. $V_3^3 = V_3^{V_1, V_2^{V_1}}$

FIGURE 8.1. a DAG represents NPSEM



- the model assumption of NPSEM (ϵ_m) can be translated into model assumption of causal DAG (causal Markov assumption)
 - ϵ_m is the unobserved **mutually independent** random error which
 - (1) implies causal Markov assumption holds

(2) is equivalent to that all common causes are on the causal DAG

- causal DAG (DAG is agnostic to counterfactual) > NPSEM (incooperates counterfactual)
 - fully randomized causally interpreted structured tree graph (weaker assumption) > NPSEM (stronger assumption)

1.5. identifying condition (counterfactual) in causal DAG.

positivity: $L \rightarrow A$; the arrow is not deterministic

consistency: $A \rightarrow Y$; the arrow is well-defined

exchangeability: the lack of paths between A and Y , other than those originating from A , that would result in an association between A and Y (figure 8.2)

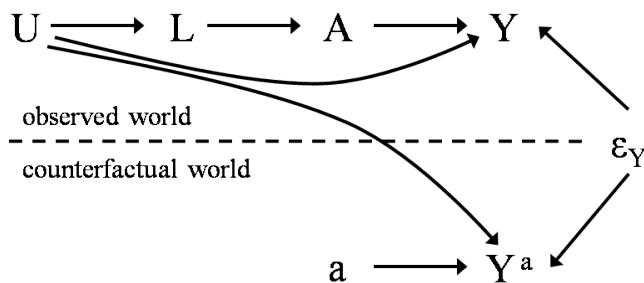


FIGURE 8.2. DAG integrating the counterfactual world $A = a$

2. Direct and Indirect Effects

2.1. Non-time varying exposure.

- in the presence of exchangeability we can calculate the net effect of exposure A on Y
 - to separate direct (not through intermediate B) and indirect effects the assumption of partial exchangeability with respect to B is required for each exposure stratum and the assumption of no interaction (figure 36)

$E[Y^{a=1,b=0}|B = 1] = E[Y^{a=1}|B = 0]$ and $E[Y^{a=0,b=0}|B = 1] = E[Y^{a=0}|B = 0]$

 - in a randomized trial to be able to separate direct and indirect effects both the exposure and an intervention on the intermediate have to be randomized
 - if **no interaction** between A and B then sum of the direct and indirect effects should equal the total effect

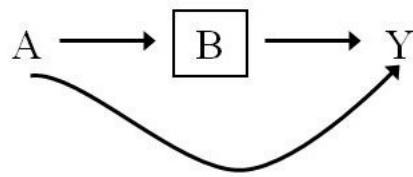


FIGURE 8.3. Causal pathway of exposure A to outcome Y with intermediate B

- in the presence of covariates that affect the intermediate and the outcome regardless of association with exposure partial exchangeability does not hold
 - *Note:* confounding of the direct effect may be present even if the total effect is unbiased
- if the covariates are not affected by exposure and data are available then the direct and indirect effects can be estimated by using conventional methods such as stratification (figure 37)

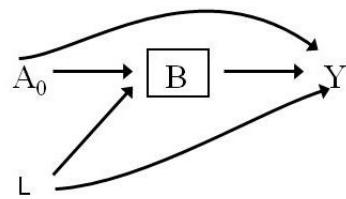


FIGURE 8.4. Causal pathway of exposure A to outcome Y with intermediate B confounded by L

- if the covariates are affected by exposure or data are unavailable, G-computation methods must be used (figure 38)

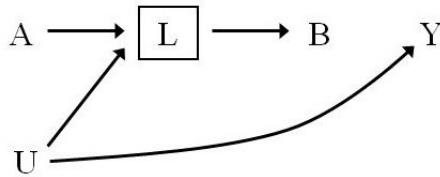


FIGURE 8.5. Causal pathway of exposure A to outcome Y with intermediate B confounded by L and U

- in the presence of interaction between exposure and intermediate the direct and indirect effects will not equal the total effects



FIGURE 8.6. Without interaction only $B^{a=1}$ is involved in any effect of A(a). In the case of interaction B^a interacts with A to cause Y as part of the direct effect of A(b)

- depending on the interaction direct and indirect effects will be super or subadditive
- Indirect effects cannot be estimated but the effect that can be prevented by intervening on the intermediate can
in the absence of an interaction this and the indirect effects were the same

2.2. Time varying exposures.

- with time varying exposures the direct (controlled) effects of A_0 on Y are the ones not mediated through A_1
if we do not specify control for A_1 then effect of A_0 is total effect.
 - with dichotomous exposures there are two such exposures:
the one with $A_1 = 0$ where the effect is the counterfactual difference

$$E[Y_{\bar{a}=1,0}] - E[Y_{\bar{a}=0,0}] = E[Y_{\bar{a}=1,0} - Y_{\bar{a}=0,0}]$$
and the one with $A_1 = 1$ where the effect is the counterfactual difference

$$E[Y_{\bar{a}=1,1}] - E[Y_{\bar{a}=0,1}] = E[Y_{\bar{a}=1,1} - Y_{\bar{a}=0,1}]$$

Note: these effects are still joint effects of the time varying exposure $[A_0, A_1]$

- must use methods of causal inference for time varying exposures to calculate these direct effects (*g-methods*)

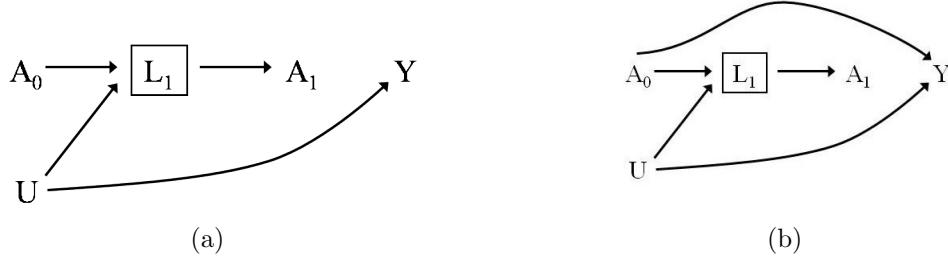


FIGURE 8.7. Time varying exposure: Direct effect of A could be partially through L_1 (b)

- we cannot determine whether any direct effect of A_0 goes through L_1 without determining causal effect of L_1 on Y
- direct effects in MSMs
 - consider the marginal structural model $E[Y_{g=(a_0, a_1)}] = \theta_0 + \theta_1 a_0 + \theta_2 a_1 + \theta_3 a_0 a_1$
 - no effect of A_0 on Y when $A_1 = 1$ then $\theta_1 + \theta_3 = 0$
 - no effect of A_0 on Y when $A_1 = 0$ then $\theta_1 = 0$
 - no effect of A_0 on Y without intervening on A_1 , then $\theta_1 = \theta_2 = \theta_3 = 0$
 θ_1 and θ_3 represent the direct effect of A_0 , while θ_2 represents the indirect effect
- direct effects in SNMs
 - consider the structural nested model
 - $$Y_{g=(a_0, a_1)} = Y_{g=(0,0)} + \beta_1 a_0 + \beta_2 a_1 + \beta_3 a_1 L_{1,g=a_0} + \beta_4 a_0 a_1 + \beta_5 a_0 a_1 L_{1,g=a_0}$$
 - no effect of A_0 on Y when $A_1 = 1$ if $\beta_1 + \beta_3 + \beta_4 + \beta_5 = 0$
 - no effect of A_0 on Y when $A_1 = 0$ if $\beta_1 = 0$
 - no effect of A_0 on Y without intervening on A_1 if all β s = 0
 β_1, β_{3-5} represent the direct effect of A_0 while β_2 represents the indirect effect

3. Static & Dynamic Observation Plans

Assumptions.

- (1) treatment $A_{i,t}$ can only change right after measurement of $L_{i,t}$
- (2) i.e. observation times are the only times of potential treatment change

3.1. Types of Observation Plans.

Static.

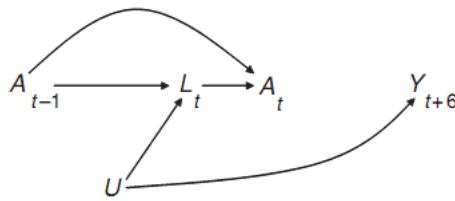
- observation plan is pre-specified
 - e.g. Nurses Health Study
- at regular intervals

Dynamic.

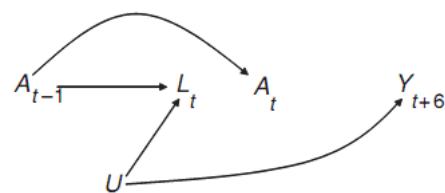
- observation plan varies across individuals
- may depend on covariate and / or treatment history
 - e.g. clinical history
 - e.g. patients in pain medication study only returning to clinic when symptomatic
- can occur in studies with static plans if individuals miss visits because of clinical history
- are generally expected in observational longitudinal studies

3.2. Static Observation Plans.

- if assigned at random at baseline, do not have to account for observation plan in analysis
- true whether one or many static plans



(a) Longitudinal study with a single static observation plan



(b) The pseudo-population with a static treatment regime

FIGURE 8.8. IPW without specifying observation plan

Inverse Probability Weighting in Static Regimes.

- we estimate numerator and denominator of SW_t^A parametrically

- only for times when treatment changes

$$SW_t^A = \prod_{k=0}^t \frac{f(A_k | \bar{A}_{k-1})}{f(A_k | \bar{A}_{k-1}, \bar{L}_k)}$$

- numerator and denominator = 1 (thus SW_t^A for times when treatment does not change)
- this estimator identifies the causal effect if the following assumptions hold
 - (1) *Consistency* $\rightarrow Y^{\bar{a}_t} = Y^{\bar{A}_t} = Y_t$ if $\bar{A}_{t-1} = \bar{a}_{t-1}$
 - (2) *Exchangeability* $\rightarrow Y^{\bar{a}_k} \perp\!\!\!\perp A_k | \bar{A}_{k-1}, \bar{L}_{k-1}$ for all \bar{a} and $t > k$
 - (3) *Positivity* $\rightarrow f(\bar{a}_{k-1}, \bar{L}_k) > 0 \implies f(a_k | \bar{a}_{k-1}, \bar{L}_k) > 0$

an implicit 4th assumption

- (4) $N_i, t = 0 \implies A_i, t = A_{i,t-1}$
 - $N_{i,t}$ is a time-varying indicator function
$$N_{i,t} = \begin{cases} 0 & \text{if subject } i \text{ not observed at time } t \\ 1 & \text{if subject } i \text{ observed at time } t \end{cases}$$
 - assumption that times of observation and treatment change coincide
 - only true if subjects have no access to treatment except at study visits and if subjects could not discontinue treatment on their own

When 4th assumption does not hold.

- bias can result
- from the fact that a covariate determined treatment time
 - usually unmeasured at the true time of change, thus acts as an unmeasured confounder
 - solution of "carrying last measurement forward"
 - * using $L^* = L_{t-1}$ as proxy for L does not remove bias
 - * IPW weights based on L^* do not create pseudo-population in which $L_t \rightarrow A - t$ is removed (as in *figure 8.8*)
 - the "coarser" the measurement plan, the larger the unmeasured confounding is expected to be

Single Static Observation Plan.

- when 4th assumption holds there is no $\rightarrow N_{i,t}$
 - also implicit: observation plan $N_{i,t}$ can only affect Y_{t+1} through treatment A_t

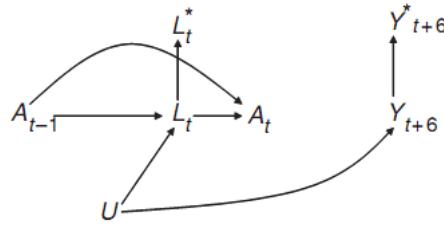


FIGURE 8.9. longitudinal study with a single static plan and a carried forward covariate L^*

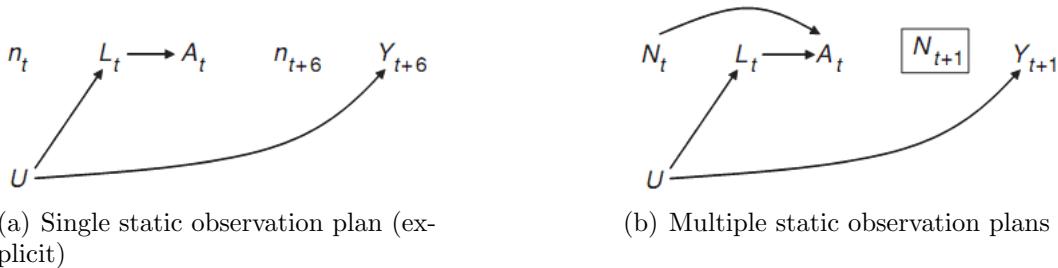


FIGURE 8.10

- thus in reality we always estimate the effect of the joint regime of two components
 - (1) observation plan \bar{n}
 - (2) treatment regime \bar{a}
- IPW with SW_t^A for static regime is implicitly conditional on the observation plan \bar{n}
- even though it is often ignored, we should **characterize causal effect according to observation plan**
 - for example, for a treatment with a narrow etiological window
 - ... finer observation plan more likely to show causal effect

Multiple Static Observation Plans.

- subjects randomly assigned to one of several observation plans
- analysis conditional on those observed at time $t + 1$ because they are the only ones with data on Y_{t+1}
 - thus conditioning on N_{t+1} is implicit in our analysis
 - see *figure 8.10*
- we can express the joint effect of treatment and observation plan as ...

$$E[Y_t^{\{\bar{n}, \bar{a}=1\}} + 1] = E[Y_t^{\{\bar{n}, \bar{a}=0\}} + 1] \implies g(\bar{a}_t; \beta)$$

- if no direct effect of $N_{i,t}$ on Y_{t+1} then β is a causal parameter

- causal effect transportable only to population with same mixture of observation plans
- two solutions: artificial censoring vs. modeling effect of observation plan

Artificial Censoring for Static Observation Plans.

- re-express our identifiability assumptions to include the observation plan
 - (1) $Y_t^{\bar{n}, \bar{a}} = Y_t^{\bar{N}, \bar{A}} = Y_t$ if $\bar{N}_t = \bar{n}_t$ and $\bar{A}_t = \bar{A}_t$
 - (2) $Y_t^{\bar{n}, \bar{a}} \perp\!\!\!\perp \{\bar{N}_k, \bar{A}_k, \bar{L}_k\}$ for all \bar{n}, \bar{a} and $t > k$
 - (3) $f(\bar{n}_k, \bar{a}_{k-1}, \bar{l}_k > 0 \implies f(n_{k+1}, a_k | \bar{n}_{k+1}, \bar{a}_{k-1}, \bar{l}_k) > 0$
- create a pseudo-population that has the mixture of static observation plans we're interested in
- censor subjects as soon as they deviate from the plan of interest
- no selection bias because observation plans randomized at baseline

Modeling Effect of Observation Plan.

- $E[Y^{\bar{a}, \bar{n}_{t+1}} | N_{t+1} = 1] = g(\bar{a}_t, \bar{n}_t; \beta)$
- in the presence of a direct effect of N_t on $Y_t + 1$, the model helps 'transport' the effect in one population with a certain mixture of static plans to another population with a particular observation plan
 - more efficient than artificial censoring if few subjects following an observation plan
 - however, requires parametric assumptions whereas artificial censoring does not
 - could lead to extrapolation beyond the data
- three cases...
 - (1) single or multiple static regimes randomized with equal probability at baseline
 - normally would use $SW_t = SW^A \times SW^N$, but is unnecessary because $SW^N = 1$ for static regimes
 - however numerator and denominator equal 1 for static regimes that are pre-specified
 - thus $SW_t = SW_t^A \times SW_t^N = SW_t^A !!!$
 - (2) multiple static regimes randomized at baseline with probabilities dependent on baseline covariate V

$$SW_t^N = \prod_{k=1}^{t+1} \frac{f(N_k | \bar{N}_{k-1}, \bar{A}_{k-1})}{f(N_k | \bar{N}_{k-1}, \bar{A}_{k-1}, \bar{L}_{k-1})}$$

- even though $SW^N \neq 1$ can add V to structural model for outcome and the conditioning event of the numerator of the weights (e.g.

$$f(N_k | N_{k-1}, A_{k-1}, V \dots)$$

- these weights for SW^N do equal 1

- (3) observation plans are dynamic (e.g. depend on former treatment but not covariates)

- SW^N still equal to 1

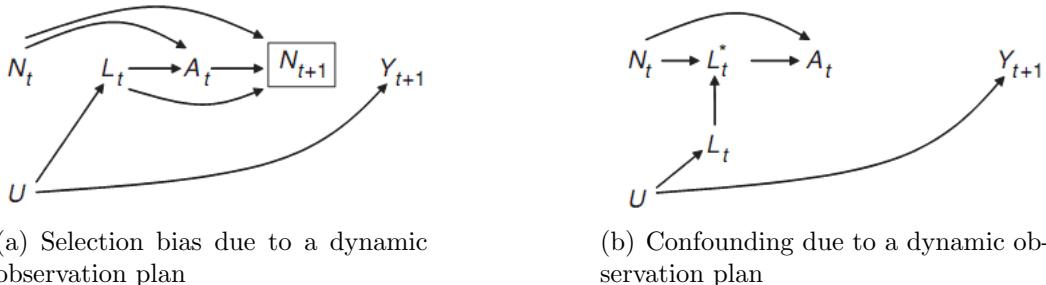


FIGURE 8.11

3.3. Dynamic Observation Plans.

- exchangeability: at every time t , one can achieve conditional exchangeability between observed and unobserved subjects by using the $\#A, L$ values measured at most recent time each subject was observed
- here we do need to take the observation plan N into account in our analysis

selection bias. (See figure 8.11(a))

- a person's observation time is affected by prior treatment history, and analysis is conditional upon the last observation time
- need to use modified weights $SW_t = SW_t^A \times SW_t^N$ to remove $L_t \rightarrow N_{t+1}$ and $L_t \rightarrow A_t$

– note that $SW_1^N \neq 1$

- censoring can be seen as an extreme case of this structure

Confounding. (See figure 8.11(b))

- for this bias N_i, t plays no role so we ignore it and all arrows into it
- here L^* is the most recently recorded value of covariates (i.e. confounders) by time t

- L^* and not L is what clinicians and patients use to decide treatment initiation/discontinuation

$$L^* = \begin{cases} L^* & \text{if } N_{i,t} = 0 \\ L & \text{if } N_{i,t} = 1 \end{cases}$$

- thus $N_{i,t} \rightarrow L^*$ is...
 - * *deterministic* if treatment can only change at observation times
 - * *predictive* if treatment can change at times other than observation times
- when treatment can only change at observation times
 - using weights $\frac{Pr[A_t | \bar{A}_{t-1}]}{Pr[A_t | \bar{A}_{t-1}, L^*_{t-1}]}$
is implicitly $\frac{Pr[A_t | \bar{A}_{t-1}]}{Pr[A_t | \bar{A}_{t-1}, L^*_{t-1}, N_t]}$
 - but we have a positivity violation because

$$Pr[A_t = 1 | A_{t-1} = 0, N_t = 0, \bar{N}_{t-1}, \bar{A}_{t-2}, L^*_{t-1}] = 0$$

- to solve we must use weights

$$SW_t^A \times SW_t^N$$

- because N is considered a 'treatment', we require it has no unmeasured confounders as well
that is...no $U \rightarrow N_t$ for all t

Issues with Exchangeability.

- conditional exchangeability holds only if L^* can suffice to block all back-door paths for treatment and observation at every time point
 - would hold if at each observation time $N_{t,i}, N_{t+1,i}$ is completely determined by \bar{A}_{k-1} and \bar{L}^*_{k-1}
 - e.g. will not hold if patients decide when to be observed based on their covariate values between visits
 - * these would be unmeasured confounders

Artificial Censoring for Dynamic Observation Plans.

- alternative to using $SW^A \times SW^N$ that adjust for L^*
- estimate effect within subjects who are always observed (i.e. $\bar{n} = \{1, 1, 1, \dots, 1\}$)
 - but in dynamic observation plans artificial censoring may be informative

- to adjust for this informative censoring, we use weights $SW_t^{A,D} = SW_t^A \times SW_t^D$

$$SW_t^D = \prod_{k=1}^{t+1} \frac{Pr[D_k = 0 | D_{k-1} = 0, \bar{A}_{k-1}]}{Pr[D_{k-1} = 0 | D_{k-1} = 0, \bar{N}_{k-1}, \bar{L}_{k-1}^*]}$$

where ...

$$D_k = \begin{cases} 1 & \text{if censored at time } k \\ 0 & \text{otherwise} \end{cases}$$

having \bar{N}_{k-1} to denominator is unnecessary

because $D_{k-1} = 0$ implies $N_m = 1$ for all observations $m > k$

- thus we fit the MSM

$$E[Y_{t+1}^{\bar{a}, \bar{d}=0} | D_{t+1} = 0] = g(\bar{a}_t; \beta)$$

$$g(\bar{a}_t; \beta) = \beta_0 + \beta_1 \sum_{k=0}^t a_k + \beta_2 t + \beta_3 t^2$$

by fitting

$$E[Y_{t+1} | \bar{A}, \bar{D} = 0] = g(\bar{a}_t; \gamma)$$

$$g(\bar{A}_t; \gamma) = \gamma_0 + \gamma_1 \sum_{k=0}^t A_k + \gamma_2 t + \gamma_3 t^2$$

weighted by $SW_t^{a,D} = SW_t^A \times SW_t^D$

- if the observation plan is still dynamic in the artificially censored data, we use

$$SW_t^{A,N,D} = SW_t^A \times SW_t^N \times SW_t^D$$

- this approach can be generalized to include only periods of follow up
where $\bar{n}_{t_k \rightarrow t_k + \delta} = \{1, 1, 1 \dots\}$
and periods $t_k \rightarrow t_k + \delta$ occur at regular, specified, intervals over follow up
- e.g. every 3 months (as opposed to every week of follow up)
- see *Hernan. Stat Methods Med Res 2008; 00:1-26* for further details

4. Static & Dynamic Treatment Regimes

4.1. Dynamic Regimes.

- treatment is dependent on \bar{L}_k and \bar{A}_{k-1} , where $k = 0, \dots, K$
- can refer to both random and non-random dynamic regimes, but usually refers to non-random dynamic regimes

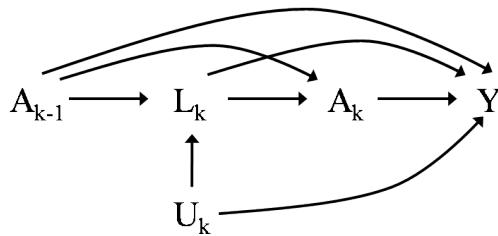


FIGURE 8.12. DAG of a dynamic regime where A_k is determined by A_{k-1} and L_k

4.1.1. Non-random Dynamic Regimes.

- treatment strategy g where treatment may depend on the evolution of time-dependent covariates \bar{L}_t and/or treatment \bar{A}_{t-1} up to time t in a *deterministic* manner
- simply put, regime looks at values of covariates L and A to determine whether to give current treatment
- example: take treatment if $RBC > 1,000$ and if treatment was not taken last week
- mathematically, $g_k\{\bar{a}_{k-1}, \bar{l}_k\}$ is a function of \bar{a}_{k-1} and \bar{l}_k that determines treatment at time k (a_k)

4.1.2. Random Dynamic Regimes.

- treatment strategy g where treatment may depend on the evolution of time-dependent covariates \bar{L}_t and/or treatment \bar{A}_{t-1} up to time t in a *probabilistic* manner
- example: assign treatment probability to 0.80 if $RBC > 1,000$ and if treatment was not taken last week, otherwise treatment probability is 0.5
- random dynamic regimes are *sequentially randomized experiments* because treatment A_k depends only on \bar{a}_{k-1} and \bar{l}_k in a known probabilistic manner at each time point k , thus $Y^g \perp\!\!\!\perp A_t | \bar{L}_t = \bar{l}_t, A_k = g_k\{\bar{A}_{k-1}, \bar{L}_k\}$

4.1.3. Optimal Treatment Regime.

- if high values of $E[Y]$ are preferred, then the optimal treatment regime g will maximize $E[Y^g]$
- should consider static treatment regimes when determining optimal regime
- should consider non-random dynamic regimes (likely optimal over static regime)
- should not consider random dynamic regimes for optimal regime because a random strategy should not be preferred as an optimal treatment strategy
- random dynamic regimes remain scientifically necessary for clinical trials when it is unknown which non-random dynamic regime is optimal (*equipoise*)

4.1.4. “Strengthened” Identifiability Conditions.

- under these strengthened conditions, a non-randomized (deterministic) dynamic treatment regime can be analyzed as a sequentially randomized trial

- that is “full” consistency and “full” exchangeability will be achieved when all parents of the treatment variables A_m are measured (no $U \rightarrow A_m$)
- “full” consistency and exchangeability only refer to non-dynamic regimes, but are met when “strengthened” consistency and exchangeability are achieved in a dynamic regime
- SW IPTW weights cannot be used with a dynamic regime, only W may be used
- the g-formula can be used if \bar{a}_{k-1} is replaced by $\bar{g}\{\bar{l}_{k-1}\}$ and \bar{a} is replaced by $\bar{g}\{\bar{l}_K\}$ (assuming only \bar{l}_k determines A_k), however, g-formula suffers from the null paradox and is less robust
- positivity condition remains unchanged

4.1.5. “*Strengthened*” Consistency.

- in addition to $Y^g = Y$, $\bar{L}_k^g = \bar{L}_k$,
- \bar{L}_k^g is the counterfactual history of L (\bar{L}) through time k under treatment regime g
- for treatment regime g where $A_k = g_k\{\bar{a}_{k-1}, \bar{l}_k\}$ (treatment depends on \bar{L} and past \bar{A}), $Y^g = Y$ and $\bar{L}_k^g = \bar{L}_k$
- for treatment regime g where $A_k = g_k\{\bar{l}_k\}$ (treatment depends only on \bar{L}), $Y^g = Y$ and $\bar{L}_k^g = \bar{L}_k$

4.1.6. “*Strengthened*” Conditional Exchangeability.

- in addition to $Y^g \perp\!\!\!\perp A_t | \bar{L}_t$, $Y^g \perp\!\!\!\perp A_t | \bar{L}_t, \bar{A}_k$
- counterfactual outcome Y^g is independent of treatment A_t given both past covariate history and treatment history
- for treatment regime g where $A_k = g_k\{\bar{a}_{k-1}, \bar{l}_k\}$,

$$Y^g \perp\!\!\!\perp A_t | \bar{L}_t = \bar{l}_t, A_k = g_k\{\bar{A}_{k-1}, \bar{L}_k\}$$

- for treatment regime g where $A_k = g_k\{\bar{l}_k\}$,

$$Y^g \perp\!\!\!\perp A_t | \bar{L}_t = \bar{l}_t, A_{t-1} = \bar{g}_{t-1}\{\bar{l}_{t-1}\}$$

Part 3

Inference without Models

CHAPTER 9

Statistical Inference

1. Hypothesis Testing

- $H_0 : x = a$ vs $H_1 : x \neq a$

$$Z^2 = \frac{(x - E[x|H_0])^2}{Var[x|H_0]} \sim \chi_1^2$$

- $(x - E[x|H_0])^2$ is a function of the magnitude and direction of effect
- $Var[x|H_0]$ is a function of statistical power
 - * power determined by study design, including sample size and balance of numbers of subjects in exposure levels (or case and control groups)
 - * can always try to get more power by study design
 - * thus, statistical significance does not always imply biological significance

- p-value: probability that a result as extreme or more extreme than the result we observed would occur due to chance variation, if the null hypothesis were true

- **interpretation of p value:** The probability that these data or any result more extreme would have been obtained if the null hypothesis were true is (plug in the p-value you have)

- “data” in the previous interpretation means *test statistics*, not effect estimates

- p-value gives no information on

- *effect magnitude*
- *direction*
- *range* of the effect that are consistent with the observed data (confidence interval)
- *statistical power of the study*

2. Confidence Interval

- $x \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(x)}$
- if we repeat the experiment many times, $100(1 - \alpha)\%$ of the time, it will include the true value of the parameter of interest

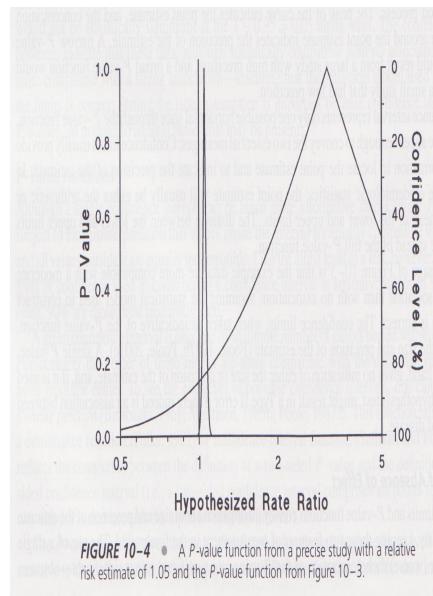


FIGURE 9.1. P-value function plot showing two rate ratios with different precision

- **interpretation of 95% confidence interval:** with 95% confidence, the true risk ratio is between a 2-fold increase and a 3-fold increase
- **interpretation of 95% confidence interval:** These data are consistent with a 2-fold to 3-fold increase in risk ratio of exposure A on outcome Y with 95% confidence, assuming there is no confounding, no selection bias, and no information bias
- confidence interval provides information on power by its range
- confidence interval can be used to do hypothesis testing (see figure 9.1)

3. Sources of Random Variability

- randomized trial
 - random treatment assignment
- observational study
 - sampling of subjects who validly represent the exposure-outcome association
 - conditional randomization by nature
- must achieve internal validity of study before statistical inference

CHAPTER 10

Traditional Methods

TABLE 10.1. Test of no exposure-disease association

$$Z^2 = \frac{(X - \hat{E}(X|H_0))^2}{\hat{Var}(X|H_0)} \sim \chi_1^2$$

		closed cohort		
		exposed	unexposed	
cases	a	b		M_1
non-cases	$N_1 - a$	$N_0 - b$		M_0
		N_1	N_0	T

		open cohort		
		exposed	unexposed	
cases	a	b		M_1
person-time	N_1	N_0		T

		case-control		
		exposed	unexposed	
cases	a	b		M_1
controls	c	d		M_0
		N_1	N_0	T

H_0	X	$\hat{E}(X H_0)$	$\hat{Var}(X H_0)$	stratified
closed cohort				
$C_1 = C_0$	a	$\frac{N_1 M_1}{T}$	$\frac{M_1 M_0 N_1 N_0}{T^3}$	unstratified
	$\sum a_i$	$\sum \frac{N_{1i} M_{1i}}{T_i}$	$\sum \frac{M_{1i} M_{0i} N_{1i} N_{0i}}{T_i^3}$	stratified
open cohort				
$I_1 = I_0$	a	$\frac{N_1 M_1}{T}$	$\frac{N_1 N_0 M_1}{T^2}$	unstratified
	$\sum a_i$	$\sum \frac{N_{1i} M_{1i}}{T_i}$	$\sum \frac{N_{1i} N_{0i} M_{1i}}{T_i^2}$	stratified
case-control				
$OR = 1$	a	$\frac{N_1 M_1}{T}$	$\frac{M_1 M_0 N_1 N_0}{T^2(T-1)}$	unstratified
	$\sum a_i$	$\sum \frac{N_{1i} M_{1i}}{T_i}$	$\sum \frac{M_{1i} M_{0i} N_{1i} N_{0i}}{T_i^2(T_i-1)}$	stratified

TABLE 10.2. Confidence intervals for ratio and difference measures

$$X \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(X)}$$

	X	w_i	$\hat{Var}(X)$
closed cohort			
risk difference	$\frac{a}{N_1} - \frac{b}{N_0}$	—	$\frac{ac}{N_1^3} + \frac{bd}{N_0^3}$
summary risk difference	$\frac{\sum w_i \left(\frac{a_i}{N_{1i}} - \frac{b_i}{N_{0i}} \right)}{\sum w_i}$	$\frac{N_{1i}^3 N_{0i}^3}{N_{0i}^3 a_i c_i + N_{1i}^3 b_i d_i}$	$\frac{1}{\sum w_i}$
risk ratio (ln)	$\ln \left(\frac{\frac{a}{N_1}}{\frac{b}{N_0}} \right)$	—	$\frac{c}{aN_1} + \frac{d}{bN_0}$
summary risk ratio (ln)	$\ln \left(\frac{\sum w_i \frac{\frac{a_i}{N_{1i}}}{\frac{b_i}{N_{0i}}}}{\sum w_i} \right)$	$\frac{b_i N_{1i}}{T_i}$	$\frac{\sum (M_{1i} N_{1i} N_{0i} - a_i b_i T_i)}{\left(\sum \frac{a_i N_{0i}}{T_i} \right) \left(\sum \frac{b_i N_{1i}}{T_i} \right)}$
open cohort			
rate difference	$\frac{a}{N_1} - \frac{b}{N_0}$	—	$\frac{a}{N_1^2} + \frac{b}{N_0^2}$
summary rate difference	$\frac{\sum w_i \left(\frac{a_i}{N_{1i}} - \frac{b_i}{N_{0i}} \right)}{\sum w_i}$	$\frac{N_{1i}^2 N_{0i}^2}{a_i N_{0i}^2 + b_i N_{1i}^2}$	$\frac{1}{\sum w_i}$
rate ratio (ln)	$\ln \left(\frac{\frac{a}{N_1}}{\frac{b}{N_0}} \right)$	—	$\frac{1}{a} + \frac{1}{b}$
summary rate ratio (ln)	$\ln \left(\frac{\sum w_i \frac{\frac{a_i}{N_{1i}}}{\frac{b_i}{N_{0i}}}}{\sum w_i} \right)$	$\frac{b_i N_{1i}}{T_i}$	$\frac{\sum (M_{1i} N_{1i} N_{0i})}{\left(\sum \frac{a_i N_{0i}}{T_i} \right) \left(\sum \frac{b_i N_{1i}}{T_i} \right)}$
case-control			
odds ratio (ln)	$\ln \left(\frac{ad}{bc} \right)$	—	$\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$
summary odds ratio (ln)	$\ln \left(\frac{\sum w_i \frac{a_i d_i}{b_i c_i}}{\sum w_i} \right)$	$\frac{b_i c_i}{T_i}$	RGB variance

TABLE 10.3. Estimate of the RGB variance

		case-control		
		exposed	unexposed	
cases		a	b	M_1
	controls	c	d	M_0
		N_1	N_0	T

For the variance of the natural log of the Mantel-Haenszel summary odds ratio we use the RGB variance. To compute the $\hat{var}(\ln(\hat{OR}_{MH}))$ the following formulas are used:

$$A_i = \frac{a_i d_i}{T_i}$$

$$B_i = \frac{b_i c_i}{T_i}$$

$$C_i = \frac{a_i + d_i}{T_i}$$

$$D_i = \frac{b_i + c_i}{T_i}$$

The derived quantities A_i, B_i, C_i , and D_i are combined and summed by stratum:

$$(AC) = \sum A_i C_i$$

$$(AD) = \sum A_i D_i$$

$$(BC) = \sum B_i C_i$$

$$(BD) = \sum B_i D_i$$

With this notation the \hat{OR}_{MH} can be written as:

$$\begin{aligned} \hat{OR}_{MH} &= \frac{\sum A_i}{\sum B_i} \\ &= \frac{\sum \frac{a_i d_i}{T_i}}{\sum \frac{b_i c_i}{T_i}} \end{aligned}$$

And $\hat{var}(\ln(\hat{OR}_{MH}))$ can be written as:

$$\begin{aligned} \hat{var}(\ln(\hat{OR}_{MH})) &= \frac{1}{2} \left(\frac{(AC)}{A^2} + \frac{(AD)+(BC)}{A \times B} + \frac{(BD)}{B^2} \right) \\ &= \frac{1}{2} \left(\frac{\sum \left(\frac{a_i d_i}{T_i} \right) \left(\frac{a_i + d_i}{T_i} \right)}{\left(\sum \frac{a_i d_i}{T_i} \right)^2} + \frac{\sum \left(\frac{a_i d_i}{T_i} \right) \left(\frac{b_i + c_i}{T_i} \right) + \sum \left(\frac{b_i c_i}{T_i} \right) \left(\frac{a_i + d_i}{T_i} \right)}{\left(\sum \frac{a_i d_i}{T_i} \right) \left(\sum \frac{b_i c_i}{T_i} \right)} + \frac{\sum \left(\frac{b_i c_i}{T_i} \right) \left(\frac{b_i + c_i}{T_i} \right)}{\left(\sum \frac{b_i c_i}{T_i} \right)^2} \right) \end{aligned}$$

1. Road Map

- Mantel-Haenszel test statistics $\sim \chi_1^2$
 - $H_0 : \hat{RR}_{MH} = 1$ or $\hat{RD}_{MH} = 0$ or “there is no association between exposure and outcome”
- Homogeneity test statistics $\sim \chi_{I-1}^2$ where I is the number of levels of stratification factor
 - $H_0 : \hat{RR}_i = RR_j$ for all i, j or “all stratum specific effect estimates are the same”
- stratum specific weights differ for absolute and relative measures of effect
 - *risk/rate difference*: inverse variance weights are most efficient ($\frac{1}{\text{var}(RD)}$)
 - *risk/rate/odds ratio*: Mantel-Haenszel weights are most efficient ($\frac{b_i N_{1i}}{T_i}$ or $\frac{b_i c_i}{T_i}$)

1.1. inference about hypothesis testing.

1.1.1. count data.

- count data refers to
 - (1) number of cases and non-cases collected in a closed cohort study
 - (2) number of cases and controls collected in a case-control study
- for count data, the 2 numbers in the 4 cells of the contingency table can be random variables
 - 2 of $a, (N_1 - a), b, (N_0 - b)$ for closed cohort study
 - 2 of a, b, c, d for case-control study
 - see closed cohort study and case-control study part of table 10.1
- however, it's difficult to make inference about 2 random variables at the same time
- thus, we pretend that we know the margin numbers of the table before we even collect the data
 - for closed cohort study, we only know how many exposed subjects (N_1) and unexposed subjects (N_0) will be recruited into the study
 - for case-control study, we only know how many cases (M_1) and controls (M_0) will be recruited into the study
- that is, we fix the margins of the contingency table (M_0, M_1, N_0, N_1)
- **key assumption:**
 - by fixing the margins, only one random variable is left for inference
 - \Rightarrow this random variable follows **hypergeometric** distribution
- hypothesis test for random variable follows hypergeometric distribution

- random variable: a
- from hypergeometric distribution, we can get expected value and variance of a

$$E[a|H_0] = \frac{N_1 M_1}{T}$$

$$Var[a|H_0] = \frac{M_1 M_0 N_1 N_0}{T^2(T-1)}$$

- construct test statistics:

$$Z = \frac{a - E[a|H_0]}{\sqrt{Var[a|H_0]}} \text{ follows normal distribution asymptotically}$$

$$\frac{(a - E[a|H_0])^2}{Var[a|H_0]} \sim \chi_1^2 \text{ asymptotically}$$

$$\frac{(a - \frac{N_1 M_1}{T})^2}{\frac{M_1 M_0 N_1 N_0}{T^2(T-1)}} \sim \chi_1^2 \text{ asymptotically}$$

- for closed cohort data, use $Var[a|H_0] = \frac{M_1 M_0 N_1 N_0}{T^3}$ for continuity correction

1.1.2. person-time data.

- person-time data refers to
 - (1) number of cases (counts) and person-times collected in a closed cohort study
 - (2) or in an open cohort study
- for person-time data, the exposed and unexposed person-times are only “units” used to standardize the counts
 - we may arbitrarily change the person-time unit for a incidence rate
 - e.g. 1 per 1000 person-year = 10 per 10,000 person-year
 - however, we need the same unit to compare counts in person-time data
 - e.g. 1 case is not less than 10 cases in the sense that the former 1 case aroused from 1000 person-year exposed person-time and the latter 10 cases aroused from 10,000 person-year exposed person-time
- because person-times are only “units” but not random variables, we do not use them in hypothesis test
- we only use counts as random variables in person-time data, which are the numbers of exposed cases (a) and unexposed cases (b)
 - see open cohort study part of table 10.1
- **key assumption**

we pretend that we know the total number of cases in the study but we do not know how many exposed cases will be there

- that is, a is a random variable follows **binomial** distribution with the total number of trials is **(a+b)**

- $Binomial(n, p)$: n is the total number of trials and p is the probability of successful trials
- hypothesis test for random variable follows **binomial** distribution
 - random variable: a
 - from binomial distribution, we can get expected value and variance of a

$$E[a|H_0] = \frac{N_1 M_1}{T}$$

$$Var[a|H_0] = \frac{M_1 N_1}{T^2}$$

- construct test statistics:

$$Z = \frac{a - E[a|H_0]}{\sqrt{Var[a|H_0]}} \text{ follows normal distribution asymptotically}$$

$$\frac{(a - E[a|H_0])^2}{Var[a|H_0]} \sim \chi_1^2 \text{ asymptotically}$$

$$\frac{(a - \frac{N_1 M_1}{T})^2}{\frac{M_1 N_1}{T^2}} \sim \chi_1^2 \text{ asymptotically}$$

1.1.3. Mantel-Haenszel test.

- once we have the two tests constructed above, it's easy to construct M-H test
- essentially ...

$$\frac{(\sum_i a_i - \sum_i E[a_i|H_0])^2}{\sum_i Var[a_i|H_0]} \sim \chi_1^2 \text{ asymptotically}$$

- use the expected values and variances correspond to the data type

1.1.4. use risk difference to perform hypothesis test for count data.

- we may also use a regular paired t-test
- we compared risk among the exposed to the risk among the unexposed
- because risks are probabilities, we take the two risks as probabilities of success in two binomial distributed random variables and compare them
- that is ...

- random variable $X = (\frac{a}{N_1} - \frac{b}{N_0})$

- X has $E[X|H_0] = 0$ and $Var[X|H_0] = \frac{M_0 M_1}{T} (\frac{1}{N_1} + \frac{1}{N_0})$

- construct test statistics:

$$t = \frac{X - E[X|H_0]}{\sqrt{Var[X|H_0]}}$$

t follows normal distribution asymptotically (when n is greater than 30...)

$$\frac{(X - E[X|H_0])^2}{Var[X|H_0]} \sim \chi_1^2 \text{ asymptotically}$$

$$\frac{\left(\frac{a}{N_1} - \frac{b}{N_0}\right)^2}{\frac{M_0}{T} \frac{M_1}{T} \left(\frac{1}{N_1} + \frac{1}{N_0}\right)} \sim \chi_1^2 \text{ asymptotically}$$

1.1.5. use rate difference to perform hypothesis test for person-time data.

- we may also use a regular paired t-test
- we compared rate among the exposed to the rate among the unexposed
- because rates are parameters of Poisson distributed random variables, we take the two rates as parameters of two Poisson distributions and compare them
- that is ...
 - random variable $X = \left(\frac{a}{N_1} - \frac{b}{N_0}\right)$
 - X has $E[X|H_0] = 0$ and $Var[X|H_0] = \frac{M_1}{T} \left(\frac{1}{N_1} + \frac{1}{N_0}\right)$
 - construct test statistics:

$$t = \frac{X - E[X|H_0]}{\sqrt{Var[X|H_0]}}$$

t follows normal distribution asymptotically (when n is greater than 30...)

$$\frac{(X - E[X|H_0])^2}{Var[X|H_0]} \sim \chi_1^2 \text{ asymptotically}$$

$$\frac{\left(\frac{a}{N_1} - \frac{b}{N_0}\right)^2}{\frac{M_1}{T} \left(\frac{1}{N_1} + \frac{1}{N_0}\right)} \sim \chi_1^2 \text{ asymptotically}$$

1.2. inference about 95% confidence interval.

- $100 \times (1 - \alpha)\%$ confidence interval for a variable follows normal distribution is $\hat{X} \pm Z_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{X})}$
- when X is a parameter (e.g. rate ratio), X usually follows normal distribution
 - Central limit theorem: the summation of values of a random variable following any distribution ($\sum_i X_i$) would follows normal distribution
- the formula for 95% confidence interval of a parameter estimate is $\hat{X} \pm 1.96 \times se(\hat{X})$
 - $se(\hat{X}) = \sqrt{Var(\hat{X})}$ given \hat{X} is a parameter estimate
- therefore, the inference of confidence interval relies entirely on $Var(\hat{X})$
- the following part is just to show how to get variance for different parameters
 - plug-in $se(\hat{X})$ into the formula given above to get 95% confidence interval
- for ratios, we **log-transform** the ratio first and then exponentiate it back to original scale after getting the confidence interval for $\log(ratio)$
 - $e^{\log(\hat{X}) \pm 1.96 \times se(\log(\hat{X}))}$

- this is purely a technical issue (it is much easier to deal with $Var(\log(x) - \log(y))$ than $Var(\frac{x}{y})$)

1.2.1. Odds ratio (OR).

- see table 10.1
- $OR = \frac{ad}{bc}$
- $\log(OR) = \log(a) + \log(d) - \log(b) - \log(c)$
- $Var(\log(OR)) = Var[\log(a) + \log(d) - \log(b) - \log(c)]$

$Var(X + Y) = Var(X) + Var(Y)$ and $Var(X - Y) = Var(X) + Var(Y)$ given X and Y are independent from each other

- $Var(\log(OR)) = Var[\log(a)] + Var[\log(b)] + Var[\log(c)] + Var[\log(d)]$

$Var[X] = E[X]$ given X is a count (Poisson distribution)

$$Var[\log(X)] = (\frac{1}{X})^2 Var(X) = (\frac{1}{X})^2 (X) = \frac{1}{X} \text{ (delta method)}$$

- $Var(\log(OR)) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$

a different way to get the variance of odds ratio

- this method is based on assuming binomial distribution, which is closer to a actual data structure
- the previous one is based assuming Poisson distribution for each count, which is not actually the case

$$\begin{aligned}
 Var(\log(OR)) &= Var(\log(\frac{a/b}{c/d})) \\
 &= Var(\log(\frac{a}{b}) - \log(\frac{c}{d})) \\
 &= Var(\log(\frac{a}{b})) + Var(\log(\frac{c}{d})) \\
 &= Var(\log(\frac{a/M_1}{b/M_1}) + Var(\log(\frac{c/M_0}{d/M_0})) \\
 &= Var(\log(\frac{\hat{P}_1}{1 - \hat{P}_1})) + Var(\log(\frac{\hat{P}_0}{1 - \hat{P}_0})) \\
 &\quad \text{set } \frac{a}{M_1} = P_1 \text{ and } \frac{c}{M_0} = P_0, \text{ both are parameters of binomial distribution} \\
 &= (\frac{1}{P_1} - \frac{-1}{1 - P_1})^2 Var(\hat{P}_1) + (\frac{1}{P_0} - \frac{-1}{1 - P_0})^2 Var(\hat{P}_0) \\
 &\quad Var(\log(\frac{\hat{P}}{1 - \hat{P}})) = Var(\log(P) - \log(1 - P)) = (\frac{1}{P} - \frac{1}{1 - P})^2 Var(\hat{P}) \\
 &= (\frac{1}{P_1(1 - P_1)})^2 Var(\hat{P}_1) + (\frac{1}{P_0(1 - P_0)})^2 Var(\hat{P}_0) \\
 &= (\frac{1}{P_1(1 - P_1)})^2 \frac{\hat{P}_1(1 - \hat{P}_1)}{M_1} + (\frac{1}{P_0(1 - P_0)})^2 \frac{\hat{P}_0(1 - \hat{P}_0)}{M_0} \\
 &= (\frac{1}{P_1(1 - P_1)}) \frac{1}{M_1} + (\frac{1}{P_0(1 - P_0)}) \frac{1}{M_0} \\
 &= (\frac{1}{\frac{a}{M_1}(1 - \frac{a}{M_1})}) \frac{1}{M_1} + (\frac{1}{\frac{c}{M_0}(1 - \frac{c}{M_0})}) \frac{1}{M_0} \\
 &= (\frac{1}{\frac{a}{M_1} \frac{b}{M_1}}) \frac{1}{M_1} + (\frac{1}{\frac{c}{M_0} \frac{d}{M_0}}) \frac{1}{M_0} \\
 &= (\frac{1}{\frac{ab}{M_1}}) + (\frac{1}{\frac{cd}{M_0}}) \\
 &= \frac{M_1}{ab} + \frac{M_0}{cd} \\
 &= \frac{a+b}{ab} + \frac{c+d}{cd} \\
 &= \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}
 \end{aligned}$$

- this method requires more algebra and is confusing ...

1.2.2. *rate ratio.*

- see table 10.1
- $IRR = \frac{a/N_1}{b/N_0}$
- $\log(IRR) = \log(\frac{a}{N_1}) - \log(\frac{b}{N_0})$
- $Var(\log(IRR)) = Var(\log(\frac{a}{N_1}) - \log(\frac{b}{N_0})) = Var(\log(\frac{a}{N_1})) + Var(\log(\frac{b}{N_0}))$
- $Var(\log(IRR)) = Var(\log(\frac{a}{N_1})) + Var(\log(\frac{b}{N_0})) = (\frac{N_1}{a})^2 Var(\frac{a}{N_1}) + (\frac{N_0}{b})^2 Var(\frac{b}{N_0})$
by delta method
- $Var(\log(IRR)) = (\frac{N_1}{a})^2 Var(\frac{a}{N_1}) + (\frac{N_0}{b})^2 Var(\frac{b}{N_0}) = (\frac{N_1}{a})^2 (\frac{1}{N_1})^2 Var(a) + (\frac{N_0}{b})^2 (\frac{1}{N_0})^2 Var(b)$

N_1 and N_0 are person-time units used to standardize the study bases from which the counts are from; therefore, N_1 and N_0 are considered as constants, not random variables (random variables are a and b)

$Var(kX) = k^2 Var(X)$ given that k is a constant

- $Var(\log(IRR)) = (\frac{N_1}{a})^2 (\frac{1}{N_1})^2 Var(a) + (\frac{N_0}{b})^2 (\frac{1}{N_0})^2 Var(b) = (\frac{N_1}{a})^2 (\frac{1}{N_1})^2 (a) + (\frac{N_0}{b})^2 (\frac{1}{N_0})^2 (b)$
 $Var(X) = X$ for a count
- $Var(\log(IRR)) = \frac{1}{a} + \frac{1}{b}$

1.2.3. *risk ratio.*

- see table 10.1
- $RR = \frac{a/N_1}{b/N_0}$
- $\log(RR) = \log(\frac{a}{N_1}) - \log(\frac{b}{N_0})$
- $Var(\log(RR)) = Var(\log(\frac{a}{N_1}) - \log(\frac{b}{N_0})) = Var(\log(\frac{a}{N_1})) + Var(\log(\frac{b}{N_0}))$
- $Var(\log(RR)) = Var(\log(\frac{a}{N_1})) + Var(\log(\frac{b}{N_0})) = (\frac{N_1}{a})^2 Var(\frac{a}{N_1}) + (\frac{N_0}{b})^2 Var(\frac{b}{N_0})$
by delta method
- $$\begin{aligned} Var(\log(RR)) &= (\frac{N_1}{a})^2 Var(\frac{a}{N_1}) + (\frac{N_0}{b})^2 Var(\frac{b}{N_0}) \\ &= (\frac{N_1}{a})^2 (\frac{a}{N_1})(1 - \frac{a}{N_1})(\frac{1}{N_1}) + (\frac{N_0}{b})^2 (\frac{b}{N_0})(1 - \frac{b}{N_0})(\frac{1}{N_0}) \\ &= (\frac{N_1}{a})^2 (\frac{a}{N_1})(\frac{N_1 - a}{N_1})(\frac{1}{N_1}) + (\frac{N_0}{b})^2 (\frac{b}{N_0})(\frac{N_0 - b}{N_0})(\frac{1}{N_0}) \\ &= \frac{N_1 - a}{aN_1} + \frac{N_0 - b}{bN_0} \end{aligned}$$

$Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$ for \hat{p} is the parameter for a binomial distribution

- $Var(\log(RR)) = \frac{N_1-a}{aN_1} + \frac{N_0-b}{bN_0}$

1.2.4. rate ratio.

- see table 10.1
- $IRD = \frac{a}{N_1} - \frac{c}{N_0}$
- $Var(IRD) = Var\left(\frac{a}{N_1} - \frac{c}{N_0}\right) = Var\left(\frac{a}{N_1}\right) + Var\left(\frac{c}{N_0}\right) = \left(\frac{1}{N_1}\right)^2 Var(a) + \left(\frac{1}{N_0}\right)^2 Var(c)$
 $Var(kX) = k^2 Var(X)$ given that k is a constant
- $Var(IRD) = \left(\frac{1}{N_1}\right)^2 Var(a) + \left(\frac{1}{N_0}\right)^2 Var(c) = \left(\frac{1}{N_1}\right)^2 (a) + \left(\frac{1}{N_0}\right)^2 (c)$
 $Var[X] = E[X]$ given X is a count (Poisson distribution)
- $Var(IRD) = \frac{a}{N_1^2} + \frac{c}{N_0^2}$

1.2.5. risk ratio.

- see table 10.1
- $RD = \frac{a}{N_1} - \frac{c}{N_0}$
- $Var(RD) = Var\left(\frac{a}{N_1} - \frac{c}{N_0}\right) = Var\left(\frac{a}{N_1}\right) + Var\left(\frac{c}{N_0}\right)$
- $Var(RD) = Var\left(\frac{a}{N_1}\right) + Var\left(\frac{c}{N_0}\right) = \frac{a}{N_1}(1 - \frac{a}{N_1})\left(\frac{1}{N_1}\right) + \frac{c}{N_0}(1 - \frac{c}{N_0})\left(\frac{1}{N_0}\right)$
 $\frac{a}{N_1}$ and $\frac{c}{N_0}$, both are parameters of binomial distribution
 $Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$ for \hat{p} is the parameter for a binomial distribution
- $Var(RD) = \frac{a}{N_1} \frac{b}{N_1} \frac{1}{N_1} + \frac{c}{N_0} \frac{d}{N_0} \frac{1}{N_0}$
- $Var(RD) = \frac{ab}{N_1^3} + \frac{cd}{N_0^3}$

2. Stratification

- marginal (unconditional) causal effect measures: causal effect for the entire population, $Pr[Y^{a=1} = 1]$ vs $Pr[Y^{a=0} = 1]$
- conditional causal effect measures: causal effect for a subset of the population, $Pr[Y^{a=1} = 1 | M = m]$ vs $Pr[Y^{a=0} = 1 | M = m]$
 - if the stratum-specific effect measures are all the same, the effect measure for an entire population would be the same as the stratum-specific effect measure
- collapsibility: population effect measure can be expressed as a weighted average of the stratum specific measures
 - risk ratio and risk difference are collapsible
- non-collapsibility
 - odds ratio and odds difference are non-collapsible

- when the stratum-defining variable is associated with outcome, the population odds ratio can be closer to the null value than any of the non-null stratum specific causal odds ratios
- under the sharp null hypothesis, all effect measures are collapsible
- restriction: form of stratification where investigator only looks at certain strata of L
 - can be used to preserve positivity
- can be used to compute average causal effects in a subset of the population, but not individual effects
- pooled stratum-specific effect measures
 - only when there is no effect measure modification
 - reduce variability of the estimate

3. Matching

- selection of a reference series that is identical to the index series *with respect to the distribution of one or more potential confounder*
- chiefly applied in case-control studies

3.1. The Matching Process.

- matching factor: factor L by which exposed and unexposed in a prospective cohort study are matched
 - mind that matching factor (and many other factors) will change over time
 - desirable characteristics of matching factors
 - (1) easily assessed in entire population
 - (2) not interested in its effect
 - (3) meet positivity assumption (strata with only treated, or untreated, individuals are excluded from the analysis)
 - (4) are associated with outcome
 - (5) not so tied up with exposure that all members of matched sets (both cases and controls) have the same value for exposure
- subjects are matched $1 : N$ (or $N : N$) to create a subset of *matched subjects* where the variables L are forced, *by design*, to follow the same distribution in both the treated and untreated
 - (1) *individual matching*: more than 4 controls per case would result in little gain in power

- relative efficiency (RE) = $\frac{m}{m+1}$, where m is the number of matched controls per case
- (2) frequency matching: may have residual confounding after frequency matching on continuous variable
 - ex: deciding to have 50 % males and 50 % females in both the exposed and unexposed groups
 - can be used to create a matched population that follows *any* arbitrary distribution of L (i.e. 10/20, 50/50, 70/30, etc.)

Matching, Confounding, & Efficiency.

- in case-control study matching will only improve the statistical efficiency in the subsequent stratified analysis but will *not* remove confounding
- if to control for confounding only through matching
 - must assume conditional exchangeability given matching factors L
 - in the subset of matched pairs, the treated and untreated are marginally exchangeable
- if matching for efficiency
 - not all covariates needed to achieve exchangeability are included
- usually always want to adjust for matching factor in analysis because it usually imposes a selection bias (except in the case of over-matching)
 - match for efficiency only: $A \rightarrow [S] \leftarrow M \rightarrow Y$
 - match for confounding: add arrow $M \rightarrow A$
 - overmatching: $A \rightarrow [S] \leftarrow M \rightarrow Y$
- matching can be used to create a matched population with *any* arbitrary distribution of L in both the treated and the untreated
 - matching can be seen as “standardization by design”
- the average causal effect in the matched population may not be the same as the average causal effect in the population
 - causal effects computed by matching are *conditional* because they apply to a subset of the population with a particular distribution of covariates L

3.2. Matching Strategies.

- there are several different matching strategies
 - effect in the untreated: match to make the study population look like (i.e. have the covariate distribution of L) the untreated group

- effect in the treated: match to make the study population look like (i.e. have the covariate distribution of L) the treated group
- in 1:1 matching, the treatment group with the smaller sample size determines the distribution of L in the final matched population (i.e. strata of L that have only treated or only treated are discarded)

4. Matched Cohort study

- prevents association between confounder and exposure of interest
- matching prevents confounding even in *crude analysis*
- stratified analysis improves efficiency
- *can* assess the effect of matching factor on the outcome
- *can* assess the effect modification by matching factor on the exposure of interest

5. Matched Case-Control study

- eliminates the association between confounder and the outcome
- matching does not prevent confounding *by itself*
- matching only improves efficiency
- matching for *efficiency* , NOT validity
 - a way to deal with potential positivity violations
 - ensure enough cases and controls within strata to allow for adjustment
 - stratified analysis will be performed for confounding with or without matching on confounder
- matching introduces *selection bias* \Rightarrow *toward the null*
 - if the matching factor is associated with the exposure, matching would make the exposure distribution in controls similar to the exposure distribution in cases
 - must perform matched analysis (stratified analysis) after matching in case-control studies
- *CANNOT* assess the effect of matching factor on the outcome
- *can* assess effect modification by matching factor on the exposure of interest
- when the level of potential confounders is *unique* each case, matching is necessary to obtain controls for each case (individual matching)
- when each matched set represents a level of confounder that is observed *repeatedly*, all study subjects with the same level of confounders can be analyzed with in a single stratum (stratified analysis)

- e.g. for 1 : 1 individually matching on gender only, may perform stratified analysis on gender instead of stratified analysis on matching set.

TABLE 10.4. Efficiency and validity of matching under different analysis schemes

		Effect of Matching in Case-control Studies		
		Design	ANALYSIS	
		Stratified	Not Stratified	
$E \leftarrow M \rightarrow D$	Appropriate Matching	Match	V PPP	BIAS
		Do not Match	V P	BIAS
$E \rightarrow M \rightarrow D$	Unneccesary Matching	Match	V PP	V PPP
		Do not Match	V PP	V PPP
$E \leftarrow M \rightarrow D$	Over-Matching	Match	V P	BIAS
		Do not Match	V P	V PPP
$E \rightarrow M \rightarrow D$	Match on Intermediate	Match	BIAS	BIAS
		Do not Match	BIAS	V PPP

6. Informative and non-informative risk sets

- note that informativeness is relative to your estimand
- For ratio measures
 - you need a case, exposed and unexposed denominator
- however for the intercept, i.e. the absolute 'baseline' rate
 - all you need is denominator in the unexposed
 - for this reason, in Poisson strata that have no cases and only unexposed person-time *are* informative for the estimation of the baseline rate, but not the effect estimate for exposure
 - in cox however, such risk-sets are thrown out completely
- also note, in a technical sense all you need is a case and variation in covariates to contribute to the ratio parameters of the model in some way
- all of this translates to matched sets in matched case-control studies
- when stratifying on time scale and additional confounders 2x2 tables per stratum represent risk sets
- using the Mantel-Haenszel estimator $RR_{MH} = \frac{\sum_i \frac{a_i N_{0i}}{T_i}}{\sum_i \frac{b_i N_{1i}}{T_i}}$ only risk sets with $M_{1i} > 0$ are considered informative so when $a_i > 0$ the risk set contributes to the numerator

TABLE 10.5. Notation for cohort and case-control data

	exposed	unexposed	
cases	a	b	M_1
non-cases	$N_1 - a$	$N_0 - b$	M_0
	N_1	N_0	T

of the estimator and when $b_i > 0$ it contributes to the denominator if both $N_{1i} > 0$ and $N_{0i} > 0$

- the same informative risk sets are the ones contributing to the RR_{MH} variance $\hat{Var}[\log(\hat{RR})] = \frac{A}{BC}$, where
$$A = \sum_i \frac{M_{1i}N_{1i}N_{0i}}{T_i^2}$$

$$B = \sum_i \frac{a_iN_{0i}}{T_i}$$

$$C = \sum_i \frac{b_iN_{1i}}{T_i}$$
- matched data
 - risk set sampling assures $M_{1i} > 0$ as risk sets are defined at the time each case occurs

		Unexposed Pair Member		Exposed totals
		Disease	No disease	
Exposed Pair Member	Disease	T	U	T+U
	No Disease	V	W	V+W
	Unexposed totals	T+V	U+W	

FIGURE 10.1. 2x2 Notation for matched pairs cohort data

- consider figure 10.1 where
 - T are pairs with an exposed and an unexposed case
 - U are pairs with an exposed case and unexposed non-case
 - V are pairs with an unexposed case and exposed non-case
 - W are pairs with an exposed and unexposed non-case
- matched case control data assure $M_{1i} > 0$ for each stratum of matching factors

- for matched cohort data the RR_{MH} simplifies to $RR_{MH} = \frac{T+U}{T+V}$

		Control Pair Member		Case Totals
		Exposed	Unexposed	
Case Pair Member	Exposed	T	U	T+U
	Unexposed	V	W	V+W
Control Totals		T+V	U+W	

FIGURE 10.2. 2x2 Notation for matched pairs case-control data

- consider figure 10.2 where
 - T are pairs with an exposed case and an exposed control
 - U are pairs with an exposed case and unexposed control
 - V are pairs with an unexposed case and exposed control
 - W are pairs with an unexposed case and unexposed control
- in matched data analysis informative risk sets are those with discordant case-control pairs with respect to exposure levels for OR estimates

$$OR_{MH} = \frac{\sum_i \frac{a_i d_i}{T_i}}{\sum_i \frac{b_i c_i}{T_i}} \text{ simplifies to } OR_{MH} = \frac{U}{V}$$

7. Analysis of Matched Case-control Studies

- matching should be followed by stratified analysis
- match set = strata
- Paired data analysis for matched case-control studies
 - (1) notation

		controls	
cases		exposed	unexposed
cases	exposed	f_{11}	f_{10}
	unexposed	f_{01}	f_{00}

- f_{ij} is the number of pairs

– i is the exposure status of the case; j is the exposure status of the control

(2) hypothesis testing (McNemar test)

H_0 : there is no association between the exposure and the outcome

H_1 : there is an association between the exposure and the outcome

$$Z^2 = \frac{[f_{10} - f_{01}]^2}{f_{10} + f_{01}} \sim \chi_1^2$$

(3) only discordant pairs provide information

(4) same as the transmission disequilibrium test (TDT) in genetics

(5) confidence interval

$$- \hat{OR}_{MH} = \frac{f_{10}}{f_{01}}$$

$$- Var[\log(\hat{OR}_{MH})] = \frac{1}{f_{10}} + \frac{1}{f_{01}}$$

• McNemar vs Mantel-Haenszel

– for 1 : 1 matching

		Layout A		Layout B		Layout C		Layout D	
		E	\bar{E}	E	\bar{E}	E	\bar{E}	E	\bar{E}
Case	1	1	0	Case	1	0	Case	0	1
Cont	1	0	0	Cont	0	1	Cont	1	0

FIGURE 10.3. paired data in stratification layout

$$\hat{OR}_{MH} = \frac{\sum_i \frac{a_i d_i}{T_i}}{\sum_i \frac{b_i c_i}{T_i}} = \frac{A[\frac{1 \times 0}{2}] + B[\frac{1 \times 1}{2}] + C[\frac{0 \times 0}{2}] + D[\frac{0 \times 1}{2}]}{A[\frac{0 \times 1}{2}] + B[\frac{0 \times 0}{2}] + C[\frac{1 \times 1}{2}] + D[\frac{1 \times 0}{2}]} = \frac{B}{C}$$

CHAPTER 11

G-Methods

1. Standardization

1.1. General Idea.

$$E[Y^a] = \sum_l E[Y|A = a, L = l] Pr[L = l]$$

- the general idea is to simulate an occurrence measure for population A as if, *counterfactually*, population A had the covariate distribution (looked like) population B
- weights equal to the proportion of individuals in the reference population within each stratum (i.e. $Pr[L = l]$)
- causal risk ratio: $\frac{\sum_l Pr[Y=1|A=1, L=l]Pr[L=l]}{\sum_l Pr[Y=1|A=0, L=l]Pr[L=l]}$
- standardized occurrence measure: weighted average of the stratum-specific occurrence measures (i.e. $Pr[Y = 1|L = l]$) where the weights are the stratum-specific probabilities of the covariate L (i.e. $Pr[L = l]$)

$$\sum_l Pr[Y = 1|A = a, L = l] \times Pr[L = l]$$

if $Pr[A = a|L = l] > 0$ for all l with $Pr[L = l] \neq 0$

- standard population: reference population used to create the weight $Pr[L = l]$
- For a subset of the population ($A = a'$)

$$\sum_l E[Y|A = a, L = l] Pr[L = l|A = a']$$

TABLE 11.1. traditional notation for case-control/cohort data

		exposed	unexposed	
cases	a	b		M_1
non-cases	$N_1 - a$	$N_0 - b$		M_0
	N_1	N_0		T

1.2. Traditional Notation.

Unified approach.

- Standardization can be expressed as a weighted average of stratum specific effect measures

- in the presence of effect modification we prefer standardization
 - * the weights meaningly refer to the structure of the population and provide a true average effect
- in the absence of effect modification we prefer mantel-haenszel / inverse variance weights
 - * the weights are arbitrary, but if **no** effect modification they provide a true average effect
 - *assuming identifiability conditions hold*

- **Direct standardization of the effect measure**

- standardized rate(risk) ratio = $\sum_I w_i \hat{R}R_i = \frac{\sum_I w'_i \hat{R}R_i}{\sum_I w'_i}$
- standardized rate(risk) difference = $\sum_I w_i \hat{R}D_i = \frac{\sum_I w'_i \hat{R}D_i}{\sum_I w'_i}$
- $w_i = \frac{w'_i}{\sum_I w'_i}$ and $\sum_I w_i = 1$
- strata I defined by collection of covariates (L vector in counterfactual notation)

- **Direct standardization**

- weighted average of stratum specific rate ratios using the distribution of unexposed cases as the weights ($\hat{I}_{0i}N_0$)
- unexposed is the standard in terms of stratum specific rates and data distribution

$$SRR = \sum_I w_i \hat{R}R_i = \sum_I \frac{\hat{I}_{0i}N_{0i}}{\sum_I \hat{I}_{01}N_{0i}} \times \frac{\frac{a_i}{N_{1i}}}{\frac{b_i}{N_{0i}}} = \sum_I \frac{b_i}{\sum_I b_i} \times \frac{\frac{a_i}{N_{1i}}}{\frac{b_i}{N_{0i}}} = \frac{\sum_I \frac{a_i}{N_{1i}} N_{0i}}{\sum_I b_i}$$

where ...

$$w_i = \frac{\hat{I}_{0i}N_{0i}}{\sum_I \hat{I}_{01}N_{0i}} = \frac{\frac{b_i}{N_{0i}}N_{0i}}{\sum_I \frac{b_i}{N_{0i}}N_{0i}} = \frac{b_i}{\sum_I b_i}$$

$$w'_i = b_i$$

- **Indirect standardization**

- weighted average of stratum-specific rate ratios using the distribution of expected exposed cases, which were calculated by the incidence rate among the

controls (\hat{I}_{0i}) and the distribution of exposed subjects (N_{1i}), as the weights

$$SMR = \sum_I w_i \hat{R} \hat{R}_i$$

where ...

$$w_i = \frac{\hat{I}_{01} N_{1i}}{\sum_I \hat{I}_{01} N_{1i}} = \frac{\frac{b_i}{N_{0i}} N_{1i}}{\sum_I \frac{b_i}{N_{0i}} N_{1i}} = \frac{E_i}{\sum_I E_i}$$

$$w'i = E_i$$

$$\begin{aligned} SMR &= \sum_I w_i \hat{R} \hat{R}_i = \sum_I w_i \frac{\hat{I}_{1i}}{\hat{I}_{1i}} = \sum_I w_i \frac{\frac{a_i}{N_{1i}}}{\frac{b_i}{N_{0i}}} \\ &= \frac{\sum_I \frac{b_i}{N_{01}} N_{1i}}{\sum_I \frac{b_i}{N_{01}} N_{1i}} \frac{\frac{a_i}{N_{1i}}}{b_i N_{0i}} = \dots = \frac{\sum_I a_i}{\sum_I \frac{b_i}{N_{0i}} N_{1i}} = \frac{O}{E} \end{aligned}$$

where ...

* O is observed cases in the exposed

* E is expected cases in the exposed under the null hypothesis for stratum specific rates

– *SMR is standardized ratio when the **exposed** are the standard in terms of data distribution*

Traditional approach.

- individual standardization of rates(risks) in the exposed and unexposed

- different weights but algebraically identical formulas

- **Direct standardization**

- weight $w_i = \frac{N_{0i}}{\sum_I N_{0i}}$

- proportion of PT in stratum i in the unexposed

- *Standardized rate in the exposed*

- expected crude rate in the unexposed if the rates were the same as the exposed

$$\hat{I}_1 = \sum_I w_i \hat{I}_1 = \sum_I w_i \frac{a_i}{N_{1i}} = \sum_I \frac{N_{0i}}{\sum_I N_{0i}} \frac{a_i}{N_{1i}} = \frac{\sum_I N_{0i} \frac{a_i}{N_{1i}}}{N_0}$$

where $N_0 = \sum_I N_{0i}$

- *Standardized rate in the unexposed*

- crude rate in the unexposed

$$\hat{I}_0 = \sum_I w_i \hat{I}_0 = \sum_I w_i \frac{b_i}{N_{0i}} = \sum_I \frac{N_{0i}}{\sum_I N_{0i}} \frac{b_i}{N_{0i}} = \frac{\sum_I b_i}{N_0}$$

- *standardized rate ratio*

$$S\hat{R}R = \frac{\hat{I}_1}{\hat{I}_0} = \frac{\sum_I N_{0i} \frac{a_i}{N_{1i}}}{\sum_I N_{0i} \frac{b_i}{N_{1i}}} = \frac{\sum_I N_{0i} \frac{a_i}{N_{1i}}}{\sum_I b_i}$$

* which is equivalent to the unified approach

- **Indirect standardization**

- weight $w_i = \frac{N_{1i}}{\sum_I N_{1i}}$

* proportion of PT in stratum i in the exposed

- *Indirectly standardized rate in exposed*

$$\hat{I}_1 = \sum_I w_i \hat{I}_1 = \sum_I w_i \frac{a_i}{N_{1i}} = \sum_I \frac{N_{1i}}{\sum_I N_{1i}} \frac{a_i}{N_{1i}} = \frac{\sum_I a_i}{\sum_I N_{1i}} = \frac{O}{N_1}$$

where ... $N_1 = \sum_I N_{1i}$ and O is the observed cases in the exposed

$$O = \sum O_i = \sum a_i$$

- *Indirectly standardized rate in unexposed*

$$\hat{I}_0 = \sum_I w_i \hat{I}_0 = \sum_I w_i \frac{b_i}{N_{0i}} = \sum_I \frac{N_{1i}}{\sum_I N_{1i}} \frac{b_i}{N_{0i}} = \frac{\sum_I \frac{b_i}{N_{0i}} N_{1i}}{\sum_I N_{1i}} = \frac{E}{N_1}$$

where E is the expected cases in the exposed if the exposed had the rates of the unexposed

- *Indirectly standardized rate ratio*

$$SMR = \frac{\hat{I}_1}{\hat{I}_0} = \frac{\sum_I a_i}{\sum_I \frac{b_i}{N_{0i}} N_{1i}} = \frac{O}{E} \text{ just like the unified approach.}$$

1.3. Interpretations of standardized measures.

- *directly* standardized measures are interpreted as the measures which would be observed in any population ...
 - with similar distribution of stratification factors as the unexposed
 - when rates are similar to those in the exposed
- *indirectly* standardized measures are interpreted as the measures which would be observed in any population
 - with a similar distribution of stratification factors as the exposed
 - rates similar to those observed in the exposed
- *Comparability*

- *directly* standardized measures are comparable from study to study if the same standard population is used
- *indirectly* standardized measures are comparable only in the absence of effect modification

1.4. Relation between Causal and Traditional Notation.

- when using the entire population $\frac{\sum_i Pr[Y=1|A=1, L=l] Pr[L=l]}{\sum_i Pr[Y=1|A=0, L=l] Pr[L=l]}$ is equivalent to $\frac{\sum_I w_i \frac{a_i}{N_{1i}}}{\sum_I w_i \frac{b_i}{N_{0i}}}$
if $w_i = \frac{N_i}{\sum_I N_i}$ $(\frac{N_i}{\sum_I N_i} = Pr[L = l])$
- when using weights based on the unexposed (direct) we have the effect among the untreated
 $\frac{\sum_i \frac{a_i}{N_{1i}} N_{01}}{\sum_i b_i}$ same as $\frac{\sum_l Pr[Y=1|A=1, L=l] Pr[L=l|A=0]}{Pr[Y=1|A=0]}$
- similarly the SMR is the effect among the treated
 $\frac{\sum_i \frac{a_i}{N_{1i}} N_{1i}}{\sum_i \frac{b_i}{N_{0i}} N_{1i}}$ is the same as $\frac{Pr[Y=1|A=1]}{\sum_l Pr[Y=1|A=0, L=l] Pr[L=l|A=1]}$

1.5. Variances.

- SRR

$$Var[S\hat{R}R] = \sum_I w_i^2 Var[R\hat{R}_i] = \sum_I w_i^2 \hat{Var}[e^{\log R\hat{R}_i}] = \sum_I w_i^2 R\hat{R}_i^2 \hat{Var}[\log R\hat{R}_i]$$

$$(\hat{Var}[\log R\hat{R}_i] = \frac{1}{R\hat{R}_i^2} Var[R\hat{R}_i])$$

- SRD

$$\hat{Var}[S\hat{R}D] = \sum_I w_i^2 \hat{Var}[R\hat{D}_i]$$

1.6. Reference Populations and Exchangeability Assumptions.

Standardization in the entire population.

$$E[Y^a] = \sum_l E[Y|A = a, L = l] Pr[L = l]$$

- requires *full* conditional exchangeability $Y^{a=1, a=0} \perp\!\!\!\perp A|L$
- Standardized Risk Ratio
 - $\frac{\sum \text{expected deaths among population if all exposed}}{\sum \text{expected deaths among population if all unexposed}}$
 - $\frac{Pr[Y^{a=1}]}{Pr[Y^{a=0}]}$
 - expected number of cases in the entire population had everyone been exposed divided by the number of cases had everyone been unexposed

$$\begin{aligned}
SRR_T &= \frac{Pr[Y^{a=1}]}{Pr[Y^{a=0}]} \\
&= \frac{\sum_l Pr[Y^{a=1} = 1, L = l]}{\sum_l Pr[Y^{a=0} = 1, L = l]} \text{ (introduce L)} \\
&= \frac{\sum_l Pr[Y^{a=1} = 1 | L = l] \times Pr[L = l]}{\sum_l Pr[Y^{a=0} = 1 | L = l] \times Pr[L = l]} \\
&= \frac{\sum_l Pr[Y^{a=1} = 1 | A = 1, L = l] \times Pr[L = l]}{\sum_l Pr[Y^{a=0} = 1 | A = 1, L = l] \times Pr[L = l]} \text{ (conditional exchangeability)} \\
&= \frac{\sum_l Pr[Y = 1 | A = 1, L = l] \times Pr[L = l]}{\sum_l Pr[Y = 1 | A = 0, L = l] \times Pr[L = l]} \text{ (consistency)}
\end{aligned}$$

Standardization in the exposed.

- requires *partial* conditional exchangeability $Y^{a=0} \perp\!\!\!\perp A | L$

$$E[Y^a] = \sum_l E[Y | A = a, L = l] Pr[L = l | A = 1]$$

- Standardized Mortality Ratio - *SMR*

- $\frac{\sum \text{observed deaths among exposed}}{\sum \text{expected deaths among exposed}}$
- $\frac{Pr[Y^{a=1} | A=1]}{Pr[Y^{a=0} | A=1]}$
- observed number of cases in the exposed divided by the number of cases had the exposed been unexposed

$$\begin{aligned}
SMR &= \frac{Pr[Y^{a=1} = 1 | A = 1]}{Pr[Y^{a=0} = 1 | A = 1]} \\
&= \frac{Pr[Y^{a=1} = 1 | A = 1]}{\frac{Pr[Y^{a=0} = 1, A = 1]}{Pr[A = 1]}} \\
&= \frac{Pr[Y^{a=1} = 1 | A = 1]}{\sum_l \frac{Pr[Y^{a=0} = 1, A = 1, L = l]}{Pr[A = 1, L = l]} \times \frac{Pr[A = 1, L = l]}{Pr[A = 1]}} \text{ (introduce L)} \\
&= \frac{Pr[Y = 1 | A = 1]}{\sum_l Pr[Y = 1 | A = 1, L = l] Pr[L = l | A = 1]} \\
&= \frac{Pr[Y = 1 | A = 1]}{\sum_l Pr[Y = 1 | A = 0, L = l] Pr[L = l | A = 1]} \text{ (conditional exchangeability)} \\
&= \frac{Pr[Y = 1 | A = 1]}{\sum_l Pr[Y = 1 | A = 0, L = l] Pr[L = l | A = 1]} \text{ (consistency)}
\end{aligned}$$

- the denominator simulates the number of cases in the unexposed as if they had the covariate distribution of the exposed

Standardization in the unexposed.

$$E[Y^a] = \sum_l E[Y|A=a, L=l] Pr[L=l|A=0]$$

- requires *partial* conditional exchangeability $Y^{a=1} \perp\!\!\!\perp A|L$

- Standardized Risk Ratio - SRR_U

- $\frac{\sum_{\text{expected deaths among unexposed}}}{\sum_{\text{observed deaths among unexposed}}}$
- $\frac{Pr[Y^{a=1}|A=0]}{Pr[Y^{a=0}|A=0]}$

- the number of cases had the unexposed been exposed divided by the observed number of cases in the unexposed

$$\frac{\sum_l Pr[Y=1|A=1, L=l] Pr[L=l|A=0]}{Pr[Y=1|A=0]}$$

- the numerator simulates the number of cases in the exposed as if they had the covariate distribution of the unexposed

Standardization in a subset of the entire population.

$$E[Y^a|M=m'] = \sum_l E[Y|A=a, L=l] Pr[L=l|A=0, M=m']$$

$$Pr[Y^{a=1}=1|M=m']$$

$$= \sum_l Pr[Y^{a=1}=1, L=l|M=m']$$

$$= \sum_l \frac{Pr[Y^{a=1}=1, L=l, M=m']}{Pr[M=m']}$$

$$= \sum_l \frac{Pr[Y^{a=1}=1|L=l, M=m'] \times Pr[L=l|M=m'] \times Pr[M=m']}{Pr[M=m']}$$

$$= \sum_l Pr[Y^{a=1}=1|L=l, M=m'] \times Pr[L=l|M=m']$$

$$= \sum_l Pr[Y^{a=1}=1|A=1, L=l, M=m'] \times Pr[L=l|M=m'] \text{ (conditional exchangeability)}$$

$$= \sum_l Pr[Y=1|A=1, L=l, M=m'] \times Pr[L=l|M=m'] \text{ (consistency)}$$

2. G-formula

2.1. definitions for estimating the causal effect of time-varying exposure.

- for real situation, use the DAG with assumptions on associations between variables
- assuming a closed cohort with no right censoring
- using “time since baseline” as time scale
- only consider exposures occurring at or after baseline
- fixed exposure
 - (1) only occur at the baseline
 - (2) do not change over time
 - (3) evolve over time in a deterministic way
- time-varying exposure
- time-dependent confounder: an independent predictor of both subsequent exposure and the outcome within strata jointly determined by baseline covariates and prior exposure ($A_{t+1} \leftarrow L_t \rightarrow Y$)
- the conventional methods would fail if ...
 - (1) time-dependent confounders present
 - (2) within strata of the baseline covariates, baseline exposure predicts the subsequent time-dependent confounders
- three g-methods will be used to deal with time-varying exposures and time-dependent confounders:
 - (1) g-computation algorithm formula (g-formula)
 - (2) inverse probability of treatment weighting (IPTW) of marginal structural models (MSMs)
 - (3) g-estimation of structural nested models (SNMs)

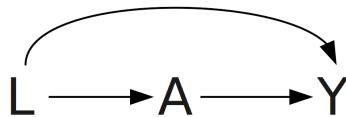
all 3 methods will give identical results **non-parametrically**

the method of choice depends on the causal contrast of interest and the robustness of the method to model misspecification

2.2. g-computation algorithm formula (g-formula).

- we may get the probability of Y^a from a causal DAGs with causal Markov assumption following the *g-computation algorithm formula*
- step by step ...
 - (1) draw the temporally ordered complete DAG

FIGURE 11.1. temporally ordered complete causal DAG

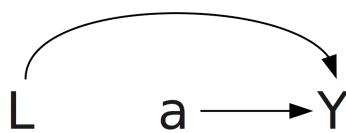


(2) write out the **joint probability** of all variables in the DAG

$$f(Y, A, L) = f(Y|A, L)f(A|L)f(L)$$

(3) draw the reduced, intervention DAG

FIGURE 11.2. intervention DAG



(4) reduce the **joint probability** of all variables in the DAG according to the intervention DAG

$$f^a(Y, a, L) = f(Y|a, L)f(L)$$

(5) sum over the nuisance parameter (usually $L_{(t)}$) and get $f(Y^a = y)$

$$f(Y^a = y) = f^a(Y = y) = \sum_l f(Y = y|a, l)f(l)$$

- Theorem

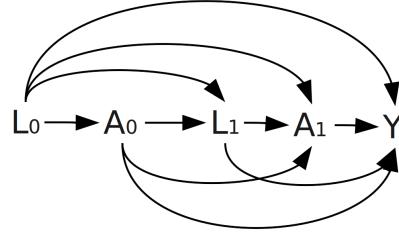
$$\text{if } Y^a \perp\!\!\!\perp A|L \text{ then } f^a(Y = y) = f(Y^a = y)$$

$$\begin{aligned} f^a(Y = y) &= \sum_l f(Y = y|a, l)f(l) \\ &= \sum_l f(Y^a = y|a, l)f(l) \text{ (consistency)} \\ &= \sum_l f(Y^a = y|a, l)f(l) \text{ (conditional exchangeability)} \\ &= \sum_l f(Y^a = y) \end{aligned}$$

- for time-varying exposure . . .

- (1) draw the temporally ordered complete DAG

FIGURE 11.3. temporally ordered complete causal DAG

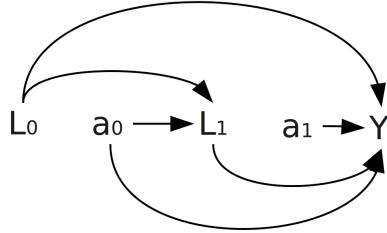


- (2) write out the **joint probability** of all variables in the DAG

$$f(Y, A_1, L_1, A_0, L_0) = f(Y|A_1, L_1, A_0, L_0)f(A_1|L_1, A_0, L_0)f(L_1|A_0, L_0)f(A_0|L_0)f(L_0)$$

- (3) draw the reduced, intervention DAG

FIGURE 11.4. intervention DAG



- (4) reduce the **joint probability** of all variables in the DAG according to the intervention DAG

$$f^{a_0, a_1}(Y, a_1, L_1, a_0, L_0) = f(Y|a_1, L_1, a_0, L_0)f(L_1|a_0, L_0)f(L_0)$$

- (5) sum over the nuisance parameter and get $f(Y^{a_0, a_1} = y)$

$$f(Y^{a_0, a_1} = y) = f^{a_0, a_1}(Y = y) = \sum_{l_0, l_1} f(Y|a_1, l_1, a_0, l_0)f(l_1|a_0, l_0)f(l_0)$$

3. Inverse Probability Weighting (IPW)

3.1. Inverse Probability Weights (W).

$$\frac{1}{Pr[A = a|L = l]} \quad \text{or} \quad \frac{1}{f[A|L]} \quad \text{or} \quad E[Y^a] = E \left[\frac{I(A = a)Y}{f[a|L]} \right]$$

- each person in the original population is weighted by the inverse probability of receiving the treatment *they actually received* given the covariates needed to identify the causal effect (i.e $1/(propensity\ score)$ for treated subjects and $1/(1 - propensity\ score)$ for untreated subjects) to create conditional exchangeability in a new pseudo-population
 - assumes conditional exchangeability can be obtained via the *measured* covariates L
 - IPW cannot create conditional exchangeability in the pseudo-population when there is confounding by unmeasured variables U ($A \leftarrow U \rightarrow Y$)
- simulation of what would have happened had all subjects in the population been treated and untreated *and* had the same distribution of L as the standard population
 - pseudo-population: hypothetical population created by applying IP weights to the observed population
 - * size of the pseudo-population is the size of the original population times the number of treatment categories (for stabilized weights size is the same as original population)
 - * the treated and untreated are unconditionally exchangeable because they are the same individuals under different treatment levels
 - * associational RR in pseudo-population equals causal RR in the original population
- analytic control: adjust for L , linguistically different from physically controlling for L in design phase
 - removes the arrow from L into A
- Non-parametric IPW and standardization are equivalent

Let's assume that A , L , and Y are all binary variables. The IP weighted risk of Y for treatment group $A = a$ is $E \left[\frac{I(A=a)Y}{f[a|L]} \right]$. The standardized risk of Y for treatment group $A = a$ is $\sum_l Pr[Y = 1|A = a, L = l]Pr[L = l]$. To show that these two things are equal, first we need to put that $E \left[\frac{I(A=a)Y}{f[a|L]} \right] = E \left[\frac{I(A=a)Y}{f[A|L]} \right]$. Then, we may write down explicitly how they are equivalent:

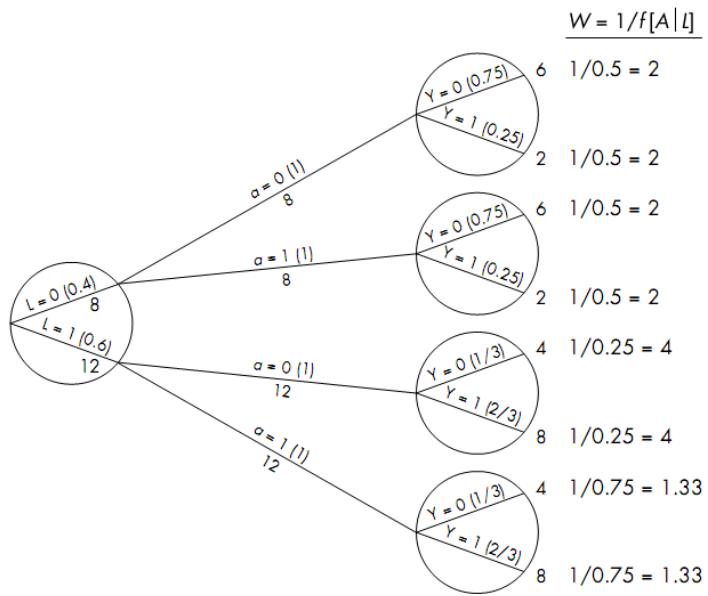


FIGURE 11.5. unstablized IP weighting creates a psuedo-population which is twice as large as the original population for dichotomous exposures

$$\begin{aligned}
E \left[\frac{I(A=a)Y}{f[A|L]} \right] &= E \left[\frac{I(A=a)}{f[A|L]} \times Y \right] \\
&= \frac{I(A=1)}{f[A|L=1]} \times f[Y, A, L=1] \times Y + \frac{I(A=1)}{f[A|L=0]} \times f[Y, A, L=0] \times Y \\
&= \frac{I(A=1)}{f[A|L=1]} \times f[Y=1, A, L=1] \times 1 + \frac{I(A=1)}{f[A|L=0]} \times f[Y=1, A, L=0] \times 1 \\
&+ \frac{I(A=1)}{f[A|L=1]} \times f[Y=0, A, L=1] \times 0 + \frac{I(A=1)}{f[A|L=0]} \times f[Y=0, A, L=0] \times 0 \\
&= \frac{I(A=1)}{f[A|L=1]} \times f[Y=1, A, L=1] + \frac{I(A=1)}{f[A|L=0]} \times f[Y=1, A, L=0] \\
&= \frac{I(A=1)}{f[A|L=1]} \times \frac{f[Y=1, A, L=1]}{f[A, L=1]} \times \frac{f[A, L=1]}{Pr[L=1]} \times Pr[L=1] \\
&+ \frac{I(A=1)}{f[A|L=0]} \times \frac{f[Y=1, A, L=0]}{f[A, L=0]} \times \frac{f[A, L=0]}{Pr[L=0]} \times Pr[L=0] \\
&= \frac{I(A=1)}{f[A|L=1]} \times f[Y=1|A, L=1] \times f[A|L=1] \times Pr[L=1] \\
&+ \frac{I(A=1)}{f[A|L=0]} \times f[Y=1|A, L=0] \times f[A|L=0] \times Pr[L=0] \\
&= f[Y=1|A, L=1] \times Pr[L=1] + f[Y=1|A, L=0] \times Pr[L=0] \\
&= Pr[Y=1|A=1, L=1] \times Pr[L=1] + Pr[Y=1|A=1, L=0] \times Pr[L=0] \\
&+ Pr[Y=1|A=0, L=1] \times Pr[L=1] + Pr[Y=1|A=0, L=0] \times Pr[L=0] \\
&= \sum_l Pr[Y=1|A=a, L=l] Pr[L=l]
\end{aligned}$$

Thus, we can show that, at least under this simplified situation, the IP weighted risk and standardized risk are equivalent.

$$\begin{aligned}
 f[Y, A, L] &= \underbrace{f[Y|A, L] \times f[A|L] \times f[L]}_{\text{Markov Factorization}} \\
 \frac{f[Y, A, L]}{f[A|L]} &= f[Y|A, L] \times f[L] \\
 \underbrace{\sum_l \left(\frac{f[Y, A, L]}{f[A|L]} \right)}_{\text{IPW Estimator}} &= \underbrace{\sum_l (f[Y|A, L] \times f[L])}_{\text{G-formula/Standardization}} \\
 \frac{f[Y, A]}{f[A]} &= f[Y|A] \\
 f[Y|A] &= f[Y|A]
 \end{aligned}$$

FIGURE 11.6. The mathematical equivalence of IPW and standardization

- in *figure 11.6* note that each method perturbs the joint distribution to make the term $f[AL]$ equal $f[A] = 1$
- that is, they simulate the data observed if $A = 1$ for all persons by making A independent of L while preserving all other relationships in the data

3.2. Stabilized Weights (SW).

$$SW = \frac{f[A]}{f[A|L]}$$

- more efficient than unstabilized weights (W)
- SW can be used wherever W is used, except with dynamic treatment regimes in which case W must be used
- for a static regime $\bar{a} = \{a_0, a_1\}$

$$SW = \frac{f[A_0]f[A_1|A_0]}{f[A_0]f[A_1|A_0, L_1]} = \frac{f[A_1|A_0]}{f[A_1|A_0, L_1]}$$

- it is important to include $f[A_0]$ in IPW weight W for dynamic regimes
- *fine point*:
 - the denominator of IPW weight removes the association between A and whatever is in the conditioning event (i.e. $1/f[A|L]$ removes $L \rightarrow A$)
 - the numerator of stabilized weight restores the association between A and whatever is in the conditioning event (i.e. $f[A|L_0]/f[A|L_0, L_1]$ removes $L_1 \rightarrow A$ but not $L_0 \rightarrow A$)

- variables should be in conditioning event of numerator **only if** they are present in the conditioning event of the denominator

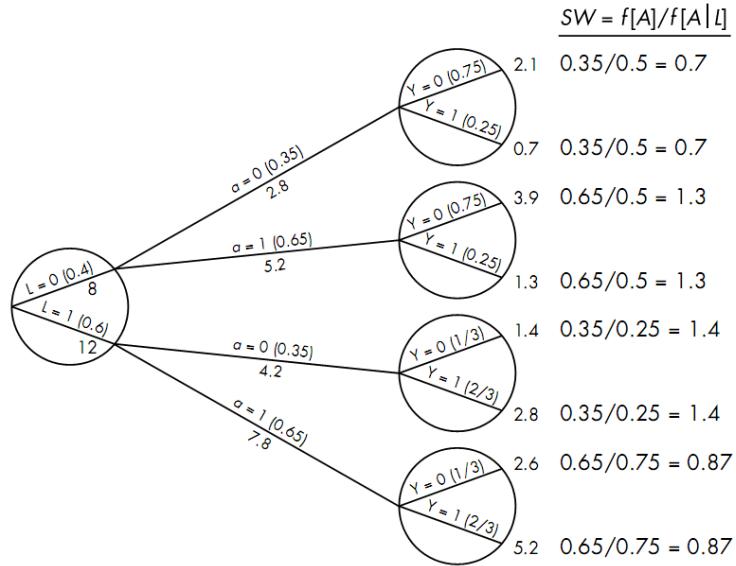


FIGURE 11.7. Stabilized IP weighting creates a psuedo-population which has the same size as the original population

3.3. Censoring Weights (CW).

$$CW = \frac{1}{f[C = 0|A, L]}$$

$$SCW = \frac{f[C = 0|A]}{f[C = 0|A, L]}$$

- the main idea is to eliminate the path (conditioned collider) that was unblocked by censoring (figure 11.8)

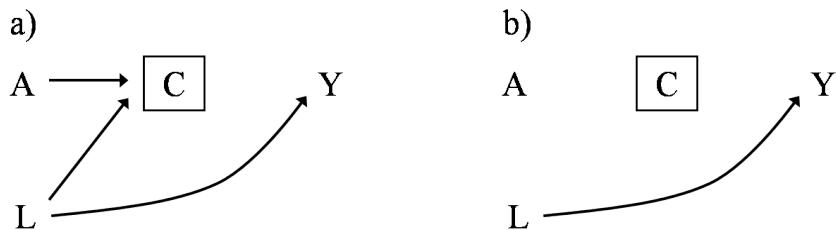


FIGURE 11.8. a) Study with selection bias due to censoring, b) Study population after censoring IP weights are applied

- we want to simulate what would have happened had everyone's outcome been observed ($Y^{a,c=1}$)

- underlying assumption: $Y^c \perp\!\!\!\perp C|A, L$
 - censoring is random within levels of A and L
 - no unmeasured variables U affecting censoring
 - within levels of A and L , the $Pr[C = 1] = Pr[C = 0]$
- censored individuals have $w = 0$ and non-censored individuals have $w = CW$
- IP weights can be created that account for *both* confounding and censoring
 - simply multiply the standardized weight for confounding by the standardized weight for censoring

$$\frac{f[A]}{f[A|L]} \times \frac{f[C = 0|A]}{f[C = 0|A, L]}$$

- L is not necessarily the same for the confounding and censoring weights and is determined by the structure of the causal diagram

3.4. SMR.

$$SMW = \frac{f[A = a'|L]}{f[A|L]}$$

- for estimation of causal RR within a subset of population (i.e. SMR for exposed group) must use modified weights

3.5. Effect modification.

4. G-estimation

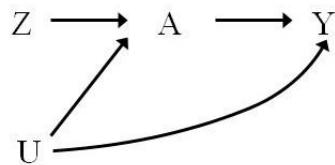
- Null Hypothesis
 - null hypothesis of no average causal effect: $E[Y^{a=1}] = E[Y^{a=0}]$
 - sharp causal null hypothesis: $Y_i^{a=1} = Y_i^{a=0} = Y$ for all subjects i .
- G-null test
 - A test of the sharp null hypothesis under the assumption that conditional exchangeability holds
 - Conditional exchangeability $Y^a \perp\!\!\!\perp A|L = l$ can be equivalently expressed as $A \perp\!\!\!\perp Y^a|L = l$
 - We can model exposure on outcome $\text{logit}\{Pr[A = 1|L, Y^{a=0}]\} = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L$
 - Then we test whether $\alpha_1 = 0$ in the model to see if sharp null hypothesis is true
- Structural mean model
 - $E[Y^a] = \beta_0 + \beta_1 a$

- This model is saturated marginal structural mean model
- This model can also be referred to as a structural nested mean model
- Structural nested model
 - Structural nested model is composed of two models
 - Structural model $E[Y^a] = \beta_0 + \beta_1 a$
 - Exposure model $\text{logit}\{Pr[A = 1|L, Y^{a=0}]\} = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L$
 - Called nested model because structural model can be nested in the exposure model
- G-estimation
 - A method to estimate the parameters of structural nested models
 - Let $H(\psi)$ be the counterfactual outcome, then the relation between $H(\psi)$ and treatment (A) can be written as $H(\psi) = Y - \psi A$
 - One could then test that the coefficient of $H(\psi)$ to be zero in a model for the regression of A on $H(\psi)$ (exposure model).
 - Grid search intervals of value of ψ
 - * The coefficient that has a P-value equals 1 is the valid parameter estimate, and the set of all coefficients that have a p-value greater than α form a $1 - \alpha$ confidence interval.
 - model:
 - $\text{logit}\{Pr[A = 1|L, H(\psi)]\} = \alpha_0 + \alpha_1 H(\psi) + \alpha_2 L$
 - $H(\psi) = Y^{a=0}$ if $\psi = \psi^*$ where ψ^* is the true value of ψ
 - ψ^* with 95 percent confidence interval has the causal interpretation for the effect of A on the counterfactual outcome
 - $H(\psi)$ can be transformed to several function forms such as log or square root to increase the efficiency of g-estimation process

5. Instrumental Variables

5.1. Instrumental Variables.

- instrumental variables can be used to estimate causal effects without exchangeability
- an instrumental variable Z
 - has a causal effect on exposure A , or shares a common cause with A
 - affects outcome Y only through A
 - there is no confounding on the effect of Z on Y



(a) causal effect of Z on A

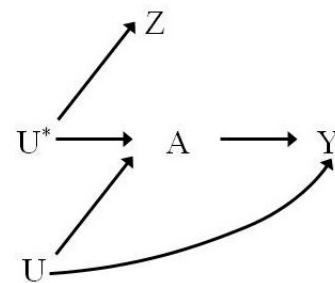
(b) common cause U^* of Z and A

FIGURE 11.9. Instrumental Variables

- examples of instrumental variables
 - cigarette prices is an instrumental variable Z, for the exposure of cigarette smoking A, and various health outcomes Y (figure 1a)
 - self reported lactose intolerance (Z) and dietary calcium intake (A) have a common cause in the lactose intolerance gene (U^*) with respect to osteoporosis (Y) (figure 1b)
- standard IV estimator $E[Y^{\alpha=1}] - E[Y^{\alpha=0}] = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[A|Z=1] - E[A|Z=0]}$
- may use IP weights for censoring before applying IV estimator

5.2. Limitations of IV.

5.2.1. Unverifiable assumptions.

- an instrument cannot be verified as one
 - non-instruments introduce bias

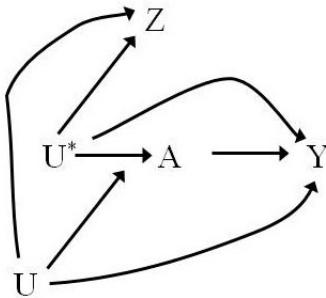


FIGURE 11.10. Instrumental Variables

- weak instruments blow up bias (weak association of Z and A leads to small denominator in the IV estimator)
- standard IV methods do not deal well with time-varying exposures
- need additional unverifiable assumptions to estimate causal effects
the IV estimator is a valid point estimate of causal effect under additional assumptions
without these assumptions an IV can only be used to calculate bounds for a causal effect
- additional assumptions are
 - constant treatment effect across all subjects
 - no interaction between instrument and exposure (on either the additive or multiplicative scale)
- neither of these assumptions can be verifiable
- can be replaced by the assumption of monotonicity

5.2.2. Non-compliance in randomized experiments.

- randomized experiments with full compliance assure the 3 identifiability assumptions of exchangeability, consistency and positivity
- if noncompliance is not random, then exchangeability achieved with effective randomization is lost
- random noncompliance leads to decreased power
- noncompliance can also lead to violations of positivity when randomization is done taking into account covariates L

intention to treat and IV.

- treatment assignment is instrumental variable Z

- causal effect of treatment A on outcome Y given by

$$E[Y^{a=1}] - E[Y^{a=0}] = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[A|Z=1] - E[A|Z=0]}$$
the denominator equals 1 if full compliance
with non-compliance the intention to treat effect in the numerator is inflated

5.2.3. Monotonicity.

- monotonicity is an alternative assumptions for use of instrumental variables to those of constant effect of treatment and no effect modification
- 4 types of people in trial with randomized treatment (A) assignment (Z)
 - always takers, $A^{Z=0} = 1, A^{Z=1} = 1$
 - never takers, $A^{Z=0} = 0, A^{Z=1} = 0$
 - compliers, $A^{Z=0} = 0, A^{Z=1} = 1$
 - defiers, $A^{Z=0} = 1, A^{Z=1} = 1$
- monotonicity if no defiers
- instrumental variable of treatment assignment measures effect in compliers
- problems
 - we cannot identify the compliers
 - in continuous exposures everyone is a 'complier'

5.3. IV as special case of g-estimation.

- standard IV does not work for time-varying exposures
- have to use g-estimation of SNMs
- using the simple $H(\psi)$ model we can do IV analysis by fitting the model

$$\text{logit}Pr[Z = 1|H(\psi)] = \alpha_0 + \alpha_1 H(\psi)$$
- advantages of g-estimation IV analysis
 - can use baseline covariates increasing efficiency and allow richer dose response function
 - use multiplicative model and assumption of no multiplicative scale interaction
 - 95% CIs
 - extension to time-varying exposures

Part 4

Inference with Models: Concepts & Methods

CHAPTER 12

Modeling Concepts

1. Modeling Concepts in Epidemiology

1.1. Estimators.

- *Estimand*: the unknown population parameter, what we are trying to find out (E)
- *Estimate*: the result of applying the estimator to a particular data set (\hat{E})
- *Estimator*: an algorithm or rule that when applied to the data estimates a parameter of interest
 - *Consistent*: $\lim_{n \rightarrow \infty} \hat{\theta} = \theta$
 - *Unbiased*: in repetitions of the study the 95% CI contain the true value 95% of the time
 - *Biased*: in repetitions of the study the 95% CI does not contain the true value 95% of the time
 - * systematically deviate from the true value of the effect measure
- *Bias of Estimator*: the average deviation of the estimator from the true value of the effect measure over study repetitions

$$E[\hat{\theta} - \theta]$$

- *Statistical Model*: a mathematical expression for a set of assumed restrictions on the possible state of nature
- *Nuisance Parameter*: any parameter (not of interest) that must be estimated in order to estimate the parameter of interest
 - estimating these results in a loss of power for estimating the parameter of interest

1.2. Model Classes & Types.

Broad Model Classes.

- *Conditional*: parameters refer to subgroups of the population with particular covariate values
- *Marginal*: parameters refer to the entire population under different covariate values
- *Structural*: parameters have a causal interpretation

- *Associational*: parameters only describe statistical relationships and may not have a causal interpretation
- *Nested*: a model can be expressed as a simplification of a larger, more complicated model,
 - the result of placing constraints on the larger model

Inference vs. Prediction.

- *Prediction*
 - only worry about validity of predicted value
 - * no requirement for conditional exchangeability, positivity, or consistency
 - concerned more about
 - * overall model efficiency
 - * transportability (external validity)
 - concerned less about validity of parameter estimates
 - covariate inclusion driven by efficiency, transportability, and clinical utility
 - * often use algorithms (stepwise or data-mining)
- *Causation* (Inference)
 - only worry about validity of parameter estimate
 - * require conditional exchangeability, positivity, and consistency
 - concerned less about efficiency of overall model
 - may or may not be concerned about transportability (external validity)
 - covariate inclusion driven by subject-matter knowledge
 - **only the parameter for exposure has a causal interpretation**
 - * modeling decisions made to meet identifiability conditions for that parameter only
 - * other parameters may only have associational interpretation
 - * separate modeling process, covariate choice, etc. for *each* exposure of interest
 - model should *not* be evaluated with goodness of fit tests
 - (1) the number of good fitting models will be large
 - (2) these models will have a wide range for the effect estimate
 - (3) the range of effect-estimates from well fitting models may have a wider range than the model-specific standard errors

Incorporating a priori information.

- *Non-Parametric:*
 - a function of the data only (no distributional assumptions)
 - allows the data to "speak for itself"
 - synonym: *saturated*
 - non-parametric estimators can be expressed in a model form if ...
 - * $\#$ parameters = $\#$ possible values for dependent variable
 - i.e. the $\#$ of quantities we want to estimate
 - * includes all main effects of covariates and all possible interactions between them
- *Semi-Parametric:*
 - distributional assumptions for *some* of the parameters
 - the unspecified parameters are allowed to follow any distribution
 - * subtle point: while the parameter itself may be unspecified, its estimation still requires the use of an estimator or algorithm
 - ex. cox-proportional hazards, splines, etc.
- *Parametric:*
 - all parameters have distributional assumptions
 - function of data and a priori information
 - * if assumptions correct, avoids estimating nuisance parameters and increases power for estimating parameter of interest
 - synonym: unsaturated

General Classes of Models.

- *Generalized Additive Model:*
 - example: $E[Y|X_1, X_2, \dots, X_p] = \alpha + f_1(X_1) + f_2(X_2) + \dots + f_p(X_p)$
 - response variable is a sum of functions of the covariates
 - not necessarily linear
 - any functions suitable (linear, non-parametric, moving average, locally weighted mean, etc)
 - * example: smoothing functions (e.g. kernel, spline, etc.)
- *Generalized Linear Models:*
 - example: $E[Y|X_1, X_2, \dots, X_p] = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)$

- the response variable is a function of a linear combination of the covariates via their *parameters*
- linearity means ...
 - * linear (additive) function of the *parameters* on a particular scale (absolute, logarithmic, etc.)
 - * the covariates themselves do not have to be linear functions
 - e.g. $x^2 \rightarrow$ still a linear model
 - e.g. $\exp(\beta) \rightarrow$ still a linear model
 - e.g. $\beta^2 \rightarrow$ not a linear model
- Need to also specify
 - (1) functional form (*Link function*: logit, log-linear, etc.)
 - (2) statistical distribution of the residuals (normal, binomial, Poisson, etc.)
 - includes linear, logistic, Poisson, etc.
 - these assumptions can be relaxed (GEE & random effects)
- *Random-Effects Models*
 - synonyms: hierarchical, multilevel, growth-curve
- *Models for failure-time data*
 - i.e. survival analysis
 - parametric approaches: exponential, weibull
 - semiparametric approaches: Cox proportional hazards model, Accelerated failure time model

2. Why we need models

2.1. Model Specification.

- the assumptions made in the modeling strategy should reflect one's prior beliefs
- causal inference requires the condition of *no model misspecification*

2.2. The curse of dimensionality.

- (1) occurs when we need to control for many variables in order to satisfy identifiability conditions
 - ex. for binary exposure, outcome and 10 binary covariates, the saturated model has $2^{11} = 2048$ parameters
 - even if assume no effect modification for the effect of exposure there are $\rightarrow 2^{10} + 1 = 1025$ parameters

- # parameters equivalent to # 2x2 tables in stratified analysis
- we only care about the parameter for exposure, the rest are nuisance parameters
 - very few strata likely to be informative (i.e. exposed and unexposed subjects)
 - without assumptions one could learn very little about the effect of exposure in any schedule
- (2) in extreme cases the effects of exposure cannot be separated from the effects of covariates
 - stratum level and exposure status are almost perfectly associated in the data
 - i.e. lack of positivity prohibits identifiability of causal effect

2.3. Supplementing the Data with A priori information.

- *Sparse Data*: highly stratified data, but can obtain useful confidence intervals (e.g. matched pair analysis)
- *Weak Data*: no unbiased estimator for the parameter of interest has a variance small enough to construct useful confidence intervals (even with no unmeasured confounding)
 - weak data requires modeling assumptions
 - * weak data = sparse data always
 - * sparse data may or may not = weak data
- **modeling assumptions supplant the data with a priori information** →
 - restricts some or most of the nuisance parameters to equal zero
 - * hence reduces number of strata & number of parameters to estimate
 - *increases efficiency* because more strata with larger numbers of exposed and unexposed subjects
 - parameter for exposure will have narrower confidence interval
- usually prior beliefs for risk factors of disease are more "sharp" than beliefs for predictors of exposure
- recall that a confounder has a structural definition based on subject matter knowledge
 - its inclusion into the model should not be subject to a statistical test

2.4. Bias and efficiency trade-off.

- saturated vs. reduced model
 - (1) a highly stratified model will have small bias (provided no unmeasured confounding) but effect estimate may have larger variance

- (2) a reduced model is more likely to be misspecified (and thus biased) but effect estimate has smaller variance
- (3) the savings in variance afforded by modeling assumptions should offset the increase in bias that would result if those assumptions were incorrect
 - true magnitude of such bias is unknown
- (4) if one's assumptions are badly contradicted by the data one should be willing to give them up
- (5) if statistical tests are used to verify assumptions, the α level should depend in part on the strength of one's beliefs
 - stronger beliefs: more liberal rejection criteria \rightarrow higher α
 - stronger beliefs: more restrictive rejection criteria \rightarrow lower α
- weak data refers to the kind of data that we cannot gain any information on the parameter of interest because the variance of the estimate was too large
- mean square error is the sum of variance of the estimate and the square of the bias
- saturated model \Rightarrow less bias, larger variance (less efficient)
- reduced model \Rightarrow more bias, smaller variance (more efficient)
- do not know whether saturated or reduced model has smaller mean square error

3. Stratification-based models

- most common adjustment method in practice (ex: linear and logistic regression)
- exposure and confounders are included as covariates in the model for outcome
- stratification (standard regression) vs. g-methods:
 - identifying assumptions (exchangeability, positivity, consistency) the same
 - no model for exposure in standard regression
 - model for the outcome is *conditional* on all covariates in standard regression
 - standard regression cannot estimate the average causal effect in the whole population, unless the assumption of no effect modification is made
 - standard regression has to adjust for effect modification because of conditional nature of the model (otherwise the model is misspecified)
 - g-methods only incorporate effect modification if it is relevant to the research question

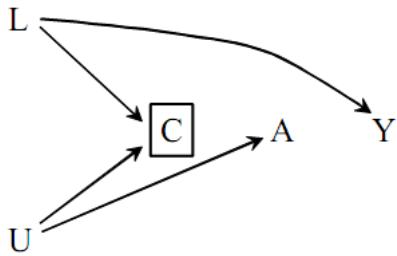


FIGURE 12.1. Fireman example where A =physical activity, Y =CHD, C =being a firefighter, L =parental SES, and U =attraction toward activities requiring physical activity

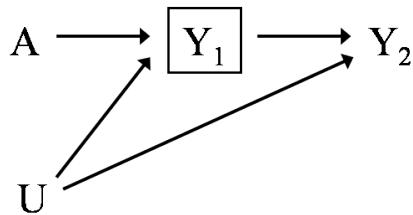


FIGURE 12.3. DAG depicting hazard ratio where A = surgery, Y_1 = death at $t = 1$, Y_2 = death at $t = 2$, and U = protective haplotype

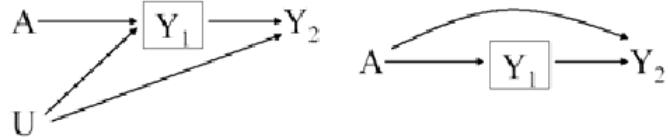


FIGURE 12.2. cannot distinguish between the two structures using hazard ratio as effect estimate

- advantages of structural classification of bias:
 - (1) structure of the problem guides choice of analytical methods
 - (2) structure may aid in study design, even when structure has no implications for data analysis, ex: firefighter study (figure 12.1)
 - (3) selection bias from conditioning on pre-exposure variables may help explain why covariates behave as confounders in one study and not another (m-bias)
 - (4) causal DAGs aid in communication among investigators

3.1. Biases in Stratified Analysis.

- a variable that affects survival at $t = 1$ always affects survival at $t = 2$
 - effect of A on the risk of Y on the *risk ratio scale*:
 - * $t = 1$: $Pr[Y_1 = 1|A = 1]/Pr[Y_1 = 1|A = 0] > 1$, assuming A is causative

- * $t = 2$: $Pr[Y_2 = 1|A = 1]/Pr[Y_2 = 1|A = 0] > 1$, even though A does not have a direct effect on Y_2
- * association of A and Y_2 arises from indirect effect of A through Y_1 (figure 12.3)
- * risk ratio is a measure of risk of A on total mortality through time t (cumulative risk)
- effect of A on the risk of Y on the *hazard ratio scale*:
 - * hazard ratio
 - hazard: probability of dying between time t and $t + 1$, given you were alive at time t , $Pr[Y_{t+1} = 1|Y_t = 0]$
 - hazard at $t=1$: $Pr[Y_1 = 1]$
 - hazard at $t=2$: $Pr[Y_2 = 1|Y_1 = 0]$
 - as opposed to hazards, risk is cumulative
 - * $t = 1$: $Pr[Y_1 = 1|A = 1]/Pr[Y_1 = 1|A = 0] > 1$
 - * $t = 2$: $Pr[Y_2 = 1|Y_1 = 0, A = 1]/Pr[Y_2 = 1|Y_1 = 0, A = 0] \neq 1$
 - * association between A and Y_2 is opened by conditioning on a collider (Y_1), figure
 - * hazard ratio neither measures the direct effect of A on Y_2 or the effect of A on total mortality through time $t = 2$
 - * crossing hazards:
 - $t = 1$: $Pr[Y_1 = 1|A = 1]/Pr[Y_1 = 1|A = 0] > 1$
 - $t = 2$: $Pr[Y_2 = 1|Y_1 = 0, A = 1]/Pr[Y_2 = 1|Y_1 = 0, A = 0] < 1$
 - after each successive time of follow-up, individuals more resistant against developing the outcome are selected for (*survival bias* where there is depletion of the susceptible individuals in the population)

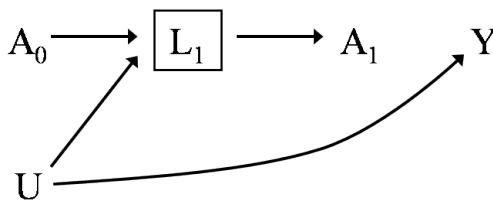


FIGURE 12.4. DAG depicting time varying confounder affected by treatment where A_t = ART at $T = t$, Y = viral load, L_t = CD4 count at $T = t$, and U = immunosuppression level

- time varying confounders L affected by exposure A

- ex: the effect of a time varying treatment ($A = A_0 + A_1 = 0, 1$, or 2) on outcome (figure 12.4)
- necessary to block all back-door paths for A , that is for A_1, A_2, \dots
- eliminates confounding for one component, but creates selection bias for another
- to stratify, or not to stratify
 - * not stratifying results in bias due to confounding
 - * stratifying introduces selection bias
- need g-methods to **remove arrows**, rather than stratification methods that put a **box around** (i.e. *condition on*) the covariate

3.2. Parameter interpretation for associational models.

3.3. Parameter interpretation for structural models.

4. Model Specification

5. Hypothesis Testing & Estimation

5.1. hypothesis testing for β .

- Wald test for $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ (H_0 : rate ratio = 1 vs H_1 : rate ratio $\neq 1$)

$$\frac{\hat{\beta}^2}{\widehat{Var}(\hat{\beta})} \sim \chi_1^2$$
- likelihood ratio test: $-2 \log \left[\frac{L(\beta)_{red}}{L(\beta)_{full}} \right] \sim \chi^2$
 - use the same data set
 - use nested model
 - the degree of freedom of the test is the difference between number of variables in two models
- test statistics is a function of sample size, study design, and the evidence in the data supporting the hypothesis under test

5.2. 95% confidence interval for β .

$$e^{\log(RR) \pm z_{1-\alpha/2} \sqrt{\widehat{Var}[\log(RR)]}} = e^{\Delta \times \{\hat{\beta} \pm z_{1-\alpha/2} \widehat{SE}(\hat{\beta})\}}$$

where Δ is the the increment to be used for comparison with a continuous variable ($\Delta = 1$ for binary variable)

- e.g. $e^{10 \times \hat{\beta}}$ is the log rate ratio of death for each 10 -year increase in age

- we make the assumption that the rate vary linearly on the log scale with age (log-linear)

6. Interaction in Regression Models

6.1. Main Points.

- effect modification can be modeled in two ways
 - by modeling a variable that defines all strata formed by the interacting variables
 - by adding product terms of two or more variables in the model
- the interaction term of continuous variable implies linearity assumption about the effect modification on the scale of the model (log scale for Cox model and Poisson model, log scale of odds for logistic model)
- the presence of effect modification can be tested
 - by testing the regression coefficient of the product term in the model using Wald test or likelihood ratio test
 - likelihood ratio test has better finite sample property than Wald test
 - one could also do a test of homogeneity (i.e. $\beta_i = \beta_j$)
- the magnitude can be informally examined by comparing the relative size of the parameters on the effect measure's scale
 - $\exp(\beta_1) vs. \exp(\beta_1 + \beta_3)$

6.2. Interpretation.

- logistic, Poisson, cox, and conditional logistic are all multiplicative models
 - in such models the baseline, or absolute risk, odds, or rate for the reference group is multiplied by the *relative* effect of being in the comparison group

let g = outcome or time to outcome
 α = some arbitrary baseline function

$$g(x) = \alpha \times \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

$$= \alpha \times e^{\beta_1 X_1} \times e^{\beta_2 X_2} \times \cdots \times e^{\beta_p X_p}$$

- thus an interaction represents a departure from (equivalent)
 - additive effects on the log scale
 - departure from purely multiplicative effect of *only* exposure

$$g(x) = \alpha \times \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2)$$

$$g(x) = \alpha \times e^{\beta_1 X_1} \times e^{\beta_2 X_2} \times e^{\beta_3 X_1 X_2}$$

- inferring interaction requires identifiability conditions to hold for both variables, otherwise we only infer effect-modification
- with multiplicative models, inferring interaction on the additive scale or sufficient cause interaction requires special conditions and assumptions
 - you can infer effect modification on the additive scale if it is qualitative (change in sign)
 - if you can infer effect modification on the additive scale, you can further infer sufficient cause interaction if you assume monotonicity
- studies often underpowered to detect effect (measure) modification or interaction
- unclear whether observed heterogeneity is real or only due to chance
 - should be considered on the basis of plausibility, magnitude of effect, and random variation
 - likelihood ratio tests often better for assessing effect modification
- unlike stratification, using regression models allows you to assume that there is either
 - interaction between all variables in model (saturated model)
 - interaction between only some variables
- the proportional hazards assumption can be assessed by examining if the hazard ratio for exposure varies over time. Tested by...
 - reclassifying exposure according to time periods and see if period specific parameters are different
 - assessing whether the product term between exposure and the time-scale is significant
- assessing effect modification in conditional logistic regression follows same logic
 - reclassifying exposure according to matching factor level and see if period specific parameters are different
 - assessing whether the product term between exposure and the matching factor is significant

6.3. Review of Interaction terms.

- fitting separate models for each level of some external variable Z is analogous to assuming interaction by Z in a unified model
- for continuous variables, the effect of X depends on the *level* of Z

- to interact categorical variables without assuming trend, need to interact the 'dummy' variables

$$g(x, z) = \alpha + \beta \sum_{i=1}^I I(X = i) + \beta \sum_{j=1}^J I(Z = j) + \beta \sum_{i=1}^I \sum_{j=1}^J I(X = i) \times I(Z = j)$$

6.4. Assessment of Interaction using Estimation.

- For all models let us assume
 - X_1 and Z_2 are two binary covariates and Y is outcome of interest
 - $\ln(\alpha)$ is the outcome for the baseline reference group
 - $I(\text{argument}) = 1$ if true, and $= 0$ if false

Wald Statistic.

- if β is the OR comparing $X = 1$ to $X = 0$ for a given level of Z
and if $\beta_{z=1}$ and $\beta_{z=0}$ are uncorrelated then...

$$\text{Wald Statistic} = \frac{(\beta_{z=1} - \beta_{z=0})^2}{\text{Var}(\beta_{z=1}) + \text{Var}(\beta_{z=0})} \sim \chi_1^2 \quad \text{where...}$$

$$\text{Var}(\beta) = \frac{(\log e^\beta - \log(\text{lower bound}))}{(Z_{1-\alpha/2})^2}$$

- this is analogous to a test-based confidence interval
 - the variance is under H_0 but should be under H_A

6.5. Assessment of Interaction using Hypothesis Testing.

Heterogeneity.

- conceptually analogous to *effect modification*
- $RR(X)_{z=1} \neq RR(X)_{z=0}$
- in regression...

$$\begin{aligned} \text{Logit} Pr[Y = 1 | X = x, Z = z] &= \alpha \exp(\beta_1 I(X = 1) \\ &\quad + \beta_2 I(Z = 1) + \beta_3 [I(X = 1) \times I(Z = 1)]) \end{aligned}$$

$$\begin{aligned}
 X = 0, Z = 0 & \quad \text{Logit}Pr[Y = 1|X = 0, Z = 0] = \alpha \\
 X = 1, Z = 0 & \quad \text{Logit}Pr[Y = 1|X = 1, Z = 0] = \alpha \exp(\beta_1) \\
 X = 0, Z = 1 & \quad \text{Logit}Pr[Y = 1|X = 0, Z = 1] = \alpha \exp(\beta_2) \\
 X = 1, Z = 1 & \quad \text{Logit}Pr[Y = 1|X = 1, Z = 1] = \alpha \exp(\beta_1 + \beta_2 + \beta_3)
 \end{aligned}$$

Under null hypothesis of no effect modification $\{e^{\beta_1} = e^{\beta_1} + e^{\beta_3}\}$ and $\{e^{\beta_2} = e^{\beta_2} + e^{\beta_3}\}$
 $\rightarrow H_0: \beta_3 = 0$ or $e^{\beta_3} = 1$

Observed vs. Expected Joint Effects.

- conceptually analogous to *interaction*

$$\frac{\text{Risk}_{X=1,Z=1}}{\text{Risk}_{X=0,Z=0}} \neq \frac{\text{Risk}_{X=1,Z=0}}{\text{Risk}_{X=0,Z=0}} \times \frac{\text{Risk}_{X=0,Z=1}}{\text{Risk}_{X=0,Z=0}}$$

Joint effect of X & Z Main effect of X Main effect of Z

- in regression, create new variable defining subgroups...

$$\begin{aligned}
 \text{Logit}Pr[Y = 1|X = x, Z = z] &= \alpha \exp(\beta_1[I(X = 1) \times I(Z = 0)] \\
 &\quad + \beta_2[I(X = 0) \times I(Z = 1)] + \beta_3[I(X = 1) \times I(Z = 1)])
 \end{aligned}$$

$$\begin{aligned}
 X = 0, Z = 0 & \quad \text{Logit}Pr[Y = 1|X = 0, Z = 0] = \alpha \\
 X = 1, Z = 0 & \quad \text{Logit}Pr[Y = 1|X = 1, Z = 0] = \alpha \exp(\beta_1) \\
 X = 0, Z = 1 & \quad \text{Logit}Pr[Y = 1|X = 0, Z = 1] = \alpha \exp(\beta_2) \\
 X = 1, Z = 1 & \quad \text{Logit}Pr[Y = 1|X = x, Z = z] = \alpha \exp(\beta_3)
 \end{aligned}$$

- $e^{(\beta_1 + \beta_2)}$ and e^{β_3} are, respectively, the expected and observed joint effect of X and Z
- under null hypothesis of no interaction $e^{\beta_3} = e^{\beta_1} \times e^{\beta_2}$
- $\rightarrow H_0: \{\beta_3 - (\beta_1 + \beta_2) = 0\}$ or $\{e^{\beta_3} \div (e^{\beta_1} \times e^{\beta_2}) = 1\}$

CHAPTER 13

Time

1. Person-Time Analysis

Definition.

- is the amount of time an individual contributes to a group's observation, usually in a homogeneous state of risk
- has units of *person × time*

Estimation.

- There are 3 equivalent ways of calculating person-time
 - in a study of n subjects i , total person-time (PT) of observation =

$$(a) \sum_{i=1}^n [time]_i \quad (b) E[(time)_i] \times \sum_{i=1}^n i$$

$$(c) \sum time \times E \left[\sum_{i=1}^n i \right]$$

- (a) for each person, identify amount of person time contributed to group's observation, then sum the times of individual persons to get total person time for the group
- (b) multiply the number of persons under observation by the average duration of observation per person
- (c) multiply the length of the period of observation by the average number of persons under observation during the period

Assumptions.

- classification by time intervals and covariates assumes that within strata
 - i.e. risk of event is approximately constant during interval of interest
 - the incidence rate is applicable to time point in the interval (i.e. it is constant)
- this assumption means that the following are equivalent
 - n persons followed during t units of time
 - t persons observed during n units of time
- strong cumulative effects and latency periods make assumption implausible

- solution is to divide follow-up period into very small intervals of homogeneous risk
 - can use multiple time scales
- assumes an induction period of zero
 - analogous to crossover randomized trials
 - no *accumulation* of risk
 - effect of exposure is instantaneous
 - * resolved by complex exposure definitions:
 - total pack-years
 - lag or latency times

Group vs. Individual Data.

- in a dynamic population, when withdrawals and events occur uniformly...
 - the incidence rate is constant
 - average incidence rate = incidence density
 - group data (average) or individual (density) level data will suffice for estimation of hazard
- corresponds to actuarial life table methods (uniformity of events over interval)
- allows the comparison of *average rate*, based on an *average population*, to an incidence density based on person-years
 - in vital statistics → mid-point estimate at July 1

Advantages when person-time is unit of analysis.

- accounts for changes in individual experience in multiple categories of exposure over time
- cumulative effects of exposure can be taken into account with more complex definitions

2. Age-Period-Cohort Effects

- age-period-cohort analyses can be applied to prevalence, incidence, and mortality data
- age, period, and cohort effects often distort the perception of how disease progresses over time or as persons age in a *cross-sectional* analysis
- cross-sectional data
 - cannot be used to make *future* projections about the burden of disease
 - can be used to assess the *current* burden of disease

- can be difficult to tease apart, usually requires complex modeling (APC models)
- note that $age = birth\ cohort + age$

2.1. Definitions.

Age Effect.

- change in the rate of a condition according to age, irrespective of birth cohort and calendar time
- a change in disease incidence that is due to a biological concomitant of aging
- Age is usually a confounder
 - (1) associated with many exposure
 - (2) a strong risk factor for many health outcomes
- is usually the time-scale of choice in *observational* studies

Period Effect.

- *Definition*
 - change in disease frequency that is specific to a calendar time
 - change in the rate of a condition affecting an entire population at some point, irrespective of age and birth cohort
- *occurs when...*
 - war, new treatment, massive migration, etc.
 - introductions of new medications or preventative interventions
 - changes in prevalence of exposures that have short induction periods
- period effects for incidence rates are more prominent for diseases where cumulative effect of exposure is unimportant
 - e.g. infectious disease, injury

Cohort Effect.

- *Definition*
 - change in disease frequency shared by all members of a group who entered follow-up at a common time
 - change in the rate of a condition according to year of birth, irrespective of age and calendar time
- *occurs when...*
 - lifetime experience of individuals born at a given point in time influence the disease or outcome of interest

- * may **or** **may not** be related to circumstances at the time of birth of a given cohort
 - also viewed as an interaction between *age* and *calendar time*
 - * or rather as a combination of age and period effect
 - usually driven by factors established in childhood or adolescence (always prior to disease)
 - cohort effects for incidence rates are more prominent for diseases where cumulative effects of exposure are important
 - * e.g. chronic diseases
 - in the absence of a cohort effect, the same age patterns for rates are found in cross-sectional and cohort curves
 - cohort effects can affect associations between disease outcomes and variables other than age
 - * e.g. case-control study matched on age
 - if cases and controls identified from different birth cohorts
 - where birth cohort is associated with exposure of interest
 - birth cohort would be a confounder even though age is not

3. Person-Time Data Structure

- Typically, cohort study data is constructed with one record per person

ID	Birth Year	Baseline Age	Baseline Year	Years of Follow-up
1	1985	19	2004	4
2	1984	18	2002	6
3	1984	20	2004	4
4	1985	18	2003	5

TABLE 13.1. Example dataset for an open cohort

- Person-time data structure disaggregate individual's follow-up period to single person-time unit and transformed the short-form ("wide") data to long-form data.
- Person-time data structure also called Anderson-Gill data structure or counting process data structure
- Advantage of person-time data structure for data analysis
 - Easily allows for time-varying exposures and confounders, with full ability to check programming
 - Easily allows for exploration of alternate time scales

ID	Birth Year	Age	Calendar Year	Year of Follow-up
1	1985	19	2004	1
1	1985	20	2005	2
1	1985	21	2006	3
1	1985	22	2007	4

2	1984	18	2002	1
2	1984	19	2003	2
2	1984	20	2004	3
2	1984	21	2005	4
2	1984	22	2006	5
2	1984	23	2007	6

3	1984	20	2004	1
3	1984	21	2005	2
3	1984	22	2006	3
3	1984	23	2007	4

4	1985	18	2003	1
4	1985	19	2004	2
4	1985	20	2005	3
4	1985	21	2006	4
4	1985	22	2007	5

TABLE 13.2. Example of Anderson-Gill data structure

- Easily allows to switch from one to the other of the four methods presented for the analysis of cohort studies (Mantel-Haenszel, Cox, Poisson and pooled logistic regression)
- Easily allows for discontinuous follow-up time; that is, it is a straightforward way to handle the situation where the same participants move in and out of follow-up
- Easily allows for multiple events per person
- Allows for left-truncation and variable duration of follow-up (though this can also be done with one-record per person in Cox analyses)
- Scenarios that the counting process data structure is particularly well suited for
 - **Time dependent covariates:** For example, the Framingham Heart Study, in which self-reported blood pressure is repeatedly assessed.
 - **Alternate time scales:** For example, the uranium miners study, in which we can use age, years since starting to work, or calendar year as the time scale.
 - **Multiple events per subject:** For example, time to recurrent infections in patients treated with immunosuppressor agents after organ transplantation in which the same subject can experience recurrent or multiple events.
 - **Discontinuous intervals of risk:** For example, in a health insurance claim database, individuals may be on the health plan for an interval of years, then

be off the health plan, then back on again.

- **Time dependent strata:** For example, in organ transplantation studies we can assess the predictive value of covariates in the patient outcome both before and after transplantation, in which transplant status (yes vs. no) is a time dependent stratification variable.

4. Time Scales & Cohorts

- time scales can be arranged in various manners based on what is optimal for the causal question of interest
- practically speaking, variables are only suitable for use as a time scale if they do not stop accruing
- allow for better control of confounding by time or periods of varying baseline risk (seasonality, secular trends, etc.)
 - however, only control for seasonality if it relates to the time scale chosen, otherwise need to model it explicitly

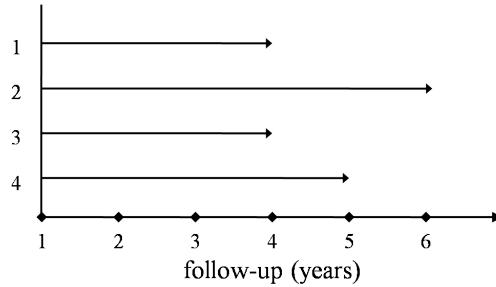


FIGURE 13.1. Example of using cohort as time scale

- using the cohort as the defining time scale (Figure 13.1) is often the time scale of choice in a randomized trial since exposure groups are exchangeable at baseline
- year of birth (birth cohort) is often used to assess exposures that may be different in birth cohorts
 - example: In 1899, maternal exposure to rubella during early pregnancy caused an increase in deafness in 10 year olds in 1909, 20 year olds in 1919, and 30 year olds in 1929
- **inception cohort:** individuals observed at the beginning of an exposure that defines a cohort, t_0
 - the cohort as it exists at some initial point of observation
- **survivor cohort:** individuals who remain under observation at some time t_j after the initial point of observation t_0
 - the cohort as it exists at some further removed point of observation

- all cohorts defined later than birth are to some extent a survivor cohort
- may result from selection bias due to loss of follow-up, competing risks, or depletion of susceptibles

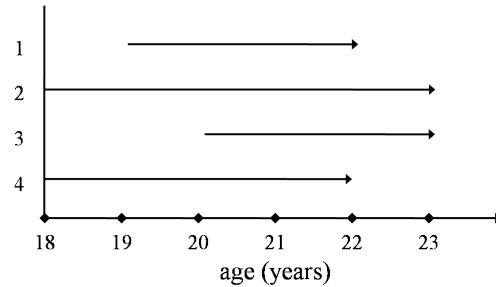


FIGURE 13.2. Example of using age as time scale

- age (Figure 13.2) is often the time scale of use in observational studies because it stratifies risk sets based on age to rule out confounding due to age (age is often associated with both exposure and outcome)
 - can further account for cohort effects by including birth (or entry) cohort into model
 - *note:* when age used as time scale, persons older than the case are excluded →

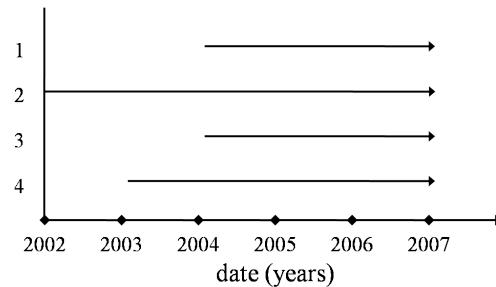


FIGURE 13.3. Example of using period as time scale

- risk sets can be used to investigate period effects that change over calendar time (Figure 13.3)
 - example: the way a disease is defined and diagnosed changes over time
 - example: a change in a manufacturing process that alters exposure levels to an industrial toxin
 - because risk sets match on time, the effect of the time scale cannot be estimated
 - * so the time scale should be something where you're not interested in estimating its effect
- remember that you can always estimate effect-modification by the time scale which is essentially testing the PH assumption

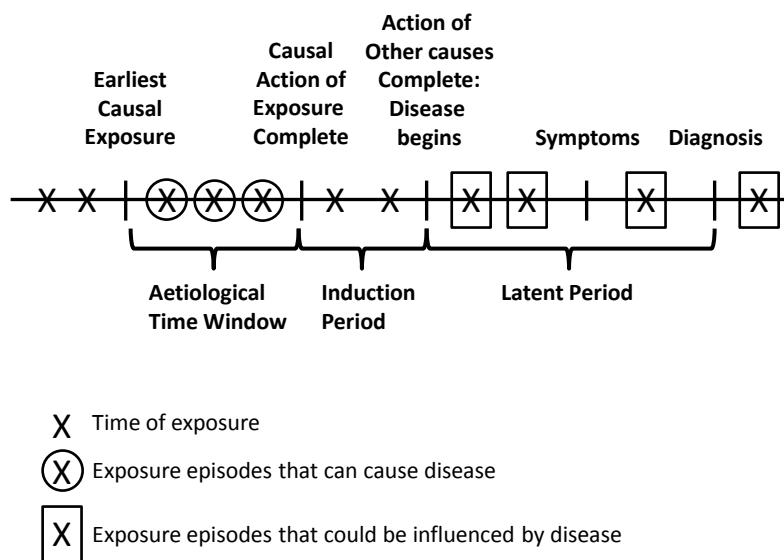


FIGURE 13.4. Type of Exposure During Follow Up

5. Time-Based Definitions of Exposure

5.1. Concepts.

- person time classified based on exposure status according to study hypothesis
- any assumptions about induction and latent periods should be incorporated to study hypothesis
- the aetiological window should be incorporated as well
- the time when exposure occurs and the time at risk of its effect are not necessarily the same
- *Induction time*
 - time from exposure to disease onset
 - disease-free period after exposure, during which "pathogenic" mechanisms work toward production of manifest disease
 - *Residual effect*
subsequent changes in disease incidence attributable to exposure
 - characteristic of a delayed rise in incidence rate after exposure
 - minimum induction period: time elapsed until summed residual effects from all prior exposures become high enough to result in observable disease
- *Latent period*
 - time from disease onset to detection (exposure has no effect during it)
 - disease is present but not manifest

- *Etiological Window*

- the interval of time in which exposure is most relevant to cause of disease
- determines intervals of susceptibility
- person-time outside the etiological window should not be counted as exposed
→ bias towards null
- Examples:
 - * in utero, infancy, childhood, adolescence, etc.
 - * time in relation to menarche, pregnancy, childbirth, menopause
 - * calendar years
 - * **can be examined statistically through two-way interaction terms with age**

- *Chronic vs Point exposure*

- in *chronic exposures*, if we assume a long induction period, we assume that exposure must accumulate to a certain level before having an effect
- same rationale for latent periods
 - * in this situation the time when exposure occurs \neq time at risk of its effect
 - * classifying person-time as exposed during the induction period has the effect of independent non-differential misclassification of exposure (expected bias towards null)
- in *point exposures* we assume negligible induction period
 - * also known as *acute* exposure
 - * exposure occurs instantaneously
 - * in this situation time of exposure = time at risk of its effect

- *Immortal person time*

- entry criteria into cohort is dependent upon survival or meeting eligibility criteria
- should be excluded from analyses because it will downwardly bias estimated disease rate for that group
- similarly for exposure categories, follow-up time should exclude time during which the exposure-category criteria are being met

- *Post-exposure events*

- allocation of follow-up time should not depend on events that occur after it has accrued
- "future should not determine the past"

- consider the randomized experiment you'd like to perform and classify person-time in your observational study accordingly

5.2. Defining exposed and unexposed person-time.

a choice between exposure definitions can be based on...

- (1) *a priori* considerations (subject matter knowledge)
- (2) empirically (data driven) via comparison log-likelihood of non-nested models
 - often not much power to determine which exposure definition is more biologically relevant
 - because they are related quantities

Exposure Definitions.

- *Categorical*
 - events and person time *must* be in the same category
- *Continuous*
 - unlike cumulative, may increase or decrease over time
- *Cumulative*
 - the summation of all exposures endured from t_0 until t_1
 - e.g. time since...
 - makes sense for "one-hit" disease, where multiple hits are proportional to risk
 - makes sense for disease where accumulated effect of exposure must meet a threshold for disease to occur
 - for cumulative dose need: (1) age began/ended (2) frequency (3) intensity
- *Composite measures*
 - e.g. pack-years
 - are necessary when exposure is a function of frequency, duration, intensity, and dose
 - often components are best modeled as separate variables
 - * allows simultaneous interpretation of components and the composite measure
- *Intensity*
 - average, maximum, median, minimum etc.
 - makes sense for disease where *current* exposure must meet a threshold for disease to occur
 - minimum → preventative (or beneficial) treatment (e.g. essential nutrients)

- maximum → causative (or harmful) treatment
- *Average lifetime exposure intensity*
 - * average exposure received by an individual where the average is taken over some specified time-frame
- *Dose*
 - available dose, active dose, absorbed dose etc.
 - be sure to account for method of exposure, route of administration, etc.
 - should be comparable between exposed and unexposed
- *Lagging exposures*
 - exposures at or up to a specified time before current time
 - to avoid confounding by subclinical disease during induction period
 - * avoidance or use of treatment/exposure ← symptoms → disease onset
 - to avoid reverse causation bias during latent period

Active Agent.

- often the active agent in an exposure depends on disease outcome and hypothesis
- i.e. in coffee → caffeine-parkinsons vs. nutrient-diabetes

CHAPTER 14

Continuous Exposures & Confounders

1. Indicator Method

1.1. Reference coding.

- take a continuous variable V with range $[0, \infty]$

let $C_k = \text{threshold } k = \{1, 2, \dots, k\}$

$$\text{define } X = \begin{cases} 1 & \text{if } V \text{ in } [0, C_1) \\ 2 & \text{if } V \text{ in } [C_1, C_2) \\ \vdots \\ k & \text{if } V \text{ in } [C_k, \infty) \end{cases}$$

- defining an indicator function $I(X)$ to equal 1 if true, 0 if false...

$$\begin{aligned} E(Y) &= \beta_0 + \beta \sum_{k=1}^{k-1} I(X = k) \\ &= \beta_0 + \beta_1 I(X = 1) + \beta_2 I(X = 2) \dots \beta_{k-1} I(k - 1) \end{aligned}$$

– the category k is the reference group to which all other categories are compared

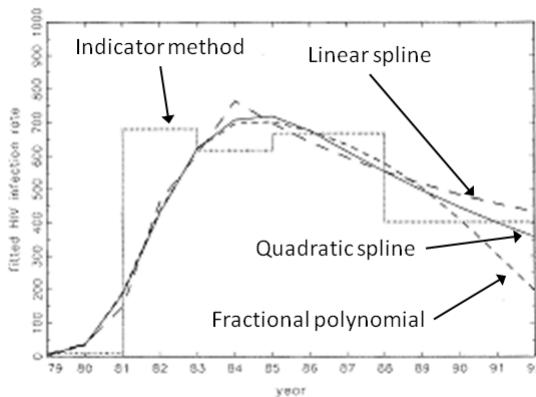


FIGURE 14.1. Comparison of Dose-Response Curves

e.g. if logistic regression...

$$\beta_1 = \left(\frac{Pr[Y = 1|X = 1]/Pr[Y = 0|X = 1]}{Pr[Y = 1|X = k]/Pr[Y = 0|X = k]} \right)$$

$$\beta_2 = \left(\frac{Pr[Y = 1|X = 2]/Pr[Y = 0|X = 2]}{Pr[Y = 1|X = k]/Pr[Y = 0|X = k]} \right)$$

- when $X = k$, $\text{logit}[Pr[Y|X]] = \beta_0$, so β_0 equals the log odds when $X = k$ and all other covariates have value $L = 0$. So $I(X = k)$ is not included in the model
- note that any category left out of the model is absorbed into the reference category
 - if have categories $X = \{1, 2, 3, 4\}$ but only include $I(X = 1)$ and $I(X = 4)$ in model, the reference category becomes $I(X = 2 \cup 3)$

1.2. Assumptions.

- the dose response curve is a step function with abrupt changes in risk at the chosen thresholds C_k
 - i.e. line with zero slope in each category
- the parameter β_k describing the relationship between outcome Y and exposure V is **homogeneous** over the range $[C_{k-1}, C_k]$
- choice of thresholds and reference group are both biologically meaningful
 - for exposure \rightarrow analogous to no misclassification
 - for confounder \rightarrow analogous to no residual confounding

1.3. Practical issues.

Implementation.

- choosing the category with the largest sample size to be the reference group is the most efficient (powerful) use of data
- when interacting a categorized variable X with a variable Z , must model *all* product terms between Z and all included $I(X)$

$$E(Y|X, Z) = \beta_0 + \beta_1 I(X = 1) + \beta_2 I(X = 2) + \beta_3 Z + \beta_4 [Z \times I(X = 1)] + \beta_5 [Z \times I(X = 2)]$$

Strengths.

- the step function characterizing the dose-response curve can assume any arbitrary pattern
 - i.e. does not assume a monotonic relationship
- very interpretable, providing effect measures for each category
- More powerful to identify non-linear or threshold effect than linear trend test

- Less sensitive to outlier
- Less influenced by extreme value in the end category where data is usually sparse

Weaknesses.

- choice of (a) reference group (b) # and location of thresholds is *arbitrary* . . . either use subject matter knowledge or evenly spaced intervals (quantiles)
 - the comparison of all categories to a common category may not be biologically relevant
 - the thresholds may have no biological meaning
 - may mask important trends in dose-response curve
- using narrower (and hence, more) categories can decrease the potential bias from above issues, but will also decrease power
- Equal-distance categories may not identify effects at extreme ends
- Unrealistic step-risk function (sudden change of risk once across the category boundary)
- Less statistical power and precision, throwing out of intra-category information and increasing degree of freedom
- Potential data dredging due to manipulation of binning
- May sometimes violate monotonic increase or decrease assumption, which is biologically implausible

2. Score-Test

- consider the simple regression model $\text{logit}Pr[Y = 1] = \beta_0 + \beta_1 A$, where A is a continuous exposure
 - the Score-Test is a test used to test the hypothesis of $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$
 - the Score-Test is based on the score statistic $U(\beta_1) = \frac{\partial(\text{log}(L(\beta_0, \beta_1)))}{\partial \beta_1} = \sum(Y - P)A$ where $P = P[Y|A]$
 - and the information matrix $I(\beta_1) = -\frac{\partial^2(\text{log}(L(\beta_0, \beta_1)))}{\partial \beta_1^2} = \sum P(1 - P)A$
 - follows a $\chi^2_{(1)}$ distribution under the null
 - a score test that fails to reject the null $H_0 : \beta_1 = 0$ does not mean no association
 - the exposure A can be replaced by any ordinal score vector S that is assigned to different exposure levels of categorical variable
 - simple likelihood ratio test for the null hypothesis for categorical predictor of no association $H_0 : \beta_1 = \dots = \beta_k = 0$ is H_a =not all β s are equal
 - using scores S an alternative hypothesis with a natural ordering of a categorical predictors the alternatives $H_a : \beta_1 < \dots < \beta_k$ or $H_a : \beta_1 > \dots > \beta_k$

- we can use a LRT test with the continuous exposure model nested within a model with categorical variables for the exposure to test if the linear model fits the data

2.1. Linear trend test.

- Advantages
 - More power and efficiency if the trend curve fit the linear function form
 - Immune to data dredging due or manipulation of binning
 - Avoid unreasonable assumption of sudden change of risk between categories
- Disadvantages
 - Make a restrictive function form assumption of the trend curve (log-linear or logistic-linear)
 - More sensitive to outlier
 - Not robust to non-linear relationship

3. Fractional Polynomials

3.1. polynomial model.

- polynomial models were often used to examine the dose-response relationship without over-simplification of the dose-response relationship
- in theory, one can approximate any smooth curve with enough polynomial terms
- in reality, the number of terms required may be large and result in numerically unstable estimates
 - models with polynomials greater than quadratic tend to be unstable and tend to produce artificial patterns
 - on contrary, polynomial model was less flexible if only quadratic terms were included in the model

3.2. Fractional Polynomial model.

- fractions and inverse powers of the variable of interest (X) were suggested in addition to linear and quadratic terms
- e.g. $E[Y] = \beta_0 + \beta_1 x^{1/2} + \beta_2 x + \beta_3 x^{-3/2} + \beta_4 x^2$ (see figure 14.1)
 - exponential and log terms (e^X and $\log(X)$) can also be used
- advantages of fractional polynomial model
 - a simple qualitative dose-response or trend analysis can be done by adding power terms
 - three different powers of X between x^{-2} and X^2 can give the model a wide range of shapes

- for logistic and log-linear model, include $\ln(X)$ term in the model
 - * use $\ln(x)$, the rate ratio, risk ratio, and odds ratio can be $e^{\beta \ln(X)} = X^\beta$
 - * X^β can increase slower than exponentially ($e^{\beta X} = [e^X]^\beta$) if $\beta \geq 1$
 - * X^β can increase slower than linearly if $1 > \beta > 0$
- Allow change of slope and make use of all information
- Increase power and precision to identify non-linear relationship
- More powerful to identify non-linear or threshold effect than linear trend test
- No assumption of step-risk function
- More parsimonious model than categorical or spline regression
- disadvantages of fractional polynomial model
 - (1) covariate X cannot be negative due to the fact that $X^{1/2}$ is not a real number if $X < 0$.
 - it is better for x to have an absolute zero level which is coded as zero and all other levels of x are greater than zero
 - may add a value to make all X are positive (however, need to decide what value to add)
 - (2) it is difficult to decide what terms to put in the model
 - it is better that the shape of the curve was specified prior to model fitting
 - otherwise, it's fishing
 - however, it requires one to know each power term's effect on the shape of the curve to pre-specify the shape of the curve
 - add quadratic term to increase the slope
 - add square-root term to decrease the slope
 - add more extreme power terms to obtain more rapid change in shape
 - (3) outliers and influential points may affect the model fitting a lot
 - regression diagnostics (e.g. residual plot, Cook's distance, goodness-of-fit, etc.) and 95% confidence interval (CI) may be helpful to assess the model fitting
 - (4) Unstable function forms, many FP functions fit equally well
 - (5) Still parametric and less flexible than non-parametric (eg. categorical) methods
 - (6) Poor fit of end effects (compared to spline)
 - (7) Parameter interpretation is difficult

4. Splines

- splines are a sequence of joined segments that produces a piecewise pattern, also known as the “broken-stick” model
- provides a more flexible way to accommodate non-linear trends that cannot be approximated by simple polynomials
- the basic idea is to divide the independent variable (X , L , etc) into a series of segments and fit piecewise trends having different functions but joined at fixed intervals
- **knots**: are locations where segments are joined together, predetermined values x^*
- **spline**: resulting piecewise linear curve
- the graph of a spline looks like a series of connected line segments, it is piecewise linear but continuous
- various spline functions can be considered
- restrictions can also be placed on the splines (ex: linearity at extremes of the distribution, constraints to produce monotonic curves)

4.1. Linear Splines.

- the simplest possible spline model is a linear spline with only one knot:

$$E(Y_{ij}) = \beta_0 + \beta_1 X + \beta_2(X - x^*)_+$$

- x^* is equal to a predetermined value of X where a knot is placed
 - $(X - x^*)_+$ is a function $g(X)$ defined as $\max(0, X - x^*)$
 - $g(X)$ equals 0 if g is negative, and $X - x^*$ otherwise
 - once $X > x^*$, the function g always equals $X - x^*$
 - more complex models may have multiple knots:
- $$E(Y_{ij}) = \beta_0 + \beta_1 X + \beta_2(X - x_1^*)_+ + \beta_3(X - x_2^*)_+ + \cdots + \beta_{k+1}(X - x_K^*)_+$$
- where $x_1^*, x_2^*, \dots, x_K^*$ are predetermined knots along the distribution of X for the linear spline model
 - β_k is the change in the slope of the dose-response going from category $k - 1$ to category k
 - linear splines improve on indicator variables, which correspond to a step function, by guaranteeing the function is connected at knots
 - kinks in the function sometimes occur at knots, which are not biologically possible
 - linear splines can also suffer from instabilities that are sensitive to the choice of knots

4.2. Quadratic and Cubic Splines.

- offer even greater flexibility over linear splines by allowing a curvilinear function for each of the splines
- address the problem of kinks at the knots by creating a curve with no sharp bends

4.2.1. Quadratic Splines.

- quadratic splines can be fit using the model:

$$E(Y_{ij}) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 (X - x_1^*)_+^2 + \beta_4 (X - x_2^*)_+^2 + \cdots + \beta_{K+2} (X - x_K^*)_+^2$$

- β_k represents the change in the quadratic term (departure from linearity) of the dose-response function from category $k - 1$ to category k
- one additional parameter ($\beta_2 X^2$) is needed to fit the quadratic spline
- quadratic splines usually require fewer knots for accuracy, thus the added parameter ($\beta_2 X^2$) is offset by the fact that fewer parameters are needed
- the quadratic spline has a smooth appearance because the same slope is shared on either side of the knots
- like higher order polynomials, quadratic spline models can exhibit odd behavior at the extremes of the exposure distribution
 - behavior of the lower tail can be improved by dropping X^2 from the model, that is restricting the fitted curve to be linear at this spline
 - behavior of the upper tail can be improved by dropping $(X - x_K^*)_+^2$ and replacing all $(X - x_k^*)_+^2$ with $((X - x_k^*)_+^2 - (X - x_K^*)_+^2)$

4.2.2. Cubic Splines.

- splines can be further generalized to have cubic terms
- cubic splines can be fit using the model:

$$E(Y_{ij}) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X - x_1^*)_+^3 + \beta_5 (X - x_2^*)_+^3 + \cdots + \beta_{K+3} (X - x_K^*)_+^3$$

- β_k represents the change in the cubic term of the dose-response function from category $k - 1$ to category k
- two additional parameters ($\beta_2 X^2$ and $\beta_3 X^3$) are required over linear spline models, and one additional parameter ($\beta_3 X^3$) is required over quadratic spline models to fit the quadratic spline
- cubic spline models are preferred by most statisticians

4.3. Practical Considerations.

- Advantages
 - Combine the advantages of categorical and power model

- Semi-parametric, flexible to capture dose response relationships of almost any shape
- require little specification of the dose-response relationship
- Smooth connection between knots no sudden change in risk across category boundary
- can get a p-value for a test of non-linearity
- Disadvantages
 - Overfitting is likely and may affect generalizability
 - may be strongly influenced by outliers
 - can be sensitive to the choice of smoothing parameter (ex: linear, quadratic, cubic)
 - difficult to get a summary relative risk for the dose-response relationship (best to look at a graph of the function)
 - Curves in the end categories (tails) may become very unstable for high power spline (can improve with restrictive spline)
 - Effects depends on the number and value of knots while the selection of knots is usually visual or empirical
 - not easily implemented in standard software packages, requires extra data manipulation
 - can be mathematically complex and computationally intensive

5. Comparisons

	Trend test	Categorical	FP	Spline
Assumption	Linear assumption	Step-risk assumption	Parametric, assumption of power term function form	Semi-parametric, assumption of knot and within-category function form
Robust to non-linearity	No	Yes	Yes	Yes
Power and precision	Good if linear relationship hold	Less	Best power and precision	Between categorical and FP
Interpretability	High	High	Low	Low
Ease of implementation	Easy	Easy	Moderate	Difficult
Influence by outlier	Moderate	Mild	Strong	Strong
Goodness of fit	Good if linear relationship hold	Poor	Good	Best
Potential of data-dredging	Low	High	Low	Moderate

FIGURE 14.2. Comparison of different dose-response analytical strategies

	Assumptions	Strengths	Limitations
Single Measure	one measurement accurately represents individual exposure, exposure remains constant	inexpensive, simple analysis, easily interpretable	susceptible to fluctuations in exposure, more susceptible to measurement error
Multiple Measure	exposure changes over time, one exposure measurement is insufficient	multiple assessments of exposure over time, can detect exposure trends over time	more resource intensive, requires complicated longitudinal analysis techniques
Cum. Dose	exposure bioaccumulates over time and is related to the outcome in a cumulative manner	able to investigate cumulative effect of an exposure on outcome	requires multiple measurements, lose information by collapsing measurements
Average Dose	effects of exposure over time are transient or trying to adjust for cumulative exposure over time	easy to analyze and interpret, compare individuals with different exposure times	requires multiple measurements, lose potential information by collapsing measurements into an average
Min/Max Dose	threshold level needed for exposure to initiate/prevent outcome, duration not important	aids in identifying biological threshold, easy analysis, easily interpretable	requires multiple measurements, only uses data from one measurement point
Exposure Duration	assumes length of exposure is an important determinant of outcome	only necessary to measure if exposed and not exposure intensity, easy to analyze and interpret	does not factor in the intensity of the exposure
Exposure Window	assumes certain periods of exposure may be more important for outcome	can detect important windows of exposure effect, powerful analysis to detect time effect	need to determine what time scale is important, need to define time periods of interest
Lag Exposure	assumes an induction / latent period defined by the investigator	avoids misclassification of exposure by adjusting for induction/latent period	ignores etiological effects of exposure during induction time, throws away data

TABLE 14.1. Comparison of Exposure Assessments

CHAPTER 15

Missing Data

1. Types of Missing Data

Define... M = missing data, Y_{miss} = unobserved data, Y_{obs} = observed data

(1) MCAR: Missing Completely at Random

- $M \perp\!\!\!\perp Y_{obs}, Y_{miss}$

$$Pr(M|Y_{obs}, Y_{miss}) = Pr(M)$$

- the missingness process does not depend on collected or uncollected information
- to obtain a valid analysis there is *no need* to model the missingness process but it may be more efficient to do so
- example
 - random data entry / processing errors
- methods that are unbiased:
 - complete case, IPW, single/multiple imputation, maximum likelihood

(2) MAR: Missing at Random

- $M \perp\!\!\!\perp Y_{miss} | Y_{obs}$

$$Pr(M|Y_{obs}, Y_{miss}) = Pr(M|Y_{obs})$$

- the missingness process depends only on information you collected
- to obtain a valid analysis you must model the missingness process
- example:
 - genetic epi study: choose to genotype only a subset of SNPs because the rest can be inferred from the linkage disequilibrium pattern in the genotyped SNPs
- methods that are unbiased:
 - IPW, single/multiple imputation, maximum likelihood
 - complete case is multiplicative

(3) MNAR: Missing Not at Random

- $M \not\perp\!\!\!\perp Y_{miss} | Y_{obs}$

Method	Advantages	Disadvantages
Complete case	• Easy	• Generally biased if data are not MCAR* • Inefficient
Missing indicator	• Easy for one variable • A little more efficient	• Biased • Difficult for more than one variable
Weighted	• Unbiased if data are MAR and missingness model correctly specified • Point estimation easy • Can be quite efficient**	• Estimating standard errors can be difficult • Can be inefficient**
Single imputation	• Easy • Can be unbiased in important situations (e.g. under the null) • Can be quite efficient**	• Generally biased • Estimating standard errors can be difficult • Can be inefficient**
Maximum likelihood	• Unbiased if missingness model correctly specified (even for MNAR) • Can be more efficient	• Very difficult to implement

*Unbiased if missingness probability is “multiplicative”

**Loss of info depends on how accurately missing data can be predicted

FIGURE 15.1. Approaches for handling missing data

$$Pr(M|Y_{obs}, Y_{miss}) = Pr(M|Y_{obs}, Y_{miss})$$

- example:
 - study of substance abuse: persons with a criminal record are more likely to refuse questions about illicit drug use. Unfortunately you did not collect information on criminal record, and the data you have don't give you information about the possibility of a criminal record.
- methods that are unbiased:
 - maximum likelihood
 - *missing indicator is always biased*

method	assumption	validity	efficiency	implementation
complete case	MCAR	unbiased if assumption holds	inefficient	easy
missing indicator	No assumption	always biased	a little more efficient than complete case analysis	difficult if more than 1 variable
IPW	MAR and no model mis-specification (or MCAR)	unbiased if assumption holds	depends on the accuracy of missing data prediction given observed data	easy to get point estimates, difficult to get variance estimates
unconditional single mean imputation	distribution of the variable (no model mis-specification)	generally biased (underestimates variance of the variable and weaken the association between the variable and the outcome)	depends the accuracy of missing data prediction given observed data	easy
unconditional single draw imputation	distribution of the variable (no model mis-specification)	generally biased (underestimates variance of the variable and weaken the association between the variable and the outcome)	depends the accuracy of missing data prediction given observed data	easy
conditional single mean imputation		It creates biased results!!!		do NOT use this method
conditional single draw imputation	distribution of the variable given observed data (no model mis-specification)	unbiased if MAR	estimates variance of imputed variable better than mean imputation	easy
maximum likelihood	distribution of the variable (no model mis-specification)	unbiased if no model mis-specification	efficient	difficult

TABLE 15.1. Comparison between methods for missing data methods

2. Complete Case

- Only use persons without missing data on any variable in model
- Unbiased in MCAR

- Example case-control study, MCAR
- $Pr((M = 0|D, E, X_1, \dots, X_p) = Pr(M = 0) = f$ for all participants, where $M = 1$ means missing confounder

D	E=1	E=0
Case	fa_i	fb_i
Control	fc_i	fd_i

FIGURE 15.2. Case-control data stratum i

- Use complete data in stratum i $OR_i = \frac{fa_i fd_i}{fb_i fc_i} = \frac{a_i d_i}{b_i c_i}$ is a valid estimate
- Efficiency decreased because sample size is reduced by $(1 - f)\%$
- Unbiased in MAR if missingness is multiplicative
 - MAR $[-] \text{ multiplicatively}$
 - $Pr((M = 0|D, E, X_1, \dots, X_p) = Pr((M = 0|D, X_1, \dots, X_p)Pr((M = 0|E, X_1, \dots, X_p) = f_{di}g_{ei}$

D	E=1	E=0
Case	$f_{1i}g_{1i} a_i$	$f_{1i}g_{0i} b_i$
Control	$f_{0i}g_{1i} c_i$	$f_{0i}g_{0i} d_i$

FIGURE 15.3. Case-control data stratum i

- $OR_i = \frac{f_{1i}g_{1i} a_i f_{0i}g_{0i} d_i}{f_{1i}g_{0i} b_i f_{0i}g_{1i} c_i} = \frac{a_i d_i}{b_i c_i}$ is a valid estimate
- In general, if each cell has a fraction of complete case $f_{11i}, f_{10i}, f_{01i}, f_{00i}$, then complete case method is unbiased if $\frac{f_{11i}f_{00i}}{f_{10i}f_{01i}} = 1$
- Efficiency reduced compared to unobserved full data
- For continuous outcome, even though the missing process is multiplicative, the complete case method is biased

3. Missing Indicator

- Recode missing data as zero, create indicator variable for missing then use it for stratified analysis or control in regression model
- Not biased if not a confounder

- Guaranteed to be biased if a confounder (effect estimates are weighted avg. of nonmissing data (unbiased) and a missing data estimates(biased)).
- For missing data on many covariates, each missing data level must have variation in the outcome → at least 1 case and control
- Underestimates variability in dataset
- Very efficient b/c get to use all of your data
 - A trade off between bias and efficiency, increase efficiency at a price of small bias if the proportion of missing data is small
 - *missing indicator is always biased*

4. Inverse Probability Weighting

4.1. Single variable.

- observed data weighted to recreate the full (including unobserved) data
- fit logistic regression for missingness
 - simple example with outcome Y, exposure A, confounder L, $M = I(L_{missing})$
 - model $logit[Pr(M|A, Y)] = \beta_0 + \beta_1 Y + \beta_2 X + \beta_3 YA$
 - set weights $w(A, Y) = \frac{1}{Pr(M|A, Y)}$
 - then run weight regression of Y on A,L (must use robust variance)
- assumptions
 - data missing at random (distribution of L the same in those where it is observed and those where it is unobserved)
 - missingness model is correctly specified
 - no missing data on A and Y

4.2. Missingness Pattern.

- with multiple variables with missing data use 'missingness pattern'
- covariates $L_1, \dots, L_k, M_i = I(L_{i missing})$
- M^* is a categorical variable with 2^k levels where k is the number of covariates with missing data
- If $M_1 = M_K = 0$, then $M^* = 1$; if $M_1 = 1$ and $M_2 = \dots = M_k = 0$, $M^* = 2$; etc.
- fit polytomous logistic regression $logit[Pr(M^* = i|A, Y)] = \beta_{i0} + \beta_{i1} Y + \beta_{i2} A + \beta_{i3} AY$
 - to minimize the number of parameters and strata for the polytomous logistic model we could make further assumptions
 - e.g. $Pr(L|M^* = i) = Pr(L|M^* = j)$ for all $i, j > 1$

- fit weighted regression of Y on A, L_1, \dots, Z_k , with weights $w(A, Y) = \frac{1}{Pr(M^*|A, Y)}$ (robust variance)
- same assumption as with single variable

5. Imputation

5.1. Concepts.

- conceptually, imputation requires that the distribution of missing values, given the observed data, be specified

$$Pr(\mathbf{X}^{miss}|\mathbf{M}, \mathbf{X}^{obs}) = \frac{Pr(\mathbf{M}|\mathbf{X}^{miss}, \mathbf{X}^{obs}, Y) \times Pr(Y|\mathbf{X}^{miss}, \mathbf{X}^{obs}) \times Pr(\mathbf{X}^{miss}, \mathbf{X}^{obs})}{\sum_{\mathbf{X}^{miss}} Pr(\mathbf{M}|\mathbf{X}^{miss}, \mathbf{X}^{obs}, Y) \times Pr(Y|\mathbf{X}^{miss}, \mathbf{X}^{obs}) \times Pr(\mathbf{X}^{miss}, \mathbf{X}^{obs})}$$

where $\mathbf{M} = \{M_1, \dots, M_p\}$ is a vector of missing indicators ($M_i = 1$ if X_i is missing, 0 otherwise); $\mathbf{X} = \{X_1, \dots, X_p\}$ is a vector of covariates

- to estimate this distribution we need three components:
 - (1) a model for outcome Y , conditional on the complete set of covariates
 - $\mathbf{X} = (\mathbf{X}^{miss}, \mathbf{X}^{obs})$
 - (2) a model for the missingness process \mathbf{M}
 - $Pr(Y = \mathbf{M}|\mathbf{X}, Y)$
 - if data is not MCAR does not have to be specified
 - ... $Pr(Y = \mathbf{M}|\mathbf{X}, Y) = Pr(\mathbf{M})$ and it cancels out
 - (3) a model for the joint distribution of the covariates \mathbf{X}
 - $Pr(\mathbf{X}) = Pr(\mathbf{X}^{miss}, \mathbf{X}^{obs})$
 - this is the part that makes imputation difficult
- this involves more assumptions than we typically make in multivariate analysis

5.2. Single Imputation.

5.2.1. Unconditional mean.

- replace missing values with crude mean of observed values

$$X_{ij}^{imputed} = \bar{X}_j = \frac{1}{\sum(1 - M_{ij})} \sum(1 - M_{ij})X_{ij}$$

- **biased** (in general) and not recommended
 - parameters are biased
 - underestimates variability in X_j
 - weakens any associations with other X 's and outcome Y
 - bias towards the null

5.2.2. Unconditional draw.

- replace missing values with realizations from a probability distributions based on the observed values

$$X_{ij}^{imputed} \sim (\bar{X}_j, s_j^2)$$

$$\text{where } \dots s_j^2 = \frac{1}{(\sum_{\{1-M_{ij}\}}-1)} \sum (1 - M_{ij})(X_{ij} - \bar{X}_{ij})^2$$

- reduces variability in X 's
- unbiased if MCAR and distributional assumption correct

5.2.3. Conditional mean.

- regress missing variable on other covariates

$$X_{ij} = \hat{\alpha} + \hat{\beta}Z_i$$

- **never** include outcome of interest as a predictor for this approach
- will correlation between imputed variable and outcome where none exists
- imputing mean values does not preserve variability in population
 - * uses entire population unlike IPW, which only uses observed data and thus preserves the original variability

5.2.4. Conditional draw.

- regress missing variable on other covariates, but include a random error term $\epsilon \sim N(0, 1)$

- steps

$$(1) \quad X_{ij}^{imputed} = \hat{\alpha} + \hat{\beta}Y_i$$

$$(2) \quad X_{ij}^{imputed} \sim N(\hat{\beta}Y_i, \hat{s}^2)$$

- final predictions are sampled from mean of original predicted values but have an added random error component

- reduces variability in imputed X 's
 - variance is constrained by distribution and modeling assumptions
- is unbiased under MAR, but still underestimates variance of parameter estimates

5.3. Multiple Imputation.

- the main idea is to “fill in” missing data multiple times by replacing an unobserved variable by a random draw from the distribution of that variable conditional on the observed data
- multiple imputation attempts to get an estimate of the β parameter for X , our covariate of interest, rather than just estimating a value for $X_{missing}$

m	Fraction of Missing Information				
	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

TABLE 15.2. Relative efficiency of M under different scenarios of missingness

- the advantage of multiple imputation over single imputation is it allows one to account for additional uncertainty due to the missingness
- multiple imputation creates M “filled in” datasets to analyze and yield estimates of the parameter of interest β ($\beta_1, \beta_2, \dots, \beta_M$)
- the overall estimate of β is the average of the M estimates

$$\hat{\beta} = \frac{1}{M} \sum_{j=1}^M \hat{\beta}_j$$

- the variance is estimated as

$$\hat{Var}(\hat{\beta}) = \frac{1}{M} \sum_{j=1}^M \hat{\sigma}_j^2 + \left(1 + \frac{1}{M}\right) \left(\frac{1}{M-1} \sum_{j=1}^M (\hat{\beta}_j - \hat{\beta})^2 \right) = A + \left(1 + \frac{1}{M}\right) B$$

- $\hat{\sigma}_j^2$ is the estimated variance of $\hat{\beta}_j$ in the j th imputation
- A = the variance of each $\hat{\beta}_j$
- B = the variance among the $\hat{\beta}_j$ ’s, this helps account for the variability in the imputation due to uncertainty from missing data

- to conduct the hypothesis test that $\beta = 0$ the following test statistic is used

$$\frac{\hat{\beta}}{\sqrt{\hat{Var}(\hat{\beta})}}$$

- this follows a t distribution with degrees of freedom = $(M - 1) \left(1 + \frac{A}{(1+M^{-1})B}\right)$
- a key question with multiple imputation is how large M should be
- it turns out that M does not need to be very large (Table 15.2)
- SAS Proc MI can be used to generate multiple imputed datasets
 - the default M is 5
 - Proc MI can use a Markov Chain Monte Carlo (MCMC) procedure to impute missingness

- * basically MCMC is a computationally intensive approach that breaks up a complicated distribution into simpler, smaller parts that can be sampled from
- * it is a chain because the algorithm randomly jumps from one point in the sampling distribution to another

6. Maximum Likelihood

- can deal with MNAR
- rely on the fact that

$$Pr[M, X_{obs}, Y] = \sum_{X_{missing}} Pr[M|X_{missing}, X_{obs}, Y] Pr[Y|X_{missing}, X_{obs}] Pr[X_{missing}, X_{obs}]$$

- we may correctly impute $\{X_{missing}, X_{obs}\}$ if we can correctly specify 3 models for each component of $Pr[M, X_{obs}, Y]$
- also, we need to integrate the function over all possible $X_{missing}$

Part 5

Inference With Models: Types

CHAPTER 16

Models for Statistical Inference

1. Stratified Analysis using Regression Models

- examples
 - linear regression: $Y = \beta_0 + \beta_1 X_1 + \epsilon$
 - logistic regression: $\text{logit}\{E[Y|X]\} = \beta_0 + \beta_1 X_1$
 - Poisson regression (for rate) : $\log\{E[Y]\} = \beta_0 + \beta_1 X_1 + \log\{\text{person-time}\}$
 - Cox proportional hazard model: $h_X(t) = h_0(t)e^{\beta_1 X}$
- exposures and confounders are both included in the model as covariates
- **must condition on all covariates in the model**
- require exchangeability, positivity, and consistency assumptions
- no average causal effect in the population was estimated
 - unless we assume that there is a constant average causal effect across strata

TABLE 16.1. Interpretation of Regression Coefficients

Form	model	Interpretation
Logistic	$\log(\text{odds}) = \beta_0 + \beta_1 X_1 + X'\beta$	Increase in the <i>log odds</i> of outcome per unit increase in X_1 , adjusted for all other variables in the model
Cox	$\log(\text{hazard}) = \beta_0 + \beta_1 X_1 + X'\beta$	Increase in the <i>log hazard</i> of outcome per unit increase in X_1 , adjusted for all other variables in the model
Poisson	$\log(\text{rate}) = \beta_0 + \beta_1 X_1 + X'\beta$	Increase in the <i>log rate</i> of outcome per unit increase in X_1 , adjusted for all other variables in the model

2. Linear Regression

3. Mantel Hanszel as a Regression Function

3.1. M-H Estimator.

- Mantel-Haenszel (MH) analysis for an open cohort study

		open cohort		M_1
		exposed	unexposed	
cases	a_i	b_i	T_i	M_1
	N_{1i}	N_{0i}		

- recall the summary Mantel-Haenszel rate ratio is equal to:

$$\hat{IRR}_{MH} = \frac{\sum_{i=1}^I \frac{a_i N_{0i}}{T_i}}{\sum_{i=1}^I \frac{b_i N_{1i}}{T_i}}$$

- with variance equal to:

$$\hat{Var} \left(\log(\hat{IRR}) \right) = \frac{A}{BC}$$

- where:

$$- A = \sum_{i=1}^I \frac{M_{1i} N_{1i} N_{0i}}{T_i^2}, \text{ the variance of the } \hat{IRR}_{MH}$$

$$- B = \sum_{i=1}^I \frac{a_i N_{0i}}{T_i}, \text{ numerator of } \hat{IRR}_{MH}$$

$$- C = \sum_{i=1}^I \frac{b_i N_{1i}}{T_i}, \text{ denominator of } \hat{IRR}_{MH}$$

3.2. M-H Equation.

- the multivariate model implicitly assumed by the Mantel-Haenszel analysis is:

$$I(t|A, L_1, L_2, L_3) = I_{0jkl}(t) \times e^{\beta_1 A(t)}$$

$$I(t) = I_0(t) \times e^{\beta_1 A(t) + \beta_2 L_1 + \beta_3 L_2 + \beta_4 L_3 + \beta_5 L_1 L_2 + \beta_6 L_1 L_3 + \beta_7 L_2 L_3 + \beta_8 L_1 L_2 L_3}$$

- if interested in a dichotomous outcome, MH can be generalized to logistic regression where an OR_{MH} is estimated, which approximates the IRR under the assumption of a rare outcome
- baseline varies over time t , but other measures of time can be included into L (ex: time of day of blood draw, season of blood draw, time sample stored in freezer, etc)

3.3. Assumptions.

- the baseline hazard $I_{0jkl}(t)$ is correctly specified (not really an assumption because it is a saturated model)
- the model is saturated for possible confounders (L_1 , L_2 , L_3) and interactions between confounders, making it robust against model misspecification
 - avoids potential bias, but not statistically efficient
- there is no effect modification between treatment A and confounders L , that is the effect of A is constant across strata of L
- there is no effect modification between treatment A and time t , that is the effect of A is constant over time
- there is no residual confounding due to L , no unmeasured confounding due to U , no selection bias, and no information bias

3.4. Parameter Interpretation.

- β_1 is the combined log M-H incidence rate ratio comparing those on treatment ($A = 1$) to those not on treatment ($A = 0$) holding the level of L_1 , L_2 , and L_3 constant
- $\beta_2 - \beta_8$ are implied by the M-H model, but not estimated in the stratified analysis

4. Unconditional Logistic Regression

4.1. Likelihood Function.

- the likelihood function for the unconditional logistic regression is:
- $L(\theta|A, x_1, \dots, x_p) = \prod_i \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}$

4.2. Regression Equation.

- $Pr(Y = 1|x_1, \dots, x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}$, or
- $\text{Logit}[Pr(Y = 1|x_1, \dots, x_p)] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

4.3. Assumptions.

- The mean of the response variable ($\text{Logit}[Pr(Y = 1)]$) is linearly related to the explanatory variables (x_1, \dots, x_p)

4.4. Parameter Interpretation.

- β can be interpreted as log odds ratio comparing $X=1$ to $X=0$ (binomial variable) or log odds ratio associated with a one unit change of X (continuous variable) controlling for other variables
- $\exp(\beta)$ is the odds ratio
- the effect estimates only has causal interpretation within the level of covariates, unless assuming constant effects across level of covariates

4.5. Relationship between log-linear model and logistic regression.

- log-linear model for number of cases in j^{th} risk set: $\log(E[X_j]) = \beta_1 + \beta_2 Z_{1j} + \log(t_j)$
- log-linear model for number of cases in j^{th} risk set based on covariate Z_1 : $\log(E[Y_j]) = \gamma_1 + \log(t_j)$
- logistic regression in j^{th} risk set:

$$\log(E[X_j]) - \log(E[Y_j]) = (\beta_1 - \gamma_1) + \beta_2 Z_{1j}$$

$$\log(E[X_j]/E[Y_j]) = \beta'_1 + \beta_2 Z_{1j}$$

$$\log(E[P_j]/(1 - E[P_j])) = \beta'_1 + \beta_2 Z_{1j}$$

5. Poisson

(Log-Linear)

- models *average count*, adjusted for length of follow up time when an offset is used
 - in this situation parameters have interpretation as average log incidence rates
 - not instantaneous incidence rate... e.g. hazard
- akin to a person-time analysis

A review of Poisson Distribution.

- Let $Y = \#$ events, $\mu =$ average $\#$ events, $T =$ follow up time, $\lambda =$ incidence rate
- if $Y \sim \text{Poisson} (\mu = \lambda T)$ then it's expected value depends on the length of follow up and a constant incidence rate...

$$E(Y) = \mu = \lambda T$$

- the probability mass function is given as...

$$Pr[Y = y] = (\mu^y e^{-\mu})/Y!$$

- also $\Delta T = \lim_{\Delta \rightarrow 0} t < k < t + \delta$
- thus this distribution assumes that (a) within fine intervals the event is rare (b) the total number of events is dependent on (i) a constant incidence rate and (ii) the length of follow up...

- (1) $P(Y = 1|\Delta T) \approx \lambda T$
- (2) $P(Y \geq 2|\Delta T) \approx 0$
- (3) $P(Y = 1|\Delta T)$ is constant over time
- (4) events occur independently over time
- (5) constant incidence rate within strata (if stratified)

5.1. Likelihood Function.

- we use the PMF to construct the likelihood function

$$L(\beta_0, \beta_1) = \prod_{i=1} P(Y_i = y_i | \lambda_i, T_i = t_i, X = x_i) = \prod_{i=1} \left((\mu_i^{y_i} e^{-\mu_i}) / y_i! \right)$$

where ...

$$\mu_i = \exp[\beta_0 + \beta_1 + \log(t_i)] \quad \text{and} \quad \log(t_i) = \text{offset for subject } i$$

- to see that the likelihood assumes a constant incidence rate we can write

$$L(\beta_0, \beta_1) = \prod_{i=1} P(Y_i = y_i | \lambda_i, T_i = t_i, X = x_i) = \prod_{i=1} \left((\lambda_i \times T_i)^{y_i} e^{-(\lambda_i \times T_i)} / y_i! \right)$$

where ...

$$\lambda_i = \exp(\beta_0 + \beta_1 X_i)$$

- note that the constant incidence rate can vary by covariates in X
 - thus we can *encode* a time-varying incidence rate by explicitly including $f(t)$ as a covariate in X
- note also that individuals can have different baseline rates that are *constant* over follow up
- the likelihood allows us to model an expected counts μ or an incidence rate λ

5.2. Regression Equation.

- (1) to model a constant incidence rate that depends only on covariate values...

$$\lambda_{A,Z} = \exp(\underbrace{\beta_0}_{\text{baseline rate}} + \beta_1 A + \beta_2 Z)$$

- in terms of a generalized linear model

link: *log* , distribution: *Poisson*

- an offset term is applied (later...)

- (2) to model a incidence rate that varies over time, we must specify it's function

(a) piecewise:

$$\lambda(t)_{A,Z} = \exp \left[\underbrace{\beta_0 + \beta_2 \sum_{k=1}^K I(G = k) + \beta_3 A + \beta_4 Z}_{\text{baseline rate}} \right]$$

where ...

$$k = \text{kth time interval} \implies \{1, 2, \dots, K\}$$

(b) log-linear:

$$\lambda(t)_{A,Z} = \exp \left(\underbrace{\beta_0 + \beta_1 t}_{\text{baseline rate}} + \beta_2 A + \beta_3 Z \right)$$

(c) fractional polynomial:

$$\lambda(t)_{A,Z} = \exp \left(\underbrace{\beta_0 + \beta_1 t + \beta_2 t^{-2} + \beta_3 t^2}_{\text{baseline rate}} + \beta_4 A + \beta_5 Z \right)$$

(d) non-parametric smoothing:

$$\lambda(t)_{A,Z} = \exp \left[\underbrace{\beta_0 + g(\theta, t)}_{\text{baseline rate}} + \beta_1 A + \beta_2 Z \right]$$

What & Where is the Offset?

- the random variable in a Poisson distribution is the count of events
- an offset is the log of the total amount of follow up time for subject i
- we add it to the model to give a rate ratio interpretation to the parameters even though we're modeling the average count of events
- it 'standardizes' the counts according to differential follow up by encoding the unit-specific amount of follow-up time
- recall that $\mu = \lambda T \implies \log(\mu) = \log(\lambda) + \log(T)$

$$\begin{aligned} \log(\lambda_i) &= \beta_0 + \beta_1 X_i \\ \log(\mu_i/t_i) &= \beta_0 + \beta_1 X_i \\ \log(\mu_i) - \log(t_i) &= \beta_0 + \beta_1 X_i \\ \log(\mu_i) &= \underbrace{\beta_0 + \beta_1 X_i}_{\text{model}} + \underbrace{\log(t_i)}_{\text{offset}} \end{aligned}$$

Parameter Interpretation.

$$h(t) = \exp(\beta_0 + \beta_1 f(t) + \beta_2 X_1)$$

- $\beta_0 + \beta_1 f(t)$ is the log baseline incidence rate at time t for group $X = x$
- β_2 is the log hazard ratio of time to outcome for a 1 unit change in X

Constant Incidence Rate Assumption.

$$\log(\mu_i) = \beta_0 + \beta_1 X_i + \log(t_i)$$

$$\log(\mu_i) = \log(\lambda_i) + \log(t_i)$$

$$\mu_i = \lambda_i \times t_i$$

$$E(Y_i) = \lambda_i \times t_i$$

5.3. Assumptions.

- the **incidence rate is homogeneous** for every time t within strata formed by covariates
 - implies proportional rate (i.e. constant β over follow up)
 - baseline incidence rate varies as a function of covariates
- ⇒ λ_0 **varies by time only if $f(t)$ is explicitly specified** in the model
 - assumes that you know the true form of the function $f(t)$
 - can cause bias if it is mis-specified
- best for Poisson processes
 - outcome is rare within strata formed by covariates

5.4. Comparisons.

- unlike cox
 - models average incidence rate (i.e. not hazard)
 - estimates the rate by counting events within strata and dividing by total person-time within strata
 - * *does not use risk sets!*
 - * *but approaches them if modeling time very finely*
 - * *is not based on ranks, so the time scale is only accounted for by specifically parameterizing it in the model*
 - estimates baseline incidence rate
- like cox
 - uninformative strata are where strata have no events

5.5. Practical Issues.

- **advantages**
 - can obtain absolute rates → estimate rate difference
 - natural extension of non-parametric person-time analysis
- **disadvantages**

- constant incidence rate over time is a strong assumption, unlikely to hold
- model is vulnerable to mis-specification of $\lambda(t)_0$ via $f(t)$ in model
 - * can induce residual confounding by covariates included in the model

6. Pooled Logistic Regression

6.1. likelihood function.

- $L(\beta_0, \dots, \beta_p | x_1, \dots, x_{p,t_{i-1}}) = \prod_{t=1}^{i-1} \frac{\exp(\beta_0 + \dots + \beta_p x_{p,t_{i-1}})}{1 + \exp(\beta_0 + \dots + \beta_p x_{p,t_{i-1}})}$

6.2. equivalence between pooled logistic model and Cox model.

- an intuitive way to show the equivalence between pooled logistic model and Cox model

$$(1) \text{ logit}[Pr(Y_{t_i} = 1 | Y_{t_{i-1}} = 0, X_{t_{i-1}})] = \underbrace{\beta_0 + \beta_{t_i} t_i}_{\text{baseline rate}} + \beta_1 X_{1,t_{i-1}} + \dots + \beta_p X_{p,t_{i-1}}$$

$$\Rightarrow \log \frac{Pr(Y_{t_i} = 1 | Y_{t_{i-1}} = 0, X_{t_{i-1}})}{1 - Pr(Y_{t_i} = 1 | Y_{t_{i-1}} = 0, X_{t_{i-1}})} = \underbrace{\beta_0 + \beta_{t_i} t_i}_{\text{baseline rate}} + \beta_1 X_{1,t_{i-1}} + \dots + \beta_p X_{p,t_{i-1}}$$

- (2) if the disease is rare in the person-time unit used in the model, the model can be reduced to the following model

$$\log[Pr(Y_{t_i} = 1 | Y_{t_{i-1}} = 0, X_{t_{i-1}})] = \underbrace{\beta_0 + \beta_{t_i} t_i}_{\text{baseline rate}} + \beta_1 X_{1,t_{i-1}} + \dots + \beta_p X_{p,t_{i-1}}$$

- (3) to see the correspondence between pooled logistic model and Cox model

$$(4) \text{ Cox model: } I(t) = I_0(t) e^{\beta_1 X_{1,t_{i-1}} + \dots + \beta_p X_{p,t_{i-1}}}$$

$$\Rightarrow \log[I(t)] = \log[I_0(t)] + \beta_1 X_{1,t_{i-1}} + \dots + \beta_p X_{p,t_{i-1}}$$

$$(a) I(t) \Rightarrow Pr(Y_{t_i} = 1 | Y_{t_{i-1}} = 0, X_{t_{i-1}})$$

$$(b) I_0(t) \Rightarrow \beta_0 + \beta_{t_i} t_i$$

- (5) $\beta_0 + \beta_{t_i} t_i$ determines the modeling of baseline rate ($I_0(t)$)

6.3. regression equation.

- $\text{logit}[Pr(Y_{t_i} = 1 | X_{t_{i-1}})] = \beta_0 + \beta_1 X_{1,t_{i-1}} + \dots + \beta_p X_{p,t_{i-1}}$

$$\log \left[\frac{Pr(Y_{t_i} = 1 | Y_{t_{i-1}} = 0, X_{t_{i-1}})}{1 - Pr(Y_{t_i} = 1 | Y_{t_{i-1}} = 0, X_{t_{i-1}})} \right] = \beta_0 + \beta_1 X_{1,t_{i-1}} + \dots + \beta_p X_{p,t_{i-1}}$$

$$\frac{Pr(Y_{t_i} = 1 | Y_{t_{i-1}} = 0, X_{t_{i-1}})}{1 - Pr(Y_{t_i} = 1 | Y_{t_{i-1}} = 0, X_{t_{i-1}})} = e^{\beta_0 + \beta_1 X_{1,t_{i-1}} + \dots + \beta_p X_{p,t_{i-1}}}$$

$$\frac{I(t_i)}{1 - I(t_i)} = e^{\beta_0 + \beta_1 X_{1,t_{i-1}} + \dots + \beta_p X_{p,t_{i-1}}}$$

$$I(t_i) = \frac{e^{\beta_0 + \beta_1 X_{1,t_{i-1}} + \cdots + \beta_p X_{p,t_{i-1}}}}{1 + e^{\beta_0 + \beta_1 X_{1,t_{i-1}} + \cdots + \beta_p X_{p,t_{i-1}}}} \approx I_0(t) \times \frac{e^{\beta_1 X_{1,t_{i-1}} + \cdots + \beta_p X_{p,t_{i-1}}}}{1 + e^{\beta_1 X_{1,t_{i-1}} + \cdots + \beta_p X_{p,t_{i-1}}}}$$

- the $\text{logit}(I(t_i))$ is a linear combination of the covariates in the model
- this is a multiplicative model on the log of the odds of the incidence rate scale
- $\log(\frac{I(t)}{1-I(t)}) \approx \log(I(t))$ when the disease is rare in the person-time unit used in the model
- we may use pooled logistic model to estimate incidence rate ratio
- we may also estimate incidence rate because the β_0 was specified in the model
 - models with different baseline rate parameter

$$\text{logit}[Pr(Y_{t_i} = 1|X_{t_{i-1}})] = \underbrace{\beta_0}_{\text{baseline rate}} + \beta_1 X_{1,t_{i-1}} + \cdots + \beta_p X_{p,t_{i-1}}$$

$$\text{logit}[Pr(Y_{t_i} = 1|X_{t_{i-1}})] = \underbrace{\beta_0 + \beta_{t_i} t_i}_{\text{baseline rate}} + \beta_1 X_{1,t_{i-1}} + \cdots + \beta_p X_{p,t_{i-1}}$$

$$\text{logit}[Pr(Y_{t_i} = 1|X_{t_{i-1}})] = \underbrace{\beta_0 + \beta_{t_{2,i}} t_{2,i} + \cdots + \beta_{t_{q,i}} t_{q,i}}_{\text{baseline rate}} + \beta_1 X_{1,t_{i-1}} + \cdots + \beta_p X_{p,t_{i-1}}$$

6.4. assumptions.

- short intervals for the grouping of outcome events
- the incidence rate is small (less than 0.1) in the interval (usually the measured person-time units)

7. Conditional Logistic Regression

7.1. likelihood function.

- the likelihood function for the conditional logistic regression for matched data set with k strata and M controls for each case is

$$\prod_{k=1}^K \frac{\exp(\alpha_k + \sum_{i=1}^I \beta_i x_{k0i})}{\sum_{j=0}^M [\exp(\alpha_k + \sum_{i=1}^I \beta_i x_{kji})]} \quad \text{which is simplified to...}$$

$$\prod_{k=1}^K \frac{1}{1 + \sum_{j=1}^M \exp[\sum_{i=1}^I \beta_i (x_{kji} - x_{k0i})]}$$

7.2. regression equation.

- the typical regression equation for a conditional logistic model of a binary outcome Y a covariate vector X and I matched strata is $\text{logit}Pr[Y = 1] = \alpha_k + \beta x$ where the case/total ratios per stratum will equal

$$\text{logit}(A_k(X)/N_k(X)) = \alpha_k^* + \beta x \quad \text{where...}$$

$\alpha_k^* = \alpha_k + \ln(\frac{f_k}{h_k})$ and f_k and h_k are the sampling fractions for cases and controls in strata k

7.3. assumptions.

- assumptions of logistic regression plus the assumption of fixed 2x2 table margins (*hypergeometric distribution*)
- data probabilities conditioned on the observed margins
- no effect modification across matched strata

7.4. parameter interpretation.

- for the simple regression with outcome Y and exposure A
 $\text{logitPr}[Y = 1] = \beta_{0k} + \beta_1 A$
 - the parameter β_{0k} in the conditional logistic model is an intercept term that varies for each matched stratum k
 - $\beta_1 = \log(OR)$ of the effect of the exposure A on outcome Y, $\exp(\beta_1) = OR$ approximates hazard ratio
 - Note that the regression equation does not include parameters for matching factors as the effects of matching factors on the outcome Y cannot be computed. The stratum-specific effects of the matching factors are included in the intercept terms.
 - however, we can assess effect modification by matching factors *can* be assessed by
 - * modeling exposure classified according to matching factors.
 - * including a product term between exposure and the matching factor

7.5. practical issues.

- the conditional maximum likelihood (CML) method is restricted to multiplicative-intercept models for rates and odds
- the conditional likelihood function is very similar to the partial likelihood function used in the Cox model
- when time is part of the matching factors as for example is the case in nested case control risk set sampling, then the conditional logistic model is algebraically equivalent to a cox-model and can use phreg and estimate hazard ratios that will approximate true HR as the matching ratios approach infinity
- conditional logistic models are also used in case-crossover studies (each individual case is a stratum) and also approximate true hazard ratios

8. Cox-Proportional Hazards Regression

8.1. counting process.

- in the Anderson-Gill formulation, the time interval is considered as $(start, end]$ (not to include left hand side, include right hand side)
- (Anderson-Gill) counting process data structure is suitable for the follow scenarios
 - (1) time dependent covariate
 - (2) recurrent event
 - (3) alternative time scale
 - (4) time dependent strata
 - (5) discontinuous intervals of risk
- all study subjects who have entered the study in the same day or earlier as the current case, and have not died from another cause (competing risk) or been lost to follow-up by the day that the case became a case, would belong to the case's risk set.

8.2. likelihood function.

$$I(t) = I_0(t)e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}$$

- the major advantage of Cox proportional hazards model is that it does not require estimating the baseline rate $I_0(t)$
- it is similar to calculating rate ratio in each strata of time and summarized these rate ratios by M-H estimate
- Cox model is a semi-parametric model
 - we do not specify the parameters of baseline incidence rate $I_0(t)$ is non-parametrically modeled)
 - we do specify the parameters for the covariates in the model (the effect of X_1, X_2, \dots, X_p are parametrically model)
 - the combination of the non-parametric model of $I_0(t)$ and the parametric model of X_1, X_2, \dots, X_p is a semi-parametric model
- the magic of Cox model is that we can estimate parameters of the covariates without estimating the non-parametric baseline rate by using partial likelihood function
- the partial likelihood function

- the likelihood function at the distinct failure time t_j when individual j fails

$$\begin{aligned}
 L_j(\beta) &= \Pr(\text{individual with covariate value } x_j \text{ fails at } t_j | 1 \text{ individual fails from the risk set at } t_j) \\
 &= \frac{\Pr(\text{individual with covariate value } x_j \text{ fails} | \text{in the risk set } R_{t_j})}{\sum_{i \in R_{t_j}} \Pr(\text{individual with covariate value } x_i \text{ fails} | \text{in the risk set } R_{t_j})} \\
 &= \frac{\Delta t \times \lambda(t_j; Z_j)}{\sum_{i \in R_{t_j}} \Delta t \times \lambda(t_j; Z_i)} \\
 &= \frac{\Delta t \times \lambda(t_j; Z_j)}{\Delta t \times \sum_{i \in R_{t_j}} \lambda(t_j; Z_i)} \\
 &= \frac{\lambda(t_j; Z_j)}{\sum_{i \in R_{t_j}} \lambda(t_j; Z_i)}
 \end{aligned}$$

- under the proportional hazards assumption $\lambda(t; Z) = \lambda_0 e^{\beta Z}$
- and the assumption that these individual likelihoods are independent from each other given the risk set

$$L^{partial}(\beta) = \prod_{j=1}^K \frac{\lambda_0(t_j) e^{\beta Z_j}}{\sum_{i \in R_{t_j}} \lambda_0(t_j) e^{\beta Z_i}} = \prod_{j=1}^K \frac{e^{\beta Z_j}}{\sum_{i \in R_{t_j}} e^{\beta Z_i}}$$

8.3. regression equation.

- $I(t) = I_0(t) e^{\beta_1 X_1} \times e^{\beta_2 X_2} \times \dots \times e^{\beta_p X_p} = I_0(t) e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}$

8.4. assumptions.

- in terms of PH model: the hazard ratio is constant over time = the hazard functions (e.g. $I_0(t)e^\beta$ and $I_0(t)$) are parallel to each other
- in terms of M-H method: there is no effect modification by time for the rate ratio
- the proportional hazards assumption can be tested by testing the interaction term between time and covariates

8.5. comparisons.

- advantages of Cox model:
 - (1) do not need to specify and estimate the baseline rate
 - (2) less vulnerable to model misspecification
- disadvantages of Cox model:
 - (1) cannot estimate baseline rate directly

8.6. parameter interpretation.

- $I(t)$ = the incidence rate at time t for a subject with values $X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$
- $I_0(t)$ = the baseline incidence rate at time t
= the incidence rate at time t for a subject with values $X_1 = 0, X_2 = 0, \dots, X_p = 0$
- β_1 = the log rate ratio for a one unit change in X_1
e.g. β_1 = the log rate ratio for death comparing male to female
e.g. β_2 = the log rate ratio for death for one unit (mmHg) increase in blood pressure
- the 95% confidence interval of β is not symmetric since that it is $e^{\hat{\beta} \pm 1.96 \times se(\hat{\beta})}$

9. Accelerated Failure Time Regression**9.1. likelihood function.****9.2. regression equation.****9.3. assumptions.****9.4. parameter interpretation.****9.5. comparisons.****9.6. practical issues.**

TABLE 16.2. Summary of Regression Models in Epidemiology

Model	Assumptions	Strengths	Weaknesses
<i>Unconditional Logistic relative risk</i>	Fixed follow up, constant risk within strata, time-to-event not important, odds a linear fxn of covariates on exp scale, independence	In matched case-control, if OR does not vary by matching factor can reduce assumptions about matching factors (collapse them) or remove them entirely; estimate baseline risks if had sampling fractions and external data	CASE-CONTROL: can be inefficient in matched case-control. COHORT: not appropriate for open cohorts (staggered entry, loss to follow up) or time trends in risk (variable induction periods)
<i>Poisson (log-linear) rate ratio</i>	Open follow up, constant rate within strata, rate a linear fxn of covariates on exp scale, Poisson process: rare outcome, fine strata, independence	Model baseline rate; censoring not an issue; good for rare outcomes; can test if baseline rate varies over time; can get effect for time-scale	Could mis-specify baseline rate, not as efficient as cox
<i>Conditional Logistic matched relative risk</i>	Within strata...(see logistic)	Efficient for matched case-control studies, analyzes in risk-sets; can still test if OR varies by matching factors	No parameter estimates for baseline odds or main effect of matching factors; harder to reduce assumptions about matching factors
<i>Pooled Logistic hazard ratio</i>	COHORT: Outcome is rare over narrowly defined intervals	Model baseline rate	Must account for censoring, could mis-specify baseline
<i>Cox hazard ratio</i>	Hazards are proportional, risk-sets are independent, uniformative censoring	Efficiently, does not model baseline rate, common outcomes	PH assumption may not hold; no absolute rates; cannot test if baseline rate varies; can't get effect for time-scale

CHAPTER 17

Models for Statistical Inference: Special Topics

1. Model Selection for Case-control Data

1.1. Incidence Density Sampling.

- unconditional logistic regression is the choice of model for typical density sampling
- the OR estimates of the exposure are approximates for the rate ratio of the underlying cohort

1.2. Cumulative Incidence Sampling.

- logistic models fit with OR approximating a risk ratio only under the rare disease assumption
- sometimes additive logistic models are used (linear OR models)

1.3. Matched Case-Control.

- conditional logistic regression is used with the OR approximating Rate Ratios
- in nested case-control studies (risk-set sampling) where participants are also matched on time the conditional logistic model and Cox model are equivalent and cox model can be fit conditioning out the stratum intercepts and directly calculate hazard ratios (OR from conditional logistic regression approximates HR as the sampling ratio approaches infinity)

1.4. Case-crossover.

- conditional logistic regression with each person representing a stratum
- effect estimates interpreted as hazard ratio (need exchangeability of case-control periods)

1.5. Case-Cohort.

- risk set method allows for cox-model (see nested-case control)

2. Covariate Selection

- Problems that will occur when the model fitted to the data is incorrect
 - standard errors will be biased
 - power to detect an exposure-response relationship will be reduced
 - the effects of extremes of the exposure distribution will be more biased than estimates of exposure effect near the center of the exposure distribution
 - transforming the exposure or adding higher-order interaction terms will not always be able to overcome the bias due to model mis-specification.
- Situations when controlling for all possible confounders cause more bias than it prevents
 - Finite samples with sparse data. Large samples can have sparse data issues as well if the number of confounders is large.
 - Incorrectly specified model, such as adjusted for intermediates
- Covariates selection
 - All known risk factors should be treated as potential confounders and included in the model.
 - Inclusion of variables associated with exposure but not outcome will not introduce bias, but will decrease efficiency (thus making it undesirable)
 - Inclusion of variables associated with outcome but not associated with exposure will increase efficiency, and will not introduce bias.
- Covariates selection algorithm
- **Deluxe algorithm**
 - Start with highly saturated model
 - Drop risk factors weakly associated with outcome ($p > 0.2$); ie. collapsing 2x2 table on the variable in stratified analysis or deleting the variable with all its higher order interaction terms
 - Check change of effect estimate, keep the variable if change $> 10\%$
- $$BIAS = \frac{OR_{reduced} - OR_{saturated}}{OR_{saturated}} * 100$$
- **Bargain basement algorithm**
 - Start by including the main effects only of all known risk factors.
 - Identify suspected risk factors with $p < 0.20$ in crude analysis with the outcome.
 - Drop the variables weakly associated with outcome ($P > 0.2$) in the multivariate model.

- Keep the variable by 10% change of effect estimate rule.
- Assumption: Higher-order interactions terms will be unimportant if each association of the potential confounders is multiplicative.
- Bias traded off with efficiency in the selection of confounders
 - A model that includes all potential confounders may get least biased results
 - When the covariates is relatively large to the sample (greater than 1 covariates for 10 events), the confidence intervals will become so large that we cannot infer anything useful about the parameter of interest
 - Remove some covariates may increase efficiency, but may increase bias if we missed important confounders
 - In either situation, mean squared error may increase because mean squared error is equal to the variance of the estimator plus the square of the bias. Thus, both efficiency and bias are taken into account in mean squared error.

3. Stratified Sampling Fractions

3.1. Formulation.

- regular log-linear model where x is expected number of cases over follow up t_j

$$\log E(X_j) = \beta_0 + \beta_1 z_{1j} + \log t_j$$
- now if we let γ_0 be sampling fraction when $Z = 0$ and $\{\gamma_0 + \gamma_1\}$ be the sampling fraction for when $Z = 1$ then...

$$\log \left[\frac{E(Y_j)}{E(t_j)} \right] = \gamma_0 + \gamma_1 z_1$$
- then we can model sampled controls (days) Y as a function of sampling fraction γ that depends on covariate z_j and the corresponding amount of follow up time

$$\log E(Y_j) = \gamma_0 + \gamma_1 z_{1j} + \log t_j$$

- subtracting we get

$$\begin{aligned} \log \left[\frac{E(X_j)}{E(Y_j)} \right] &= (\beta_0 - \gamma_0) + (\beta_1 - \gamma_1) z_{1j} \\ &= \beta'_0 + \beta'_1 z_{1j} \end{aligned}$$

- because Y_j is number of controls, we have essentially a logit model
 - where p_j is observed fraction of subjects at covariate level i who are cases
 - if we condition on the total number of cases and controls at the covariate level...

$$\log \left[\frac{E(P_j)}{1 - E(P_j)} \right] = \beta'_0 + \beta'_1 z_{1j}$$

- β_0 is not baseline odds of outcome, but we can recover it if we know the sampling fractions
- note: if one uses stratified sampling fractions, should always **include the stratifying factors in the analysis**
- to simulate the entire cohort, can either
 - include the sampling fraction as an the offset in a Poisson model as shown above
 - weight the cases and controls by the inverse of their covariate-specific sampling fraction

CHAPTER 18

Models for Causal Inference

1. Parametric G-Formula

$$E[Y^a] = \sum_l E[Y|A = a, L = l] Pr[L = l]$$

- estimate $E[Y|A, L]$ with parametric model
- null paradox (always get effect)
- even if the causal null hypothesis is true, effect estimate will not be null
- model misspecification
- need model for each time point and thus the risk of model misspecification is high
- Strength
 - (1) Appropriately adjusts for time-varying confounding affected by prior exposures although subject to the g-null paradox.
 - (2) Naturally handles joint interventions and dynamic interventions
 - (3) Estimates multiple parameters (including risk ratios and risk differences) and yields population estimates.
- Weakness
 - (1) Requires models for covariates as well as outcome
 - (2) More sensitive to violations of assumptions of model misspecification (A violation in one of the multiple models may reverberate throughout the others)
 - (3) Not implemented in packaged statistical software

2. Marginal Structural Models

- Conditional versus Marginal Models
 - conditional: $E[Y|A, L] = \theta_0 + \theta_1 A + \theta_2 L$
 - marginal: $E[Y^a] = \beta_0 + \beta_1 a$
 - marginal *and* conditional: $E[Y^a|L] = \beta_0 + \beta_1 a + \beta_2 L + \beta_3 aL$, used to assess effect modification

- *marginal*:
 - model the marginal distribution of counterfactual random variables
 - not modeling observed associations
- *structural*:
 - model the probabilities of counterfactual variables
- *saturated*:
 - unless exploring effect modification
 - can be regarded as a weighted crude analysis of the study population
- parameters are estimated via inverse probability weighted regression

$E[Y^a|a] = \gamma_0 + \gamma_1 a$ is estimated by ...

$$E[Y|A] = \beta_0 + \beta_1 A \text{ with weights } \frac{1}{f[A|L]}$$

- when identifiability conditions hold...

$$\beta_0 = \gamma_0 \text{ and } \beta_1 = \gamma_1$$

2.1. MSMs for Constant Exposures.

Estimation.

$$W^A = \frac{1}{Pr[A = a|L = l]}$$

- we need to simulate the population in the counterfactual world using IP weight, then use the marginal structural model with the *simulated data* (pseudo-population)
- for exposures as *dichotomous* variables, estimate $Pr[A = a|L = l]$ using logistic regression. e.g....

$$\begin{aligned} \text{logit}[Pr[A = a|L, M, N]] &= \theta_0 + \theta_1 L + \theta_2 M + \theta_3 N + \theta_4 LM \\ &\quad + \theta_5 LN + \theta_6 MN + \theta_7 LMN \quad (\text{saturated model}) \end{aligned}$$

- estimate weights W^A with logistic regression

$$\begin{aligned} \text{logit } Pr[A = 1|L] &= \theta_0 + \theta_1 L \\ Pr[A = 1|L] &= \left(\frac{e^{\theta_0 + \theta_1}}{1 + e^{\theta_0 + \theta_1}} \right) \\ Pr[A = 0|L] &= \left(1 - \frac{e^{\theta_0 + \theta_1}}{1 + e^{\theta_0 + \theta_1}} \right) \\ W^A &= \frac{1}{Pr[A = a|L = l]} \end{aligned}$$

- model $E[Y|A]$ in study population with weights W^A

- use robust variance (sandwich estimator) or bootstrap because counterfactual outcomes are correlated within individuals (recall ϵ_{Y^a} in twin DAG)

Parameter Interpretation.

$$\begin{aligned}
 E[Y^a] &= \beta_0 + \beta_1 a \\
 E[Y^{a=0}] &= \beta_0 \\
 E[Y^{a=1}] &= \beta_0 + \beta_1 \\
 E[Y^{a=1}] - E[Y^{a=0}] &= \beta_1
 \end{aligned}$$

- β_0 is the average outcome if the entire population had not been treated
- $\beta_0 + \beta_1$ is the average outcome if the entire population had been treated
- β_1 = causal effect of treatment
- depends on model form and specification
 - *linear* \rightarrow causal difference in outcome per unit change in exposure
 - *logistic* \rightarrow causal odds ratio
 - *log-linear* link \rightarrow causal incidence rate ratio

Stabilized Weights.

$$SW^A = \frac{Pr[A = a]}{Pr[A = a | L = l]}$$

- stabilizes variance of causal effect
- by preventing persons with $Pr[A|L] \approx 0$ from dominating the pseudo-population
- **necessary** for continuous exposures

Subset of Population Treated with $A = a'$.

$$SW^{A'} = \frac{Pr[A = a' | L]}{Pr[A = a | L]}$$

- note: when $a' = 1$, causal effect is analogous to *SMR*
 - for treated $\rightarrow SW^{A'} = \frac{Pr[A=1|L]}{Pr[A=1|L]=1} \rightarrow 1$
 - for untreated $\rightarrow SW^{A'} = \frac{Pr[A=0|L]}{Pr[A=1|L]} = 1 \rightarrow P/(1-P)$
- note: when $a' = 0$, causal effect is analogous to *SRR*
 - for treated $\rightarrow SW^{A'} = \frac{Pr[A=0|L]}{Pr[A=1|L]} = 1 \rightarrow (1-P)/P$
 - for untreated $\rightarrow SW^{A'} = \frac{Pr[A=0|L]}{Pr[A=0|L]} = 1 \rightarrow 1$

Effect Modification.

- we can model effect modification of causal effect estimates from marginal structural models

$$E[Y^a, M] = \beta_0 + \beta_1 a + \beta_2 M + \beta_3 aM$$

$$\begin{aligned}
 E[Y^{a=0}|M=0] &= \beta_0 \\
 E[Y^{a=0}|M=1] &= \beta_0 + \beta_2 \\
 E[Y^{a=1}|M=0] &= \beta_0 + \beta_1 \\
 E[Y^{a=1}|M=1] &= \beta_0 + \beta_1 + \beta_2 + \beta_3
 \end{aligned}$$

$$\begin{aligned}
 E[Y^{a=1}|M=1] - E[Y^{a=0}|M=1] &= \beta_1 + \beta_3 \\
 E[Y^{a=1}|M=0] - E[Y^{a=0}|M=0] &= \beta_1
 \end{aligned}$$

Adjustment for Censoring.

$$SW^C = \frac{Pr[C=0|A=a]}{Pr[C=0|A=a, L=l]}$$

- where $C = 1$ if the subject was lost to follow-up
- censoring is conceptualized as just another treatment
- interest is in the causal effect of treatment A had all subjects remained uncensored ($C = 0$)
- requires no unmeasured confounders of both treatment A and “treatment” C (censoring)
- weighted regression model is restricted to uncensored subjects where the subject specific weight $w_i = SW_i^A \times SW_i^C$
- $IPW = \text{standardization non-parametrically}$

these two methods use different part of the joint probability of random variables on a causal DAG represented by Makcov factorization:

$$\begin{aligned}
 f(Y, A, L) &= f(Y|A, L)f(A|L)f(L) \\
 \text{IPW : } &\frac{1}{f(A|L)} \\
 \text{standardization : } &f(Y|A, L)f(L)
 \end{aligned}$$

- assumptions needed for parametric MSM:
 - (1) identifying conditions (especially *positivity* . . .)
 - structural violation of positivity: subjects with certain confounder values cannot possibly be exposed/unexposed ($Pr[A=1|L=l] = 0$ or $Pr[A=0|L=l] = 0$)
 - random violation of positivity: sparse data problem ($\hat{Pr}[A=1|L=l] = 0$ or $\hat{Pr}[A=0|L=l] = 0$)

- using parametric model to estimate IP weight is essentially **assuming random non-positivity**
- (2) assumptions implied by the IPW model (e.g. interaction between Ls)
 - (3) assumptions implied by the marginal structure model (e.g. $cum(\bar{a}_i)$)

2.2. MSMs for Time-Varying or Joint Exposures.

Estimation.

- estimate weights W^A with pooled logistic regression

$$\begin{aligned} \text{logit } Pr[A = 1 | \bar{L}_k, A_{\bar{k}-1}] &= g(\bar{L}_k, \bar{A}_{k-1}; \theta) \\ &= \theta_0 f(k) + \theta_i \sum_{k=0}^k L_k + \theta_j \sum_{k=0}^{k-1} A_i \\ Pr[A_k = 1 | \bar{L}_k, A_{\bar{k}-1}] &= \left(\frac{e^{g(\bar{L}_k, \bar{A}_{k-1}; \theta)}}{1 + e^{g(\bar{L}_k, \bar{A}_{k-1}; \theta)}} \right) \\ Pr[A_k = 0 | \bar{L}_k, A_{\bar{k}-1}] &= \left(1 - \frac{e^{g(\bar{L}_k, \bar{A}_{k-1}; \theta)}}{1 + e^{g(\bar{L}_k, \bar{A}_{k-1}; \theta)}} \right) \end{aligned}$$

$$W^A = \frac{1}{\prod_{k=0}^K Pr[A_k = a_{ki} | \bar{A}_{k-1} = \bar{a}_{(k-1)i}, \bar{L}_k = \bar{l}_{ki}]}$$

- estimation for each unit of person-time
- only need to estimate for times k where treatment changes because otherwise $Pr[A = 1 | \bar{L}_k, A_{\bar{k}-1}] = 1$
- subtle point: only need to establish identifiability conditions for times where treatment changes, all other times follow suit
- often simplify $f(\theta_j \sum_{k=0}^{k-1} \bar{A}_i)$ as $f(\bar{A}_k)$ and $h(\theta_i \sum_{k=0}^k L_k)$ as $h(\bar{L}_k)$
 - * requires subject matter considerations
 - * often necessary for estimation (i.e. positivity violation s & efficiency)
- model $E[Y | f(\bar{A})]$ in study population with weights W^A
- use robust variance or bootstrap because counterfactual outcomes are correlated within individuals (recall ϵ_{Y^a} in twin DAG)

Parameter Interpretation.

- as before *depends on model form and specification*

$$E[Y^{\bar{a}_k} | \bar{a}_k] = \beta_0 + \beta_1 f(\bar{a}_k)$$

- β_0 is the average counterfactual outcome if entire population untreated over entire follow up
 $\bar{a}_k = (A_0 = 0, A_1 = 0, \dots, A_k = 0)$ for all k
- $\beta_0 + \beta_1$ is the average counterfactual outcome if the entire population had been treated with regime \bar{a}_k
- interpretation of causal contrast depends on how $f(\bar{a}_k)$ is specified

Stabilized Weights.

$$SW^A = \frac{\prod_{k=0}^K Pr[A_k = a_{ki} | \bar{A}_{k-1} = \bar{a}_{(k-1)i}]}{\prod_{k=0}^K Pr[A_k = a_{ki} | \bar{A}_{k-1} = \bar{a}_{(k-1)i}, \bar{L}_k = \bar{l}_{ki}]}$$

- estimate of β_1 is unbiased even if the numerator model is misspecified
- all others must hold (exposure, censoring, and outcome)

Subset of Population: $\bar{A}_k = \bar{a}'_k$.

$$SW^A = \frac{\prod_{k=0}^K Pr[A_k = a'_{ki} | \bar{A}_{k-1} = \bar{a}'_{(k-1)i}, \bar{L}_k = \bar{l}_{ki}]}{\prod_{k=0}^K Pr[A_k = a_{ki} | \bar{A}_{k-1} = \bar{a}_{(k-1)i}, \bar{L}_k = \bar{l}_{ki}]}$$

Effect Modification (by V).

$$SW^A = \frac{\prod_{k=0}^K Pr[A_k = a_{ki} | \bar{A}_{k-1} = \bar{a}_{(k-1)i}, V = v]}{\prod_{k=0}^K Pr[A_k = a_{ki} | \bar{A}_{k-1} = \bar{a}_{(k-1)i}, \bar{L}_k = \bar{l}_{ki}]}$$

- V is a subset of $L_{k=0}$ (i.e. *pre-treatment* baseline covariates)
- assess effect modification by including V in marginal model for outcome
- efficiency gain if include in numerator of weights (recommended)
- **cannot** assess effect-modification by time-varying covariates (need SNM)

Adjustment for Censoring.

$$SW^C = \frac{\prod_{k=0}^{K+1} Pr[C_k = 0 | \bar{C}_{k-1} = 0, \bar{A}_{k-1} = \bar{a}_{(k-1)i}]}{\prod_{k=0}^{K+1} Pr[C_k = 0 | \bar{C}_{k-1} = 0, \bar{A}_{k-1} = \bar{a}_{(k-1)i}, \bar{L}_k = \bar{l}_{ki}]}$$

- where $C_k = 1$ if the subject was lost to follow-up
- interest is in the causal effect of treatment history \bar{A} had all subjects remained uncensored ($C = 0$)
- requires no unmeasured confounders of both treatment \bar{A} and “treatment” \bar{C} (censoring)
- weighted regression model is restricted to uncensored subjects where the subject specific weight equals $SW^A \times SW^C$

Artificial Censoring.

- for **static** regimes, weights for artificial censoring = treatment weights W^A
 - *more on this in chapter: advanced topics in causal inference*
- similar to ‘per-protocol’ analysis

2.3. practical issues.

- In practice, unconditional pooled logistic regression will be used to approximate a Cox proportional hazards model, because the PROC PHREG is unable to deal with the time-dependent weights required for the MSM.
- For pooled logistic regression, the baseline hazard (unspecified in Cox model) has to be specified.
- A function of time to model the baseline hazard, either single variable or polynomials, has to be included in pooled logistic models.
- MSMs use weights and create a correlated data structure of treatment and covariates, so GEE with sandwich variance estimator (PROC GENMOD with the option repeated) has to be used to obtain robust confidence intervals.
- To obtain both censoring and treatment stabilized weight, four four pooled logistic models are fit.
- In two of them the outcome will be censor and in the other two the outcome will be treatment
- For each outcome, the numerator model contains exclusively baseline covariates; and a denominator model contains baseline and time-dependent covariates.
- By using the predicted values of the models, we can estimate, for each person-day, the probability of being uncensored and the probability of being treated or untreated up to that time.
- The final weight SW for each person-day will be the product of the probabilities, for treatment up to time K, for treatment to time K+1

positivity violations.

- nonpositivity can be
 - random: result of a non-infinite sample and multiple confounders stratification different zero cells in different samples
 - structural (non random): a certain confounder profile cannot possibly be exposed (structural zero probability of being exposed)
always zero cells for these confounder values
example men will always be unexposed in terms of hormone replacement therapy
- for structural nonpositivity causal inference has to be restricted to subset of population for which positivity holds
- with random nonpositivity we can use parametric models to smooth over zero cells with parametric IPW model weights for zero cells estimated using parametric model for $Pr[A|L]$
- keep in mind that parametric IPW and standardization are not equivalent different models that are only equivalent if saturated and in parametric situations $IPW \neq \text{standardization}$ in high-dimensional data
IPW models estimate $\hat{f}[A|L]$ while standardization models estimate $f[L]$, $f[Y|A, L]$
- high-dimensional data occur in the cases of
 - many categorical variables
 - continuous variables
 - time-varying variables
 - combinations of the above

TABLE 2. Effect of HAART* versus no HAART on change in HIV-1* RNA viral load under a series of models using increasingly fine categorization of time-varying CD4 count and viral load in construction of inverse probability weights, Multicenter AIDS* Cohort Study and Women's Interagency HIV* Study, 1996–2005†

No. of categories‡	Estimated weights		Difference in viral load, \log_{10} copies/ml	
	Mean (SD*)	Minimum/maximum	Estimate	SE*, \$
1	1.00 (0)	1.00/1.00	-1.59	0.089
3	1.01 (0.96)	0.15/33.5	-1.73	0.103
5	1.00 (1.42)	0.06/59.1	-1.79	0.125
7	1.03 (1.61)	0.06/74.2	-1.74	0.392
9	536.7 (8,037.3)	0.05/1.6 \times 100,000	-¶	-

FIGURE 18.1. Increasing categories of a covariate increases non-positivity bias

- by increasing the number and levels of covariates for which weights can be calculated we better control for confounding bias at the expence of nonpositivity bias (even if nonpositivity is random)
 - might want to consider not controlling for a weak confounder that does not reduce confounding bias by much but increases the probability of non-positivity bias
- nonpositivity also increases with increasing time points in time-varying exposure situations but non-positivity bias decreases with the use of appropriate stabilized weights
- when parametrically estimating weights with statistical software useful hints of non-positivity are weights with $mean \neq 1$ and an inflated standard deviation
- in general the parameters of MSMs can be estimated even with non-positivity using parametric weights but the parameters are no longer causal

strength and weakness.

- strength
 - (1) Easy to understand and easy to fit with standard software that allows for weights
 - (2) Can also handle dynamic regimes by dynamic MSM
- weakness
 - (1) Pseudo bias and imprecision, the product of probability of treatment became very small in large number of time periods and may inflate the IPTW inappropriately (the drawback can be ameliorated by bounded doubly robust estimators, stabilized weights, truncating large influential weights)
 - (2) MSM cannot be used to estimate causal effects when treatment is confounded, but instrument is available
 - (3) MSMs do not allow modeling of interaction between treatment and time-varying covariates and cannot directly quantify effect modification

3. Structural Nested-Models

3.1. SNMs for Constant Exposures.

3.1.1. *Some things to remember...*

- **conditional exchangeability:**
 - conditional exchangeability $Y \perp\!\!\!\perp A | L$ can be parameterized in a regression model
 - $\text{logit} Pr[A = 1 | L, Y^{a=0}] = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L$
 - where $\alpha_1 = 0$
- **Sharp Causal Null:**

- no causal effect for any individual i
- $Y = Y_i^{a'} = Y_i^{a''}$ for all i and all a
- **rank-preservation:** *not necessary but useful for understanding*
 - each individual's counterfactual outcome, in the presence of treatment, is raised (if causative) or lowered (if preventative) by the amount equal to the causal effect
 - ignores heterogeneity and random variability in outcome and is biologically implausible
 - is **unnecessary** mathematically but it helps us understand g-estimation

3.1.2. Sharp G-null Test.

- (1) Parameterize the assumption of conditional exchangeability

$$\text{logit}Pr[A = 1|L, Y^{a=0}] = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L$$

- (2) Assume sharp causal null

$$Y_i^{a=0} = Y_i^{a=1} = Y \text{ for all } a$$

- (3) **g-null test:** if both sharp causal null and conditional exchangeability are true...

- $\alpha_1 = 0$ in the model $\text{logit}Pr[A = 1|L, Y^{a=0}] = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L$

- (4) in fact, $\alpha_1 \neq 0$ then either...

- we have conditional exchangeability but sharp causal null is not true
- we do not have conditional exchangeability but sharp causal null is true
- neither conditional exchangeability or sharp causal null are true

- (5) thus only under the assumption of conditional exchangeability can we use g-estimation to reject the sharp causal (i.e. when $\alpha_1 = 0$)

3.1.3. G-estimation of SNMM with Rank Preservation.

- only can be used for
 - continuous outcomes
 - time to event outcomes
 - dichotomous outcomes only if use log-linear model with rare-disease assumption
- in the presence of a point exposure g-estimation and marginal structural models use exactly the same models, albeit in a different way
- thus if paramaterized equivalently the approaches make exactly the same modeling assumptions

(1) Begining with a mean structural model, specify an individual level structural model

$$\begin{aligned}
 E[Y^a] &= \beta_0 + \beta_1 a & & \\
 E[Y^{a=1}] &= E[Y^{a=0}] + \beta_1 a & \beta_0 &= E[Y^{a=0}] \\
 E[Y^{a=0}] &= E[Y^{a=0}] - \psi^* a & \beta_1 &= \psi^* \\
 Y^{a=0} &= Y^{a=1} - \psi^* a & & \text{by rank preservation} \\
 Y^{a=0} &= Y - \psi^* A & & \text{by consistency} \\
 H(\psi^*) &= f(Y - \psi^* A) & & \text{for efficiency}
 \end{aligned}$$

note that up to this point we have not used conditional exchangeability
also, $f(x)$ can equal $\ln(x)$, x^2 , $\frac{1}{x}$, etc.

(2) Parameterize the assumption of conditional exchangeability

- use either of the following...

$$\begin{aligned}
 Pr[A = 1 | L, Y^{a=0}] &= \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L \\
 Pr[A = 1 | L, Y^{a=0}] &= \alpha_0 + \alpha_1 H(\psi^*) + \alpha_2 L
 \end{aligned}$$

(3) Find the true ψ^* using a "grid search" to test all candidate ψ

- choose evenly space ψ over a range, usually bounded by subject matter knowledge
- translate all candidate ψ into $H(\Psi)$ using individual level structural model

$$H(\psi) = Y - \psi A$$

- for each $H(\psi)$, fit the model

$$Pr[A = 1 | H(\psi), L] = \alpha_0 + \alpha_1 H(\psi) + \alpha_2 L$$

- Point estimate:** iteratively, narrow the window of candidate ψ until find ψ^* such that when $\psi = \psi^* \implies \alpha_1 = 0$
- 95% confidence interval:** using the p-value function, find the values of $H(\psi)$ that yield $p < 0.05$ for the wald test of $H_0: \alpha_1 = 0$

- translate $H(\hat{\psi}^*)$ and $H(\psi^*)_{L,U} \rightarrow \hat{\psi}^*$ and $\psi^*_{L,U}$ using structural model

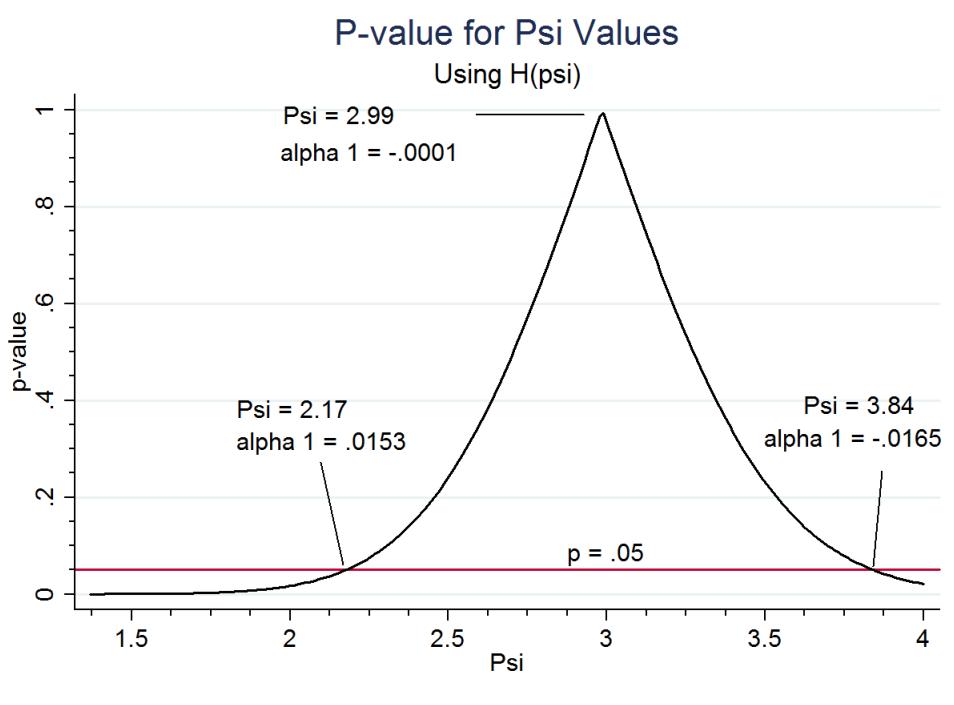


FIGURE 18.2. 95% ci and point estimate

3.1.4. Censoring.

- (1) use $IPW = SW^C$ to create a pseudo-population in which no one has been censored
- (2) g-estimation in pseudopopulation equates to weighting both the treatment model and structural model by SW^C

3.2. SNMMs for Time-Varying Exposures.

3.2.1. Structural Nested Mean Model.

- consider the general form of the additive structural nested mean model where m are treatment times $m = 0 \dots K$

$$\begin{aligned} E[Y_{g=\bar{\alpha}(m-1), \alpha(m), 0(m+1)} | \bar{L}_{\bar{\alpha}(m-1)}(m) = l(m), \bar{A}(m-1) = \bar{\alpha}(m-1)] \\ = E[Y_{g=\bar{\alpha}(m-1), 0(m)} | \bar{L}_{\bar{\alpha}(m-1)}(m) = l(m), \bar{A}(m-1) = \bar{\alpha}(m-1)] + \alpha(m)\gamma_m[\bar{\alpha}(m-1), \bar{l}(m), \beta^*] \end{aligned}$$

for nondynamic regimes $g = \bar{\alpha}(m-1), \alpha(m), 0(m+1)$ and $g = \bar{\alpha}(m-1), 0(m)$, where both have treatment $\bar{\alpha}(m-1)$ through $m-1$, treatment $\alpha(m)$ and 0 at m respectively, and both have treatment 0 from $m+1$ onward, where β^* is an unknown parameter vector,

$\gamma_m[\bar{\alpha}(m-1), \bar{l}(m), \beta^*]$, is a known function that $\gamma_m[\bar{\alpha}(m-1), \bar{l}(m), 0] = 0$ for $\beta^* = 0$ under the null hypothesis of no effect treatment

- for $K=1$ the model becomes $E[Y_{g=\bar{\alpha}(0), \alpha(1), 0(2)} | \bar{L}_{\bar{\alpha}(0)}(1) = l(1), \bar{A}(0) = \bar{\alpha}(0)] = E[Y_{g=\bar{\alpha}(0), 0(1)} | \bar{L}_{\bar{\alpha}(0)}(1) = l(1), \bar{A}(0) = \bar{\alpha}(0)] + \alpha(1)\gamma_1[\bar{\alpha}(0), \bar{l}(1), \beta^*]$ and for a saturated model

$\gamma_0[\bar{\alpha}(-1), \bar{l}(0), \beta^*] = \beta_0^*$, and $\gamma_1[\bar{\alpha}(0), \bar{l}(1), \beta^*] = \beta_{1,1}^* + \beta_{1,2}^* l(1) + \beta_{1,3}^* a(0) + \beta_{1,4}^* a(0)l(1)$ so that we get the SNM

$$Y_{g=(\alpha_0, \alpha_1)} = Y_{g=(\alpha_0, 0)} + \beta_{1,1} \alpha_1 + \beta_{1,2} \alpha_1 L_{1,g=a_0} + \beta_{1,3} \alpha_0 \alpha_1 + \beta_{1,4} \alpha_0 \alpha_1 L_{1,g=\alpha_0}$$

- g-stimation for estimating the unknown β^*

– define

$$Y_m(\beta) = Y - \sum_{j=m}^K A(j) \gamma_j [\bar{A}(j-1), \bar{L}(j), \beta]$$

- calculate $Y_m(\beta)$ for range of β containing the true β^* and a CI.
- fit a pooled logistic model $\text{logit} Pr[A(m) = 1 | \bar{L}(m), \bar{A}(m-1), Y_m(\beta)] = \alpha^T W(m) + \theta Y_m(\beta)$, where $W(m)$ is a vector for covariates and treatment history $[\bar{L}(m), \bar{A}(m-1)]$, α^T is a transposed row vector of unknown parameters
- the model gives an estimate of β^* by the value of β for which test score for $\theta = 0$ is 0, ($p\text{-value} = 1$)
- also 95% CI are given by the values of β for which the null hypothesis of $\theta = 0$ cannot be rejected ($p\text{-value} > 0.05$)
- this procedure is valid as if $\beta = \beta^*$ then $Y_m(\beta) = Y_{g=\bar{A}(m-1), 0m}$ and also $Y_{g=\bar{A}(m-1), 0m} \prod A(m)$ so for $\beta = \beta^*$ $\theta = 0$
- for a vector β then the pooled logistic model will contain a vector parameter θ corresponding to a vector of an adequate number of covariate functions needed to estimate the number of parameters in the vector β

Structural Accelerated Failure Time model.

- consider the model $T_\alpha = T_0 \exp[-\psi^* \alpha]$, where T_α and T_0 are survival times under treatment regime α and no treatment respectively
 - $\exp[-\psi^*]$ is the ratio $T_{\alpha=1}/T_{\alpha=0}$
 - it is the expansion or contraction of survival time attributable to treatment
 - $\psi^* > 0$ corresponds to shorter survival time for the treated compared to their survival time had they remained untreated and $\psi^* < 0$ corresponds to longer survival times when treated.
 - no causal effect when $\psi^* = 0$
- g-estimation of the unknown ψ^*
 - transform model to $T_0 = T_\alpha \exp[\psi^* \alpha]$
 - * model is also rank preserving
 - for time varying exposures where the exposure effect on survival time is averaged over duration through time t then the model becomes $T_0 = \int_0^{T_\alpha} \exp[\psi^* \alpha(t)] dt$
 - using observational data then we have $T_0 = \int_0^T \exp[\psi^* A(t)] dt$

- substitute T_0 with $H(\psi)$ and $H(\psi) = \int_0^T \exp[\psi^* A(t)] dt$ where $T_0 = H(\psi^*)$
- fit pooled logistic model $\text{logit} \Pr[A(m) = 1 | \bar{A}(m-1), \bar{L}(m), T > u(m), H(\psi)] = \theta_0(m) + \theta_1 A(m-1) + \theta_2 L(m) + \theta_3 H(\psi)$
- we know that $H(\psi^*) \prod A(m) | \bar{A}(m-1)$ for all m so for $\psi = \psi^*$ in the logistic model $\theta_3 = 0$
- for estimate of ψ^* and 95% CI using grid search score test for $\theta_3 = 0$

3.3. strength and weakness.

- strength
- Can estimate causal effects when treatment is confounded, and instrument is available
- Allow modeling of interaction between treatment and time-varying covariates
- weakness
- SNMs cannot be easily used to compare non-dynamic regimes when there is effect modification by time-dependent covariates
- Logistic SNMs, unlike log-linear model, cannot be fit by g-estimation
- MSM cannot be used to estimate causal effects when treatment is confounded, but instrument is available
- SNMs for failure time data should be based on accelerated failure time-like models, that cannot handle censoring properly

4. Propensity Scores

- **Definition of propensity score** : Probability of receiving exposure A conditional on the confounders L
- $PS = \Pr [A=1 | L]$
- Rosenbaum and Rubin (1983):
- If there is conditional exchangeability given the confounders, then there is conditional exchangeability given the PS
- It enables controlling for all measured confounders and provides more accurate estimates than can conventional methods.
 - Include all measured confounders in traditional regression model may decrease the statistical efficiency or have sparse data problems
 - Conventional variable-selection algorithms (eg. 5)
- **Assumptions of propensity score methods**

- Causal inference assumption: Conditional exchangeability, positivity, and consistency
- Model assumption: No misspecification of PS model, no misspecification of outcome model conditional on PS

- **Propensity score can be used to control confounding by**

- Stratification
- Regression
- Matching
- Weighting (marginal structural model)

- **Stratification**

- Create a categorical PS variable using the continuous PS (eg. decile) and stratify subjects by of the decile PS
- Combine the decile PS stratum-specific effect estimate by MH weight or inverse of variance weight
- Limitation: may have sparse data problems, have to make an assumption of homogenous effect across PS categories

- **Regression**

- Use continuous or categorical PS as a regressor in the regression model
- Limitation: continuous PS have to make a strong linearity assumption, categorical PS may have residual confounding

- **Matching**

- Matched each exposed subject to unexposed subject with same propensity score
- Can be 1:N matching
- Most clean type of covariate control (each pair is a strata)
- Limitation: less efficiency than regression because throw out data, cannot always find a matched unexposed subject

- **Weighting**

- Inverse probability of treatment weight is the inverse of PS
- Exposed group 1/PS ; unexposed group 1/(1-PS)
- Can also get SMR or SRR effect estimate
 - * SMR weight: Exposed 1, unexposed PS/(1-PS)
 - * SRR weight: Unexposed 1, exposed (1-PS)/PS

- **Stratification, regression, and matching are stratification based method and have the limitation of**

- Have to make assumption of no effect modification across PS-defined strata
- No average causal interpretation among the whole population, only within the levels of PS
- Result in bias if stratified on time-varying exposure

- **Other PS problems**

- A PS that perfectly predict exposure may suffer variance inflation by selecting nonconfounding but strong exposure predictors
- A PS includes variables predicting outcome but not exposure increases precision. (However, such model is no longer $\Pr[A \rightarrow L]$)
- A hybrid method is to combine exposure and outcome modeling score and used them for controlling confounding and increasing precision.
- Propensity score for continuous or multiple category outcomes is not well established
- The use of propensity score in case-control study is not well established

5. Instrumental Variables

6. G-Methods Analysis Example

- the following is an example of how data from a non-randomized dynamic treatment regime can be analyzed by the g-methods: g-formula algorithm, IPTW (MSM), and g-estimation (SNM)

TABLE 18.1. Example dataset for g-methods computation

A_0	L_1	A_1	N	Y
0	0	0	6,000	50
0	0	1	2,000	70
0	1	0	2,000	200
0	1	1	6,000	220
1	0	0	3,000	230
1	0	1	1,000	250
1	1	0	3,000	130
1	1	1	9,000	110

- the structure of the data is depicted in Figure 18.3, where treatment A_1 is dynamically determined by past treatment A_0 and covariate L_1

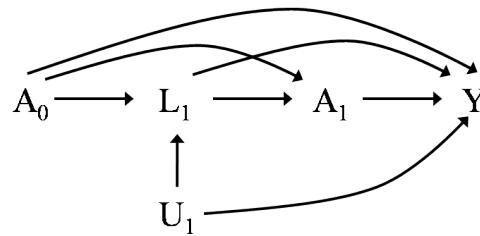


FIGURE 18.3. DAG of a dynamic regime where A_1 is determined by A_0 and L_1 .

- in order to appropriately analyze the data, g-methods are necessary to remove arrows into A_1 to make the study appear as a sequentially randomized study (Figure 18.4)

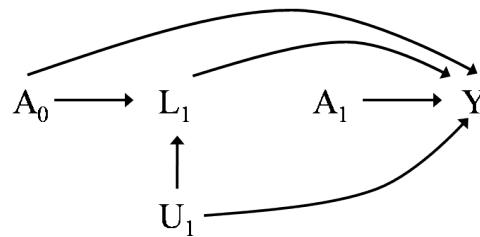


FIGURE 18.4. DAG of a dynamic regime where arrows into A_1 have been removed by g-methods

- the following is an example of how to make valid causal inferences from the data using the g-formula algorithm, IPTW, and g-estimation

6.1. G-Formula. Example of using of the g-formula algorithm

TABLE 18.2. Data for use with the g-formula

A_0	L_1	A_1	N	Y	$Pr[L_1 A_0]$
0	0	0	6,000	50	0.50
0	0	1	2,000	70	0.50
0	1	0	2,000	200	0.50
0	1	1	6,000	220	0.50
1	0	0	3,000	230	0.25
1	0	1	1,000	250	0.25
1	1	0	3,000	130	0.75
1	1	1	9,000	110	0.75

- in this example, L is a binary variable so the g-formula can be explicitly written as:

$$E[Y^g] = E[Y|A_0 = a_0, A_1 = a_1, L_1 = 0] \times Pr[L_1 = 0|A_0 = a_0] \\ + E[Y|A_0 = a_0, A_1 = a_1, L_1 = 1] \times Pr[L_1 = 1|A_0 = a_0]$$

- the causal outcomes of interest $Y^{g\{a_0, a_1\}}$ are estimated as follows:

- $Y^g\{0,0\} = 125$

$$Y^g\{0,0\} = 50 \times 0.50 + 200 \times 0.50 = 125$$
- $Y^g\{0,1\} = 145$

$$Y^g\{0,1\} = 70 \times 0.50 + 220 \times 0.50 = 145$$
- $Y^g\{1,0\} = 155$

$$Y^g\{1,0\} = 230 \times 0.25 + 130 \times 0.75 = 155$$
- $Y^g\{1,1\} = 145$

$$Y^g\{1,1\} = 250 \times 0.25 + 110 \times 0.75 = 145$$

- causal risk differences for the joint effects of treatments A_0 and A_1 are calculated as follows:

- $Y^g\{1,0\} - Y^g\{0,0\} = 155 - 125 = 30$
- $Y^g\{0,1\} - Y^g\{0,0\} = 145 - 125 = 20$
- $Y^g\{1,1\} - Y^g\{1,0\} = 145 - 155 = -10$
- $Y^g\{1,1\} - Y^g\{0,1\} = 145 - 145 = 0$
- $Y^g\{1,1\} - Y^g\{0,0\} = 145 - 125 = 20$

6.2. IPTW (MSM). Example of IPTW

TABLE 18.3. Data from a Non-random Dynamic Regime with *unstabilized* IPTW

A_0	L_1	A_1	N	Y	$f\{A_0\}$	$f\{A_1 A_0, L_1\}$	W	<i>Pseudo N</i>
0	0	0	6,000	50	0.50	0.75	$\frac{8}{3}$	16,000
0	0	1	2,000	70	0.50	0.25	8	16,000
0	1	0	2,000	200	0.50	0.25	8	16,000
0	1	1	6,000	220	0.50	0.75	$\frac{8}{3}$	16,000
1	0	0	3,000	230	0.50	0.75	$\frac{8}{3}$	8,000
1	0	1	1,000	250	0.50	0.25	$\frac{8}{3}$	8,000
1	1	0	3,000	130	0.50	0.25	8	24,000
1	1	1	9,000	110	0.50	0.75	$\frac{8}{3}$	24,000

- the first step is to calculate *unstabilized* IPTW weights to create a pseudo population without confounding (Table 18.3)
- the goal is to make $Y^g \perp\!\!\!\perp A_1|A_0, L_1$, thus removing any arrows into A_1
- data can then be collapsed over strata of L_1 to create Table 18.4
- the Marginal Structural Model (MSM) we want to fit is:

$$E[Y^g=\{a_0, a_1\}] = \eta_0 + \eta_1 a_0 + \eta_2 a_1 + \eta_3 a_0 a_1$$

TABLE 18.4. Data collapsed over strata of L

A_0	A_1	Pseudo N	$E[Y A_0, A_1]$	Parameter
0	0	32,000	125	γ_0
0	1	32,000	145	$\gamma_0 + \gamma_2$
1	0	32,000	155	$\gamma_0 + \gamma_1$
1	1	32,000	145	$\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3$

- the associational model we fit is:

$$E[Y|a_0, a_1] = \gamma_0 + \gamma_1 a_0 + \gamma_2 a_1 + \gamma_3 a_0 a_1$$

- however, if we fit an appropriately IPTW weighted regression model in the actual population, the result is an associational model fitted to a pseudo population where there is no confounding
- thus association in the pseudo population *is* causation in the actual population, $\eta_i = \gamma_i$
- the model parameters are estimated as:

$$\begin{aligned} - \gamma_0 &= \eta_0 = 125 \\ \gamma_0 &= 125 \\ - \gamma_1 &= \eta_1 = 20 \\ \gamma_0 + \gamma_2 &= 125 + \gamma_2 = 145 \\ - \gamma_2 &= \eta_2 = 30 \\ \gamma_0 + \gamma_1 &= 125 + \gamma_1 = 155 \\ - \gamma_3 &= \eta_3 = -30 \\ \gamma_0 + \gamma_1 + \gamma_2 + \gamma_3 &= 125 + 20 + 30 + \gamma_3 = 145 \end{aligned}$$

- the causal outcomes of interest $Y^{g=\{a_0, a_1\}}$ are estimated as follows:

$$\begin{aligned} - E[Y^{g=\{0,0\}}] &= 125 \\ E[Y^{g=\{0,0\}}] &= \eta_0 = 125 \\ - E[Y^{g=\{0,1\}}] &= 145 \\ E[Y^{g=\{0,1\}}] &= \eta_0 + \eta_2 = 145 \\ - E[Y^{g=\{1,0\}}] &= 155 \\ E[Y^{g=\{1,0\}}] &= \eta_0 + \eta_1 = 155 \\ - E[Y^{g=\{1,1\}}] &= 145 \\ E[Y^{g=\{1,1\}}] &= \eta_0 + \eta_1 + \eta_2 + \eta_3 = 145 \end{aligned}$$

- MSMs cannot assess effect modification by non-baseline covariates

6.3. G-Estimation (SNM). Example of g-estimation

- Structural Nested Models (SNMs) can be used for any static or dynamic regime g
- SNMs need one equation for each treatment time

TABLE 18.5. Data from a Non-random Dynamic Regime for G-estimation

A_0	L_1	A_1	N	Y	$Y^{g=\{A_0,0\}}$	$Y^{g=\{0,0\}}$
0	0	0	6,000	50	50	50
0	0	1	2,000	70	$70 - \beta_{1,1}^*$	$70 - \beta_{1,1}^*$
0	1	0	2,000	200	200	200
0	1	1	6,000	220	$220 - \beta_{1,1}^* - \beta_{1,2}^*$	$220 - \beta_{1,1}^* - \beta_{1,2}^*$
1	0	0	3,000	230	230	$230 - \beta_0^*$
1	0	1	1,000	250	$250 - \beta_{1,1}^* - \beta_{1,3}^*$	$250 - \beta_0^* - \beta_{1,1}^* - \beta_{1,3}^*$
1	1	0	3,000	130	130	$130 - \beta_0^*$
1	1	1	9,000	110	$110 - \beta_{1,1}^* - \beta_{1,2}^* - \beta_{1,3}^* - \beta_{1,4}^*$	$110 - \beta_0^* - \beta_{1,1}^* - \beta_{1,2}^* - \beta_{1,3}^* - \beta_{1,4}^*$

– time 0: $Y^{g=\{a_0,0\}} = Y^{g=\{0,0\}} + \beta_0^* a_0$

* where:

$\beta_0^* = Y^{g=\{1,0\}} - Y^{g=\{0,0\}}$ is the direct effect of treatment a_0 on outcome Y when a_1 is withheld

– time 1: $Y^{g=\{a_0,a_1\}} = Y^{g=\{a_0,0\}} + \beta_{1,1}^* a_1 + \beta_{1,2}^* a_1 L_1^{g=\{a_0\}} + \beta_{1,3}^* a_0 a_1 + \beta_{1,4}^* a_1 a_0 L_1^{g=\{a_0\}}$

* where β_1^* parameterizes the effect of a_1 on Y within the 4 possible levels of past treatment and covariate history:

$\beta_{1,1}^*$ is the effect of a_1 on Y when a_0 is withheld among the subset of individuals with $L_1^{g=\{0\}} = 0$

$\beta_{1,1}^* + \beta_{1,2}^*$ is the effect of a_1 on Y when a_0 is withheld among the subset of individuals with $L_1^{g=\{0\}} = 1$

$\beta_{1,1}^* + \beta_{1,3}^*$ is the effect of a_1 on Y when a_0 is taken among the subset of individuals with $L_1^{g=\{0\}} = 0$

$\beta_{1,1}^* + \beta_{1,3}^* + \beta_{1,4}^*$ is the effect of a_1 on Y when a_0 is withheld among the subset of individuals with $L_1^{g=\{0\}} = 1$

- SNMs are locally rank preserving models
- the assumption of conditional exchangeability ($Y^{g=\{0,0\}} \perp\!\!\!\perp A_0$ and $Y^{g=\{a_0,0\}} \perp\!\!\!\perp A_1 | A_0, L_1$) is needed to estimate the SNM parameters
 - under $Y^{g=\{a_0,0\}} \perp\!\!\!\perp A_1 | A_0, L_1$, the counterfactual outcome $Y^{g=\{A_0,0\}}$ is equal for $A_1 = 0$ and $A_1 = 1$ in the same strata of A_0 and L_1 (denoted by sections in between dashed lines in Table 18.5)
 - thus, the vector of parameters β_1^* can be estimated using the SNM at time 1:
 - * $\beta_{1,1}^* = 20$:
 - $50 = 70 - \beta_{1,1}^*$ (Table 18.5, rows 1 and 2)

- * $\beta_{1,2}^* = 0$:
 $200 = 220 - \beta_{1,1}^* - \beta_{1,2}^*$ (Table 18.5, rows 3 and 4)
- * $\beta_{1,3}^* = 0$:
 $230 = 250 - \beta_{1,1}^* - \beta_{1,3}^*$ (Table 18.5, rows 5 and 6)
- * $\beta_{1,4}^* = -40$:
 $130 = 110 - \beta_{1,1}^* - \beta_{1,2}^* - \beta_{1,3}^* - \beta_{1,4}^*$ (Table 18.5, rows 5 and 6)

– under $Y^{g=\{0,0\}} \amalg A_0$, the mean of $Y^{g=\{0,0\}}$ is the same for $A_0 = 0$ as for $A_0 = 1$

- * $\beta_{1,1}^*, \beta_{1,2}^*, \beta_{1,3}^*$, and $\beta_{1,4}^*$ can be plugged in to calculate $Y^{g=\{0,0\}}$

– thus, β_0^* can be estimated using the SNM at time 0:

- * $\beta_0^* = 30$:
 $50 \times \frac{6,000+2,000}{16,000} + 200 \times \frac{2,000+6,000}{16,000} = 125$ (Table 18.5, rows 1-4)
 $230 \times \frac{3,000+1,000}{16,000} + 130 \times \frac{3,000+9,000}{16,000} = 155 - \beta_0^*$ (Table 18.5, rows 5-8)
 $125 = 155 - \beta_0^*$

– finally, $E[Y^g]$ can be estimated using the locally rank-preserving SNMs

- * $E[Y^{g=\{0,0\}}] = 125$
estimated from SNM at time 0 (above)
- * $E[Y^{g=\{1,0\}}] = 155$
 $E[Y^{g=\{1,0\}}] = E[Y^{g=\{0,0\}}] + \beta_0^* = 125 + 30 = 155$, estimated from SNM at time 0
- * $E[Y^{g=\{0,1\}}] = 145$
 $E[Y^{g=\{0,1\}}] = E[Y^{g=\{0,0\}}] + \beta_{1,1}^* a_1 + \beta_{1,2}^* a_1 L_1^{g=\{0\}} = 125 + 20 + 0 = 145$, estimated from SNM at time 1
- * $E[Y^{g=\{1,1\}}] = 145$
 $E[Y^{g=\{1,1\}}] = E[Y^{g=\{1,0\}}] + \beta_{1,1}^* a_1 + \beta_{1,2}^* a_1 L_1^{g=\{1\}} + \beta_{1,3}^* a_0 a_1 + \beta_{1,4}^* a_1 a_0 L_1^{g=\{1\}}$
 $E[Y^{g=\{1,1\}}] = 155 + 20 + 0 + (0 - 40) \times \frac{3}{4} = 145$, estimated from SNM at time 1
- * where $L_1^{g=\{A_0=1\}} = \frac{3}{4}$, because association is causation for the effect of A_0 on L_1 , hence the $Pr[L_1 = 1 | A_0 = 1] = \frac{3,000+9,000}{3,000+1,000+3,000+9,000} = \frac{12,000}{16,000} = \frac{3}{4}$

- these estimates of $E[Y^g]$ are the same as those from the g-formula and the IPTW MSM
- SNM allow for the additional investigation of effect modification of treatment regime g by time-varying covariates