Prospective/Retrospective Studies

In many studies, the population can be categorized according to two binary variables:

$$D = \begin{cases} 1: & \textit{presence of a disease } (e.g., \textit{ lung} - \textit{cancer}); \\ 0: & \textit{absence of a disease}. \end{cases}$$

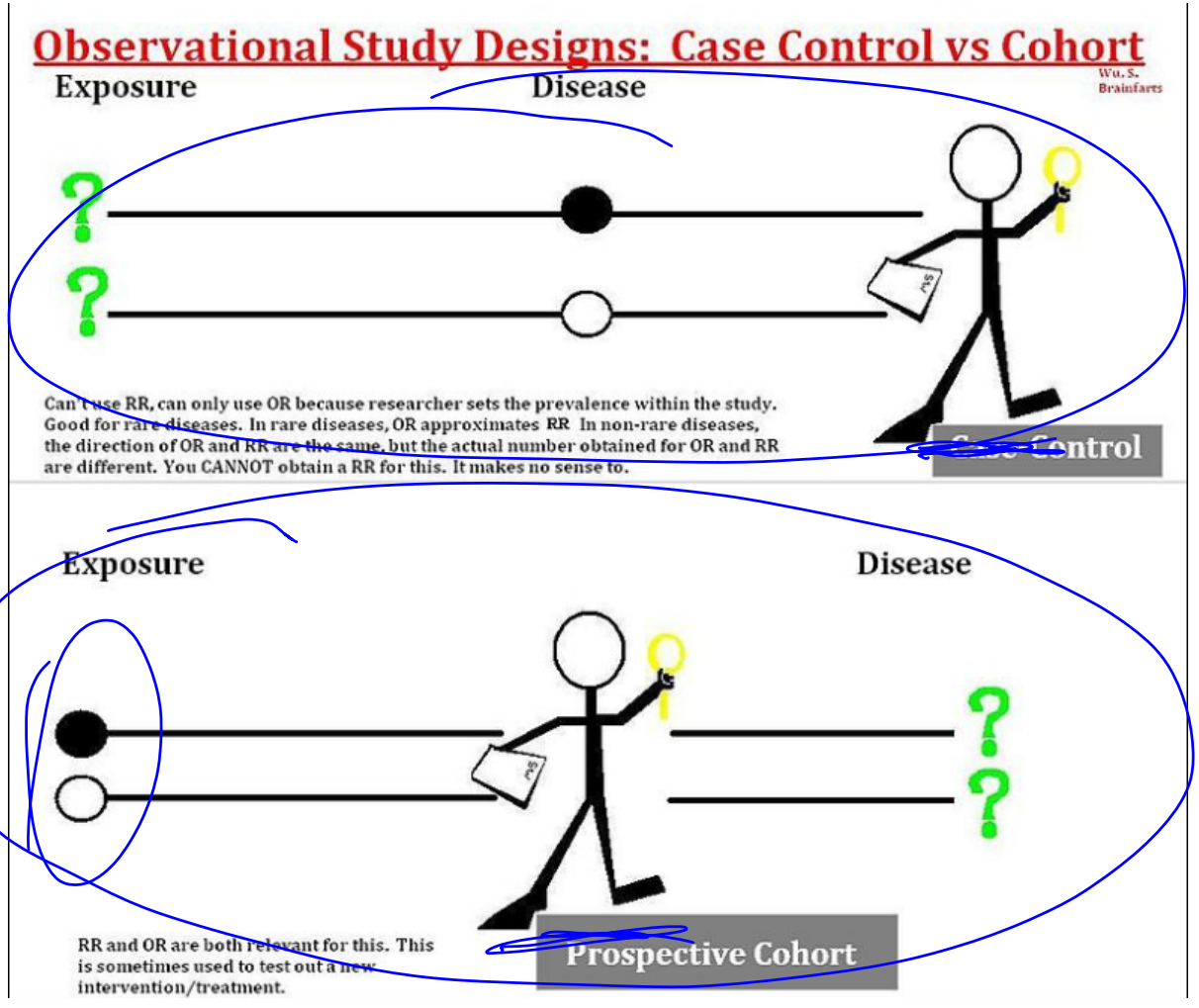
$$X = \begin{cases} 1: & \textit{exposure to a certain toxin } (e.g., \textit{ smoking}); \\ 0: & \textit{non} - \textit{exposure to a certain toxin}. \end{cases}$$

		D	
		1	0
X	1	Y_{11}	Y_{10}
	0	Y_{01}	Y_{00}

There exist two sampling schemes to obtain data for the study, Prospective and Retrospective.

Observational Study Designs: Case Control vs Cohort

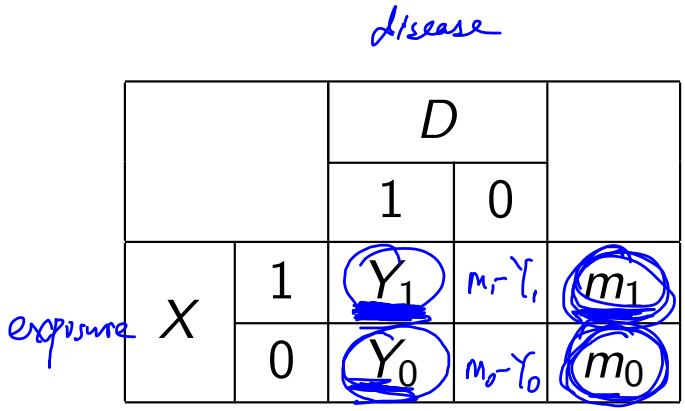
Wu. S.
Brainfarts



retro

prop

Prospective sampling:



- an exposed group is selected together with a non-exposure group.
- both groups are monitored over a prolonged period to compare the incidence of diseases in the two groups
- row totals are fixed, column totals are random.
- Example: randomized clinical trials

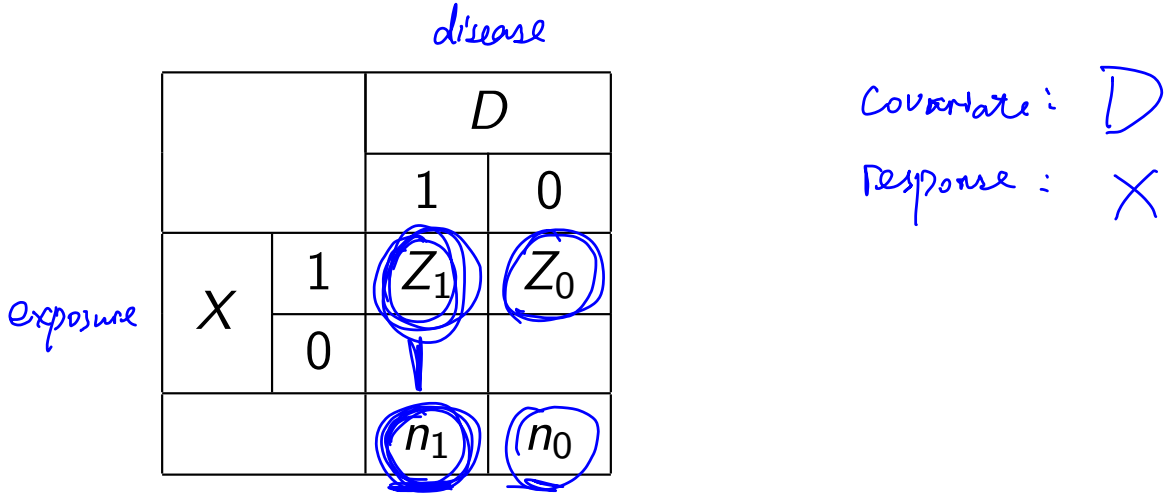
- Model: disease status is response; exposure (and other covariates) are predictors
  - $Y_1 \sim Bin(m_1, \pi_1)$
  - $Y_0 \sim Bin(m_0, \pi_0)$
  - Logistic model:

$$logit(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 X_i, \ (X_i = 0, 1)$$

$$\beta_0 = log \frac{\pi_0}{1-\pi_0}, \ \beta_1 = \log \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

- We are interested in relative risk (RR) π_1/π_0 and odds ratio (OR) $\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$

Retrospective sampling:



- Suitable when disease incidence rate is low and/or disease takes time to occur.
- Two groups of subjects are selected, one with disease, the other without disease (e.g., case-control study).
- Exposure information is obtained retrospectively.
- The difference in exposure history between cases and controls reveals disease-exposure relation.

Two ways to model retrospective data:

- Retrospective model: exposure history as response ((Z_1, n_1) and (Z_0, n_0)) , disease status as predictor

$$Z_1 \sim Bin(n_1, \rho_1), \; Z_0 \sim Bin(n_0, \rho_0)$$

$$\log \frac{\rho_i}{1-\rho_i} = \alpha_0 + \alpha_1 D_i, \; (D_i = 0, 1)$$

$$\alpha_0 = \log \frac{\rho_0}{1-\rho_0}, \; \alpha_1 = \log \frac{\rho_1/(1-\rho_1)}{\rho_0/(1-\rho_0)}$$

- Can estimate $\mathbb{P}(Exposure|Disease)$
- Cannot estimate $\mathbb{P}(Disease|Exposure)$
- α_0 is the log odds of exposure given no disease (in the control group)
- α_1 is the log odds ratio of exposure between case and control groups

- **Invariance**: The odds ratio of exposure in case vs control is equal to the odds ratio of disease in the exposed group vs the non-exposed group (Bayes rule)

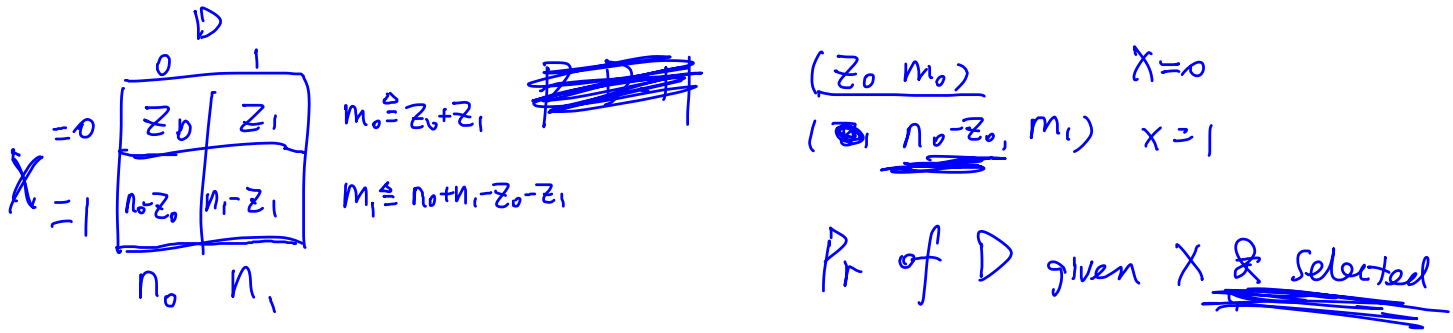
$$\begin{aligned} OR &= \frac{odds(exposure|case)}{odds(exposure|control)} \\ &= \frac{odds(disease|exposure)}{odds(disease|non\text{-}exposure)} \end{aligned}$$

- But retrospective data CANNOT be used to evaluate the relative risk of disease in the exposed group vs the non-exposed group.

$$RR = \frac{\mathbb{P}(disease|exposure)}{\mathbb{P}(disease|non\text{-}exposure)}$$

- Exception: for rare disease ($\mathbb{P}(disease) \approx 0$), $RR \approx OR$.

$RR = \frac{\pi_1}{\pi_0}$ $\quad OR = \frac{\pi_1(1-\pi_1)}{\pi_0/(1-\pi_0)}$ $\leftarrow$ $\pi_1 \approx 0$, $\pi_0 \approx 0$



- Prospective model: despite the fact that data are collected retrospectively, we treat it as a prospective study.
- Treat disease status (D) as response, and exposure (X) as predictor (as in the prospective study), and fit a logistic model.
- Based on a variant of the invariance theory, the coefficient for X is the same as in the retrospective model, i.e., the desired log odds ratio!

Multiple Covariates

- We are interested in how the change of $\mathbf{X}$ affects $\mathbb{P}(D)$
- Assume

$$\mathbb{P}(D = 1|\mathbf{X}) = \frac{\exp(\alpha + \mathbf{X}\beta)}{1 + \exp(\alpha + \mathbf{X}\beta)}$$

- If prospective data are collected, we can fit a logistic regression model and estimate α and β.

- If we only have retrospective data, we can still use a prospective model.
- Essential assumption: the selection criteria of the case-control study is independent of covariates
- Using the prospective model on the retrospective data, we are really modeling $\mathbb{P}(D = 1|\boldsymbol{X}, S = 1)$
- From Bayes theorem, we can derive

$$\mathbb{P}(D = 1|\mathbf{X}, S = 1) = \frac{\exp(\alpha^* + \boldsymbol{X}\boldsymbol{\beta})}{1 + \exp(\alpha^* + \boldsymbol{X}\boldsymbol{\beta})}$$

- Thus we can estimate $\boldsymbol{\beta}$, but not α (α^* is a nuisance parameter)
- $\exp(\boldsymbol{\beta})$ provides the odds ratios of disease corresponding to unit change in different covariates.

Recap

- Data may be collected prospectively or retrospectively.
- We can[1] always fit a prospective model to retrospective data by treating disease status as response, and having multiple covariates.
- OR is invariant.
- In general, RR can only be evaluated for prospective data.
- For rare diseases, RR$\approx$OR.

[1] Sampling procedure must be independent of covariates.