Week 3: Stratified Case-Control Data

Video 1: Introduction and Notation

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Objectives

- Recognize notation used for the analysis of stratified case-control data
- Construct hypothesis tests of association and interpret p-values for stratified case-control data
- Calculate and interpret point estimates and confidence intervals for the summary (common) odds ratio in stratified case-control data
- Construct hypothesis tests to evaluate homogeneity of the odds ratio in stratified case-control data and interpret the results

Notation for Unstratified Case-Control Data

- Recall our notation for unstratified case-control data:

Case-Control Data

	E	$\bar{E}$	
Cases	a	b	M_1
Controls	c	d	M_0
	N_1	N_0	T

Notation for Stratified Case-Control Data

- Now stratify the data by confounding variables, so that each stratum consists of subjects who have, on average, the same risk of disease, with the possible exception of the exposure effect.
- We have i = 1,...I of these strata, which are formed by unique combinations of levels of the confounding variables for which there are data.

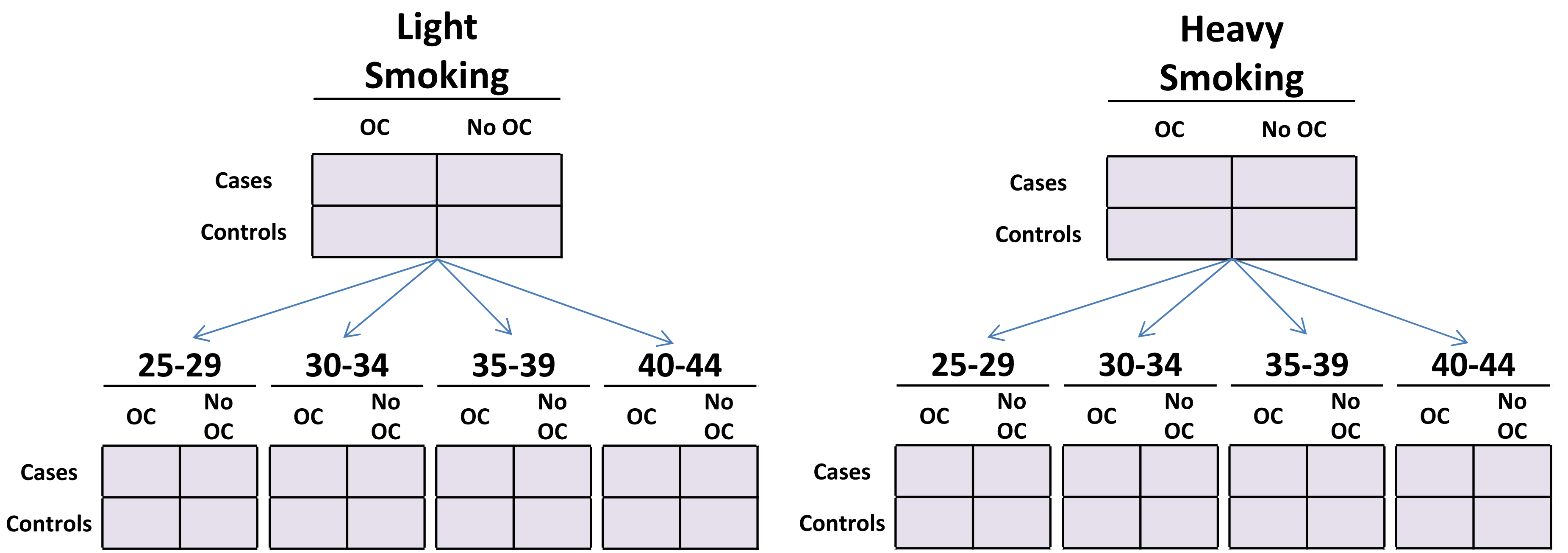
Stratified Case-Control Data

	E	$\bar{E}$	
Cases	a_i	b_i	M_{1i}
Controls	c_i	d_i	M_{0i}
	N_{1i}	N_{0i}	T_i

Stratification on Several Confounders

- If cigarette smoking is a risk factor for MI and is also associated with OC use in these data, then we may wish to control for smoking status (light/heavy) as well as age group in 4 categories.

- In this case, there will be **8** strata:



OC Use and MI Study Revisited

Crude

	OC	No OC	
Cases	23	112	135
Controls	130	1306	1436
	153	1418	1571

Stratified by Age

Age	25-29			30-34			35-39			40-44		
	OC	No OC		OC	No OC		OC	No OC		OC	No OC	
Case	4	2	6	9	12	21	4	33	37	6	65	71
Control	62	224	286	33	390	423	26	330	356	9	362	371
Total	66	226	292	42	402	444	30	363	393	15	427	442
ÔR	7.2			8.9			1.5			3.7		

Week 3: Stratified Case-Control Data

Video 2: Hypothesis Tests

EPI202 – Epidemiologic Methods II

Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Hypothesis Test for Unstratified Data

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

We saw previously that in case-control studies with no confounding,

- X = number of exposed cases,

- $E(X|H_0)$ =
  - number of exposed cases expected under H_0
  - T * Pr[D] * Pr[E]
  - T * (M_1/T) * (N_1/T)
  - $M_1(N_1/T)$

- $Var(X|H_0) = \dfrac{M_1 M_0 N_1 N_0}{T^2 (T-1)}$

Hypothesis Test for Stratified Data (1)

- We first stratify the data on all confounding variables to form I strata and then calculate the **Mantel-Haenszel χ^2 test statistic**:

$$Z^2 = \frac{\left[\sum_{i=1}^{I} X_i - \sum_{i=1}^{I} E_i\left(X_i | H_0\right)\right]^2}{\sum_{i=1}^{I} Var_i\left(X_i | H_0\right)} \sim \chi_1^2$$

Hypothesis Test for Stratified Data (2)

- The summations are over each of the I = 1,…,I strata
- Each component of the test statistic has the same stratum-specific form as in the crude analysis:
  - □ X_i = number of exposed cases in stratum i = a_i
  - □ $E_i(X|H_0)$ = number of exposed cases expected under H_0 in stratum I
    $= M_{1i}(N_{1i}/T_i)$
  - □ $Var_i(X|H_0) = \dfrac{M_{1i}\,M_{0i}\,N_{1i}\,N_{0i}}{T_i^2(T_i-1)}$ in stratum I
- and Z^2 is distributed χ^2 with one degree of freedom

Hypothesis Test for Stratified Data
Null and Alternative Hypothesis

- H_0: There is no association between OC use and MI incidence
  - $OR_{MH} = 1 \leftrightarrow IRR_{MH} = 1 \leftrightarrow ln(OR_{MH}) = 0$

- H_A: There is an association between OC use and MI incidence
  - $OR_{MH} \neq 1 \leftrightarrow IRR_{MH} \neq 1 \leftrightarrow ln(OR_{MH}) \neq 0$

OC Use and MI
Mantel-Haenszel Test Statistic (1)

$$Z^2 = \frac{\left[\sum_{i=1}^{I} X_i - \sum_{i=1}^{I} E_i(X_i|H_0)\right]^2}{\sum_{i=1}^{I} \mathrm{Var}_i(X_i|H_0)} \sim \chi_1^2$$

- $\Sigma X_i = \Sigma a_i = 4 + 9 + 4 + 6 = 23$
- $\sum E_i(X \mid H_0) = \sum \frac{M_{1i}\, N_{1i}}{T_i} = \frac{6*66}{292} + \frac{21*42}{444} + \frac{37*30}{393} + \frac{71*15}{442} = 8.577$
- $\sum \mathrm{Var}_i(X \mid H_0) = \sum \frac{M_{1i}\, M_{0i}\, N_{1i}\, N_{0i}}{T_i^2\,(T_i - 1)}$

$$= \frac{6*286*66*226}{292^2*291} + \frac{21*423*42*402}{444^2*443} + \frac{37*356*30*363}{393^2*392} + \frac{71*371*15*427}{442^2*441}$$

$$= 7.08$$

OC Use and MI
Mantel-Haenszel Test Statistic (2)

$$Z^2 = \frac{(23 - 8.577)^2}{7.08} = 29.40$$

- $\Pr[\chi^2 > 29.40] = 0.000001$

- After conditioning on age, the data are not very compatible with the state of nature described by the null. If we were interested in a hypothesis testing framework with a pre-specified 2-sided alpha of 0.05, as in the crude analysis, we would reject the null hypothesis and conclude that there is a statistically significant association between OC use and MI incidence in these data.

- Compare the crude *p*-value (0.003) to the age-adjusted *p*-value (0.000001). After conditioning on age, there is much stronger evidence against the null hypothesis.

Week 3: Stratified Case-Control Data

Video 3: Point Estimation

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Summary Odds Ratio
Point Estimate

- To calculate the estimate of the summary odds ratio, we use a weighted sum of the stratum-specific estimates:

$$\hat{O}R = \frac{\sum_{i=1}^{I} w'_i \, \hat{O}R_i}{\sum w'_i} = \sum_{i=1}^{I} w_i \, \hat{O}R_i$$

$$Where \;\; w_i = \frac{w'_i}{\sum w'_i} \quad and \quad \sum w_i = 1$$

- The exact value of the summary odds ratio depends upon the chosen weights.
- In the absence of effect measure modification, any weights will yield an unbiased estimate of the common odds ratio
- Weights are chosen to optimize the statistical precision of the estimate

Mantel-Haenszel Weights

$$w_i' = \frac{b_i c_i}{T_i}$$

- Mantel-Haenszel weights provide a valid, efficient estimate of the common odds ratio under the assumption of no effect modification of the odds ratio by the stratification factor(s)
- Unlike inverse variance weights $\left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)$, which become undefined if there is a single cell with a zero cell count, the Mantel-Haenszel weights perform well whether or not there are sparse data within-strata

Mantel-Haenszel Estimator of the Odds Ratio

- Computational formula for the **Mantel-Haenszel summary odds ratio**:

$$\hat{OR}_{MH} = \frac{\sum_{i=1}^{I} w_i' \hat{OR}_i}{\sum_{i=1}^{I} w_i'} = \frac{\sum_{i=1}^{I} \frac{b_i c_i}{T_i} \frac{a_i d_i}{b_i c_i}}{\sum_{i=1}^{I} \frac{b_i c_i}{T_i}} = \frac{\sum_{i=1}^{I} \frac{a_i d_i}{T_i}}{\sum_{i=1}^{I} \frac{b_i c_i}{T_i}}$$

- A given stratum may contribute to the numerator, the denominator, both or neither

Week 3: Stratified Case-Control Data

Video 4: Confidence Intervals

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

95% CI for $ln(OR_{MH})$

$$X \pm Z_{1-\alpha/2}\sqrt{\hat{V}ar\,(X)}$$

- Where:
  - $X = \ln(\hat{O}R_{MH})$
  - $Z_{1-\alpha/2}$ = 100(1-α/2)th percentile of the standard normal distribution
  - $\hat{V}ar\,(X)$ = estimated variance of $\ln(\hat{O}R_{MH})$

Robins, Greenland, & Breslow (RGB)
Estimated Variance Formula

- RGB formula for the estimated variance of $\ln(\hat{\text{OR}}_{\text{MH}})$:

$$\frac{1}{2}\left[\frac{\sum\left(\frac{a_i d_i}{T_i}\right)\left(\frac{a_i + d_i}{T_i}\right)}{\left(\sum\frac{a_i d_i}{T_i}\right)^2} + \frac{\sum\left(\frac{a_i d_i}{T_i}\right)\left(\frac{c_i + b_i}{T_i}\right) + \sum\left(\frac{b_i c_i}{T_i}\right)\left(\frac{a_i + d_i}{T_i}\right)}{\sum\left(\frac{a_i d_i}{T_i}\right)\sum\left(\frac{b_i c_i}{T_i}\right)} + \frac{\sum\left(\frac{b_i c_i}{T_i}\right)\left(\frac{c_i + b_i}{T_i}\right)}{\left(\sum\frac{b_i c_i}{T_i}\right)^2}\right]$$

- Accurate when the point estimate is near the null and when it is far from the null

- Works for both sparse and large strata

- Computationally cumbersome -- use the computer, especially with a large number of strata

RGB Variance of $ln(\mathrm{OR}_{\mathrm{MH}})$

- Using Stata or the EPI202 calculator, the RGB variance for the log of the Mantel-Haenszel odds ratio is 0.0750. The 95% confidence interval for the Mantel-Haenszel odds ratio is thus

$$e^{\ln[3.99] \pm 1.96\sqrt{0.075}} = e^{(0.847,\, 1.921)} = (\, 2.33,\, 6.83\,)$$

- Assuming no residual confounding by age, no confounding by other variables, no selection bias and no information bias, these data are consistent with odds ratios ranging from 2.3 to 6.8 with 95% confidence.

Week 3: Stratified Case-Control Data

Video 5: : Effect Modification and Test of homogeneity

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

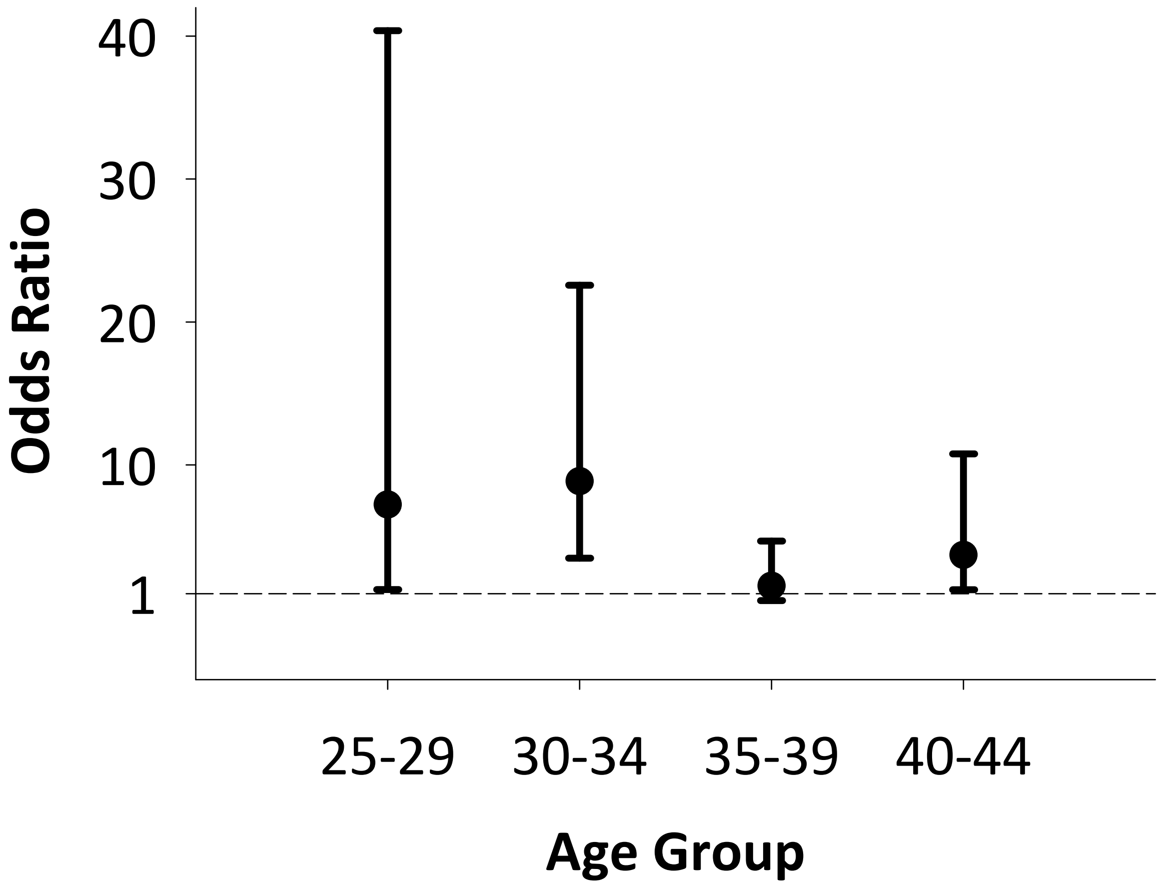
OC Use and MI

Evaluating Effect Measure Modification

- We wish to evaluate these data for evidence of effect measure modification by age

Age	25-29			30-34			35-39			40-44		
	OC	No OC		OC	No OC		OC	No OC		OC	No OC	
Case	4	2	6	9	12	21	4	33	37	6	65	71
Control	62	224	286	33	390	423	26	330	356	9	362	371
Total	66	226	292	42	402	444	30	363	393	15	427	442
$\hat{OR}$	7.2			8.9			1.5			3.7		

OC Use and MI
Stratum-specific Estimates: “Eyeball Test”



Evaluating Effect Measure Modification
Hypothesis Test

- Specify the null hypothesis and its alternative.
- Form a test statistic that has a known probability distribution under the assumptions implied by the null hypothesis.
- Find the p-value which corresponds to the observed value of the test statistic under the null.
- Interpret the results.

Test of Homogeneity (OR)
Null and Alternative Hypotheses

- H_0: The odds ratio is the same across all I levels of the stratification variable(s)
  - ◻ ↔ There is no effect measure modification by the stratification variable(s)
  - ◻ ↔ The odds ratio is *homogeneous* across the strata
  - ◻ ↔ $OR_1 = OR_2 = \ldots = OR_I$ ↔ $IRR_1 = IRR_2 = \ldots = IRR_I$
  - ◻ ↔ $OR_i = OR_j$ for all i,j

- H_A: The odds ratio is not the same across all I levels of the stratification variable(s)
  - ◻ ↔ There is effect measure modification by one (or more) of the stratification variable(s)
  - ◻ ↔ The odds ratio is *heterogeneous* across the strata
  - ◻ ↔ At least one of the ORs does not equal at least one of the others
  - ◻ ↔ At least one of the IRRs does not equal at least one of the others
  - ◻ ↔ $OR_i \neq OR_j$ for some i,j

Test for Homogeneity
Odds Ratio

- The test for homogeneity of the odds ratio has the following form

$$H = \sum_{i=1}^{I} \frac{[\ln(\hat{OR}_i) - \ln \hat{OR}_{MH}]^2}{\hat{Var}_i[\ln(\hat{OR}_i)]} \sim \chi^2_{I-1}$$

- Where:
  - $\hat{OR}_i$ is the stratum specific odds ratio
  - $\hat{OR}_{MH}$ is the Mantel-Haenszel summary odds ratio
  - $\hat{Var}[ln(\hat{OR}_i)]$ is the (large-sample) variance of the stratum specific ln(OR);

- H is distributed χ^2 with I-1 degrees of freedom

ad hoc Adjustment Methods For Zero Cells

- If the number of exposed or unexposed cases in a stratum is 0, the weight for that stratum is undefined.

- Two *ad hoc* methods are often used:
  - Add 0.5 to every cell in that stratum
  - Collapse adjacent sparse strata to eliminate 0 cells

- Neither of these options is entirely satisfactory.

OC Use and MI

Age	25-29			30-34			35-39			40-44		
	OC	No OC		OC	No OC		OC	No OC		OC	No OC	
Case	4	2	6	9	12	21	4	33	37	6	65	71
Control	62	224	286	33	390	423	26	330	356	9	362	371
Total	66	226	292	42	402	444	30	363	393	15	427	442
$\hat{O}R$	7.2			8.9			1.5			3.7		

- We know from previous calculations that $\hat{O}R_{MH} = 3.99$
- In the first stratum (age 25-29), the variance of the estimate of the stratum specific log odds ratio is

  1/4 + 1/2 + 1/62 + 1/224 = **0.771**

OC Use and MI

Test of Homogeneity

	25-29	30-34	35-39	40-44
$\hat{\mathrm{Var}}\left[ln\left(\hat{\mathrm{OR}}_i\right)\right]$	0.771	0.227	0.322	0.295
$\dfrac{[ln(\hat{\mathrm{OR}}_i) - ln\,\hat{\mathrm{OR}}_{MH}]^2}{\hat{\mathrm{Var}}_i\,[ln(\hat{\mathrm{OR}}_i)]}$	0.4578	2.803	2.822	0.0175

- Thus,
  - H = [0.4578 + 2.803 + 2.822 + 0.0175] = 6.10
  - $\Pr[\chi^2_{I-1=3} > 6.10] = 0.12$

- There is not sufficient evidence in these data to reject the null hypothesis of homogeneity of the odds ratio across strata.

Test of Homogeneity
Limitations

- Share the same limitations as other hypothesis tests
  - Fail to summarize data with respect to their consistency with alternative hypotheses
  - Provide no indication of the power to detect alternative hypotheses of interest
- Considerably more data are needed to detect and characterize effect measure modification than to detect statistical evidence of an association under the assumption of no effect measure modification.
- Most studies, unless explicitly designed to detect and characterize effect measure modification, will not have sufficient data to do so.
- Failure to reject the null hypothesis of no effect measure modification may often be explained by a lack of power.

Test of Homogeneity

- When the null hypothesis is rejected, the pooled estimate is not meaningful

- It is often best to report stratum specific estimates

- If a summary measure is needed, it is best to use weights that do not reflect arbitrary features of the study design.
  - Standardization
  - Inverse Probability Weighting

Summary

- Point estimates of the summary odds ratio are constructed using weighted averages of stratum-specific estimates

- In the absence of effect measure modification, Mantel-Haenszel weights provide valid and efficient estimates of the summary (common) odds ratio after adjusting for confounding by one or more stratification factors

- The test of homogeneity can be used to evaluate whether there is statistical evidence for effect measure modification on the multiplicative scale