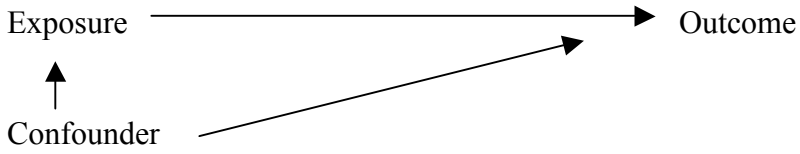
Matching

Matching in a cohort study usually involves selecting non-exposed subjects to have the same the distribution of the matching factor that exists in the exposed group. For example, matching by age may involve selecting a non-exposed subject with the same age as each exposed subject in the study. If the same number of matched non-exposed subjects is enrolled for each enrolled exposed subject (**fixed matching ratio**), then the distribution of the matching factor among the non-exposed will be identical to that of the exposed subjects.

Matching in a case control study involves selecting a control with the same value for the matching factor as each case. If the same number of matched controls is enrolled for each case, then the distribution of the matching factor among the controls will be identical to that of the cases.

The main motivation for matching is typically to avoid confounding in the resulting data that are collected. The simple DAG in the following figure shows the anticipated associations in a study in the presence of confounding and displays of the required relationship of the confounder with the exposure and the outcome. These are the targets of matching in order to avoid confounding in the resulting data.

Directed Acyclic Graph (DAG) reflecting the relationship between a confounder and the exposure and the outcome in the absence of matching in a study.



Matching in Cohort Studies with a fixed matching ratio guarantees that the matching factor will have identical distribution among the exposed and among the non-exposed subjects in the data. This entails eliminating the arrow from the confounder (matching factor) and the exposure in this figure, thereby eliminating the backdoor pathway for confounding. It follows that matching in a Cohort Study, with a fixed matching ratio, avoids confounding by the matching factor in the resulting data.

Matching with a fixed matching ratio in case control studies forces cases and controls to have the same distribution of the matching factor. For example, matching on age would mean that the percentage of elderly people would be the same for cases and controls. However, the causal diagram shows that there are two factors that influence the age distribution of the cases: the direct influence of age (depicted by the arrow connect age with the outcome) and the direct influence of the exposure (which is related to age). Therefore, matching in a case control study does not directly relate to only the pathway from the confounding factor to the outcome. As a result, matching in a Case Control

Study does not block the backdoor pathway and avoid confounding by the matching factor.

Matching by a confounder in a Case Control Study builds similar distributions of the matching factor among the cases and among the matched controls. It will also build similar distributions of other factors that are correlated with the matching factor, including the exposure of interest. Therefore, matching in a Case Control Study tends bias the value for the crude Odds Ratio by towards its null value (1.0). This is demonstrated in the following example.

Example

Large Population:

A. Sex distribution among exposed and among non-exposed subjects in the source population

	Exposed	Non-exposed
Males	8,000 (80%)	2,000 (20%)
Females	2,000 (20%)	8,000 (80%)
Total	10,000	10,000

B. Exposure and sex-specific risks of outcome

	Exposed	Non-exposed
Males	.06	.02
Females	.03	.01

C. Expected number of outcomes cases (# subject x risk)

	Exposed	Non-exposed
Males	480	40
Females	60	80
Total	540	120

D. Expected sex-specific data

Males	Outcome +	-	Total
Exposed	480	7520	8000
Non-exposed	40	1960	2000

RR = 3.0

Females	Outcome +	-	Total
Exposed	60	1960	2000
Non-exposed	80	7920	8000

RR = 3.0

E. Expected crude data

	Outcome		
	+	-	Total
Exposed	540	9460	10000
Non-exposed	120	9880	10000

RR = 4.5

In these data, sex is a confounder since it is associated with the exposure in the source population (Panel A) and is an independent determinant of the outcome (Panel B). Moreover, the common stratum-specific value for the Risk Ratio (RR = 3.0, Panel D) differs from the crude value of the Risk Ratio (RR = 4.5, Panel E).

Matched Cohort Study

A Cohort Study that matches on sex with a fixed matching ratio will enroll a sample of exposed subjects and a sample of non-exposed subjects whose sex distributions are identical. The following table presents the expected results from a matched cohort study that enrolled 1000 exposed subjects selected at random from all exposed subjects in the original large population described in Panel A of the original data. These data also contain 1000 non-exposed subjects who are matched by sex to the exposed subjects

A. Sex distribution among exposed and among matched non-exposed subjects

	Exposed	Non-exposed
Males	800 (80%)	800 (80%)
Females	200 (20%)	200 (20%)
Total	1000	1000

B. Exposure and sex-specific risks of outcome

	Exposed	Non-exposed
Males	.06	.02
Females	.03	.01

C. Expected number of outcomes cases

	Exposed	Non-exposed
Males	48	16
Females	6	2
Total	54	18

D. Expected sex-specific data

	Males				Females		
	Outcome				Outcome		
	+	-	Total		+	-	Total
Exposed	48	752	800	Exposed	6	184	200
Non-exposed	16	784	800	Non-exposed	2	198	200
	RR = 3.0				RR = 3.0		

E. Expected crude data

	Outcome		
	+	-	Total
Exposed	54	946	1000
Non-exposed	18	982	1000
	RR = 3.0		

Panel A of this table shows that the impact of matching in a cohort study is to create a study population to one that cannot support confounding due to the lack of association between the matching factor and the exposure. As a result, the crude Risk Ratio (3.0 from Panel E) is equal to the adjusted values (from Panel D).

Matched Case Control Study

Matching in a Case Control Study impacts the selection of controls. Although a fixed matching ratio will guarantee the lack of a crude association between the matching factor and the outcome, unless the exposure has no effect on the outcome (Odds Ratio =1.0), matching on a confounder will not result in the lack of a conditional association between the matching factor and the outcome . This is demonstrated by the example in the following table. The data are from matched (by sex) case control study based on all 660 outcome cases that developed from the original large population and 660 sex-matched controls.

A. Sex distribution among cases and matched controls

	Cases	Controls
Males	520 (79%)	520 (79%)
Females	140 (21%)	140 (21%)
Total	660	660

B. Prevalence of exposure among cases and matched controls

Cases (Panel C of table for original large population):

	Exposed	Non-exposed	Total
Males	480 (92%)	40 (8%)	520
Females	60 (43%)	80 (57%)	140
Total	540 (82%)	120 (18%)	660

Control (Based on Panel A for original large population):

	Exposed	Non-exposed	Total
Males	416 (80%)	104 (20%)	520
Females	28 (20%)	12 (80%)	140
Total	444 (67%)	216 (33%)	660

C. Expected sex-specific data

Males

	Exposure +	-	Total
Case	480	40	520
Control	416	104	520

OR = 3.0

Females

	Exposure +	-	Total
Case	60	80	140
Control	28	112	140

OR = 3.0

D. Expected crude data

	Exposure +	-	Total
Case	540	120	660
Control	444	216	660

OR = 2.2

E. Expected exposure-specific data:

Exposed

	Male Sex +	-	Total
Case	480	60	540
Control	416	28	444

OR = .54

Non-exposed

	Male Sex +	-	Total
Case	40	80	120
Control	104	112	216

OR = .54

This example demonstrates that matching by sex in this case control study did not avoid confounding. The association between sex and exposure among the controls in

Panel B suggests that the matching factor (sex) satisfies the first criterion for confounding. More importantly, Panel E demonstrates that despite the equal sex distribution among cases and controls caused by matching, conditional on exposure status sex remains an independent determinant of the outcome, although its measure of effect, OR = .54, is different from that suggested in the original large population, RR = 2.0). Therefore by the results of Panels B and E, sex satisfies the two criteria for confounding. In addition, the stratum-specific measure of the effect of the exposure in Panel C (OR=3.0) is different from the crude measure in Panel D (OR = 2.2), suggesting that sex is a confounder by the change-in-estimate criterion for confounding.

Because of its independent relationship to the outcome, stratifying by a confounder in an unmatched Case Control Study would show varying ratios of cases to controls over the strata. If the factor is a strong determinant of the outcome, then some of these strata may contain many cases but few controls (or vice versa), suggesting an inefficient basis to try to measure the effect of the exposure. This is especially true for small studies or when the confounder is nominal is scale with many distinct categories. Since matching tends to make this ratio constant over strata, one would expect that matching and stratification on a confounder might lead to a more efficient analysis than not matching but stratifying by a confounder.

Matched Analysis

The basic principle that underlies a matched analysis is that the association between the exposure and the outcome is first performed within each matched group and then pooled over groups to obtain a summary average. As an example of a matched analysis, the following table presents the basic layout of the results from a Case Control study with a one-one matching design.

		Exposure Status of Control	
		+	-
Exposure Status of Case	+	A	B
	-	C	D

Matched pairs (A and D) with identical exposure status for both the case and the matched control are referred to as concordant pairs. These pairs provide no information on the relationship between exposure and outcome. Matched pairs (B and C) show different exposure status for the case and the matched control, and thus provide information about the relationship between the exposure and the outcome. These matched pairs are referred to as discordant pairs and are the basis for estimates of the effect of the exposure on the outcome and for tests of significance concerning this effect.

The relative sizes of the two types of discordant pairs provide the basis for the measure of the effect of the exposure. An estimate for the odds ratio (OR) from 1-1 match data is

$$OR = B/C$$

This estimate is identical to the Mantel-Haenszel estimator from a stratified analysis with each stratum corresponding to a separate matched group containing one case and 1 control. This demonstrated by the following example presenting an analysis of Case Control Study (Hosmer and Lemeshow. *Applied Logistic Regression Second Edition*. John Wiley & Sons; New York: 2000) that examined risk factors for low birth weight babies. The data for this analysis pertain to 56 matched pairs. Each matched pair contained one case (low birth weight infant) and one control (normal birth weight infant), with the controls matched by the age of mothers of the cases.

		Maternal Smoking of Control	
		+	-
Maternal Smoking of Case	+	8	22
	-	8	18

$$OR = 22/8 = 2.75$$

The following table displays the stratified analysis from the 56 strata.

Strata		Frequency	AD/T	BC/T
D+	D-			
E+ 1	1	8	0	0
E- 0	0			
E+ 1	0	22	1/2	0
E- 0	1			
E+ 0	1	8	0	1/2
E- 1	0			
E+ 0	0	18	0	0
E- 1	1			

$$OR_{MH} = 22(1/2) / 8(1/2) = 22/8 = 2.75$$

Effect Modification

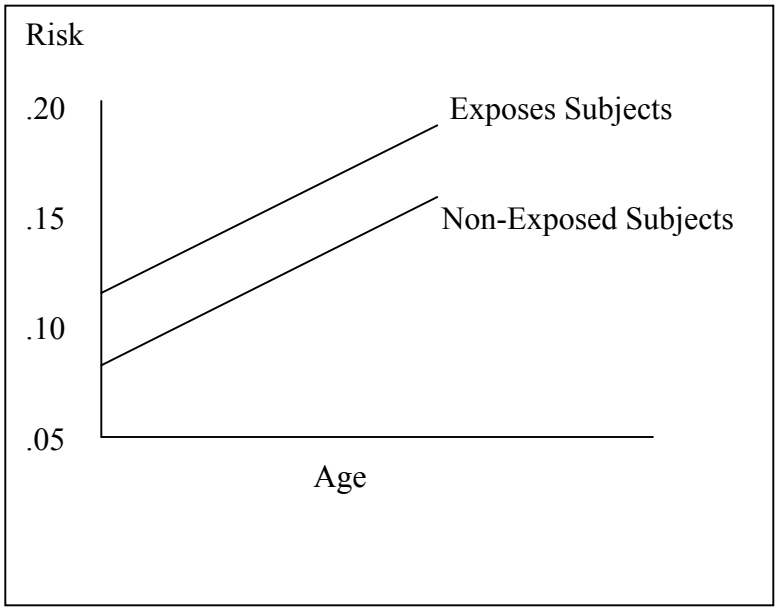
Effect Modification refers to the situation where the effect of the exposure is modified or changed according to the value or level of another factor. Effect Modification is detected by examining sub-group analyses, examining the association between and exposure and an outcome with sub-groups defines by categories of the candidate effect modifier.

For example, the following data are from a Retrospective Cohort Study examining the relationship between perioperative beta-blocker use and in-hospital mortality among 663,535 patients undergoing non-cardiac surgery at 329 Hospitals (Lindenauer et al: N Engl J Med 2005;353:349-61) The data are stratified by the Revised Cardiac Risk Index Score (RCRI), a measure of a patient's risk of developing a cardiac complication during surgery.

RCRI Score	Odds Ratio	Confidence Interval
0	1.36	(1.27,1.45)
1	1.09	(1.01,1.19)
2	0.88	(0.80,0.98)
3	0.71	(0.63,0.80)
≥ 4	0.58	(0.50,0.67)

These data shows the effect of beta-blocker use on mortality depends on the level of the RCRI. Among patients with low risk of a cardiac complication (RCRI scores of 0 or 1), beta-blocker use tends to increase the risk of in-hospital death (RR = 1.36 for patients with RCRI=0 and RR= 1.09 for patients with RCRI=1). On the other hand, for patients with higher values of RCRI, beta-blocker use tends to decrease the risk of in-hospital death (RR= 0.88 for patients with RCRI=2, RR= 0.71 for patients with RCRI=3 and RR= 0.58 for patients with RCRI $\geq$ 3).

The detection of effect modification depends on the choice of the measure of association. For example, the following figure suggests that age does not modify the value for the Risk Difference (the distance between points on the two lines is the same for any value for age), but does modify the value for the Risk Ratio (the ratio of the points on the two lines becomes less with increasing age.)



The detection of Effect Modification is challenging. Before concluding that results like those presented in the previous numerical example reflect the presence of effect modification, an investigator must rule out the roles of

1. Bias
2. Confounding
3. Chance

In addition, as with subgroup analyses in experimental studies, the investigator should have a clinical argument to justify examining for effect modification by a risk factor

Tests for detecting effect modification are based on comparing stratum-specific estimates of effect, which are each based on only a portion of the original data. Therefore, they may have limited power. A comparison of confidence intervals around measures of association for various sub-group analyses also provides evidence about chance being the explanation for the data suggesting the existence of effect modification. For example, each of the RCRI-specific Odds Ratio in the beta- blocker example is not contained the confidence intervals for any of the other sub-group analysis.

Presenting the Effect of an Exposure in the Presence of Effect Modification

When effect modification exists, the single average measures of association cannot be expected to estimate the differing values for the measure of association that exist in the various strata. In this situation, the presentation of stratum-specific results or a method of standardization is superior to the method of weighted averaging (Mantel-Haenszel Estimate) as described in the previous lecture notes.

When effect modification exists, probably the best manner to present the effect of the exposure is by displaying the different measures of association for different subgroups or strata of the effect modifier. For example, if age is an effect modifier, then one might display the separate effects of the exposure for young, middle-aged, and old subjects.

An alternative to presenting stratum-specific estimates of effect is to present a summary (average) measure of association that is linked to a specified population with a known distribution of the effect modifier. This involves the method of **direct standardization** and compares two standardized measures of disease frequency: one under the assumption that everyone in the standard population has the stratum-specific risks of the exposed subjects, and the other under the assumption that everyone in the standard population has the stratum-specific risks of the non-exposed subjects. Therefore, standardization estimates the counterfactual outcomes that were described in the previous series of lecture and estimates the average causal effect of the exposure in the standard population.

The formula for the standardized risk ratio is

$$SRR = [\Sigma\{n_i R_{1i}\}] / [\Sigma\{n_i R_{0i}\}]$$

$$= [\Sigma\{w_i RR_i\}] / [\Sigma\{w_i\}]$$

where

n_i = number of subjects in the standard population in the i^{th} stratum

R_{1i} = risk (rate) of the outcome among the exposed subjects in the i^{th} stratum

R_{0i} = risk (rate) of the outcome among the non-exposed subjects in the i^{th} stratum

$w_i = n_i R_{0i}$

If the exposed subjects are chosen as the standard population (i.e. $n_i = N_{1i}$), then this formula simplifies to

$$SRR = a / [\Sigma\{N_{1i} R_{0i}\}]$$

where a = total number of exposed subjects who develop the outcome.

In this special case where the exposed subjects are taken as the standard population, the standardized risk ratio becomes the ratio of the observed number of exposed cases to the expected number of exposed cases. This ratio is usually referred to as the SMR

(standardized mortality ratio or standardized morbidity ratio) and is a component of the method of indirect standardization.

A standardized risk ratio reflects the overall, unconfounded effect of the exposure in a specific population (the chosen standard) with a specific distribution for the effect modifier. Choosing a different standard with a different distribution for the effect modifier should result in a different value for the standardized risk ratio, reflecting the overall effect of the exposure in the new standard population.

Example

The following data can be used estimate the age-standardized risks of death among smokers and non-smokers in the FHS teaching data set. The standard population is the total number of study subjects (4434) in the data and the age strata are the same that were used in the previous lecture notes to describe stratification.

Stratified Analysis

Age ≤ 40	Died	Survived	Total
Smokers	67	385	452
Non-Smokers	25	277	302
Total	92	662	754

40 < Age ≤ 50	Died	Survived	Total
Smokers	266	689	955
Non-Smokers	110	574	684
Total	376	1263	1639

50 < Age ≤ 60	Died	Survived	Total
Smokers	286	281	567
Non-Smokers	312	500	812
Total	598	781	1379

Age > 60	Died	Survived	Total
Smokers	169	38	207
Non-Smokers	315	140	455
Total	484	178	662

The following table displays the calculation of the expected number of deaths under the two scenarios that all members of the population smoked and that none of the members of the population smoked. The expected number of deaths for each age group (columns 4 and 6) are estimated by multiplying the number of subjects in the standard population who are in that age group by the correspond risk of dying for that age group.

Standard Population(Age Group)	Number	Risk if all were Exposed (Smoker)	Expected. # Cases	Risk if all Non-Exposed (Non-Smoker)	Expected # of Cases
≤ 40	754	67/452 = .1482	754(.1482) = 111.74	25/302 = .0828	754(.0828) = 62.43
(40, 50]	1639	266/955 = .2785	1639(.2785) = 456.46	110/684 = .1608	1639(.1608) =263.55
(50, 60]	1379	286/567 = .5044	1379(.5044) = 695.57	312/812 = .3842	1379(.3842) = 529.82
> 60	662	169/207 = .8164	662(.8164) = 540.52	315/455= .6923	662(.6923) = 458.30
Total	4434		1804.29		1314.10

The Standardized Risks of Death (estimated counterfactual outcomes) and Standardized Risk Ratio from this table are

Smokers: 1804.29/4434 = 0.4069

Non-Smokers: 1314.10/4434 = 0.2964

Standardized Risk Ratio: .4069/.2964 = 1.37

Inverse Probability Weighting

Standardization involves inverse probably weighting. For example, the previous table shows that the expected number of deaths in the youngest age group if all 754 subjects smoked is111.74. The following calculation shows two formulas for the calculation of this value.

111.74 = (smokers risk) x (size of standard population)
= (67/452) x (754)
= (67/452) x (452) x (754/452)
= (smokers risk) x [(# smokers) x (weight)]

The bottom equation implies that the expected number of death can be obtained by multiplying the number of young smokers by a weight and then multiplying this value by the risk of death among the young smokers. The weight (754/452) equals

754/452 = 1/P(Smoking| Age $\leq$ 40)
= 1/P(Exposure| Age $\leq$ 40)
= 1/P(Exposure|Confounder)

The P(Exposure | Confounder(s)) is called the **Propensity Score.** This topic will be discussed in the next series of lecture notes. The previous calculation demonstrated that the estimation of the number of deaths if everyone in the standard population has the exposure involves weighting the exposed subjects by the inverse of their propensity scores. Similarly, the estimation of the number of deaths if no one in the standard population has the exposure involves weighting the non-exposed subjects by the inverse of (1- propensity score).