Concept of Confounding

Suppose an investigator performed a Cohort Study to investigate the association between smoking and the risk of developing Coronary Heart Disease and found that the incidence of CHD among smokers was twice that of non-smokers (Risk Ratio = RR = 2.0). As noted previously before we can conclude that this measure of association reflects the causal effect of smoking, we need to rule out the alternative explanations of

1. Bias
2. Confounding
3. Chance

Suppose that the investigator is confident that there is little potential for bias in this study and that the large sample size limits the role of chance as possible reason for this association. However, a closer look at the data reveals that the smokers are much older than the non-smokers. Does the measure of association (RR = 2.0) reflect the effect of

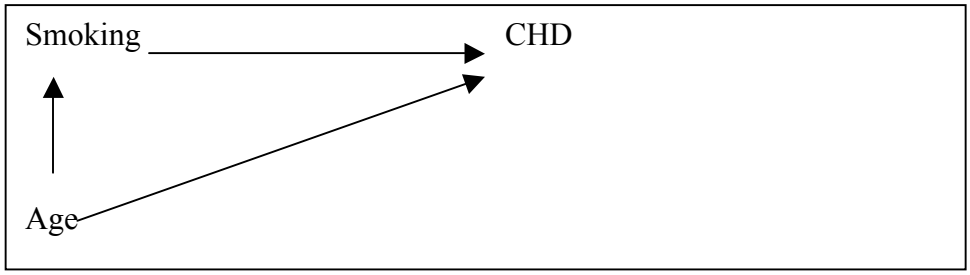
1. Smoking ?
2. Older Age ?
3. Both ?

Recall that the **causal effect of a risk factor for a person** reflects the change in the outcome that is observed when that person is exposed to that risk factor and when that person is not exposed to that risk factor, **under identical situations**. These two outcomes are referred to as **counterfactual outcomes** (Hernan and Robins. J *Epidemiol Community Health* 2006;60:578-586), but only one of them can be observed in reality. For example, the causal effect of lifetime smoking starting at age 20 on the development of CHD for a person is the difference in the CHD counterfactual outcome when that person spends a lifetime as a smoker, compared to the CHD counterfactual outcomes when that person spends a lifetime as a non-smoker. If a person is a lifetime smoker then the observed CHD outcome (factual outcome) matches the counterfactual outcome following a lifetime of smoking. However, we are unable to observe the other CHD counterfactual outcome for that person had that person never smoked. Hence, it is not possible to measure the causal effect of a risk factor for a specific person.

The **average causal effect of a risk factor for a population** is the difference in average counterfactual outcomes when all members of the population receive the risk factor and when none of the members of the population receive it. If the population is comprised of some members who receive the risk factor and others who do not, then it may be possible to estimate each counterfactual outcome and the average causal effect of the risk factor. For example, in the absence of confounding and bias, the incidence of coronary heart disease from a group of non-smokers (**a factual outcome**) may be a valid estimate for the counterfactual outcome for a group of smokers had they not smoked. This assumes an ability to "exchange" factual outcomes of the non-smokers with the counterfactual outcomes for the smokers (Greenland and Robins, *Intl J Epidemiol*

1986;15:412-419). This implies that the observed difference (ratio) in incidence of coronary heart disease in the two comparison groups (**a measure of association)**, is an estimate of the causal effect of smoking.

Causal diagrams in epidemiology are called **Directed Acyclic Graphs (DAGS).** They are directed in that they contains arrows that reflect causal assumptions between potential risk factors and outcomes and they acyclic in that the direction of the arrows do not contains loops from outcomes back to their causes. For example, the following DAG describes the Cohort Study mentioned above and shows two pathways (explanations) that could account for the crude association between smoking and CHD in a data set.



The top arrow reflects of potential direct effect that smoking may have on the risk of developing CHD. However, the other two arrows suggest that there is also a second, **backdoor pathway** to explain this association. Smokers might be older than non-smokers and older age also influences the risk of developing CHD. The existence of a backdoor pathway involving a third factor (e.g. age) provides a **DAG-based definition of confounding.** A risk factor, whose control blocks a backdoor pathway, is a **confounder (confounding factor)**. The challenge of an epidemiologist is to avoid confounding in a study though aspects of the study design (e.g. randomization, restriction, or matching) or to adjust (control) for confounding in the analysis (through stratification, standardization, regression modeling, and the various methods involving propensity scores).

Adjusted measures of association control for factors (confounders) that account for a difference in the factual outcomes in the non-exposed group and the counterfactual outcomes in the exposed group. The lack of **exchangeability** is the **counterfactual-based definition of confounding**. For example, if the smokers are older than the non-smokers in our study, then the non-observed counterfactual outcomes of the smokers (e.g. their incidence of disease under the condition that they did not smoke) would be greater than the factual outcome of the non-smokers (e.g. their observed incidence of disease). The reason for this difference in incidence is the influence of the older age distribution of the smokers. The crude measure of association would mix (confound) the two causal influences (the effect of smoking and the effect of older age).

The existence of a backdoor pathway in a causal diagram, leading to a lack of exchangeability leads to a commonly sited definition of confounding.

1. The confounder must be associated with the exposure (in this example smokers have an older age distribution than non-smokers).

2. The confounder must be associated with the disease, independent of the exposure (in this example, older age increases the risk of disease among the non-smokers).

3. The confounder must not be part of the causal pathway connecting the exposure to the disease.

Suppose that the following data reflect the associations found in the Cohort Study mentioned above.

Crude Analysis

	CHD		
	+	-	Total
Smokers	240	760	1000
Non-Smokers	120	880	1000

Stratified analysis (by Age)

	Young			Old		
	CHD			CHD		
	+	-	Total	+	-	Total
Smokers	60	340	400	180	420	600
Non-Smokers	80	720	800	40	160	200

Do these data reflect confounding by age? The first criterion for confounding states that the confounder must be associated with the exposure. This holds in these data as the prevalence of old age, P(old age), among smokers and non-smokers differ.

P(Old|Smoker) = 600/1000 = 60%
P(Old|Non-Smoker) = 200/1000 = 20%

The second criterion for confounding refers to the relationship between age and CHD, independent of smoking. This can be examined by examining the relationship between age and CHD among the non-smokers. The data support such a relationship as

P(CHD|Young Non-Smoker) = 80/800 = 10%
P(CHD|Old Non-Smoker) = 40/200 = 20%

The third criterion for a confounder states that age should not be in the causal pathway that link smoking with CHD. This criterion can not be examined by the data but can be logically ruled out as smoking does not cause old age.

When confounding exists, the causal graph implies that the value for the crude measure of the association reflects both the effect of the exposure and of the confounder (direct pathway and backdoor pathway). On the other hand, adjusting for a confounder blocks the backdoor pathway and the value for the adjusted measure of association reflects only the direct effect of the exposure. Therefore, when confounding exists, one would expect to observe different values for the crude and adjusted measures of association. This results in commonly used working definition of confounding:

> A confounder is a factor that when adjusted in the analysis results in a value for the adjusted measure of association that is meaningfully different from the value for the crude measure of association.

Although easy to implement and often used in practice, this **"change-in-estimate" definition** of confounding is a necessary but not a sufficient property of confounding and examples have shown it may lead to incorrect conclusions. Nevertheless it is often taken as a method for detecting confounding. This is demonstrated by the following analyses performed on the above data:

Crude Analysis:

	CHD		
	+	-	Total
Smokers	240	760	1000
Non-Smokers	120	880	1000

$$RR_{Crude} = (240/1000)/(120/1000)$$
$$= .24/.12$$
$$= 2.0$$

Stratified analysis (by Age)

	Young			Old		
	CHD			CHD		
	+	-	Total	+	-	Total
Smk	60	340	400	180	420	600
Non-Smk	80	720	800	40	160	200

$$RR_{Young} = (60/400)/(80/800)$$
$$= .15/.10$$
$$= 1.5$$

$$RR_{Old} = (180/600)/(40/200)$$
$$= .30/.20$$
$$= 1.5$$

$$RR_{adjusted} = RR_{Young} = RR_{Old} = 1.5$$

The change-in-estimate method is practical for detecting confounding and displays the "result" of confounding, while the conceptual definition of confounding describes the "mechanism" for the change in estimates.

Stratification

The implication of confounding is that the crude measure of association reflects a mixture of the effect of the exposure and the effect of the confounding factor(s). When confounding exists in a data set, analytical **methods of adjustment** must be used to separate the effect of the exposure from the effect(s) of the confounding factor(s). There are two general approaches for adjusting for confounding factors in the analysis:

1. **Stratification**,
2. **Regression Modeling**.

Regression modeling is the more common method for controlling confounding and stratification can be considered as a special case of modeling. However, because of its intuitive appeal, controlling confounding by stratification will be discussed initially.

Stratification involves dividing the data set into disjoint subgroups (strata) based on categories/values of the confounder(s). There are two methods for adjustment based on stratification:

1. **Pooling (weighted averaging)**
2. **Standardization**.

Stratification with pooling involves the following steps:

1. Create subgroups (strata) defined by categories or sub-ranges of the confounding factor, which are free of residual confounding by that factor,

2. Estimate the value for the measure of association within each stratum, and

3. **When appropriate**, average (pool) these estimates over strata to determine the adjusted measure of association.

The justification for this method is reflected in its first step. If all subjects within a stratum possess (essentially) a common value for a risk factor, then that factor cannot satisfy either of the first two criteria for confounding defined above within that stratum. For a non-continuous confounder, strata defined by distinct categories of the confounder automatically satisfy this situation. For example, when stratifying by sex, the exposed subjects and the non-exposed subjects will have the same sex distribution within the male stratum (all will be males). However, stratification by a continuous confounder requires the specifications of sub-ranges of the confounder to define the strata.

Depending on how sub-ranges of a continuous are defined, there may still be residual confounding by the stratifying factor within a stratum. Suppose that age is a confounder in a study. On one hand, narrowly defined sub-ranges (for example, one-year age intervals) are more homogenous and are less prone to contain residual confounding, but this approach may result in a large number of strata, with little information (individuals) contained within each strata. On the other hand, broadly defined sub-ranges (for example, decades of age) result in fewer strata containing more information, but also have the potential for within-stratum residual confounding by the stratifying factor. In the extreme, the broadest sub-range will result in a single stratum containing the entire data (crude analysis), with no control of confounding.

Given the creation of strata that are free of residual confounding by the stratifying factor, the second step in the stratification calls for estimating the chosen measure of association within each stratum. Averaging (when appropriate) the stratum-specific measures of association into a single number (adjusted measure of association) is usually not based on a simple arithmetic mean, but is based on a method of **weighted averaging** or **pooling** that takes into account the amount of information associated with each stratum-specific estimate.

The most commonly used method for averaging stratum-specific estimates of effect is the method proposed by Mantel and Haenszel (JNCI 22:719-748, 1959). Suppose that the following tables displays the data for the i^{th} stratum

	Disease		
	+	-	Total
Exposure+	a_1	b_i	N_{1i}
Exposure-	c_i	d_i	N_{01}
Total	M_{1i}	M_{01}	T_i

The formula for the **Mantel-Haenszel Weighted Average** is:

Risk Ratio estimate:

$$RR_{MH} = [\Sigma\{a_i N_{0i}/T_i\}]/[\Sigma\{c_i N_{1i}/T_i\}]$$

$$= [\Sigma\{w_i RR_i\}]/[\Sigma\{w_i\}] \text{ if } w_{i.} \neq 0$$

where $w_i = c_i N_{1i}/T_i$

Odds Ratio estimate:

$$OR_{MH} = [\Sigma\{a_i d_i/T_i\}]/[\Sigma\{b_i c_i/T_i\}]$$

$$= [\Sigma\{w_i OR_i\}]/[\Sigma\{w_i\}] \text{ if } w_i \neq 0$$

where $w_i = b_i c_i / T_i$

The weight (w_i) for RR_{MH} can be re-expressed as

$$w_i = [c_i/N_{0i}][(N_{1i}/T_i)(N_{0i}/T_i)][T_i]$$

From this representation, it follows that the value for the weight reflects the amount of information contained within the stratum by it being a function of the frequency of the outcome among non-exposed subjects (the first bracketed term), the balance of the relative sizes of the comparison groups (second bracketed term), and the overall size of the strata (the third bracketed term). These are three components that reflect the amount of information in the table.

Example # 1

The following tables show the crude and age-adjusted measures of association between sex and mortality among patients diagnosed with trigeminal neuralgia (Rothman. Modern Epidemiology Little Brown and Company 1986 and Rothman KJ, Monson RR. J Chron Dis 1973;26;303-309).

Crude Analysis:

	Deaths	Person-yrs	Mort. Rate (per 100 py)
Males	90	2465	3.65
Females	131	3946	3.32

$$RR_{Crude} = 3.65/3.32 = 1.10$$

Stratified (by aged) Analysis

	age < 65		age 65+	
	Deaths	py	Deaths	py
Males	14	1516	76	949
Females	10	1701	121	2245
Total	24	3217	197	3194

$RR_{age<65} = (14/1516)/(10/1701) = 1.57$

$RR_{age65+} = (76/949)/(121/2245) = 1.49$

$$RR_{MH} = [(14)(1701)/3217 + (76)(2245)/3194] / [(10)(1516)/3217 + (121)(949)/3194]$$

$$= 1.50$$

These data suggest that age is a confounder. Age satisfies the first criterion for confounding (1516/2465 = 62% of the male person-years are in the younger group, compared to 1701/3946 = 43% of the female person-years). Age also appear to satisfy the second criterion for confounding (the mortality rate among old females, 5.39 deaths/100py, is much greater than that for young females, .59 deaths/100py). This confounding by age is reflected be the difference between the crude (RR_{Crude} = 1.10) and adjusted (RR_{MH} = 1.50) measures of association.

Example # 2

The following tables show the crude and age-adjusted association between smoking and the 24-year risk of death in the FHS teaching data set.

Crude Analysis

	Died	Survived	Total
Smokers	788	1393	2181
Non-Smokers	762	1491	2253
Total	1550	2884	4434

$$RR_{Crude} = (788/2181) / (762/2253) = 1.07$$

Stratified Analysis

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

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

$$RR_{MH} = [\ \sum(a_i)(N_{0i})/Ti\ ] / [\ \sum\ (c_i)(N_{1i})/T_i]$$

$$= [(67)(302)/754 + (266)(684)/1639 + (286)(812)/1379 + (169)(455)/662] / [(25)(452)/754 + (110)(955)/1639 + (312)(567)/1379 + (315)(207)/662$$

$$= 1.38$$

Age appears to be a confounder in these data. The following table shows that age satisfies the first criterion for confounding (associated with the exposure).

	Non-Smokers (N= 2253)		Smokers (N=2181)	
Age	N	%	N	%
Age ≤ 40	302	13.40	452	20.72
40 < Age ≤ 50	684	30.36	955	43.79
50 < Age ≤ 60	812	36.04	567	26.00
Age > 60	455	20.20	207	9.49

The following table suggests that age satisfies the second criterion for confounding (associated with the outcome independent of the exposure).

Age	Estimated Risk Among Females (Non-Exposed)
Age $\leq$ 40	25/302 = .0828
40 < Age $\leq$ 50	110/684 = .1608
50 < Age $\leq$ 60	312/812 = .3842
Age > 60	315/455 = .6923

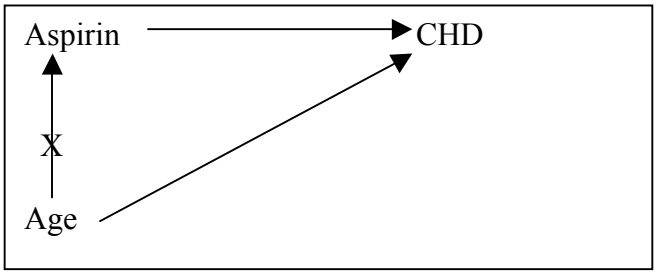
Confounding by age is reflected by the difference between the crude (RR_{Crude} = 1.07) and adjusted (RR_{MH} = 1.38) measures of association

Standardization

Standardization is a second method for adjusting for confounding through stratification. It is also used as a method for summarizing the effect of an exposure when there is **Effect Modification**. Standardization and Effect Modification will be discussed in the next set of lecture notes.

Design Methods of Avoiding Confounding

Randomization in experimental studies reduces for the potential for confounding. For example, in a large RCT examining the effect of aspirin on the risk of Coronary Heart Disease, age should not be a confounder as it would be expected to have very similar distributions in the aspirin and non-aspirin groups as depicted by the following DAG



Restriction is one way to avoid confounding in observational studies. For example, enrolling only subjects of a certain narrow age range in a Cohort Study would avoid confounding by age in a study comparing the incidence of CHD among aspirin and non-aspirin users. However, it may be difficult to generalize the result of this study to other age groups.

Matching is a less rigid for of restriction and may avoid confounding in a Cohort Study. For example, suppose for every aspirin user enrolled in the study, the investigator enrolled a non-aspirin user of the same age. As a result of this matching, the age distribution of the aspirin users would be identical to that of the non-aspirin users and age would not satisfy the first criterion for confounding (as depicted in the DAG above). The topic of matching will be covered in the next sequence of lecture notes.