# Confounding (Distortion) in Epidemiology

## الخلط أو الالتباس

Prof. Ashraf Zaghloul

# Aim in Epidemiology

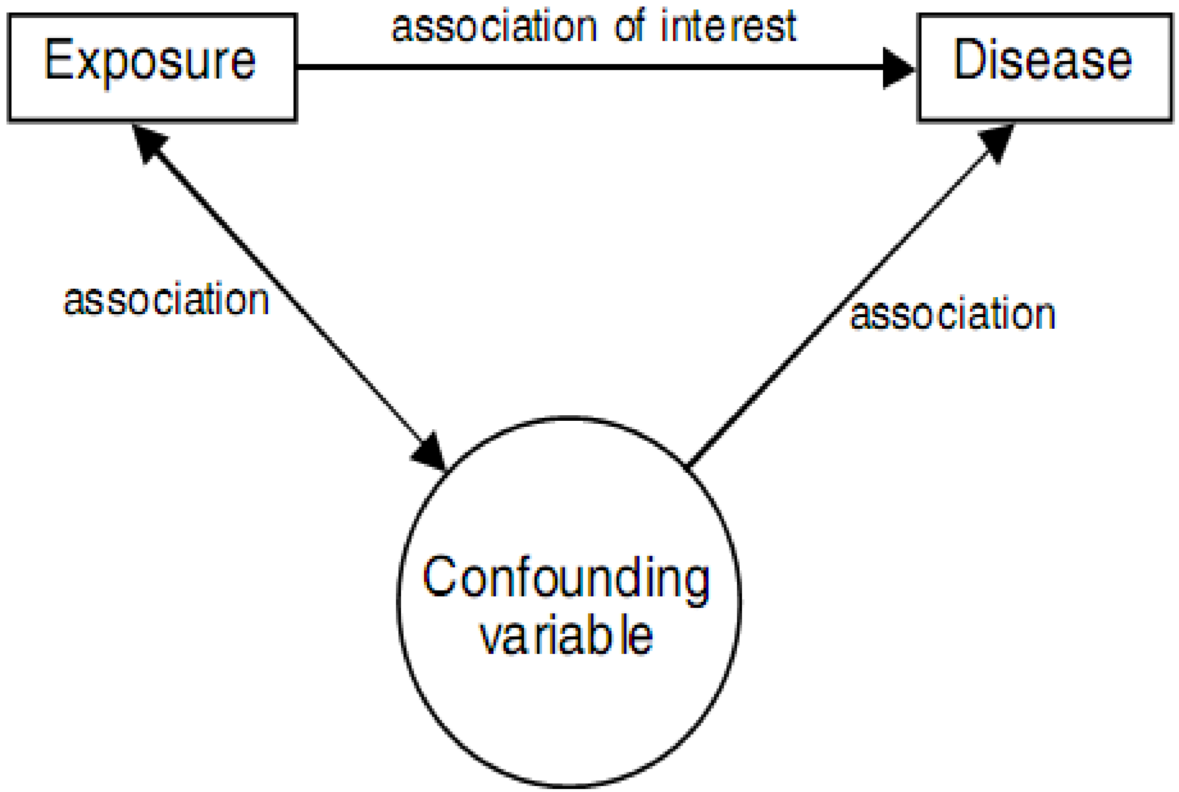
***At the heart of epidemiology is the notion of causality***

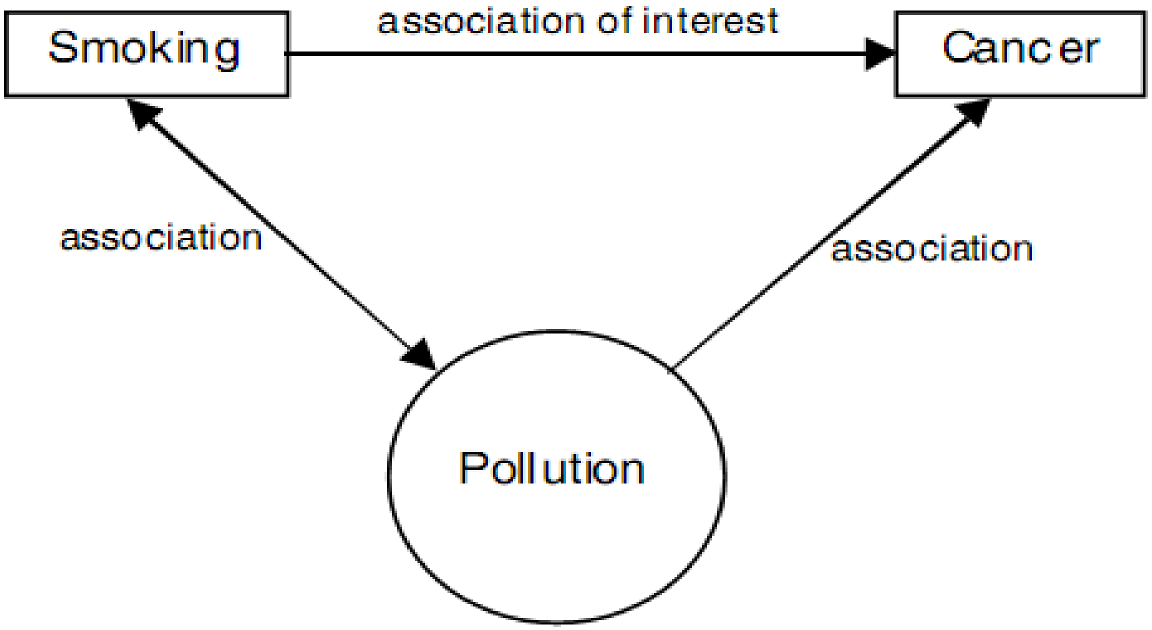
- *The idea is that when a causal association is established*, a protection and control attitude can occur rather than a mere reaction to the public health crises

# Definition

**A confounding variable is a variable (say, pollution) that can cause the disease under study (cancer) and is also associated with the exposure of interest (smoking)**

*The existence of confounding variables in smoking studies made it difficult to establish a clear causal link between smoking and cancer unless appropriate methods were used to adjust for the effect of the confounders*


Exposure
association of interest
Disease
association
association
Confounding
variable


Smoking
association of interest
Cancer
association
association
Pollution

- How do epidemiologists identify potential confounding variables? When such variables are suspected

- How can epidemiologists construct causal arguments in the face of these possible confounding variables?

# First Step

## Quantifying the Association Between Exposure and Disease

The search for factors that might be causally related to a disease (outcome) begins with the idea that people who have the exposure should have a different frequency of the disease from those who do not have the exposure


DANGER
CONTAINS
BENZENE
CANCER HAZARD

**Age:**

**Gender:**

**Income:**

**Occupation:**

**Working in a gas station:** **1- Yes** **2- No**

**X- ray chest:** **1- Positive lung cancer**

**2- No lung cancer**

| Gas station | Chest x-ray |
|---|---|
| 1.00 | 1.00 |
| 1.00 | 2.00 |
| 1.00 | 2.00 |
| 2.00 | 1.00 |
| 2.00 | 2.00 |
| 2.00 | 1.00 |
| 1.00 | 2.00 |
| 2.00 | 1.00 |
| 1.00 | 2.00 |
| 2.00 | 1.00 |
| 1.00 | 2.00 |
| 2.00 | 1.00 |

# Workers exposed to benzene vapours in gas stations

| | **Yes**<br>lung cancer | **No**<br>lung cancer | **Row Total** |
|---|---|---|---|
| **Yes**<br>Workers Exposed | **40** | **172** | **212** |
| **No**<br>Workers Unexposed | **18** | **253** | **271** |
| **Total** | **58** | **425** | **483** |

**Calculate the incidence of lung cancer in both groups**

| Exposed | Disease: Yes | Disease: No | Totals by Exposure status |
| --- | --- | --- | --- |
| Yes | a | b | (a + b) |
| No | c | d | (c + d) |
| Totals by Disease status | (a + c) | (b + d) | |

The proportions calculated

(i.e., $a/(a + b)$, $c/(c + d)$) are called risks

**Represent the risk that a person has of developing the disease**

If the proportion of those **exposed** who develop the disease is greater than **(>)** the proportion of those **not exposed** who develop the disease, we would say that the exposure and the disease are **positively associated**

$$\frac{a}{a+b} > \frac{c}{c+d}.$$

If the exposure is to a protective factor, the proportion of those exposed who develop the disease is less than (<) the proportion of those not exposed who develop the disease, and we would say that the exposure and the disease are negatively associated

$$\frac{a}{a+b} < \frac{c}{c+d}$$

If the exposure is unrelated to the onset of the disease, we would expect the proportions to be equal, in which case we would say that there is no association

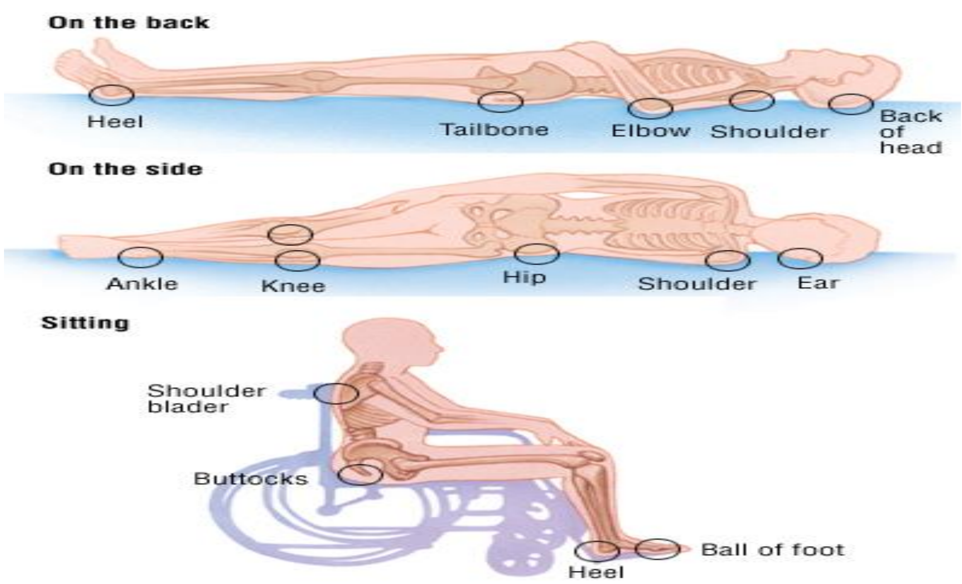
$$\frac{a}{a+b} = \frac{c}{c+d}$$

# Example

**Bedsores and Mortality**

- What is the risk of hip fracture patients of dying due to bed sores (study the association between bedsores and death among elderly hip fracture patients)

9,400 patients aged 60 and over were selected. The patients' medical charts were reviewed by research nurses to obtain information about whether they developed a bedsore during hospitalization and whether they died while in hospital



# Results of Bedsores Study, with Totals

| | Died | Did Not Die | Total |
|---|---|---|---|
| Bedsore | 79 | 745 | 824 |
| No Bedsores | 286 | 8,290 | 8,576 |
| Total | 365 | 9,035 | 9,400 |

- **What is the exposure in this example?**

- **What is the disease?**

| | |
|---|---|
| Number of people with a bedsore who died | 79 |
| Number of people with a bedsore who did not die | |
| Total number of people with a bedsore | |
| Number of people without a bedsore who died | |
| Number of people without a bedsore who did not die | |
| Total number of people without a bedsore | |
| Proportion of people with a bedsore who died | |
| Proportion of people without a bedsore who died | |

| | |
|---|---|
| Number of people with a bedsore who died | 79 |
| Number of people with a bedsore who did not die | **745** |
| Total number of people with a bedsore | **824** |
| Number of people without a bedsore who died | **286** |
| Number of people without a bedsore who did not die | **8,290** |
| Total number of people without a bedsore | **8,576** |
| Proportion of people with a bedsore who died | **79/824 = 9.6%** |
| Proportion of people without a bedsore who died | **286/8,576 = 3.3%** |

## Calculate the relative risk of death due to bed sores complications

**The likelihood of death was 2.9 times as high in people with bedsores as in people without bedsores**

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{79/824}{286/8576} = 2.9$$

Therefore, it would seem that there is a **fairly strong association** between bedsores and death

- Can we conclude, that **bedsores cause people to die?**

- Isn't it also possible that people with bedsores are more likely than people without bedsores to have some other characteristic, and that it is the other characteristic that is the cause of the higher death rate?

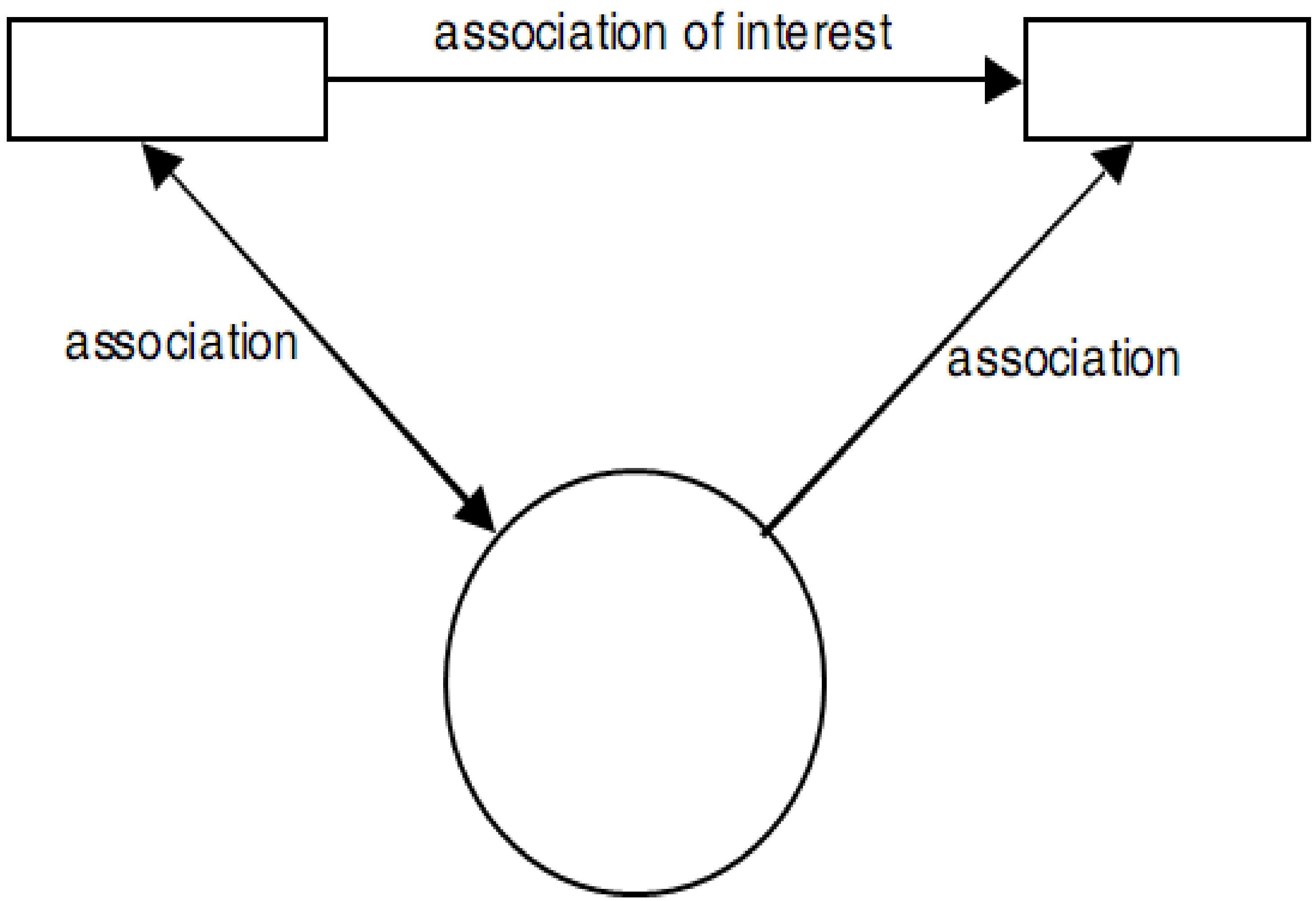
# Second Step

Can you think of **any possible confounding variables**?

A confounding variable in this example would be **a characteristic** that is more common in people with bedsores than in people without bedsores and that is associated with a higher death rate

The investigators suspected that people who have **lots of medical problems** aside from their hip fracture are more likely to get bedsores and are also more likely to die.

If that were the case, the **severity of medical problems** would be a confounding variable


association of interest
association
association

The epidemiologists who did this study obtained information about the patients' other medical problems.

The information was summarized into a score based on information about the patients' diseases when they were admitted to hospital.

To create two groups, the researchers classified everyone who had a score of 5 or more into the **high medical severity group** and everyone whose score was less than 5 into the **low medical severity group**

Of the 79 people who had bedsores and died, 55 had high medical severity and 24 had low medical severity.

Of the 745 people who had bedsores and did not die, 51 had high medical severity and 694 had low medical severity.

Of the 286 people who had no bedsores and died, 5 had high medical severity and 281 had low medical severity

Of the 8,290 people who had no bedsores and did not die, 5 had high medical severity and 8,285 had low medical severity

# Results of Bedsores Study, High Medical Severity Group

| | Died | Did Not Die | Total |
|---|---|---|---|
| Bedsore | | | |
| No Bedsores | | | |
| Total | | | |

| | Died | Did Not Die | Total |
|---|---|---|---|
| Bedsore | 55 | 51 | 106 |
| No Bedsores | 5 | 5 | 10 |
| Total | 60 | 56 | 116 |

# Results of Bedsores Study, Low Medical Severity Group

| | Died | Did Not Die | Total |
|---|---|---|---|
| Bedsore | | | |
| No Bedsores | | | |
| Total | | | |

| | Died | Did Not Die | Total |
|---|---|---|---|
| Bedsore | 24 | 694 | 718 |
| No Bedsores | 281 | 8,285 | 8,566 |
| Total | 305 | 8,979 | 9,284 |

# Calculate the relative risk in the high medical severity group

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

$$= \frac{55/106}{5/10}$$

$$= 1.04$$

# Calculate the relative risk in the low medical severity group

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

$$= \frac{24/718}{281/8566}$$

$$= 1.02$$

In answering the two preceding questions, you created strata based on categories of the suspected confounding variable and examined the exposure–disease association within each stratum

**This procedure is called stratification**

In each stratum, the association between bedsores and death cannot be explained by medical severity because in each stratum medical severity is held constant (as long as we can assume that all those with low severity have similar severity and all those with high severity have similar severity)

By looking at the relative risks separately in the high and low medical severity groups, **we have effectively adjusted for medical severity**

That is, we have estimated the bedsores–mortality association in a way that eliminates the effect of medical severity on that association

**When we stratify, we neutralize the effect of the confounder**

**Stratification is one of several ways to adjust for a confounding variable**

# Note

## Both relative risks are very close to 1

**INTERPRETATION**

- This means that the relative risk of death comparing those with and without bedsores, and adjusted for medical severity, is about 1
- This is quite different from the original relative risk of 2.9 that we found when we looked at the overall 2 X 2 table

# So

If the effect is observed to be as strong in the strata as in the overall table, it cannot be due to the confounding variable

Conversely, if the effect seen in the overall table is **reduced** or **eliminated** when we stratify (as in the example), ***the change in effect must be due to the confounding variable***

# Another way to state it

If the unadjusted and adjusted relative risks had been **similar**, it would mean that medical severity did not confound the association between bedsores and death

**The fact that the unadjusted and adjusted relative risks are different means that there is confounding by medical severity**

When there is confounding, **the confounder is associated with both the disease and the exposure**

# Conclusion

- When the effect of an exposure is mixed with the effect of another variable (the confounding variable), *we may incorrectly conclude that the disease is caused by the exposure*

- We might then attempt to eliminate the exposure in the hope that the disease could be prevented

If, however, the association between the exposure and the disease is due to confounding and is not causal, elimination of the exposure will not have any effect on the incidence of the disease

If there is an association between an exposure and a disease (outcome), but the association is entirely due to confounding, what will happen if you develop an intervention to eliminate the exposure?

# Will you have an impact on the prevention of the disease?

**There would be no impact of the intervention on the disease because the exposure is not causally related to the disease.**

**If A does not cause B, then eliminating A will have no effect on the occurrence of B**

If epidemiologists did not take great care to identify and control for **confounding**, *incorrect conclusions would be drawn, and time and resources would be unnecessarily expended with little hope of improving the well-being of the population*

# Exercise

To determine whether baldness causes coronary heart disease (CHD) in **men**, a hypothetical study was carried out.

The epidemiologist in charge of the study recruited 10,000 bald men and 10,000 men with hair into the study and followed all of them for 10 years to determine whether they developed CHD

| | CHD | | |
| --- | --- | --- | --- |
| | Yes | No | Total |
| Bald | 775 | 9,225 | 10,000 |
| Hairy | 190 | 9,810 | 10,000 |
| Total | 965 | 19,035 | 20,000 |

1. Calculate the risk of CHD among bald men.

2. Calculate the risk of CHD among hairy men.

3. What is the relative risk of CHD associated with baldness? Briefly explain how to interpret the value you calculated.

1- 0.0771

2- 0.19

3- 4.07

4. Does this result suggest that baldness may be a cause of CHD? What alternate explanation can you provide?

5. The investigator thought that the results might be confounded by age. What is meant by "confounded by age" in the context of this example?

The investigator stratified the results, displaying them separately for the older subjects (aged 65 and over) and the younger subjects (aged 40–64)

# CHD in Older Subjects

| | CHD | | |
|---|---|---|---|
| | Yes | No | Total |
| Bald | 750 | 6,750 | 7,500 |
| Hairy | 100 | 900 | 1,000 |
| Total | 850 | 7,650 | 8,500 |

# CHD in Younger Subjects

| | CHD | | |
|---|---|---|---|
| | Yes | No | Total |
| Bald | 25 | 2,475 | 2,500 |
| Hairy | 90 | 8,910 | 9,000 |
| Total | 115 | 11,385 | 11,500 |

6. Calculate the risk of CHD in the older men. Calculate the risk of CHD in the younger men. Does there appear to be an association between age and CHD? Why or why not?

7. What proportion of the older men are bald? What proportion of the younger men are bald? Does there appear to be an association between age and baldness?

8. Based on your answers to Questions 6 and 7, does the association between baldness and CHD appear to be confounded by age? Why or why not?

9. Calculate the relative risk of CHD associated with baldness in the stratum of older men. Briefly explain how to interpret the value you calculated.

6- Older group 0.1, 0.1

RR= 1

Younger group 0.01, 0.01

RR= 1

10. Calculate the relative risk of CHD in the stratum of younger men. Briefly explain how to interpret the value you calculated.

11. Compare the relative risks in Questions 9 and 10 with each other and with the overall relative risk you calculated in Question 3. Do these results suggest that the association between baldness and CHD is confounded by age? Why or why not?

# Simpson's Paradox

Simpson's paradox is an **extreme form of confounding**, where the association between two variables in a full group is in the opposite direction of the association found within every subcategory of a third variable

Consider a new drug treatment that initially appears to be effective, with 54 percent of treated patients recovering, as compared to 46 percent of placebo patients. However, when the sample is divided by gender, it is found that 20 percent of treated males recover compared to 25 percent of placebo males, and 75 percent of treated females recover as compared to 80 percent of placebo females.

So the apparent paradox is that the drug is found to be more effective than the placebo in the full group, but less effective than the placebo in each of the two gender-specific subgroups that fully comprise the combined group

| | Treatment | Placebo |
|---|---|---|
| Male | 10/50 (20%) | 20/80 (25%) |
| Female | 60/80 (75%) | 40/50 (80%) |
| All patients | 70/130 (54%) | 60/130 (46%) |

It would be concluded that the subgroup-specific results – the drug is not effective – and that the apparent effectiveness found in the combined group is merely a statistical artifact of the study design due to the gender confound

# Thank You