GEA1000 QUANTITATIVE REASONING WITH DATA
TUTORIAL 2

Please work on the problems before coming to class. In class, you will engage in group work.

Case Study 1: COVID in Italy and China

The following figures were obtained from a study[1] investigating the death rate among the COVID-19 patients (Case fatality rate) in Italy and China during the early COVID stages. Observe that for all age groups, death rates among Covid-19 patients in Italy are lower than those in China, but the overall death rate among the Covid-19 patients in Italy is higher than that in China.

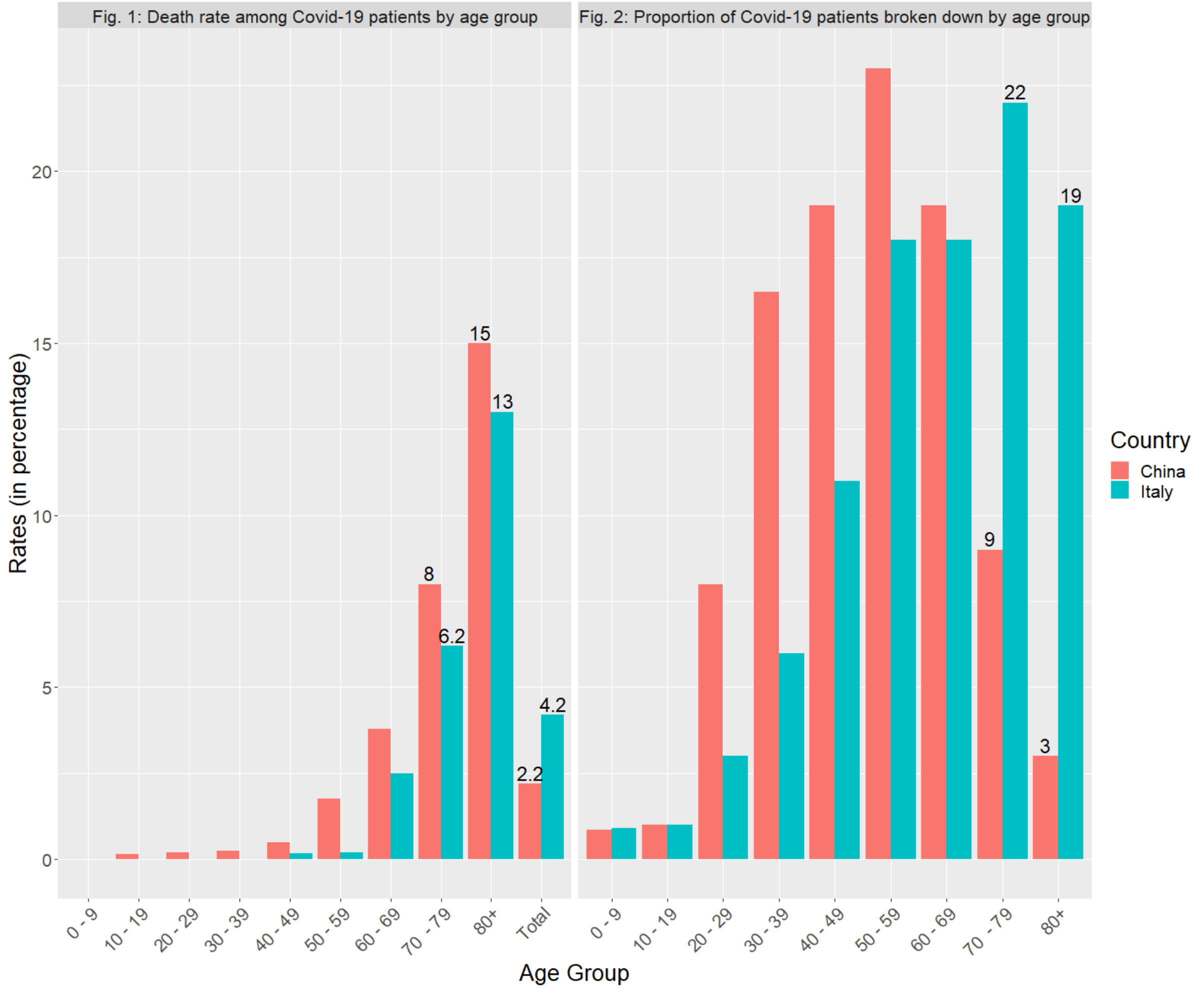


Fig. 1: Death rate among Covid-19 patients by age group

Fig. 2: Proportion of Covid-19 patients broken down by age group

1. Let us designate Covid-19 patients aged **70 years and above** as “Old”, and all other Covid-19 patients as “Young”. Let “D” represent death from Covid-19.

    a. What proportion of patients in Italy are old? How about China? Which country, Italy or China, is positively associated with old patients?

    Proportion of old patients in Italy: rate(Old | Italy) = 22%+19% = 41% or 0.41.
    Proportion of old patients in China: rate(Old | China) = 9%+3% = 12% or 0.12.

[1] Source: https://arxiv.org/pdf/2005.07180.pdf

There is a higher proportion of old patients in Italy. In other words, Italy is positively associated with having old patients, since rate(Old | Italy) > rate(Old | China)

b. In Italy, what is the death rate amongst the old patients, rate(D|Old)? Give your answers to 2 decimal places.
In Italy, rate(D|Old) = [13%(0.19) + 6.2%(0.22)]/[0.19+0.22] = 9.35%.
Note that we need to take the **weighted average** of the death rate in each age group.

**As to why you take weighted averages: You can't be sure that each age group occupies an equal slice of the pie – the "pie" here being the contracted-Covid population.*

c. From Fig 1, the overall death rate in Italy is 4.2%. Using the basic rule of rates, what must be the possible range of the death rate amongst the young patients in Italy? Is there an association between age and death among Covid-19 patients in Italy?
From Fig 1, we are given that rate(D) = 4.2%. By comparing to the answer in (b), we know that rate(D) < rate(D|Old), since 4.2% < 9.35%.
By the basic rule of rates: rate(D|Young) < rate(D) < rate(D|Old).
Thus, the possible range is: **0 < rate(D|young) < 4.2%.**

**As to why >0: Sanity check – rates cannot be negative (ever) nor zero (in this case)*

Since rate(D|Young) < rate(D|Old), this means that being old is positively associated to death in Italy.

d. Repeat parts (b) and (c) for China.
In China, Rate(D|Old) = [15%(0.03) + 8%(0.09)]/[0.03+0.09] = 9.75%.
From Fig 1. rate(D) = 2.2% in China.
Hence **0 < rate(D|Young) < 2.2%** and rate(D|Young) < rate(D|Old).

Being old is also positively associated with death in China.

e. Let's assume the following rough estimates from Fig 1:
In Italy, rate(D|Young) = 0.621%. In China, rate(D|Young) = 1.17%.
Using the information from Q1(a) to (d), explain how it is possible for the overall death rate in Italy to be higher than that in China, despite Italy having a lower death rate in China for every age group, as shown in Fig 1.

Hint: You may use the following table to help you:

	Italy	China
rate(D|Old)		
rate(D|Young)		
rate(D)		
rate(old)		

The table can be filled in using the answers from the previous parts and the information given. We can then use the table to explain the Simpson's paradox observed.

	Italy	China
rate(D|Old)	(b) = 9.35%	(d) = 9.75%
rate(D|Young)	0.621%	1.17%
rate(D)	**4.2%**	**2.2%**
rate(old)	(a) = 0.41	(a) = 0.12

From the basic rule of rates, we know that the overall rate(D) must be between the two individual rates, rate(D|Old) and rate(D|Young). However, the closer rate(Old) is to 100%, the closer the overall rate of death, rate(D), is to rate(D|Old), for each respective country.

Thus, the overall death rate in Italy is higher than China because there is a much higher proportion of old patients in Italy, and old patients are more likely to die from COVID.

From the table, we can also see the following reversal in rates when the data is sliced by age:
Amongst all old patients: rate(D|Italy) < rate(D|China)
Amongst all young patients: rate(D|Italy) < rate(D|China)
Amongst all patients : rate(D|Italy) > rate(D|China)

This is an example of how Simpson's paradox can occur.

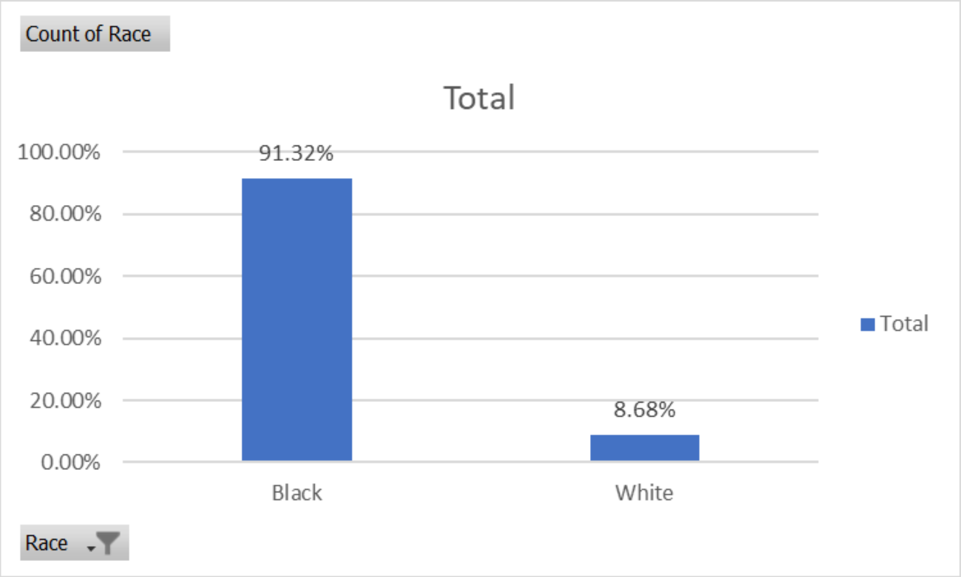
Case Study 2: Confounders and Simpsons Paradox

This question is based off a South African longitudinal study of growth of children, referred to as the Birth to Ten study (BTT). Census data of children born during a seven-week period between April and June 1990 were collected in the Johannesburg/Soweto metropolitan area of South Africa. The information collected included the **child's race (White / Black)** and whether they **received medical aid or not (Aid / No Aid)**. Having medical aid for the child is like having health insurance.

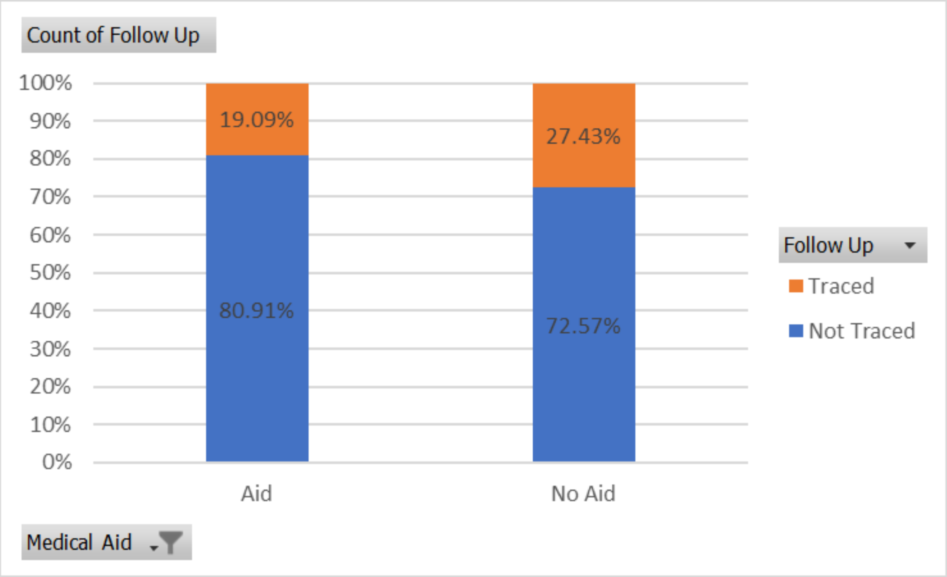
Five years later, a follow-up study on the same cohort of children was conducted. However, only 416 out of 1590 of the participants responded, despite medical screening being provided to the children as part of the follow-up study. These 416 children are labelled as "**Traced**", whereas those that did not respond are labelled as "**NotTraced**". Refer to the dataset: *Africa_study.xls*.

2. Use appropriate software to answer the following questions. Give your answers to 2 decimal places.
   a. What can you say about the proportions of whites and blacks in this study?

There are 91.32% blacks and 8.68% whites.



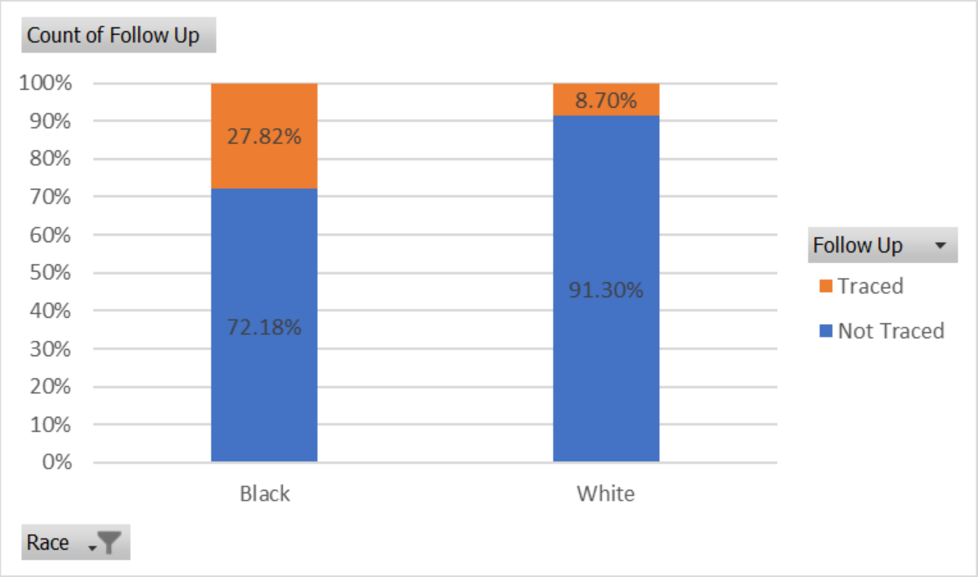
b. Determine if there is any association between the variables “Follow Up” and “Medical Aid”. Do children with medical aid tend to not follow up?



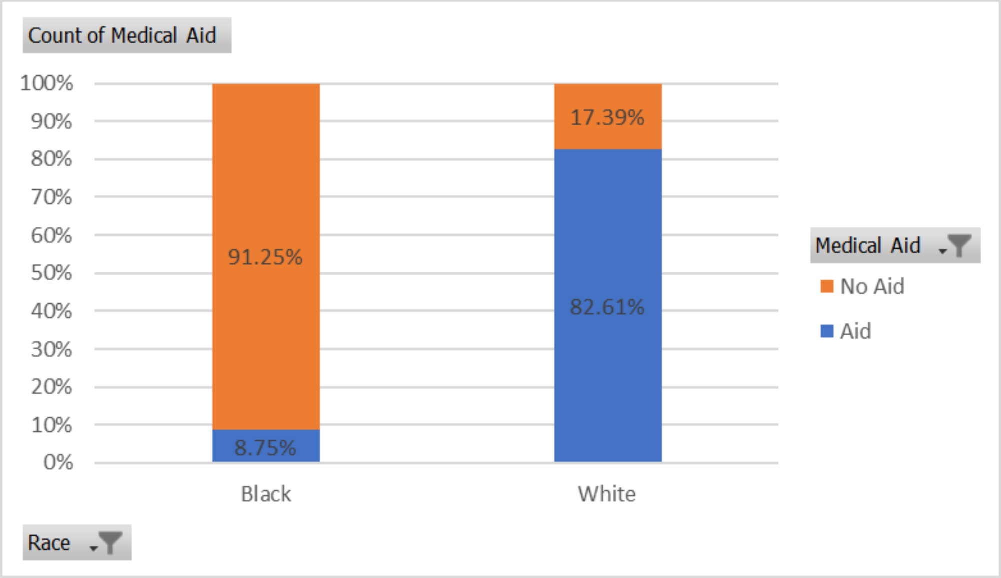
Since we can see that the rate(Traced | Aid) = 19.09% < 27.43% = rate(Traced | No Aid), we can conclude that there is a **negative association** between being traced and receiving medical aid. So children with medical aid tend not to come for the follow-up study, possibly because they are less incentivised by the free screening provided in the follow-up study.

c. Determine whether “Race” is a confounder in examining the association between the variables "Follow Up" and "Medical Aid".

For race to be a confounder, it must be associated with both "Medical Aid” and Follow Up”.

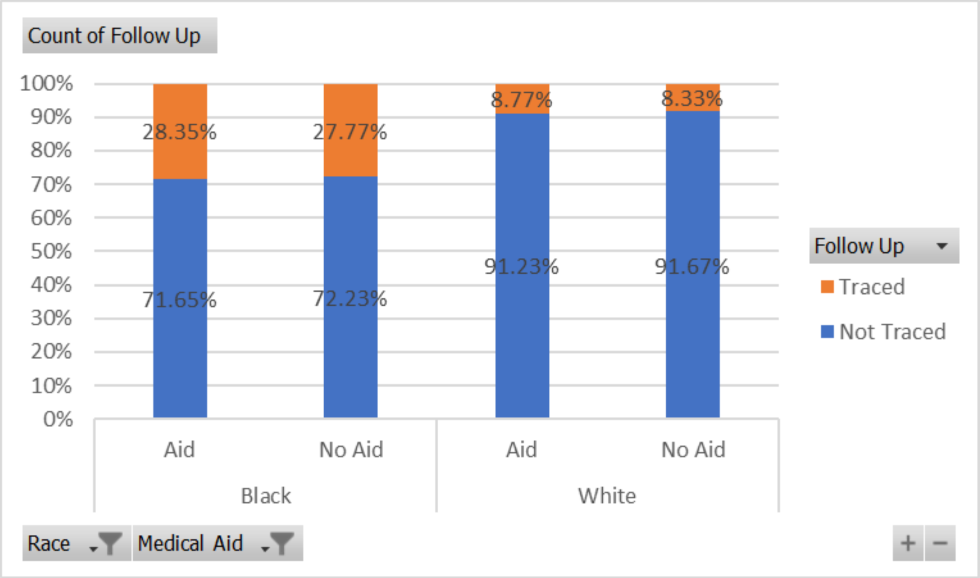


From the above, we can see that rate(Traced | Black) = 27.82% > 8.70% = rate(Traced | White). Hence, blacks are **positively associated** with getting Traced. It could be possible that whites felt that they had little to gain from participating in the study so less of them came, while the blacks may have valued the medical check-up and screening provided to the children in the follow up study as a replacement for (or in addition to) a regular medical screening.



We can also see that rate(Aid | Black) = 8.75% < 82.61% = rate(Aid | White). Hence, whites are **positively associated** with receiving medical aid.

d. In relation to (c), do we observe Simpson's Paradox when investigating the association between the variables "Follow Up" and "Medical Aid"?



Recall that in **part (b)**, we see that there is a **negative association** between being traced and receiving medical aid.

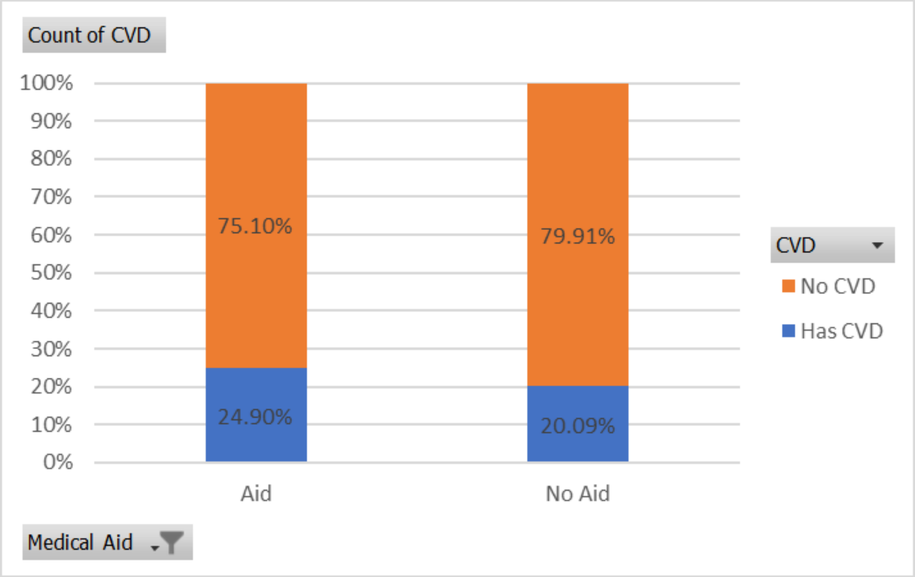
By **slicing according to "Race"**, we now see that rate(Traced | Aid) = 28.35% > 27.77% = rate(Traced | No Aid) for the blacks. Similarly, rate(Traced | Aid) = 8.77% > 8.33% = rate(Traced | No Aid) for the whites.

Since **the trend has reversed** when we are looking at the sliced subgroups, we have a case of Simpson's Paradox over here.

The aim of the study was to identify factors related to the emergence of cardiovascular disease (CVD) risk factors in children living in an urban environment in South Africa. Suppose that hypothetically, we have collected the data of these children many years later in a survey, to determine whether they ended up with cardiovascular disease or not, shown under the CVD (Yes/No) column. We are now interested in knowing if medical aid helps to mitigate the risk of CVD.

Thus, the outcome of interest (aka *response* variable) is now "CVD", and the treatment variable (aka *exposure* variable) is "Medical Aid".

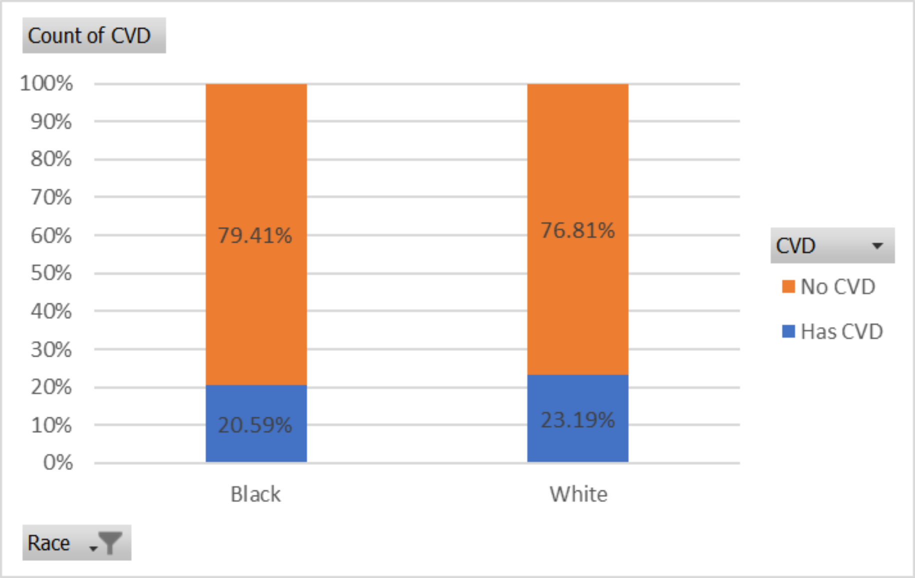
e. Describe the association between the variables "Medical Aid" and "CVD".



From the above, rate(Having CVD | Aid) = 24.90% > 20.09% = rate(Having CVD | No Aid). Hence, having CVD is **positively associated** with receiving medical aid.

f. Determine whether "Race" is a confounder in examining the association between "Medical Aid" and "CVD".

For race to be a confounder, it must be associated with **both** "Medical Aid" and "CVD".

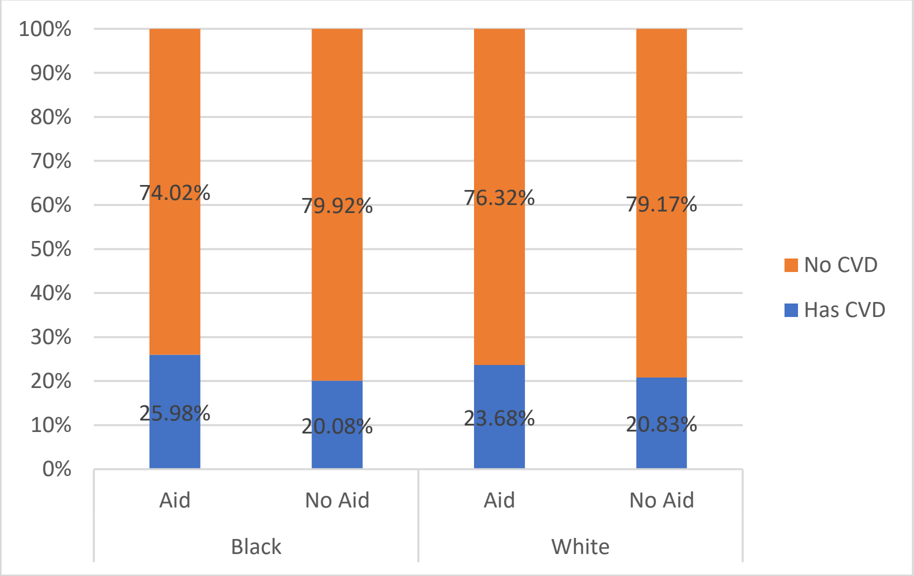


From the above, we can see that rate(Having CVD | Black) = 20.59% < 23.19% = rate(Having CVD | White). Hence, whites are **positively associated** with developing CVD.

Recall from **part (c)** that whites are **positively associated** with receiving medical aid.

Hence, “Race” is a confounder when studying the association between "Medical Aid” and “CVD”.

g. In relation to (f), do we observe Simpson’s paradox when investigating the association between the variables “Medical Aid” and “CVD”?



From **part (e)**, we know that receiving medical aid is **positively associated** with CVD. Looking at the sliced bar graph above, we see that having medical aid

remains positively associated with CVD for both whites and blacks, thus there is no Simpson's Paradox.

Even though, in **part (d)**, we found that race is a confounder, this does not necessarily mean that Simpson's Paradox will be observed in this particular comparison.

As part of data analysis, we should try to interpret the results based on the context of the study. It might seem counter-intuitive that people receiving medical aid are more likely to develop cardiovascular disease. A possible reason could be that there are many people with hidden cardiovascular diseases, but they only find out about it when they undergo medical screening. Thus, frequent medical screenings are more likely to uncover such conditions.

Case Study 3: Confounders in an experimental study

Background: Polio, also known as infantile paralysis, is an infectious disease that strikes young children, often causing permanent paralysis. It spreads through person-to-person contact. In the 1950's, American scientist Jonas Salk developed a vaccine that protected monkeys from polio and was safe when injected into human subjects in the laboratory in the 1950's. By 1954, the vaccine was ready to be tested in the real world.

In order to determine if the polio vaccine reduces the risk of polio infection, a cohort of children were invited to take part in a study. However, only some children had parental consent to receive the vaccine, which posed a problem for researchers.

Two different study designs are conducted:

First design: Children with parental consent were divided into two groups – treatment and control. The treatment group was vaccinated, and **no placebo was given to the control group**. **All children without parental consent were placed in the control group as well**, since they were not allowed to take the vaccine. This was known as the NFIP study *(you may read up more on this if you wish).*

Second design: Only children with parental consent were considered in this study. Those without consent were excluded. Children with parental consent were assigned into the treatment and control groups based on a 50-50 randomised procedure. The control group received a placebo injection of salt dissolved in water. Doctors involved in the diagnosis were not told which group the children belonged to.

The results of the two studies are tabulated as shown below.

NFIP study	Sample size	Polio +ve	Polio -ve
Vaccinated (with parental consent)	225000	56	224944
Control (parental consent or no parental consent)	725000	391	724609

Randomized controlled trial	Sample size	Polio +ve	Polio -ve
Treatment	200000	56	199944
Control	200000	142	199858

Furthermore, some important characteristics of the different groups were discovered:

- Families who provided consent tended to be of a higher income group and as a result lived in more hygienic conditions.
- Children living in more hygienic conditions were more susceptible to polio as they were not exposed to the virus since young and lacked immunity to the virus.

3a) Calculate the following for both study designs. For rates, leave your answers in percentages corrected to 3 s.f.

(i) Find the conditional rate of getting polio given that they are vaccinated.

NFIP study: 56/225000 x 100% = 0.025%

RCT: 56/200000*100% = 0.028%

(ii) Find the conditional rate of getting polio given that they are in control group.

NFIP study: 391/725000 x 100% = 0.054%

RCT: 142/200000*100% = 0.071%

(iii) Comment on the appropriateness of using rates rather than absolute numbers for comparison in the NFIP study.

Since the vaccinated group and the control group are unequal in size, rather than looking at the absolute numbers of polio cases, the use of rates allows for comparison between the two groups as it accounts for **unequal group sizes** (bases).

Although we can deal with unequal group sizes by using rates, we should still be careful if the sample size is too small for each group, which might give us unreliable results. (ie. changes in rates might result in only a small change in absolute numbers)

3b) Compare the two study designs by answering the questions below.

(i) Were the children randomly assigned into treatment and control groups?

NFIP study	Randomized controlled trial (RCT)
No random assignment was done control group - mixture of children with and without consent treatment group - children with consent.	Random Assignment - E.g. a random number generator was done to assigned children with parental consent into the treatment or control group.

(ii) Discuss how the assignment in 3b(i) might affect the results of the studies.

NFIP study	Randomized controlled trial (RCT)
- Children with consent are **more prone to polio** since they are more likely to come from a higher income family with a better hygiene in living conditions *(the logic being that they do not get opportunities to "train" their immune system).* - The control group (consisting of children with and without parental consent) would be at a lower risk overall than the treatment group. - The NFIP study is biased against the vaccine – it might cause the vaccine to **seem less effective**.	- With large enough samples, all the important characteristics of both treatment and control group subjects will resemble each other very closely - Allows for only the differences to be due to the response to the vaccine.

(iii) Were the subjects/assessors blinded?

NFIP study	Randomized controlled trial (RCT)
It was not stated if the doctors are blinded but the children knew if they were vaccinated or not as no placebo was given.	**Double-blinding**. - Children were blinded to the group they are in via the use of a placebo – in the form of the saltwater injection - Doctors involved in diagnosing the children are blinded to prevent making biased judgement.

(iv) Discuss how the blinding in b(iii) might affect the results of the studies.

NFIP study	Randomized controlled trial (RCT)
Double-blinding was not done for the NFIP study, which may lead to bias. For example, children in the control group who know that they are not getting the vaccine might result in them taking their own measures to reduce their risk of getting the disease. On the other hand, vaccinated individuals might think they now have full immunity and would be less stringent in health safety.	- With the use of placebo in the RCT, children do not know whether they were in treatment or in control and their response will thus be to the vaccine rather than the idea of the treatment. - Since many forms of polio are hard to diagnose, in borderline cases, doctors involved in diagnosing the children might be affected by knowing whether the children were vaccinated or not.
Remark: In general, the importance of blinding depends on the context of the study. In some cases, whether subjects are blinded or not may not affect the results of the study. Nonetheless, blinding and the use of placebos act as important precautionary measures, especially in situations where it is difficult to tell if blinding can have an impact on the results or not.	

(v) According to the table of results, by how much does the vaccine reduce the polio rate? If we consider the discussions in b(i) to (iv), do you think the vaccine more or less effective than what is stated in the table of results? Consider each study separately.

NFIP study	Randomized controlled trial (RCT)
- The vaccine seems to reduce polio rate by 0.054% – 0.025% = 0.029% - The **actual effect** of the vaccine is likely to be **larger than 0.029%**. - As discussed above, since no random assignment was done, the treatment group only consists of children with parental consent but the control group had a mixture of children with and without parental consent. This leads to the NFIP study being biased against the vaccine, causing the vaccine to **seem less effective than the results obtained**.	- The vaccine reduces the polio rate by 0.071% - 0.028% = 0.043% - The actual effect is likely to be similar to 0.043% as the study was conducted with randomised assignment, double blinding and large sample size.