Simpson's Paradox and Mediation Analysis in Covid-19

Zishan Cheng, Kehui Zhao

Abstract

In this final project, we perform data analysis on Covid-19 case fatality rates regarding Simpson's paradox as well as mediation analysis of age-related causal effects. Our inspiration comes from the works done by Kugelgen, Gresele and Scholkopf (2020). This paper proceeds as follows: in first section we review the inspiration paper. In second section we reproduce the Simpson's paradox analysis on China and Italy, and supplement analysis on Sweden and Switzerland. Section three we did mediation analysis on China and the US, and among three states in the US. Section four is discussion and thoughts on our project.

1. Reference review

In this paper, the authors look at the case fatality rates (CFR) of Covid-19, which indicates the proportion of confirmed cases which end fatally. In addition to total CFR, CFRs are often also reported separately by age since they differ significantly across different age groups, with older people at higher risk.

1.1 Simpson's paradox

When comparing CFR in China and Italy, the authors observe Simpson's paradox, suggesting opposite conclusions depending on whether the data is analyzed in aggregate of age-stratified form. By comparing China data on Feb 17 and Italy data on March 9, they found that if look at aggregated CFR, Italy is higher; but in each age group, Italy actually has lower CFRs.

1.2 Mediation model

The authors then build a causal model based on the assumption that age acts as a mediator of the effect of country on mortality, as shown in figure 1.

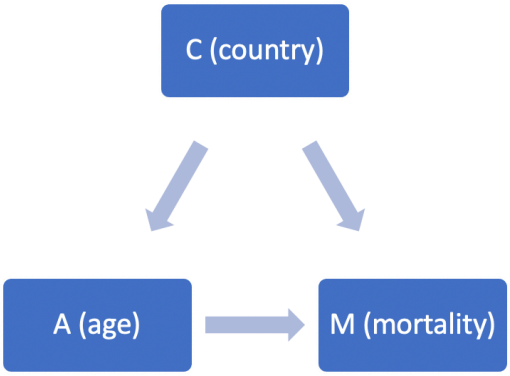

Figure 1: Assumed causal graph

In particular, this view encompasses at least the following influences:

• (C → A) encodes that the age distribution of cases is country- dependent. This difference might be due to a general difference in age demographic between countries, but other mechanisms such as inter- generational mixing or age-targeted social distancing may also play a role.
• (A → M) reflects the notion that the disease is more dangerous for the elderly, i.e., age appears to have a causal effect on mortality.
• (C → M) summarizes country-specific influences on mortality other than age, e.g., approaches to testing, lockdown strategy and other non- pharmaceutical interventions, air pollution levels, and medical infrastructure, e.g., availability of hospital beds and ventilators. We will refer to the combination of all these effects as a country's approach.
The authors then define different causal effects of interest:

1) Total causal effect (TCE), which simply given by the difference in total CFRs;
2) Controlled direct effect (CDE), which is given by the difference of CFRs for a given age group.
3) Natural direct effect (NDE), which corresponds to asking about the effect of switching country without affecting the age distribution across the confirmed cases, a direct path C → M.
4) Natural indirect effect (NIE), which isolates the indirect effect that a country exhibits on mortality only via age, C → A → M.

Their empirical formula can be found defined as follows:

$$\mathrm{TCE}^{\mathrm{obs}}_{0\to 1} = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0],$$
$$\mathrm{CDE}^{\mathrm{obs}}_{0\to 1}(x) = \mathbb{E}[Y|T=1, X=x] - \mathbb{E}[Y|T=0, X=x].$$
$$\mathrm{NDE}^{\mathrm{obs}}_{0\to 1} = \sum_x P\big(X=x|T=0\big)\big(\mathbb{E}[Y|T=1, X=x] - \mathbb{E}[Y|T=0, X=x]\big),$$
$$\mathrm{NIE}^{\mathrm{obs}}_{0\to 1} = \sum_x \big(P(X=x|T=1) - P(X=x|T=0)\big)\mathbb{E}[Y|T=0, X=x].$$

In their analysis the authors tracing causal effects over time in comparing Italy data ranging from March 9 to May 26 with China data on Feb 19, and discover that 1) from TCE we can tell that Italy has a higher total CFR that increases rapidly; 2) NDE tells that initially China would benefit from changing approach to Italy's, but after mid-March when there were an overwhelmed heath care system reported, switching to the Italian approach would increase the total CFR; 3) a stable higher NIE in Italy tells that switching to the case demographic in Italy would lead to increase in total CFR.

In summary, while indirect age-related effects considerably contribute to differences in total CFR— especially initially, when the instance of Simpson's paradox is reflected in the opposite signs of NDE and NIE—it is mainly the direct effect that drives the observed changes over time. Results shown in Figure 2. They then moved forward to compare NDEs and NIEs among switching countries from control to treatment. Results shown in Figure 3.

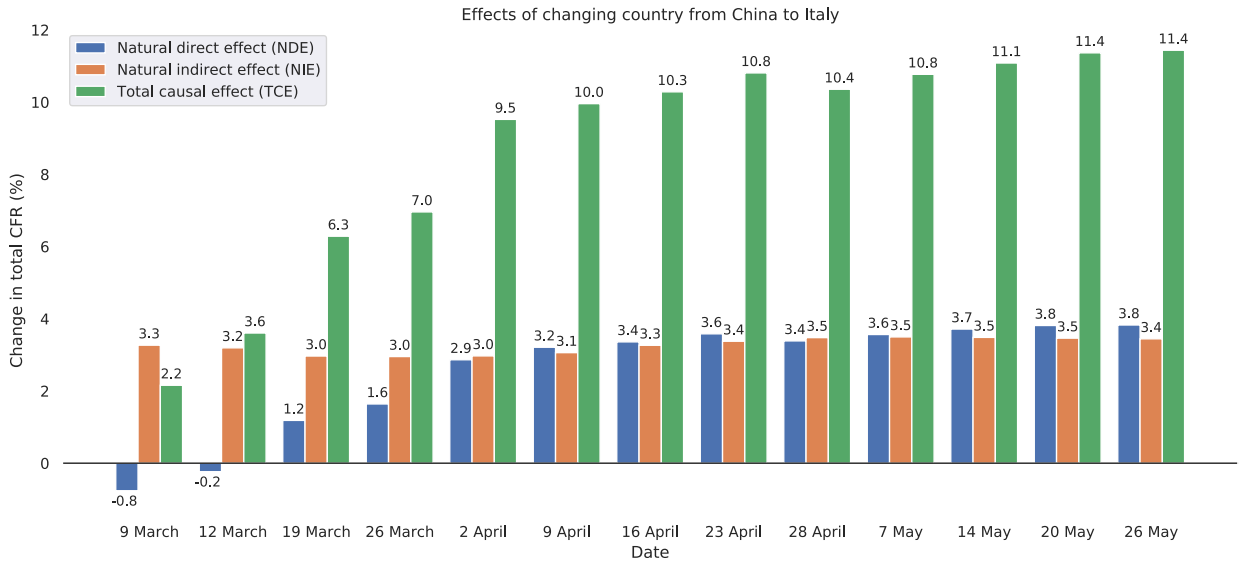

Figure 2: Evolution of TCE, NDE, and NIE of changing country from China to Italy on total CFR over time.

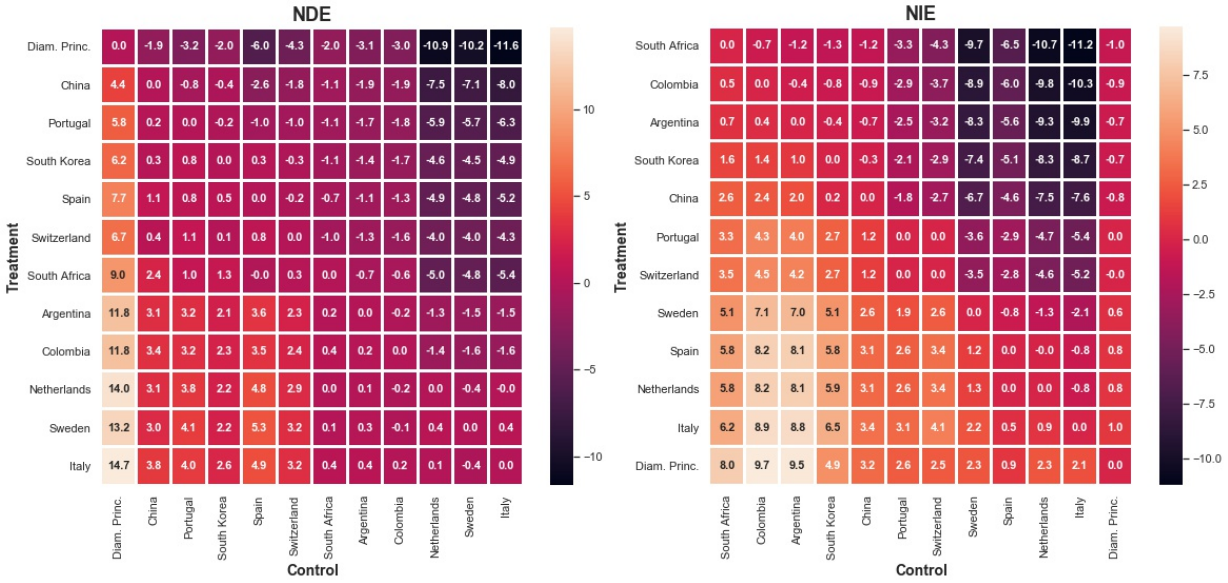

Figure 3: NDEs (left) and NIEs (right) for switching from the control country (columns) to the treatment country (rows)

Their final conclusion is that it is safe to assume that the virus is ultimately agnostic to the notion of different "countries" and that the influence of country on mortality C → M is not actually a direct one, but instead mediated by additional variables Xi. Candidates for such additional mediators Xi include, e.g., non-pharmaceutical interventions and critical healthcare infrastructure.

1.3 Dataset

They curated a dataset involving 756004 confirmed cases and 68508 fatalities, separated into age groups of 10-year intervals (0-9, 10-19, etc.), reporting from 11 different reported from 11 different countries from Africa, Asia, Europe and South America and the Diamond Princess cruise ship. We make use of their dataset in our project, and also add our own data from the US in total and three states: Minnesota, Illinois and DC. All of our data and codes can be found in the github.

2. Simpson's paradox in Covid-19 case fatality rates

There's an interesting Simpson's paradox in Covid-19 case fatality rates, which is the rate to die when you are infected by coronavirus. And it's calculated as the proportion of confirmed Covid-19 cases within a given group which end fatal.

When comparing Covid-19 CFRs for different age groups reported by the Chinese Center for Disease Control and Prevention with preliminary CFRs from Italy as reported on March 9 by the Italian National Institute of Health, a seemingly strange pattern can be observed: for all age groups, CFRs in China are larger than those in Italy, but the total CFR in China is lower than that in Italy. We can see this amazing conclusion clearly from the graph below: in every age bucket, you are more likely to survive when you are in Italy than in China because of its lower CFRs. However, when you aggregate them all up, it's better off to be China instead of Italy.

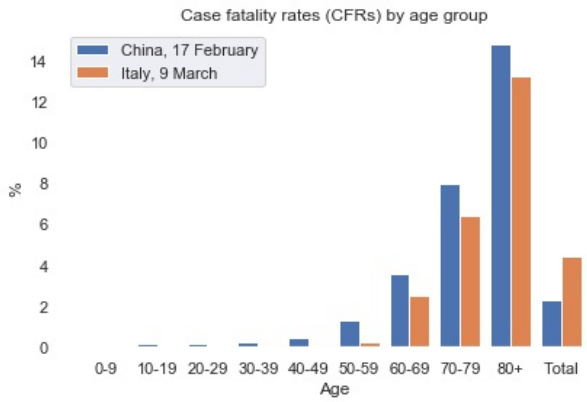

Figure 4: CFRs by age groups in two countries

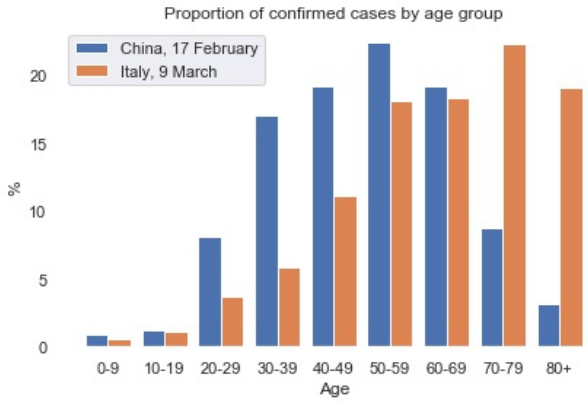

Figure 5: the prop of confirmed cases by age groups

Why we encounter Simpson's paradox here? The key to explain it is to distinguish the rate and the number of absolute death cases. When it comes to the proportion of confirmed cases by age group, we can see the majority patients in China is of age 40 – 60, while that in Italy is of age 60 – 79, which is at the high risk of dying from Covid-19. From Figure 5 Italy reported a much higher proportion of confirmed cases in older patients compared to China. And it's obvious that younger people are more likely to survive. Based on this fact, since Italy has this high proportion of elder generation cases, its total case fatality rates are pulled down by the high CFRs of elder people.

The followings are two pie charts of demographic of China and Italy population. From the figures, we can see Italy is experiencing severe ageing, and the country has a large number of elderly people, who are of higher risk of contracting the coronavirus. Compared with Italy, China has more younger people who can recover quickly from the Covid-19. And most of youngers don't need to go to hospital and they can recover by their immune system, which is a great relief for the medical system of China so that more beds can be provided for critically ill patients, leading to lowering the CFRs of China.

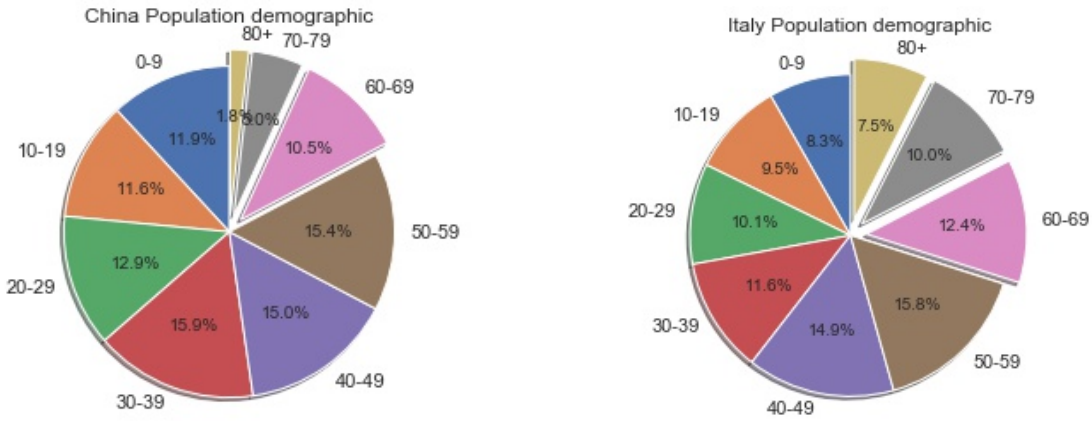

Figure 6: pie chart of China (left) and Italy (right) population demo

To prove that the population demographic is the key for Simpson's paradox in Covid-19 CFRs, let's do the same comparison for two countries of similar age structure. Here we compare two North Europe countries: Sweden and Switzerland. We can see from figure 7, Sweden and Switzerland have close proportion of elder people. Based on similar age structure, we expect there's no Simpson's paradox in Covid-19 fatality rates here.

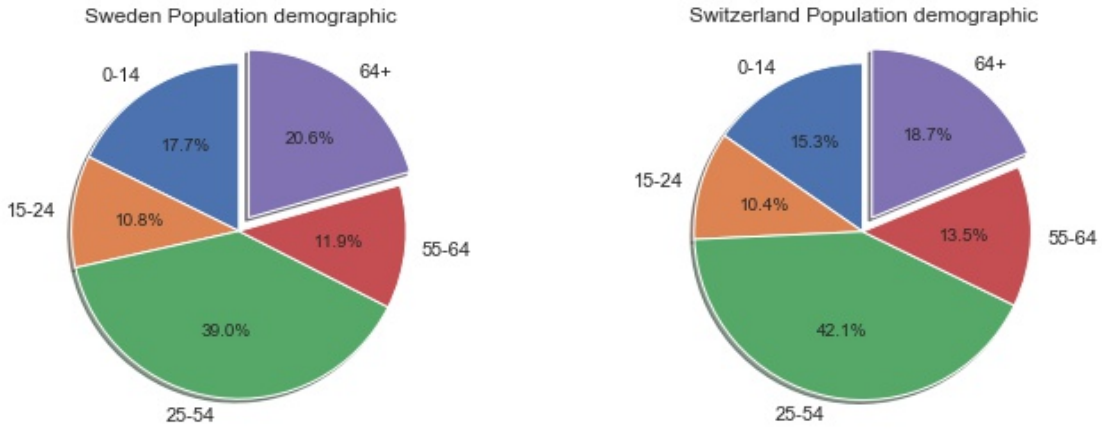

Figure 7: pie chart of Sweden (left) and Switzerland (right) population demo

From Figure 8 and 9, we can see the total CFRs are corresponding with the age-stratified CFRs. For each age group, Sweden has higher CFRs and thus gives rise to higher total CFRs. When

exploding its confirmed cases, we found Sweden has a high proportion of confirmed cases of age 80+, leading to a higher CFRs in total.

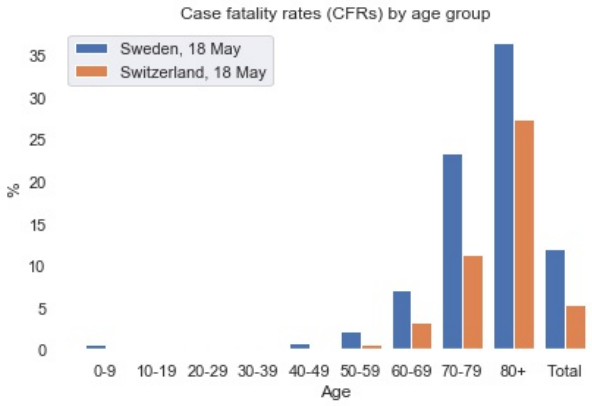

Figure 8: CFRs of Sweden and Switzerland

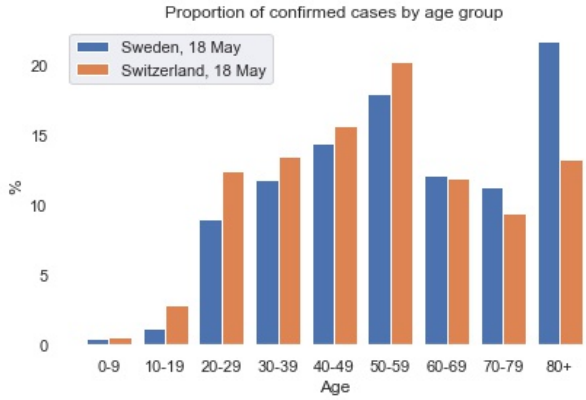

Figure 9: prop of confirmed cases of two countries

In summary, the larger share of confirmed cases among elderly people in Italy shown in Figure 4, and the fact that the elder people are generally at higher risk when contracting Covid-19, explains the inconsistency between total CFRs and CFRs by age groups shown in Figure 4 and thus leads to Simpson's paradox.

If we look at the CFRs by age group and case demographic for different countries, we can find some interesting facts. For example, the total CFRs is mainly corresponding with CFRs of 70+. What's more, there's also some mismatch between CFRs and proportions of confirmed cases. Such as Diamond Princess Cruise Ship, it has a very high proportion of confirmed cases in age group 60-79, but very low CFRs, nearly close to zero. What's more, the proportion for adults to contract Covid-19 in their age groups doesn't vary much from group to group, but the CFRs do!

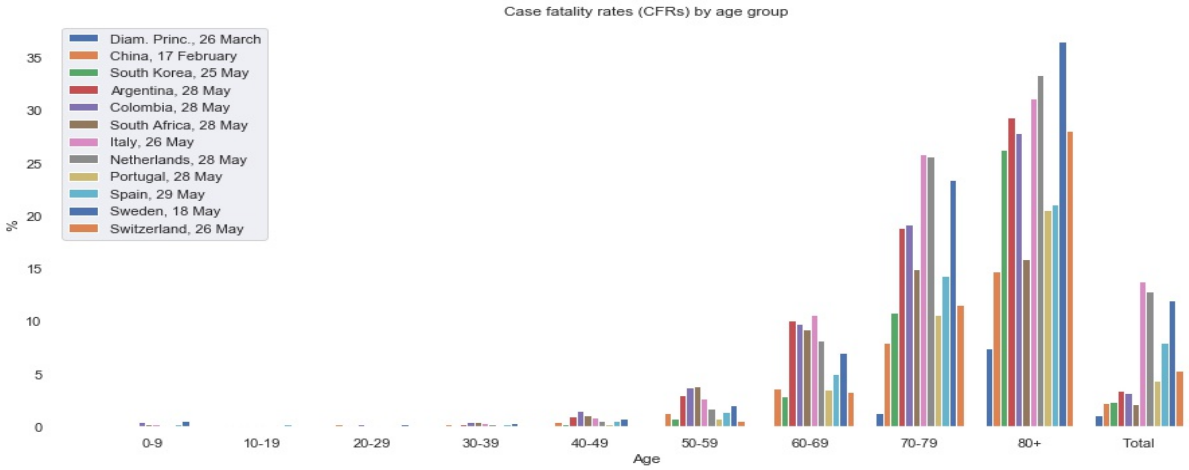

Figure 10: CFRs of different countries by age group

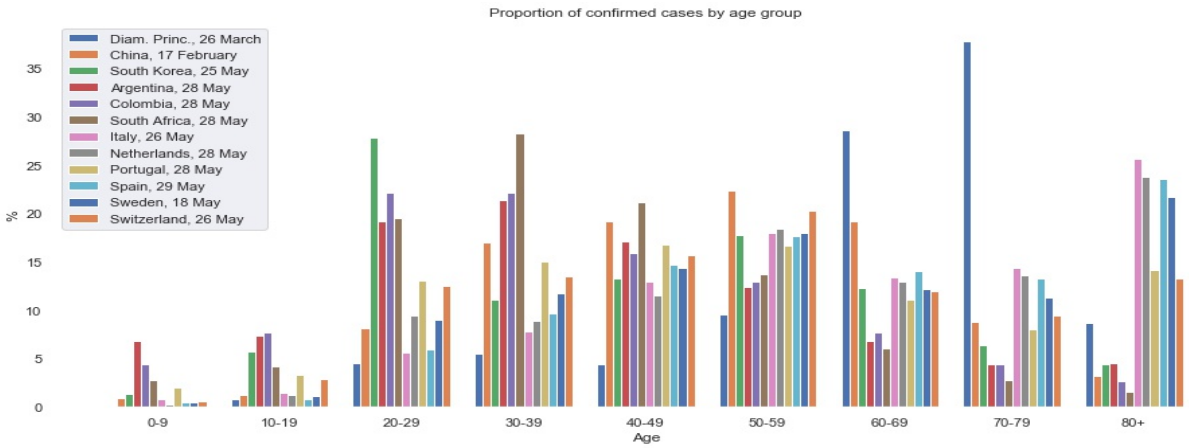

Figure 11: Proportion of confirmed cases for different countries

3. Mediation models: using US data

We perform similar mediation analysis as the authors, but instead utilizing our data downloaded from CDC in the US (updated to Nov 19) and three states' (MN, IL, DC) Healthcare department dataset (updated to Dec 10).

3.1 Tracing causal effects over time

First, we investigate the temporal evolution of direct and indirect (age-mediated) causal effects on mortality. The result of tracing TCE, NDE, and NIE of changing from China to US over a period of 10 months using data from CDC is shown in Figure 12. Note that case and fatality numbers for China remain constant in the figure, so that any changes over time can be attributed to US.

First look at TCE. We observe a pattern of first increasing to 4.9% in end of April, and then decreasing to -0.1% towards end of November. Remember TCE measures what would happen to the total CFR if both CFRs by age group and case demographic in China were changed to those from US. Thus our results implies that the difference between the two countries' total CFR becomes more pronounced at first, but then drops slowly.

To understand what drives the difference, we consider the NDE. NDE first increase rapidly from 0.4% to up to 3.6% as TCE, and around the same turning point it began to slowly drop down to 0.3%. It captures what would happen to the total CFR if the case demographic were kept the same, while only the approach (CFRs per age group) was changed. Therefore, the stable positive results imply that US approach is outperformed by China's, but after May the US is trying better approach. This seems to correspond to Trump's refusal to take the Covid-19 seriously and make restrictions on control its development at the beginning of the pandemic before May, when the outbreak number kept rising and US becomes world's no.1 country of confirmed cases. Only after that did Trump gradually apply more strict methods.

Last, we consider NIE, which measures what would happen to total CFR if the approach were kept the same, while the case demographic was changed to that in US. It seems the effect varies slightly over time with highest point at 0.6% but then drops to negative very soon. As can be seen from

population structure data, the US has a population with more younger residents. China is suffered more from the problem of population aging. Another issue may attribute here is the demographic structure. Maybe the more diverse races components in the US makes its average population has better health status. It can't be denied that there exists differences in natural endowments on body or health status among different races.

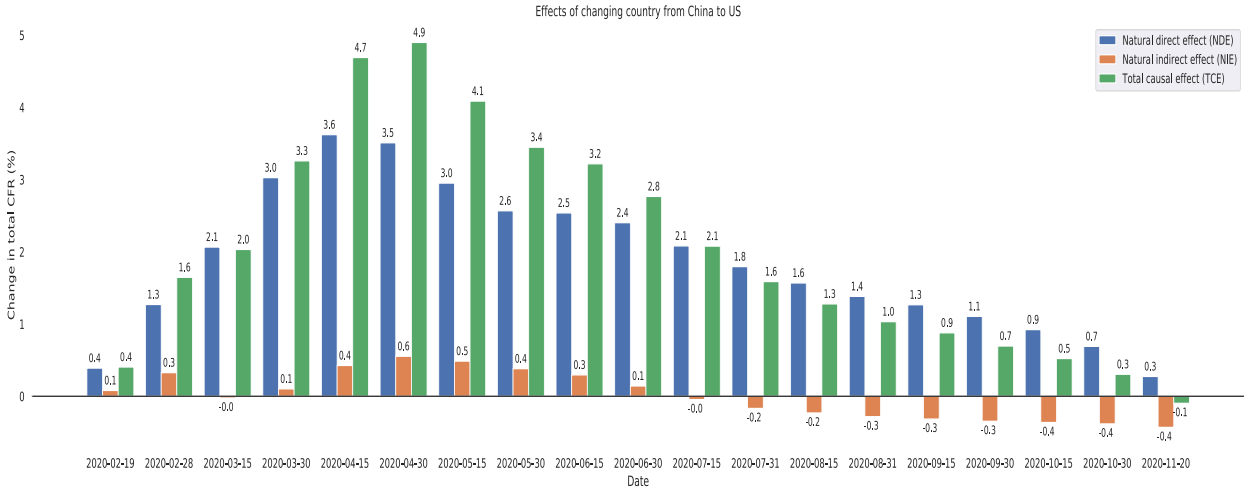

Figure 12: Evolution of TCE, NDE, and NIE of changing country from China to US on total CFR over time

3.2 Comparison between several different states

Since the US is consists of rather independent states which has their distinct population structure and applies their own strategies, it makes sense to compare the CFRs on state level. The results are shown in figure 13. In particular we compare three states: Minnesota in the further North, Illinois in between and DC in the East. We see that regarding NDE, DC performed the worst in its approach, since switching to the two other states will reduce significantly the total CFR, while MN is slightly better than IL. In terms of NIE, we see that DC also has a worst performed case demographic, which may relate to an aging population there.

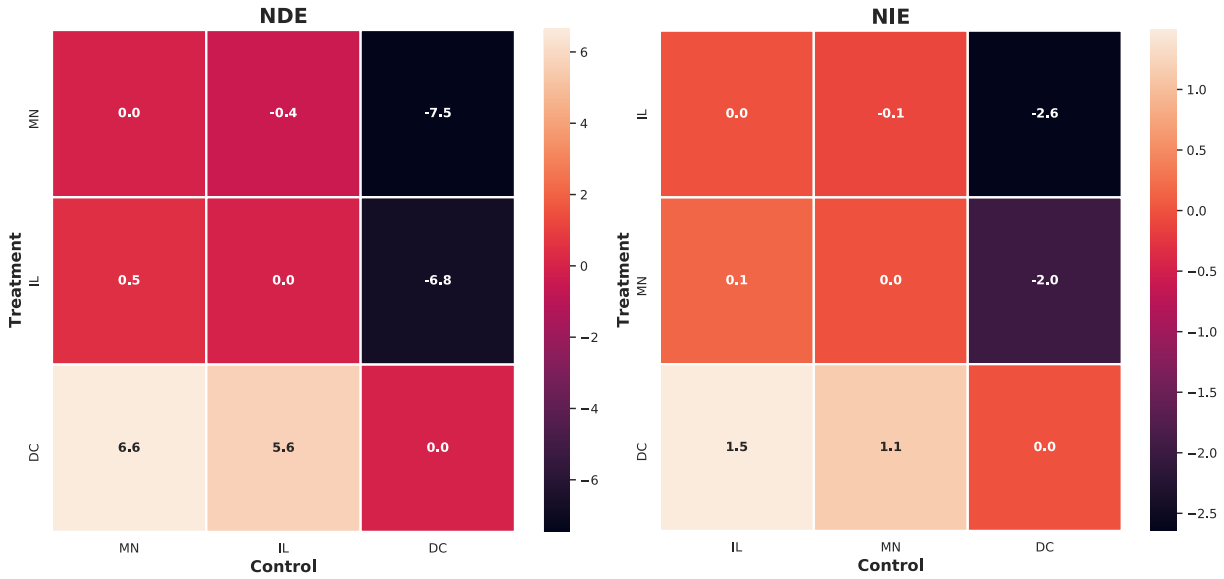

Figure 13: NDE (left) and NIE for for switching from the control state (columns) to the treatment state (rows)

4. Discussion

We successfully reproduce the authors works in our reference paper, and managed to add new data in the US to perform mediation analysis. We made reasonable interpretation on our results, which may give some inspirations to future analysis. Some limitations that can be promoted by future analysis includes more data from different states to create more comparable results. The difficulty of collecting US data is that different states using different format when reporting: some only has age group stratified data for deaths but not for cases; many use different grouping of age so that it's hard to compare each other; tests methods also varies in states…But it would be really interesting to consider such problem. Another inspiration might be including more mediator factors as our reference paper suggests. At least in the US we can look at CFRs among different races.

Reference

[1] arXiv:2005.07180 [stat.AP] Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects (Julius von Kügelgen, Luigi Gresele, Bernhard Schölkopf)
[2]data resources: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf
https://www.health.state.mn.us/diseases/coronavirus/situation.html#ageg1
https://coronavirus.dc.gov/data
http://www.dph.illinois.gov/covid19/statistics
[3] Parts of data and code used by this project are from: github.com/Juliusvk/Covid19-age-related-causal-effects
[4] Our data and code can be found: https://github.com/zcheng233/Causal_effect_final_project