Department of Economics, University of Warwick

Monash Business School, Monash University

as part of

Monash Warwick Alliance

The Role of Social Contact in the Infectious Disease Spreading: Evidence from the 1918 Influenza in Sweden

Xinghua Qi

Warwick-Monash Economics Student Papers

March 2023 No: 2023/49

ISSN 2754-3129 (Online)

The Warwick Monash Economics Student Papers (WM-ESP) gather the best Undergraduate and Masters dissertations by Economics students from the University of Warwick and Monash University. This bi-annual paper series showcases research undertaken by our students on a varied range of topics. Papers range in length from 5,000 to 8,000 words depending on whether the student is an undergraduate or postgraduate, and the university they attend. The papers included in the series are carefully selected based on their quality and originality. WM-ESP aims to disseminate research in Economics as well as acknowledge the students for their exemplary work, contributing to the research environment in both departments.

"*We are very happy to introduce the Warwick Monash Economics Student Papers (WM-ESP). The Department of Economics of the University of Warwick and the Economics Department at Monash University are very proud of their long history of collaboration with international partner universities, and the Monash Warwick Alliance reflects the belief in both Universities that the future will rely on strong links between peer Universities, reflected in faculty, student, and research linkages. This paper series reflects the first step in allowing our Undergraduate, Honours, and Masters students to learn from and interact with peers within the Alliance.*"

Ben Lockwood (Head of the Department of Economics, University of Warwick) and Michael Ward (Head of the Department of Economics, Monash University)

Recommended citation: Qi, X. (2023). The Role of Social Contact in the Infectious Disease Spreading: Evidence from the 1918 Influenza in Sweden. *Warwick Monash Economics Student Papers* 2023/49

.

[1] Warwick Economics would like to thank Lory Barile, Gianna Boero, and Caroline Elliott for their contributions towards the selection process.

The Role of Social Contact in the Infectious Disease Spreading: Evidence from the 1918 Influenza in Sweden [*]

Xinghua Qi[*]

Abstract

Infectious disease has always been a concern to people, especially under the current COVID-19 pandemic. The aim of this paper is to find a causal relationship between social interaction and disease spreading. This paper takes the 'Spanish Flu' in 1918 in the background of Sweden rather than COVID to rule out some uncertainty in transmission tunnels and use railway access as proximity to social contact. Using Diff-in-Diff identification, combined with a short-term event-study design, I show that localities that have railway stations nearby are likely to have more death cases during the influenza period. I use exogenous variation in railway station emergence from initial railway plans in addition and verifying that railway indeed facilitates the disease transmission and mortality rate as well but only with limited effects.

Keywords: disease spreading, railways, 1918 Influenza, Sweden

JEL: Y40

[*]I want to extend a sincere and heartfelt obligation towards all the personages without whom the completion of the project was not possible. I express my profound gratitude and deep regard to Prof. Dr. Sonia Bhalotra, the University of Warwick; Prof. Dr. Daniel Kühnle, the University of Duisburg-Essen; Prof. Dr. Martin Karlsson, the University of Duisburg-Essen and Prof. Dr. Eric Melander, the University of Birmingham for their guidance, valuable feedback, and constant encouragement throughout the project. Their valuable suggestions were of immense help. I sincerely acknowledge their constant support and guidance during the project.

[*]Email address: qixinghua990512@163.com

Dropbox link for data and additional files: https://www.dropbox.com/s/fmneaay769uq3ud/online%20appendix.zip?dl=0

1 Introduction

Since the 14th century Black Plague, infectious diseases were considered as a symbol of death, which has become one of the major threats to humans, as well as the global economy and society. In the past, most literature discussed viral diseases in a theoretical way regarding the relationships between disease severity, government policies and individual behaviours (Perra 2021). The interest in viral diseases has been renewed by the pandemic of COVID-19. Brodeur et al. (2021) note that there were many articles that almost discussed every aspect of COVID-19, which also pointed out the importance and effects of social contact and activities in the spread of COVID-19. However, it is hard to exempt endogeneity since many factors may affect the transmission of the novel disease.

Therefore, I prefer to use great influenza in 1918 as a research object. I would like to use the high-frequency data collected in Sweden during the influenza period to contribute to our understanding empirically regarding the relationship between social contacts and disease spreading. The reasons for selecting Sweden are contextual factors as it is much easier to specify the channels of disease spreading and there are a large number of mortality data available. Because of the scarcity of medical treatment during the pandemic, many non-pharmaceutical interventions' effectiveness in the 1918 influenza such as school closure and isolation has been examined (Ager et al. 2020, Markel et al. 2007). Thus, the level of social contact will be represented by a new idea: "the Swedish Railway network" in our research. To be specific, by comparing the mortality in parishes that has built a railway station when the pandemic started, with parishes that have no station before the pandemic, we could answer the question that whether accessing railway services causally affects the transmis-

sion of disease This analysis applies the difference-in-difference identification method to minimize the endogeneity and extends with an instrumental variable approach inspired by Melander (2020)[3] as well, which captures the impact of social contacts on disease transmission in the form of railway and mortality in Sweden and in an empirical perspective.

This research is trying to answer several questions, like whether increasing social interaction among individuals will accelerate the spreading of infectious disease, whether the effects differ among people with different characteristics, and the potential reasons behind it. This analysis may not only contribute to our understanding of the importance of interactions of people under the circumstance of diseases such as Spanish Flu and COVID but also be helpful in the future.

2 Literature Review

2.1 Global Spreading of the Great Influenza

During the last time of World War I (WWI) in the 20th century, the world experienced a devastating H1N1 viral disease, the so-called 'Spanish Flu'. This deadly pandemic was even worse than armed conflicts during the war as it not only caused more deaths but also depressed the global economy and social harmony. For one year period started to form the spring of 1918, there were nearly 30 million people died, while this number was renewed to more than 50 to 100 million cases

[3]More details in Section 2.4 and Section 3.2.4.

around the world[4] (Johnson & Mueller 2002, Patterson & Pyle 1991). At that time, nearly half a billion people were infected and the death toll accounted for about 1 to 6 % of the world population in 2 years. Compared with the COVID-19 pandemic which also caused half a billion people infected but only 6 million deaths in 2 years, the Spanish Flu was more dangerous at that time and has been considered as one of the deadliest pandemics. Despite the number of death, its transmission speed is more noticeable, as this lethal disease killed such a large amount of population in a very short period.

Although this disease is known as 'Spanish Flu'[5], there is much debate regarding its origination, and the exact time of its emergence remains uncertain. Several potential channels accelerated the transmission of this novel virus, one was the military transportation networks. During WWI, in the spring of 1918, the US increased their deployment of soldiers, regardless of their health conditions, to the battlefield in Europe through sea transport, which transmitted the virus to Europe[6] (Patterson & Pyle 1991). In August, the more deadly second wave started from France and Spain and quickly diffused to entire Europe and nearly every continent of the world (Chandra et al. 2020, Patterson & Pyle 1991).

Another possible channel was the colonial system, by which Europe retained frequent contact with other continents like the Americas, Africa, and Asia. Together

[4] Patterson & Pyle (1991) made their calculation through various sources of yearbooks in different countries and resulted in a total of 24.7-39.3 million deaths during influenza, while Johnson & Mueller (2002) reviewed the number of death and updated to 50-100 million in total.

[5] The name may be coming from that the virus has not been detected before its variation initiated, and Spain is the first Country that announced the appearance of such disease (Taubenberger et al. 2019).

[6] Patterson & Pyle (1991) states that in July 1918, wounded soldiers returned homes from the front also contributed to the spreading of disease that reached Africa, China, India and Australia.

with WWI, the colonial system built bridges between different areas[7](Chandra et al. 2020). Moreover, despite the population movement, the importance of shipping cargo through the colonial network in disease transmission was not neglectable (Chandra et al. 2020). Although some of the colonial channels overlapped with military transport, they also contributed to the global transmission of the 1918 pandemic. After the second wave in the autumn of 1918, the death numbers sharply increased, where around 26 million people died in Asia and more than 30 million people died throughout the world couple of months later (Johnson & Mueller 2002).

2.2 Effects on Public Health

Since the disease was reported in Spain, medical and historical researchers have agreed on that the 1918 pandemic had several waves. Starting from the spring of 1918, the first wave seemed to have little impact on human health and only caused a few death cases. However, the second wave that was reborn in the autumn of 1918 was considered the most serious and deadly period of this pandemic. Compared with normal influenza mortality of which only 0.1 % infected people died, this virus has a more than 10-20 times bigger effect[8] (Karlsson et al. 2014, Morens & Fauci 2007). After 1919, during the late pandemic period, the occurrence of influenza was seldom noticed. However, according to Almond (2006), who studies the long-run effect of influenza, the long-term impact on individuals is not neglectable, especially in their

[7]For instance, the disease was first detected in the Bombay form British India about one month after the virus was reported in Britain, and the virus also reached the Dutch-administered Sumatra from British Malaya. Similarly, it was visible that the French empire and Japanese colonial networks introduced the virus to West Africa and the primary nations in Asia.

[8]Morens & Fauci (2007) propose the main reason for this aggressive mortality rate to be that besides the bronchus, this virus also attacked patients' lungs, which led to a large number of pneumonia deaths.

social, educational and economic performance[9].

Another characteristic that must be noted is that this novel virus has its unique incidence pattern. Different from other infectious diseases, it was observed in many countries and regions that adults aged between 20 to 40 were more likely to be affected by the 1918 influenza. However, there are numerous explanations for this phenomenon. In recent years, Gagnon et al. (2013) has partly explained the reason behind the specific incidence population and extended our understanding of the such disease. In their research, they found a remarkable relationship between the '1918 Influenza' and the '1889 Russia Flu' and pointed out that individuals that were exposed to the '1889 Russia Flu' early are more likely to be affected by 'Spanish Flu' either[10].

Furthermore, except for the certain susceptible population to 1918 influenza, the overall mortality was also differentiated among people with various socio-economic statuses. Tuckel et al. (2006) using Cox regression analysis find in their research in American cities that, though biological factors such as age matter in the disease infection, some social factors like ethnicities play an essential role in contracting virus[11]. In addition, Galletta & Giommoni (2022) pay attention to the 1918 influenza and

[9] Almond (2006) access US Census historical data and found that those most likely to have been exposed in utero to the pandemic would perform poorer than others born in different time spans in many aspects in their later lives such as educational attainment, income and social status.

[10] Gagnon et al. (2013) observed the peak death age of 1918 influenza is 28-year old, which inline with their argument that immune system of an individual may be subverted due to the development of immunological memory resulted from an infection of antigenically dissimilar influenza subtype (the Russi a Flu in 1889-90) in his or her early lives, even for individuals during their infancy periods.

[11] (Tuckel et al. 2006) using Cox regression analysis based on individual data in Hartford, Connecticut, abnormally high mortality and incidence speed was seen on immigrants from southern and eastern Europe who lived in neighbourhoods that did not reflect their ethnic backgrounds, while local people showed a low probability of being infected. Even though the number of immigrants with the same background is not large, they must have maintained contact frequently, which also helped the transmission

inequality of personal income in Italy and illustrate a positive relationship between influenza severity and income inequality[12]. At the same time, medical resources that could be used by the poor is also limited. Thus, the loop arises that the poor become much poorer and are more likely to be affected by infectious disease, as well as healthcare issues persist not only in the short run but also after one century.

In addition, climate differences have also been considered an important determinant of the spread pattern in different regions. Influenza is a seasonal disease, and in the northern hemisphere, flu season usually begins in October and lasts several months until late spring a year after (Chandra et al. 2020). Evidence also showed that the disease was unlikely to be prevalent in tropical or subtropical regions like Australia, while the virus spread quickly in northern areas like America and Europe due to the suitable weather condition (Chandra et al. 2020). In Canada, America and European countries including Sweden, and Russia, where the latitude is high, the prevalence period of 1918 influenza was quite consistent with the period of normal flu. This also verifies the importance of seasonality of the spreading.

To counter this deadly disease, governments around the world have carried out many approaches aimed at reducing mortality and slowing down the transmission as well. However, in the 20th century, to deal with this sudden global shock, medical and biological treatment was unrealistic. Therefore, they pay more attention to Non-Pharmaceutical interventions (NPI) such as social distancing, public gathering cancellations, school closure, isolation periods and so on. Markel et al. (2007) study the association between NPI and disease diffusion and notice a strong beneficial effect

[12]Galletta & Giommoni (2022) indicate that in Italy, the pandemic deteriorated the income equality level, which was mainly because of the reduction in the share of income held by poor people rather than top earners. And this effect seems to be a long-term influence since those seriously affected localities by influenza still experiencing a high level of income inequality.

of such interventions on pandemic consequences[13].

2.3 Social Contact in Disease Transmission

2.3.1 Importance of Connections

Overseas transportation brought the virus from one continent to another, while inland connectivity dominated the disease spreading in different locations. As we discussed before, human beings were the primary carriers of viruses, and evidence showed that cutting down the physical contact between people indeed helped prevent individuals from being infected. Under the current COVID-19 pandemic, loads of literature have verified the connections between COVID-19 infection spread and social contact and activities (Brodeur et al. 2021). Before the effective anti-virus vaccine has been developed and put into use, self-quarantine and reduction in public activities played an essential role. Analogously, during the time of 1918 influenza pandemic, the effect of social contact is also non-negligible.

Chandra et al. (2020) argue that in the 1918 influenza pandemic, the deadly infection emerged primarily due to interactions of people[14]. Also, Adda (2016) studies the association between economic activities and infectious diseases transmission states that interpersonal contact matters in the prevalence of viruses, and concludes that for those places with above average mortality, NPIs that aim at reducing mortality

[13] Markel et al. (2007) carried research among US cities using historical data, aimed at determining the effectiveness of NPI in the circumstance of disease spreading. Evidence shows that approaches such as school closure and bans of public gatherings were usually used together in many cities in the US, and they did lead to lower overall mortality and lower peak mortality rates.

[14] Chandra et al. (2020) review and demonstrate the global perspective regarding the transmission of 1918 influenza. To be specific, in the early stage of the pandemic which was during the WWI period, the huge amount of human interaction played an essential role in distributing the virus.

such as school closure and transportation shutdown are cost efficient in places with abnormal higher mortality[15]. Inspired by Adda (2016), this research will more precisely focus on the role of railway networks as a form of social interaction in disease transmission.

2.3.2 Evidence in Railway Transportation

The transmission of the disease was largely affected by the emergence of transportation decades. The railway is one of the most common transportation methods, not only at present but also 100 years ago. In the 20th century, industrialization accelerated the construction of railways in many places of the world. Evidence showed that in India and America, the expansion of railways reduced trading costs, boosted the economy and enhanced individuals' satisfaction (Donaldson 2018, Donaldson & Hornbeck 2016). Although trains during that period did convenience individuals and capture excess profit through trading, they also enhanced the connectivity among locations and extended human contacts thus facilitating the transmission of diseases. And in the pandemic of 1918 influenza, plenty of evidence has been seen around the world about the effect of railways on disease spreading.

In Nigeria, more than 500,000 out of 18 million people died in 6-month during the 1918 influenza period. Ohadike (1991) studies the transmission of this virus throughout Nigeria. In his research, the inland diffusion of the virus was taken advantage of the normal transportation methods such as highways, roads and railway

[15] Adda (2016) using high-frequency historical data from France together with OLS and IV approaches, taking three types of diseases (flu-likely disease, chickenpox and acute diarrhoea) into the analysis. In her research, school closure and transportation networks such as railway expansions, which were instrumented by the lagged weather episodes were taken as representations of social contact and were investigated regarding their influences on virus spread respectively.

lines. Evidence showed that in Nigeria, at the beginning of influenza, incidence cases were found in locations that were closer to the existing railways during the pandemic earlier than other places that were less connected[16](Olapoju 2020). Besides, different infection numbers were seen among places, where the central cities which were highly urbanized were more likely to be a stroke, and this could be mainly due to the more crowded population density compared with villages and towns Ohadike (1991). Similarly, Reyes et al. (2018) who study the spatial diffusion of the 1918 influenza in British India also show that human mobility such as travel via railroads was able to predict the observed transmission of disease across the country[17].

However, alongside railway lines, the influenza virus might also spread through other possible channels, such as inland routes and riverways. Therefore, studies in South Africa would be more convincing. Hogbin (1985) done similar research regarding the relationship between railway and disease in South Africa and show a similar pattern that the virus travelled followed the railway lines, and major city centres seemed to suffer more 1918 influenza[18]. To be specific, returning soldiers carried by the trains that departed from Cape Town dispersed the virus to each station and remote municipality at each railway stops (De Kadt et al. 2020).

[16]For instance, after the vessel called *S.S.Bida* that carried infected passengers stopped at Lagos, the disease travelled at a shocking speed into the hinterland through railway networks.

[17]Reyes et al. (2018) use historical death data in India find that both long-distance travel(railroads) and local transportation display a strong prediction of spatial diffusion of the 1918 influenza.

[18]The results are reliable to a large extent as in South Africa where the scarcity of waterways made long-distance railway transport the dominant approach during the pandemic time in 1918.

2.4 In the Context of Sweden

At the beginning of the 20th century, Sweden was characterised by industrialization. Although over half of the total population still made their living through agriculture, individuals in urban areas were largely employed in manufacturing factories, thereby boosting the forming of an open economy. However, devastating influenza hit Sweden in 1918 and caused nearly 38,000 deaths accounting for 1 % of its population, which was considered the severest shock in Swedish history. The figure B.1 is the mortality rate between 1918 and 1920 in each county of Sweden. It is clear to see that the mortality varied widely across the country, where central and northern cities suffered the most while the southern areas were less stroked. This surprising finding has led to numerous discussions. Some argued that there were more young populations in these counties, and others also noted that the difference in fatality among counties is partly due to the remoteness, that is individuals in these areas were less likely exposed to the previous viruses, thereby observing a lower mortality pattern[19] (Karlsson et al. 2014). A similar picture was acquired by Rogers (1968) those who observe that during the 1918 influenza pandemic, though the overall mortality rate in America is quite low (6.5 deaths per 1,000 individuals), over 10 times larger effects could be seen in some centre cities like Boston and New York.

Moreover, the suitability of studying infectious disease and social contact in Sweden has been verified. Despite its similarities to COVID-19 regarding its public effects, the more important factor is that it is much more valid to be considered as a natural experiment. First, influenza only lasts several months starting in late

[19] Also consistent with the opinion from Gagnon et al. (2013) that the previous Russia Flu matters in a medical perspective.

1918 and it is an unexpected event, which means there was a lower probability of behaviour variations among individuals. Besides, even during the world war period, there might be disturbances in the mortality pattern, compared with nowadays where population flow and transmission are fast in speed and various channels, a disease from one century ago could rule out some endogenous problems.

In addition, the railway networks also played a fundamental part in Sweden. Before the 1918 influenza, the construction of railway lines has already facilitated the social movement of inhabitants across cities. In the late 19th century, the traditional waterways were gradually replaced by high-speed railway lines, which reduced public travel costs and thereby increased passenger volume. Swedish started planning their railway construction in the 1840s, and the first plan was proposed by Count Adolf von Rosen. The first short railway line was built in 1856, and afterwards, major cities in the south part were nearly all connected (Heckscher et al. 1954). Colonel Nils Ericson carried a second draft regarding railway plans that aimed at connecting major municipalities. Although their plans failed ultimately, right before *World War I* started, the railway has already extended throughout nearly the entire Sweden.

Melander (2020) studies the social movement and railway networks in Sweden and finds that the expansion of the Swedish railway not only spread the idea across municipalities but also has a great impact on the emergence and diffusion of social activities[20]. Besides, Karlsson et al. (2022) study the determinants of the excess mortality of 1918 influenza in Sweden. They used data from 2,500 municipalities in Sweden during the pandemic period, and evidence suggested that despite the population density, the ability to access railway services was also positively correlated

[20]He noted that there was an adverse relationship between the minimum distance to railways of a parish and the emergence of movements, suggesting that railway was a key factor that improved the interaction of people.

with excess mortality.[21]

The literature discussed above has investigated many aspects of the transmission of 1918 influenza through railway infrastructures in many places, which enriched our understanding of the social interaction in disease diffusion and provided the basics for this analysis.

[21]Inspired by Melander (2020) and Karlsson et al. (2022), this analysis will focus on studying the causal impact of railway accessing on mortality on different groups of individuals and look insight to its effect on total population using instrument variable. See more detail in Section 3.2.4.

3 Data and Methodology

3.1 Data

The analysis relies on several sources of data.[22] The main data set used in this paper is a panel which covering 2,441 Swedish parishes over eight years from 1914 to 1921, for a total of 1,015,872 observations.

3.1.1 Mortality in Sweden

The individual dataset contains detailed information on inhabitants who died between years 1914 and 1921. The original information includes the exact date of the individual's birth and death, as well as their personal details such as gender, individual socio-economic status, and in which parish and county they died. Therefore, it is able to determine the number of deaths for certain population groups of each parish each week.

We can observe from Table 1[23] the difference between genders is small, while the gaps among different ages and individual socio-economic groups are significant. Also, the higher maximum death toll in urban may be due to more interactions among people, since cities have a larger population and the emergence of railway

[22]Thanks for data provided by prof. Dr. Daniel Kühnle and Prof. Dr. Melander, which includes historical Swedish individual and parish level data sets.

[23]The table displays the overall death tolls from 1914 to 1921 in entire Sweden for different groups of people, and here I made a comparison for urban and rural areas respectively. The SES1 - SES5 in the first column represent the economic status of the individual, from the lowest (SES1) to the highest (SES5). This classification comes from the historical Swedish individual dataset provided by Prof. Dr. Daniel Kühnle.

Table 1: Mortality in Sweden urban and rural parishes: 1914-1921

VARIABLES	(1) Rural Obs.	(2) mean	(3) sd	(4) min	(5) max	(6) Urban Obs.	(7) mean	(8) sd	(9) min	(10) max
Mortality										
Total	975,104	0.486	0.948	0	37	44,096	4.093	12.20	0	312
Female	975,104	0.243	0.584	0	17	44,096	2.106	6.333	0	171
Male	975,104	0.243	0.595	0	26	44,096	1.987	6.045	0	166
0-20	975,104	0.112	0.400	0	17	44,096	0.964	2.873	0	68
20-40	975,104	0.0646	0.305	0	22	44,096	0.756	3.383	0	163
40-60	975,104	0.0610	0.261	0	9	44,096	0.681	2.368	0	52
>60	975,104	0.249	0.556	0	11	44,096	1.692	4.652	0	83
SES1	975,104	0.175	0.474	0	19	44,096	0.992	2.840	0	97
SES2	975,104	0.117	0.377	0	10	44,096	0.804	2.292	0	62
SES3	975,104	0.0755	0.297	0	10	44,096	0.651	2.007	0	50
SES4	975,104	0.0574	0.256	0	8	44,096	0.659	2.205	0	54
SES5	975,104	0.0618	0.266	0	10	44,096	0.986	3.670	0	101

networks also started across major cities. The following figures could show more intuitively.

Figure B.2(a) shows the total number of deaths in Sweden from 1914 to 1921. It clearly displays the seasonality as the death toll reaches a peak in winter times. And it is also observable that at the peak of 1918 influenza death toll was over 4,000 individuals per week which were more than 2 times the number in other years. Figure B.2(b) shows that the overall mortality in rural areas is more than 2 times higher, which may be because of the large number of rural parishes and thereby a larger rural population. However, from the Table 1, the maximum death toll during the pandemic is much higher in urban areas. This picture might consistent with our hypothesis that urban areas have more human interactions due to the appearance of

Table 2: Mortality in Sweden high and low SES parishes: 1914-1921

	(1) low SES	(2)	(3)	(4)	(5)	(6) high SES	(7)	(8)	(9)	(10)
VARIABLES	Obs.	mean	sd	min	max	Obs.	mean	sd	min	max
Mortality										
Total	766,288	0.473	0.923	0	33	252,912	1.155	5.353	0	312
Female	766,288	0.238	0.576	0	21	252,912	0.583	2.791	0	171
Male	766,288	0.235	0.576	0	21	252,912	0.572	2.674	0	166
0-20	766,288	0.105	0.380	0	12	252,912	0.280	1.311	0	68
20-40	766,288	0.0610	0.292	0	16	252,912	0.196	1.470	0	163
40-60	766,288	0.0581	0.254	0	7	252,912	0.178	1.048	0	52
> 60	766,288	0.249	0.557	0	11	252,912	0.501	2.080	0	83
SES1	766,288	0.177	0.473	0	14	252,912	0.310	1.300	0	97
SES2	766,288	0.113	0.370	0	10	252,912	0.247	1.056	0	62
SES3	766,288	0.0713	0.287	0	9	252,912	0.189	0.915	0	50
SES4	766,288	0.0537	0.247	0	8	252,912	0.174	0.982	0	54
SES5	766,288	0.0576	0.255	0	7	252,912	0.236	1.595	0	101

a railway and high population density.

Besides, Figures B.2(c) and (d) are the mortality numbers for different ages and socioeconomic status people respectively. It could be observed that during normal times, elders maintain higher average mortality of around 800 deaths per week, while during the 1918 influenza pandemic, the death toll for young adults aged 20-40 increased sharply and reached a peak of over 1,500 cases a week. Besides, individuals with low socioeconomic status[24] shows persistent higher mortality than others, and the gap enlarged to about 3 times higher than the death tolls of high-status individuals. These findings in Sweden are quite consistent with the literature regarding the specific susceptible population. The same pattern could be found in Table 2 where

[24]The socio-economics status are ranked into 5 levels, with SES1 as the lowest while SES5 represents the highest.

a comparison between high and low SES parishes is showed[25]. The overall average mortality is generally higher in parishes that were richer.

3.1.2 Swedish Railway Networks

The main treatment variable in this analysis is the emergence of railway stations. Thus, I use the historical railway and parish information to construct a panel indicating whether a parish has a station in 1908, or 1918, and coded them to be used in Diff-in-Diff identification[26].

For an additional part of the analysis, I use the data of distance between parishes and the initial railway plans[27] for an instrument variable design. The data used is collected and processed by Melander (2020). The coloured lines in the Figure B.3 display the different initial plans.

[25]High SES and low SES parishes are determined by the pre-epidemic average government revenue amount documented in the Swedish 1910 Census. To be specific, those parishes with above-average government revenue are classified as the so-called high SES parishes, and the rest parishes with below-average revenue are low SES parishes subsequently.

[26]To use the diff-in-diff identification and an event-study in this analysis, I generated a continuous time variable in week frequency from 1914 week 1 to 1921 week 52, and I determine the week 28 in the year 1918 as the beginning of the second wave of influenza which is the exogenous event of our interest. In addition, based on the historical latitude and longitude data of each railway station and parish, I calculated the distance between each of them and thereby identified the emergence of the railway station by checking whether the parish was near the railway station. Therefore, I can determine which parish could access to the railway when the pandemic began, thus determining the treatment and control groups. More discuss in Section 3.2.1

[27]The initial plans are based on sources that discussed in Section 2.4, and more discussion is in Section 3.2.4 below.

3.1.3 Other data

I also include pre-pandemic parish level characteristics as controls, including local government and demographic factors. In the regressions, I controlled factors such as population density, the share of poor houses, the share of people aged 20-40, parish taxable income and so on. More details are included in Table A.1. These data are from the 1910 Sweden Census and are constant, thereby I interacted with them with a continuous time indicator to include them in the panel analysis.

3.2 Methodology

In order to find the impact of railways on disease mortality, the target effects are mortality changes over time between parishes that has a railway station and those has no station at all. Thus, it is suitable for the application of difference-in-difference identification. By contrasting different locations at the parish level, we are able to exclude overall trends in mortality. This section will discuss the treatment variable, possible outcome variables as well as potential threats and limitations.

3.2.1 Define the treatment and outcome variables

The treatment variable is defined as a dummy variable, denoting D_i^{1918} which is an interaction term of two dummy variables $Rail_i^{1918}$ (whether a parish has a station in 1918) and $Post$ (whether the pandemic started)[28]. The basic outcome variable is

[28]Here $Rail_i^{1918}$ equals 1 if the parish is near at least one railway station, and equals 0 otherwise. Remember I generated a continuous weekly time variable, and we assume the second wave of influenza started at the beginning of September (the 28th week in 1918 in our dataset), therefore it is able to define $Post$ that equals 1 if the time is after 1918 week 28 and equals 0 otherwise.

the number of death in each parish in week frequency.

Thus, interacting $Rail_i^{1918}$ and $Post$ enables us to determine the difference in weekly mortality between parishes near the stations and parishes that were not after influenza began. However, to determine a more intuitive effect, it may be replaced by excess mortality. The excess mortality will be expressed as the death toll that exceeds the average weekly death numbers in each parish from 1914 to 1917.

3.2.2 Identification and Econometric approach

To carry out this research, we may want to capture the average treatment effect on the treated:

$$ATET = E[Y_i^1 - Y_i^0 | D_i^{1918} = 1],$$

where Y_i^1 is the mortality in parish i if it has a railway station before 1918, while Y_i^0 is the counterfactual. D_i^{1918} is the diff-in-diff term. For simplicity, the equation could be expressed by:

$$Y_i = \alpha + \beta \times D_i^{1918} + \epsilon_i, \tag{1}$$

where β is the effect that we are interested in. The analysis relies on the exogenous assumption that the shock of influenza is random in each parish. Although it is not able to test directly, we can take several indirect tests to check its availability.

Nevertheless, the regression 1 above cannot estimate consistently as the construction of the railway may be affected by other factors. Therefore, we can narrow the difference among parishes by conditioning on $Rail_i^{1908}$ which is a binary variable controlling for the unobserved factors in parishes that have a station in 1908-1918, and combining it with a two-way fixed effect model. To be specific, $Rail_i^{1908}$ is similar

to $Rail_i^{1918}$, which equals 1 if a parish has a station in 1908 and 0 otherwise. By taking $Rail_i^{1908}$ into account, we thereby identify the difference in mortality between parishes that has a station in 1918 and other parishes that have no stations at all more flexibly as this variable makes sure parishes that have a railway station in 1918 (the treatment group) and parishes have no railway stations in 1918 (the control group) do not differ significantly in unobserved aspects prior to the pandemic began, and also allows parishes having station earlier to have flexible impact over time. So, the baseline equation can be expressed as:

$$Y_{it} = \lambda_i + \gamma_t + \beta \times D_i^{1918} + \delta_i \times Rail_i^{1908} \times I[Year] + Z'_{it} \times \Phi + \epsilon_{it}, \tag{2}$$

where Y_{it} could be the mortality or excess mortality for each parish in each week and β is our interesting effect. λ_i represents the parish fixed effect that contains time-invariant factors in parishes while γ_t represents the year fixed effect that controls the time-variant factors. Z'_{it} contains a complete set of controls. And I interacted $Rail_i^{1908}$ with a continuous year indicator $I[year]$ to create a panel form.

Furthermore, by assuming the influenza pandemic hit Sweden at a different time, say 1916, we can verify the exogenous assumption. If we re-run the regression to assume influenza began in 1916, and get a statistically insignificant coefficient, we can conclude that the common trend assumption is satisfied, otherwise, it is violated[29]. In addition, if the assumption is violated, the time auto-correlation problem may arise which may bias downward the standard error and increase the probability of False positives. Besides, I remove those parishes that already have stations in 1908 to create a sub-sample before rerunning the equation to check the robustness. This approach aims at providing more comparable treatment and control groups. If it

[29] A form of placebo test.

displays a similar impact as before, we thereby verify the results.

Moreover, I extend this formula to see whether there are different effects of railway networks on different groups of people such as different ages and SES individuals. I replace the outcome variable with the mortality number for each group of people to find distinct effects. And this could be easily extended to study the inequality between parishes like the difference between urban and rural, between high SES and low SES areas[30].

3.2.3 An Event Study Design

I also propose an event-study design with the same identification as before. This approach extends the logic as before, extends the above equation and compares the impact of influenza development between parishes that has a railway station constructed before the pandemic and those parishes that have no railway station before the pandemic[31]. The estimation equation could be expressed as:

$$Y_{it} = \lambda_i + \gamma_t + \sum_{j=-15}^{20} \alpha_j \times I[t-k=j] + \sum_{j=-15}^{20} \beta_j \times D_i^{1918} \times I[t-k=j] + Z'_{it} \times \Phi + \epsilon_{it} \quad (3)$$

[32]

[30] Urban and rural are determined by the historical Swedish data set, while high SES and low SES parishes are determined by the pre-epidemic average government revenue amount documented in the Swedish 1910 Census. To be specific, those parishes with above-average government revenue are classified as the so-called high SES parishes, and the rest parishes with below-average revenue are low SES parishes subsequently. I make a such split of sample in order to see different impacts among different individuals and parishes.

[31] Similarly, this approach should also satisfy the common trend assumption which could be checked graphically. See Figure B.4.1 and Figure B.4.2.

[32] Y_{it} is the weekly mortality or excess mortality in each parish during each week. The coefficient β_j captures the treatment effect over time. The parish and year fixed effect is controlled and represented by λ_i and γ_i respectively. $I[t-k=j]$ is an indicator variable which measures the number of weeks from the pandemic started (k), and our estimation takes 15 weeks before the pandemic started as a comparison. Z_{it} contains all other controls. Besides, the error is clustered

However, these baseline regressions may be exposed to problems such as endogeneity that biased the estimation. Firstly, there could be omitted variable problems, some unobserved factors may also affect the estimation. The endogeneity challenge this analysis face is that maybe the mortality of this infectious disease is correlated with the probability of social contacts without any causal relationship between social interaction and disease mortality. For example, there may be individuals of whom with poor baseline health who are more likely to interact with people and are also more likely to die. Besides, some parish leaders might want the municipal to be better connected with others but also tried to lower the mortality rate to make the local area more 'harmonious'. Besides, the inaccuracy of historical data collected might also be a concern. If there exist measurement errors in the independent variable, the estimation will face attenuation bias which underestimates the absolute value of the estimation. Similarly, the outcome variable in this analysis is also likely to be exposed to measurement error. For instance, if the number of death or excess mortality is not precisely counted, the estimation will be biased either. Overall, it may be difficult for us to rule out all potential unobserved factors and produce actual causal effects.

3.2.4 Instrumental Variable Design

The purpose of using IV is that there were many other factors we cannot observe that could affect the outcome and treatment effect are contained in the error term when using the original Diff-in-Diff equation. To deal with this endogeneity, I propose to use the distance of each parish to the initial railway plans as an instrument. The idea is inspired by Melander (2020) who studies the association between railway

at the parish level (ϵ_{it}).

networks and social movement in Sweden[33]. Despite the two initial plans discussed in Section 2.4, he also constructed straight lines based on the nodal destinations in the former proposals[34]. Overall, the instrument to be used is the minimum value of the distance from each parish to each of the three planned lines:

$$PlanDist_i = min\{Dist.\ von\ Rosen_i; Dist.\ Ericson_i; Dist.\ Straight\ Lines_i\}^{35}$$

Note that the distance to railway plans is time-invariant thereby I transform it into natural log form and interact it with a year fixed effect $I[Year = d]$, where we take the year 1914 as the base year. Formally, the first stage equation can be expressed as:

$$D_i^{1918} = \lambda_i + \gamma_t + \sum_d \delta_d \times ln(PlanDist_i) \times I[Year = d] + Z'_{it} \times \Phi + \nu_{it} \qquad (4)$$

First, the instrument should be relevant. The distance to initial railway plans should be able to predict the emergence of railway stations in parishes[36]. Second, the instrument should not have any correlation with the error term ϵ_{it} in the equation 2 to satisfy the exclusion restriction. To be specific, the proximity to railway plans should affect the mortality only through individuals accessing railway services. This

[33] Melander (2020) uses the minimum distance to railway networks for each parish as the treatment and uses the expected distance of parishes to the initial railway plans as the instrument. By using the two-stage least squares framework, Melander (2020) captured the local average treatment effect of accessing railways to social movement outcomes.

[34] Melander (2020) created straight lines based on several destinations. The destinations are: Gothenburg, Malmö, Östersund, Korsvinger and Stockholm. And in this analysis, the distance between each parish and these straight lines is calculated.

[35] Melander (2020) has verified the validity of this instrument while it needs to be retested due to different research period and different interested effects.

[36] This condition seems to be satisfied as construction of national infrastructures always has plans, and at localities where the railway plans are made for, the probability of construction of a railway station is higher.

condition is satisfied intuitively since the railway plans were basically formed and proposed to improve the nation's infrastructures, which has no reason to be connected with increased mortality. We can run a regression with D_i^{1918} replaced by the instrument $ln(PlanDist_i) \times I[Year = d]$ to check the robustness. Besides, as the pandemic that happened in 1918 was as good as random, the instrument is not likely to be dependent on the treatment or outcome. Finally, the monotonicity also seems to be satisfied since it is more likely for the government to build a railway station in parishes that were near the already established railway plans, rather than in remote parishes.

4 Main Results

I start by exploring the perceptual causal relationship between the emergence of railway stations and the influenza mortality level through Diff-in-Diff identification. Then I move on to observe a more intuitive impact of social interaction on excess mortality through event-study plots. Finally, I extend the above logic by using instrument variable design to improve the robustness.

4.1 Baseline Evidence

I begin by estimating the equation 2, in which the total weekly death number for each different group of individuals is regressed on the emergence of railway stations[37]. Table A.2.1 is the estimation among all parishes. I also make a comparison of railway

[37]I use Difference-in-Difference framework in this panel analysis to eliminate some of the endogenous problems and to yield the *ATET*.

impacts for rural, urban and high/low SES parishes. Table A.2.2 and Table A.2.3 display the impact of social contact on disease spreading in rural and urban areas respectively, and Table A.2.4 and Table A.2.5 represent the results from a parish with high and low socio-economic characteristics. In these tables, each column represents a category of mortality number as dependent variables[38].

Overall, the railway has a statistically significant positive effect on influenza mortality. From Table A.2.1 , through the two-way fixed effects model, the emergence of railway stations in 1918 caused the weekly death toll to increase an average of 0.157 cases per week more than in parishes that are less likely to access railways. It is the same as our expectation and literature as well that the effect on the 20-40 age group is much larger than other age groups which are consistent with Gagnon et al. (2013), and the same larger impact is also seen among low SES individuals compared with higher SES groups which makes sense as Tuckel et al. (2006) has discussed the possible transmission path from oversea immigrants[39]. The differences between rural and urban areas are noticeable in Table A.2.2 and A.2.3. The overall impact of railway in urban parishes from the coefficient estimated is more than 5 times higher than that in rural parishes, which is probably due to the higher population density, higher than average family size and that the emergence of the railway was earlier in developed regions, thereby facilitated the mobility speed of people. The appearance of the railway increased the total weekly death toll to 0.556 cases in urban but only 0.072 in non-urban regions. By comparing results from the Table A.2.4 and A.2.5 between high and low economic characteristic parishes, we find a similar pattern

[38]Each table contains three panels of regression output, of which I show different results by clustering standard error, controlling parish and year fixed effect. The covariates are fully controlled. See more in Table A.1

[39]As Tuckel et al. (2006) discussed, southern and eastern European immigrants went into the US shows a higher infectious rate, and they usually has lower social status

that the mortality in parishes with greater development such as higher government expenditure and taxable income were more likely to be inflated by the railway. The effect on total weekly mortality is 0.29 in rich parishes but only 0.078 in others.

For the robustness check, I report the result in Table A.2.6. I initially run the placebo test by assuming influenza started at the beginning of the year 1916. The results in Panel A show that in 1916, the hypothetical influenza was not likely to induce any impact on the mortality number, which is our expectation. Besides, Panel B I displays the results of baseline DID regression by removing parishes that have stations earlier[40]. The similar significant coefficients indicate that the original output is robust. This evidence shows that overall, there is a significant causal effect between railway emergence and disease mortality.

4.2 Event-Study Result

As an additional empirical analysis, the event-study design enables us to look more precisely into the overtime effects of the increase in social contact on disease transmission. In the analysis, I replace the dependent variable with the excess mortality number in each week. I report two sets of common trend plots for urban and rural parishes as shown in Figure B.4.1 and Figure B.4.2 respectively. These plots display the trend line of each subgroup and show the treatment effects over time.

The results of this additional exercise are in line with the baseline analysis, where railways have a larger impact on urban areas. Figure B.4.1(a) shows that 10 weeks after the pandemic started, the death toll reached the peak from 0 to nearly 9 excess

[40]I use a sub-sample that only include parishes has no station before 1908 and regressed it using the equation 2, which made the treatment and control group more comparable.

death cases per week at a shocking speed, while Figure B.4.2(a) indicates a smaller peak number among rural areas, around 0.9 weekly excess death. Regardless of the differences, the results are economically intuitive. Since influenza started, there appeared a lag effect for nearly 10 weeks before the excess mortality number shot up. It could be interpreted as with the constructed railway and railway stations in 1918, people travelled as usual through railway transportation and the influenza shock was as good as random which did not put restrictions on social contacts. And the inverted line may be due to the introduction of non-pharmaceutical interventions like social interaction restrictions and public area closures, which significantly decreased the probability of interaction among individuals. Therefore, as it took some time, maybe several weeks, from being infected to death, the trend line plots then make sense. And the insignificant zero effects prior to the treatment indicate that the observed effects of the railway on disease mortality are not driven by differential pre-trends.

In addition, these graphs for each sub-outcome group also show that the impact magnitude for different genders, different age groups and different socio-economic status populations are consistent as we displayed in the previous regressions and literature in both urban and rural regions. In summary, the event analysis suggests a remarkable effect of railway stations' emergence on the excess mortality triggered by influenza in 1918.

4.3 Instrument Variable Results

However, despite the above Diff-in-Diff identification could rule out some endogenous issues, there still exist many unobserved factors that affected the construction

of the railway as well as the changes in the death toll, which make the explanation tough. I perform an instrument variable design combined with the baseline Diff-in-Diff identification to get a more reliable causal relationship between social contact and disease spreading using two-stage least squares.

I display the first-stage regression corresponding to equation 3 in Table A.3.1. Here I make a comparison by changing the standard in identifying whether a railway station is near enough to a parish[41]. In the first column for each dependent variable, I only report the parish and year fixed effect combined with only demographic controls in addition to the 7 instruments[42]. As expected, the coefficient of instruments is highly statistically significant. In the other two columns for each dependent variable, I add different controls into the model to produce a more stable result. In the second column, parish characteristics are added. And in the last column, I add the land area interacts with year fixed effect for each parish.

The results in Table A.3.1 indicate that the instruments satisfy the condition of relevance. And the validity condition that the instrument only impacts the death numbers through the emergence of railway stations also seems to be satisfied. I regress the instruments directly on total weekly mortality numbers with a complete set of controls independent of the emergence of the railway station. The statistically insignificant coefficients suggest that the initial planning of railway constructions is not likely systematically influence the death rate in regions. In addition, I make several tests including unidentifiable tests, weak instrument tests and over-identification tests on the instrument. I report the test statistics in Table A.3.1 as well and they

[41]To be specific, I replace the original 1 km threshold with 5 km and 10 km respectively.

[42]7 instruments including $ln(PlanDist_i) \times I[Year = 1915], ..., ln(PlanDist_i) \times I[Year = 1921]$ represent the constant distance value multiplied by year fixed effect correspondingly, and leaves the year 1914 as the baseline year.

all suggest that by fully controlling for the parish and demographic characteristics, the minimum distance to initial railway plans as instruments are appropriate.

Knowing that the proposed instruments are valid and exhibit a strong first stage, I now proceed to compare the main DID results and instrument variable results. I intend to focus mainly on total mortality in each parish and report only the main treatment variable in Table A.3.2 where the first three columns show OLS results and others for IV, and controls are input in the same succession as for Table A.3.1. Panel A to C is specified by different measurement thresholds of the distance between the parish and station. Beginning with OLS results, the positive coefficients are supportive of the positive relationship between railway access and mortality number. Statistically significant coefficient throughout all specifications suggests that parishes that are able to increase their individual social contact probabilities are more likely to cause more deaths during the 1918 influenza period.

For reasons that were discussed before, these estimates are subject to biases thereby I now turn to the IV results in the last three columns. The coefficients remain positive but not in strong statistical significance. Take column 6 as an example where covariates are richly controlled. Among three Panels, when choosing 5km and 10 km as thresholds in determining parishes with railway stations, the estimates show a 10% significance level smaller impact of railway emergence on disease mortality than OLS results, compared with choosing 1km as a threshold. It makes sense that maybe 1km is a too strict rule to justify whether a parish has a station nearby. Overall, the IV estimates have produced results similar to previous ones but with lower statistical significance.

Throughout the results reported, a feature is that for those where the railway has

a significant effect on mortality, the coefficients are always smaller than OLS results. This is consistent with expectations for several reasons. First, IV could rule out some endogenous factors such as unobserved factors that may increase the mortality rate that biases the OLS results. In particular, since railways constructions are aimed at connecting major municipalities where the population base is large and has generally high mortality than in hinterlands, such consideration would generate a spurious positive relationship between railway access and mortality number, thereby leading to a biased larger OLS estimation. By making use of the fact that some parishes that were near the initially designed railway network plans were better connected, my instrument avoids such endogeneity.

Secondly, the instrument variable estimation captures the local average treatment effect of "compilers". It suggests that the estimated effect of railway emergence on parish mortality obtained from IV estimation comes from those parishes which were near enough to the railway stations only if they were proximate to the initial railway plans networks, while would not have a nearby railway station otherwise. In reality, there may have some exceptions such as some parishes might have constructed railway stations "by accident" that also contribute to the OLS results. For instance, stations might be constructed near to parishes that are not very close to planned routes. In addition, the choice of threshold in measuring whether the station appears could also have some deviation towards actual effects.

5 Conclusion and Remarks

In this paper, I document the impact of increasing social interactions on infectious disease spreading. Specifically, this analysis base on Swedish history uses railway station emergence and mortality in 1918 influenza to represent the treatment and outcome variables and estimate the causal effects. Through Diff-in-Diff identification, combined with an event study, I show that accessing railways indeed facilitated influenza spreading and inflated the death tolls. Furthermore, I provide evidence that different groups of individuals display distinct mortality patterns in the 1918 influenza. Additionally, I show a more robust estimation through the instrument variable approach, which is in line with the baseline evidence that enhancing connectivity of localities, and increasing people contacts are able to predict the infectious rate during the pandemic circumstance.

Although there are limitations to this analysis based on historical background, this research is distinct in its choice of target, proxy and interesting effects and sheds light on the essential role of social contact in disease transmission as well. I thereby contribute to our understanding of epidemic disease and how individual interaction accelerates its diffusion.

Bibliography

Adda, J. (2016), 'Economic activity and the spread of viral diseases: Evidence from high frequency data', *The Quarterly Journal of Economics* **131**(2), 891–941.

Ager, P., Eriksson, K., Karger, E., Nencka, P. & Thomasson, M. A. (2020), 'School closures during the 1918 flu pandemic', *The Review of Economics and Statistics* pp. 1–28.

Almond, D. (2006), 'Is the 1918 influenza pandemic over? long-term effects of in utero influenza exposure in the post-1940 us population', *Journal of Political Economy* **114**(4), 672–712.

Barry, J. M. (2004), 'The site of origin of the 1918 influenza pandemic and its public health implications', *Journal of Translational medicine* **2**(1), 1–4.

Brodeur, A., Gray, D., Islam, A. & Bhuiyan, S. (2021), 'A literature review of the economics of covid-19', *Journal of Economic Surveys* **35**(4), 1007–1044.

Chandra, S., Christensen, J. & Likhtman, S. (2020), 'Connectivity and seasonality: The 1918 influenza and covid-19 pandemics in global perspective', *Journal of Global History* **15**(3), 408–420.

De Kadt, D., Fourie, J., Greyling, J., Murard, E., Norling, J. et al. (2020), *The causes and consequences of the 1918 influenza in South Africa*, Department of Economics, University of Stellenbosch.

Donaldson, D. (2018), 'Railroads of the raj: Estimating the impact of transportation infrastructure', *American Economic Review* **108**(4-5), 899–934.

Donaldson, D. & Hornbeck, R. (2016), 'Railroads and american economic growth: A "market access" approach', *The Quarterly Journal of Economics* **131**(2), 799–858.

Gagnon, A., Miller, M. S., Hallman, S. A., Bourbeau, R., Herring, D. A., Earn, D. J. & Madrenas, J. (2013), 'Age-specific mortality during the 1918 influenza pandemic: unravelling the mystery of high young adult mortality', *PloS one* **8**(8), e69586.

Galletta, S. & Giommoni, T. (2022), 'The effect of the 1918 influenza pandemic on income inequality: Evidence from italy', *Review of Economics and Statistics* **104**(1), 187–203.

Heckscher, E. F., Heckscher, G. et al. (1954), *An economic history of Sweden*, number 95, Harvard University Press.

Hogbin, V. (1985), 'Railways, disease and health in south africa', *Social Science & Medicine* **20**(9), 933–938.

Johnson, N. P. & Mueller, J. (2002), 'Updating the accounts: global mortality of the 1918-1920" spanish" influenza pandemic', *Bulletin of the History of Medicine* pp. 105–115.

Karlsson, M., Kühnle, D. & Prodromidis, N. (2022), The 1918–1919 influenza pandemic in economic history, *in* 'Oxford Research Encyclopedia of Economics and Finance'.

Karlsson, M., Nilsson, T. & Pichler, S. (2014), 'The impact of the 1918 spanish flu epidemic on economic performance in sweden: An investigation into the consequences of an extraordinary mortality shock', *Journal of Health Economics* **36**, 1–19.

Markel, H., Lipman, H. B., Navarro, J. A., Sloan, A., Michalsen, J. R., Stern, A. M. & Cetron, M. S. (2007), 'Nonpharmaceutical interventions implemented by us cities during the 1918-1919 influenza pandemic', *Jama* **298**(6), 644–654.

Melander, E. (2020), *Transportation technology, individual mobility and social mobilisation*, University of Warwick, Centre for Competitive Advantage in the Global Economy, Department of Economics.

Morens, D. M. & Fauci, A. S. (2007), 'The 1918 influenza pandemic: insights for the 21st century', *The Journal of Infectious Diseases* **195**(7), 1018–1028.

Ohadike, D. C. (1991), 'Diffusion and physiological responses to the influenza pandemic of 1918–19 in nigeria', *Social Science & Medicine* **32**(12), 1393–1399.

Olapoju, O. M. (2020), 'Estimating transportation role in pandemic diffusion in nigeria: A consideration of 1918-19 influenza and covid-19 pandemics', *Journal of Global Health* **10**(2).

Patterson, K. D. & Pyle, G. F. (1991), 'The geography and mortality of the 1918 influenza pandemic', *Bulletin of the History of Medicine* **65**(1), 4–21.

Perra, N. (2021), 'Non-pharmaceutical interventions during the covid-19 pandemic: A review', *Physics Reports* **913**, 1–52.

Phillips, H. (1984), 'Black october: The impact of the spanish influenza epidemic of 1918 on south africa'.

Reyes, O., Lee, E. C., Sah, P., Viboud, C., Chandra, S. & Bansal, S. (2018), 'Spatiotemporal patterns and diffusion of the 1918 influenza pandemic in british india', *American Journal of Epidemiology* **187**(12), 2550–2560.

Rogers, F. B. (1968), 'The influenza pandemic of 1918-1919 in the perspective of a half century.', *American Journal of Public Health and the Nations Health* **58**(12), 2192–2194.

Taubenberger, J. K., Kash, J. C. & Morens, D. M. (2019), 'The 1918 influenza pandemic: 100 years of questions answered and unanswered', *Science Translational Medicine* **11**(502), eaau5485.

Tuckel, P., Sassler, S., Maisel, R. & Leykam, A. (2006), 'The diffusion of the influenza pandemic of 1918 in hartford, connecticut', *Social Science History* **30**(2), 167–196.

Appendix A

Descriptive Data and Tables

Table A.1: Descriptive Data [1]

VARIABLES	Obs.	Mean	SD	Min	Max
Panel A: Mortality and excess mortality					
Total	1,019,200	0.642	2.8	0	312
Female	1,019,200	0.324	1.485	0	171
Male	1,019,200	0.319	1.43	0	166
0-20	1,019,200	0.149	0.735	0	68
20-40	1,019,200	0.0945	0.777	0	163
40-60	1,019,200	0.0879	0.569	0	52
>60	1,019,200	0.311	1.148	0	83
SES1	1,019,200	0.21	0.769	0	97
SES2	1,019,200	0.146	0.619	0	62
SES3	1,019,200	0.1	0.522	0	50
SES4	1,019,200	0.0835	0.536	0	54
SES5	1,019,200	0.102	0.828	0	101
Total (excess)	1,019,200	0.0232	1.192	-45.16	223.4
Female (excess)	1,019,200	0.0109	0.717	-26.01	121.0
Male (excess)	1,019,200	0.0123	0.724	-25.14	128.3
0-20 (excess)	1,019,200	0.00275	0.450	-16.94	45.06
20-40 (excess)	1,019,200	0.0170	0.585	-11.90	151.0
40-60 (excess)	1,019,200	0.00291	0.317	-10.68	32.32
>60 (excess)	1,019,200	0.000585	0.603	-27.59	45.41
SES1 (excess)	1,019,200	0.00663	0.514	-13.50	80.50
SES2 (excess)	1,019,200	0.00526	0.422	-12.36	50.85
SES3 (excess)	1,019,200	0.00448	0.346	-9.899	35.10
SES4 (excess)	1,019,200	0.00418	0.322	-11.83	36.17
SES5 (excess)	1,019,200	0.00265	0.388	-18.30	69.70
Urban	44,096	4.093	12.2	0	312
Rural	975,104	0.486	0.948	0	37
Panel B: Distance to Railway Plans					
Von Renes's Plan	1,019,200	56,390	102,751	0.741	892,594
Ericson's Plan	1,019,200	64,679	102,627	15.13	893,258
Nodal plan	1,019,200	51,090	62,497	6.435	523,118
Min. Dist.	1,019,200	34,494	60,715	0.741	523,118

ln(Min. Dist.)	1,019,200	9.429	1.628	-0.300	13.17
Panel C: Railway Indicator					
within 1 km					
$Rail_i^{1908}$	1,019,200	0.0224	0.148	0	1
$Rail_i^{1918}$	1,019,200	0.0453	0.208	0	1
D_i^{1918}	1,019,200	0.0197	0.139	0	1
within 5 km					
$Rail_i^{1908}$	1,019,200	0.265	0.441	0	1
$Rail_i^{1908}$	1,019,200	0.409	0.492	0	1
D_i^{1918}	1,019,200	0.178	0.383	0	1
within 10 km					
$Rail_i^{1908}$	1,019,200	0.550	0.497	0	1
$Rail_i^{1908}$	1,019,200	0.685	0.465	0	1
D_i^{1918}	1,019,200	0.298	0.457	0	1
Panel D: Other variables					
debt	1,019,200	8.681	4.300	0	15.68
num. poorhouses	1,016,704	45.00	57.57	0	627.8
taxable income	1,019,200	11.37	5.071	0	17.32
family size	1,018,368	16.89	11.28	0	49.24
share of middle aged (20-40)	1,018,368	1.164	0.77	0	3.975
population density	1,015,872	2.076	9.341	0	433.0
breadwinner in HISCO category 1	1,017,952	0.0253	0.0227	0	0.289
breadwinner in HISCO category 2	1,017,952	0.0205	0.0246	0	0.391
breadwinner in HISCO category 3	1,017,952	0.0137	0.0251	0	0.303
breadwinner in HISCO category 4	1,017,952	0.0260	0.0370	0	0.753
breadwinner in HISCO category 5	1,017,952	0.155	0.145	0	3.176
breadwinner in HISCO category 6	1,017,952	0.658	0.508	0	2.797
breadwinner in HISCO category 7	1,017,952	0.0937	0.121	0	1.653
breadwinner in HISCO category 8	1,017,952	0.0617	0.0815	0	2.349
breadwinner in HISCO category 9	1,017,952	2.434	1.613	0	6.617

[1]**Note:** Summary statistics for key variables. **Panel A** contains summary statistics for the complete parish-week (2,441 weeks, 416 weeks). Each mortality outcome indicates the weekly number of death for different groups of individuals in each parish. It is decomposed into 12 subgroups by gender, age and socio-economic status, and also by urban and rural areas. **Panel B** contains summary statistics for initial railway plans used in IV identification. The first three variables capture the distance (in kilometres) between each parish and the nearest railway station measured by three different plans. And the following two variables measure the minimum value among three plans in the actual distance and natural log form respectively.

Table A.2.1 : Baseline Diff-in-Diff Output in all Parishes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Dependant variable: Mortality for different groups of people											
	Total	Female	Male	0-20	20-40	40-60	>60	SES1	SES2	SES3	SES4	SES5
$Rail_i^{1918}$	0.270	0.139	0.130	0.083	0.034	0.046	0.106	0.055	0.064	0.049	0.042	0.058
	(1.15)	(1.16)	(1.14)	(1.56)	(0.70)	(1.03)	(1.17)	(1.04)	(1.56)	(1.35)	(1.05)	(0.85)
Post	0.296***	0.145***	0.151***	0.088***	0.152***	0.033***	0.023***	0.096***	0.068***	0.048***	0.040***	0.044***
	(60.20)	(48.97)	(50.52)	(47.83)	(63.53)	(24.87)	(9.28)	(45.11)	(39.18)	(33.63)	(29.91)	(27.33)
D_i^{1918}	0.043***	0.014*	0.029***	0.025***	0.012**	0.007**	-0.003	0.010**	0.008**	0.005	0.007**	0.011***
	(3.70)	(1.95)	(4.11)	(5.78)	(2.21)	(2.30)	(-0.54)	(2.08)	(1.99)	(1.42)	(2.27)	(2.90)
Cluster SE												
Parish FE												
Year FE												
$Rail_i^{1918}$	0.270*	0.139*	0.130*	0.083**	0.034	0.046*	0.106*	0.055	0.064**	0.049**	0.042*	0.058*
	(1.95)	(1.96)	(1.92)	(2.22)	(1.46)	(1.95)	(1.86)	(1.52)	(2.24)	(2.16)	(1.88)	(1.83)
Post	0.296***	0.145***	0.151***	0.088***	0.152***	0.033***	0.023***	0.096***	0.068***	0.048***	0.040***	0.044***
	(12.01)	(12.00)	(11.80)	(22.96)	(9.47)	(7.36)	(8.51)	(15.02)	(14.40)	(12.59)	(9.38)	(6.47)
D_i^{1918}	0.043**	0.014	0.029**	0.025**	0.012	0.007*	-0.003	0.010	0.008	0.005	0.007*	0.011*
	(2.14)	(1.28)	(2.48)	(2.01)	(1.19)	(1.82)	(-0.33)	(1.28)	(1.55)	(0.97)	(1.80)	(1.86)
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Parish FE												
Year FE												
D_i^{1918}	0.157***	0.065***	0.092***	0.058***	0.074***	0.015***	0.009	0.047***	0.035***	0.024***	0.022***	0.029***
	(5.62)	(4.96)	(5.49)	(4.54)	(5.60)	(2.97)	(0.82)	(4.82)	(5.31)	(3.79)	(4.19)	(3.93)
$Rail_i^{1908} \times\ Year\ FE$	-0.028***	-0.011***	-0.017***	-0.008***	-0.014***	-0.001	-0.006**	-0.009***	-0.006***	-0.005***	-0.003**	-0.005***
	(-4.98)	(-3.62)	(-4.68)	(-2.92)	(-5.54)	(-0.61)	(-2.02)	(-4.11)	(-3.49)	(-2.84)	(-2.46)	(-2.87)
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Parish FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456
Number of Parish	2,441	2,441	2,441	2,441	2,441	2,441	2,441	2,441	2,441	2,441	2,441	2,441

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: regression outputs among all 2,441 parishes using Diff-in-Diff identification of the form: $Y_{it} = \lambda_i + \gamma_t + \beta \times D_i^{1918} + \delta_i \times Rail_i^{1908} \times I[Year] + Z'_{it} \times \Phi + \epsilon_{it}$. Each column represents mortality for different groups of people. $Rail_i^{1908}$ is a binary variable that equals 1 if a parish has a railway station in 1908, and it is interacted with the year indicator variable to make a panel. All regressions are conditioned on a full set of control variables. Demographic characteristics from the Swedish 1910 Census including the share of middle-aged individuals (20-40 years old), population density, average family size and breadwinners in 9 major HISCO categories interacted with continuous year indicators from 1914 to 1921. Parish pre-epidemic characteristics such as share of poor houses, debt amount and taxable income also interact with year indicator. The baseline land areas for each parish interact with year fixed effect.

Table A.2.2: Baseline Diff-in-Diff Output in Non-urban Parishes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Dependant variable: Mortality for different groups of people											
	Total	Female	Male	0-20	20-40	40-60	>60	SES1	SES2	SES3	SES4	SES5
$Rail_i^{1918}$	-0.016	-0.004	-0.012	-0.004	-0.002	-0.002	-0.008	-0.015	0.005	0.002	-0.003	-0.005
	(-0.31)	(-0.14)	(-0.48)	(-0.25)	(-0.29)	(-0.23)	(-0.37)	(-0.87)	(0.36)	(0.25)	(-0.46)	(-0.66)
Post	0.226***	0.110***	0.116***	0.077***	0.106***	0.023***	0.021***	0.080***	0.055***	0.038***	0.028***	0.027***
	(69.62)	(50.53)	(52.49)	(51.34)	(86.76)	(21.71)	(9.60)	(43.22)	(36.94)	(32.05)	(26.85)	(24.94)
D_i^{1918}	0.000	-0.007	0.007	0.003	-0.001	0.002	-0.004	-0.006	-0.001	0.003	0.003	0.002
	(0.06)	(-1.20)	(1.28)	(0.93)	(-0.45)	(0.93)	(-0.81)	(-1.41)	(-0.34)	(1.04)	(1.14)	(0.65)
Cluster SE												
Parish FE												
Year FE												
$Rail_i^{1918}$	-0.016	-0.004	-0.012	-0.004	-0.002	-0.002	-0.008	-0.015	0.005	0.002	-0.003	-0.005
	(-0.33)	(-0.14)	(-0.54)	(-0.35)	(-0.40)	(-0.25)	(-0.34)	(-0.88)	(0.36)	(0.27)	(-0.56)	(-0.83)
Post	0.226***	0.110***	0.116***	0.077***	0.106***	0.023***	0.021***	0.080***	0.055***	0.038***	0.028***	0.027***
	(29.80)	(29.25)	(26.03)	(27.26)	(29.67)	(17.19)	(8.74)	(25.55)	(24.27)	(22.07)	(20.50)	(18.16)
D_i^{1918}	0.000	-0.007	0.007	0.003	-0.001	0.002	-0.004	-0.006	-0.001	0.003	0.003	0.002
	(0.04)	(-1.19)	(0.93)	(0.89)	(-0.28)	(0.99)	(-0.78)	(-1.37)	(-0.34)	(0.82)	(1.17)	(0.62)
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Parish FE												
Year FE												
D_i^{1918}	0.072***	0.025***	0.047***	0.027***	0.035***	0.007**	0.002	0.018***	0.019***	0.014***	0.010***	0.011***
	(4.43)	(3.53)	(4.17)	(5.45)	(4.78)	(2.21)	(0.30)	(3.21)	(3.50)	(2.60)	(3.34)	(2.76)
$Rail_i^{1908} \times Year\ FE$	-0.014***	-0.005**	-0.009***	-0.005***	-0.006***	-0.000	-0.003	-0.005***	-0.005***	-0.002	-0.001	-0.002**
	(-3.68)	(-2.34)	(-3.60)	(-3.59)	(-4.44)	(-0.04)	(-1.27)	(-3.34)	(-3.30)	(-1.38)	(-0.65)	(-2.11)
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Parish FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	971,360	971,360	971,360	971,360	971,360	971,360	971,360	971,360	971,360	971,360	971,360	971,360
Number of Parish	2,335	2,335	2,335	2,335	2,335	2,335	2,335	2,335	2,335	2,335	2,335	2,335

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: regression outputs among all 2,335 rural parishes using Diff-in-Diff identification of the form: $Y_{it} = \lambda_i + \gamma_t + \beta \times D_i^{1918} + \delta_i \times Rail_i^{1908} \times I[Year] + Z'_{it} \times \Phi + \epsilon_{it}$. Each column represents mortality for different groups of people. $Rail_i^{1908}$ is a binary variable that equals 1 if a parish has a railway station in 1908, and it is interacted with the year indicator variable to make a panel. All regressions are conditioned on a full set of control variables. Demographic characteristics from the Swedish 1910 Census including the share of middle-aged individuals (20-40 years old), population density, average family size and breadwinners in 9 major HISCO categories interact with continuous year indicators from 1914 to 1921. Parish pre-epidemic characteristics such as share of poor houses, debt amount and taxable income also interact with year indicator. The baseline land areas for each parish interact with year fixed effect.

Table A.2.3: Baseline Diff-in-Diff Output in Urban Parishes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Dependant variable: Mortality for different groups of people											
	Total	Female	Male	0-20	20-40	40-60	>60	SES1	SES2	SES3	SES4	SES5
$Rail_i^{1918}$	-0.984	-0.514	-0.470	-0.199	-0.203	-0.192	-0.391	-0.183	-0.145	-0.142	-0.167	-0.347
	(-0.35)	(-0.36)	(-0.35)	(-0.33)	(-0.35)	(-0.35)	(-0.36)	(-0.30)	(-0.31)	(-0.34)	(-0.35)	(-0.41)
Post	1.963***	0.980***	0.983***	0.320***	1.268***	0.258***	0.117***	0.500***	0.393***	0.301***	0.331***	0.438***
	(21.98)	(19.94)	(19.81)	(11.89)	(25.98)	(13.20)	(3.52)	(17.97)	(16.50)	(14.70)	(15.78)	(15.10)
D_i^{1918}	-0.013	-0.039	0.026	0.202***	-0.146**	-0.008	-0.066	0.037	0.009	-0.032	-0.017	-0.013
	(-0.11)	(-0.60)	(0.39)	(5.71)	(-2.27)	(-0.30)	(-1.51)	(1.02)	(0.29)	(-1.19)	(-0.61)	(-0.34)
Cluster SE												
Parish FE												
Year FE												
$Rail_i^{1918}$	-0.984	-0.514	-0.470	-0.199	-0.203	-0.192	-0.391	-0.183	-0.145	-0.142	-0.167	-0.347
	(-0.71)	(-0.73)	(-0.70)	(-0.57)	(-0.81)	(-0.74)	(-0.73)	(-0.60)	(-0.59)	(-0.67)	(-0.71)	(-0.87)
Post	1.963***	0.980***	0.983***	0.320***	1.268***	0.258***	0.117***	0.500***	0.393***	0.301***	0.331***	0.438***
	(3.68)	(3.68)	(3.66)	(7.07)	(3.38)	(2.70)	(2.97)	(3.79)	(4.16)	(3.78)	(3.69)	(2.91)
D_i^{1918}	-0.013	-0.039	0.026	0.202*	-0.146	-0.008	-0.066	0.037	0.009	-0.032	-0.017	-0.013
	(-0.09)	(-0.41)	(0.38)	(1.74)	(-1.18)	(-0.25)	(-0.86)	(0.74)	(0.33)	(-1.07)	(-0.63)	(-0.28)
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Parish FE												
Year FE												
D_i^{1918}	0.556***	0.229***	0.328***	0.275**	0.256***	0.048*	-0.024	0.183***	0.116***	0.067***	0.081***	0.109***
	(5.13)	(3.60)	(4.72)	(2.55)	(3.92)	(1.98)	(-0.31)	(4.57)	(4.69)	(2.64)	(3.36)	(3.16)
$Rail_i^{1908} \times Year\,FE$	-0.063**	-0.025	-0.038*	0.012	-0.048***	0.001	-0.027*	-0.016	-0.007	-0.016**	-0.014*	-0.010
	(-2.17)	(-1.52)	(-1.86)	(0.38)	(-3.91)	(0.09)	(-1.86)	(-1.30)	(-0.72)	(-2.10)	(-1.78)	(-0.95)
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Parish FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	44,096	44,096	44,096	44,096	44,096	44,096	44,096	44,096	44,096	44,096	44,096	44,096
Number of Parish	106	106	106	106	106	106	106	106	106	106	106	106

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: regression outputs among all 106 urban parishes using Diff-in-Diff identification of the form: $Y_{it} = \lambda_i + \gamma_t + \beta \times D_i^{1918} + \delta_i \times Rail_i^{1908} \times I[Year] + Z'_{it} \times \Phi + \epsilon_{it}$. Each column represents mortality for different groups of people. $Rail_i^{1908}$ is a binary variable that equals 1 if a parish has a railway station in 1908, and it is interacted with the year indicator variable to make a panel. All regressions are conditioned on a full set of control variables. Demographic characteristics from the Swedish 1910 Census including the share of middle-aged individuals (20-40 years old), population density, average family size and breadwinners in 9 major HISCO categories interact with continuous year indicators from 1914 to 1921. Parish pre-epidemic characteristics such as share of poor houses, debt amount and taxable income also interact with year indicator. The baseline land areas for each parish interact with year fixed effect.

Table A.2.4: Baseline Diff-in-Diff Output in High SES Parishes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Dependant variable: Mortality for different groups of people											
	Total	Female	Male	0-20	20-40	40-60	>60	SES1	SES2	SES3	SES4	SES5
$Rail_i^{1918}$	0.412	0.198	0.214	0.141	0.039	0.068	0.166	0.093	0.106	0.069	0.061	0.088
	(0.54)	(0.50)	(0.58)	(0.82)	(0.24)	(0.47)	(0.56)	(0.55)	(0.78)	(0.59)	(0.46)	(0.39)
Post	0.572***	0.275***	0.297***	0.131***	0.330***	0.070***	0.041***	0.162***	0.125***	0.093***	0.085***	0.107***
	(34.26)	(29.13)	(30.90)	(23.62)	(37.25)	(17.69)	(5.95)	(27.84)	(25.12)	(22.05)	(20.34)	(19.48)
D_i^{1918}	0.092***	0.038**	0.054***	0.068***	0.016	0.012	-0.007	0.037***	0.026***	0.001	0.006	0.020*
	(2.78)	(2.04)	(2.80)	(6.17)	(0.93)	(1.54)	(-0.50)	(3.22)	(2.64)	(0.17)	(0.69)	(1.84)
Cluster SE												
Parish FE												
Year FE												
$Rail_i^{1918}$	0.412	0.198	0.214	0.141	0.039	0.068	0.166	0.093	0.106	0.069	0.061	0.088
	(1.20)	(1.14)	(1.26)	(1.57)	(0.61)	(1.13)	(1.20)	(1.09)	(1.53)	(1.26)	(1.09)	(1.02)
Post	0.572***	0.275***	0.297***	0.131***	0.330***	0.070***	0.041***	0.162***	0.125***	0.093***	0.085***	0.107***
	(5.92)	(5.77)	(6.02)	(12.06)	(4.96)	(4.12)	(5.35)	(6.68)	(7.07)	(6.39)	(5.16)	(4.00)
D_i^{1918}	0.092**	0.038	0.054**	0.068*	0.016	0.012	-0.007	0.037**	0.026**	0.001	0.006	0.020
	(2.08)	(1.48)	(2.16)	(1.86)	(0.58)	(1.22)	(-0.31)	(1.98)	(2.16)	(0.13)	(0.62)	(1.37)
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Parish FE												
Year FE												
D_i^{1918}	0.290***	0.129***	0.161***	0.120***	0.133***	0.029**	0.008	0.093***	0.067***	0.037***	0.037***	0.057***
	(5.10)	(4.54)	(4.89)	(3.42)	(4.52)	(2.43)	(0.36)	(4.36)	(4.94)	(2.97)	(3.21)	(3.69)
$Rail_i^{1908} \times Year\ FE$	-0.048***	-0.021***	-0.028***	-0.013	-0.025***	-0.001	-0.008	-0.013***	-0.009*	-0.010**	-0.008**	-0.009**
	(-4.03)	(-2.79)	(-3.47)	(-1.47)	(-4.67)	(-0.49)	(-1.27)	(-2.65)	(-1.89)	(-2.44)	(-2.39)	(-2.00)
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Parish FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	250,003	250,003	250,003	250,003	250,003	250,003	250,003	250,003	250,003	250,003	250,003	250,003
Number of Parish	619	619	619	619	619	619	619	619	619	619	619	619

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: regression outputs among all 619 parishes with higher economic status using Diff-in-Diff identification of the form: $Y_{it} = \lambda_i + \gamma_t + \beta \times D_i^{1918} + \delta_i \times Rail_i^{1908} \times I[Year] + Z'_{it} \times \Phi + \epsilon_{it}$. Each column represents mortality for different groups of people. $Rail_i^{1908}$ is a binary variable that equals 1 if a parish has a railway station in 1908, and it is interacted with the year indicator variable to make a panel. All regressions are conditioned on a full set of control variables. Demographic characteristics from the Swedish 1910 Census including the share of middle-aged individuals (20-40 years old), population density, average family size and breadwinners in 9 major HISCO categories interact with continuous year indicators from 1914 to 1921. Parish pre-epidemic characteristics such as share of poor houses, debt amount and taxable income also interact with year indicator. The baseline land areas for each parish interact with year fixed effect.

Table A.2.5: Baseline Diff-in-Diff Output in Low SES Parishes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Dependant variable: Mortality for different groups of people											
	Total	Female	Male	0-20	20-40	40-60	>60	SES1	SES2	SES3	SES4	SES5
$Rail_i^{1918}$	0.057	0.038	0.019	0.013	0.007	0.011	0.024	0.005	0.014	0.014	0.011	0.012
	(1.00)	(1.33)	(0.65)	(0.71)	(0.79)	(1.49)	(0.96)	(0.24)	(0.96)	(1.38)	(1.39)	(1.30)
Post	0.210***	0.105***	0.105***	0.074***	0.098***	0.021***	0.018***	0.076***	0.051***	0.034***	0.026***	0.024***
	(58.34)	(42.96)	(43.21)	(45.04)	(73.33)	(17.66)	(7.62)	(36.40)	(30.77)	(26.40)	(22.98)	(20.16)
D_i^{1918}	0.005	-0.005	0.010	0.002	0.003	0.002	-0.003	-0.006	-0.003	0.004	0.006**	0.002
	(0.56)	(-0.75)	(1.57)	(0.51)	(0.83)	(0.85)	(-0.42)	(-1.06)	(-0.67)	(1.29)	(2.27)	(0.82)
Cluster SE												
Parish FE												
Year FE												
$Rail_i^{1918}$	0.057	0.038	0.019	0.013	0.007	0.011	0.024	0.005	0.014	0.014	0.011	0.012
	(0.55)	(0.66)	(0.41)	(0.50)	(0.52)	(0.70)	(0.53)	(0.15)	(0.67)	(0.84)	(0.64)	(0.61)
Post	0.210***	0.105***	0.105***	0.074***	0.098***	0.021***	0.018***	0.076***	0.051***	0.034***	0.026***	0.024***
	(27.21)	(26.63)	(23.35)	(24.60)	(27.10)	(14.86)	(6.98)	(22.57)	(21.53)	(19.50)	(18.38)	(15.64)
D_i^{1918}	0.005	-0.005	0.010	0.002	0.003	0.002	-0.003	-0.006	-0.003	0.004	0.006*	0.002
	(0.29)	(-0.57)	(0.91)	(0.48)	(0.37)	(0.90)	(-0.31)	(-0.90)	(-0.68)	(0.92)	(1.90)	(0.54)
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Parish FE												
Year FE												
D_i^{1918}	0.078***	0.027**	0.051***	0.025***	0.039***	0.007*	0.007	0.019**	0.017***	0.015**	0.014***	0.012*
	(2.92)	(2.30)	(3.07)	(3.93)	(3.32)	(1.82)	(0.59)	(2.50)	(2.72)	(2.28)	(2.72)	(1.69)
$Rail_i^{1908} \times Year\ FE$	-0.015***	-0.005*	-0.010***	-0.004***	-0.007***	0.000	-0.004	-0.006***	-0.004***	-0.002	-0.001	-0.002
	(-2.82)	(-1.91)	(-2.95)	(-2.90)	(-3.22)	(0.03)	(-1.51)	(-2.91)	(-2.79)	(-1.49)	(-0.91)	(-1.56)
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Parish FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	765,453	765,453	765,453	765,453	765,453	765,453	765,453	765,453	765,453	765,453	765,453	765,453
Number of Parish	1,871	1,871	1,871	1,871	1,871	1,871	1,871	1,871	1,871	1,871	1,871	1,871

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: regression outputs among all 1,871 parishes with lower economics status using Diff-in-Diff identification of the form: $Y_{it} = \lambda_i + \gamma_t + \beta \times D_i^{1918} + \delta_i \times Rail_i^{1908} \times I[Year] + Z'_{it} \times \Phi + \epsilon_{it}$. Each column represents mortality for different groups of people. $Rail_i^{1908}$ is a binary variable that equals 1 if a parish has a railway station in 1908, and it is interacted with the year indicator variable to make a panel. All regressions are conditioned on a full set of control variables. Demographic characteristics from the Swedish 1910 Census including share of middle-aged individuals (20-40 years old), population density, average family size and breadwinners in 9 major HISCO categories interacted with continuous year indicators from 1914 to 1921. Parish pre-epidemic characteristics such as share of poor houses, debt amount and taxable income also interact with year indicator. The baseline land areas for each parish interact with year fixed effect.

Table A.2.6: Robustness Check results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Dependant variable: Mortality for different groups of people											
	Total	Female	Male	0-20	20-40	40-60	>60	SES1	SES2	SES3	SES4	SES5
Panel A: robustness test 1												
$D_i^{hypothesis}$	0.008	-0.005	0.013	0.029**	-0.004	0.001	-0.018*	0.005	0.000	-0.009	0.005	0.007
	(0.44)	(-0.46)	(1.12)	(2.26)	(-0.59)	(0.16)	(-1.75)	(0.66)	(0.01)	(-1.28)	(0.91)	(1.34)
$Rail_i^{1908} \times Year\,FE$	-0.003	0.000	-0.003	-0.002	-0.001	0.002	-0.002	-0.002	-0.001	0.000	-0.000	-0.001
	(-0.70)	(0.00)	(-1.13)	(-0.75)	(-0.69)	(1.53)	(-0.67)	(-0.71)	(-0.34)	(0.29)	(-0.34)	(-0.60)
Observations	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456	1,015,456
Number of Parish	2,441	2,441	2,441	2,441	2,441	2,441	2,441	2,441	2,441	2,441	2,441	2,441
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Company FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Panel B: robustness test 2												
$D_i^{hypothesis}$	0.124***	0.050***	0.074***	0.061***	0.050***	0.011*	0.002	0.037***	0.025***	0.018**	0.019***	0.026***
	(4.03)	(3.49)	(3.97)	(3.56)	(4.22)	(1.87)	(0.19)	(3.50)	(3.49)	(2.38)	(3.11)	(3.11)
Observations	992,576	992,576	992,576	992,576	992,576	992,576	992,576	992,576	992,576	992,576	992,576	992,576
Number of Parish	2,386	2,386	2,386	2,386	2,386	2,386	2,386	2,386	2,386	2,386	2,386	2,386
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Company FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Robust t-statistics in parentheses												
*** p¡0.01, ** p¡0.05, * p¡0.1												

Note: results from two types of robustness check regressions. In **Panel A**, we assume the 1918 influenza happened in 1916. We rerun the baseline regression $Y_{it} = \lambda_i + \gamma_t + \beta \times D_i^{hypothesis} + \delta_i \times Rail_i^{1908} \times I[Year] + Z'_{it} \times \Phi + \epsilon_{it}$ by replacing D_i^{1918} to $D_i^{hypothesis}$, where $D_i^{hypothesis}$ represents the main treatment variable assuming the hypothesis pandemic began in 1916. In **Panel B**, I remove those parishes that already have a station in 1908 from the sample, and run the regression $Y_{it} = \lambda_i + \gamma_t + \beta \times D_i^{1918} + Z'_{it} \times \Phi + \epsilon_{it}$. All regressions are clustered at the parish level, combined with a two-way fixed effect model, and a full set of controls.

Table A.3.1: First stage: proximity to railway plans and railway stations emergence

	Dependent variable: Emergence of railway (D_i^{1918})								
	within 1 km			within 5 km			within 10 km		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$ln(Plandist) \times 1915$	0.001***	0.001***	0.001***	0.004***	0.004***	0.004***	0.004***	0.004***	0.004***
	(4.28)	(3.01)	(3.93)	(7.31)	(8.61)	(6.51)	(7.89)	(9.31)	(7.06)
$ln(Plandist) \times 1916$	0.002***	0.002***	0.001***	0.008***	0.008***	0.008***	0.008***	0.008***	0.008***
	(4.28)	(6.07)	(3.89)	(7.31)	(17.65)	(7.48)	(7.89)	(18.85)	(7.92)
$ln(Plandist) \times 1917$	0.003***	0.001***	0.002***	0.012***	0.013***	0.012***	0.012***	0.013***	0.012***
	(4.28)	(9.08)	(3.87)	(7.31)	(26.21)	(7.55)	(7.89)	(27.86)	(8.03)
$ln(Plandist) \times 1918$	-0.003*	-0.003**	-0.003*	-0.019***	-0.018***	-0.018***	-0.032***	-0.031***	-0.030***
	(-1.76)	(-10.31)	(-1.88)	(-5.79)	(-27.62)	(-5.25)	(-10.33)	(-40.27)	(-9.67)
$ln(Plandist) \times 1919$	-0.009***	-0.009***	-0.009***	-0.054***	-0.052***	-0.050***	-0.080***	-0.079***	-0.074***
	(-2.46)	(-29.15)	(-2.75)	(-8.79)	(-91.91)	(-7.99)	(-13.98)	(-149.06)	(-12.97)
$ln(Plandist) \times 1920$	-0.008***	-0.008***	-0.008***	-0.050***	-0.048***	-0.046***	-0.076***	-0.074***	-0.070***
	(-2.46)	(-26.41)	(-2.49)	(-7.86)	(-83.44)	(-7.14)	(-12.87)	(-139.96)	(-11.96)
$ln(Plandist) \times 1921$	-0.007***	-0.007***	-0.007***	-0.046***	-0.044***	-0.042***	-0.072***	-0.070***	-0.066
	(-2.46)	(-23.62)	(-2.22)	(-6.94)	(-74.27)	(-6.29)	(-11.76)	(-129.70)	(-10.94)
Observations	1,015,872	1,015,456	1,015,457	1,015,872	1,015,456	1,015,457	1,015,872	1,015,456	1,015,457
Parishes	2,442	2,441	2,442	2,442	2,441	2,441	2,442	2,441	2,441
Parish FE	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES
Cluster SE	YES	YES	YES	YES	YES	YES	YES	YES	YES
demographic chars. Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES
parish chars. Controls		YES	YES		YES	YES		YES	YES
baseline parish chars.$\times YearFE$			YES			YES			YES
Kleibergen-Paap rk LM	29.61***	3038***	28.89***	159***	26000***	147***	284***	61000***	247***
Cragg-Donald Wald F	616***	634***	614***	4056***	3869***	3391***	8333***	8134***	6955***
Kleibergen-Paap Wald rk F	4.646***	440***	4.458***	28.32***	4117***	25.46***	53.30***	10000***	44.90***
Hansen J p-value	0.085*	0.000***	0.130	0.077*	0.000***	0.115	0.08*	0.000***	0.104

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: First stage regressions of the form: $D_i^{1918} = \lambda_i + \gamma_t + \sum_d \delta_d \times PlanDist_i \times I[Year = d] + Z'_{it} \times \Phi + \nu_{it}$. The natural log form of minimum distance to railway plan interacts with fixed year effects from 1914 to 1921, and we take 1914 as a base year thus omitted. The dependent variable is the emergence of railway stations (the main independent variable in Diff-in-Diff identification). I make a comparison by using 1km, 5km and 10km as hurdles for identifying whether a parish was near the railway station. Constant Demographic characteristics from the Swedish 1910 Census including share of middle-aged individuals (20-40 years old), population density, average family size and shares of breadwinners in 9 major HISCO categories interacted with continuous year indicator from 1914 to 1921 to make into a panel. Parish pre-epidemic characteristics such as share of poor houses, debt amount and taxable income also interact with year indicator. The baseline land areas for each parish interact with year fixed effect.

Table A.3.2: Main results: railway station emergence and mortality number

	Dependent variable: weekly mortality number					
	OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Railway station within 1 km						
D_i^{1918}	0.157***	0.157***	0.157***	0.565	0.504	0.803
	(0.0280)	(0.0279)	(0.0278)	(0.522)	(0.330)	(0.523)
Observations	1,015,872	1,015,456	1,015,456	1,015,872	1,015,456	1,015,456
Parishes	2,442	2,441	2,441	2,442	2,441	2,441
K-P Wald F				4.646	440.501	4.458
Panel B: Railway station within 5 km						
D_i^{1918}	0.152***	0.154***	0.157***	0.0810	0.0821	0.143*
	(0.0258)	(0.0257)	(0.0257)	(0.0747)	(0.0562)	(0.0806)
Observations	1,015,872	1,015,456	1,015,456	1,015,872	1,015,456	1,015,456
Parishes	2,442	2,441	2,441	2,442	2,441	2,441
K-P Wald F				28.319	4117.366	25.459
Panel C: Railway station within 10 km						
D_i^{1918}	0.224***	0.227***	0.233***	0.0459	0.0450	0.0880*
	(0.0260)	(0.0259)	(0.0261)	(0.0471)	(0.0377)	(0.0498)
Observations	1,015,872	1,015,456	1,015,456	1,015,872	1,015,456	1,015,456
Parishes	2,442	2,441	2,441	2,442	2,441	2,441
K-P Wald F				53.302	10000	44.903
Parish FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
Cluster SE	YES	YES	YES	YES	YES	YES
demographic chars. Controls	YES	YES	YES	YES	YES	YES
parish chars. Controls		YES	YES		YES	YES
baseline parish chars.*year FE			YES			YES

robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: OLS and IV regressions output of the form: $Y_{it} = \lambda_i + \gamma_t + \beta \times D_i^{1918} + \delta_i \times Rail_i^{1908} \times I[Year] + Z'_{it} \times \Phi + \epsilon_{it}$. Dependent variables are defined as the weekly mortality number in the parish. The independent variables are defined as follows. Panel A: indicate whether a railway station is near enough to the parish (within 1 km) when influenza started. Panel B: indicate whether a railway station is near enough to the parish (within 5 km) when influenza started. Panel C: indicate whether a railway station is near enough to the parish (within 10 km) when influenza started. Constant Demographic characteristics from the Swedish 1910 Census including share of middle-aged individuals (20-40 years old), population density, average family size and shares of breadwinners in 9 major HISCO categories interacted with continuous year indicator from 1914 to 1921 to make into a panel. Parish pre-epidemic characteristics such as share of poor houses, debt amount and taxable income also interact with year indicator. The baseline land areas for each parish interact with year fixed effect.

Appendix B

Figures and Graphs

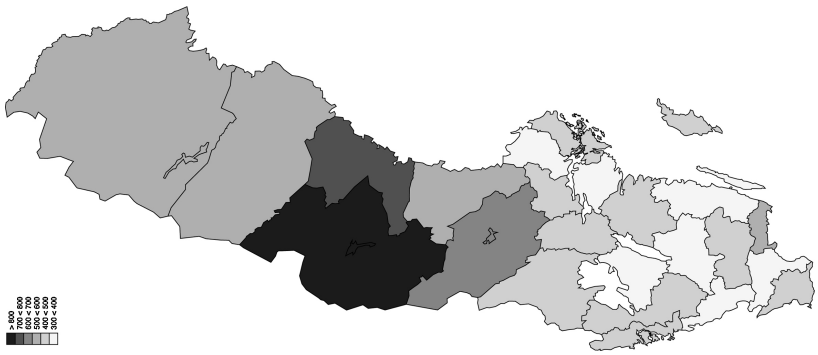
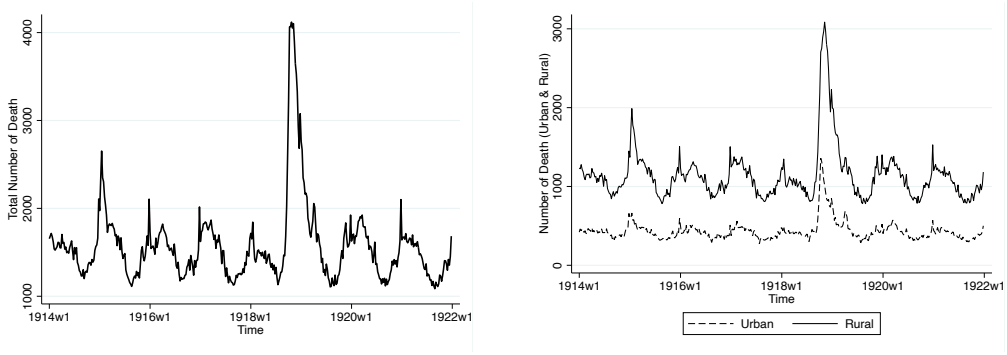


Figure B.1: influenza mortality rates in Swedish counties 1918–1920 (per 100,000 inhabitants);
source: Karlsson et al. (2014)



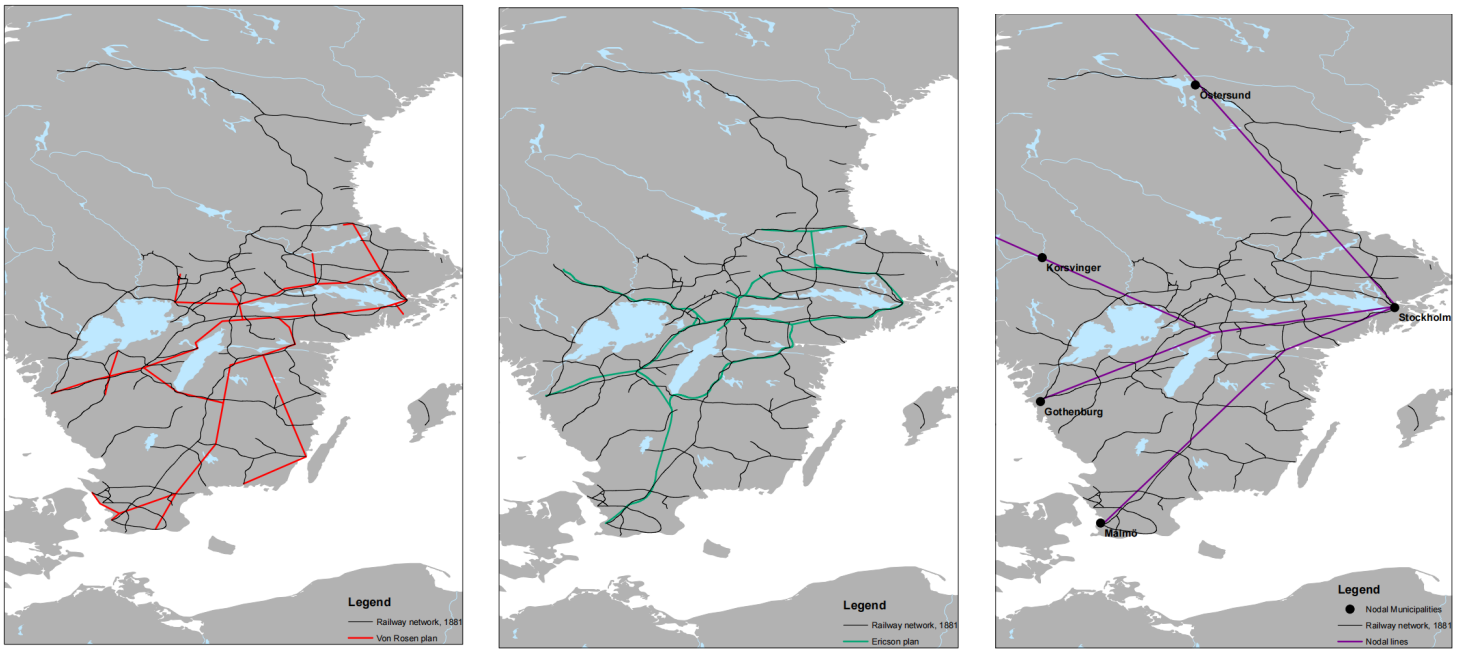
(a) total population

(b) Urban and Rural

(c) age groups

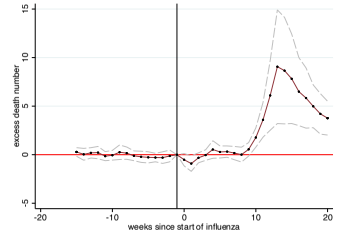
(d) SES groups

Figure B.2: Mortality trend in Sweden: 1914-1921

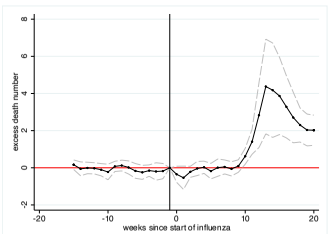


(a) von Rosen plan (b) Ericson plan (c) Nodal plan

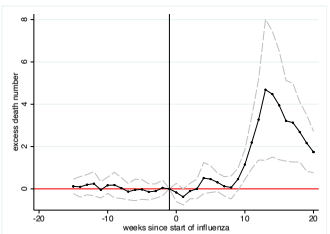
Figure B.3: Initial railway plans in Sweden;
source: Melander (2020)



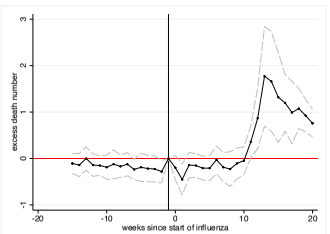
(a) total excess death



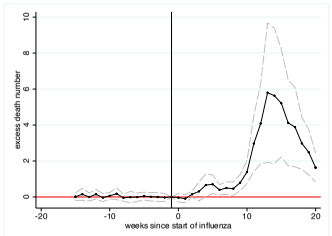
(b) female excess death



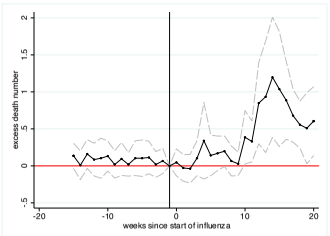
(c) male excess death



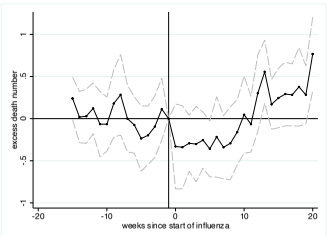
(d) 0-20 excess death



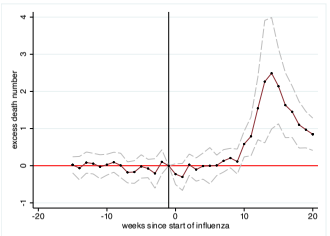
(e) 20-40 excess death



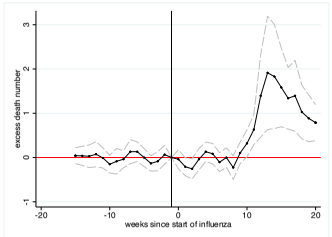
(f) 40-60 excess death



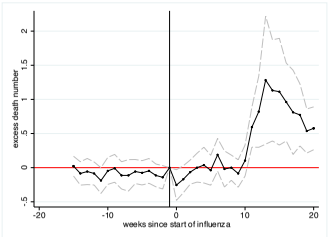
(g) over 60 excess death



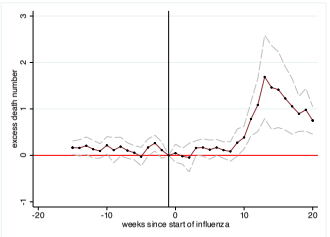
(h) ses1 excess death



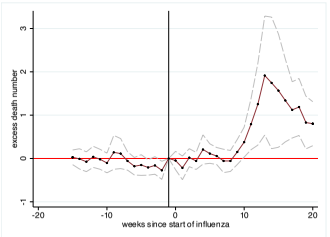
(i) ses2 excess death



(j) ses3 excess death

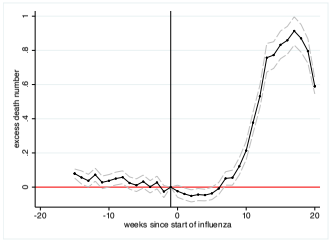


(k) ses4 excess death

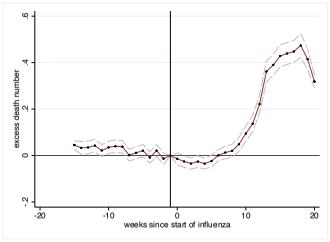


(l) ses5 excess death

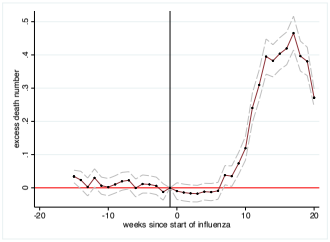
Figure B.4.1: Common Trend Plots (urban)



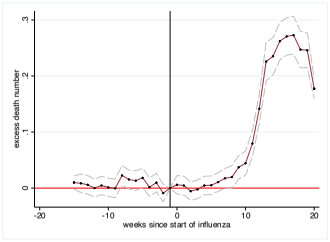
(a) total excess death



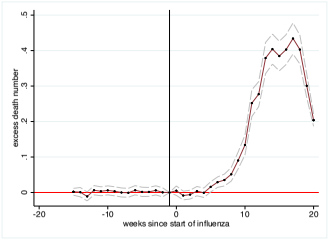
(b) female excess death



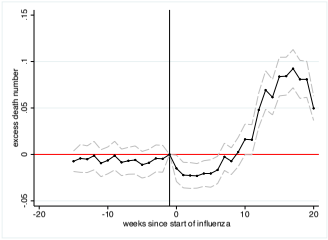
(c) male excess death



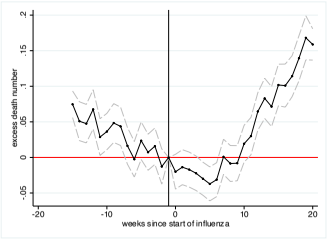
(d) 0-20 excess death



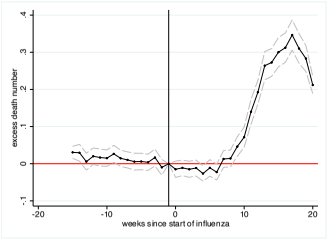
(e) 20-40 excess death



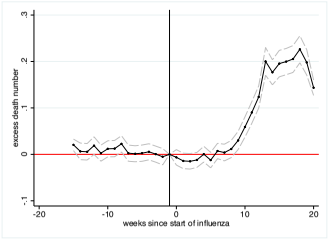
(f) 40-60 excess death



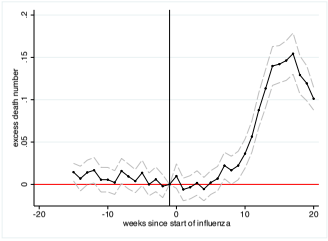
(g) over 60 excess death



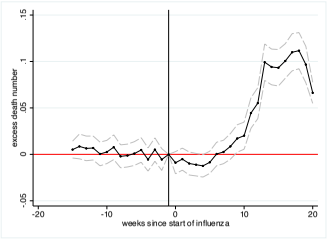
(h) ses1 excess death



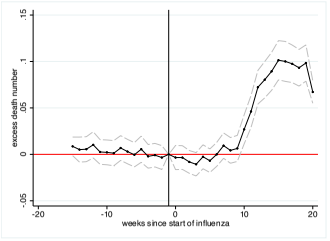
(i) ses2 excess death



(j) ses3 excess death



(k) ses4 excess death



(l) ses5 excess death

Figure B.4.2: Common Trend Plots (rural)