Who are you sneezing at: Role of networks in spreading the flu

Abstract

Epidemics can have devastating health and economic consequences. This paper studies the diffusion of flu through social and economic networks. Using almost two decades of weekly, county-level infection and mortality data from Poland, it studies within and across-counties flu transmission. Firstly, it evaluates the causal effect of school closures on viral transmission. The results show that closing schools for two weeks decreases the number of within county flu cases by 30-40%. The decline in infections extends to elderly and pre-school children. In addition, flu-related hospitalizations drop by 7.5%, and mortality related to respiratory diseases among the elderly drops by 3%. Secondly, the paper demonstrates the significant contribution of economic links to diffusion across counties. The disease follows the paths of workers commuting between home and workplace. Together with the structure of the labor mobility networks, these results highlight the central role of regional capitals in sustaining and spreading the virus.

Keywords: influenza, school closure, mortality, commuting network

1 Introduction

Infectious diseases impose a significant health and economic tolls on society. Every year in the USA 20,000 to 60,000 people die and 140,000 to 810,000 are hospitalized due to flu infection[1]. The global pandemic of COVID-19 stopped economic activity, causing a plunge in stock markets and massive lay-offs. The negative consequences of epidemics may persist for generations, as children born to mothers infected with flu receive inferior wages in their adult lives (Schwandt [2018], Almond [2006]).

This paper uses Polish data to causally investigate the role of within-county and cross-county networks in spreading the flu, and evaluates the impact of policies limiting interactions. The analysis relies on high-frequency incidence data with geographical granularity exceeding previous work.

The first part of my paper investigates the effect of school closures on the number of infections, hospitalizations, and deaths. I take advantage of the winter vacation schedule, which varies across regions and years. The observed effects are causal as the academic calendar is set in the year prior to the recess. This study follows the framework of Adda [2016] who uses an event study to explore the impact of school winter break on flu, gastroenteritis, and chickenpox in France. His paper additionally investigates the impact of transportation and trade networks on the diffusion of flu. Similar in spirit is work by Stoecker et al. [2016]. He shows that flu mortality increases in counties with teams participating in the Superbowl. It comes as an aftermath of increased socialization among the elderly, who are at high risk of influenza complications. Besides economics, epidemiologists produced evaluations of the impact of school closures. Their analysis usually focuses on a particular season of influenza, and the impact on flu varies from none to 40% decline in cases. (Wheeler et al. [2010],Cowling et al. [2008], Cauchemez et al. [2009], Cauchemez et al. [2008]). My analysis uses winter recess as an exogenous drop in social interactions. The event study provides further evidence of the success of school closure in slowing the diffusion. The incidence drops about 30-50% in a week following the vacation. Moreover, doctors report fewer cases for up to three weeks after the recess. Furthermore, my work contributes to the literature

[1] https://www.cdc.gov/flu/about/index.html

by showing externalities of the policy to groups with a high risk of complications. School closure causes a significant drop in infections among elderly and pre-school children. The closure also frees up healthcare resources by decreasing the number of flu-related hospitalizations. Finally, fewer deaths related to respiratory diseases occur after the school recess. Closing schools comes with a cost in terms of human capital and productivity of care-takers. This paper discusses only health implications, a reader interested in costs may refer to Lempel et al. [2009] and Adda [2016].

The second part contributes to the literature exploring the role of economic activity in spreading diseases(Oster [2012], Adda [2016], Markowitz et al. [2010]). It uses information on inter-county commuters to examine the role of labor mobility in spreading the flu. A unique administrative data set provides the number of work commuters between each pair of Polish counties. To obtain casual estimates, I instrument prior stock of infections with environmental conditions influencing the spread. In particular, I use temperature, humidity and sunlight, which affect past infections through their impact on immunity, ease of transmission and behavioral patterns. The IV estimates show a significant contribution of these economic linkages to the diffusion of the flu. By using higher geographical resolution and IV method, my study adds to the results obtained by Pei et al. [2018]. Their work demonstrates that interstate labor mobility helps to predict the peak of the flu epidemic. My results and the network structures highlight the central role of regional capitals in sustaining and spreading the flu. The high connectivity of these counties sustains the basic reproductive number [2] above one. Limiting connections to the regional capitals is a crucial step in halting the diffusion.

The remainder of the paper is structured as follows. First, I discuss the data. Next, I show the model guiding the remaining analysis. Third, I use the event study to demonstrate the impact of winter vacation on flu diffusion and mortality. In the last section, I investigate the role of labor mobility in spreading the virus.

2 Data

The study is set in Poland and leverages detailed data available in the country.

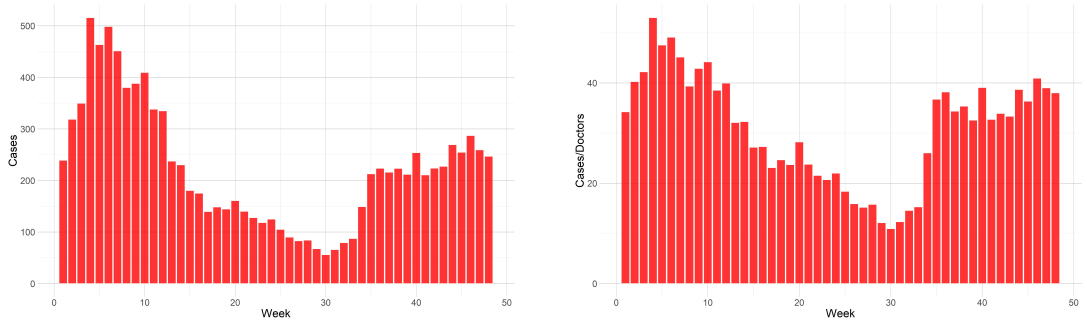
Infections

Data related to the flu infections originate in the weekly reports submitted by first contact doctors. The reporting system is part of the national flu surveillance program. Each report specifies the number of patients with flu symptoms and the number of flu-related hospitalizations. These numbers are divided into four age groups: 0-4, 5-14, 15-65, 65+ years old. The reports are submitted voluntarily to the local epidemiological station of a "powiat". "Powiat" is an administrative entity that corresponds to a county, and I will be using the word "county" henceforth. County and regional ("wojewódzkie") epidemiological stations enabled me to access the report data. Figure 8 in the appendix shows counties that made the data available. My data covers 168 out of 380 counties.

Although I cannot see every report, I can observe the total number of cases reported and the number of doctors submitting reports for each county and each week. Nationally, about 1000 health practitioners are taking part in the program. For most counties, I can also observe the division of cases into age groups and the number of hospitalizations. The data spans the years 2005-2019. The number of years covered varies across counties. However, at least five years of data is available for each county. Each month is divided into four weeks, so each year has 48 weeks.

Figures 1a and 1b summarize the average weekly incidence in a county. The left panel shows the mean number of infections reported in a county in a given week. This quantity can increase because either new infections occurred or more doctors submitted reports. To isolate the first channel, I divide the number of infections by the number of doctors submitting reports. This variable is defined as "cases per doctor". It can be interpreted as a draw from a sampling distribution of a number of infections reported by a doctor. As such, it has a true value as its expectation and its variance decreases in the number of doctors submitting the reports. The right panel of figure 1 displays the mean of cases per doctor for each week. An epidemic usually peaks in February-March,

[2] It is a number determining whether the epidemic can be self-sustaining



(a) Total incidence

Note: Each bar corresponds to an average number of cases across years and counties

(b) Cases per doctor

Note: Each bar corresponds to an average number of cases reported per doctor across years and counties

Figure 1: Flu evolution

it plummets during summer, and it picks up again in September when the school year starts. The number of hospitalizations exhibits a similar pattern, as demonstrated in figure 9 in the appendix.

Deaths

Data on deaths come from administrative sources and cover all Polish counties in the years 2000-2018. In particular, the dataset provides the number of deaths by cause, age group, county, week, and year. This paper uses deaths due to respiratory system diseases that correspond to category J in ICD-10. For a placebo check, I use deaths due to external causes corresponding to category V in ICD-10. To calculate mortality per 100 000, I use yearly population data by age from Polish Statistical Office (GUS).

Weather

Environmental data, serving as instrumental variables and controls, come from ERA5 Land (Muñoz Sabater [2019]). It is a climate reanalysis for the global land surface. It provides hourly values of atmospheric variables at a grid of 9km resolution. This paper uses information on temperature, humidity, and sunlight. These variables have been shown to impact flu transmission (White [2017],Barreca and Shimshack [2012], Slusky and Zeckhauser [2018]). Their impact on stock of infections can come through easier transmission (humidity), impact on immunity (sunlight), and changes in whether people stay indoors or outdoors (temperature). The temperature comes in Kelvins, and I translate it to Celcius. Humidity comes as the dew point temperature, which I transform to absolute humidity by using the information on the actual temperature. Sunlight comes as the downward solar radiation (measured in Joules per m^2) accumulated during the day. I average the values to obtain mean temperature, humidity, and sunlight for each county and each year-week. Figure 10 in the appendix show averages of these variables for each week across counties and years.

Labor mobility

Information on labor mobility between counties is based on the tax and social security data from 2011 (Basarbowicz et al. [2014]). For each pair of counties i and j, this data specifies how many people living in a county i commuted to work in county j. Figure 2 shows the labor mobility connections amongst counties.

Nodes represent counties, their color corresponds to the region, and their size is proportional to the number of incoming workers. A link from i to j represents labor mobility between the two counties. Its width is proportional to the number of workers living in a county i and working in county j.

Three main features arise. First, Warsaw, the capital of the country, is the most central county. Second, most labor mobility occurs from counties to their respective regional capitals. Third, there is significant mobility between regional capitals.

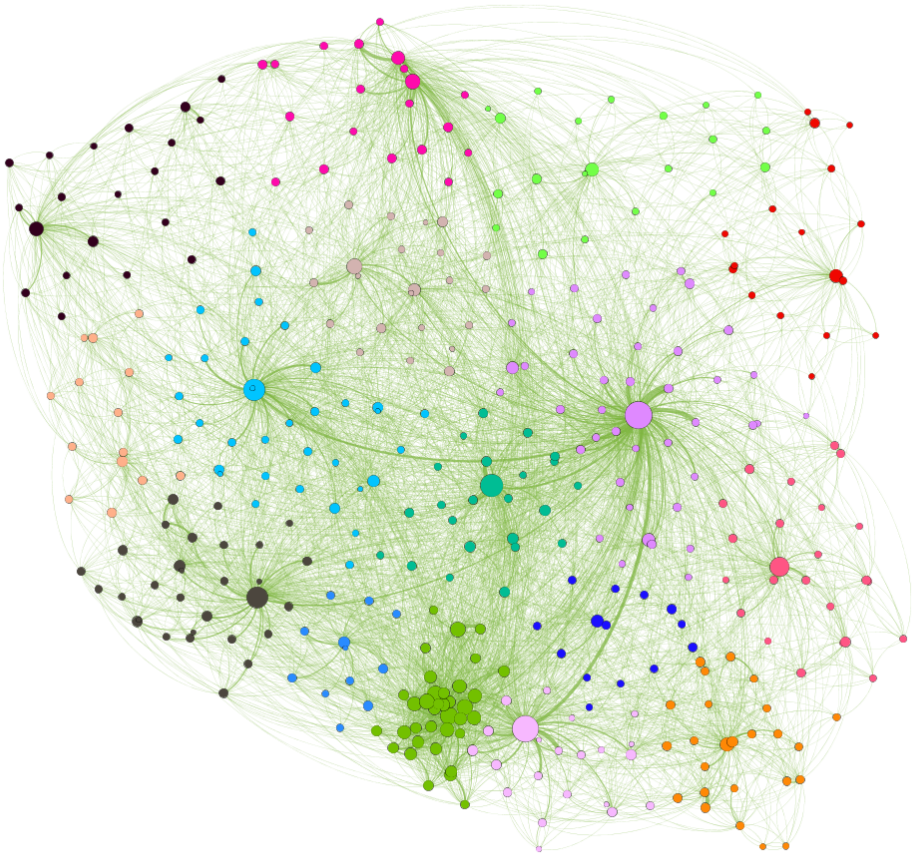


Figure 2: Labor mobility network

Note: Nodes represent counties and are located at counties' centroids. The color of nodes corresponds to the region ("wojewodztwo"). The width of the green links is proportional to the number of workers commuting between the two counties.

3 Model

This section discusses the epidemiological model which guides the analysis. I assume that infections follow the "Susceptible-Infected-Recovered" model (Anderson et al. [1992]). An agent can be in either of the three states: *Susceptible*, *Infected* or *Recovered*. *Susceptible* individuals can transfer to the *infected* state if a transmission occurs. Upon recovery, they move from *infected* to *recovered* and stay there until the loss of immunity. Then they become *susceptible* again. Figure 3 illustrates this process.

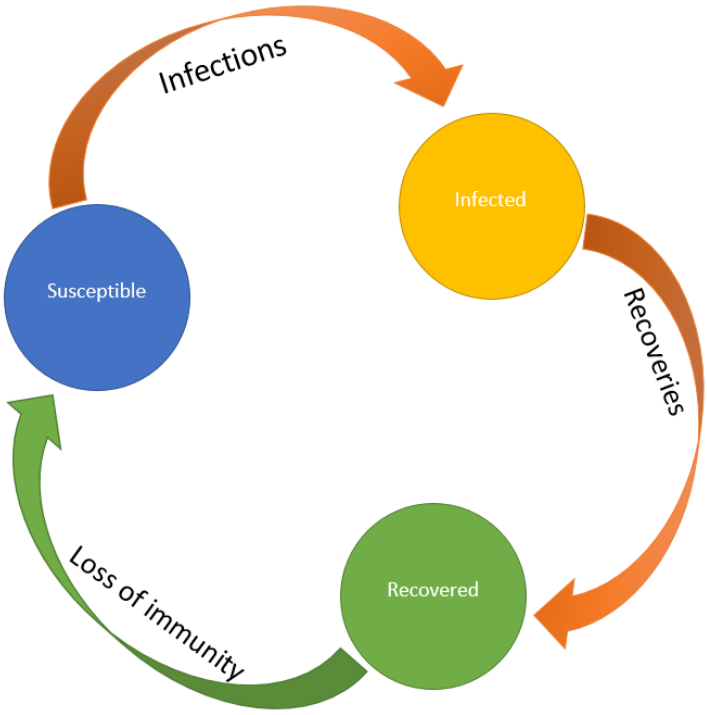


Figure 3: SIR model

Note: Circles represent states in the SIR model and arrows represent possible transitions between states.

The number of infected individuals changes for two reasons. First, there is an inflow as susceptible individuals get infected. Second, there is an outflow as sick agents recover. Equation 1 describes the evolution of the infected population.

$$I_{t,i} = \alpha d_{ii} \frac{S_{t-1,i}}{P_{t-1,i}} I_{t-1,i} + \sum_{j \neq i} \alpha d_{ij} \frac{S_{t-1,i}}{P_{t-1,i}} I_{t-1,j} + (1-\beta) I_{t-1,i} + \varepsilon_{i,t} \tag{1}$$

Here $I_{i,t}$ represents the number of infected individuals at time t in a county i. Three elements determine $I_{i,t}$.

First, there are new infections from within the county. They are equal to $\alpha d_{ii} \frac{S_{t-1,i}}{P_{t-1,i}} I_{t-1,i}$. Infections result from interactions between infected and susceptible individuals, hence this term is proportional to the product of individuals infected in the last period and the share of population which is still susceptible. This share is $\frac{S_{t-1,i}}{P_{t-1,i}}$ where $S_{t-1,i}$ [3] is the susceptible population in a county i at time $t-1$ and $P_{t-1,i}$ is the total population in the same period. It is further multiplied by d_{ii} which is the number of contacts of an individual from county i with other agents in this county. Consequently, $d_{ii} \frac{S_{t-1,i}}{P_{t-1,i}}$ represents the number of susceptible contacts of an individual

[3] It is defined as total population minus the number of cases since the beginning of the season, that is since last September, see appendix for robustness to different measures of susceptibility

in county i with other inhabitants of county i and $d_{ii}\frac{S_{t-1,i}}{P_{t-1,i}}I_{t-1,i}$ stands for the total number of susceptible contacts of infected individuals in county i with inhabitants of county i. Hence this last term comprises all opportunities for transmission within a county.The remaining term α is per contact probability of transmission.

Second, there are new infections from other counties represented by $\sum_{j\neq i}\alpha d_{ij}\frac{S_{t-1,i}}{P_{t-1,i}}I_{t-1,j}$. Analogously to the previous term, $d_{ij}\frac{S_{t-1,i}}{P_{t-1,i}}$ stands for the number of susceptible contacts in a county i of an individual in a county j. Hence, $\alpha d_{ij}\frac{S_{t-1,i}}{P_{t-1,i}}I_{t-1,j}$ are new infections in a county i coming from county j. The sum is taken over all counties, even though d_{ij} may be close to 0 for distant locations. Third, some individuals did not recover from the previous period, which is equal to $(1-\beta)I_{i,t-1}$, where β stands for the recovery rate.

Lastly, $\epsilon_{t,i}$ includes all determinants of infections not included in previous terms.

Several assumptions underlie this model. Firstly, it assumes random mixing. Namely, the probability that contact is infected equals the share of infected agents in their county population. That assumption would be satisfied if the matching process was random. As the flu spreads through droplets in the air, it does not require any intentional contact [4]. In other words, it spreads even through random meetings. Hence flu diffusion is relatively consistent with random mixing.

Secondly, one needs to define the time period of the model. I assume it to be one week, as influenza has an incubation period of no more than one week (Lessler et al. [2009]) .

Thirdly, I assume that agents return to a doctor if they do not recover after a week. Such a defined model guides the analysis in the remainder of the paper.

4 Role of the school interactions: an event study

This section discusses event studies that demonstrate the school closures' impact on new infections, hospitalizations, and mortality.

Interactions among students at school provide an excellent platform for the spread of the virus. Children spend from 5 to 8 hours in a classroom daily, with an average class size of around 30 students. Consequently, one sick pupil can easily transmit the disease to several colleagues. Insufficient hygienic practices further facilitate the diffusion of the flu. Additionally, infected children bring the virus home and transmit it to their siblings, parents, and grandparents. Hence, students can act as a link between the pathogen and vulnerable populations.

To measure the importance of in-school interactions for the incidence of the flu, I exploit variation winter vacation timing. At the beginning of each year, all schools close for two weeks. The timing of the closure depends on the region. In addition to the geographic variation in a given year, the order in which regions close schools varies across years. The school calendar is usually set in June preceding the winter vacation (approximately seven months in advance). At that time, decision-makers do not have any information about the course of the future epidemic. Hence the timing of the recess across regions is independent of the geographical differences in the incidence level.

Equation 1 sheds light on the impact of this policy on the level of infections. As schools close, the contact rate between children plummets. The intervention, therefore, corresponds to a temporary decline in the parameter governing the interactions within the county denoted as d_{ii}. The model predicts a decrease in incidence in the weeks following the start of the vacation. However, the effects can last longer than the vacation period because the pool of infected individuals shrinks. Consequently, fewer new infections may arise even 2 to 3 weeks after the recess.

Equation 2 represents the empirical framework of the event studies:

$$y_{i,t} = \sum_{T=-5}^{10} \gamma_T 1_{i,t-T} + \delta X_{it.} + e_{i,t} \tag{2}$$

The first event study investigates the impact of winter recess on the number of infections per doctor. The outcome variable $y_{i,t}$ is $log(\frac{I_{i,t}}{L_{i,t}}+1)$, log of cases per doctor in a county i in a week

[4] Contrary to, for instance, sexually transmitted diseases

t plus 1. I add 1 to avoid the issue of $log(0)$. The treatment is modelled as a set of indicator functions. $1_{i,t-T}$ equals 1 if a county i has winter vacation at time $t-T$, and 0 otherwise. T starts at -5 and finishes at 10, so γ_Ts capture the effects ranging from 5 weeks prior to vacation to 10 weeks after the break. The main event corresponds to $T=0$, which is the week when the recess starts. The sixth week before the vacation is excluded, and hence it serves as a reference point [5]. X_{it} includes a set of dummies controlling for the county-year, week-year of the year-week fixed effects. These fixed effects should effectively control for different patterns of seasonality and reporting across counties. Standard errors are clustered at the year-region level. The results for the incidence are shown in the figure 4. The line represents the coefficients γ_Ts. The shaded region covers 95% confidence interval of γ_Ts.

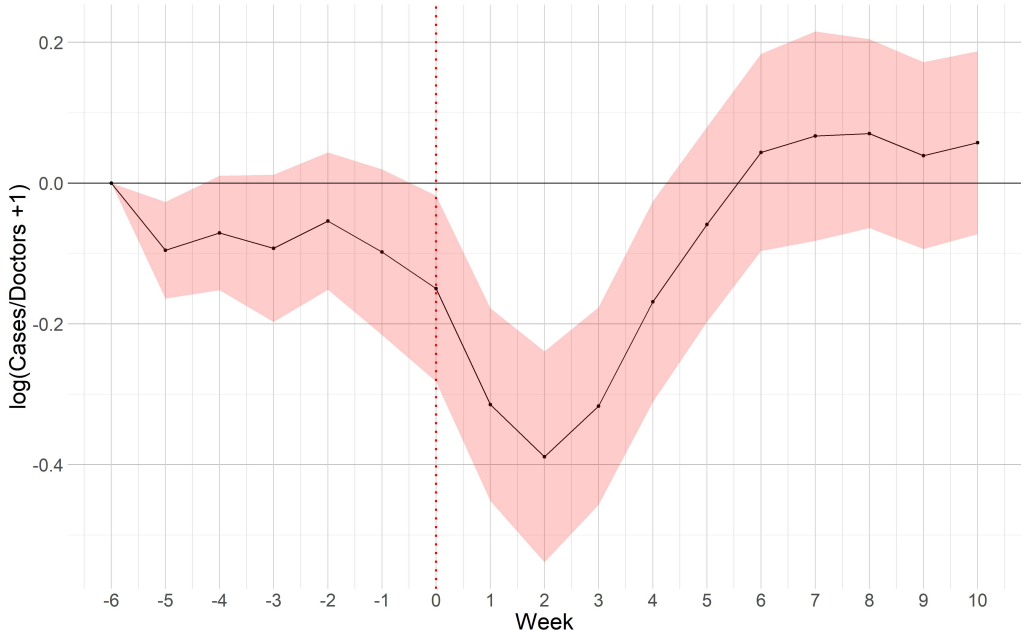


Figure 4: Winter Vacation Event Study: Infections

Note: Line represents estimates of the coefficients γ_T from equation 2 where the outcome is $log(\frac{I_{i,t}}{L_{i,t}}+1)$. Shared area represents 95% confidence interval for the estimates.

The figure shows a significant decrease in the number of cases per doctor starting during the vacation[6]. However, the most significant decline occurs two weeks after the recess, with 40% fewer cases per doctor. The effects persist for three weeks after the end of the break. These results are consistent with the numbers reported by Adda [2016] who observes a 20%-30% drop in incidence.

The figure corroborates the hypothesis that the decline results from the lower level of interactions among students. In the appendix (figure 12) I show that winter vacation decreases the transmission rate (αd_{ii}) in equation (1) . Undeniably, the boost in travel and related absence may contribute to the decrease in reporting during the recess. However, the subsequent drop cannot be attributed to an absence during the vacation, as children need to be back in school in the following week. The persistence is consistent with an initial decrease in the pool of infected individuals due to fewer interactions, and it is inconsistent with a temporary decrease in reporting. It is also worth noting that the impact is evident as the winter vacation happens during the typical peak of the epidemic [7].

The policy effectively stops the diffusion of flu among school-age children, and it reduces the exposure of groups at high risk of complications. School-age childrens' symptoms are usually

[5]Note that the overlap between 48 weeks of the epidemic calendar and the weeks of the school year calendar is not perfect. Hence sometimes the vacation may start a couple of days before the week with $T=0$, so I do not exclude the dummy $T=-1$

[6]There is a slight decrease before the start of the vacation due to non-overlapping calendars

[7]A similar exercise is not feasible for summer vacation because the flu level is negligible and all counties have summer recess at the same time every year

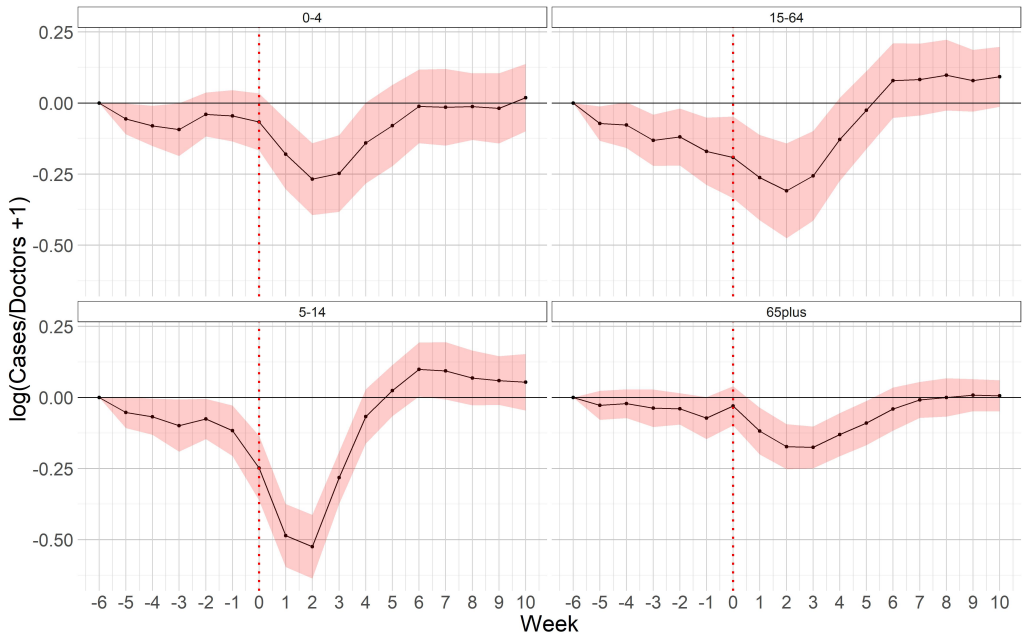


Figure 5: Event study: Flu by age

Note: Each sub-figure represents results of the estimation of equation 2 on a sub sample restricted to age groups 0-4,5-14,15-64, 65+ Line represents estimates of the coefficients γ_T from equation where the outcome is $log(\frac{I_{i,t}}{L_{i,t}} + 1)$. Shared area represents 95% confidence interval for the estimates.

mild and recovery fast. In contrast, very young children and the elderly are at risk of influenza-related complications. Hence, the latter cases are more costly from the healthcare perspective. Fortunately, school closures lower the incidence in high-risk groups too. Figure 5 demonstrates the effect of school closure by age. It shows the decrease in the number of cases among all age groups following the recess. While the largest decrease occurs among school-age children (5-14), we still see a drop in infections among pre-school children (0-4) and the elderly (65+) by about 20-25%. These infections are likely averted by lower within-household transmission from school-age children to other age groups. As the flu in vulnerable populations declines, one should observe a similar drop in flu-related hospitalizations.

The next event study explores the impact of the winter vacation on flu related hospitalizations. It follows the framework of equation 2 where the outcome variable is log of hospitalizations plus 1 $log(H_{i,t} + 1)$. Figure 6 illustrates the estimated coefficients.

While the results are less precise due to the smaller sample size, there is a statistically significant decline in hospitalizations after the winter recess. Doctors send about 5-10% fewer patients to hospitals due to the flu. These results provide evidence of externalities of the policy to groups at risk of complications. Moreover, they show that school closure is effective in limiting healthcare expenditures related to flu.

Furthermore, school closures lead to a decline in deaths related to respiratory diseases among the elderly (65+). As the infections drop in all groups, we would expect the mortality also to go down. Nonetheless, most flu-related deaths are concentrated among the elderly. Hence one expects to see the most prominent effect in this group. To test this hypothesis, I re-estimate the equation 2 where the outcome is the number of deaths due to respiratory causes per 100 000 elderly inhabitants.

The results are presented in the figure 7. Only elderly citizens are affected by the school closures. There is a small yet statistically significant decline in mortality due to respiratory reasons. There are 0.6 fewer deaths per 100 000 in the third and fourth week after the recess. As the time from infection to death may take 1-3 weeks, the timing of the effect is consistent with the impact of school closure. In the figure 11 in the appendix, I show results of a placebo where I estimate the impact of winter vacation on mortality due to external causes. As expected, the effects of the placebo are null.

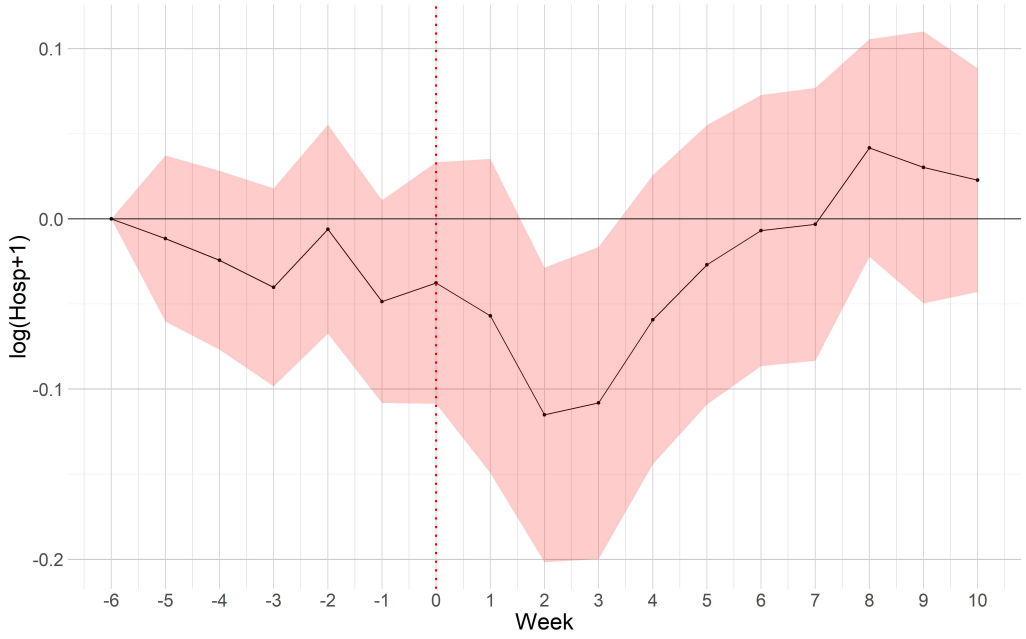


Figure 6: Event Study: Hospitalizations

Note: Line represents estimates of the coefficients γ_T from equation 2 where the outcome is $log(H_{i,t}+1)$. Shared area represents 95% confidence interval for the estimates.

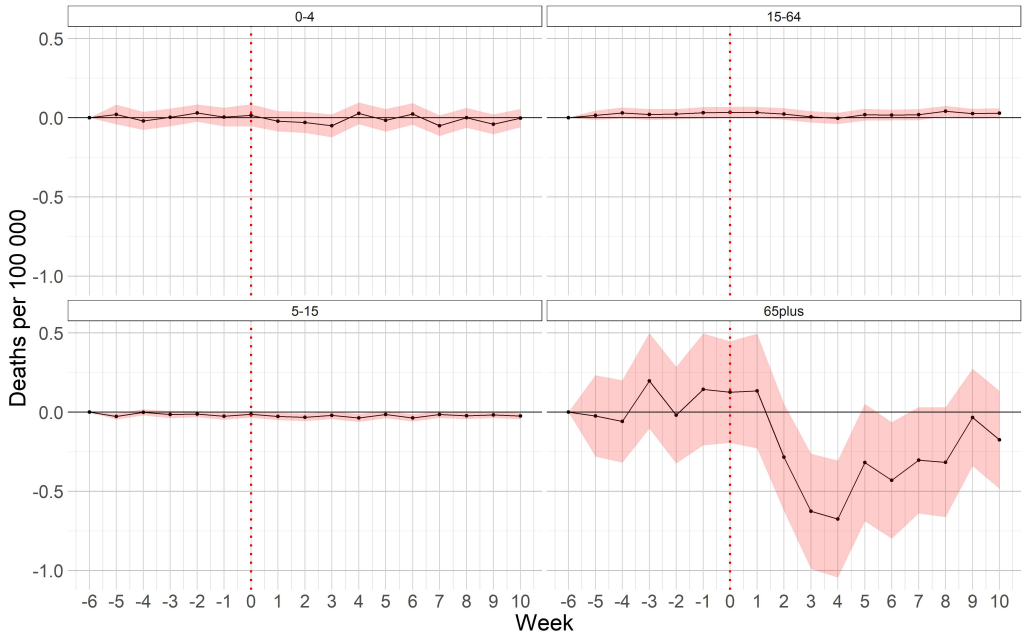


Figure 7: Event study: Mortality by age

Note: Each sub-figure represents results of the estimation of equation 2 on a sub sample restricted to age groups 0-4,5-14,15-64, 65+. The outcome is the number of deaths due to respiratory causes in the given age category in a county i and week t per 100 000 of inhabitants (of given age).

Overall, back of envelope calculation shows that winter vacation help to avert about 635 000 infections, 860 hospitalizations, and 78 deaths in a year[8].

These event studies demonstrate that students' interactions at school contribute considerably to the diffusion of the flu. Moreover, they provide additional evidence that closing schools are an

[8]Results take that there are 11 369 first contact doctors (specialized in family medicine) in Poland who could report flu

effective policy to contain the viral spread and lower healthcare spending. While school closure limits the density of the social network within the county, it does not answer how to stop geographic spreading. Therefore, the following section discusses the role of economic ties in the diffusion of the disease across counties.

5 Role of economic linkages in spreading the flu

The previous section demonstrated the spread of infection within the counties. This section, in turn, is intended to explain how the virus diffuses across counties.

While most daily social interactions concentrate within counties, several routes are enabling high-distance geographic spread. One of them consists of employees commuting to work in distant locations. Two channels arise from this labor mobility. Firstly, workers can get infected in the county of their work and bring the disease to their home county. Secondly, workers can catch the flu in their home county and carry it to their workplace. I will explore empirically whether the diffusion of the virus follows these paths.

The use of the labor mobility data presents some limitations. First, inter-county work commutes may be strongly correlated with other economic linkages such as trade. In the following analysis, I am unable to distinguish between these various channels. Hence the labor mobility serves as a proxy for broadly defined economic interactions. Secondly, the data relied on commutes in 2011. I assume that there is a persistence in the economic linkages between counties.

I adapt the model in equation 1 to investigate the spread of the flu on the labor mobility network [9]. The resulting equation is presented in 3

$$\frac{I_{i,t}}{L_{i,t}} = \widetilde{\alpha}_w \widetilde{S}_{i,t-1} \widetilde{\frac{I_{i,t}}{L_{i,t}}} + \sum_{j \neq i} (\widetilde{\alpha}_{out} w_{ij}^{out} + \widetilde{\alpha}_{in} w_{ij}^{in}) \widetilde{S}_{i,t-1} \widetilde{\frac{I_{j,t-1}}{L_{j,t-1}}} + \delta X_{it} + \varepsilon_{i,t} \tag{3}$$

Here w_{ij}^{out} stands for the number of workers that live in county i and commute to work in county j, defined as outgoing workers. Conversely, w_{ij}^{in} stands for the number of workers that live in county j and commute to work in county i, defined as incoming workers. X_{it} stands for a week, year, region x week, and county fixed effects. Consequently, the parameter $\widetilde{\alpha}_w$ approximates how many susceptible individuals from the county i become infected by one sick person living in the same county. Analogously, $\widetilde{\alpha}_{in} w_{ij}^{in}$ represents the number of new infections in county i produced by one sick person from county j if there are w_{ij}^{in} workers commuting from county j to work in county i. Symmetrically, $\widetilde{\alpha}_{out} w_{ij}^{out}$ represents the number of new infections in county i produced by one sick person in county j if there are w_{ij}^{out} workers commuting from county i to work in county j. The controls X_{it} contain county-week, week-year, year-county fixed effects as well as dummies for winter vacation.

Panel A of table 1 presents results of the OLS estimation of equation 3. One infected person from county i produces about 0.272 new infections within their county. Furthermore, employees commuting to work in other counties seem to bring the disease back home. In contrast, workers do not carry the virus from their homes to their workplace, as the coefficient $\widetilde{\alpha}_{in}$ is statistically indistinguishable from 0.

OLS estimates may be plagued with biases linked to imperfect reporting of the disease incidence and the dynamic specification. In addition to the standard measurement error, reporting may respond to increasing infections in neighboring counties. This would result in non-classical bias. Therefore, I use instrumental variables to remove these potential biases. In particular, I follow Adda [2016], and I instrument the lagged cases per doctor with past environmental condition. There are two types of instruments. The first set of instruments utilized are the temperature, absolute humidity, and sunlight in the week before a report. The second set of instruments measures the number of weeks with low and high temperature, humidity, and sunlight from the beginning of the season. Namely, six instruments are counting the weeks with a temperature below 0,5,10,15,20 and 25 degrees celsius, six instruments are counting the weeks with absolute humidity below 2, 5,

[9] See appendix for the derivation of this equation from the original model. Note that the last term from equation 1 is dropped. This term represents infections from the previous week which are still active. Data shows reports of new infections, so I cannot estimate β

7, 9, 11, and 15 grams of water per m3, and five instruments are counting the weeks with sunlight below 10th, 30th, 50th, 70th, and 90th percentile of sunlight distribution. These instruments aim to capture the evolution of the susceptible population [10]. The term $\widetilde{S}_{i,t-1}\widetilde{\frac{I_{i,t}}{L_{i,t}}}$ is instrumented with past weather as specified above. For the terms $\sum_{j\neq i} w_{ij}^{out}\widetilde{S}_{i,t-1}\widetilde{\frac{I_{j,t-1}}{L_{j,t-1}}}$ and $\sum_{j\neq i} w_{ij}^{in}\widetilde{S}_{i,t-1}\widetilde{\frac{I_{j,t-1}}{L_{j,t-1}}}$ I calculate averages of weather variables counties j weighted by the number of commuters w_{ij}^{in} and w_{ij}^{in}[11]. Subsequently, I re-estimate equation 3 as a two-stage least square. The results are presented in panel B of table 1.

Table 1: Flu and the labor mobility network

	Cases per doctor	
	(A) OLS	(B) IV
$\widetilde{\alpha}_w$	0.272***	1.023***
	(0.04)	(0.05)
$\widetilde{\alpha}_{out}$	0.000 003*	0.000 014***
	(0.000 001 6)	(0.000 005)
$\widetilde{\alpha}_{in}$	0.000 000 5	0.000 002 9
	(0.000 002)	(0.000 007)
Observations	73824	73824

Note: Standard errors in parentheses
$^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

One can notice a considerably higher level of infectivity within the county. One sick person from a county i produces around 1.02 new infections within the same county. Secondly, the transmission from commuters is more pronounced. While the IV coefficient on incoming workers has larger magnitude than its OLS counterpart, it remains statistically insignificant.

These results and the structure of the labor mobility network highlight the role of the regional capitals in spreading the flu. Their central position means that flu readily diffuses to provincial counties along paths of employees commuting to the capital. Moreover, there is an interesting difference between reproduction numbers in capitals and other counties. Reproduction number or R_0 describes the number of secondary cases[12] produced by an infected person. It is possible to infer (a lower bound of) the R_0 from the above estimation. At the median, a sick individual from a provincial county i transmits the virus to:

- 1.02 individuals within the county
- 0.013 individuals in counties which send employees to the county i (median number of incoming worker: 929)

Instead, a sick individual from a county k containing regional capital transmits the virus to:

- 1.02 individuals within the county
- 0.34 individuals in counties which send employees to the county k (median number of incoming worker: 24696)

Hence, an infected individual in a local county produces 1.033 new cases in aggregate while an infected individual in a regional capital produces 1.36 new cases. It may be easy to bring R_0 below 1 in provincial counties. However, it is more difficult in regional capitals. Thus, they play an indispensable role in sustaining the epidemic. Cutting connections to these central counties may be necessary to halt the diffusion.

[10] See appendix tables 2 and 3 for the first stage regression
[11] See appendix for the specification
[12] Secondary cases are infections transmitted by a newly infected individual

6 Conclusion

This paper investigated the role of networks in spreading the flu. It demonstrated a significant contribution of in-school interactions to the diffusion of the virus. Closing schools is an adequate policy in halting the epidemics. They decrease the incidence among students and vulnerable groups such as preschool children and the elderly. Consequently, the hospitalizations and mortality related to flu decline. Hence, school closure appears as a particularly effective response to slow the epidemic and avoid an overload of the healthcare system. Furthermore, the paper highlighted the role of regional capitals. As central actors in a nationwide network of interactions, they are instrumental in sustaining and spreading the disease. Investing in healthcare in these central locations has significant positive externalities. Limiting mobility between these counties and more provincial locations can help to limit the geographic spread of infections.

While this paper investigated the diffusion of the flu, its conclusions apply to other diseases as well. As long as the transmission mechanism is airborne and does not require an intentional act, the discussed diffusion routes and suggested policies curbing the spread are still effective. Furthermore, the COVID-19 epidemic demonstrated that slowing the spread of disease may be the most important goal to avoid an economic downturn.

References

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

D. Almond. 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, 2006.

R. M. Anderson, B. Anderson, and R. M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

A. I. Barreca and J. P. Shimshack. Absolute humidity, temperature, and influenza mortality: 30 years of county-level evidence from the united states. *American journal of epidemiology*, 176 (suppl_7):S114–S122, 2012.

K. Basarbowicz, S. Filas-Przybył, M. Kaźmierczak, K. Kruszka, D. Pawlikowski, D. Stachowiak, and J. Kowalewski. Dojazdy Do Pracy W Polsce, Wyniki NSP 2011. Technical report, Główny Urzad Statystyczny, Warsaw, Poland, 2014. URL `https://stat.gov.pl/statystyka-regionalna/badania-regionalne/dojazdy-do-pracy-w-polsce--wyniki-n`

S. Cauchemez, A.-J. Valleron, P.-Y. Boëlle, A. Flahault, and N. M. Ferguson. Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, 452(7188):750–754, Apr. 2008. ISSN 1476-4687. doi: 10.1038/nature06732. URL `https://www.nature.com/articles/nature06732`.

S. Cauchemez, N. M. Ferguson, C. Wachtel, A. Tegnell, G. Saour, B. Duncan, and A. Nicoll. Closure of schools during an influenza pandemic. *The Lancet infectious diseases*, 9(8):473–481, 2009.

B. J. Cowling, E. H. Lau, C. L. Lam, C. K. Cheng, J. Kovar, K. H. Chan, J. M. Peiris, and G. M. Leung. Effects of school closures, 2008 winter influenza season, hong kong. *Emerging infectious diseases*, 14(10):1660, 2008.

H. Lempel, J. M. Epstein, and R. A. Hammond. Economic cost and health care workforce effects of school closures in the us. *PLoS currents*, 1, 2009.

J. Lessler, N. G. Reich, R. Brookmeyer, T. M. Perl, K. E. Nelson, and D. A. Cummings. Incubation periods of acute respiratory viral infections: a systematic review. *The Lancet infectious diseases*, 9(5):291–300, 2009.

S. Markowitz, E. Nesson, and J. Robinson. The effects of employment on influenza rates. Technical report, National Bureau of Economic Research, 2010.

J. Muñoz Sabater. ERA5-Land hourly data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2019. 10.24381/cds.e2161bac.

E. Oster. Routes of infection: exports and hiv incidence in sub-saharan africa. *Journal of the European Economic Association*, 10(5):1025–1058, 2012.

S. Pei, S. Kandula, W. Yang, and J. Shaman. Forecasting the spatial transmission of influenza in the united states. *Proceedings of the National Academy of Sciences*, 115(11):2752–2757, 2018.

H. Schwandt. The lasting legacy of seasonal influenza: In-utero exposure and labor market outcomes. 2018.

D. Slusky and R. J. Zeckhauser. Sunlight and protection against influenza. Technical report, National Bureau of Economic Research, 2018.

C. Stoecker, N. J. Sanders, and A. Barreca. Success is something to sneeze at: Influenza mortality in cities that participate in the super bowl. *American Journal of Health Economics*, 2(1):125–143, 2016.

L. Su, B. Kidd, A. Han, J. Kotzin, and M. Davis. Virus-Specific CD4+ Memory-Phenotype T Cells Are Abundant in Unexposed Adults. *Immunity*, 38(2):373–383, Feb. 2013. ISSN 1074-7613. doi: 10.1016/j.immuni.2012.10.021. URL `https://www.sciencedirect.com/science/article/pii/S1074761313000526`.

C. C. Wheeler, L. M. Erhart, and M. L. Jehn. Effect of school closure on the incidence of influenza among school-age children in arizona. *Public health reports*, 125(6):851–859, 2010.

C. White. The dynamic relationship between temperature and morbidity. *Journal of the Association of Environmental and Resource Economists*, 4(4):1155–1198, 2017.

7 Appendix

Derivation of the model of the diffusion through the labor mobility network

This section demonstrates how to transform equation (1) into equation (4). Estimation of the equation (1), presented for the ease of reading as equation (8), is impossible

$$I_{t,i} = \alpha d_{ii} \frac{S_{t-1,i}}{P_{t-1,i}} I_{t-1,i} + \sum_{j \neq i} \alpha d_{ij} \frac{S_{t-1,i}}{P_{t-1,i}} I_{t-1,i} - (1-\beta) I_{t-1,i} + \varepsilon_{i,t} \tag{4}$$

The total number of infections is unknown and hence needs to be replaced by the number of cases per doctor. Divide both sides by number of doctors in county i at time t. Moreover, I have only data on new infections, hence I drop the last term.

$$\frac{I_{i,t}}{L_{i,t}} = \alpha d_{ii} \frac{S_{t-1,i}}{P_{t-1,i}} \frac{I_{i,t-1}}{L_{i,t}} + \sum_{j \neq i} \alpha d_{ij} \frac{S_{t-1,i}}{P_{t-1,i}} \frac{I_{j,t-1}}{L_{i,t}} + \varepsilon_{i,t} \tag{5}$$

If one knows the number of doctors, this equation is directly estimable. However, the data for the number of doctors in a county is unavailable. I assume that it is proportional to the population of the county at time t, that is: $L_{it} = cP_{it}$, where c is a constant of proportionality. Using this relationship I can transform (9) into (10):

$$\begin{aligned} \frac{I_{i,t}}{L_{i,t}} = \alpha d_{ii} \frac{S_{t-1,i}}{P_{t-1,i}} \frac{P_{t-1,i}}{P_{t,i}} \frac{I_{i,t-1}}{L_{i,t-1}} + \\ + \sum_{j \neq i} \alpha d_{ij} \frac{S_{t-1,i}}{P_{t-1,i}} \frac{P_{t-1,j}}{P_{t,i}} \frac{I_{j,t-1}}{L_{j,t-1}} + \varepsilon_{i,t} \end{aligned} \tag{6}$$

Note that all the terms in the above equations are in the data and the equation (10) is directly estimable. Now for the ease of notation let

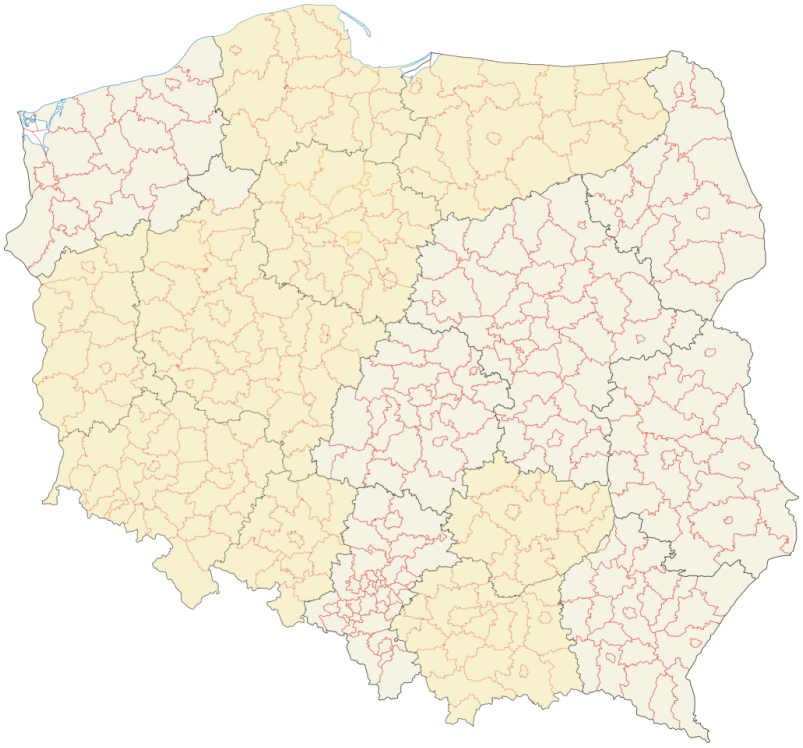


Figure 8: Counties with data available

Note: Shaded areas represent counties which made their data available

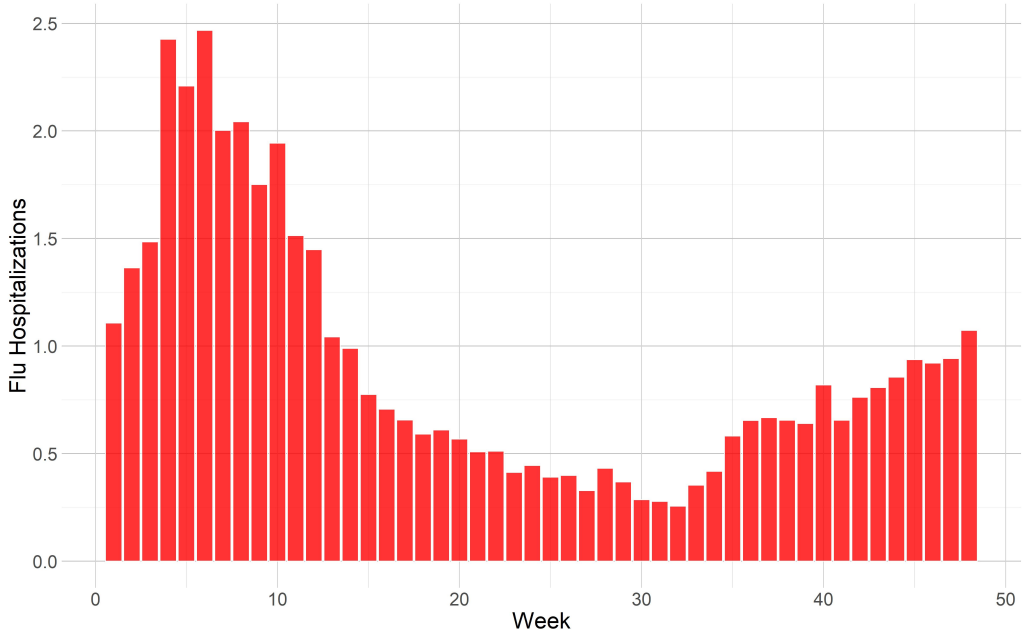
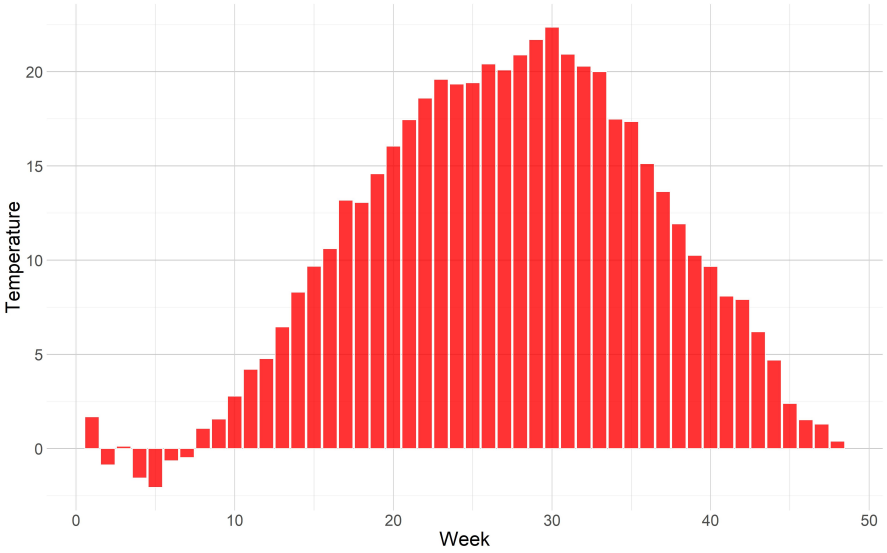


Figure 9: Flu related hospitalizations

Note: Each bar corresponds to an average number of flu related referrals for hospitalization across years and counties

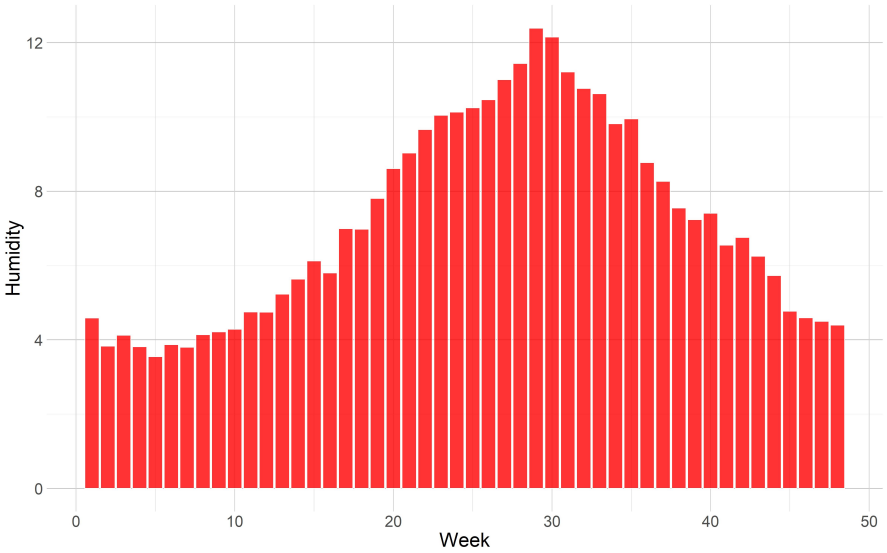
- $\widetilde{\frac{I_{t,i}}{L_{t,i}}} = \frac{P_{t-1,i}}{P_{t,i}} \frac{I_{t-1,i}}{L_{t-1,i}}$
- $\widetilde{S_{t-1,i}} = \frac{S_{t-1,i}}{P_{t-1,i}}$

These quantities are known. Hence we can rewrite (10) as:



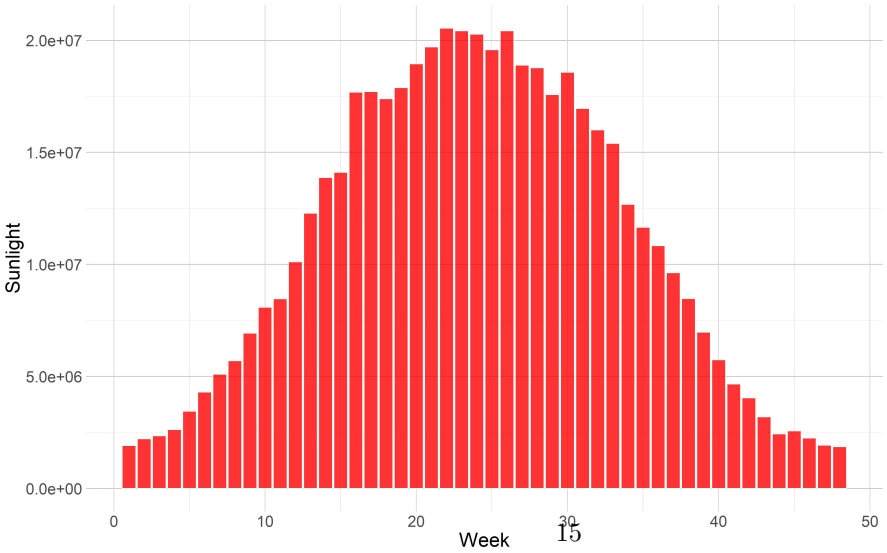
(a) Sunlight

Note: Each bar corresponds to an average daily temperature (measured in Celcius) in a given week



(b) Humidity

Note: Each bar corresponds to an average daily absolute humidity g/m^3 in a given week



(c) Sunlight

Note: Each bar corresponds to an average daily sunlight J/m^2 in a given week

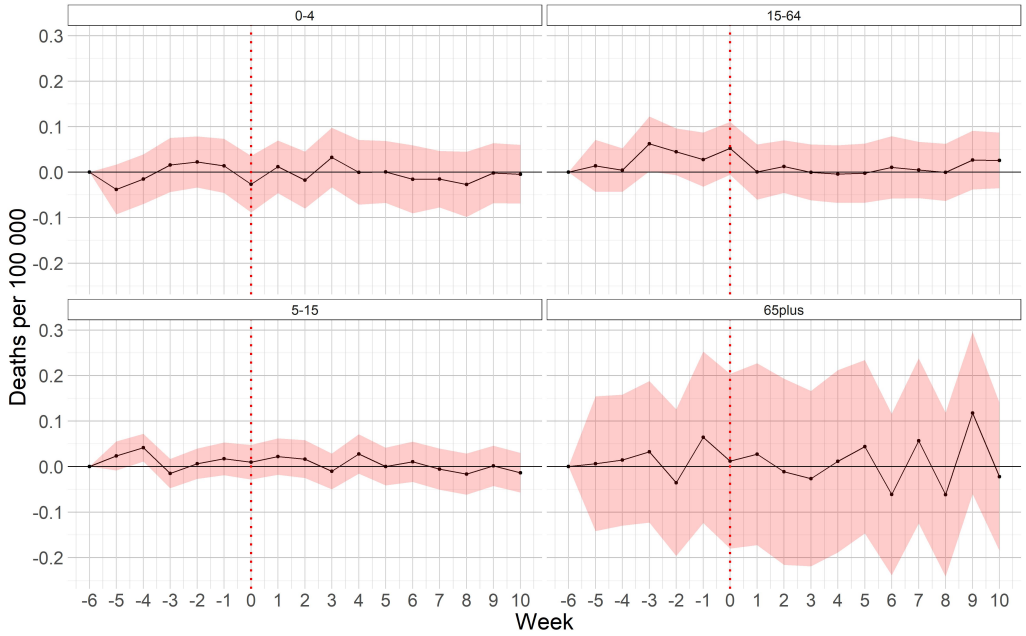


Figure 11: Event study: Mortality due to external causes by age

Note: Each sub-figure represents results of the estimation of equation 2 on a sub sample restricted to age groups 0-4,5-14,15-64, 65+. The outcome is the number of deaths due to external causes in a county i and week t per 100 000. Line represents estimates of the coefficients γ_T from equation where the outcome is $log(\frac{I_{i,t}}{L_{i,t}} + 1)$. Shared area represents 95% confidence interval for the estimates.

$$\begin{aligned}\frac{I_{i,t}}{L_{i,t}} = \alpha d_{ii} \widetilde{S_{t-1,i}} \widetilde{\frac{I_{t-1,i}}{L_{t-1,i}}} + \\ + \sum_{j \neq i} \alpha d_{ij} \widetilde{S_{t-1,i}} \widetilde{\frac{I_{t-1,j}}{L_{t-1,j}}} + \varepsilon_{i,t}\end{aligned} \tag{7}$$

To obtain equation (4), one needs to make the following substitutions in notation: $\alpha d_{ii} = \widetilde{\alpha}_w$ and $\alpha d_{ij} = \widetilde{\alpha}_{out} w_{ij}^{out} + \widetilde{\alpha}_{in} w_{ij}^{in}$. As a consequence we obtain equation (4).

First stage

This section presents the regressions of the number of cases per doctor on the past weather. It serves as the first stage for the estimation in section 5. The first stage for own infections $\widetilde{S_{t,i}} \widetilde{\frac{I_{t,i}}{L_{t,i}}}$ follows the equation (12):

$$\begin{aligned}\widetilde{S_{t,i}} \widetilde{\frac{I_{t,i}}{L_{t,i}}} = \sum_{TL \in t\_lev} \lambda_{TL} weeks\_temp\_below\_TL_{t,i} + \sum_{HL \in h\_lev} \beta_{HL} weeks\_hum\_below\_HL_{t,i} + \\ \sum_{TL \in t\_lev} \lambda_{TL} weeks\_sun\_below\_SL_{t,i} + \sum_{k=1}^{2} \gamma_{Tk} Temp_{t-1,i}^{k} + \sum_{k=1}^{2} \gamma_{Hk} Hum_{t-1,i}^{k} \\ + \sum_{k=1}^{2} \gamma_{Sk} Sun_{t-1,i}^{k} + \alpha X_{t,i} + e_{i,t}\end{aligned} \tag{8}$$

The variable $weeks\_temp\_below\_TL_{t,i}$ counts weeks with temperatures below TL starting from the beginning of the season (September) up to time t. Analogously, $weeks\_hum\_below\_HL_{t,i}$ counts

weeks with humidity below HL and $weeks\_sun\_below\_SL_{t,i}$ counts weeks with sunlight below SL over the same period.

The next six instruments pertain to the weather in the previous week. They stand for temperature, humidity and sunlight during the last week as well as they squared values. Finally, $X_{t,i}$ represents controls and fixed effects. The first stage for the inter-county transmission follows the same logic, except that instruments are weighted averages. That is:

$$\begin{aligned}\sum_{j\neq i} w_{ij}^{out}\widetilde{S_{t,i}}\widetilde{\frac{I_{t,j}}{L_{t,j}}} = &\sum_{TL\in t\_lev}\lambda_{TL}\sum_{j\neq i} w_{ij}^{out} weeks\_temp\_below\_TL_{t,j}+\\ &\sum_{HL\in h\_lev}\beta_{HL}\sum_{j\neq i} w_{ij}^{out} weeks\_hum\_below\_HL_{t,j}+\\ &\sum_{TL\in t\_lev}\lambda_{TL}\sum_{j\neq i} w_{ij}^{out} weeks\_sun\_below\_SL_{t,j} + \sum_{k=1}^{2}\gamma_{Tk}\sum_{j\neq i} w_{ij}^{out} Temp_{t-1,j}^{k}+\\ &\sum_{k=1}^{2}\gamma_{Hk}\sum_{j\neq i} w_{ij}^{out} Hum_{t-1,j}^{k} + \sum_{k=1}^{2}\gamma_{Sk}\sum_{j\neq i} w_{ij}^{out} Sun_{t-1,j}^{k} + \alpha X_{t,i} + e_{i,t}\end{aligned} \tag{9}$$

and

$$\begin{aligned}\sum_{j\neq i} w_{ij}^{in}\widetilde{S_{t,i}}\widetilde{\frac{I_{t,j}}{L_{t,j}}} = &\sum_{TL\in t\_lev}\lambda_{TL}\sum_{j\neq i} w_{ij}^{in} weeks\_temp\_below\_TL_{t,j}+\\ &\sum_{HL\in h\_lev}\beta_{HL}\sum_{j\neq i} w_{ij}^{in} weeks\_hum\_below\_HL_{t,j}+\\ &\sum_{TL\in t\_lev}\lambda_{TL}\sum_{j\neq i} w_{ij}^{in} weeks\_sun\_below\_SL_{t,j} + \sum_{k=1}^{2}\gamma_{Tk}\sum_{j\neq i} w_{ij}^{in} Temp_{t-1,j}^{k}+\\ &\sum_{k=1}^{2}\gamma_{Hk}\sum_{j\neq i} w_{ij}^{in} Hum_{t-1,j}^{k} + \sum_{k=1}^{2}\gamma_{Sk}\sum_{j\neq i} w_{ij}^{in} Sun_{t-1,j}^{k} + \alpha X_{t,i} + e_{i,t}\end{aligned} \tag{10}$$

Table 1A demonstrates the F test on the significance of the excluded instruments.Table 2A shows the coefficients from the first stage. For column 2 and 3, covariates are weighted averages as in equation 13 and 14.

Table 3: F test: joint significance of excluded instruments

	F statistic	p-value	Df	Res. Df
$\widetilde{S_{t,i}}\widetilde{\frac{I_{t,i}}{L_{t,i}}}$	9.6^{***}	0	23	63764
$\sum_{j\neq i} w_{ij}^{out}\widetilde{S}_{i,t}\widetilde{\frac{I_{j,t}}{L_{j,t}}}$:	30.09^{***}	0	23	63764
$\sum_{j\neq i} w_{ij}^{in}\widetilde{S}_{i,t}\widetilde{\frac{I_{j,t}}{L_{j,t}}}$:	34.6^{***}	0	23	63764

The first stage for the section 6 follows a different formulation. It predicts the number of infections per doctor in each county as in equation 15. Then, these predictions appear in the regressions in section 6.

Table 2: First stage regression

	Dependent variable:		
	$\widetilde{S_{t,i}\frac{I_{t,i}}{L_{t,i}}}$	$\sum_{j\neq i} w_{ij}^{out}\widetilde{S}_{i,t}\widetilde{\frac{I_{j,t}}{L_{j,t}}}$	$\sum_{j\neq i} w_{ij}^{in}\widetilde{S}_{i,t}\widetilde{\frac{I_{j,t}}{L_{j,t}}}$
	(1)	(2)	(3)
Temp(K)_{t-1}	1.602	1.049^{***}	0.321^{***}
	(4.256)	(0.239)	(0.057)
Sun_{t-1}	-0.00000^{*}	3.142^{**}	-0.00000
	(0.00000)	(1.417)	(0.00000)
Hum_{t-1}	0.185	-0.004^{***}	1.135^{***}
	(1.193)	(0.001)	(0.288)
Temp(K)^2_{t-1}	-0.003	-0.107^{*}	-0.001^{***}
	(0.008)	(0.062)	(0.0002)
Sun^2_{t-1}	0.000	-0.000^{***}	-0.055^{***}
	(0.000)	(0.000)	(0.012)
Hum^2_{t-1}	-0.005	0.00000^{***}	0.000
	(0.058)	(0.00000)	(0.000)
Weeks Sun$< 10^{th}$	-0.011	3.631^{***}	0.137
	(0.212)	(0.411)	(0.107)
Weeks Sun$< 30^{th}$	0.240	1.412^{***}	-0.044
	(0.197)	(0.424)	(0.097)
Weeks Sun$< 50^{th}$	-0.523^{**}	0.049	0.701^{***}
	(0.205)	(0.374)	(0.081)
Weeks Sun$< 70^{th}$	-0.084	1.530^{***}	1.005^{***}
	(0.224)	(0.379)	(0.085)
Weeks Sun$< 90^{th}$	0.194	-4.173^{***}	0.614^{***}
	(0.274)	(0.580)	(0.144)
Weeks Hum< 2	-3.604^{***}	1.478	1.405^{***}
	(0.610)	(1.008)	(0.199)
Weeks Hum< 5	-0.199	2.334^{***}	0.703^{***}
	(0.178)	(0.339)	(0.084)
Weeks Hum< 7	1.026^{***}	1.664^{***}	-0.125
	(0.202)	(0.414)	(0.104)
Weeks Hum< 9	-0.267	-2.731^{***}	-0.035
	(0.183)	(0.363)	(0.089)
Weeks Hum< 11	-0.193	1.835^{***}	-0.578^{***}
	(0.230)	(0.438)	(0.104)
Weeks Hum< 15	0.642	-0.421	-1.369^{***}
	(0.721)	(1.284)	(0.290)
Weeks Temp(C)< 0	-0.281	-1.063^{***}	-0.219^{***}
	(0.241)	(0.276)	(0.059)
Weeks Temp(C)< 5	0.363^{*}	-0.499	0.046
	(0.216)	(0.389)	(0.087)
Weeks Temp(C)< 10	-0.220	0.008	-0.520^{***}
	(0.249)	(0.473)	(0.103)
Weeks Temp(C)< 15	0.768^{***}	1.425^{***}	0.052
	(0.198)	(0.361)	(0.095)
Weeks Temp(C)< 20	-0.073	3.591^{***}	0.239^{**}
	(0.251)	(0.441)	(0.109)
Weeks Temp(C)< 25	-0.327	-1.808	-0.025
	(0.586)	(1.139)	(0.283)
Note:			*p<0.1; **p<0.05; ***p<0.01

$$\frac{I_{t,i}}{L_{t,i}} = \sum_{TL \in t\_lev} \lambda_{TL} weeks\_temp\_below\_TL_{t,i} + \sum_{HL \in h\_lev} \beta_{HL} weeks\_hum\_below\_HL_{t,i} + \sum_{TL \in t\_lev} \lambda_{TL} weeks\_sun\_below\_SL_{t,i} + \sum_{k=1}^{2} \gamma_{Tk} Temp^{k}_{t-1,i} + \sum_{k=1}^{2} \gamma_{Hk} Hum^{k}_{t-1,i} + \sum_{k=1}^{2} \gamma_{Sk} Sun^{k}_{t-1,i} + \alpha X_{t,i} + e_{i,t} \quad (11)$$

7.1 Robustness checks

7.1.1 Initial immunity

The preferred specification assumes that everyone is susceptible at the beginning of a new season (i.e. at the end of a summer). However, some studies (ex: Su et al. [2013]) point out that individuals can form immunity without being exposed to the disease[13]. Hence, as a robustness check, I test the transmission equation with some levels of pre-existing immunity. Namely, I assume that 5%,10%, or 15% of the population has immunity at the beginning of the season.

Table 4: Transmission with non-zero initial immunity

Initial immunity	Cases per doctor 5%		10%		15%	
	OLS	IV	OLS	IV	OLS	IV
$\widetilde{\alpha}_w$	0.28***	1.078***	0.30***	1.14***	0.32***	1.21***
	(0.35)	(0.054)	(0.38)	(0.06)	(0.04)	(0.65)
$\widetilde{\alpha}_{out}$	0.000003**	0.000014***	0.000004*	0.000014**	0.000004*	0.000015**
	(0.0000017)	(0.0000052)	(0.0000018)	(0.0000057)	(0.0000019)	(0.0000061)
$\widetilde{\alpha}_{in}$	0.0000006	0.000003	0.0000006	0.000003	0.0000006	0.000003
	(0.0000026)	(0.000008)	(0.0000027)	(0.000009)	(0.0000028)	(0.00001)
Observations						73284

Note: Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3A demonstrate the results of this robustness check. The parameters increase slightly with the pre-existing immunity level, but the qualitative conclusions remain unchanged.

7.1.2 Event study: impact on transmission

This event study looks at the impact of the winter vacation on the transmission of the virus. It measures the impact of the break on the interactions which fuel the spread of the virus. School recess disrupts interactions in weeks 0 and 1 (recess lasts two weeks). Hence the virus will have a lower reproduction rate from week 0 to 1 and from week 1 to 2. The event study follows the equation:

$$\frac{I_{i,t}}{L_{i,t}} = \sum_{T=-5}^{10} \alpha_T 1_{i,t-T} * \widetilde{S_{t-1,i}} \widetilde{\frac{I_{t-1,i}}{L_{t-1,i}}} + \alpha_0 \widetilde{S_{t-1,i}} \widetilde{\frac{I_{t-1,i}}{L_{t-1,i}}} + \delta X_{it}. + e_{i,t} \quad (12)$$

Figure 15 shows the results. Vacation clearly disrupts the transmission of the flu, as evidenced by the high decline in the reproduction of the virus in weeks 1 and 2. In other words, the same amount of infections results in fewer secondary cases in weeks 1 and 2 because of fewer interactions

[13] I thank Jeffery Shaman for that remark

among children in respective previous weeks. As expected, the parameter returns to the pre-vacation value three weeks after the vacation as children start interacting again at a usual rate in week 2.

Figure 12: Impact of winter vacation on transmission

Note: Line represents estimates of the coefficients α_T from equation 12. Shared area represents 95% confidence interval for the estimates.