Vax Populi: the Social Costs of Online Vaccine Skepticism[*]

Matilde Giaccherini[1,4], Joanna Kopinska[2,1], and Gabriele Rovigatti[3]

[1]*CEIS, University of Rome "Tor Vergata"*

[2]*University of Rome "La Sapienza"*

[3]*Bank of Italy*

[4]*CESifo*

June 2, 2023

Abstract

This paper quantifies the impact of online vaccine skepticism on pediatric vaccine uptake and health outcomes. We propose a methodology that combines Natural Language Processing and an instrumental variable strategy that leverages the complex structure of social networks. By matching Italian vaccine-related tweets for 2013-2018 with vaccine coverage and preventable hospitalizations data, we find that a 10pp increase in anti-vaccine sentiment causes a 0.43pp decrease in Measles-Mumps-Rubella vaccine coverage, additional 2.1 hospitalizations per 100,000 residents, and an 11% increase in relevant healthcare expenses. The results of a simulated model further show the importance of targeted interventions to counter misinformation.

JEL Classification: I18, D85, L82, C81

Keywords: Social network, Twitter, vaccines, controversialness, polarization, text analysis.

[*]Corresponding author - Giaccherini: matilde.giaccherini@uniroma2.it. Kopinska: joanna.kopinska@uniroma1.it. Rovigatti: gabriele.rovigatti@bancaditalia.it. We would like to thank Sofia Amaral-Garcia, Tiziano Arduini, Federico Belotti, Vincenzo Carrieri, Francesco Decarolis, Ruben Durante, Ludovica Gazze, Leonardo Giuffrida, Arthur Lewbel, Sauro Mocetti, Tommaso Orlando, Francesco Pierri, Monou Prem, Francesco Sobbrio, Lucia Rizzica, Lorenzo Rovigatti, Federico Rossi, Gianluca Russo, for useful suggestions and the participants of the ASSA Annual Meeting 2023, CESifo Area Conference in Economics of Digitization 2022, EuHEA Seminar 2022, SEHO Conference 2022. Special thanks go to Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini for sharing their code before its publication. The views expressed in this paper are those of the authors and do not involve the responsibility of the Bank of Italy. The usual disclaimers apply. The authors acknowledge the financial support of the Einaudi Institute for Economics and Finance (EIEF) and the computational support of ISCRA/CINECA.

1 Introduction

The phenomenon of misinformation is deeply ingrained in contemporary society, impacting political, economic, and social well-being (Vosoughi et al., 2018). Before the COVID-19 era, a supposed link between pediatric vaccines and autism was one of the most propagated fake news, stemming from A. Wakefield's 1998 Lancet article on the trivalent Measles-Mumps-Rubella (MMR) vaccination (Jolley and Douglas, 2014, Leask et al., 2006, Opel et al., 2011). Although the article has been retracted, and despite overwhelming evidence supporting the safety and efficacy of vaccines, this piece of disinformation remains widespread (see among others Allcott et al., 2019, Chiou and Tucker, 2018).

The diffusion of the internet and, more recently, the rise of social media have provided an unparalleled platform for the dissemination of similar viewpoints on vaccines (Burki, 2019).[1] These platforms have granted virtually unlimited access to information that is not subjected to fact-checking or editorial judgment. As a result, the ability of consumers to discriminate between true and "fake" (or unsubstantiated) news has decreased (Lazer et al., 2018, Sunstein, 2001, 2017, 2018). Additionally, the dynamics of social networks tend to foster the formation of ideological "echo chambers" (Cinelli et al., 2021, Flaxman et al., 2016), which fuel polarization (Azzimonti and Fernandes, 2022), ideological self-segregation (Berinsky, 2017, Gentzkow and Shapiro, 2011, Mullainathan and Shleifer, 2005), and misinformation diffusion (Allcott and Gentzkow, 2017).

Due to the spread of false claims about the vaccine-autism link, an increasing number of parents choose not to vaccinate their children, relying instead on the herd immunity provided by vaccinated peers (Esposito et al., 2014, Smith et al., 2017). In Italy, as in other countries, decreasing vaccination rates have led to reemergence of previously controlled diseases such as measles. This has sparked a policy debate and prompted the implementation of legal measures that impose costs on individuals who choose not to vaccinate. Although vaccine mandates curtailing individual freedom of choice have always been controversial, the Italian healthcare department has argued that falling uptake poses a risk not only to those who are eligible for vaccination but also to vulnerable individuals who are unable to receive vaccines, such as infants aged 0-12 months, pregnant women, and immunosuppressed patients.

In this ongoing conflict between personal interests and public health endeavors, a crucial aspect to consider is the influence of online vaccine skepticism on vaccination rates and vaccine-preventable diseases. If skepticism spread through social media has a sizeable impact on vaccine hesitancy, addressing it could help individuals make better decisions in their own best interest. Furthermore, given that communicable diseases can impose significant externalities on society, including health risks, increased hospitalization rates and costs

[1] In addition, the fact-checking standards on social media are often lax, and the emotional appeal of such messages can contribute to their rapid and widespread dissemination (Zhuravskaya et al., 2020).

for individuals who are not targeted by vaccination campaigns, a comprehensive analysis should consider these additional burdens placed on the community as a whole.

We empirically quantify the effects of exposure to vaccine skepticism on local public health outcomes such as vaccination rates, vaccine-preventable hospitalizations, and their relative costs. We focus on Italy during the 2013-2018 period, which coincided with a major reform in 2017 that expanded the pediatric vaccine mandate to include the MMR vaccine.[2] The implementation of the new law followed a period of declining vaccine coverage and was preceded by intense debates both offline and online.

To examine the spread of vaccine skepticism, we utilize Twitter data, specifically focusing on Italian tweets related to vaccines. By employing a Natural Language Processing (NLP) algorithm similar to that used by Polignano et al. (2019), we develop an anti-vaccine (hereafter *anti-vax*) classifier to determine the stance expressed in the tweets. Twitter data has been shown to accurately reflect public attitudes towards policy-relevant topics across different locations and over time, as demonstrated by Grossman et al. (2020), Jin et al. (2021), Kim (2022). Using this dataset, we calculate the average anti-vax sentiment expressed in geolocated tweets as a proxy for the relevance and spread of the anti-vaccine movement within Italian municipalities. This allows us to examine the relationship between vaccine-related attitudes and public health outcomes, such as vaccination rates, vaccine-preventable hospitalizations, and associated costs at the local level.

When estimating the causal relationship between exposure to anti-vaccine views on social media and the health outcomes associated with vaccine hesitancy, we encounter two challenges: i the endogeneity inherent in this relationship, and ii the lack of data on individual-level vaccine hesitancy.

To address the first issue of endogeneity and formalize its sources, we employ a model of opinion dynamics in social networks, inspired by the work of Baumann et al. (2020), to examine the evolution of individual vaccine stances on Twitter. We show that even moderate degrees of homophily, when combined with highly controversial topics, endogenously result in the formation of echo chambers and opinion polarization. Our dynamic model emphasizes two complementary effects: the "*link formation effect*", whereby users are more inclined to share content and establish connections with individuals who hold similar beliefs, and the "*exposure effect*", whereby users' stances are influenced by the opinions they are exposed to, particularly those at the extreme ends of the spectrum. While the former effect determines the extent of exposure to anti-vaccine content endogenously based on users' own stances, the latter effect is the primary focus of our investigation.

To estimate the exposure effect, we employ an instrumental variables approach informed by the network interaction literature (see e.g. Bramoullé et al., 2020, Cagé et al., 2022, De Giorgi et al., 2010). Our approach

[2] Until 2017, Italy required only four vaccines, and the mandate was rarely enforced. Vaccines for MMR, chickenpox, meningococcal, and pneumococcal were strongly recommended but not mandatory, leaving the decision to parents. The legal enforcement of mandatory pediatric vaccines upon school enrollment was implemented in late 2017, with a one-year transitional period to facilitate parental compliance with the new regulations.

relies on the existence of a subset of network connections that are unlikely to be influenced by homophily in vaccine stances. First, for each user i we construct the complete network of her *followings* at the end of the sample period, when all link formation effects have taken place.[3] Second, we focus on the connections that were not formed due to homophily in vaccine stances, referred to as "passive connections" (i.e., connections that do not engage in vaccine-related discussions). Finally, we define each user i's followings-of-passive-followings (FoPF) network as the set of users followed by their passive connections j.[4] Under standard assumptions, such FoPF average exposure constitutes an exogenous source of variation that we exploit as an instrumental variable for individuals' stances.

To overcome the lack of individual data on vaccine hesitancy and its outcomes, we pair aggregated vaccine-related Twitter stances with disease-specific vaccine coverage rates, vaccine-preventable hospitalizations, and relative costs at the municipal level. This allows us to harness both the power of the individual-level Twitter data and the highly detailed municipal data on vaccinations, hospitalizations and health-related costs through a Mixed Two-Stage Least Squares (M2SLS) approach (Dhrymes and Lleras-Muney, 2006). Building on the individual-level first-stage regression, we aggregate the instrumented variable at the municipal level to obtain a valid causal estimate of the exposure effect.

Our estimates show that exposure to online vaccine skepticism causes a significant reduction in vaccination rates for the MMR, particularly targeted by anti-vax misinformation. We find no statistically significant impact on vaccines not affected by fake news (Hexavalent, Meningococcal, Pneumococcal). A 10 pp increase in average vaccine skepticism at the municipality level leads to a 0.43 pp decrease in vaccination coverage (mean value 89.50). Furthermore, vaccine skepticism leads to higher rates of hospitalization for vaccine-preventable diseases and increased healthcare costs. Specifically, a 10 pp increase in the average stance leads to 2.1 additional hospitalizations per 100,000 residents (mean value 22) and an excess expenditure of 7,311 euros, representing an 11% increase in the relevant healthcare expenses.

To ensure the robustness of our findings, we control for the impact of Twitter algorithm changes,[5] local vaccine campaigns, and the impact of populist votes, finding virtually unchanged results. In addition to the baseline analysis, we propose an alternative estimation strategy to address potential concerns about the exogeneity of our preferred instrument. The results are comparable to the baseline, both in terms of magnitude and statistical significance.

[3] There are two categories of relationships on Twitter: "followers" and "followings". "Followers" refers to Twitter users who follow user i, whereas "followings" represents the Twitter users that user i follows and whose content she is directly exposed to. It's important to note that these connections can be unilateral and do not necessarily have to be reciprocated.

[4] Among the FoPF network, we exclude users who have a mutual relationship with user i, as well as direct connections with any "active" following.

[5] Twitter introduced an "ampliphication algorithm" in 2016. As argued by Acemoglu et al. (2021), such algorithms are aimed at maximizing engagement and tend to create more homophilic communication patterns, or "filter bubbles."

Finally, we examine the implications of our findings for policymakers and public health organizations in terms of actions on social media to effectively engage with the public. We first investigate the potential non-linearity in the exposure effects on individual user stances. Specifically, we examine whether pro- or anti-vaccine individuals react differently to the exposure to FoPF content. We find that pro-vaccine users exhibit a stronger "persuasion" effect compared to anti-vaccine users. This implies that interventions aimed at retaining individuals with doubts about vaccines may be more effective than efforts focused on convincing ardent anti-vax supporters. Second, in the spirit of Athey et al. (2022), we exploit the exogenous timing of events like epidemics, scientific breakthroughs, court rulings, legislation, and news to test whether the "type" of event influences the strength of the exposure effect.[6] Our results show that political events and news originating from national or international institutions (considered trustworthy sources) support pro-vaccine stances and weaken the exposure effect of anti-vax content.

Additionally, we conduct Monte Carlo counterfactual analyses by simulating two alternative scenarios of our dynamic model. The first scenario involves implementing a *Censorship* policy targeting anti-vax content, while the second scenario involves running vaccine *Informative Campaigns*. We find that informative campaigns are the most efficient approach to counteract the effects of online misinformation and reduce polarization. These findings imply that social media vaccine awareness campaigns may be a practical and scalable intervention to increase understanding of public health issues and contain the spread of misinformation.

While a growing body of literature examines the effects of fake news on vaccine hesitancy (Carrieri et al., 2019, Chiou and Tucker, 2018), anti-vaccine beliefs and behavior (Allam et al., 2014), and improving immunization (Alatas et al., 2019), to the best of our knowledge, this is the first paper that jointly i uses detailed data at a fine-grained geographical level on vaccination rates and hospitalizations, ii provides a data-driven approach to proxy users' stances toward vaccine-related topics, iii implements a causal identification strategy at the user level, and most importantly, iv quantifies the monetary costs of online vaccine skepticism, distinguishing between the target population and the externalities for the vulnerable individuals not subject to the vaccination campaigns.

We also contribute to the small but growing literature on the tangible effects of social media content on offline communities, pioneered by Bond et al. (2012). Prominent contributions include Enikolopov et al. (2020), who show that social media actually alleviate the collective action problem by lowering coordination costs, Bursztyn et al. (2019), who estimate the causal effect of social media penetration on ethnic hate crimes and xenophobic attitudes, and Allcott et al. (2020), who run a large-scale experiment showing that reducing online activity increases wellbeing. More recently, Guriev et al. (2021) investigate the effects of the mobile broadband

[6] We classify events into four broad categories: vaccine efficacy, statements from trustful institutions, politics and mandates, and allegations of vaccine unsafety.

expansion on confidence in the government, and argue that the access to social media platforms might play a prominent role in empowering anti-establishment politicians. In addition, social media platforms expose voters to false information and increase their overconfidence, deteriorating the quality of democratic choice (Kartal and Tyran, 2022). Finally, Cagé et al. (2022), Draca and Schwarz (2021), Müller and Schwarz (2020), Qin et al. (2017) show that social media are associated with real-life social effects spanning from news production and spread to hate crimes.

This work also contributes to the literature on the effects of vaccine mandates. Previous research has shown that mandates can significantly impact vaccination uptake and decrease the incidence of infectious diseases, such as pertussis, smallpox, chickenpox, and hepatitis A, with large long-term effects on affected individuals (Abrevaya and Mulligan, 2011, Carpenter and Lawler, 2019, Holtkamp et al., 2021, Lawler, 2017). Our results suggest that counteracting the spread of pediatric vaccine skepticism can have a significant impact on immunization. Forced medical interventions are often seen as curtailments of individual freedom, which can lead to controversy and unintended consequences. Athey et al. (2022) have recently shown that social media had a significant impact on self-reported beliefs and knowledge about COVID-19 vaccines through public health organization campaigns on Facebook and Instagram. The results of the study conducted by Larsen et al. (2022) showed that using a counterstereotypical messenger on social media[7] can be a powerful catalyst in encouraging COVID-19 vaccine uptake among the hesitant. Additionally, Breza et al. (2021) have found that mobility and COVID-19 infection rates decreased as a result of randomly assigned exposure to Facebook messages encouraging preventive health behaviors. Bailey et al. (2020) have also shown that Facebook users with friends exposed to COVID-19 were more likely to support social distancing and other public health behavior measures. Our findings provide direct evidence of the potential benefits of policies aimed at raising awareness of the risks of communicable diseases and promoting preventive immunization to combat the effects of vaccine skepticism on public health.

2 Institutional background

The advances in vaccine technology have been major contributors to the increases in life expectancy that characterized the 19^{th} and 20^{th} centuries. Paradoxically, due to the past success of collective vaccination efforts, individuals tend to underestimate the value of immunization and are more willing to risk being unprotected. Additionally, the "self-eroding" nature of vaccination can lead to fluctuations in vaccine coverage for newborns, which, in turn, affects the level of protection for the entire community when herd immunity is not achieved (Siegal et al., 2009). In this sense, vaccine uptake can be seen as an example of a free-riding problem, where

[7] They used a Youtube "public service" announcement featuring Donald Trump encouraging his supporters to get vaccinated.

individuals may prioritize their own interests over those of the community when deciding whether or not to get vaccinated. Consequently, this can lead to cycles of suboptimal participation in vaccination campaigns.

One of the turning points in the history of Italian vaccine campaigns was the eradication of smallpox between 1978 and 1998, followed by the introduction of hepatitis B and anti-pertussis vaccines. In the early 2000s, the first national vaccination plans were introduced under the National Plan of Vaccine Prevention (PNPV). The PNPV establishes a vaccine calendar and offers eligible individuals free vaccines at Local Health Authorities (LHAs).[8]

Until 2017, four vaccines were mandatory for children: polio, diphtheria, tetanus, and hepatitis B. They were often combined with haemophilus influenzae type b and whooping cough into a 6-in-1 vaccine known as hexavalent. Vaccines for the trivalent MMR, chickenpox, meningo- and pneumo-coccal diseases were only strongly recommended. In 2012, a local court in Rimini issued a sentence against the Health Ministry, falsely claiming a causal link between the MMR vaccine and autism. This decision had a detrimental impact on immunization rates, leading to a decline that reached its lowest point in 2015, a year in which the local court of Bologna reversed the controversial Rimini sentence (Carrieri et al., 2019). In response to the declining immunization rates and a significant rise in measles cases in Italy, there was a strong political commitment to counter anti-vaccine movements. This commitment resulted in the approval of a new PNPV in 2017, known as the "*Lorenzini*'s Decree", which extended the scope of mandatory pediatric vaccines by making them a requirement for school enrollment and introduced stricter penalties for doctors who promoted anti-vaccine views.

Under the 2017 PNPV, the number of mandatory vaccinations increased from four to ten (adding whooping cough, Haemophilus influenzae type b, measles, mumps, rubella, and chickenpox). Although vaccine mandates curtailing individual freedom have always been disputed, the PNPV proposers argued that the declining vaccine coverage rate, driven by anti-vaccine sentiment, resulted in significant negative externalities, increasing the risk of infection not only for the eligible population but particularly for vulnerable individuals not targeted by the vaccination.

3 Data

We gathered data on vaccine skepticism and relevant discussions from Twitter. To do this, we used the Twitter Application Programming Interface (API) to retrieve all publicly available tweets written in Italian that con-

[8] The regional authorities implement public health policies through their health departments, while health protection and promotion fall under the responsibility of the Departments of Prevention within the 101 LHAs. LHAs cover on average 590,000 individuals each and are divided into 711 districts with an average population of 84,000. LHAs manage and deliver vaccinations free of charge to eligible populations, including the pediatric population, the elderly, and other protected categories.

tained vaccine-related keywords and a wide range of information on users for the period from 2013 to 2018. Furthermore, we hand-collected news-related data from newspapers and official sources of information covering various topics related to vaccines, such as vaccine-preventable disease outbreaks, legal cases, court rulings, and regulatory interventions at both local and national levels.

In terms of health data, we rely on two main sources. The first dataset contains annual information on disease-specific vaccination rates provided by the LHAs, aggregated at the municipal level for the period 2013-2018. The second source is an administrative dataset on all hospital admissions in Itay, provided by the Italian Ministry of Health. This dataset allows us to examine vaccine-preventable conditions in both the target population and the population exempted from the vaccination plan, such as infants of 0-12 months, pregnant women, and immunosuppressed patients, aggregated at the municipality/year level for the period 2013-2016.

3.1 Twitter data

Twitter ranks as the fourth most popular social media platform in Italy, following Facebook, Instagram, and LinkedIn, with 8 million unique users in 2018. Alongside TikTok, it boasts the fastest-growing user base. The demographics of Twitter users tend to skew towards older age groups, with 39% of users identifying as female. Notably, Twitter serves as a hub for news outlets, TV channels, and blogs, emphasizing its role in disseminating information. Furthermore, Twitter's influence extends beyond its own platform, as its content is often shared across other social media platforms. In fact, a significant portion of Twitter users is also present on Facebook (84%), YouTube (80%), and Instagram (88%).[9]

To harness the significance of Twitter data in information dissemination, we leverage the *Academic Research product track* to access the complete archive of tweets that have not been deleted. In addition to the tweet text, the API provides information about the tweets and the associated users. Apart from textual analysis, we focus on two aspects: geolocation mapping of users and analysis of their online networks. Geolocation data provides insights into the contexts in which target populations reside (Martinez et al., 2018). Through the API, we retrieve the complete list of users that each vocal user follows and is followed by, enabling the construction of user-specific online networks.[10] This approach facilitates the study of the interplay between users' Twitter conversations and their local environments.

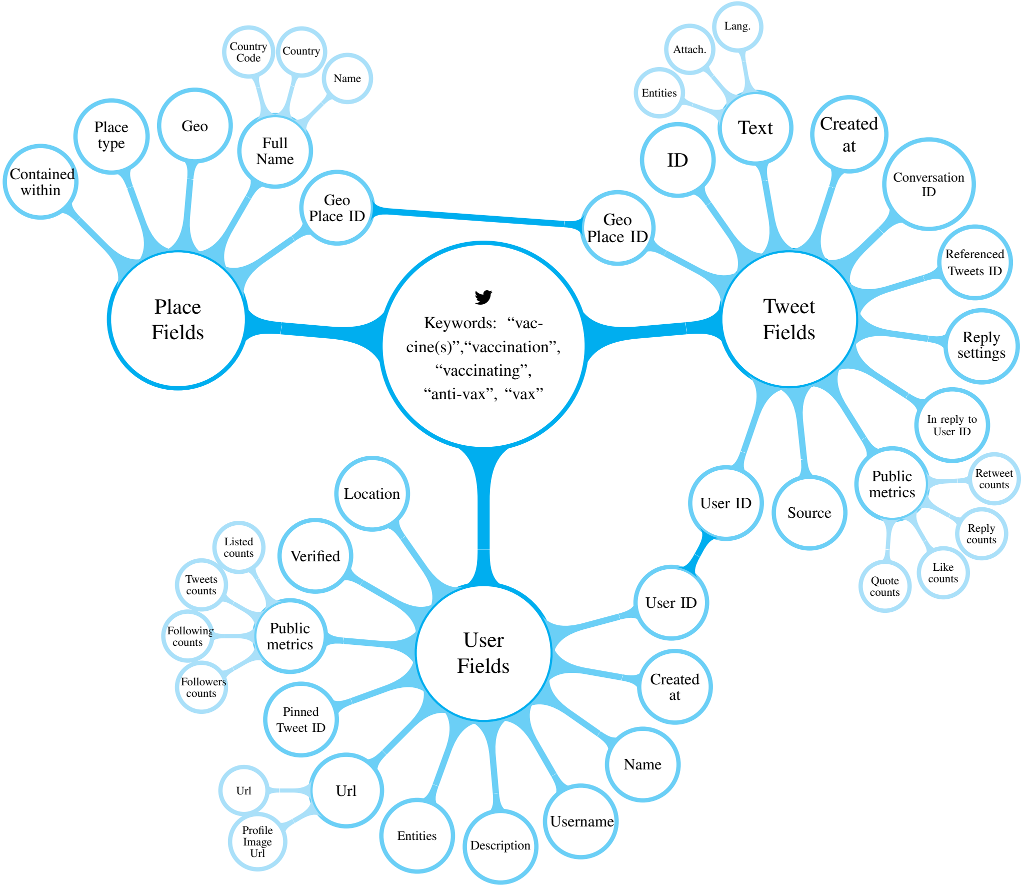
[9] Data from the *Authority for Communications Guarantees* (AgCom).

[10] As of now, Twitter API v.2 permits retrieval of the following/follower structure at the date of download, which, in our case, spans from May to September 2021. Consequently, we establish network-related variables based on the "equilibrium" network, encompassing all endogenous interactions among users during the 2013-2018 analysis period.

Download and filtering. We collected all tweets that contained the Italian equivalents of the following keywords: “vaccine(s)”, “vaccination”, “vaccinating”, “anti-vax”, “vax”, for a total of 2,031,448 observations.[11] The current version of the dataset was downloaded on April 23^{rd}, 2021.

Each retrieved object consists of the following information: i) the plain text of the tweet; ii) the unique tweet ID, creation date, counts of associated replies, likes, mentions, retweets, hashtags, and multimedia contents, as well as tweet-specific location if available, and iii) user details including ID, Twitter handle, display name, short bio, and several metrics such as the number of followings, followers, and tweets posted, the verified status of the account, date of Twitter account creation, and user location if available (see Figure 1).[12]

Figure 1: Twitter objects



Notes: Structure of Twitter objects returned by the API. The structure includes: i) Place fields, ii) Tweet fields and iii) User fields. The former two are matched through the Geo place ID, the latter two are univocally connected through the User ID.

We also collect information on the *followers* and *followings* of users. *Followers* refer to Twitter users who follow a specific user, while *followings* are the Twitter users that a specific user follows and whose content she is

[11] To ensure data relevance, we excluded tweets referring to cow milk (“latte vaccino” in Italian). The query used was “(vaccino OR vaccini OR vaccinazione OR vax OR novax OR vaccinarsi OR vaccinato OR vaccinati) -mozzarella -latte lang:it”.

[12] This study does not include any personally identifiable information.

directly exposed to. We discuss specific aspects regarding the followings group in more detail in subsection 5.1.

Data cleaning. To extract relevant content, we apply filters to the tweets, excluding hashtags, special characters, emojis, and multimedia items. Furthermore, we remove tweets that solely consist of links or mentions[13] and tweets from temporarily unavailable accounts due to violation of Twitter's media policy.[14] We also disregard tweets referring to pet vaccinations, tweets where the string "vax" is only found in a URL within the tweet, and tweets written in languages other than Italian. In total, we excluded 13,909 tweets.

Within our Twitter sample, we geocode the tweets through a three-step process. Firstly, we use the tweet-specific geo-tag information ("Place fields" in Figure 1). Secondly, for the remaining tweets, we rely on the users' geo-tag information ("location" within the "User Fields" in Figure 1). Lastly, we leverage Twitter users' profile information with place-name dictionaries (e.g. "live in Rome"). To map the geocoded tweets, we use geospatial shapefiles and match the latitude and longitude to Italian municipalities.[15] Figure A1 in Appendix A shows the distribution of tweets across municipalities over time.

We distinguish between original tweets, retweets, and mentions. Original tweets refer to the first occurrence of a particular content, retweets are copies of the original tweet, and mentions are copies of the original tweet accompanied by a comment.[16].

Descriptive Statistics. After the cleaning process, we are left with a sample of 2,017,539 tweets related to 227,182 unique users out of the initial 2.03 million tweets. The geolocalization narrows down the sample to 830,253 tweets written by 80,471 unique users across 4,220 municipalities from January 2013 to December 2018. This user-specific sample is strongly unbalanced, with only 4.04% of unique users present throughout the entire 6-year period, 7.13% for 5 years, 9.56 % for 4 years, 15.38% for 3 years, 25.35 % for 2 years, and 38.54% for 1 year only. Panel (a) in Table 1 provides an overview of the main characteristics of the users, panel (b) presents information about the tweets, and panel (c) focuses on user activity. On average, users created their accounts in 2012 and tweeted about vaccines ten times. Only 0.7% of users have a verified account.[17]

[13] A tweet containing another user's username, indicated by "@" symbol.

[14] Since 2021, Twitter has implemented measures such as labeling potentially misleading COVID-19 vaccine-related tweets and removing the most harmful misleading information from the platform.

[15] Roughly 5% of tweets or users' locations fall outside the Italian territory and are excluded.

[16] We screen tweets' contents for the prefix "RT @", indicating the reposting of an original tweet. We extract the Twitter handles of the creators of the original tweets by identifying the content following "@" preceding the main text. Through this procedure, we also identify replies and mentions to both original and retweeted versions of the content

[17] During the analysis period, a verified Twitter user was typically an account of public interest, often associated with well-known individuals in various domains such as music, acting, fashion, politics, religion, news, sports, and business.

Table 1: Summary statistics: Twitter data

	median	mean	sd	min	max
Panel a: User characteristics					
Tweets about vaccine	1.00	6.24	32.82	1.00	3,720
Total *tweets*	5,586.00	19,793.54	50,699.13	1.00	1,825,203
Total *followers*	335.00	3,692.14	51,951.40	0.00	3,262,940
Total *followings*	462.00	970.31	2,759.93	0.00	189,582
Account's date of creation		2012	2.49	2006	2018
Verified accounts		0.007	0.084	0	1
Panel b: Tweets' characteristics					
Length of the tweet (number of characters)		102.42	42.05	0	306
Number of words		16.13	6.96	0	62
Retweets (%)		0.60	0.49	0	1
Replies (%)		0.10	0.30	0	1
Panel c: Original Tweets' metrics					
Retweet count		2.59	35.85	0.00	6696
Reply count		0.73	7.10	0.00	1106
Quote count		0.06	1.31	0.00	341
Like count		5.71	90.44	0.00	14,188

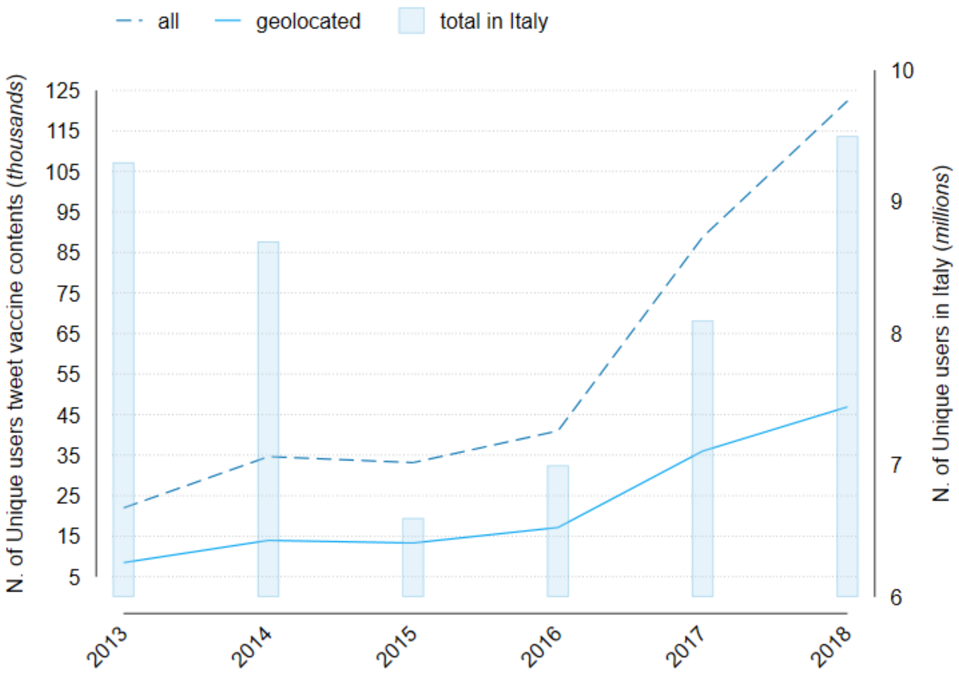
Notes: *(a)* summary statistics of 80,471 geotagged users tweeting on vaccines (2013-2018); *(b)* summary statistics of 830,253 geotagged tweets cleaned by hashtag, "RT @", "@", url and emoji; *(c)* Tweet-related popularity metrics of 328,879 original tweets.

Within the sample, retweets or mentions account for 60% of the tweets, while replies make up 10%. On average, original tweets are retweeted 2.5 times, receive 0.7 replies, 1.6 likes, and 0.06 quotes (Table 1).

Figure 2 plots the temporal distribution of unique Twitter users in Italy. The bars show the total number of Twitter users, the dashed line represents the number of users engaged in the Twitter vaccine debate, and the solid line shows the subset of geolocalized users within the previous group. The number of users in all three categories exhibits an upward trend, reaching its peak towards the end of our analysis sample, reflecting the increasing popularity of Twitter in recent times.

Among the geolocalized tweets, approximately 1% consists of municipalities where only one user tweets about vaccines in a year on average. For our analyses, we disregard this first percentile of municipalities and assess the impact of this sample restriction on our results in Table A.10 in Appendix A.

Figure 2: Number of unique users



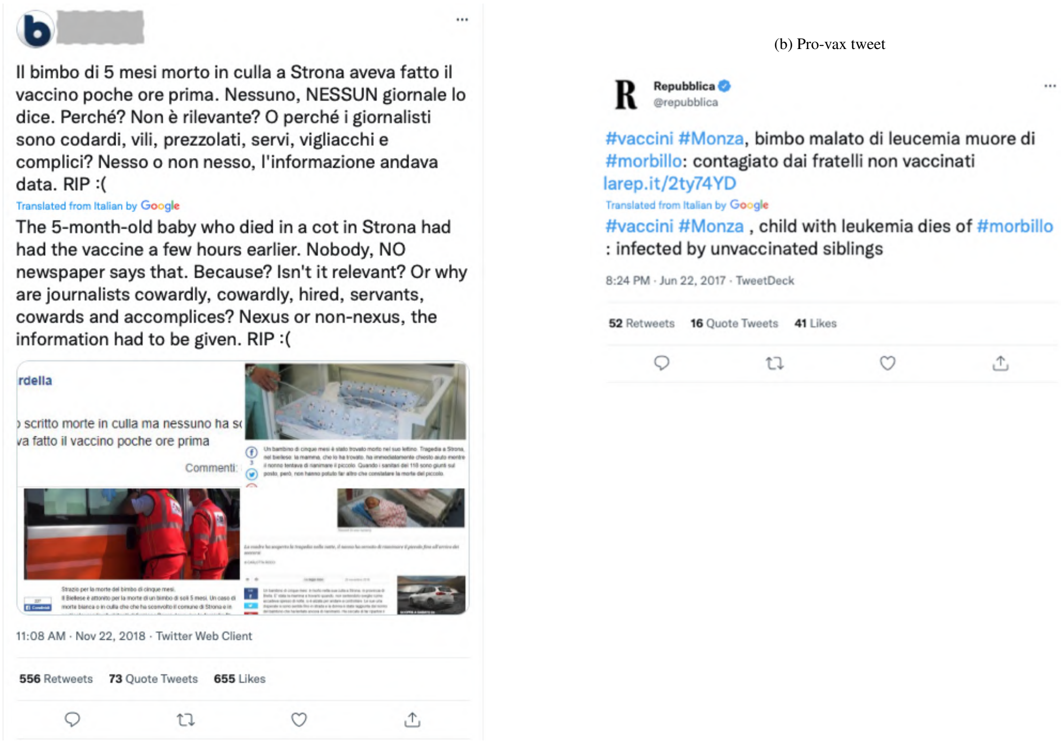
Notes: The figure shows the absolute and geotagged users who tweeted vaccine contents in Italy (left-hand axis) and the total number of Unique users in Italy as reported by AgCom (right-hand axis).

Anti- and pro-vax stances To label vaccine-skeptic tweets, we build a Natural Language Processing (NLP) transfer learning model called VaxBERTo. This anti-vax tweet classifier is developed on top of a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018a) trained in Italian (similar to the approach proposed by Polignano et al., 2019), providing the necessary final step in the data processing pipeline.

To construct the training set for our model, we pre-label tweets as 0 or 1, with 1 indicating vaccine skeptic content. Following Pierri et al. (2020), we curate the training set using tweets from renowned fake news spreaders and vaccine-skeptic users (labeled as 1), as well as pro-vaccine activists and mainstream media outlets (labeled as 0). The training sample consists of 43,472 tweets, with 20,422 pro-vaccine tweets (46.98%) and 23,050 anti-vaccine tweets (53.02%) from 108 unique users. We divide the sample into a training set of 39,124 tweets ($\approx$ 90% of the total) and a validation set of 4,348 tweets to fine-tune the training. Additionally, we build a labeled test set of 4,830 tweets to evaluate the model's performance on a different set of users. For technical details, please refer to Appendix C.

In NLP applications, the choice of the training sample directly affects the model's performance. In our approach, we construct the training sample based on pro- and anti-vax *users*, assuming that: i) users in our training sample consistently express the same stance on vaccination within their tweets, unlike "fringe" users whose stance may change over time, and ii) there are identifiable language characteristics, syntax, and structure that distinguish pro-vaccine and anti-vaccine tweets. As an example, consider the two tweets shown in Figure 3. In panel (a), a popular Italian fake news outlet spreads false information about a baby's death related to a vaccine. The tweet employs linguistic constructs commonly found in fake news, including conspiracy allusions, attacks on mainstream media, and the expression of doubts and mysteries.[18] Conversely, panel (b) shows a tweet from a mainstream media outlet reporting the death of a pediatric leukemia patient due to measles contracted from unvaccinated siblings. The tweet lacks emotional language and does not contain any conspiracy allusions.

Figure 3: Example of anti-vax (left) and pro-vax (right) tweets used for training



(a) Anti-vax tweet

[redacted]

Il bimbo di 5 mesi morto in culla a Strona aveva fatto il vaccino poche ore prima. Nessuno, NESSUN giornale lo dice. Perché? Non è rilevante? O perché i giornalisti sono codardi, vili, prezzolati, servi, vigliacchi e complici? Nesso o non nesso, l'informazione andava data. RIP :(

Translated from Italian by Google

The 5-month-old baby who died in a cot in Strona had had the vaccine a few hours earlier. Nobody, NO newspaper says that. Because? Isn't it relevant? Or why are journalists cowardly, cowardly, hired, servants, cowards and accomplices? Nexus or non-nexus, the information had to be given. RIP :(

rdella

scritto morte in culla ma nessuno ha s
va fatto il vaccino poche ore prima

Commenti:

Un bambino di cinque mesi è stato trovato morto nel suo lettino. Tragedia a Strona, nel biellese: la mamma, che lo ha trovato, ha immediatamente chiesto aiuto mentre il nonno tentava di rianimare il piccolo. Quando i sanitari del 118 sono giunti sul posto, però, non hanno potuto far altro che constatare la morte del piccolo.

Strazio per la morte del bimbo di cinque mesi.
Il Biellese è attonito per la morte di un bimbo di soli 5 mesi. Un caso di morte bianca o in culla che che ha sconvolto il comune di Strona e in

11:08 AM · Nov 22, 2018 · Twitter Web Client

556 Retweets 73 Quote Tweets 655 Likes

(b) Pro-vax tweet

Repubblica
@repubblica

#vaccini #Monza, bimbo malato di leucemia muore di #morbillo: contagiato dai fratelli non vaccinati larep.it/2ty74YD

Translated from Italian by Google

#vaccini #Monza , child with leukemia dies of #morbillo : infected by unvaccinated siblings

8:24 PM · Jun 22, 2017 · TweetDeck

52 Retweets 16 Quote Tweets 41 Likes

Notes: The translation from Italian is provided by Google.

Using the trained model, we label all the tweets in our sample, assigning them a leaning $l \in \{0,1\}$. Next, following Cinelli et al. (2021), for each user i, who produces a_{it} tweets on vaccines in year t, we define her leanings as $L_{it} = \{l_1, l_2, \ldots, l_{a_{it}}\}$. The individual stance s of user i in year t is then determined by their average vaccine leaning during that period. This average is calculated as the fraction of tweets with an anti-vaccine label ($l = 1$) and is given by:

$$s_{it} \equiv \frac{\sum_{j=1}^{a_{it}} l_j}{a_{it}} \tag{1}$$

To enhance the interpretability of an individual user's stance, we rescale it to a range between 0 and 100.

[18] For a detailed linguistic analysis see Michaels (2008).

For instance, a user with $s_{it} = 50$ has an equal number of pro- and anti-vaccine tweets, while a user with $s_{it} = 100$ has only anti-vax tweets. Within our estimation sample, the average stance value is 32.46, and the overall standard deviation is 40.03. The between-user and within-user standard deviations are 38.57 and 20.36, respectively.

3.2 Vaccination data

The data on disease-specific vaccination rates at the municipal/year level from Italian LHAs were gathered through a Freedom of Information Act (FOIA) request.[19] The vaccination rates indicate the share of the target population that has received the first dose of a vaccine recommended in the national vaccination schedule. The data cover all vaccines included in the Italian routine pediatric immunization schedule: diphtheria*; hepatitis B*, tetanus*, polio*, haemophilus influenzae type B (HIB)**, pertussis** (included in the hexavalent conjugate vaccine); measles**, mumps**, rubella** (included in the trivalent conjugate MMR vaccine), meningococcal, and pneumococcal.[20]

Table 2 shows the population-weighted average vaccination rates in the study period, along with their median, standard deviation, minimum, and maximum values. As expected, the conjugated vaccines exhibit a strong correlation in vaccination rates, while the pairwise correlation between hexavalent and MMR is 0.657, and their levels vary substantially. The hexavalent vaccine shows the highest average vaccination rates (approximately 94%), likely due to its inclusion of four mandatory shots, while the meningococcal vaccine has the lowest coverage rate (81%).

[19] The Freedom of Information Act (FOIA) provides access to public data while ensuring compliance with data protection regulations.

[20] * denotes vaccines that were compulsory in Italy between 2013 and 2017, while ** denotes vaccines included in the compulsory list under the "Lorenzin's Law" (Law Decree 73, 2017). We exclude chickenpox vaccination from our analysis, as a significant portion of the eligible population acquires immunity through natural infection and is exempted from the vaccine mandate.

Table 2: Descriptive statistics of vaccination rates (2013-2018)

		Median	Mean	SD	Min	Max	N
Hexavalent	Diphteria*	94.97	94.29	3.15	54.69	100.00	44,750
	Hephatitis B*	94.80	94.15	3.19	54.69	100.00	44,750
	Polio*	95.00	94.31	3.14	54.69	100.00	44,750
	Tetanus*	95.00	94.38	3.13	54.69	100.00	44,777
	Pertussis**	94.94	94.29	3.14	54.69	100.00	44,750
	HIB**	94.64	94.04	3.17	54.69	100.00	44,749
Hexavalent		94.53	94.09	3.12	54.69	100.00	44,779
MMR	Measles**	91.05	89.52	5.97	10.72	100.00	44,750
	Rubella**	91.00	89.50	5.97	10.72	100.00	44,750
	Mumps**	91.00	89.48	5.96	10.72	100.00	44,750
MMR		91.00	89.55	5.57	10.72	100.00	44,752
Meningococcus		87.40	81.48	15.39	0.17	99.61	43,219
Pneumococcus		91.46	87.26	11.94	.17	100	43,167

Notes: Hexavalent and MMR vaccination rates across 7,929 Italian municipalities for the period 2013-2018. Average values are weighted by the municipality population size. * marks 2013-2017 set of compulsory vaccinations, ** indicates additional mandatory shots introduced by the 2017 Law Decree 73.

3.3 Hospitalization data

The Hospital Discharge Data (SDO) from the Italian Ministry of Health provides information on the universe of hospitalizations in public and publicly-funded private hospitals from 2013 to 2016. Italy's universal public healthcare system ensures equitable access to care without significant barriers. In addition, there are no cost differentials that could affect vaccine uptake. The dataset includes socio-demographic information (age, gender, nationality, place of birth and residence, educational attainment), clinical data (diagnoses, procedures, hospital transfers, discharges), and hospitalization details (type and specialty). Hospital discharge records report information on the primary diagnosis leading to the hospitalization, along with up to five secondary diagnoses.

We focus on the diagnosis of vaccine-preventable diseases in two distinct populations: the vaccine-target population and vulnerable groups not targeted by vaccines (newborns, pregnant women, and patients with immunosuppressing conditions). The identification of the relevant diagnoses is based on the International Statistical Classification of Diseases and Related Health Problems v.9 (ICD-9) codes.[21] Based on the SDO data, we construct municipality-level annual hospitalization rates and costs per 100,000 residents for both the target and non-target populations.

Table 3 provides a detailed overview of hospitalizations and costs for different population groups. Additionally, Figure A2 in Appendix A plots the monthly trends in hospitalizations in the vaccine-target population and in the vulnerable population not targeted by vaccines.

[21] ICD-9 codes for vaccine-preventable diseases include Rubella (056 and 6475), Measles (055), Diphtheria (032), Pertussis (033 and 4843), Meningococcal (036), Tetanus (037 and 7713), Polio (045–049), Hepatitis B (070[2-3]), Mumps (072), HIB (4822), and Pneumococcal (320[1-3] and 481).

Table 3: Descriptive statistics of hospitalizations due to vaccine-preventable diseases (2013-2016)

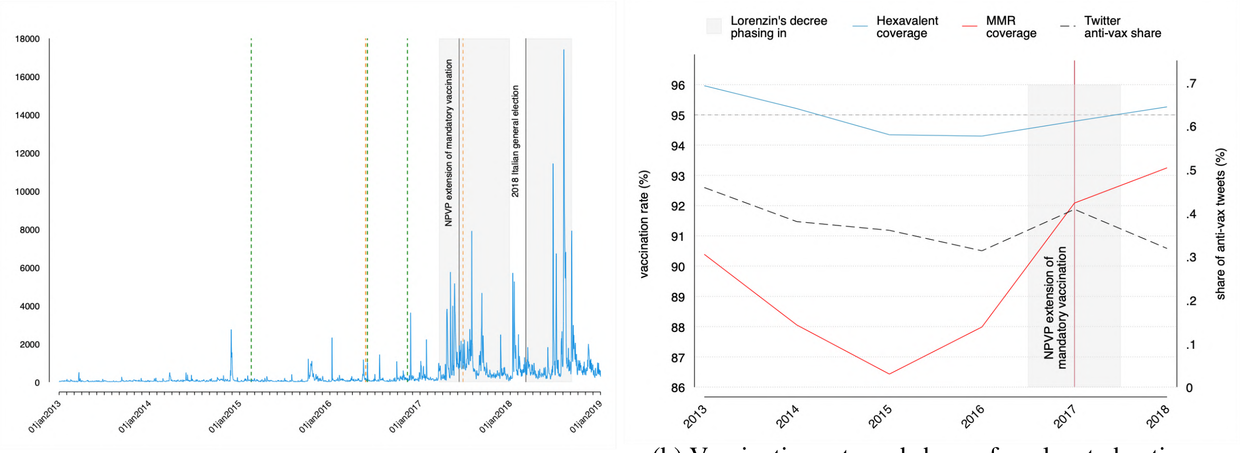
	Median	Mean	sd	Min	Max	N
Panel a: Hopitalizations						
non-target population	14.71	22.21	30.95	0.00	3,202.85	31,760
non-target population (MMR)	0.00	4.99	17.58	0.00	2,846.98	31,760
non-target population (Hexav.)	10.40	16.99	22.02	0.00	355.87	31,760
non-target population (Meningo.)	0.00	0.02	0.26	0.00	29.02	31,760
non-target population (Pneumo.)	0.00	0.88	2.25	0.00	155.04	31,760
Children age 1-10 (MMR)	0.00	2.96	6.87	0.00	1,617.25	31,760
Children age 1-10 (Hexav.)	0.00	1.27	2.70	0.00	152.44	31,760
Children age 1-10 (Meningo.)	0.00	0.04	0.41	0.00	26.21	31,760
Children age 1-10 (Pneumo.)	0.00	0.50	1.76	0.00	132.04	31,760
Panel b: Healthcare costs						
non-target population	38,581.69	66,477.60	116,320.65	0.00	59,880,842.11	31,760
non-target population (MMR)	0.00	15,381.55	96,931.58	0.00	59,880,842.11	31,760
non-target population (Hexav.)	46,275.59	83,151.57	119,925.38	0.00	14,819,697.72	31,760
non-target population (Meningo.)	0.00	150.92	3,976.38	0.00	411,341.22	31,760
non-target population (Pneumo.)	0.00	2,332.30	9,004.03	0.00	1,941,927.83	31,760
Children age 1-10 (MMR)	0.00	4,749.99	25,506.58	0.00	2,274,286.39	31,760
Children age 1-10 (Hexav.)	0.00	2,545.85	9,407.74	0.00	759,286.31	31,760
Children age 1-10 (Meningo.)	0.00	190.58	3,185.72	0.00	409,748.10	31,760
Children age 1-10 (Pneumo.)	0.00	1,255.36	5,365.51	0.00	259,504.65	31,760

Notes: The statistics refer to 7,940 municipalities for the time period between 2013-2016 and are weighted by the municipality population size.

4 Twitter stances and user interactions

Social media platforms play an active role in news creation and distribution, involving users and influencing opinions beyond the platform itself. User-to-user sharing is a key factor in disseminating content online and potentially offline, impacting opinions and real-world behaviors. Social media discussions often exhibit patterns of attention, with periods of low interest or *controversialness* interrupted by sudden spikes of activity. These spikes can be triggered by exogenous events or fueled by platform algorithms designed to enhance user engagement (Lorenz-Spreen et al., 2019). Twitter, in particular, has employed algorithmic amplification since 2016 to maximize exposure to captivating content(Huszár et al., 2022).

Figure 4: Number of tweets, vaccination rates and anti-vax sentiment in Italy



(a) Number of tweets over time

(b) Vaccination rate and share of geolocated anti-vax tweets

Notes: Panel (a) shows the time series of the number of tweets on vaccinations, 2013-2018. The dashed reference lines report notable (i.e., covered by national media) events regarding vaccination. In particular, they flag $i)$ verdicts (green): the reversal of the Rimini's Court sentence by the Bologna's Appeal Court - February 15^{th}, 2015; the recognition of the inconsistency of the link between the MMR vaccination and autism by the prosecutor of Trani - June 1^{st}, 2016; the dismissal by the court of Milan of the appeal against a sentence establishing the causal link between the vaccine and the severe encephalopathy developed by in an infant - November 10^{th}, 2016; $ii)$ death (orange) of an infant following a mandatory vaccination - May 25^{th}, 2016 and of another infant affected by leukemia of measles contracted from non-vaccinated siblings - June 23^{rd}, 2017. The first grey shaded area marks the period of the debate, which preceded and ensued the approval of "Lorenzin's Law" (June 7^{th}, 2017, solid black line). The second grey area followed the general elections (March 4^{th}, 2018) until the upcoming school starting date - a symbolic moment that created political clashes between the Italian populist parties then ruling the government due to the vaccine mandate's enforcement on school enrollment. Panel (b) reports the yearly average values of hexavalent (solid blue) and MMR (solid red) vaccine coverage rates, as well as the average Twitter anti-vax sentiment (dashed black) as computed in Figure 3.1 recorded between 2013 and 2018.

Based on our data, Figure 4, panel (a) illustrates the daily dynamics of vaccine-related tweets in our sample from 2013 to 2018. The average activity remained relatively stable until 2017, when the implementation of the *Lorenzin's law* sparked longer and more intense debates. This resulted in a peak of approximately 8,000 tweets per day around the approval date. The vaccine debate became highly politicized during the 2018 general election campaign, with populist politicians expressing skepticism towards the vaccine mandate.

Figure 4, panel (b), shows the aggregate trends in coverage rates for the hexavalent and MMR vaccines, along with the average anti-vax Twitter sentiment from 2013 to 2018. Over this period, there has been a progressive decline in the coverage of both vaccinations. Coverage rates started to increase in 2015 due to the reversal of the court ruling on the vaccine-autism link, and a significant rise in measles cases in Italy. In 2017, the expansion and legal enforcement of mandatory vaccines under the new law led to an increase in MMR coverage.

It is important to note that the fluctuations in average Twitter anti-vax sentiment may not immediately translate into lower vaccination rates. Vaccination decisions are influenced by individuals' risk perceptions, which can be shaped by cognitive biases and local epidemiology. As a result, the correlation between coverage rates and anti-vax stances may be distorted due to simultaneity and omitted variables.

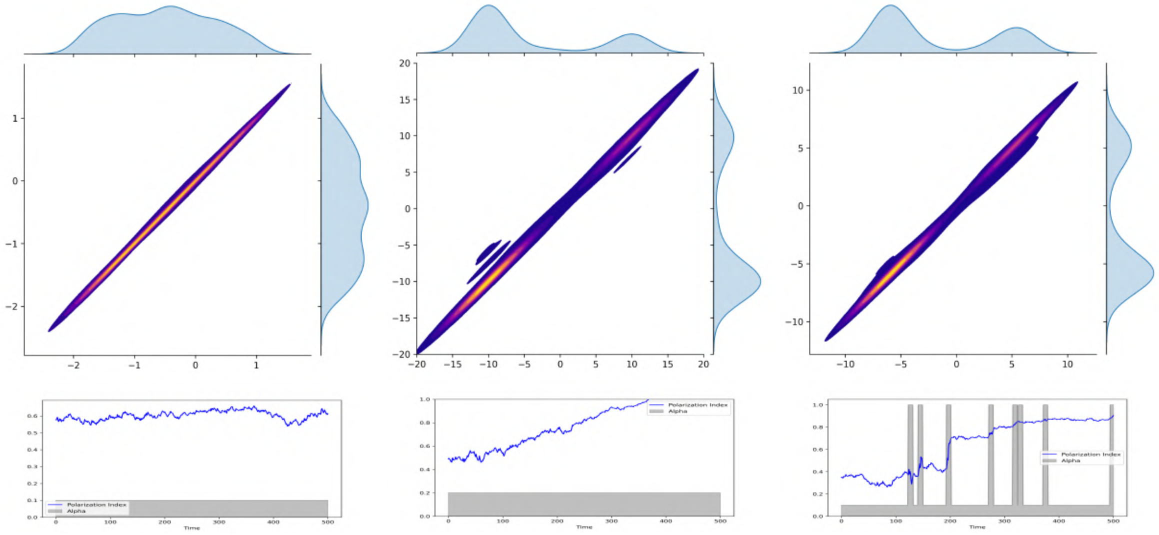
Echo chambers formation. We rationalize the evolution of anti-vax views on social media in Italy using a model of opinion dynamics in social networks based on Baumann et al. (2020) (see Appendix B for the complete model description). The model combines exogenous and endogenous drivers of attention and interactions, which can result in the polarization of opinions when controversy surrounding a topic intensifies. Users with divergent views tend to form links and cluster in echo chambers, amplifying the impact of external shocks and potentially leading to longer periods of intense activity and more extreme positions. These endogenous dynamics contribute to the complex relationship between the spread of anti-vax opinions on social media and vaccine hesitancy.

The model identifies two channels through which users influence each other's opinions on a controversial topic. On the one hand, users are influenced by exposure to views that differ from their own, with more divergent perspectives having a greater impact (*exposure effect*). On the other hand, the controversial nature of the vaccine-related topic intensifies polarization by shaping the formation of network connections (*link formation effect*). Importantly, the first channel captures the impact of anti-vax stances expressed on social media on vaccine hesitancy, which is the focal point of our analysis. However, due to the *link formation effect*, when a topic becomes controversial, leading to a non-random network structure that deviates from the assumption of random content exposure.

In the model, the dynamics of opinion within the social network are driven by interactions among agents, where each agent's stance (s_i) and the level of controversy of the topic influence the stances of others. Importantly, the influence of individual stances on other users is modulated by the controversialness, which is modeled as a hyperbolic function. Even moderate opinions can effectively capture the beliefs of peers. Each agent has a propensity to interact with a certain number of other agents, and the probability of interaction depends on the degree of homophily, modeled as a decreasing function of the opinion distance between agents (Bessi et al., 2016). Links in the network represent the medium through which information flows. For example, if user i is linked to user j, user i is exposed to the content produced by user j, resulting in an information flow from node j to node i in the network. The network's topology reveals the presence of echo chambers when a large portion of users are tied to peers with similar views, increasing the likelihood of exposure to similar content. From a network perspective, this means that a node i with a given stance s_i is more likely to be connected to nodes with stances close to s_i.

The model generates different predictions for the converged state (Figure 5) depending on the level and time-varying nature of the exogenous topic controversialness. When the controversialness remains consistently low (panel a), the tendency of users to connect with like-minded peers is counterbalanced by their limited influence, resulting in convergence to a non-polarization state. Conversely, when the controversialness is permanently high (panel b), strongly polarized echo chambers form. Similar results are observed when the level of controversialness varies over time (panel c). In our simulations, we replicate the occurrence of short periods of high controversy interspersed with prolonged periods of low controversialness, resembling the patterns observed in the data. Notably, these brief periods of high controversy have long-lasting effects on the overall level of polarization due to the link formation effect. For further details on the model and simulation results, please refer to Appendix B).

Figure 5: Simulated distribution of stances



(a) Low controversialness

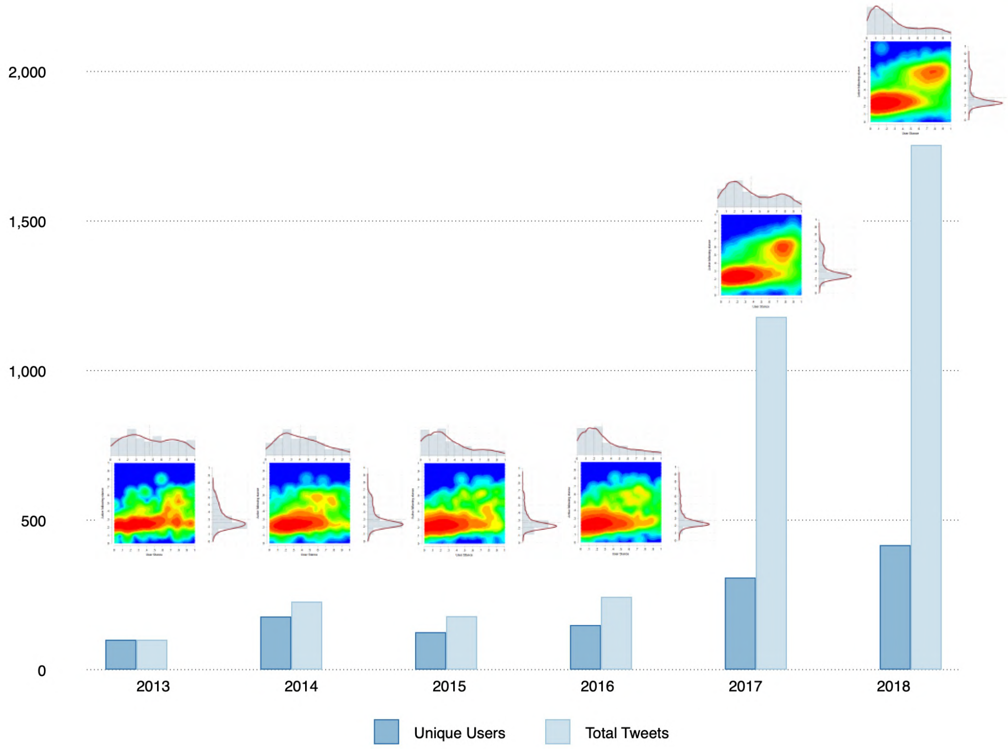
(b) High controversialness

(c) Time-varying

Notes: user (x-axis) and average followings' (y-axis) distribution of stances in a simulated model when controversialness is low ($\alpha = .1$ in panel a), high ($\alpha = .2$ in panel b), and low with short-lived outbursts ($\alpha = 0.1$ and $\alpha = 1$ in panel c). In all models, the number of individuals is $N = 500$ and the periods are $T = 5$ - divided into 100 subperiods. Initial values (s_0) are randomly drawn from a gaussian distribution with $\mu = -0.2$ and $\sigma = 0.5$ to match the asymmetry of the initial opinions in the data. The time series report the degree of polarization and the controversialness parameter observed in each subperiod.

In Figure 6, bars plot the annual count of unique users and tweets in our dataset (normalized to 2013=100). Additionally, the heatmaps depict the joint distribution of users' average stances and the average stances of the users they follow. In line with Figure 4, the period from 2013 to 2016 exhibits a sustained low level of activity. In 2017, when the vaccination mandate was extended, there was a significant increase in the number of users and tweets. Interestingly, the number of tweets increased much more (10×+) compared to the corresponding number of users (3×+), suggesting that users already interested in the topic engaged more often in vaccine-related debates (indicating a heightened controversy around the topic). Accordingly, the heatmaps show the formation of echo chambers, with two distinct clusters representing users, suggesting the radicalization of opinions among users. In this context, it is likely that the higher controversialness of vaccine-related topics was further reinforced by the Twitter amplification algorithm introduced in 2016, which magnifies the exposure to topics that engage users' attention.

Figure 6: Dynamics of Twitter activity on vaccination - 2013/2018



Notes: yearly number of vaccine-related total tweets (grey bars) and unique users (blue bars) between 2013 and 2018 (2013=100). The contour plots report the joint distribution of users' and average followings' stances on vaccination. Colors represent the density of users: the stronger the red hue, the larger the number of agents. The marginal distribution of users' opinions and their followings' are plotted on the x and y-axis, respectively. To construct the figure, we exclude the users with less than 15 friends and 10 tweets/year in the sample to avoid social bots, as their inclusion would artificially generate echo chambers - see, e.g., (Shao et al., 2018).

The persistent effect of endogenous link formation resulting in echo chambers poses a challenge for causal inference. Without adjusting for the systematic tendency towards homophily, naive estimates of the exposure to online anti-vax content on vaccine hesitancy will inevitably be biased. Hence, these model predictions motivate the use of an IV identification strategy to estimate the empirical counterpart of the exposure effect.

5 Empirical strategy

Ideally, our goal would be to estimate the impact of exposure to anti-vax content on vaccination decisions at the individual (parent) level.[22] However, there are two challenges that make this goal difficult to achieve: i

[22] In an ideal model, this would be represented by the following linear relationship at the individual (parent) level:

$$\boldsymbol{v}_{-it} = \beta s_{it} + X_i + Z_c + \Omega_t + \varepsilon_{it} \tag{2}$$

the endogeneity inherent in the relationship between exposure and stance, and ii the lack of individual-level vaccination data.

Firstly, as discussed in section 4, the presence of homophily and the controversial nature of the vaccine topic contribute to the formation of echo chambers. Consequently, any observed correlation in vaccine stances among users may be driven by the endogenous selection of connections. For instance, user i may choose to connect with user j because they both hold similar views on vaccines, leading to exposure to similar content. To address the endogeneity in the formation of social connections, also well-known in the literature on peer effects in social networks, we use an IV approach.[23] In our study, we build an instrument for the exposure to anti-vax content by leveraging the Twitter network structure. Our approach is similar in spirit to the local-average model proposed by Bramoullé et al. (2009) and others, but with a focus on the subset of connections that are plausibly exogenous. Further details on the construction of the instrument can be found in subsection 5.1.

A second challenge rises from the unavailability of individual-level data on vaccine hesitancy (v_{it}). We thus rely on the most granular data available on pediatric vaccinations based on the coverage rates at the municipal/year level in Italy. To bridge the mismatch between individual Twitter stances and vaccination rates, we use a mixed two-stage least squares (M2SLS) strategy proposed by Dhrymes and Lleras-Muney (2006) for grouped data. We explain details of this approach in subsection 5.2.

5.1 The "Followings of Passive Followings" (FoPF) network

To construct a two-step neighborhood for each user i, we gather two groups of users: the set of users they follow ($j \in J_i$) and the users followed by each of their followings ($k \in K_{j|i}$, where $K_{j|i}$ represents the set of users followed by user j). We define the set K_i as the union of all these "followings-of-followings" (FoF) users, i.e., $K_i = \bigcup_{j \in J_i} K_{j|i}$.[24] A user k in the FoF network, denoted as $k \in K_i$, can fall into two categories. First, if k is a direct connection of user i (i.e., $k \in J_i$), it is excluded from the FoF network to avoid "circles in friendships". Second, k can be an incidental connection not chosen endogenously, meaning that user i does not follow k ($k \notin J_i$), but k is connected to user i through another user j. In this second case, k still has the potential to

where $\boldsymbol{v}_{-it}$ reflects vaccine hesitancy of peers of individual i at time t, s_{it} is the stance of individual i, X_i and Z_c are individual and characteristics, and Ω_t is the amount of information available in each period, including policy interventions (e.g., vaccine mandates), new scientific knowledge, and news related to vaccine-preventable diseases outbreaks. The proposed model assumes a one-to-one mapping between vaccine hesitancy and the observed behavior towards vaccination, i.e., there is a threshold value $v^\star = \mu + \alpha$ above which parents do not vaccinate their children. The parameter of interest β would capture the influence that individual i's stance has on her peers' decision on vaccinations. An assumption underlying the above model is that the extent of anti-vaccination persuasion on Twitter is representative of the pressure exerted by vaccine skeptic activists on parents exposed to other media outlets, both online and offline.

[23] Various frameworks have been developed to identify peer effects in the presence of endogeneity, by e.g. Goldsmith-Pinkham and Imbens (2013) and (Johnsson and Moon, 2021). A second strand of literature exploits field experiments to estimate peer pressure effects (Bursztyn et al., 2014, 2019).

[24] In building the network, we use end-of-the-sample-period data, accounting for the endogenous effects of link formation that occurred during the analysis period.

influence user i's exposure to vaccine-skeptic content through interactions with j, such as retweeting, liking, or replying to j's posts. Even within the FoF network, there exists some degree of endogenous link formation, as users may interact or not based on the vaccine stance of their direct connections.[25] To address this issue, we distinguish between two types of connections: i) active followings (J_i^A), who actively generate original vaccine-related tweets, and ii) passive followings (J_i^P), who never create vaccine content but can retweet or like other users' posts. Although we cannot directly measure the vaccine stance of passive followings (J_i^P) since they do not produce original content, their lack of engagement ensures that the connection with user i is not driven by endogenous factors such as homophily (at least not specifically related to the vaccine debate).[26]

Our aim is to assess the impact of passive exposure to vaccine-related content on individual vaccine attitudes. To do so, we focus on the exposure mediated by passive users J_i^P. Specifically, we define the set of their followings as $K_i^P = \bigcup_{j \in J_i^P} K_{j|i}$, which represents the followings-of-passive-followings (FoPF) network. We utilize this network to instrument users' exposure to vaccine-related content. While the stance of FoPFs users may influence individual i's opinion, the formation of connections between them is not affected by endogeneity.[27] To ensure that the links in the FoPF network were established before the user's involvement in the vaccine debate on Twitter, we impose a restriction. Specifically, we consider only the FoPF that were engaged in any vaccine-related debate before user i.[28]

Figure 7 illustrates the selection of followings and FoPF for every "ego" user, represented by the Twitter handle @Jane. The nodes represent users, with their size indicating their distance from @Jane (first- or second-degree connections), and their color indicating their engagement level - yellow for active users, gray for passive ones. The edges represent the connections between users, which are driven by endogenous factors.

@Jane's first-degree network (J_{jane}) includes one active following (@Julie) and two passive followings (@John and @Bob). The FoF network consists of four users: one linked to @Julie (@Miriam), two connected to @John (@David and @Anne), and one user (@Tony) who is connected to both @Julie and @John. We build the user-centered FoPF network, and by considering their exclusive connections (e.g., excluding @Tony due to his connection to an active first-degree user), we build the measure of indirect and exogenous exposure

[25] For example, let's consider an anti-vaccine journalist who shares her views on Twitter. She collects information and insights from her own network, which likely consists of individuals who share her viewpoint. She then creates original content that resonates with and establishes connections with like-minded users. In this case, the formation of links is influenced by the journalist's connections, whose vaccine stance cannot be considered exogenous.

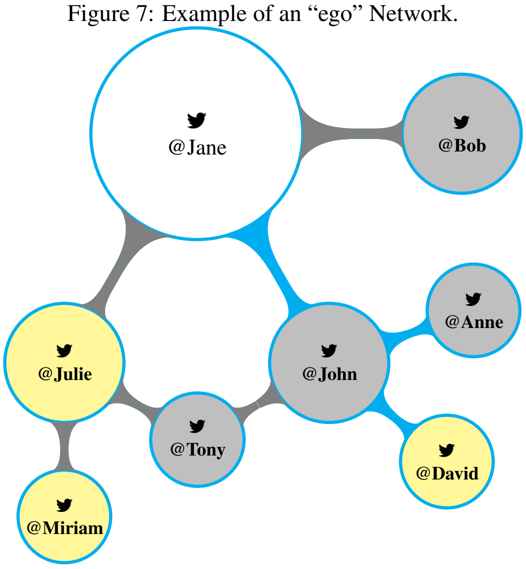
[26] It is possible for connections to form between users due to various reasons, including shared interests or offline networks such as supporting the same soccer team or working together. As long as these connection formations are unrelated to their stance on vaccines, the indirect connections established through these links provide an exogenous source of variation.

[27] To further clarify our approach, we construct the FoPF network using the active followings of the passive users, denoted as $K^A_{j^P|i}$, where j^P represents a passive user. Since the passive FoPFs ($K^P_{j^P|i}$) do not express their opinion and we cannot define their stance, we focus on the active followings to build the network.

[28] We exclude those who created their account after user i or whose first tweet about vaccines was published after user i posted her first vaccine-related tweet.

to anti-vax content of @Jane, represented by the blue edges. In this example, the measure is determined solely by the stance of @David, who is the only active user within the FoPF network.

Figure 7: Example of an “ego” Network.



Notes: The figure plots the architecture of the network on Twitter. The white node (@Jane) is the “ego” user, the gray nodes denote the passive users, in yellow the active ones. The node size depends on the distance from the ego user (lag-1 or lag-2), which is informed by the endogenously-generated links, described by the edges. These are either gray - i.e., passing through an active user in lag-1 or not connected to any lag-2 user - or blue - valid connections to build the measure of indirect exposure.

We examine a sample of 65,673,913 followings’ nodes. Among these, we identify 8,176,261 unique passive followings. As is typical on social media, we observe significant variations in the number of followings and followers. While the majority of users have only a few followings, there are some users who serve as central nodes in the network. The final sample of the second-degree network consists of approximately 2 billion nodes, corresponding to an average of 12,556 FoF per user. The median user has 469 passive followings and a median of 7,687 FoPF. These connections collectively generate an on average of 142,261 tweets about vaccines (see Table 4).

Table 4: Descriptive statistics of users’ networks

	Median	Mean	sd	Min	Max
Followings	469	973.46	2,717.55	1.00	189,433
FoPF	7,687	12,556.24	14,078.73	1.00	139,508
Total FoPF’ tweets with vaccine contents	59,535.50	142,261.09	186,460.83	1.00	1,685,355
FoPF’ stance (ffs_{it})	28.829	29.353	6.750	0	100

Notes: The networks refer to 80,471 geotagged unique users who tweeted on vaccines in Italian (2013-2018).

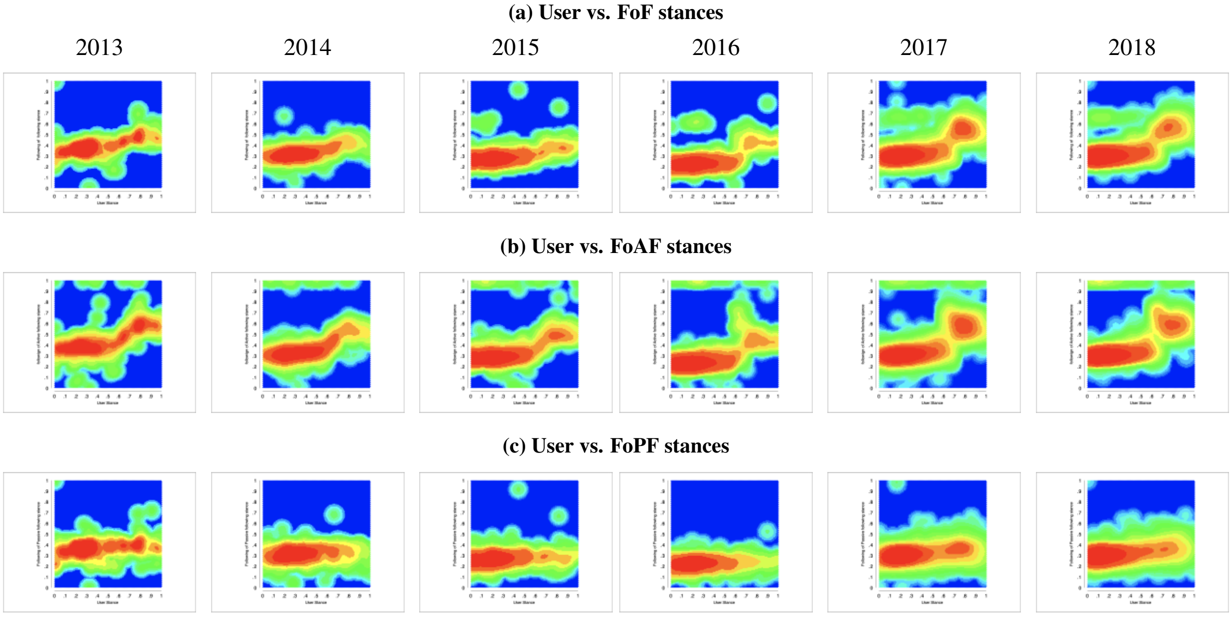
Finally, for each FoPF network, we compute the average anti-vax stance. For user i in year t, we define their indirect exposure to anti-vax stance as $ffs_{it} = \frac{\sum_{j=1}^{N_{it}} s_{jt}}{N_{it}}$, where N_{it} represents the number of FoPFs for user i in year t. This measure ranges from 0 to 100. We utilize this measure as an instrumental variable for each user's own stance.

Validation To ensure the validity of the average stance of the FoPF network as an instrument, it is important that the connections between user i and her J_i^P were not driven by endogenous factors related to the stance of their connections. Although we cannot formally test this hypothesis, we provide suggestive evidence in support of our claim in Figure 8. We plot users' stance on the x-axis against the average stance of their FoF network (row 1), followings-of-active-followings $K_{j^A|i}$ (FoAF, row 2), and FoPF network $K_{j^P|i}$ (row 3) for each year in our sample.

We observe that the FoF network exhibits a significant degree of polarization in later years, similar to the ones plotted in figures 5 and 6 for users and their direct connections. This suggests the presence of endogenous links that affect their network beyond their direct connections. Similarly, the FoAF network also displays a strong polarization pattern.

Conversely, when we focus on the FoPF network we find that their stance is uncorrelated with the stances of the users, and their distribution follows a Gaussian pattern centered around the unconditional mean in each period. This finding is consistent with passive followings being connected to users independently of their own stance on vaccines, as well as the stance of their connections.

Figure 8: Dynamics of Twitter activity on vaccination



(a) User vs. FoF stances

(b) User vs. FoAF stances

(c) User vs. FoPF stances

5.2 The Mixed two-stage least squares

In a naive OLS estimation, without taking into account endogeneity, we would measure the impact of online anti-vax skepticism and health outcomes at the municipality level as follows:

$$V_{mt} = \beta \bar{s}_{mt} + T'_{mt}\zeta + C'_{mt}\phi + \gamma_m + \theta_t + \varepsilon_{mt} \tag{3}$$

where V_{mt} is a vaccination rate, or the vaccine-preventable hospitalizations/costs in municipality m in year t, $\bar{s}_{mt}$ is the average vaccine-related stance at municipality/year level, T' represents vectors extracted from the Twitter corpus and the followings' network (i.e. the sum of tweets per municipality/year and the sum of FoPF tweets per users' municipality/year), C's are socioeconomic characteristics (income per capita at the municipality level, birth rate, the share of lower secondary school attainment, the mean age of women at the birth of their first child at province level, and health costs per capita at the regional level).[29] Additionally, as there might be strong political components to vaccination rates, in C' we include an indicator variable for the rule of *populist* parties at the local level. Several populist parties have raised concerns about vaccine safety (Guriev and Papaioannou, 2022, Kennedy, 2019).[30] We also include municipality and year fixed effects (γ_m and θ_t, respectively). Finally, as public health measures and compliance with these measures might vary at the regional level, we include a set of region-specific time trends $\rho_r \times t$ (region×year).

The simple OLS fixed effects estimation in this context would be prone to bias, which cannot be predetermined. This bias stems from potential bidirectional effects, where a stronger anti-vax sentiment may lead to lower vaccination rates, and conversely, lower vaccination rates may result in more severe health complications, potentially influencing the anti-vax sentiment.To mitigate these biases, we require an instrument that introduces exogenous, independent variation in the average sentiment.

To address this, we leverage user-level data from Twitter, which offers higher resolution, to enhance the accuracy of the first stage. However, since the outcome measures are only available at the municipality level, we employ the M2SLS approach proposed by Dhrymes and Lleras-Muney (2006). In the first stage of M2SLS, estimated using weighted least squares, we specify the following equation:

First stage - (individual level)

$$s_{it} = \alpha + \beta ffs_{it} + \mathbf{T}'_{it}\zeta + \mathbf{C}'_{mt}\phi + \gamma_m + \rho_r \times t + \theta_t + \varepsilon_{it} \tag{4}$$

where s_{it} is the Twitter stance on vaccines of user i in year t, ffs_{it} denotes her indirect exposure to anti-vax content (as described in subsection 5.1), while $\mathbf{T}_{it}$ and $\mathbf{C}_{mt}$ represent Twitter and municipal characteristics. Both s_{it} and ffs_{it} range between 0 and 100, with 100 indicating the maximum level of vaccine skepticism. In

[29] Birth rate, the percentage of people with at least lower secondary school, the mean age of females at first birth, and health costs per capita data come from the Italian National Institute of Statistics. Per-capita income data comes from the Ministry of Economy and Finance. Descriptive statistics are reported in Table A.1 in Appendix subsection A.1

[30] Following Albanese et al. (2022), parties coded as populist in Italy are the Movimento Cinque Stelle (Five Stars Movement) and Lega Nord (Northern League). The data comes from the Ministry of the Interior.

our setting, there is a one-to-one mapping between geotagged users and the municipality they reside in or tweet from. Equation (4) allows us to compute the instrumented value $\widehat{s}_{it}$, which is then aggregated at the municipal level to obtain the main regressor for the second stage, which reads:

Second stage - (municipal level)

$$V_{mt} = \alpha + \lambda\overline{\widehat{s}}_{mt} + \overline{\mathbf{T}}'_{mt}\xi + \mathbf{C}'_{mt}\phi + \gamma_m + \rho_r \times t + \theta_t + \eta_{mt} \tag{5}$$

where the outcome of interest (V_{mt}) is represented by vaccination rates, the number of vaccine-preventable hospitalizations in the targeted and non-targeted populations, or their cost. $\overline{\widehat{s}}_{mt}$ is the averaged instrumented regressor computed in the first stage, weighted by the number of observations in the original cell (number of users at municipality/year level), $\overline{T}_{mt}$ is the average value of Twitter's control variables (T'_{it}), C'_{mt} is the vector of socioeconomic characteristics, γ_m, and θ_t are municipality and year fixed effects that account for time-invariant differences between municipalities and $\rho_r \times t$ (region×year) controls for region-specific trends. All estimates are weighted by municipality population size. We correct the variance-covariance matrix throughout the analysis by bootstrapping the standard errors. In the main specification, the parameter of interest λ captures the causal effect of anti-vax stances on the vaccination rate at the municipality level.

6 Results

In our presentation of the results, we begin by examining the baseline estimates for vaccination rates, categorizing them based on the type of vaccine. This allows us to analyze the differential impact of vaccine skepticism on mandatory or recommended vaccines, including MMR vaccine, which is the specific target of online disinformation. Subsequently, we present the findings on hospitalizations. We examine the number of hospitalizations for vaccine-preventable diseases and the associated costs, adjusted per 100 thousand residents. Furthermore, we differentiate between hospitalizations of the vaccine-targeted pediatric population and those of non-target populations consisting of vulnerable individuals, such as newborns, pregnant women, and immunocompromised patients.

To assess the random assignment of the IV with respect to the contextual features of the user's geolocalized municipality, we conduct a series of regression tests. Specifically, we regress municipality characteristics, such as income per capita, birth rates, public healthcare expenditure per capita, and education attainment, on the average Twitter stance on vaccines that user i in municipality m is indirectly exposed to through her FoPF network (ffs_{it}), while controlling for municipality and year fixed effects. The results presented in Table A.2 support the assumption that our model specification identifies a source of variation unrelated to municipality characteristics, as none of the estimated correlations deviate significantly from zero.

The results of the M2SLS first stage, presented in Table Table 5, provide strong evidence that the vaccine-related stances of a user's followings-of-followings network significantly influence the user's own stance on vaccines. This finding highlights the substantial impact of indirect exposure to anti-vaccination sentiments on individuals.

The M2SLS first stage results, shown in Table 5, suggest that the vaccine-related stances of a user's followings-of-followings network significantly influence the user's own stance on vaccines.[31] A one-unit increase in the anti-vaccination stance on the 0-100 scale leads to a 0.7-unit increase in the individual's vaccine-related stance, indicating that indirect exposure to anti-vaccination stances can lead users to engage in anti-vaccination activism.

Table 5: M2SLS Individual - First stage.

	(1)	(2)	(3)	(4)	(5)	(6)
	s_{it}	s_{it}	s_{it}	s_{it}	s_{it}	s_{it}
	(30.31)	(30.31)	(30.31)	(30.31)	(30.31)	(30.31)
ffs_{it} (28.77)	0.799***	0.751***	0.703***	0.703***	0.704***	0.704***
	[0.021]	[0.021]	[0.017]	[0.017]	[0.017]	[0.017]
N	127,754	127,754	127,754	127,754	127,754	127,754
CONTROL (Twitter)				✓		✓
CONTROL (socioeconomics)					✓	✓
YEAR FE	✓	✓	✓	✓	✓	✓
CITY FE		✓	✓	✓	✓	✓
Reg × Year			✓	✓	✓	✓
F-stat	1,501.16	1,288.96	1,765.22	1,763.52	1,755.84	1,757.86

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: The numbers refer to the sample of 830,253 tweets and to a population of 80,471 unique users across 4,220 municipalities. All estimates include municipal and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level. The average values of s_{it} and ffs_{it} in parentheses are weighted by population size.

6.1 Vaccination rates

Table 6 reports the baseline IV results alongside those of the naive OLS model that does not account for endogeneity. Given that hexavalent and MMR vaccines are almost always administered through a single shot, the disease-specific vaccination rates are identical and we report the pooled figure for both. Their average coverage rates are given in parentheses, and the table reports the most demanding specifications, including all controls and fixed effects.

The coefficient estimated for mandatory vaccines (hexavalent) is not statistically distinguishable from zero, and there is no detectable difference between the OLS and the M2SLS approaches. Similarly, we estimate no statistically significant effect of anti-vax stance on the other recommended vaccines against meningococcal and

[31] Among the geolocalized tweets, 1% has an average of 1 user only tweeting about vaccines in a year. In the baseline analysis (Table 5), we drop the first percentile of municipalities. We test the results obtained on the full sample in Appendix A, Table A.10.

pneumococcal diseases. The magnitudes of the coefficient estimates, in this case, are comparable to the one on MMR, which is likely to reflect the non-compulsory nature of these shots. Yet, the precision of the estimates is scant. On the other hand, when we look at the vaccine most targeted by the fake news, the MMR shot, we find i) a significant effect on coverage rates, and ii) a sizeable difference with respect to the OLS specification. We find that a 10 percentage point increase in the municipality-level anti-vaccination stance leads to a 0.43 percentage point decrease in the MMR coverage rate.[32]

Table 6: Results of the OLS and the Second stage of the M2SLS - Vaccination rates

	(1)	(2)
	OLS	M2SLS
	V_{mt}	V_{mt}
Panel a: Hexavalent (94.06)		
s_{mt}	-0.001	-0.023
	[0.002]	[0.015]
N	7,239	7,239
Panel b: MMR (89.53)		
s_{mt}	-0.005	-0.043**
	[0.003]	[0.021]
N	7,238	7,238
Panel c: Menigococcal (81.32)		
s_{mt}	-0.002	-0.040
	[0.008]	[0.054]
N	7,061	7,061
Panel d: Pneumococcal (82.64)		
s_{mt}	-0.0001	-0.029
	[0.008]	[0.052]
N	7,066	7,066
Controls (Twitter)	✓	✓
Controls (socioeconomics)	✓	✓
City and year FE	✓	✓
Reg × year	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates as well as averages of V_{mt} are weighted by the municipality population size.

6.2 Hospitalizations

We also estimate the effect on hospitalizations due to vaccine-preventable conditions. We distinguish between two groups: the target pediatric population and non-target vulnerable individuals. In fact, the number of hospitalizations for vaccine-preventable diseases among non-targeted patients measures the extent of negative externalities of suboptimal immunization rates on local communities. Quantifying these externalities provides an objective argument in the policy debate on vaccine mandatesmat.

[32] Table A.3 in Appendix A reports the reduced form estimates. Table A.5 in Appendix A reports the full set of estimates for the different models as specified in Table 5.

Table 7: Results of the OLS and the Second stage of the M2SLS - Hospitalizations .

	(1) OLS V_{mt} non-target pop.	(2) M2SLS V_{mt} non-target pop.	(3) OLS V_{mt} non-target pop.(MMR)	(4) M2SLS V_{mt} non-target pop.(MMR)	(5) OLS V_{mt} Children age 1-10 (MMR)	(6) M2SLS V_{mt} Children age 1-10 (MMR)
Panel a: Hopitalizations						
s_{mt}	0.0211	0.213*	0.018**	0.234***	0.007	0.145**
	[0.0159]	[0.113]	[0.00841]	[0.0601]	[0.008]	[0.065]
Panel b: Healthcare costs						
s_{mt}	129.8*	731.1**	71.96**	722.1***	47.13*	366.9**
	[66.39]	[353.8]	[30.92]	[243.1]	[25.95]	[161.1]
N	3,331	3,331	3,331	3,331	3,331	3,331
Controls (Twitter)	✓	✓	✓	✓	✓	✓
Controls (socioec.)	✓	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates are weighted by the municipality population size.

Also in Table 7, we estimate the effect on hospitalizations and their cost for the two populations, expressed per 100 thousand residents. For vulnerable individuals (the non-target population), we find that a 1 percentage point increase in the municipality-level anti-vaccination stance leads to an additional 0.21 hospitalizations per 100 thousand residents (the baseline average being 22.21). This corresponds to an additional healthcare expenditure of 731.1 euros, representing a 1.1% increase relative to the baseline. Specifically, in terms of hospitalizations due to MMR, the same increase in vaccine skepticism is associated with an additional 0.23 hospitalizations per 100 thousand residents (the baseline average being 4.99) and an additional expenditure of 722.1 euros, corresponding to a 4.6% increase. When looking at hospitalizations among the target pediatric population, our estimates (column 5) suggest that a 1 percentage point increase in the municipality-level anti-vaccination stance leads to an additional 0.145 hospitalizations per 100 thousand residents (the baseline average being 2.96) and an excess expenditure of 366.9 euros, corresponding to a 7.7% increase.[33]

In line with the baseline results, Table A.6 in subsection A.1 shows no significant results for the non-target population and target pediatric population hospitalized for diseases preventable by hexavalent, meningococcus and pneumococcus vaccines, respectively.

To evaluate the efficacy of vaccinations in reducing the probability of mortality from vaccine-preventable diseases, Table A.8 in Appendix A shows the result of our estimates on the mortality cases of patients hospitalized with these diseases.

[33] Table A.4 in Appendix A reports the reduced form estimates.

6.3 Robustness checks

We first check the robustness of our results to three potential confounders that could affect either the first stage (the introduction of a homophily-enhancing algorithm on Twitter) or the second stage (pre-existing vaccine mandates or the influence of strong populist parties). Finally, we propose a reweighting of our estimates to account for the number of second-degree links between each user and her FoPF network.

First, in 2016, Twitter introduced an algorithmic timeline that rearranges users' feeds based on relevance rankings. This is likely to amplify the impact of indirect exposure on user stance formation. To account for this change, we interact our instrumental variable (ffs_{it}) with a dummy variable ($TWalg$) equal to 1 from 2016 onwards. We also provide an additional check, where we restrict the sample to the 2013-2016 period only.

Moreover, we control for the adoption of a vaccine mandate in the Emilia-Romagna region since November 25th, 2016 (Regional Law n.19), which followed several outbreaks of infectious diseases affecting non-vaccinated individuals (Gori et al., 2020). The mandate required vaccination certificates for enrollment in public schools and kindergartens. This is captured by an interaction term between ffs_{it} and an indicator variable (ER), which equals 1 for individuals in Emilia-Romagna after the regional law was implemented.

Finally, Italian populist parties have occasionally expressed concerns about vaccine safety (Guriev and Papaioannou, 2022, Kennedy, 2019). Our estimates could thus be capturing a differential effect of political stances rather than disinformation spread. We control for this potential confounder by interacting ffs_{it} with an indicator variable (PP) for a populist party ruling at the municipal level.

Table 8 reports the first-stage results of the above exercises alongside the baseline model (column 1). While we find a significant impact of the introduction of the algorithm (column 2), neither the pre-existing mandate (column 3) nor the influence of populist parties (column 4) seems to play a significant role in affecting users' stance through indirect exposure. For all models, the first-stage results are significant and strongly relevant.

Table 8: M2SLS Individual - First stage.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Main	Twitter	Emilia Romagna	Populist	Network	Excluding FoF	2013-2016
		algorithm	Law	party	distance	geolocated in	
						user's municipality	
	s_{it}	s_{it}	s_{it}	s_{it}	s_{it}	s_{it}	s_{it}
	(30.33)	(30.33)	(30.33)	(30.33)	(30.33)	(30.33)	(30.33)
ffs_{it}	0.704***	0.528***	0.706***	0.691***	0.611***	0.731***	0.512***
	[0.017]	[0.035]	[0.017]	[0.022]	[0.021]	[0.016]	[0.031]
$ffs_{it}\times$ TWalg		0.251***					
		[0.039]					
$ffs_{it}\times$ ER			0.005				
			[0.0742]				
$ffs_{it}\times$ PP				0.048			
				[0.043]			
N	127,754	127,754	127,754	127,754	127,754	127,746	48,180
Controls (Twitter)	✓	✓	✓	✓	✓	✓	✓
Controls (socioec.)	✓	✓	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓	✓	✓
F-stat	1,757.86	998.690	870.815	943.98	875.82	2102.95	266.18

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: The numbers refer to an initial sample of 830,253 tweets to a population of 80,471 unique users across 4220 municipalities. All estimates include city, region and year fixed effects and region-specific time trends fixed effects. Standard errors (in brackets) are clustered at the municipality level. Averages of s_{it} in parentheses is weighted by population size.

Additionally, in order to address any remaining concerns about the exogeneity of the FoPF network, we propose two alternative estimation strategies. Firstly, the network has a hierarchical structure with "ego" users, their passive followings, and the relative FoPF connections. If a FoPF user is linked to the ego user through multiple passive followings, this can weaken the network exogeneity assumption underlying the validity of the IV. To account for this, we reweight the estimates with the inverse of the number of connections that a FoPF shares with the ego user with the following equation:

$$w_i = \frac{1}{\sum_{j=1}^{n} f_{ij}} \tag{6}$$

where f_{ij} is the number of shared nodes between user i and each FoPF j. This weight can be regarded as a proxy for how long the new content takes to spread across the network. The first stage results (column 5) show a slightly decreased coefficient estimate, which however remains comparable to the original one in terms of both magnitude and statistical significance.

Secondly, we exclude all followings-of-followings geolocated in the user's municipality, in order to rule out the possibility that the network might be influenced by common local offline vaccine views. The respective first-stage results (column 6) remain virtually unchanged.

Table 9: M2SLS Municipal - Second stage (Vaccination rate, hospitalizations and healthcare costs).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Main	Twitter algorithm	Emilia Romagna Law	Populist Party Law	Network distance	Excluding FoF geolocated in user's municipality	2013-2016
	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}
Panel a: MMR vaccination rate (89.53)							
s_{mt}	-0.043**	-0.047**	-0.048**	-0.055**	-0.050**	-0.042**	-0.087**
	[0.021]	[0.022]	[0.021]	[0.026]	[0.023]	[0.021]	[0.044]
N	7,238	7,238	7,238	7,238	7,238	7,238	3,137
Panel b: Non-target population							
Hopitalizations							
s_{mt}	0.213*	0.231*	0.204*	0.215*	0.220*	0.205*	0.319*
	[0.113]	[0.121]	[0.112]	[0.112]	[0.115]	[0.108]	[0.167]
Healthcare costs							
s_{mt}	731.1**	821.3**	712.8**	746.5*	794.0**	909.9**	-162.2
	[409.8]	[434.7]	[406.6]	[412.2]	[411.0]	[402.0]	[952.1]
N	3,331	3,331	3,331	3,331	3,331	3,331	3,331
Panel c: Non-target population (MMR)							
Hopitalizations							
s_{mt}	0.234***	0.256***	0.233***	0.231***	0.242***	0.211***	0.320**
	[0.0601]	[0.0675]	[0.0596]	[0.0603]	[0.0621]	[0.0578]	[0.128]
Healthcare costs							
s_{mt}	722.1***	716.7***	725.1***	734.0***	743.7***	713.8***	422.6*
	[243.1]	[250.6]	[242.8]	[247.7]	[247.1]	[235.6]	[214.3]
N	3,331	3,331	3,331	3,331	3,331	3,331	
Panel d: Children age 1-10 (MMR)							
Hopitalizations							
s_{mt}	0.145**	0.150**	0.145**	0.146**	0.142**	0.115*	0.184*
	[0.0650]	[0.0664]	[0.0651]	[0.0653]	[0.0659]	[0.0619]	[0.096]
Healthcare costs							
s_{mt}	366.9**	428.7**	366.5**	363.6**	390.2**	375.5**	233.8*
	[161.1]	[171.8]	[160.9]	[163.9]	[163.7]	[162.3]	[117.1]
N	3,331	3,331	3,331	3,331	3,331	3,331	3,331
Controls (Twitter)	✓	✓	✓	✓	✓	✓	✓
Controls (socioec.)	✓	✓	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: All estimates include city and year fixed effects and region-specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates, as well as averages of V_{mt}, are weighted by the municipality population size.

Table 9 reports the second stage results for all the checks relative to the MMR vaccination rates (panel a), the hospitalization rates and costs for non-target population (panel b), the specific MMR non-target population (panel c) and for children aged 1 to 10 (panel d).[34] All estimates are qualitatively and quantitatively in line with the baseline.

[34] Table A.9 in subsection A.1 reports the (null) results on all vaccination types.

7 Non-linear effects and policy implications

To explore the potential policy implications of our results, we investigate whether there is any non-linearity in the effect of indirect exposure on user stances. Specifically, we look at whether the influence channeled through the followings-of-followings network varies depending on where a user falls in the stance distribution (i.e., whether they are vaccine supporters or skeptics).

Hence, we first re-run our main model specification while classifying user stances into two binary categories: pro-vax users (those with an average anti-vax stance of zero), and anti-vax users (those with an average anti-vax stance of 100). This allows us to better understand the factors that influence vaccine attitudes among these two sub-groups.

Table 10: M2SLS for pro-vax vs. anti-vax users - First stage.

	(1) Pro_{it} (0.495)	(2) $Anti_{it}$ (0.204)
ffs_{it} (28.77)	-0.0076 ***	0 .0046***
	[0 .0003]	[0.0001]
N	127,754	127,754
Controls (Twitter)	✓	✓
Controls (Socioec.)	✓	✓
City and year FE	✓	✓
Reg × year	✓	✓
F-stat	1,765.22	1,763.52

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: The numbers refer to an initial sample of 830,253 tweets to a population of 80,471 unique users across 4220 municipalities. All estimates include city, region and year fixed effects and region specific time trends fixed effect. Standard errors (in brackets) are clustered on municipalities level. Mean values of Pro_{it}, $Anti_{it}$ and $\bar{ffs}_{it}$ in parentheses are weighted by population size.

According to the magnitude of the coefficient estimates presented in Table 10, the exposure to followings-of-followings stances has a stronger effect on pro-vax users compared to anti-vax users. Hence, each unit change in the exposure stance is more likely to increase hesitancy among pro-vax users rather than reduce it among anti-vax users.

Table 11: Results of the Second stage of the M2SLS for pro-vax vs. anti-vax users - Vaccination rates

	(1)	(2)
	M2SLS	M2SLS
	Pro_{mt}	$Anti_{mt}$
	V_{mt}	V_{mt}
Panel a: Hexavalent (94.06)		
	0.4567	0.0674
	[1.4333]	[2.1973]
N	7,239	7,239
Panel b: MMR (89.53)		
	3.9086*	-6.6162*
	[2.1978]	[3.5315]
N	7,238	7,238
Panel c: Menigococcal (81.32)		
	0.5034	-1.6496
	[4.8856]	[8.2071]
N	7,061	7,061
Panel d: Pneumococcal (82.64)		
	2.7584	-4.2443
	[5.3633]	[8.4350]
N	7,066	7,066
Controls (Twitter)	✓	✓
Controls (Socioec.)	✓	✓
City and year FE	✓	✓
Reg × year	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level. Estimates as well as averages of V_{mt} are weighted by the municipality population size.

In fact, the second stage results (Table 11) confirm that the effect on vaccine coverage is more strongly channeled through a shift of users towards anti-vax stances, rather than pro-vax ones. In turn, this suggests that policy interventions aimed at discouraging vaccine hesitancy should be targeted toward reducing the exposure to and the flow of anti-vax content, rather than increasing pro-vax campaigns. However, social media censorship - although effective - has a number of political, social, and ethical implications that go beyond the debate around vaccinations.[35] In addition, recent contributions have shown that these measures can backfire, leading to a larger spread of censored information (Hobbs and Roberts, 2018).

We also consider the role of random events related to epidemics, scientific discoveries, court sentences, policies, and news in mitigating or reinforcing the influence of exposure on user stances. To do that, we manually collected all of the significant events related to vaccines that were discussed in the media during the period of our analysis. These topics include issues such as deaths of children allegedly caused by vaccines or lack of vaccination, court rulings in favor of anti-vax or pro-vax views, the dissemination of scientific evidence for or against vaccines, and political debates about pro- and anti-vax stances. Following Athey et al. (2022),

[35] Twitter acts on complaints by third parties, including governments, to remove illegal content from the platform. In addition, it runs its own content moderation policy, which includes actions like user suspension, content removal, and permanent bans in response to violations of the terms of use (https://help.twitter.com/en/rules-and-policies/twitter-rules). Current allegations against Twitter policies include partisan implementation of moderation rules and arbitrary or politically biased use of bans.

we manually classify these online debates into four broad domains: vaccine efficacy, statements from trustful sources, politics and mandates, and allegations that vaccines are unsafe.[36]

Table 12: User exposure to FoPF stances and the role of online debates topics.

	(1) s_{it} (30.31)	(2) Pro_{it} (0.495)	(2) $Anti_{it}$ (0.204)
ffs_{it}	0.2884***	-0.3309***	0.2295***
	[0.0693]	[0.0757]	[0.0728]
$ffs_{it} \times Efficacy$	-0.3425	0.3765	-0.3548
	[0.2724]	[0.2754]	[0.2961]
$ffs_{it} \times Trustful\, Source$	-0.3136***	0.2656**	-0.3805***
	[0.0992]	[0.1127]	[0.1057]
$ffs_{it} \times Politics\, and\, Mandate$	-0.1749***	0.0660	-0.3899***
	[0.0530]	[0.0408]	[0.0589]
$ffs_{it} \times Vaccines\, Unsafe$	-0.0697	0.1369	-0.0387
	[0.2292]	[0.2442]	[0.2495]
N	531,352	531,352	531,352
User FE	✓	✓	✓
Date FE	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Notes: The numbers refer to an initial sample of 830,253 tweets to a population of 80,471 unique users across 4220 municipalities. All estimates include individual and daily date fixed effects. Standard errors (in brackets) are clustered at the individual. Mean values of s_{it}, Pro_{it}, and $Anti_{it}$ in parentheses are weighted by population size.

Table 12 presents estimates of daily-level user stances on vaccines, conditional on user and daily date fixed effects. These estimates show how individual stances fluctuate as a function of their followings-of-followings' stances on regular days, and on days when specific events related to vaccines are debated on Twitter. Column 1 shows that, after controlling for individual fixed tendencies and day-specific features of Twitter activity, individual stances tend to be influenced by the stances of their followings-of-followings, consistent with our baseline first-stage result. Exposure to anti-vax content tends to make individuals more lenient towards such stances. However, this relationship is moderated (and even reversed) on days when credible sources, such as the World Health Organization, the academic or research community, the European Commission, or a court, issue statements in favor of vaccines. Similarly, on days when political debates about the effectiveness of vaccines are discussed on Twitter, the influence of exposure to anti-vax stances is mitigated.

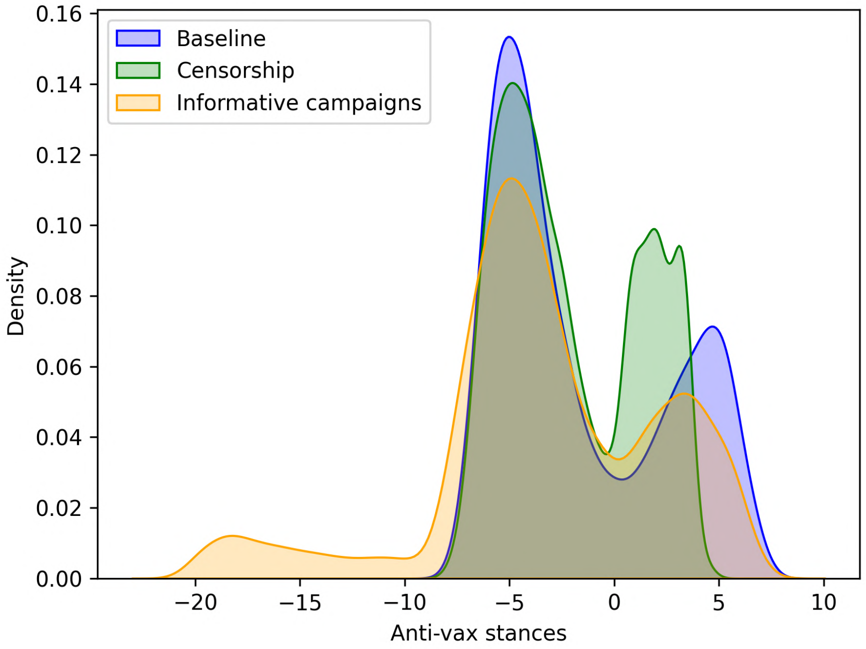
When we classify user stances into binary categories (pro-vax and anti-vax), we find that the mitigating effect of exposure to anti-vax content is more pronounced in the anti-vax category (column 3). Events involving statements from trustworthy sources and political debates generally have the ablility to offset the influence of exposure to anti-vax stances or reinforce the influence of exposure to pro-vax content.

We use our sketched model to assess the potential effects of interventions on social media platforms. Building on the previous findings, we conduct two types of counterfactual exercises that are symmetrical in imple-

[36] The full list of events are reported in Table A.11 in Appendix A.

mentation but differ in their interpretation. On the one hand, we look at the censorship effect on anti-vax stances by exogenously reducing the activity rate of users whose stance exceeds the 90^{th} percentile by half (*Censorship*). On the other hand, we investigate the effects of broadening the reach of pro-vax activists by doubling the number of contacted users for pro-vax activists in the first decile of the distribution (*Informative campaigns*).[37] The results of the exercises are reported in Figure 9, where we plot the converged density distribution of users' stances for the baseline exercise (panel (c) of Figure 5) as well as for the Censorship and for the Informative campaign counterfactuals. The figure illustrates two key findings, consistent with our estimates: firstly, both interventions result in a reduction in the peak and the number of anti-vax users; secondly, the informative campaigns, more so than the censorship intervention, contribute to a decrease in overall polarization. It is important to note that the varying effects are influenced by the initial distribution's asymmetry, which aligns with the observed data.

Figure 9: Policy counterfactuals - Monte Carlo results



Notes: converged density distributions of users' stances (N=T=500) - average over 100 Monte Carlo runs. We report the baseline model (blue), the "Censorship" counterfactual exercise (green) in which we halve the activity rate of users in the upper decile of the stance distribution, and the "Informative campaigns" counterfactual exercise (orange) in which we double the activity rate for users in the first decile of the stance distribution.

Indeed, informative campaigns targeting vaccines have the potential to be a highly effective and scalable intervention for promoting public health awareness. Their impact is particularly significant when these campaigns

[37] With these exercises, we simulate two distinct policies. First, we replicate a scenario where the platform flags and reduces the visibility of tweets based on their content. Second, we simulate an informative pro-vax campaign, potentially sponsored by the government or another public entity, which by design allows for increased reach.

are perceived as originating from trusted sources and are supported by political interventions.

8 Conclusions

Between 2013 and 2018, Italy experienced significant changes in pediatric vaccine coverage rates, partly due to the dissemination of misinformation regarding the safety of MMR vaccines. The vaccine hesitancy has contributed to outbreaks of infectious diseases, leading to the implementation and legal enforcement of a mandate for a wide range of pediatric vaccines in 2017. In this study, we utilize a rich dataset of online interactions and employ state-of-the-art natural language processing techniques to quantitatively examine the tangible costs that misinformation and disinformation impose on society.

The negative consequences of spreading fake or unverified news have been largely discussed in academic, political and media circles, especially in the context of the COVID-19 crisis. While it is known that clusters of conspiracy theories serve as fertile ground for the proliferation of fake news, and that online activities, particularly on social media platforms, can have harmful effects such as hate crimes (Müller and Schwarz, 2021) and influence electoral outcomes (Fujiwara et al., 2021), this paper makes several contributions to the ongoing debate. Firstly, we develop a method to estimate the causal effects of individual-level online interactions on observable aggregate outcomes. Secondly, we estimate the actual costs incurred by healthcare systems due to online anti-vaccine activity. Lastly, we provide data-driven insights and simulations on how to mitigate the spread of anti-scientific or unsubstantiated content on social media.

In relation to the latter point, our findings indicate that individuals who advocate for vaccination are more influenced by exposure to vaccine-related content compared to their anti-vaccine counterparts. Conversely, anti-vaccine individuals are responsive to statements from trusted sources. These results suggest that informative campaigns targeting vaccines, both online and offline, have the potential to effectively combat vaccine hesitancy. Even though it may be challenging to change the views of vaccine skeptics, such campaigns can counteract the persuasive impact of anti-vaccine content on pro-vaccine individuals.

In conclusion, our policy insights offer a viable approach to addressing the decline in vaccine coverage without resorting to coercive measures like vaccine mandates. Our findings suggest that while the legal enforcement may address the immediate effects of vaccine hesitancy on coverage rates and associated health costs, it also leads to polarization and radicalization of opinions, which are long-lasting and can perpetuate themselves when coupled with echo chambers. Therefore, policymakers must consider these potential consequences to prevent vaccine-enhancing measures from backfiring once legal enforcement is lifted.

Baumann et al. (2021) suggest that when debated topics overlap thematically, increases in controversialness can lead to the emergence of ideological states where multiple stances align within a common, “political” stance. In their model, ideology emerges endogenously from uncorrelated polarization, achieved by relaxing the unrealistic assumption of topic orthogonality. In this paper’s analysis of pediatric vaccines from 2013 to 2018, fake news related to vaccinations was limited to the debate on the vaccine-autism causation. However, today the topic is no longer uncorrelated to other salient debates. The controversy surrounding the COVID-19 pandemic has created an ideological state that covers a wide range of topics including vaccines, face masks, mobility restrictions, and ultimately political opinions. Finding a way to deescalate the debates around scientifically grounded topics can prove to be a viable way to reduce the polarization and foster constructive discussions.

References

Abrevaya, J. and K. Mulligan (2011). Effectiveness of state-level vaccination mandates: evidence from the varicella vaccine. *Journal of health economics 30*(5), 966–976.

Acemoglu, D., A. Ozdaglar, and J. Siderius (2021). A model of online misinformation. Technical report, National Bureau of Economic Research.

Alatas, V., A. G. Chandrasekhar, M. Mobius, B. A. Olken, and C. Paladines (2019). When celebrities speak: A nationwide twitter experiment promoting vaccination in indonesia. Technical report, National Bureau of Economic Research.

Albanese, G., G. Barone, and G. de Blasio (2022). Populist voting and losers' discontent: Does redistribution matter? *European Economic Review 141*, 104000.

Allam, A., P. J. Schulz, and K. Nakamoto (2014). The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating google output. *Journal of medical internet research 16*(4), e100.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). The welfare effects of social media. *American Economic Review 110*(3), 629–676.

Allcott, H. and M. Gentzkow (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives 31*(2), 211–36.

Allcott, H., M. Gentzkow, and C. Yu (2019). Trends in the diffusion of misinformation on social media. *Research & Politics 6*(2), 2053168019848554.

Athey, S., K. Grabarz, M. Luca, and N. C. Wernerfelt (2022). The effectiveness of digital interventions on covid-19 attitudes and beliefs. Technical report, National Bureau of Economic Research.

Azzimonti, M. and M. Fernandes (2022). Social media networks, fake news, and polarization. *European Journal of Political Economy*, 102256.

Bailey, M., D. M. Johnston, M. Koenen, T. Kuchler, D. Russel, and J. Stroebel (2020). Social networks shape beliefs and behavior: Evidence from social distancing during the covid-19 pandemic. Technical report, National Bureau of Economic Research.

Baumann, F., P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini (2020). Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters 124*(4), 048301.

Baumann, F., P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini (2021). Emergence of polarized ideological opinions in multidimensional topic spaces. *Physical Review X 11*(1), 011012.

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science 47*(2), 241–262.

Bessi, A., F. Petroni, M. D. Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi (2016). Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics 225*(10), 2047–2059.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012). A 61-million-person experiment in social influence and political mobilization. *Nature 489*(7415), 295–298.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of econometrics 150*(1), 41–55.

Bramoullé, Y., H. Djebbari, and B. Fortin (2020). Peer effects in networks: A survey. *Annual Review of Economics 12*, 603–629.

Breza, E., F. C. Stanford, M. Alsan, B. Alsan, A. Banerjee, A. G. Chandrasekhar, S. Eichmeyer, T. Glushko, P. Goldsmith-Pinkham, K. Holland, et al. (2021). Doctors' and nurses' social media ads reduced holiday travel and covid-19 infections: A cluster randomized controlled trial. Technical report, National Bureau of Economic Research.

Burki, T. (2019). Vaccine misinformation and social media. *The Lancet Digital Health 1*(6), e258–e259.

Bursztyn, L., F. Ederer, B. Ferman, and N. Yuchtman (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica 82*(4), 1273–1301.

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2019). Social media and xenophobia: evidence from russia. Technical report, National Bureau of Economic Research.

Bursztyn, L., G. Egorov, and R. Jensen (2019). Cool to be smart or smart to be cool? understanding peer pressure in education. *The Review of Economic Studies 86*(4), 1487–1526.

Cagé, J., N. Hervé, and B. Mazoyer (2022). Social media and newsroom production decisions.

Carpenter, C. S. and E. C. Lawler (2019). Direct and spillover effects of middle school vaccination requirements. *American Economic Journal: Economic Policy 11*(1), 95–125.

Carrieri, V., L. Madio, and F. Principe (2019). Vaccine hesitancy and (fake) news: Quasi-experimental evidence from italy. *Health economics*.

Chiou, L. and C. Tucker (2018). Fake news and advertising on social media: A study of the anti-vaccination movement. Technical report, National Bureau of Economic Research.

Cinelli, M., G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences 118*(9), e2023301118.

De Giorgi, G., M. Pellizzari, and S. Redaelli (2010). Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics 2*(2), 241–275.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018a). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018b). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhrymes, P. J. and A. Lleras-Muney (2006). Estimation of models with grouped and ungrouped data by means of "2sls". *Journal of econometrics 133*(1), 1–29.

Draca, M. and C. Schwarz (2021). How polarized are citizens? measuring ideology from the ground-up. *Working Paper*.

Enikolopov, R., A. Makarin, and M. Petrova (2020). Social media and protest participation: Evidence from russia. *Econometrica 88*(4), 1479–1514.

Esposito, S., P. Durando, S. Bosis, F. Ansaldi, C. Tagliabue, G. Icardi, E. V. S. Group, et al. (2014). Vaccine-preventable diseases: from paediatric to adult targets. *European journal of internal medicine 25*(3), 203–212.

Flaxman, S., S. Goel, and J. M. Rao (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly 80*(S1), 298–320.

Fujiwara, T., K. Müller, and C. Schwarz (2021). The effect of social media on elections: Evidence from the united states. Technical report, National Bureau of Economic Research.

Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics 126*(4), 1799–1839.

Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*(3), 253–264.

Gori, D., C. Costantino, A. Odone, B. Ricci, M. Ialonardi, C. Signorelli, F. Vitale, and M. P. Fantini (2020). The impact of mandatory vaccination law in italy on mmr coverage rates in two of the largest italian regions (emilia-romagna and sicily): an effective strategy to contrast vaccine hesitancy. *Vaccines 8*(1), 57.

Grossman, G., S. Kim, J. M. Rexer, and H. Thirumurthy (2020). Political partisanship influences behavioral responses to governors' recommendations for covid-19 prevention in the united states. *Proceedings of the National Academy of Sciences 117*(39), 24144–24153.

Guriev, S., N. Melnikov, and E. Zhuravskaya (2021). 3g internet and confidence in government. *The Quarterly Journal of Economics 136*(4), 2533–2613.

Guriev, S. and E. Papaioannou (2022). The political economy of populism. *Journal of Economic Literature (forthcoming).*.

Hobbs, W. R. and M. E. Roberts (2018). How sudden censorship can increase access to information. *American Political Science Review 112*(3), 621–636.

Holtkamp, N. C. et al. (2021). *The Economic and Health Effects of the United States' Earliest School Vaccination Mandates*. Ph. D. thesis.

Huszár, F., S. I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, and M. Hardt (2022). Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences 119*(1), e2025334119.

Jin, Z., Z. Peng, T. Vaidhya, B. Schoelkopf, and R. Mihalcea (2021). Mining the cause of political decision-making from social media: A case study of covid-19 policies across the us states. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 288–301.

Johnson, K. and D. Goldwasser (2016, November). Identifying stance by analyzing political discourse on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, Austin, Texas, pp. 66–75. Association for Computational Linguistics.

Johnsson, I. and H. R. Moon (2021). Estimation of peer effects in endogenous social networks: Control function approach. *The Review of Economics and Statistics 103*(2), 328–345.

Jolley, D. and K. M. Douglas (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PloS one 9*(2), e89177.

Kartal, M. and J.-R. Tyran (2022). Fake news, voter overconfidence, and the quality of democratic choice. *American Economic Review 112*(10), 3367–97.

Kennedy, J. (2019). Populist politics and vaccine hesitancy in western europe: an analysis of national-level data. *European journal of public health 29*(3), 512–516.

Kim, T. (2022). Measuring police performance: Public attitudes expressed in twitter. In *AEA Papers and Proceedings*, Volume 112, pp. 184–87.

Larsen, B., T. J. Ryan, S. Greene, M. J. Hetherington, R. Maxwell, and S. Tadelis (2022). Counter-stereotypical messaging and partisan cues: moving the needle on vaccines in a polarized us. Technical report, NBER Working Paper, Stanford University, Palo Alto, CA, 2022). https://www. nber ....

Lawler, E. C. (2017). Effectiveness of vaccination recommendations versus mandates: Evidence from the hepatitis a vaccine. *Journal of health economics 52*, 45–62.

Lazer, D. M., M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. (2018). The science of fake news. *Science 359*(6380), 1094–1096.

Leask, J., S. Chapman, P. Hawe, and M. Burgess (2006). What maintains parental support for vaccination when challenged by anti-vaccination messages? a qualitative study. *Vaccine 24*(49-50), 7238–7245.

Lorenz-Spreen, P., B. M. Mønsted, P. Hövel, and S. Lehmann (2019). Accelerating dynamics of collective attention. *Nature communications 10*(1), 1–9.

Martinez, L. S., S. Hughes, E. R. Walsh-Buhi, and M.-H. Tsou (2018). "okay, we get it. you vape": an analysis of geocoded content, context, and sentiment regarding e-cigarettes on twitter. *Journal of health communication 23*(6), 550–562.

Michaels, D. (2008). *Doubt is their product: how industry's assault on science threatens your health*. Oxford University Press.

Mullainathan, S. and A. Shleifer (2005). The market for news. *American Economic Review 95*(4), 1031–1053.

Müller, K. and C. Schwarz (2020). From hashtag to hate crime: Twitter and anti-minority sentiment. *Working Paper*.

Müller, K. and C. Schwarz (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association 19*(4), 2131–2167.

Opel, D. J., J. A. Taylor, R. Mangione-Smith, C. Solomon, C. Zhao, S. Catz, and D. Martin (2011). Validity and reliability of a survey to identify vaccine-hesitant parents. *Vaccine 29*(38), 6598–6605.

Perra, N., B. Gonçalves, R. Pastor-Satorras, and A. Vespignani (2012). Activity driven modeling of time varying networks. *Scientific reports 2*(1), 1–7.

Pierri, F., A. Artoni, and S. Ceri (2020). Investigating italian disinformation spreading on twitter in the context of 2019 european elections. *PloS one 15*(1), e0227821.

Polignano, M., P. Basile, M. De Gemmis, G. Semeraro, and V. Basile (2019). Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, Volume 2481, pp. 1–6. CEUR.

Qin, B., D. Strömberg, and Y. Wu (2017). Why does china allow freer social media? protests versus surveillance and propaganda. *Journal of Economic Perspectives 31*(1), 117–140.

See, A., P. Liu, and C. Manning (2017). Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics*.

Shao, C., G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer (2018). The spread of low-credibility content by social bots. *Nature communications 9*(1), 1–9.

Siegal, G., N. Siegal, and R. J. Bonnie (2009). An account of collective actions in public health. *American Journal of Public Health 99*(9), 1583–1587.

Smith, L. E., R. Amlôt, J. Weinman, J. Yiend, and G. J. Rubin (2017). A systematic review of factors affecting vaccine uptake in young children. *Vaccine 35*(45), 6059–6069.

Sunstein, C. R. (2001). *Republic. com*. Princeton university press.

Sunstein, C. R. (2017). *# Republic: Divided democracy in the age of social media*. Princeton: Princeton University Press.

Sunstein, C. R. (2018). *The cost-benefit revolution*. MIT Press.

Vosoughi, S., D. Roy, and S. Aral (2018). The spread of true and false news online. *Science 359*(6380), 1146–1151.

Zhuravskaya, E., M. Petrova, and R. Enikolopov (2020). Political effects of the internet and social media. *Annual Review of Economics 12*(1), 415–438.

Appendix A

A.1 Additional Tables

Table A.1 provides an overview of the statistical data pertaining to the characteristics of the municipality.

Table A.1: Descriptive statistics of municipality's characteristics

	Median	Mean	sd	Min	Max
Avg. mother's age at birth	31.92	31.82	0.31	30.32	32.81
Health public cost pc (€)	1,911.00	1,903.89	56.37	1,662.00	2,515.00
Income pc (€)	9,183.32	10,854.95	3,786.64	1,986.88	84,253.34
Lower secondary school attainment (%)	86.41	85.30	2.22	74.36	87.73
Birth rate (%)	7.30	7.38	0.64	5.40	10.70
Populist party	1.00	0.58	0.49	0.00	1.00

Notes: The statistics are weighted by the municipality population size.

A.2 Balance tests

Table A.2 shows the results of an instrument balance test run by regressing municipal characteristics on the instrument. Each column refers to a variable related to financial (health expenditure, income), social (education, average mother's age at birth, birth rate) or political (populist parties vote share) characteristics.

Table A.2: Instrument balance tests

	(1)	(2)	(3)	(4)	(5)	(6)
	Health public cost per capita (€)	Income per capita (€)	Lower secondary school att. (%)	Avg. mother's age at birth	Birth rate	Populist party
Panel a: geolocated in the same user's municipality						
ffs_{it}	-0.0211	-0.403	0.0001	0.0001	-0.0002	0.0002
	[0.0246]	[0.442]	[0.0002]	[0.0001]	[0.0002]	[0.0002]
N	110,639	110,639	110,639	110,589	110,639	110,639
Panel b: geolocated in municipalities different from the user's municipality						
ffs_{it}	-0.0001	-0.447	-0.0001	-0.0001	-0.00002	0.0001
	[0.0126]	[0.337]	[0.0004]	[0.0001]	[0.0001]	[0.0001]
N	131,003	131,003	131,003	130,817	131,003	131,003
Panel c: not geolocated						
ffs_{it}	0.0037	1.001	-0.00004	-0.00001	0.0001	0.0002
	[0.0121]	[0.912]	[0.0002]	[0.00003]	[0.0001]	[0.0002]
N	130,977	130,977	130,977	130,791	130,977	130,977
CITY and YEAR FE	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: The figures in each row are estimated coefficients from regressions of municipal characteristics on FoPF stances. Standard errors (in brackets) are clustered at the municipality level.

A.3 Reduced Form

Table A.3 and Table A.4 reports the reduced form results for the vaccination rates, hospitalizations and average annual costs for vaccine preventable diseases, respectively.

Table A.3: Reduced form - Vaccination rates

	(1) V_{mt} Hexavalent	(2) V_{mt} MMR	(3) V_{mt} Meningococcus	(4) V_{mt} Pneumococcus
ffs_{mt}	-0.007	-0.038**	-0.022	-0.071
	[0.012]	[0.019]	[0.049]	[0.055]
N	7,239	7,238	7,061	7,066
Controls (Twitter)	✓	✓	✓	✓
Controls (Socioec.)	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: All estimates include city, region and year fixed effects and region-specific time trends fixed effects. Standard errors (in brackets) are clustered on the municipality level. Estimates are weighted by municipality population size.

Table A.4: Reduced form - Hospitalizations.

	(1)	(2)	(3)
	V_{mt}	V_{mt}	V_{mt}
	non-target pop.	non-target pop.(MMR)	Children age 1-10 (MMR)
Panel a: Hopitalizations			
s_{mt}	0.123**	0.104***	0.0603*
	[0.0550]	[0.0309]	[0.0323]
Panel b: Healthcare costs			
s_{mt}	383.7*	326.7**	147.0*
	[203.9]	[146.2]	[78.60]
	(4)	(5)	(6)
	(Hexav.)	(Meningo.)	(Pneumo.)
	Non-target population		
Panel c: Hopitalizations			
s_{mt}	0.0266	-0.0002	-0.005
	[0.0432]	[0.0005]	[0.0079]
Panel d: Healthcare costs			
s_{mt}	-138.3	-5.515	-17.42
	[340.2]	[8.761]	[21.84]
	Children age 1-10		
Panel e: Hopitalizations			
s_{mt}	0.0005	0.00008	0.006
	[0.0097]	[0.0019]	[0.0074]
Panel f: Healthcare costs			
s_{mt}	-32.78	5.163	0.478
	[26.40]	[7.744]	[23.39]
N	5,136	5,136	5,136
CONTROL (Twitter)	✓	✓	✓
CONTROL (Socioec.)	✓	✓	✓
CITY and YEAR FE	✓	✓	✓
Reg × Year	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Notes*: All estimates include city and year fixed effects and region-specific time trends. Standard errors (in brackets) are clustered at the municipality level. Estimates are weighted by the municipality population size.

A.4 Results - Vaccination Rates (full set of estimates)

Table A.5 shows second stage estimates related to vaccination rates under several specifications.

Table A.5: Results of the Second stage of the OLS and the M2SLS - Vaccination rates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	OLS	M2SLS	M2SLS	M2SLS	M2SLS	M2SLS	M2SLS
	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}
Panel a: Hexavalent (94.06)							
s_{mt}	-0.001	-0.033	-0.005	-0.004	-0.003	-0.008	-0.023
	[0.002]	[0.026]	[0.015]	[0.015]	[0.015]	[0.016]	[0.015]
N	7,239	7,601	7,239	7,239	7,239	7,239	7,239
Panel b: MMR (89.53)							
s_{mt}	-0.005	-0.157***	-0.045*	-0.037	-0.041*	-0.048*	-0.043**
	[0.003]	[0.044]	[0.026]	[0.024]	[0.024]	[0.027]	[0.021]
N	7,238	7,600	7,238	7,238	7,238	7,238	7,238
Panel c: Menigococcal (81.32)							
s_{mt}	-0.002	-0.470***	-0.030	-0.001	-0.013	-0.009	-0.040
	[0.008]	[0.128]	[0.066]	[0.062]	[0.064]	[0.062]	[0.054]
N	7,061	7,438	7,061	7,061	7,061	7,061	7,061
Panel d: Pneumococcal (82.64)							
s_{mt}	-0.0001	-0.206**	-0.060	-0.032	-0.046	-0.071	-0.029
	[0.008]	[0.086]	[0.072]	[0.063]	[0.067]	[0.069]	[0.052]
N	7,066	7,429	7,066	7,066	7,066	7,066	7,066
CONTROL (Twitter)	✓				✓		✓
CONTROL (socioeconomics)	✓					✓	✓
YEAR FE	✓	✓	✓	✓	✓	✓	✓
CITY FE	✓		✓	✓	✓	✓	✓
Reg × Year	✓			✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates as well as averages of V_{mt} are weighted by the municipality population size.

A.5 Results - Hexavalent, Meningococcus and Pneumococcus Hospitalizations

Table A.6 reports the estimates for the second stage, indicating the number of hospitalizations and the average annual costs associated with administering Hexavalent, Meningococcal, and Pneumococcal vaccinations.

Table A.6: Results of the OLS and the Second stage of the M2SLS - Hospitalizations.

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	M2SLS	OLS	M2SLS	OLS	M2SLS
	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}
	(Hexav.)	(Hexav.)	(Meningo.)	(Meningo.)	(Pneumo.)	(Pneumo.)
	Non-target population					
Panel a: Hopitalizations						
s_{mt}	0.009	0.025	-0.0001	-0.0003	-0.0006	-0.021
	[0.012]	[0.092]	[0.0002]	[0.0009]	[0.002]	[0.015]
Panel b: Healthcare costs						
s_{mt}	102.0	-628.4	-4.756	-20.81	-10.53*	-46.519
	[100.6]	[700.3]	[3.976]	[16.46]	[6.103]	[37.26]
	Children age 1-10					
Panel a: Hopitalizations						
s_{mt}	-0.0001	0.002	0.0001	0.0003	-0.002	0.009
	[0.003]	[0.016]	[0.001]	[0.004]	[0.002]	[0.011]
Panel b: Healthcare costs						
s_{mt}	12.74	-66.18	-0.528	10.36	-3.788	-37.99
	[18.45]	[49.21]	[2.887]	[14.90]	[6.229]	[42.28]
N	3,331	3,331	3,331	3,331	3,331	3,331
Controls (Twitter)	✓	✓	✓	✓	✓	✓
Controls (Socioecon.)	✓	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates are weighted by the municipality population size.

A.6 Results - Mortality among hospitalized

Descriptive statistics and second stage estimates for mortality cases of hospitalized patients with vaccine-preventable diseases are presented in Table A.7 and Table A.8, respectively.

Table A.7: Descriptive statistics of mortality among hospitalized due to vaccine-preventable diseases (2013-2016)

	Median	Mean	sd	Min	Max	N
Panel a: Hopitalizations						
Non-target population	0.00	1.00	2.08	0.00	97.56	31,760
Non-target population (Hexav.)	0.00	0.93	1.98	0.00	97.56	31,760
Non-target population (Meningo.)	0.00	0.00	0.15	0.00	29.02	31,760
Non-target population (MMR)	0.00	0.04	0.38	0.00	28.92	31,760
Non-target population (Pneumo.)	0.00	0.14	0.68	0.00	97.56	31,760
Children age 1-10	0.00	0.01	0.12	0.00	7.28	31,760
Children age 1-10 (Hexav.)	0.00	0.00	0.06	0.00	5.53	31,760
Children age 1-10 (Meningo.)	0.00	0.00	0.07	0.00	6.67	31,760
Children age 1-10 (MMR)	0.00	0.00	0.05	0.00	7.28	31,760
Children age 1-10 (Pneumo.)	0.00	0.00	0.04	0.00	2.81	31,760

Notes: The statistics refer to 7,940 municipalities for the time period between 2013-2016 and are weighted by the municipality population size.

Table A.8: Results of the Second stage of the M2SLS - Mortality among Hospitalized.

	(1) V_{mt} Main	(2) V_{mt} Hexavalent	(3) V_{mt} MMR	(4) V_{mt} Meningococcus	(5) V_{mt} Pneumococcus
Panel a: Non-target population					
Hospitalizations					
s_{mt}	-0.013	-0.019	0.005	-0.0006	-0.006
	[0.014]	[0.013]	[0.003]	[0.0004]	[0.005]
N	3,331	3,331	3,331	3,331	3,331
Panel a: Children age 1-10					
Hospitalizations					
s_{mt}	-0.000005	0.0002	0.00006	-0.0002	-0.0001
	[0.0005]	[0.0002]	[0.00005]	[0.0003]	[0.0002]
N	3,331	3,331	3,331	3,331	3,331
Controls (Twitter)	✓	✓	✓	✓	✓
Controls (socioec.)	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: All estimates include city and year fixed effects and region-specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates are weighted by the municipality population size.

Results - robustness checks

For completeness, we report here the full tables of robustness checks, including the second-stage results on all vaccinations.

Table A.9: M2SLS Individual - Second stage (Vaccination rate)

	(1) Main	(2) Twitter algortithm	(3) Emilia Romagna Law	(4) Populist Party Law	(5) Network distance	(6) Excluding FoF geolocated in user's municipality	(7) 2013-2016
	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}
Panel a: Hexavalent (94.06)							
s_{mt}	-0.023	-0.021	-0.024	-0.022	-0.023	-0.018	-0.041
	[0.015]	[0.016]	[0.015]	[0.019]	[0.015]	[0.016]	[0.032]
N	7,239	7,239	7,239	7,239	7,239	7,239	3,139
Panel b: MMR (89.53)							
s_{mt}	-0.043**	-0.047**	-0.048**	-0.055**	-0.050**	-0.042**	-0.087**
	[0.021]	[0.022]	[0.022]	[0.027]	[0.023]	[0.021]	[0.0441]
N	7,238	7,238	7,238	7,238	7,238	7,238	3,137
Panel c: Menigococcus (81.32)							
s_{mt}	-0.040	-0.044	-0.041	-0.026	-0.038	-0.043	0.001
	[0.054]	[0.057]	[0.054]	[0.069]	[0.0559]	[0.056]	[0.0964]
N	7,061	7,061	7,061	7,061	7,061	7,061	3,023
Panel d: Pneumococcus (82.64)							
s_{mt}	-0.029	-0.035	-0.031	-0.104	-0.027	-0.004	-0.193
	[0.057]	[0.056]	[0.057]	[0.083]	[0.060]	[0.054]	[0.155]
N	7,066	7,066	7,066	7,066	7,066	7,066	3,029
Controls (Twitter)	✓	✓	✓	✓	✓	✓	✓
Controls (Socioec.)	✓	✓	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: All estimates include city and year fixed effects and region-specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates, as well as averages of V_{mt}, are weighted by the municipality population size.

Results - full sample

For the sake of clarity, we present the findings for the full sample of geolocated tweets, including the first percentile of municipalities that were left out of our primary findings.

Table A.10: Results of the M2SLS - Vaccination rate, Hospitalizations and Costs (full sample)

Panel a: First stage

	s_{it} (30.33)	s_{it} (30.33)	s_{it} (30.33)	s_{it} (30.33)
ffs_{it} (28.77)	0.709***	0.709***	0.710***	0.710***
	[0.0164]	[0.0164]	[0.0164]	[0.0164]
N	130,896	130,896	130,896	130,896
Controls (Twitter)		✓		✓
Controls (socioec.)			✓	✓
City and year FE	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓
F-stat	1,875.96	1,874.25	1,865.62	1,868.15

Panel b: Second stage - Vaccination Rate

	(1) V_{mt} Hexavalent	(2) V_{mt} MMR	(3) V_{mt} Meningococcus	(4) V_{mt} Pneumococcus
OLS				
s_{mt}	-0.0003	-0.003	-0.005	-0.006
	[0.001]	[0.002]	[0.005]	[0.005]
M2SLS				
s_{mt}	-0.014	-0.030**	-0.047	-0.012
	[0.010]	[0.014]	[0.035]	[0.034]
N	10,281	10,275	9,978	9,994

Panel c: Second stage - Non-target Population

	(1) V_{mt} Main	(2) V_{mt} Hexavalent	(3) V_{mt} MMR	(4) V_{mt} Meningococcus	(5) V_{mt} Pneumococcus
Hopitalizations					
s_{mt}	0.181**	0.0440	0.150***	-0.000370	-0.00788
	[0.0779]	[0.0610]	[0.0436]	[0.000677]	[0.0111]
Healthcare costs					
s_{mt}	585.2**	-191.6	464.3**	-8.797	-24.01
	[286.3]	[478.9]	[205.1]	[11.74]	[30.78]
N	5,136	5,136	5,136	5,136	5,136

Panel d: Second stage - Children age 1-10

		(1) V_{mt} Hexavalent	(2) V_{mt} MMR	(3) V_{mt} Meningococcus	(4) V_{mt} Pneumococcus
Hopitalizations					
s_{mt}		-0.0003	0.0846*	-0.0001	0.008
		[0.0137]	[0.0452]	[0.00266]	[0.0104]
Healthcare costs					
s_{mt}		-48.82	206.8*	6.494	-2.958
		[37.29]	[110.0]	[10.97]	[33.27]
N		5,136	5,136	5,136	5,136
Controls (Twitter)	✓	✓	✓	✓	✓
Controls (socioec.)	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Notes: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates, as well as the s_{it} are weighted by the municipality population size.

List of events and classification

Table A.11: List of events

date	Classification	Description
14jan2013	Efficacy	Vaccino anti-meningite B,protezione in 95% vaccinati
28jan2013	Efficacy	Bimba di 3 anni muore di setticemia forse provocata da pneumococco
17apr2013	Efficacy	In Italia migliaia di casi di morbillo e rosolia evitabili
02may2013	Efficacy	Epidemia di morbillo nel Regno Unito: dobbiamo preoccuparci?
15may2013	Efficacy	Sanita': Napoli; vaccini gratis contro rosolia congenita
10jun2013	Efficacy	Veneto: casi di complicanze da morbillo e varicella in persone non vaccinate
13jun2013	Efficacy	Sanita': polmonite, in Lombardia costa 68 mln di euro l'anno
26jun2013	Efficacy	Arriva vaccino contro meningite ceppo B
27jun2013	Efficacy	Meningite: Veneto; e' 10% tutte malattie invasive da batteri
24sep2013	Efficacy	In Lombardia campagna vaccinazione Hpv anche per uomini
12may2016	Efficacy	Morbillo in Campania: allerta dei medici per le basse coperture vaccinali nella Regione
08jul2016	Efficacy	Documento sui vaccini della Fnomceo
23jun2017	Efficacy	Un bimbo malato di leucemia morto per il morbillo: "Contagiato dai fratelli non vaccinati"
21may2013	Politics and Mandate	Proposta di Legge d'iniziativa del deputato Burtone Istituzione della Giornata in ricordo delle persone decedute o rese disabili a causa di vaccinazioni
01jul2016	Politics and Mandate	introduzione obbligo vaccinale per gli asili nido in Emilia Romagna
23nov2016	Politics and Mandate	introduzione obbligo vaccinale per gli asili nido in Emilia Romagna
26jan2017	Politics and Mandate	accordo stato-regioni per una legge nazionale sui vaccini
03may2017	Politics and Mandate	Tutte le volte del Movimento 5 Stelle contro i vaccini, la Rete smentisce Grillo
20may2017	Politics and Mandate	Veneto Vaccini, torna l'obbligo. Zaia: «Misura inefficace». I virologi: «Salva la vita»
07jun2017	Politics and Mandate	decreto legge n 73 /2017 "legge lorenzin"
08jun2017	Politics and Mandate	Alto Adige, dove il consiglio provinciale ha approvato all'unanimità una mozione che chiede "lo stralcio delle misure coercitive previste dal decreto sui vaccini e una campagna di sensibilizzazione ampia ed equilibrata
28jul2017	Politics and Mandate	Il Decreto vaccini è legge, tutte le novità
10jan2018	Politics and Mandate	Vaccini, Salvini: "Con noi al governo via l'obbligo". Lorenzin: "Per qualche voto gioca con la salute dei bambini"
22jun2018	Politics and Mandate	Vaccini, Salvini: 'Inutili 10 vaccini obbligatori'. Burioni: 'Bugie pericolosissime'. Alt Di Maio e della Grillo
05aug2018	Politics and Mandate	Taverna, la sciamannata vicepresidente del Senato: i vaccini? Come i marchi alle bestie
12aug2018	Politics and Mandate	Salvini sa che i soldati devono vaccinarsi? Mettetevi d'accordo" Ironia social sulla # LevaObbligatoria
06sep2018	Politics and Mandate	Vaccini a scuola, colpo di scena: emendamento ripristina l'autocertificazione
08jan2013	Trustful Source	Pneumococco. L'EU estende l'uso di Prevenar 13 a bambini e adolescenti fino a 17 anni
17sep2013	Trustful Source	Oms, nessun legame tra vaccini e autismo
26nov2014	Trustful Source	Lorenzin condanna il giudice del tribunale del lavoro:"Quella sentenza sul vaccino è un attentato alla salute pubblica
17feb2015	Trustful Source	sentenza 1767/14 della corte d'Appello di Bologna nella causa d'appello alla sentenza 15.03.2012 Rimini
10jan2016	Trustful Source	La battaglia dei vaccini - presadiretta
27mar2016	Trustful Source	Robert De Niro ritira il film sul legame tra vaccini e autismo dal Tribeca Film Festival di New York
01jun2016	Trustful Source	Procura di Trani ha riconosciuto l'inconsistenza del presunto legame tra la vaccinazione trivalente MPR (contro morbillo, parotite e rosolia) e autismo
21apr2017	Trustful Source	L'Ordine dei medici di Treviso ha radiato Roberto Gava, considerato uno dei paladini dei no-vax in Italia
02may2017	Trustful Source	new york times pubblica "Populism, Politics and Measeles"
07sep2017	Trustful Source	TAR Lazio – decreto 7 settembre 2017: respinto il ricorso del Codacons riguardante le misure adottate per ottemperare agli obblighi di documentazione vaccinale
22nov2017	Trustful Source	la sentenza della Corte costituzionale considera legittimo l'obbligo dei vaccini nel contesto attuale definito dal Decreto 73/2017 e respinge i ricorsi presentati dalla Regione Veneto
05jul2018	Trustful Source	Giulia Grillo: «Vaccini, a scuola con autocertificazione. L'obbligo cambierà. Io incinta, vaccinerò mio figlio»

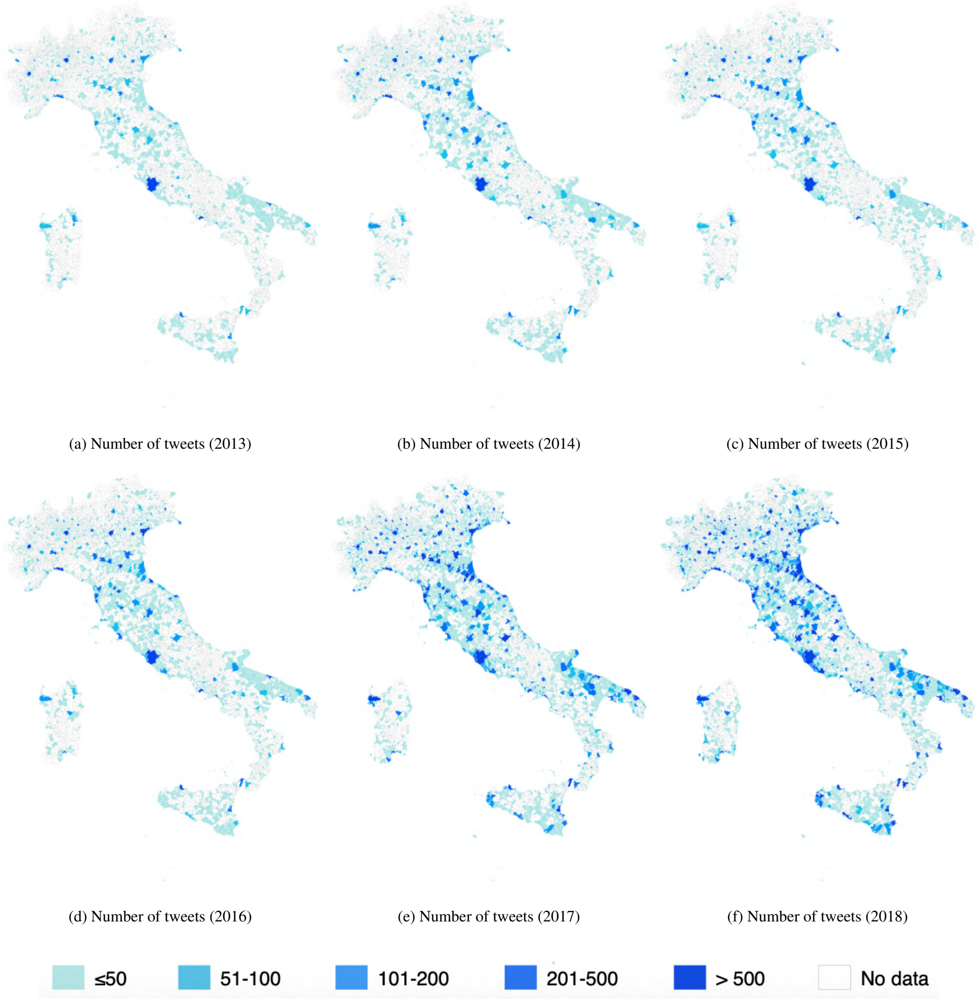
List of events - *continued from previous page*

date	Classification	Description
01jul2013	Vaccine Unsafe	Tribunale di Pesaro, 1 luglio 2013
07jul2013	Vaccine Unsafe	“Vaccine adverse events reporting system" pubblica uno studio dove Heidi Stevenson parla di migliaia di morti per colpa di vaccini, probabilità di morte aumenta del 50 per cento – e con ogni dose di vaccino supplementare"
10oct2013	Vaccine Unsafe	Vaccini:1 italiano su 2 è contrario, "inutili e poco sicuri"
11nov2013	Vaccine Unsafe	Tribunale di Pesaro, 11 novembre 2013
09jan2014	Vaccine Unsafe	Venti nuovi casi di danno da vaccini alla settimana per l’avvocato di Rimini
17mar2014	Vaccine Unsafe	Si vaccina poco? Big Pharma fa pressing sulle Asl e telefona alle famiglie
02jul2014	Vaccine Unsafe	Tribunale di Rimini, 2 luglio 2014, n. 217
23sep2014	Vaccine Unsafe	Tribunale del lavoro di milano: vaccino esavalente Infanrix Hexa Sk causa l’autismo
20oct2014	Vaccine Unsafe	La presenta di DNA fetale umano nei vaccini é una possibile causa di autismo
28nov2014	Vaccine Unsafe	Aifa: “Tredici casi di morte sospetta” vacccino antiinfluenzale
11mar2015	Vaccine Unsafe	ll ministero riconosce l’indennizzo per un bimba Catanzaro
01jul2015	Vaccine Unsafe	Jim Carrey tweet sul mercurio nei vaccini
03jul2015	Vaccine Unsafe	Jim Carrey causa l’autismo
25may2016	Vaccine Unsafe	Bimba muore a 2 mesi in culla dopo il vaccino: l’Asl sostituisce tutti i lotti
10nov2016	Vaccine Unsafe	Corte d’Appello di Milano, 10 novembre 2016, n.1255
01feb2017	Vaccine Unsafe	Corte di Cassazione, Sezione 6 civile, 1 febbraio 2017, n. 2684
17apr2017	Vaccine Unsafe	report vaccino HPV

Additional Figures

Figure A1 shows tweets’ distribution across municipalities over time.

Figure A1: Tweets over time (2013-2018)



(a) Number of tweets (2013)

(b) Number of tweets (2014)

(c) Number of tweets (2015)

(d) Number of tweets (2016)

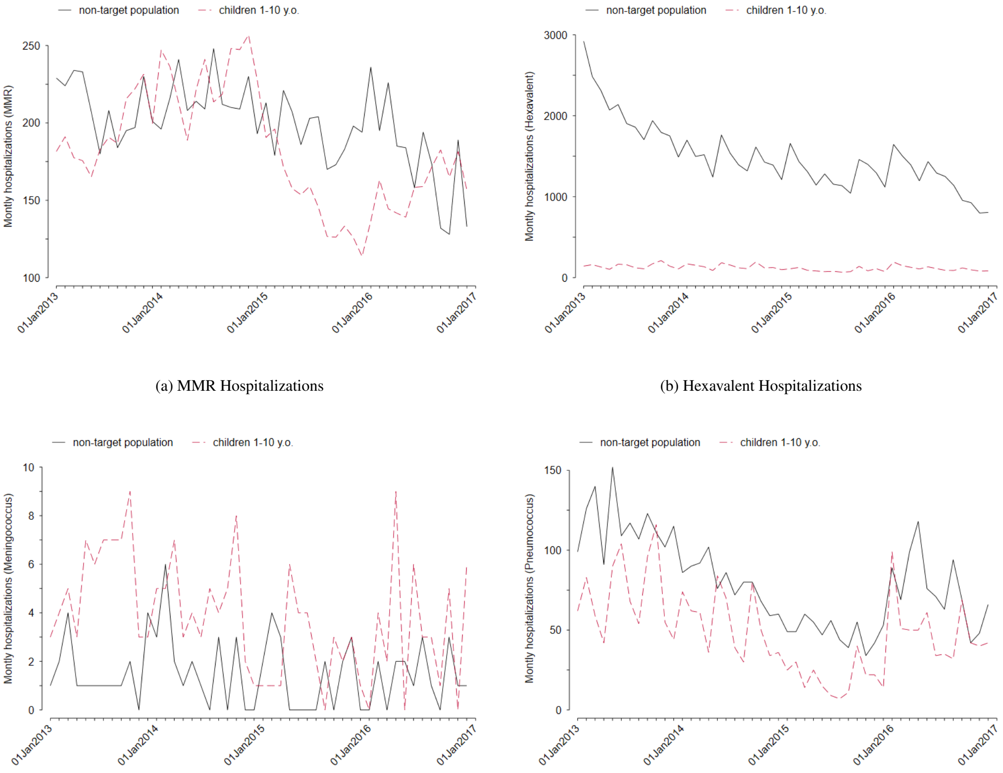
(e) Number of tweets (2017)

(f) Number of tweets (2018)

Notes: The sample consists of 830,253 tweets relative to a population of 80,471 unique users across 4,220 municipalities.

Figure A2 plots the monthly trends in hospitalizations among the vaccine-target population and in vulnerable populations that are not targeted by vaccines. The trends for hexavalent, pneumococcus, and meningococcus are generally comparable for the two groups. For diseases covered by the hexavalent vaccine, hospitalisation rates for the vaccine targeted population was stable for the entire study period, while hospitalization rates for the non-targeted population decreased from 3000 to 1000 during the study period. For Meningococcus, we see a higher overall number of hospitalizations for target population than for target population. Instead, the trend for pneumococcus is similar across groups. However, for the MMR vaccine, the hospitalization rate trends were opposite between January 2015 and January 2017, which was a period marked by several measles epidemic outbreakshe hospitalization rate for the group that received the vaccine decreased from 250 in 2015 to 170 in 2016, with the lowest point being in 2016. During the study period, the hospitalization rate for the vaccine non-targeted population remained steady at an average of 200 hospitalizations until july 2016, but then started to decrease after August 2016.

Figure A2: Hospitalization trends (2013-2016)



(a) MMR Hospitalizations

(b) Hexavalent Hospitalizations

(c) Meningococcus Hospitalizations

(d) Pneumococcus Hospitalizations

Notes: The hospitalization rate for the vaccine-targeted population is represented by the red dashed line, while the vaccine non-targeted vulnerable groups are represented by the black solid line.

Appendix B

The Model of Opinion Dynamics and Network Formation.

The model builds on Baumann et al. (2020)'s work on endogenous polarization dynamics in social networks. In the model we consider a continuum of individuals in a discrete, infinite time setting $[t = 0, 1, .., \infty]$. Each individual i has a stance on vaccinations $s_i^t = [\underline{s}, \bar{s}]$ which spans from unconditional support to hesitancy. We assume that the stance reflects individuals' opinions on the overall utility of vaccinations a one-to-one mapping between parents' and children's (perceived) utility.

Individual stances evolve over time from initial positions s_i^0, drawn from a distribution $S^0 \sim F_s(0)$, with finite first and second moments; in particular, $\mu^0 = \mathbb{E}(s_i^0)$, stands for the average initial stance in the society. To reflect the observed distribution of initial stances - on average pro-vaccines - in the baseline simulations $\mu^0 \leq 0$ and initial stances are drawn from a Gaussian distribution. We obtain qualitatively equivalent results when we move to a case where the initial distribution of opinions is centered around zero (i.e., $\mu^0 = 0$).

The opinion dynamics within the social network are entirely driven by the interactions among agents and are described by a system of N coupled differential equations:

$$\dot{s}_i = -s_i + \mathbb{I}\sum_{j=1}^{N} W_{ij}(t)tanh(\alpha_t s_j) \tag{B.1}$$

In Equation (B.1) $\mathbb{I}$ measures the strength of the interaction among users of the platform, $W(t)$ is a time-varying spatial contiguity matrix, whose i^{th}, j^{th} elements represent every link between individuals in the network - i.e., $w_{ij}(t) = 1$ if i interacts with j, $w_{ij}(t) = 0$ otherwise. The function $tanh(\cdot)$ is the hyperbolic tangent function, which provides a sigmoidal influence function of peers on individuals' stances, ensuring that i) an agent's i stance influences others monotonically and that ii) such influence "flattens" in the extremes. Finally, α_t is the degree of controversialness of the topic.

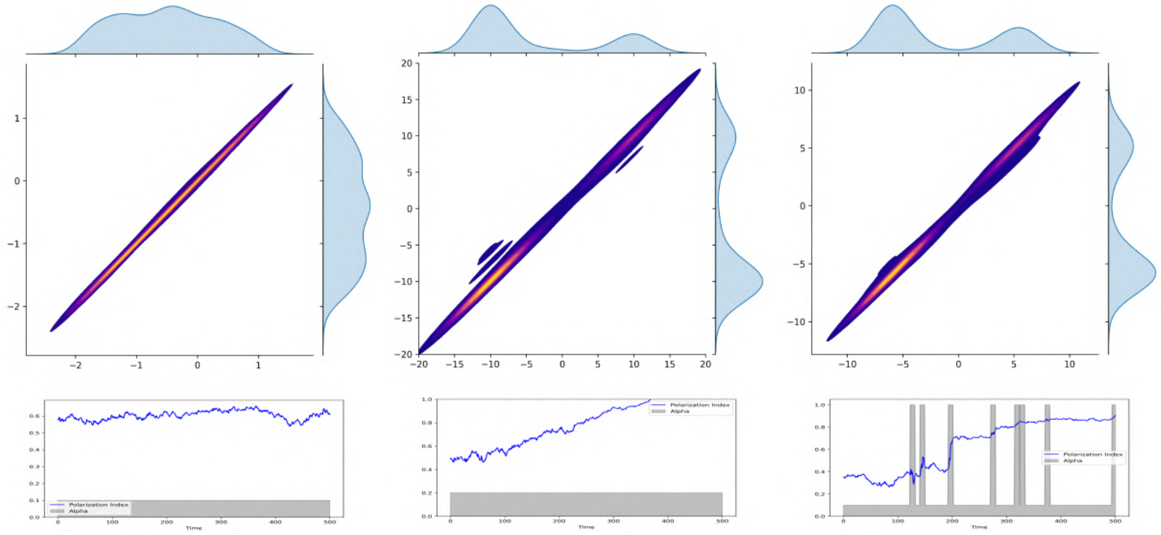
The contiguity matrix $W(t)$ evolves according to an activity-driven (AD) temporal network (Perra et al., 2012), where each agent is characterized by the propensity to interact with a share $\omega_i \in [\epsilon, 1]$ of other agents, and the probability of an interaction is driven by homophily (Bessi et al., 2016) - that is to say, individuals are more likely to interact with like-minded peers, and we model it as a decreasing function of the (absolute) distance between i and j's opinions, $p_{ij}(t) = \frac{|s_i(t)-s_j|^{-\beta}}{\sum_j |x_i - x_j|^{-\beta}}$. Note that the β parameter that informs the power law decay of interaction probability includes effects as diverse as the endogenous preferences for homophily (i.e., to what extent individuals dislike the interaction with people of different stances) or the exogenous settings embedded in the social networks' algorithms - e.g., how likely one's content is to appear in a like-minded peer's home newsfeed.

Figure 3 shows the predictions of the simulated models. The heatmaps show the distribution of stances for the users and their followings in a simulation for low controversialness ($\alpha = 0.1$ in panel a), relatively higher controversialness ($\alpha = 0.2$ in panel b), and time-varying controversialness (long periods of $\alpha = 0.1$ with short-lived outbursts of $\alpha = 1$ in panel c). The colors in the heatmaps represent the density of users, with lighter colors indicating a higher number of users. The marginal distribution of users' opinions and their followings' opinions are plotted on the x- and y-axis, respectively. The simulation shows that users are more likely to connect with peers who share similar opinions due to homophily.

In addition to homophily, higher controversialness strengthens the influence of peers' opinions on users who tend to form homogeneous groups. At the network level, this results in a correlation between users' and their followings' average opinions. When controversialness is low (panel a), the model converges to a bivariate Gaussian distribution centered at approximately (-.5,-.5); on the other hand, when the model is characterized by higher controversialness (panel b), it converges to a bivariate bimodal distribution with a high density of users with like-minded followings, resulting in two echo chambers corresponding to opposite stances on vaccinations. In a more realistic simulation where long periods of low controversialness are interrupted by short-lived, high-controversialness outbursts (panel c), the model also generates echo chambers.

The figures below the heatmaps show the degree of polarization during the simulations. When controversialness is low, there is no trend in polarization within the population, but polarization increases with relatively high controversialness. Interestingly, with time-varying controversialness, polarization increases during the outbursts and remains stable at the new, higher level until the next shift.

Figure 3: Simulated distribution of stances



(a) Low controversialness (b) High controversialness (c) Time-varying controversialness

Notes: user (x-axis) and average followings' (y-axis) distribution of stances in a simulated model when controversialness is low ($\alpha = .1$ in panel a), high ($\alpha = .2$ in panel b), and low with short-lived outbursts ($\alpha = 0.1$ and $\alpha = 1$ in panel c). In all models, the number of individuals is $N = 500$ and the periods are $T = 5$ - divided in 100 subperiods. Initial values (s_0) are randomly drawn from a gaussian distribution with $\mu = -0.2$ and $\sigma = 0.5$ to match the asymmetry of the initial opinions in the data.

Appendix C - not for publication

Data Colleciton and classification procedure

In this section we describe the process by which we obtain the Twitter data used in this paper as well as the classification procedure.

Data collection. We access these data through the Twitter Academic API, which provides the full archive public Twitter posts to users upon requests.[38]

First, we registered for an API key through Twitter's Developer program to obtain a set of keys (API key, API secret key, Access token, Access token secret) that allow us to access the data.[39] Users are required to agree to Twitter's conditions in order to use these data. Second, we wrote a Python script using the Full Archive endpoint of the Twitter developer site for academic researchers to gather publicly available Twitter posts in Italian that contained the words "vaccine(s)", "vaccination", "vaccinating", "novax", and "vax" (and their feminine and plural forms).[40] We excluded tweets that were advertisements for *"mozzarella"* or *"cow milk"* ("*latte vaccino*" in Italian). The most recent version of the dataset was downloaded on April 23, 2021.

We collected 2,031,448 tweets. We omit *ex-post* all tweets containing only links or mentions[41] and those produced by accounts that are temporarily unavailable due to violation of the Twitter media policy.[42] We also disregard all tweets referring to pets' vaccinations, those where the string "vax" is only retrieved in a URL contained in the tweet, and those written in other languages. After this initial screening process we leaved a sample of 2,017,539 posts.

```
query = "(vaccino OR vaccini OR vaccinazione OR vaccinarsi
          OR vaccinato OR vaccinata OR vaccinati OR novax OR vax
          OR vaccinare -latte vaccino) lang:it"
start_date = "01-01-2013T00:00"
end_date = "01-01-2019T00:00"
```

[38] Twitter's license agreement for this API forbids sharing the raw Twitter data publicly.

[39] Accessible here: https://developer.twitter.com/en/products/twitter-api/academic-research

[40] We used the Full Archive endpoint search/all at the following URL: https://api.twitter.com/2/tweets/search/all. This endpoint allows us to access up to 10 million tweets per month. Please note that Twitter's license agreement for this API prohibits the public sharing of raw Twitter data.

[41] A tweet containing another user's username, preceded by "@".

[42] Since 2021, Twitter has applied labels to tweets that may contain misleading information about COVID-19 vaccines and removed the most harmful misleading information from the service.

VaxBERT classifier. To determine the stance of a tweet on vaccines, we created an *anti-vax* tweet classifier called `vaxBERTo` using `BERT` (*Bidirectional Encoder Representations from Transformers*), a state-of-the-art Natural Language Processing (NLP) algorithm.[43] BERT is a pre-trained model that can be used to analyze text for a variety of tasks, as demonstrated in research such as See et al. (2017) and Johnson and Goldwasser (2016). NLP allows us to analyze text using various methods, from human analysis to automatic processes using built-in libraries, and to categorize tweets as positive or negative based on their main points. The polarity of the tweets is then determined based on these categories.

The following subsections provide a full description of each stage in the suggested model.

Data cleaning. The tweets are pre-processed and normalized, we remove all URLs and hashtag symbols (#) and retain only the text that could provide useful information. All tags (e.g. @*username*) were also removed so that the labeling process could focus solely on the content and not on the people mentioned.[44] An example of original tweet and the clean version is reported below (Table C.1).

Table C.1: Example of tweet

Original Tweet	Clean Text
rt @nextquotidiano: la storia di nadja, la bambina serba morta di #morbillo (per colpa dei #novax) https://t.co/6lfsl2mxcs	la storia di nadja, la bambina serba morta di morbillo (per colpa dei novax)

$Notes$: Cleaning of tweet ID n. 1007*************203. English translation: *"Nadja's story, the Serbian girl who died from measles (for the fault of the anti-vax)"*

Training phase. We divided our database into two categories: original tweets (which includes quotes and replies) and plain retweets. Since retweets' text is identical to the original tweet, which was still preserved, we didn't include them in the sample. Indeed, using the same lines repeatedly during the training phase would merely increase noise and not offer any new information. Additionally, as there would be no justification for the algorithm to treat the labels of retweets in the prediction phase as if they were original tweets, we remove the retweets' subsample also in that phase. Instead, we label retweets of their corresponding original tweets with the same label. In order to reduce bias in the training sample, we followed Pierri et al. (2020) and categorize tweets from well-known Italian fake news outlets, pro-vax activists, and mainstream media outlets as either "anti-vax" ($label_{\tau} = 1$), "pro- or neutral" ($label_{\tau} = 0$). A list of mainstream media outlets and fake news sources can be found in Table 1.

[43] `BERT` was developed for Google by Devlin et al. (2018b).

[44] Notice that according to Polignano et al. (2019), `BERT` is able to efficiently parse the social media language, and that mentions, links and hashtags do not affect model training. We decided in favor of the cleaning process in order to avoid any interference between mentions, unhelpful information like urls, some symbols, and the outcome, which we wanted to be focused on context only. However, results hold without the additional cleaning.

Table C.2: List of renown fake news outlets as in Pierri et al. (2020)

mag24.es	saper-link-news.com
notiziarioromacapitale.blogspot.com	tg24-ore.com
ilmessangero.it	tuttiicriminidegliimmigrati.com
mondodiannunci.com	voxnews.info
skynew.it	webtg24.com
zapping2017.myblog.it	accademiadellaliberta.blogspot.com
daily-screen.com	altrarealta.blogspot.com
il-quotidiano.info	aurorasito.wordpress.com
adessobasta.org	compressamente.blogspot.com
catenaumana.it	freeondarevolution.wordpress.com
ilvostropensiero.it	ilsapereepotere2.blogspot.com
lettoquotidiano.it	laveritadininconaco.altervista.org
interagisco.net	madreterra.myblog.it
5stellenews.com	neovitruvian.wordpress.com
breaknotizie.com	olivieromannucci.blogspot.com
byoblu.com	pianetax.wordpress.com
comedonchisciotte.org	terrarealtime.blogspot.com
direttanews24.com	corrieredelcorsaro.it
essere-informati.it	siamonapoletani.org
il-giornale.info	ilfattoquotidaino.it
il-quotidiano.info	conoscenzealconfine.it
informarexresistere.fr	disinformazione.it
informazionelibera.eu	ecplanet.org
jedanews.it	effervescienza.org
italianosveglia.com	filosofiaelogos.it
lonesto.it	hackthematrix.it
silenziefalsita.it	ilpuntosulmistero.it
sostenitori.info	libreidee.org
tankerenemy.com	liberamenteservo.com
ununiverso.it	nibiru2012.it
skytg24news.it	pandoratv.it
ilprimatonazionale.it	tmcrew.org
List of the mainstream media outlets	
ANSA	L'Avanti
Adnkronos	Liberazione
AGI	L'Osservatore Romano
ASCA	Il Sole 24 Ore
Reuters	ItaliaOggi
Press Association	Milano Finanza
Bloomberg	La Gazzetta dello Sport
News.cn	open
La Repubblica	Panorama
La Stampa	L'Espresso
Il Tempo	Micromega
La Nazione	Le Scienze
Il Messaggero	Focus
Il Giornale	Galileo
Il Fatto Quotidiano	Universinet
Il Foglio	Famiglia Cristiana
Il Giorno	Le scienze
Il Manifesto	

We manually labeled 43,472 tweets from 108 unique users for the training set, with 23,909 tweets (55% of the total) classified as 1 and 19,563 as 0. During the training process, the machine further divided the sample into a training set of 39,124 tweets (around 90% of the total) and a validation set of 4,348 tweets to fine-tune the training based on the performance of the trained neurons. Finally, we created a labeled test set of 4,830 tweets to evaluate the model's performance on a set of users different from those in the training set.[45]

First, we made the model run for 4 epochs and 32 batches, in accordance with Devlin et al. (2018a)[46]. In Table C.3 we report the training and the validation losses alongside the accuracy reached.[47] While the latter increases with the epochs, the validation loss increases at the end of the routine - indicating a possible over-fitting that we exclude by using the model run for two epochs only.

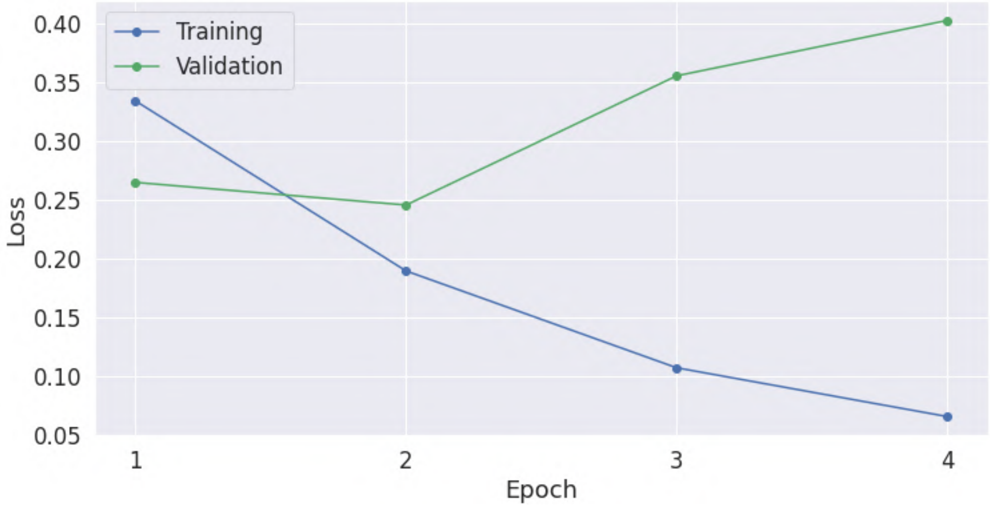
Table C.3: vaxBERTo last layer training

epoch	Training Loss	Valid. Loss	Valid. Accur.	Training Time	Validation Time
1	0.3342	0.2650	0.8885	0:05:50	0:00:13
2	0.1897	0.2456	0.9072	0:05:47	0:00:13
3	0.1074	0.3554	0.9023	0:05:47	0:00:13
4	0.0660	0.4025	0.9055	0:05:46	0:00:13

Notes: training and validation losses (columns 2 and 3), accuracy (4) and computing time (5 and 6) for each vaxBERTo training epoch.

After every batch an error is calculated comparing the predictions with the expected output. The network updates the weights before moving to the next batch. We plot the performance in terms of training loss in Figure C.1.

Figure C.1: Training and validation loss



Prediction phase. We used the Matthews correlation coefficient (MCC) to evaluate the performance of our model. MCC is a useful metric because it ranges from -1 to 1, with -1 indicating that the model made no correct predictions, 0 indicating random guessing, and 1 indicating perfect accuracy. When we ran our model,

[45] Training was carried out with platform Google Colab using the GPU *Tesla P100-PCIE-16GB*

[46] Devlin et al. (2018a) found the range of possible values to work well across all tasks: Batch size: 16, 32; Learning rate (Adam): 5e-5, 3e-5, 2e-5; Number of epochs: 2, 3, 4.

[47] Training time and Validation time may vary widely depending on the hardware used to perform the training.

the MCC was 0.749, which suggests that our model is reliable. We then applied the model to label 781,337 original tweets in our dataset (2,017,539 including the retweets, that we label back using the original labels).