Thèse n° 7482

Contact networks across ages: computational analysis of epidemiological dynamics

Présentée le 24 avril 2020

à la Faculté des sciences de la vie
Unité du Prof. Salathé
Programme doctoral en biologie computationnelle et quantitative

pour l'obtention du grade de Docteur ès Sciences

par

Gianrocco LAZZARI

Acceptée sur proposition du jury

Prof. P. De Los Rios, président du jury
Prof. M. Salathé, directeur de thèse
Dr C. Cattuto, rapporteur
Prof. M. Bochud, rapporteuse
Prof. J. Fellay, rapporteur

École
polytechnique
fédérale
de Lausanne

2020

"Whether it occurs by a quirk of nature or at the hand of a terrorist, epidemiologists say a fast-moving airborne pathogen could kill more than 30 million people in less than a year. And they say there is a reasonable probability the world will experience such an outbreak in the next 10-15 years."

Bill Gates, Business Insider, 2017

EPFL

Abstract

School of Life Science
Global Health Institute

Doctor of Philosophy

Contact networks across ages: computational analysis of epidemiological dynamics

by Gianrocco LAZZARI

For decades mathematical modeling in epidemiology has helped understanding the dynamics of infectious diseases, as well as describe possible intervention scenarios to prevent and control them. However, such models were relying on several assumptions, such as the ones on the structure of the underlying contact networks. The robustness of their predictions was therefore limited by this lack of knowledge. About a decade ago, with the advent of digital epidemiology, scientist have finally started to try to corroborate those assumptions, for instance with the use of wearable sensors to measure indeed contact networks.

In this thesis, together with collaborators, I try to combine the digital collection of public health data with computational tools, in order to have a more realistic understanding of the phenomena under consideration. In two projects it was possible to finalize such marriage, thanks also to fruitful collaborations with other researchers who provided the data. This is for instance the case for the two chapters respectively on the modeling of influenza and plague outbreaks. Although they involve different technologies for the data collection, historical epochs and data types, the traditional epidemiological modeling allowed us to derive interpretable conclusions, capable for instance to inform public health interventions. In other projects, either the relevant data collection is still ongoing in the lab (like for the FoodRepo project), or the data collection has not started yet (like for the project on measles). Nevertheless, our work provides insights on the importance of such data collection for future studies.

In the first chapter, we explore different mechanistic interpretations compatible with our data on the 1630 plague outbreak in Venice, collected through the digitization of parish books from the historical Patriarchal Archives of Venice. The data shows a non trivial temporal structure, which led us to propose few different epidemiological explanations. Further data collection will be needed to better constrain such interpretations.

In the second chapter, we use contact data previously recorded in a high-school to assess the relative effect of ventilation on influenza spread, with respect to vaccination. Our result suggest the usefulness of non pharmacological interventions such as indeed improved ventilation, which become even more meaningful in the context of vaccination hesitancy and low vaccine efficacy, due for instance to the high mutation rate of viruses like influenza virus.

In the third chapter, we propose a simple network generation model to try to explain differences in the incidence of highly infectious diseases such as measles, across countries with similar vaccination coverages. Such differences are indeed one of the main open questions in public health, which are not yet fully understood even considering social phenomena such as recent anti-vax movements.

In the last chapter, we present our open database of barcoded food products, FoodRepo. This database represents on the one hand, the first piece of a large study ongoing in our lab, in the field of nutritional epidemiology, that aims to assess the variability of glycemic response in a healthy cohort. On the other hand, important features such as its openness and programmatic accessibility make it an important digital tool at the service of any private or public actors in the field of nutrition.

The projects presented in this thesis clearly stand on different progress stages, within their own broad research picture. Nevertheless, they represent different examples of how combining 'traditional' methods (such as stochastic modeling on networks) with more recent data collection techniques, one can tackle still largely open question in public health. In each work, we have tried to provide results which are easily interpretable and translatable by domain experts, while at the same time, trough the openness of our analysis and datasets, provide as well tools on top of which future computational research can be built.

Abstract (IT)

Per decenni, i modelli matematici in epidemiologia hanno aiutato a comprendere le dinamiche delle malattie infettive e a descrivere possibili interventi per prevenirle e controllarle. Tuttavia, questi modelli si basavano su varie assunzioni, come quella sulla struttura della rete dei contatti sottostante. La robustezza di queste predizioni era pertanto limitata dalla mancanza di questo dato. Circa un decennio fa, con l'avvento dell'epidemiologia digitale, gli scienziati hanno finalmente iniziato a cercare di corroborare queste assunzioni, per esempio con l'uso di sensori indossabili, per misurare appunto la rete di contatti.

In quest tesi, insieme ai nostri collaboratori, ho provato a combinare la raccolta digitale di dati relativi ai problemi di sanità pubblica, con strumenti computazionali, per cercare di avere una comprensione più realistica dei fenomeni in considerazione. In due progetti, é stato possibile realizzare questo connubio, grazie alla fruttuosa collaborazione con altri ricercatori che hanno fornito i dati. Questo é per esempio il caso dei due capitoli sulla modellizzazione delle epidemie rispettivamente di influenza e peste. Sebbene abbiano richiesto differenti tecnologie per la raccolta dei dati, epoche storiche e tipo di dati, i modelli epidemiologici tradizionali ci hanno permesso di

ottenere delle conclusioni interpretabili, e capaci di informare misure di sanità pubblica. In altri progetti l'attinente raccolta dati o non é ancora terminata (come per il progetto FoodRepo), o non é ancora iniziata (come per il progetto sul morbillo). Tuttavia, il nostro lavoro fornisce introspezioni sull'importanza di tale raccolta dati per studi futuri.

Nel primo capitolo, esploriamo diverse interpretazioni compatibili con i nostri dati, sull' epidemia di peste a Venezia del 1630, raccolti grazie alla digitalizzazione dei libri parrocchiali dell'archivio patriarcale di Venezia. I dati mostrano una struttura temporale non banale, che ci ha portato a proporre alcune diverse interpretazioni epidemiologiche. Un'ulteriore raccolta dati sarà necessaria per costringere meglio tali interpretazioni.

Nel secondo capitolo, usiamo i dati sui contatti raccolti precedentemente in un liceo, per valutare l'effetto della ventilazione sulla diffusione dell'influenza, relativamente a quello della vaccinazione. I nostri risultati suggeriscono l'utilità di interventi non farmacologici come appunto una migliorata ventilazione, che diventano ancora più significativi nel contesto della reticenza a vaccinarsi o della bassa efficacia del vaccino, causata per esempio dall'alto tasso di mutazione di virus come quello dell'influenza.

Nel terzo capitolo, proponiamo un semplice modello generativo per reti, per cercare di spiegare le differenze nell'incidenza di malattie altamente infettive come il morbillo, tra Paesi con tassi di vaccinazione simili. Tali differenze sono infatti una delle principali domande ancora aperte in sanità pubblica, non ancora interamente spiegata anche considerando fenomeni come i recenti movimenti no-vax.

Nell' ultimo capitolo, presentiamo il nostro database di prodotti alimentari con codice a barre, FoodRepo. Questo database rappresenta da un lato, il primo pezzo di un largo studio in corso nel nostro laboratorio, nel campo dell' epidemiologia nutrizionale, che mira a valutare la variabilità della risposta glicemica in una coorte sana. Dall'altro lato, caratteristiche importanti come la sua gratuità e accessibilità programmatica, ne fanno un importante strumento digitale al servizio di ogni attore pubblico o privato, nel campo della nutrizione.

I progetti presentati in questa tesi, hanno raggiunto chiaramente diversi gradi di avanzamento, ognuno all'interno del proprio dominio di ricerca. Tuttavia, essi rappresentano diversi esempi di come, combinando metodi tradizionali (come i modelli stocastici su rete) con tecniche più recenti di raccolta dati, si possono affrontare questioni ancora aperte in sanità pubblica. In ogni lavoro, abbiamo cercato di fornire risultati facilmente interpretabili e traducibili da esperti del settore, e allo stesso tempo, grazie alla trasparenza delle nostre analisi e datasets, fornire anche strumenti sui quali ulteriore ricerca futura può essere costruita.

Acknowledgements

The main acknowledgment goes to my PhD supervisor, prof. M. Salathé. In these four years and a half, we have been sharing the joy of successes and engaging in demanding discussions that helped crossing narrow bridges, without losing the broad picture. His open-mindedness has allowed me to explore different projects that would match our common research interests and led us to the results presented in the different chapters of this thesis.

Another important thanks goes to my PhD mentor, prof. P. de Los Rios. His advice has been very helpful several times for both technical and non-technical aspects of the research life.

Strictly after follow the acknowledgments to all the people with witch I co-authored the works presented in this dissertation: Giovanni Colavizza, prof. Frédéric Kaplan, Timo Smieszek, Yannis Jaquet, Djilani Kebaili, Laura Symul. I am grateful to them, as they provided the data and the technical support on which the analysis and results were built and presented. Furthermore, as they mostly come from different backgrounds than mine, I can definitely say I have learnt how to work in a interdisciplinary environment, also thanks to them. This I believe will be very helpful for my future career, as today's world of research and innovation is indeed increasingly more interdisciplinary.

A great thank goes also the other members of our lab, with which I spent plenty of lunch discussions and few Friday happy hours. Particular acknowledgements goes to Marina, the admin of our lab (without her, clearly much of the lab's everyday life would not be possible) and to my colleague Martin, with which we shared hundreds of laughs and complaints about everything on and off our computer monitors.

As life goes on also outside working hours, important thanks are due as well to all those people with which I have been sharing fears and joys in my free time. Naming them singularly would probably take another page, so I apologize and I thanks them all in this sentence, in particular the ones I met at Cubaliente (EPFL latin dance association), Club Montagne (EPFL mountain sports association), EPFL School of Life Science (ADSV), the tennis and sailing university sport center and at the Campus Biotech, in Geneva. Another thanks goes also to all student representatives of the different EPFL PhD programs that like (or more) than me have committed part of their time to improve PhD students' and researchers' working life conditions, on the different EPFL campuses.

Another important thanks goes to my brother, with which I have been enjoying increasingly frequent discussions on clinical research, and my parents, endless source of affection, tarallini, pasticciotti and olive oil (the last one unfortunately turned out not to be endless, due to the *Xylella Fastidiosa* epidemic in the Puglia region, in south Italy). At the last position of a circular space goes a special person who helped me grow as a human being[1], in the last three years. I have definitely learnt from Joana

[1] which might be measured in number of white hair – $w(t)$, a non-decreasing function, usually, so far.

quite some geochemistry (especially stuff around serpentinization, which is nothing to do with snakes apparently), together with Portuguese culture, how to improve meowing, how to feel a hero for a second by getting rid of (not so) big spiders, and much more... simply, thanks!

Contents

Abstract iv

Acknowledgements vii

1 Introduction 1

2 Death in Venice: A Digital Reconstruction of a Large Plague Outbreak During 1630-1631 9

3 Assessing the Dynamics and Control of Droplet- and Aerosol-Transmitted Influenza Using an Indoor Positioning System 31

4 Breaking Apart Contact Networks with Vaccination 49

5 FoodRepo: An Open Food Repository of Barcoded Food Products 59

6 Conclusions and perspectives 71

7 Curriculum vitae 75

Bibliography 79

Chapter 1

Introduction

Background

Introduction to epidemiological modeling

About a century ago, Anderson McKendrick in a series of works (see for instance the famous paper (Kermack and McKendrick, 1927) started to lay down the mathematical framework which today constitutes the basis for the modeling of infectious diseases dynamics. These tools became the standard practice to inform public health policies, for instance during the HIV epidemics in the 1980s or the modern influenza pandemics, starting from 2000s. This modeling framework goes under the general name of compartmental models. Such compartments are inspired by the health condition of each individual of the population in which the infectious disease is spreading. A pathogen (often a virus or a bacterium) enters a susceptible (S-compartment) host (humans for us – animal or plants, for other public health domains) in order to proliferate. If the hosts gets exposed (E-compartment) to enough pathogenic particles, s/he will start to develop symptoms (after the so called *incubation* period) and to be infective (I-compartment) (after the so called *latency* period). Depending on the diseases, *incubation* can be longer[1] or shorter than *latency*. After being sick, the infected host usually recovers or dies, which means that s/he is not in any of the previous states (R/D-compartment). In this case, epidemiologists will use indeed the well-known SIR or $SEIR$ models. For those diseases in which the hosts can become again susceptible, a SIS or $SEIS$ model will be used instead. Here are the equations describing the typical SIR model[2]:

$$\begin{aligned} dS/dt &= -\beta S * I/N \\ dI/dt &= \beta S * I/N + \gamma I \\ dR/dt &= \gamma I \end{aligned}$$

The epidemiologist's work usually consists in estimating the transition rates between compartments (for instance β, γ in the case of SIR). In particular a key

[1] in which case one has to take into account the presence of asymptomatic carriers, like in the case of HIV or typhoid fever (remember the sad case of 'Typhoid Mary', in the late 1800s).

[2] Note that, as no internal population dynamics in included, $dS/dt + dI/dt + dR/dt = 0$

quantity to be estimated is the so called basic reproduction number R_0, defined as the expected number of secondary infection cases caused by a single typical infective case during his/her entire period of infectivity, in a wholly susceptible population (see for instance Anderson, Anderson, and May, 1992). It is important to keep in mind few caveats. First of all, a proper estimation of R_0 can be performed only with data from the very early stage of the epidemic, namely until when the assumption of "wholly susceptible population" is still valid. Furthermore, one should also keep in mind that the mathematical expression of R_0 is of course model-dependent (for instance $R_0 = \beta/\gamma$ in the case of SIR model). In addition, when the model includes also immunization, R_0 becomes $R_v = R_0(1-x)$, with x being the fraction of vaccinated susceptibles. By definition of R_0, an outbreak will occur if $R_0 > 1$, from which one can derive the minimal vaccination coverage required to prevent an outbreak, $x_v = 1 - 1/R_0$.

The estimation of the critical vaccination coverage (x_v) is a good example of the limitations of the modeling considered until here[3]. The issues emerge from the assumption of homogeneous population, with respect to different aspects, in particular the mixing, namely the fact that all hosts are connected in the same way to each other. In reality hosts with their connections represent a *network*, whose nodes' health state can be modeled with the same approach as before, namely to be in the one of above mentioned states (S, E, I, R) at a given time. The study of epidemics spread on networks has been a very active field of research in the last almost 20 years, that led to remarkable advances in our understanding of infectious disease dynamics. An important feature of networks, which has a large influence on the disease spread is the distribution $P(k)$ of a node's number of connections, also called node's degree k. The heterogeneity in the nodes' degree distribution contributes for instance to increase the critical vaccination. In particular, the above-mentioned expression of x_v becomes $x_v = 1 - 1/R_0 * (\langle k \rangle / \langle k^2 \rangle)$, for a heterogeneous network. A remarkable example are scale-free networks, characterized by a heavy-tail degree distribution $P(k)$ which takes the form of a power-law $P(k) \sim k^{-\gamma}$ (with $\gamma > 0$). Since such degree distributions tend to have very large variability ($\langle k_2 \rangle \rightarrow \infty$), critical vaccination levels approaches the whole population ($x_v \rightarrow 1$). This example shows the importance of the details of the host's network in the spread and control of an infectious disease, in addition of course to the biology of the host-pathogen interaction. For this reason, epidemiologists realized the importance of actually trying to measure real social networks, which could then better constrain network models of epidemics.

About a decade ago, few researchers started to make use of wearable sensors to track interactions in humans (Cattuto et al., 2010; Stehlé et al., 2011; Salathé et al., 2010), as well as in animals (Wilson-Aggarwal et al., 2019), in order to better quantify the amount and the temporal structure of 'contacts' networks. Such networks can then

[3] Mathematically described by a set of deterministic (or mean-field) equations (or ODEs, Ordinary Differential Equations)

be used to inform for instance models of air-born diseases (Voirin et al., 2015). This was indeed the goal in one of the works presented in this dissertation (chapter 3), where we used previously recorded data (Salathé et al., 2010) to better understand the spread of influenza in a school-setting (more details are given in the next section).

Clearly, the level of resolution needed to reconstruct contact networks is reachable only in relatively recent contexts, when records of occupancy in private/public spaces (such as houses, schools or working spaces) can be collected, either from digital devices or simple registers. Trying to study epidemics up to only few decades ago, implies that those information will likely not be available. The older one digs in the past, the less granular data one has to expect. In particular, thanks to the rich and interdisciplinary research environment of EPFL, we had the opportunity to work with 400-years-old records of a plague outbreak in Venice, in collaboration with the Digital Humanities lab, at EPFL. In the next subsection follows a brief overview on the history and etiology such notorious diseases.

History of plague

One of the most deadly diseases in the human history has been the (in)famous plague. Before digging in the past, here follows an overview of the biological details and current incidence of the disease, as reported by WHO (*WHO | News | Plague* 2020). Although in the past people used to refer to "plague" as a generic wide-spread, high-incidence and high-mortality disease, today the word is used to the specific medical conditions caused by the bacteria *Yersinia Pestis*, identified by Alexandre Yersin, in the 1894 . There are mainly two forms of plague. The bubonic one, which derives its name from the 'buboes', namely the patient's swollen lymph nodes and the pneumonic one, where the infection mainly resides in the lungs. Plague is actually not only a human disease, but occurs also within and among other mammal species, such as rats and marmots. Transmission can indeed also happen from animal to human (zoonotic disease). Therefore the whole picture of transmission routes becomes quite more complex than the one of a usual (human) infectious disease, as reported in fig. 1.1. In particular, the animal-animal and animal-human transmission is mainly mediated by fleas. Once inside the flea, the bacterium Y. Pestis settles in the digestive tract of the insect and creates a biofilm that makes it hard for the food (host's blood) to get absorbed. The blood, which now contains the bacteria gets then regurgitated in the host. Besides that, as it has troubles in food absorption, the flea tries harder to bite the host, facilitating even more the contagion. Transmission can happen as well via aerosolized particles (for instance in the case of pneumonic plague) or contact with infected blood. Concerning treatment and prevention, although no effective and safe vaccine have been produced yet (Jefferson, Demicheli, and Pratt, 1998), several antibiotics are proven to be effective, if provided within 24 hours after symptoms appearance.

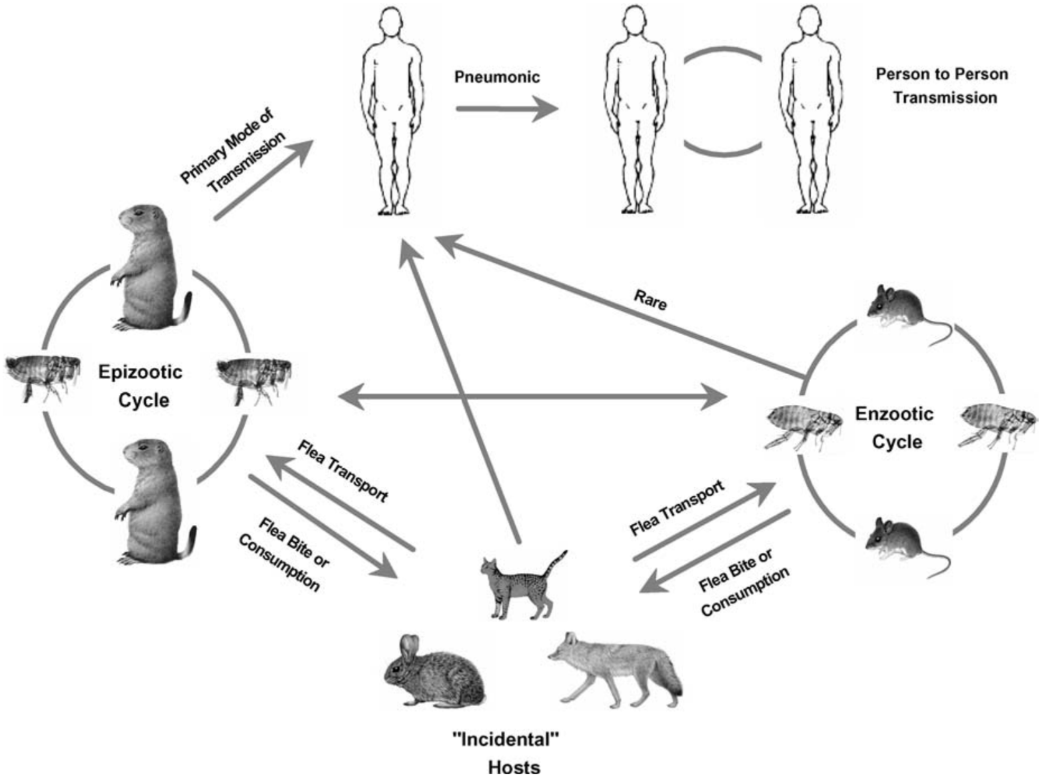


FIGURE 1.1: Transmission scheme of plague (Gage and Kosoy, 2005a)

Given its complex contagion network (fig. 1.1, endemicity of plague is explained by the presence of animal reservoirs of mammals such as marmots and rabbits, which contribute still nowadays to keep active several loci in different countries (see for instance the new cases reported last year in Mongolia and China (*Bubonic plague confirmed in China* 2019; *Mongolian couple die of bubonic plague after eating marmot* 2019), although major epidemics have been recently observed only in Madagascar, Peru and DRC (see fig. 1.2).

Plague has sadly spanned throughout the whole human history, although the ancient epidemics were likely caused by other pathogens, rather than by Yersinia

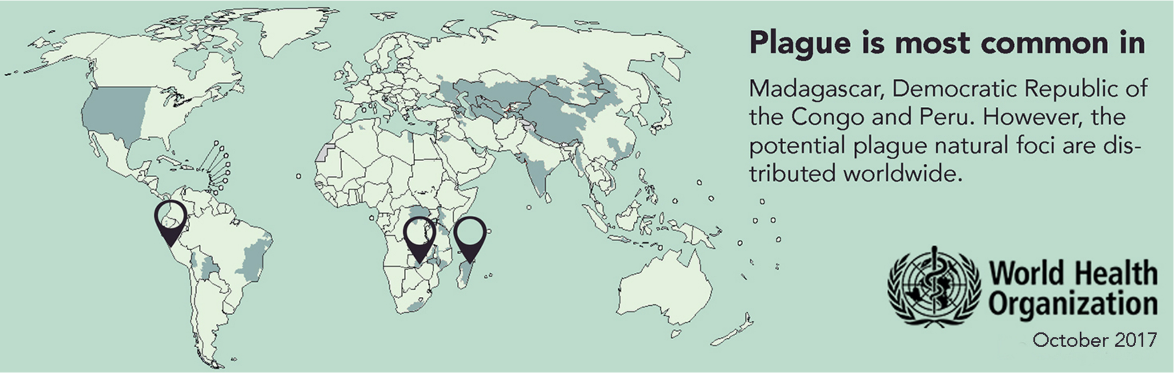
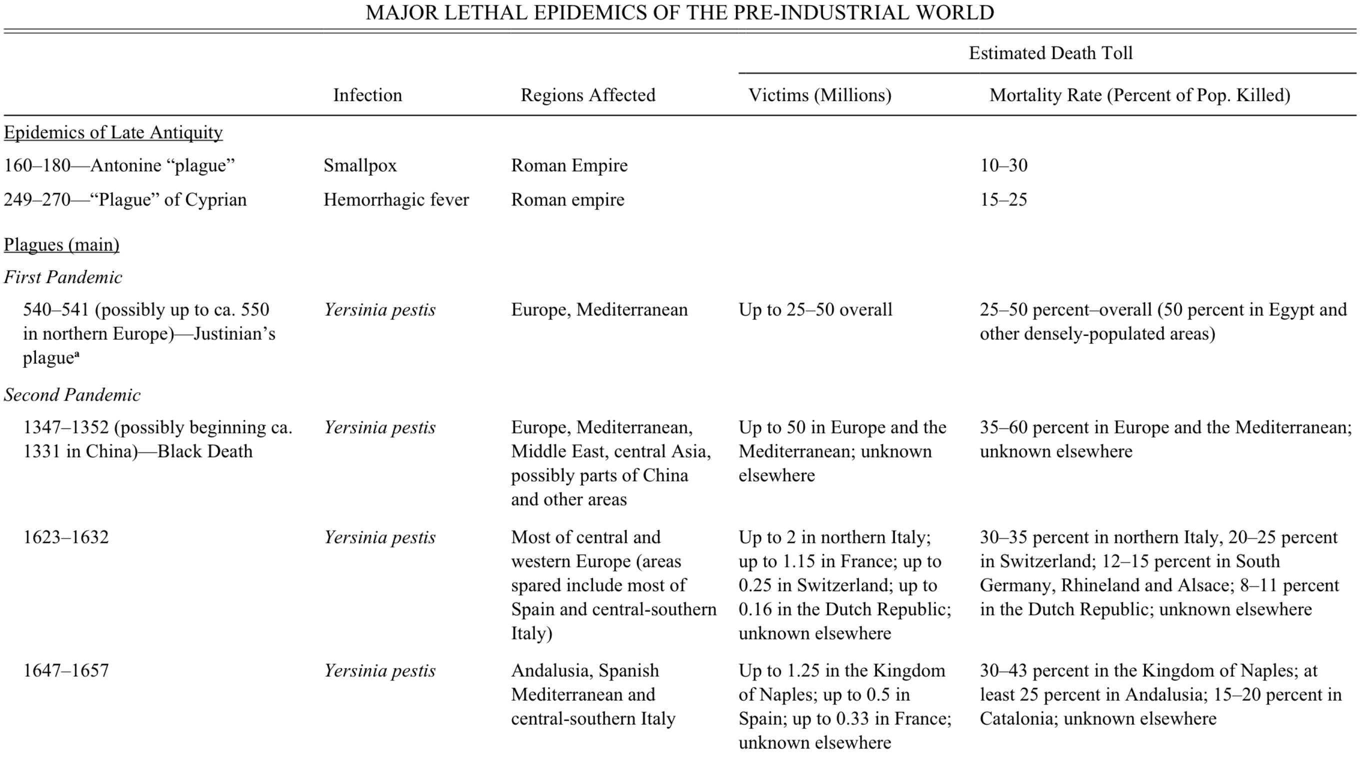

Plague is most common in
Madagascar, Democratic Republic of the Congo and Peru. However, the potential plague natural foci are distributed worldwide.
World Health Organization
October 2017

FIGURE 1.2: Space distribution of active loci and countries with main plague epidemics, as of 2017 (*WHO | News | Plague* 2020)



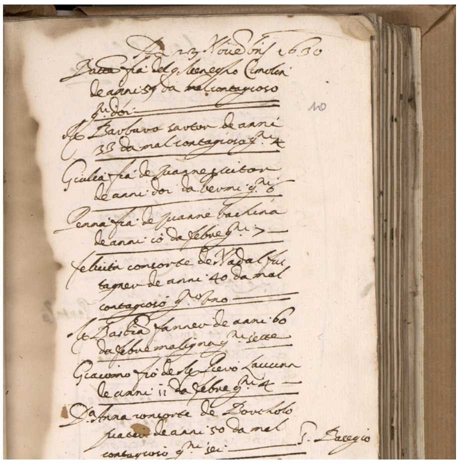
MAJOR LETHAL EPIDEMICS OF THE PRE-INDUSTRIAL WORLD

			Estimated Death Toll	
	Infection	Regions Affected	Victims (Millions)	Mortality Rate (Percent of Pop. Killed)
Epidemics of Late Antiquity				
160–180—Antonine "plague"	Smallpox	Roman Empire		10–30
249–270—"Plague" of Cyprian	Hemorrhagic fever	Roman empire		15–25
Plagues (main)				
First Pandemic				
540–541 (possibly up to ca. 550 in northern Europe)—Justinian's plague[a]	*Yersinia pestis*	Europe, Mediterranean	Up to 25–50 overall	25–50 percent–overall (50 percent in Egypt and other densely-populated areas)
Second Pandemic				
1347–1352 (possibly beginning ca. 1331 in China)—Black Death	*Yersinia pestis*	Europe, Mediterranean, Middle East, central Asia, possibly parts of China and other areas	Up to 50 in Europe and the Mediterranean; unknown elsewhere	35–60 percent in Europe and the Mediterranean; unknown elsewhere
1623–1632	*Yersinia pestis*	Most of central and western Europe (areas spared include most of Spain and central-southern Italy)	Up to 2 in northern Italy; up to 1.15 in France; up to 0.25 in Switzerland; up to 0.16 in the Dutch Republic; unknown elsewhere	30–35 percent in northern Italy, 20–25 percent in Switzerland; 12–15 percent in South Germany, Rhineland and Alsace; 8–11 percent in the Dutch Republic; unknown elsewhere
1647–1657	*Yersinia pestis*	Andalusia, Spanish Mediterranean and central-southern Italy	Up to 1.25 in the Kingdom of Naples; up to 0.5 in Spain; up to 0.33 in France; unknown elsewhere	30–43 percent in the Kingdom of Naples; at least 25 percent in Andalusia; 15–20 percent in Catalonia; unknown elsewhere

FIGURE 1.3: source: (Alfani and Murphy, 2017)

Pestis, as show in fig. 1.3. The largest plague outbreak ever was notoriously the 1347–1352 epidemic, which in Europe killed about 50 millions people (corresponding to about a third of the population) (*WHO | Plague* 2020). This pandemic seems to have started actually in north-eastern China around 1331. From there, it arrived to the Black Sea area during the following 10 years, and then it got shipped along the historical trading sea routes to Italy and south France, by 1347 (Alfani and Murphy, 2017). The year after it already spanned the whole Mediterranean area, and then moved to central/North Europe, during the following four years. The aftermath was infamously terrible, and its death toll still (fortunately) remains second only to the huge Spanish flu pandemic of 1917-1919. After two centuries of no major outbreaks reported, plague stroke again in Europe in the seventeenth century, with several smaller epidemics, which are traditionally linked by historians to the Black Death, under the umbrella of the so called 'Second Pandemic'. Unlike for the Black Death, this new deadly wave started from the North Europe in the 1620s, then moved to England and stopped in South Europe, about 40 years later. In particular, the 1630-1631 outbreak in Venice, is the one analysed in our paper (chapter 2), in collaboration with the Digital Epidemiology lab, here at EPFL.

Unfortunately, the death toll of plague continued in the beginning of the last century, starting the so-called 'Third Pandemic', running still nowadays (see again fig. 1.2). The pandemic started once again from China, in the last decade of the 19th century. From there it reached Hong-Kong and India, killing more than 12 millions people. Minor outbreaks followed in the Western countries, in particular Glasgow (1900 –



Adì 23 Novembrio 1630
[illegible] fia del q. Meneghin Comedin
de anni 57 da mal contagioso
g.ni doi
10
M. Barbaro savoner de anni
33 da mal contagioso g.ni 4
Giulia fia de Zuanne [illegible]
de anni doi da vermi g.ni 6
Anna fia de Zuanne Barbina
de anni 10 da febre g.ni 7
Felicita consorte de Nadal [illegible]
tagner de anni 40 da mal
contagioso g.ni otto
M. Bastian [illegible] de anni 60
da febre maligna g.ni sette
Giacomo fio de M. Piero [illegible]
de anni 11 da febre g.ni 4
D.a Anna consorte de Bortholo
[illegible] de anni 50 da mal
contagioso g.ni sei
S. Baseggio

FIGURE 1.4: Example of necrology book from the Venice archives.

16 deaths), San Francisco (1904 – 113 deaths) and Paris (1920 – 39 deaths). As mentioned above, plague outbreaks are still happening in the last few years. The largest one was the 2017 outbreak in Madagascar, where in the second half of 2017, 209 deaths were reported (mortality rate at 9%). Fortunately, no antibiotic *resistance* has been found[4] and no exported cases were reported (*Plague outbreak situation reports | WHO | Regional Office for Africa* 01/21/2020).

Outline

Since already about 20 years, historians have started digitizing ancient manuscripts (Abbott, 2001), therefore creating larger and larger corpora of computer-readable information (Hand, 2011), finally ready to be analyzed with computational techniques, traditionally belonging to natural scientists. Digitalization has therefore fostered new truly interdisciplinary collaborations between different scientific disciplines and humanities (Pormann, 2015). Here in EPFL, we had the great opportunity to work together with the Digital Humanities lab[5] (DHlab) that took care of the digitization of some of the 17th century parish archives of the ancient Republic of Venice (an example is provided in fig. 1.4). Such registers contain detailed records of daily deaths for each of the city's parishes. From an epidemiologist's perspective they therefore represent an important source of information on the 1630-31 plague epidemic. We performed spatio-temporal analysis on the dataset, in a close dialog with the DHlab. This informed *in silico* models of epidemic spread which let us propose few different interpretations for the patterns observed in the data. This work (Lazzari et al., 2019) is currently under peer-review and it is reported in chapter 2.

Our thirst for data, make us leap 400-years forward, back to our age. As mentioned before, measurements of physical contact networks in real-time has started

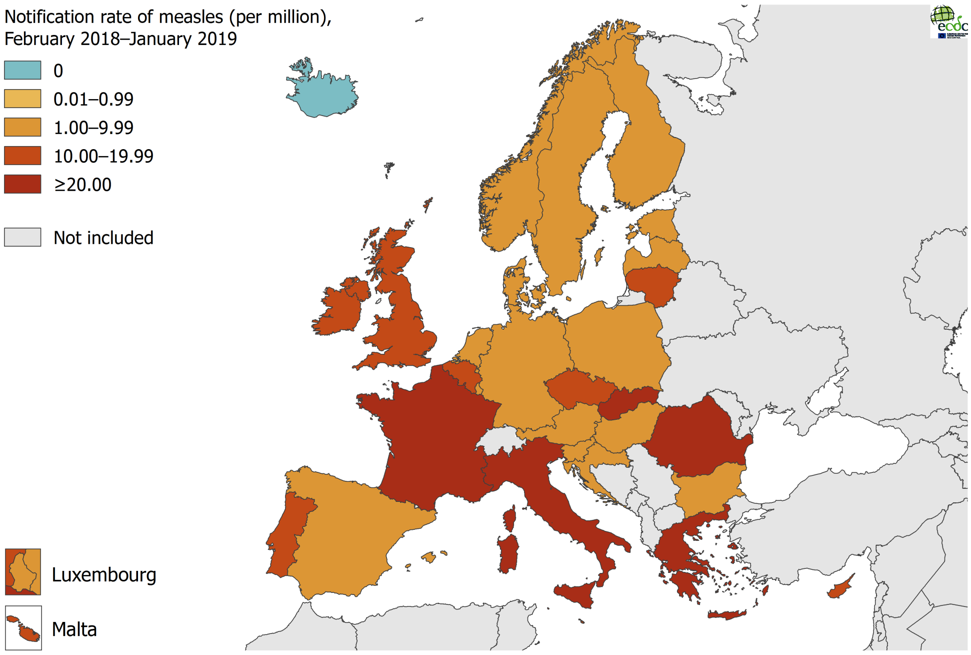
[4] In general, the appearance of antibiotic resistance poses important threats also in the case of *re*emerging infectious diseases (Cassell and Mekalanos, 2001).

[5] https://dhlab.epfl.ch

FIGURE 1.5: Example of RFID device used for tracking person-to-person interaction.

up to roughly ten years ago. One of these early successful attempts constituted indeed the starting point for our project on influenza. We built an in silico model of influenza spreading, on top of contacts data collected with RFID devices in a Californian high-school, in a previous study (Salathé et al., 2010) (see an example of such devices in fig. 1.5). Furthermore, we also considered the recent evidence that not only droplets, but also aerosol particles account for an important fraction ($\sim$ 50%) of influenza infections (Cowling et al., 2013). We therefore included both these transmission routes in our model for influenza spread, in order to assess the effectiveness of improved ventilation in public spaces, such as indeed school classrooms, as compared to pure vaccination strategies. This work has been already published (Smieszek, Lazzari, and Salathé, 2019) and it is reported in chapter 3 of this dissertation.

Although Y. Pestis has been eradicated in most of high-income countries at least, other 'plagues' still strike yearly also those countries with high-quality public health systems, and even score human, especially young lives. Indeed, despite the large efforts of governments and WHO to increase immunization coverages and fight against anti-vax movements, incidence of diseases such as measles has been on the rise in the last few years, especially in Europe, in countries like France, Bulgaria, Italy, Poland and Romania (*Monthly measles and rubella monitoring report, March 2019*) (see fig. 1.6). However, large differences in disease incidence are measured across countries with similar vaccination coverages, which still puzzles public health experts. We tried to explain such differences using a sociological argument. Namely, we started from observing that especially between countries like European and North-American, people tend to aggregate in different way, throughout their usual daily social life. We therefore built an simple algorithm that reproduces such dissimilarities in social distancing on networks, using a stochastic model of segregation inspired by the famous Schelling model (Schelling, 1971). We were indeed able to observe differences in the expected epidemic size for highly infectious diseases, as a result of changes in the networks' structural properties, induced by our segregation algorithm. More details on this work (Lazzari and Salathé, 2019, under peer-review)


Notification rate of measles (per million),
February 2018–January 2019
0
0.01–0.99
1.00–9.99
10.00–19.99
≥20.00
Not included
Luxembourg
Malta
ecdc

Produced 28 Feb 2019 using ECDC map maker: https://emma.ecdc.europa.eu

FIGURE 1.6: Measles notification rate per million population, in the EU/EEA area, from 1 February 2018 to 31 January 2019 – source: ECDC.

can be found in chapter 4.

The advantages of digital data collection clearly does not limit to the field of infectious diseases. Indeed, in the last ten years, we have assisted to a fast increase of studies proposing the use of mobile phones and wearable bio-sensors (Kim et al., 2019) to track various physiological parameters. Such digital approaches are believed to fundamentally change the way we think about health, so that the new term of Digital Medicine was coined (Elenko, Underwood, and Zohar, 2015). In particular, a field that could likely benefit from this methodological switch is nutritional epidemiology, where several phone apps have been already proposed to asses for instance dietary intake (Sharp and Allman-Farinelli, 2014). Such apps, in order to be more accurate and scalable would clearly benefit from open food databases, that would keep track of food products available on retail markets. We therefore built FoodRepo[6], a database of barcoded food products for Switzerland, currently expanding as well in Germany and Italy. The database contains digitalized information of the products package, such as nutrients, ingredients and size, and is also constantly being updated, including now more than 40000 items. More details on the rationale behind the database, as well as all steps regarding its construction, accessibility and maintenance are described in the related paper (Lazzari et al., 2018), which is here reported in chapter 5.

At last, a final chapter (chap. 6) presents a summary of the main contributions and possible future developments for each of the work presented in this thesis.

[6] https://www.foodrepo.org

Chapter 2

Death in Venice: A Digital Reconstruction of a Large Plague Outbreak During 1630-1631

Gianrocco Lazzari[1,*,+] , Giovanni Colavizza[2,*,+], Davide Drago[3], Francesca Zugno[3], Fabio Bortoluzzi[3], Andrea Erboso[3], Frédérick Kaplan[3], Marcel Salathé[1]

1. Digital Epidemiology Laboratory, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
2. University of Amsterdam, Amsterdam, Netherlands
3. Digital Humanities Laboratory, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

*gianrocco.lazzari@epfl.ch, g.colavizza@uva.nl
+ these authors contributed equally to this work

> *"Sia laudato il signor Iddio non ci sono stati morti."*
> Bless the Lord, there have been no deaths [today].
>
> December 24th 1630, in *Sant'Eufemia*, Venice.

Abstract

The plague, an infectious disease caused by the bacterium *Yersinia pestis*, is widely considered to be responsible for the most devastating and deadly pandemics in human history. Starting with the infamous *Black Death*, plague outbreaks are estimated to have killed around 100 million people over multiple centuries, with local mortality rates as high as 60%. However, detailed pictures of the disease dynamics of these outbreaks centuries ago remain scarce, mainly due to the lack of high-quality historical data in digital form. Here, we present an analysis of the 1630-31 plague outbreak in the city of Venice, using newly collected daily death records. We identify the presence of a two-peak pattern, for which we present two possible explanations based on computational models of disease dynamics. Systematically digitized historical records like the ones presented here promise to enrich our understanding of historical phenomena of enduring importance. This work contributes to the recently renewed interdisciplinary foray into the epidemiological and societal impact of pre-modern epidemics.

Introduction

Disease outbreaks of the plague in the past centuries have been so devastating throughout Eurasia that the term *plague* has become synonymous with a terrible disease. By killing a substantial proportion of the human population, which took multiple generations to recover, plague pandemics have had enormous impacts on the development of Eurasia. Correspondingly, historical questions, such as the role of institutions and the socioeconomic impact of plague outbreaks (Alfani and Murphy, 2017), as well as epidemiological questions, such as the causes, nature and interactions of vectors (Keeling and Gilligan, 2000a; Drancourt, Houhamdi, and Raoult, 2006; Hufthammer and Walløe, 2013; Dean et al., 2018), seasonality and climatic patterns (Welford and Bossak, 2009; Schmid et al., 2015) and even the distinction between plague and the Black Death (Christakos, Olea, and Yu, 2007), are still being investigated. While previous studies have highlighted some common traits to plague epidemics (Gage and Kosoy, 2005b), such as the high impact on densely-inhabited cities acting as hotspots (Gómez and Verdú, 2017; Yue, Lee, and Wu, 2017), the importance of human-to-human transmission (Whittles and Didelot, 2016) and the effect of the plague on different sexes (Curtis and Roosen, 2017), little is known about local outbreaks, due to the lack of detailed historical data.

We analyze high-quality data from death records created during the 1630-31 plague epidemic in Venice, whose initial investigation is limited and by now dated (Ulvioni, 1989). This epidemic was part of the so-called "Second Pandemic", which started with the Black Death and lasted until the early 19th century. Originated in northern Europe (modern France and the Rhineland) in 1623, this epidemic crossed the Alps approximately in 1629, in the case of the territories of the Republic of Venice likely carried by imperial armies on their way to Mantua. The cause of this specific outbreak in Venice has been linked to the bacterial species *Yersinia pestis* (Tran et al., 2011), and with a set of surprising results, including an uneven and unexpected impact on different cohorts by sex and age, a high parallel increase of mortality due to a synchronous smallpox epidemic and a raise in public violence (Ell, 1989).

Venetian death records from this period, also referred to as *necrologies*, are organized by parish and contain the systematic registration of every death among the resident population. These necrologies, edited by the parson, were established by decree since 1504 and kept in the archives of the responsible magistracy (Bamji, 2016). While death records were commonplace in all Christendom since the late Middle ages, and are commonly used for demography studies including on the plague (Alfani and Cohn Jr, 2007; Alfani and Murphy, 2017), Venetian records were particularly detailed. In the Patriarchal Archives of Venice, 54 out of more than 70 existing parishes at the time possess at least part of the registrations for the plague year (September 1630 to September 1631), while in the State Archive of Venice, the extant records for the plague year are few and scattered. Based on our assessments, these record series are overlapping and one (the former) constitutes the source for the other (the latter). We thus focus our efforts on the Patriarchal records. An example page from a necrology record is shown in Figure 2.1. Necrology records were kept in tiny and oblong books, with entries grouped chronologically by day. Typically, the most recurring details given for every entry were: the name, profession, sex and age of the person, the cause of death, approximate length of illness and whether a doctor attended them or not. The main dataset we use in what follows contains the number of daily deaths per parish. Data were collected following the work-flow illustrated in Figure 2.1; more details are given in the **SI**.

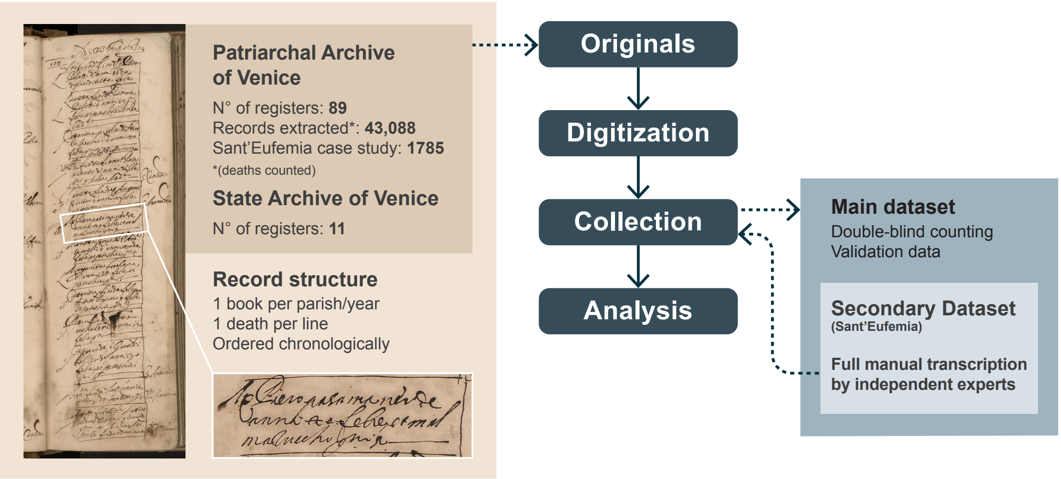


FIGURE 2.1: Illustration of the data collection workflow and datasets, including an example page from a death records book. The zoomed-in registration reads as follows: "*Messer Piero pasamaner de anni 40 febre et mal mazuccho giorni 5*", which roughly translates to "Mister Piero passementerie's weaver aged 40 fever and plague 5 days." What is meant is that Mister Peter, a passementerie's weaver forty of age (approximately), died of fever and plague after five days of sickness. It was the 23rd of October, 1630 (as it can be read at the top of the page).

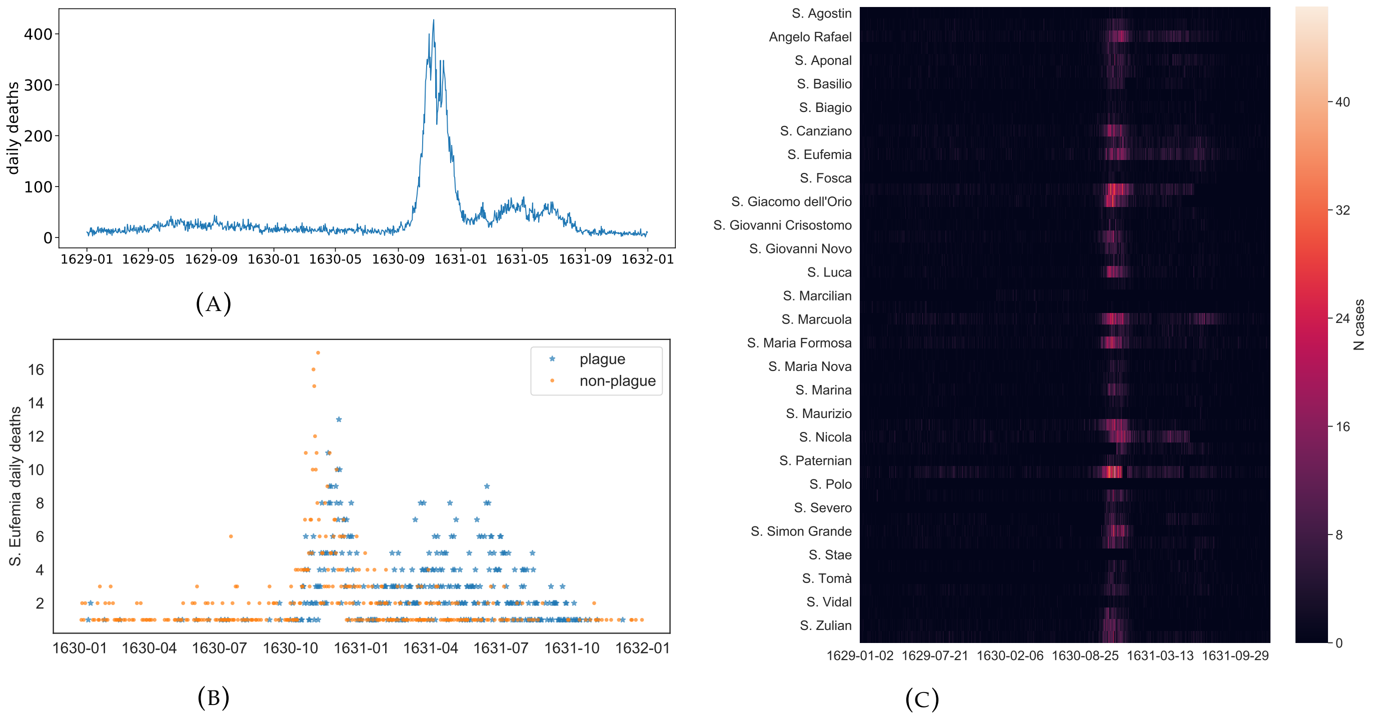
Results

Our data aggregated over all parishes clearly shows the massive outbreak which took place between the September and December of 1630, as detailed in Figure 2.2a. The death counts are staggering: 20,923 deaths between September and December 1630 alone, followed by 10,430 between January and August 1631. In total, 43,088 deaths were recorded over just three years. These numbers are in line with the 35% estimated mortality in northern Italy during the same epidemic outbreak (Alfani and Murphy, 2017), and should be compared to an estimated average annual mortality between 3.7 and 2.7% (but 29.7% for newly-born infants) during the whole seventeenth century (Beltrami, 1954; Gordon M., 1970). We stress that not all death records survived, therefore these numbers must be taken to represent a lower bound of the actual death toll. Historical demographic sources, even though uncertain (Favero et al., 1991), report a population of 141,625 inhabitants for Venice in 1624 and of 102,243 in 1633, a reduction of 27,81% (Beltrami, 1954; Gordon M., 1970).

The presence of a single peak of deaths is common in plague outbreaks within densely populated regions and cities (Welford and Bossak, 2009; Dean et al., 2018). Its presence in Venice indicates that the authorities' best efforts to contain the epidemic – for example by gathering all sick people in public hospitals or in their houses (Ell, 1989) – simply failed. The city was too densely populated and well-connected to leave any margin for containment. In fact, as it can be seen in Figure 2.5 (and especially 2.5c), the outbreak in 1630 swept through the parishes practically in sync, as no discernible space correlation is present. However, while the outbreak in 1630 is known, the subsequent 1631 long tail of high mortality has not been described in the literature before.

In order to gain a better understanding of the disease dynamics, we investigated another dataset taken from the records of a specific parish: *Sant'Eufemia*. This was a populous parish, with a significant amount of deaths in the 1631 tail and whose necrology records are well-preserved in their entirety. We transcribed all the information available in its necrologies, i.e. the name, sex and age at death of each person, together with the cause of death and the length of sickness. This transcription includes 1785 deaths registered between January 1630 and December 1631. The identification of deaths due to plague appears to be deceptively simple, as they were usually registered as fatalities caused by suspicious illness ("*mal sospetto*"), or with visible buboes. Nevertheless, previous studies have taken a more inclusive approach, considering also deaths not clearly caused by other factors as due to plague (Ell, 1989). We take the more conservative approach in what follows – see Tables 2.1, 2.2 and 2.3 for details on which causes of death were considered to be plague.

The statistics of the causes of death give us a first insight. In Figure 2.6a we show the distribution of deaths grouped by cause and (conservatively) classified as related to the plague or not. One can see how the two distributions are skewed, meaning that a small fraction of causes (5%) contributes to a large fraction of deaths (63%).

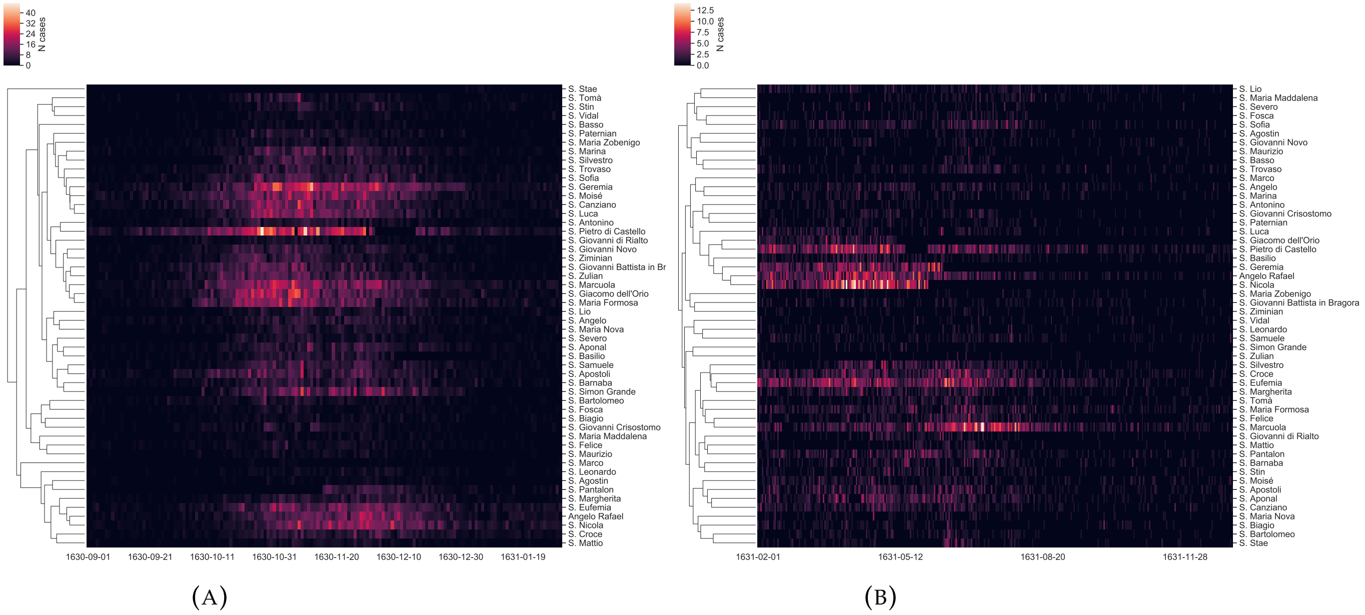


(A)

(B)

(C)

FIGURE 2.2: An overview of the full plague outbreak (main dataset): (a) Cumulative daily deaths for the whole recorded period (1095 days in total). A total number of 43,088 deaths were reported. One can clearly see the presence of a two-stage process, spanning until fall 1631. (b) Daily deaths recorded in the parish of *Sant'Eufemia*, almost surely due to plague (blue stars – $N_{plague} = 1007$) and possibly to other causes (orange circles – $N_{not\ plague} = 778$). Only days when someone died are considered. (c) A heatmap view of the dataset; for the sake of clarity, not all parishes names are plotted.



(A)

(B)

FIGURE 2.3: Hierarchical clustering of parishes zoomed on the main late-1630 peak (a) and on the 1631 outbreaks (b).

However, while the number of deaths clearly due to plague ($N_{plague} = 1007$) and possibly non-plague are similar ($N_{not\ plague} = 778$), only 56 out of 156 causes could be clearly attributed to plague, leaving more vagueness around the non-plague causes (in Figure 2.6b the causes with more than 50 deaths are listed). This seems to suggest that our plague-death counts likely constitute a lower bound of the total number of deaths directly linked to plague, which we cannot further refine from the records. In Figure 2.2b we show the time-series of deaths belonging to the *Sant'Eufemia* parish, distinguishing between those caused by the plague and the ones possibly due to other causes. Surprisingly, the first peak of the epidemic begins with few references to the common symptoms of the plague (October to November), when the records point instead to more generic and common illnesses, such as fever or spasms (Bamji, 2016). Only afterwards the records start to extensively mention the plague as the cause of death, well into the Fall of 1631. This might indicate an initial reticence to acknowledge the epidemic outbreak, as well as a subsequent possible overemphasis of it. This reticence might be caused by the public authorities' practice to quarantine the whole household in their house when someone from it died of plague. It might also be due to a surveillance issue generating a bias in the records: while many deaths were occurring, medical examination was no longer taking place and the registrations of the causes of death were not happening regularly, but instead in batches, leading to approximations. Furthermore, several people were moved to quarantine areas (*lazzaretti*) and died there, while their registration happened subsequently, possibly by reporting generic causes of death. It is thus likely that these deaths are also in large part attributable to plague. However, other explanations are also possible, such as a known epidemic of smallpox co-occurring during the main peak (Ell, 1989). Despite these limitations and open questions, *Sant'Eufemia*'s causes of death confirm the duration of the epidemic well into the autumn of 1631.

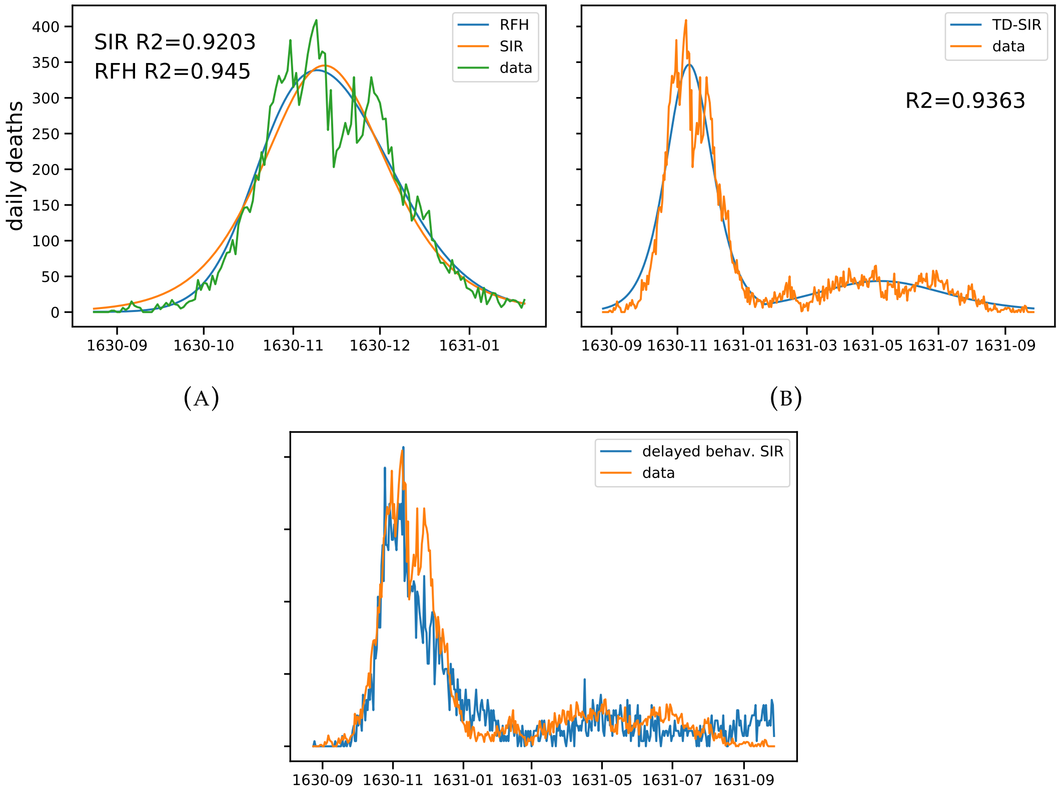
We further verify that deaths by plague were not significantly affected by sex, under the reasonable assumption that sexes were equally distributed in the population of Venice at the time (Gordon M., 1970). Indeed, the male to female deaths ratio was close to one ($N_{male}/N_{female} = 865/917 \sim 0.94$), a result confirmed by the majority of the literature (Alfani and Murphy, 2017; Whittles and Didelot, 2016; DeWitte, 2009; Bradley, 1977; Scott and Duncan, 2001; Schofield, 1977), with few exceptions (Ell, 1989; Curtis and Roosen, 2017). Furthermore, the distribution of illness duration and of age at death did not significantly change with sex (see Figure 2.7a and 2.7b respectively). Assessing the effect of the plague on age is challenging, as assumptions on the age distribution of population at that time are quite difficult to make and historical statistics are hard to find. Furthermore, the literature on the effect of the plague on different age cohorts is still ambiguous. Nevertheless, our data are in line with previous studies (Abrate, 1972; Manfredini, De Iasio, and Lucchetti, 2002; Alfani and Cohn Jr, 2007; DeWitte, 2010; Alfani and Murphy, 2017) indicating that the plague had higher relative impact among age cohorts of typically low mortality, in particular adolescents and adults between 14 and 44 years of age, as shown in

Figure 2.7c and Figure 2.7d.

Figure 2.2c shows the heatmap of reported cases, for each of the parishes of Venice, for the entire time window ($N_{tot\_deaths} = 43088$). One can see that while the main outbreak occurring in the last four months of 1630 shows good synchronization across all parishes, the second, smaller outbreak occurring until fall 1631 seems to have peaked at rather different time points within each parish, between February and July 1631. We therefore investigate whether space patterns are present, especially in the 1631 outbreaks ($N_{deaths\_tail} = 10363$). In order to assess the presence of spatial patterns, we simply plot the pairwise correlation among cases for all couples of parishes, against the distance between parishes (Figure 2.8). The resulting scatter plots show no spatial patterns. Nevertheless, the secondary outbreak in 1631 does not seem to have peaked as homogeneously as the first large outbreak in 1630 (Figure 2.2c). We hence performed a clustering analysis to highlight possible groups of rather synchronous parishes (Figure 2.3). The analysis on the main 1630 outbreak (Figure 2.3a) appears instead to be in sync across parishes.

The clustering on the 1631 outbreaks (Figure 2.3b) shows clusters of parishes with more spread-out peaks, across the first half of 1631, with tails reaching the fall of the same year. The main cluster is the one led by the three populous parishes of *S. Geremia*, *Angelo Rafael* and *S. Nicola*, with peaks between March and May 1631 (central part of Figure 2.3b). Another cluster is the one led by the *S. Eufemia* and *S. Marcuola* parishes (bottom part of Figure 2.3b), a more heterogeneous group, with peaks occurring mostly in June/July 1631.

Even though these clusters seems to be well separated in time, there is no clear evidence of a specific process or event in the history of the city that might have driven this spatial distribution of localized epidemics in different parishes during 1631. We therefore assess epidemiological models on data aggregated over all parishes. The plague is generally modeled as a zoonosis, in which the transition from an epizootic (typically, in rodents) to a human epidemic is mediated by animal fleas, the vector carrying *Yersinia Pestis* (Keeling and Gilligan, 2000b; Monecke, Monecke, and Monecke, 2009). From here on, we refer to this model as the Rats-Fleas-Humans (RFH) model. At the same time, other studies suggest that these models are not always preferable to explain the outbreaks dynamics, especially due to the 'efficacy and speed' of some historical plague outbreaks (Alfani and Murphy, 2017), if compared to the typical dynamics of RFH models. We first confirm that neither a deterministic RFH nor a deterministic Susceptible-Infected-Removed (SIR) model can explain the presence of the 1631 secondary outbreaks (see Figure 2.9b). We then investigate the transmission nature of the Venice plague, by considering separately the main 1630 outbreak and the one in 1631. In both cases, we find that the RFH model did not perform much better than a simple SIR model, as shown in Figure 2.4a (main 1630 outbreak), and Figure 2.9c (1631 outbreaks). We therefore implement a time-dependent SIR and find that it can better explain the dynamics over the



(A)

(B)

(C)

FIGURE 2.4: (a) Best fit comparison of a simple SIR model against the model from (Monecke, Monecke, and Monecke, 2009) on the main outbreak peak (150 days time window). (b) Best fit of an explicit time-dependent SIR; parameters are shown in Figure 2.9d. (c) Example realization of a stochastic delayed behavioral SIR; the evolution of transmission rate $\beta(I)$ is shown in Figure 2.9e.

entire time window (Figure 2.4b), with an increase in the basic reproduction number that could indicate a change in the transmission mechanism of pathogen (for clarity, fitted parameters are reported in Figure 2.9d). In particular this might suggest a transition from bubonic to pneumonic plague, a shift already hypothesized for other historical plague epidemics (Schofield, 2016). However, a change in the effective transmission rate might also be due to people's behavioral response to the outbreak. In order to investigate the fitness of such hypothesis, we implement a stochastic delayed behavioral SIR (details can be found in the `Methods`). In Figure 2.4c we show one example of such model's stochastic realizations, which presents both a main peak and a long tail dynamics. This shows that a change in pathogen's transmission route is not necessarily required in order for the epidemic to show a non-trivial temporal pattern, such as the one present in our data. For the sake of completeness we also check whether a deterministic delayed behavioral SIR would fit our data. In Figure 2.10b, we show that it cannot actually reproduce the 1631 tail, in spite of a good fit of the first part of the 1630 outbreak.

Although a change in diffusion parameters seems to provide a reasonable explanation of the two-peak structure, we investigate the possibility of having two-peak outbreaks similar to the observed one, as a result of the stochastic nature of the disease spread combined with structural properties of the host network. It is indeed known that the community structure of a network can strongly impact epidemic dynamics (Salathé and Jones, 2010). We therefore perform a series of stochastic simulations of a simple SIR process on top of a small-word graph, a network model which is likely to resemble the modular structure of social contacts (Schnettler, 2009) (further details on the simulations are given in the `Methods`). We find that few simulated epidemics do resemble the data, as shown in Figure 2.10a. However, as this happens in only about 0.1% of the simulations, such alternative interpretation of the 1631 tail based on pure stochastic effects and network structure, although reasonable, remains very unlikely.

Discussion

In summary, we find a novel epidemic pattern of two peaks in the 1630-31 plague outbreak in Venice. The first peak in 1630 was very high, and the outbreak highly synchronized among all parishes; the second peak in 1631 shows temporal variability, and was much less pronounced in strength. Most previous recorded cases show a single main peak (Welford and Bossak, 2009; Monecke, Monecke, and Monecke, 2009; Dean et al., 2018) of varying duration (Alfani and Cohn Jr, 2007; Whittles and Didelot, 2016), with possible cyclical recurrence (Welford and Bossak, 2009). Relying on fine-grained daily death records (Alfani and Murphy, 2017), we are able to confirm that the plague spanned both the main peak and the long tail, over a period of more than a year and caused the death of approximately 30% of the city's population.

Providing an interpretation of the two-stage process remains challenging with the evidence at our disposal. Firstly, not all deaths could be clearly attributed to the plague during the early weeks of the main peak. Generic causes of death such as fever and spasms might indicate plague deaths as well as deaths due to other causes. A first hypothesis is therefore that the same plague epidemic went on for more than a year, while being aggravated by other concomitant causes during the main peak. An alternative hypothesis is that two distinct plague epidemics took place instead, one during the main peak and another during the long tail. Previous studies suggest the possibility of a transition from a mainly bubonic to a mainly pneumonic plague, for example. Furthermore, we show that it is also possible that such temporal pattern could be generated by the adaptation of hosts' behavior to the increase of the number of infected, effectively decreasing the transmission rate, as the outbreak advances. Lastly, social factors such as the timing and effectiveness of public containment policies could have played a role.

Further investigations will be needed in order to fully qualify the Venetian 1630-31 plague outbreak, as well as the Second Pandemic overall. Indeed, as we have shown, historical records contain information which has so far been relied upon only to study few episodes but, when digitized and made available at scale and systematically, can help cast new light on these long-lasting research issues. For an understanding of detailed local dynamics, but also of global patterns of disease spread, modern human data and animal research can now be complemented with digital data collection driven by the digital and medical humanities.

Methods

Data collection

The main dataset we consider consists of the daily number of deaths per parish, from January 1629 to December 1631. We have first proceeded with a full double-blind counting, then compared the two series, checking and correcting all discrepancies. Secondly, two different co-authors have counted again all deaths from a sample of 20 parishes out of 70 (8 and 12 each), to further assess our main dataset, with the following results:

- 1629: 22 errors over 2395 assessed registrations (0,91%).
- 1630: 60 over 8989 (0,66%).
- 1631: 16 over 3730 (0,42%).

Confirming that the main dataset was already of high quality. Eventually, all remaining errors were checked again and corrected in the final dataset, which we analyze in this contribution.

We note that the parson of every parish was supposed in principle to a) get a medical inspection of every dead body to rule out contagious causes, b) report all deaths every morning to the magistrate called *Provveditori alla Sanità*, c) get burial licenses from this magistracy before inhumation. Steps a and c usually were not taking place during the months of peak mortality at the end of the year 1630. It is important to clarify that our death records include deaths which occurred in the main care institutions in Venice: the four *Ospedali Grandi* (main hospitals), as well as minor ones, with respect to residents in the available parishes. They also include all deaths occurred at the *lazzaretti*: temporary locations setup for quarantine or inhumation of persons affected by the plague. They do not include foreigners. We finally note that the parish of *S. Nicola* is to be identified with *San Nicola dei Mendicoli*.

Data analysis and modeling

All data analysis and modeling are done in Python. For the general data cleaning we use the `pandas` package. The distance between two parishes is defined as the geodetic distance between the centers of the corresponding polygons, defining the jurisdiction of the same parishes. The `geodesic` function from the `geopy.distance` module is used for this task.
All dendrograms (Figure 2.3) are plotted using the `seaborn.clustermap` package. In particular, we use the metric `correlation`[1] and the method `complete` to build the linkage matrix, needed to compute the clusters. The compartmental epidemic models are integrated using the `odeint` function from the `scipy.integrate` module. The parameters estimations are then obtained using the `curve_fit` and `differential_evolution` function from the `scipy.optimize` package (Jones, Oliphant, and Peterson, 2001). In order to account for false positives, we estimate a baseline of deaths very likely to be unrelated to the plague outbreak, by fitting a sinusoidal signal from the beginning of the recordings, until the end of August, as shown in Figure 2.9a. In the time-dependent SIR model we assumed a simple step function dependence for both $\beta(t)$ and $\gamma(t)$, leading to a total of five fitted parameters: $\beta_1, \beta_2, \gamma_1, \gamma_2$ and the transition time τ (see again fig 2.9d).
Stochastic simulations in fig. 2.4c and 2.10a were done using the `ndlib` package (Rossetti et al., 2018), on graphs generated with the `networkx` package (Hagberg, Schult, and Swart, 2008).
The delayed behavioral SIR model (fig. 2.4c) was defined using the following expression for the transmission rate $\beta(t) = \beta_0 e^{-I(t-\tau)/I^*}$, where β_0, τ and I^* were fitted parameters, together with the usual (constant) death rate γ and initial number of infected I_0 ($\beta_0 = 0.06429$, $I^* = 72$, $\tau = 32$, $\gamma = 0.02859$, $I_0 = 3$). For its stochastic implementation we used a Erdos-Renyi graph, with an edge creation probability $p = 4/N_{nodes}$ ($N_{nodes} = 20000$).

[1] For more details, find here the description of possible metrics: scipy.spatial.distance.pdist.

Data availability

All code and data needed to reproduce plots and analysis presented in the manuscript will be made available in a dedicated GitHub repository before publication.

Acknowledgements

We would like to thank the support of the Patriarchal Archive of Venice and the State Archive of Venice during data collection. We thank Paolo de los Rios and Giulio Rossetti for useful discussions on diffusion processes on networks, and gratefully acknowledge the help of Laurent Bolli in designing Figure 2.1.

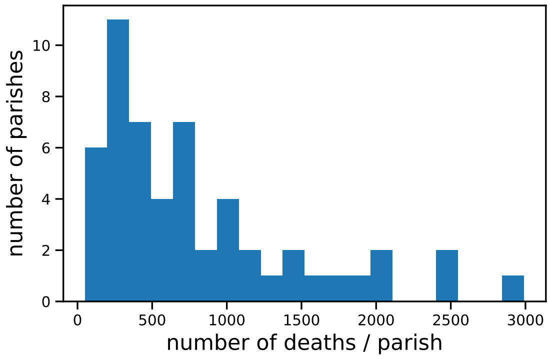
Author contributions statement

G.L. and G.C. performed the analysis. D.D., F.Z., F.B., A.E. and G.C. performed data collection. G.L., G.C., M.S. and F.K. wrote the paper. M.S. and F.K. supervised the study.

Additional information

The authors declare no conflict of interests.

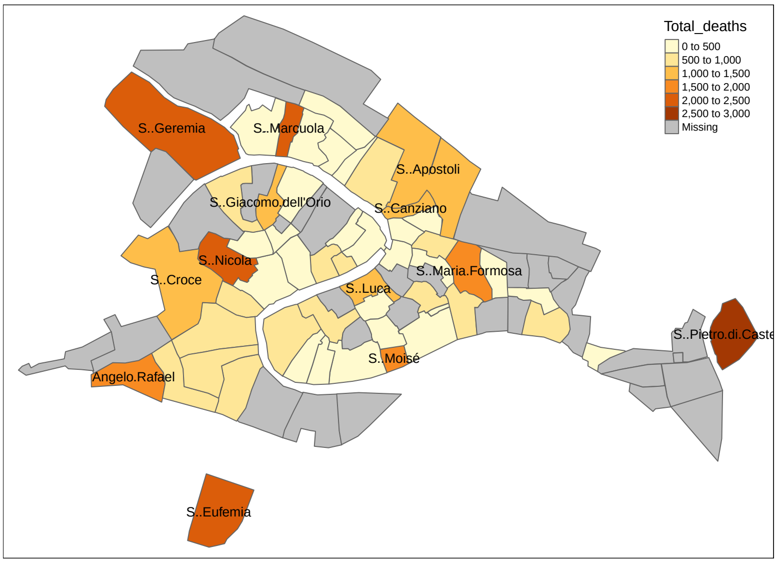
Supplementary Information



(A)

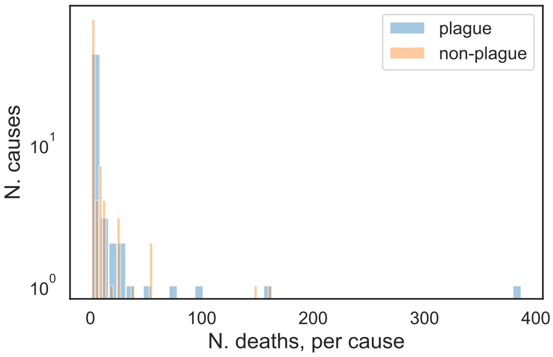
Parish	N deaths
S. Eufemia	2089
S. Nicola	2097
S. Geremia	2405
S. Marcuola	2491
S. Pietro di Castello	2990

(B)



(C)

FIGURE 2.5: Distribution of number of deaths by parish. (a) One can clearly see the skewed distribution, with the top 24% of parishes accounting for about 55% of the total deaths. (b) For the sake of clarity, only parishes with more than 2000 deaths are listed. (c) Map of Venice parishes, color-coded by total recorded deaths, summed over the entire time-window. For clarity, only the names of parishes with more than 1000 total deaths are shown.

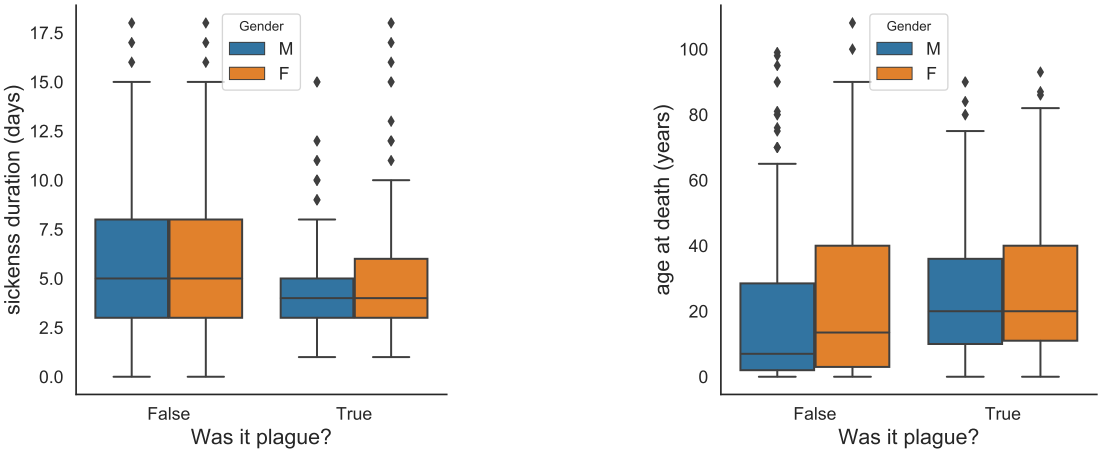


(A)

Cause of death	N deaths
mal sospetto	387
febbre	163
petecchie nere	157
spasimo	148
contagio	100
mal contagioso	73
vermi	56
nascente	54

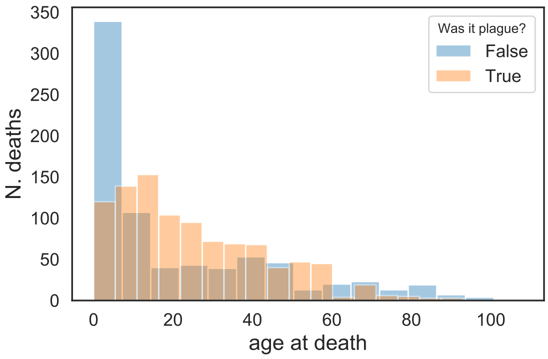
(B)

FIGURE 2.6: Distribution of number of deaths by cause, for the parish of *Sant'Eufemia*. In the records, 156 unique causes of death are found, of which 56 were attributed to plague. One can clearly see the skewed distribution, with the top 5% of causes accounting for about 63% of the overall deaths (a). For the sake of clarity, only causes with more than 50 deaths are listed; in bold the ones attributed to plague (b).



(A)

(B)

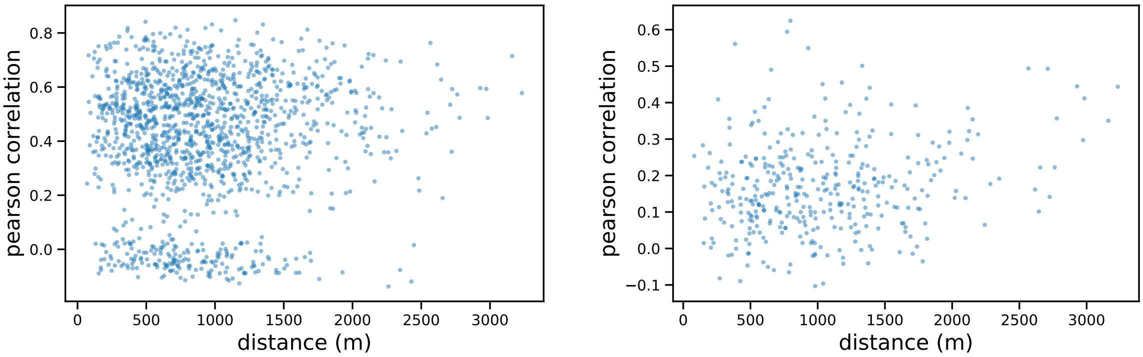


(C)

Age cohort (years old)	Plague	Not plague	Aggregated
Infant (0-2)	3.38%	23.52%	12.16%
Child (3-13)	32.01%	29.2%	30.42%
Young (14-23)	20.95%	8.74%	15.63%
Adult (24-44)	29.2%	16.45%	23.64%
Old (45+)	17.28%	19.28%	18.15%

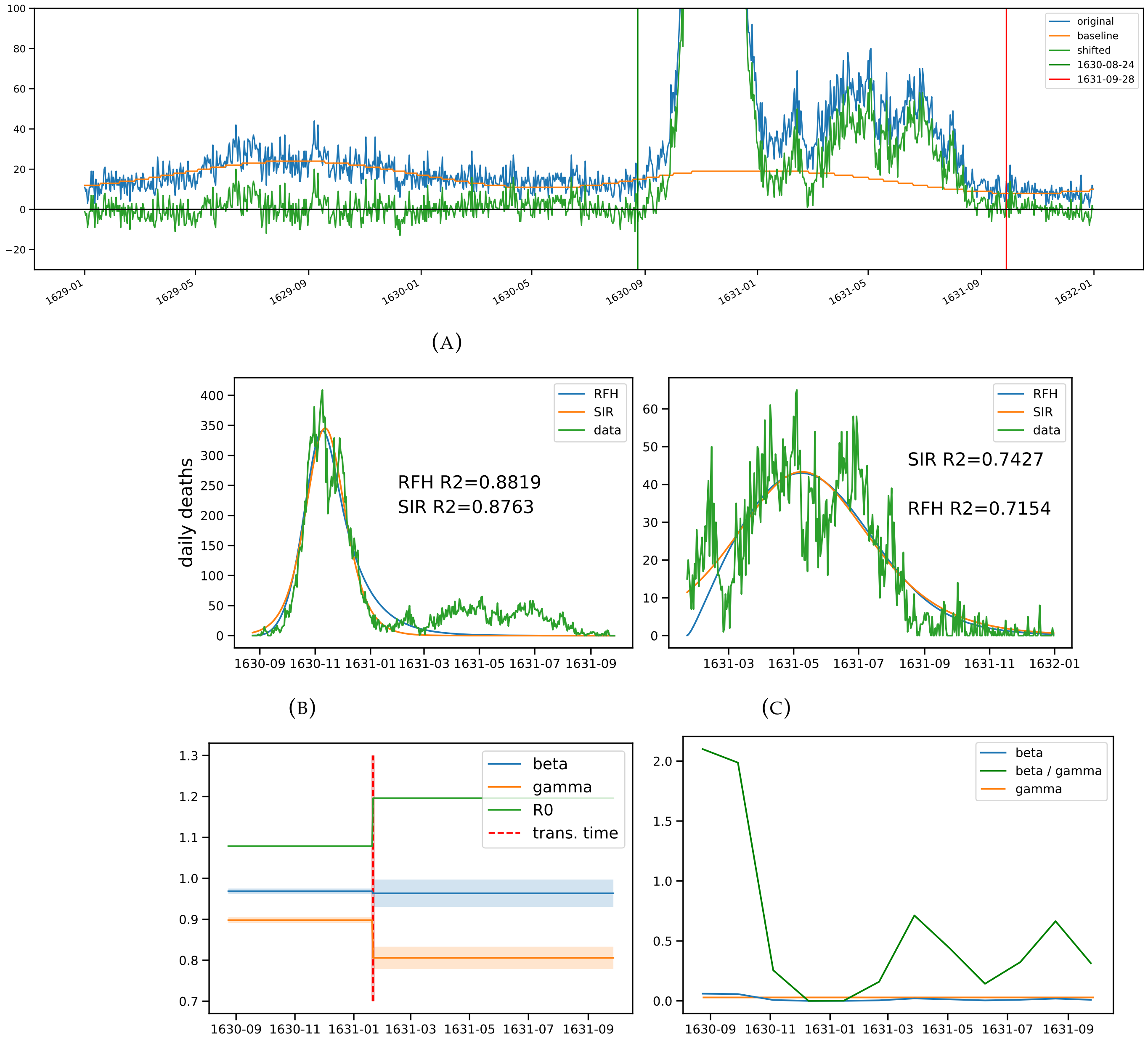
(D)

FIGURE 2.7: Demographic statistics for the parish of *Sant'Eufemia*. (a) Distribution of sickness duration, by sex and cause of death for *S. Eufemia* death records – for sake of clarity the boxplots include only cases with a sickness spanning less than 20 days (this still covers about 88% of the total sickness duration distribution). (b,c) Distributions of age at death, stratified by cause of death. No significant age difference emerges due to sex ($p > 0.001$ on two samples KS test, for both causes of death) (b), while a significant one appears between the plague VS non-plague deaths, aggregated over sex ($p < 10^{-20}$ on two samples KS test) (c). (d) Table with the same numbers of deaths, divided into age groups. Note that an infant mortality at birth between 20 and 30% was common at the time (Gordon M., 1970).



(A) (B)

FIGURE 2.8: Pairwise Pearson correlation between cases time-series of each couple of parishes, as function of distance between the two parishes. The same scatter plot for all parishes ($N_{parishes} = 54$) for the entire time-windows (a) and for the largest parishes, for the 1631 outbreaks only (b). The largest parishes are defined as those reporting more than 500 deaths ($N_{parishes} = 28$).



(A)

(B)

(C)

(D)

(E)

FIGURE 2.9: (a) Estimation of baseline cases due to other causes: a sinus function (orange line) is fitted from the beginning of the data until the beginning of the fit (green vertical line) to estimate the mortality rate, which is then applied to the original data (blue) to get the data used for the fit (green). (b-c) Comparison between a simple SIR and the more complex RFH model, in the 400 days window (b) and zoomed on the second part of the epidemic (c). (d) Fitted parameters for the time-dependent SIR model, as in Figure 2.4b. (e) Evolution of fitted $\beta(I)$ and $\beta(I)/\gamma$ for the delayed behavioral SIR shown in Figure 2.4c.

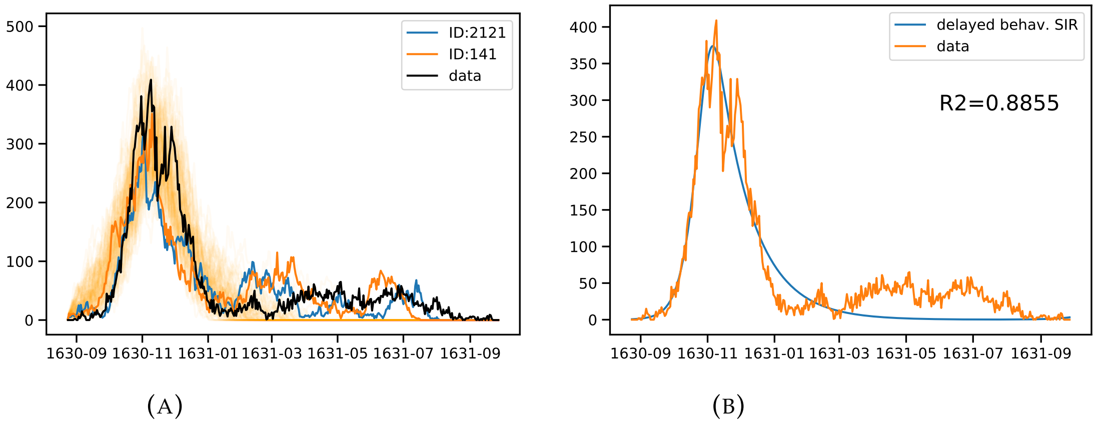


FIGURE 2.10: (a) Selected stochastic simulations of a simple SIR model on top of a small-word network. Two particular epidemics are highlighted, in order to show the possibility of having a large peak followed by a long tail, as present in the data. In shaded orange we only show, for sake of clarity, the simulated epidemics with lowest deviation from the data ($RMSE < 50 - N = 93$) (b) Best fit of deterministic behavioral delayed SIR. Although the model can fit very well the first part of the epidemic, it does not show a secondary outbreak, in 1631.

Death causes	**Related to plague ?**
annegato	False
apoplessia	False
brusco	False
bubbone pestilenziale all'inguine	**True**
caduta	False
caduta	False
caduta apoplettica	False
caduto da una scala	False
caita nella gola	False
cancro alla bocca	False
carbone	**True**
carboni	**True**
carboni e parto	**True**
carbonie e petecchie nere	**True**
catare nella gamba	False
catarro	False
contagio	**True**
contagio e petecchie nere	**True**
convertito etico	False
croplesion	False
disperso	False
doglia di testa	False
doglia di testa e mazzucco	**True**
doglia di testa e vermi	False
doglia e spasimo	False
dolor di vita	False
febbre	False
febbre	False
febbre continua	False
febbre continua e altre indisposizioni	False
febbre e catarro	False
febbre e doglia	False
febbre e doglia di schiena	False
febbre e doglie di testa	False
febbre e ferita	False
febbre e flusso	False
febbre e gotta	False
febbre e lepra(?)	False
febbre e mazzucco	**True**
febbre e nosella	False
febbre e petecchie	**True**
febbre e petecchie nere	**True**
febbre e petecchie rosse	**True**
febbre e petecchie rosse non pestilenziali	False
febbre e ponta	False
febbre e spasimo	False
febbre e suspetto	**True**
febbre e un brusco	False
febbre e una doglia in un fianco	False
febbre e una postiema	False
febbre e una scorencia	False
febbre e variole	False
febbre e vecchiezza	False
febbre e vermi	False
febbre etica	False
febbre etica e catarro	False
febbre ferita e flusso	False
febbre galica e catarro	False
febbre maligna	**True**
febbre maligna e mal sospetto	**True**
febbre maligna e punti	**True**
febbre senza sospetto	False
ferita	False
ferita	False
ferita dietro l'orecchio	False
ferite	False
ferite da peste	**True**

TABLE 2.1: Manual classification of all 156 death causes as reported in necrologies, as associated to plague or not – part 1.

Death causes	**Related to plague ?**
ferito	False
flusso	False
fracassato la testa	False
illegibile	False
incinta da febbre e doglie	False
infermo	False
ipolesia	False
ipoplessia e febbre	False
macchie nel petto giudicate pestilenziali	**True**
mal caduco	False
mal caduco e vermi	False
mal contagioso	**True**
mal di febbre	False
mal di gotta	False
mal di mare e mal sospetto	**True**
mal di mazzucco	**True**
mal di pietra	False
mal di reni	False
mal mazzucco	**True**
mal sospetto	**True**
mal sospetto e petecchie	**True**
mazzucco	**True**
morto improvvisamente	False
n.d.	False
nascente	False
non aver latte	False
non si sa il male	False
nosella	False
nosella di mal contagioso	**True**
nosella nel cuore	False
paralitico senza contagio	False
parto	False
parto e febbre	False
parto e ferita	False
parto e spasimo	False
partorito morto	False
patimento	False
per non aver avuto latte	False
percossia	False
peste	**True**
peste e petecchie nere	**True**
peste e strupiata	**True**
petecche paonazze	**True**
petecchi nere	**True**
petecchie	**True**
petecchie	**True**
petecchie e febbre maligna	**True**
petecchie e mazzucco	**True**
petecchie e spasimo	**True**
petecchie e un brusco	**True**
petecchie et un brusco	**True**
petecchie nere	**True**
petecchie nere	**True**
petecchie nere contagiose	**True**
petecchie nere e carbone	**True**
petecchie nere e rosse	**True**
petecchie nere pestilenziali	**True**
petecchie pestilenziali	**True**
petecchie rosse	**True**
petecchie rosse e alcune nere	**True**
petecchie rosse e parto	**True**
petecchie rosse verso il nero	**True**

TABLE 2.2: Manual classification of all 156 death causes as reported in necrologies, as associated to plague or not – part 2.

Death causes	**Related to plague ?**
punta	False
rogna	False
sconosciuta	False
sempre infermo	False
senza peste esterna	False
senza sospetto	False
spasimo	False
spasimo e infermitá	False
spasimo e mazzucco	**True**
spasimo e petecchie nere	**True**
spasimo e sturioli	False
spasimo e vecchiezza	False
spasimo e vermi	False
stroppiata su la palada	False
strupiata	False
tumore	False
tumore alla gola	False
un carbon	**True**
variole	False
variole e sturioli	False
vecchiezza	False
vecchiezza e febbre	False
vermi	False
vermi e petecchie	**True**
vermi e spasimo	False

TABLE 2.3: Manual classification of all 156 death causes as reported in necrologies, as associated to plague or not – part 3.

Chapter 3

Assessing the Dynamics and Control of Droplet- and Aerosol-Transmitted Influenza Using an Indoor Positioning System

Timo Smieszek[1,2,3,+], Gianrocco Lazzari[4,+], Marcel Salathé[4,*]

[1]Modelling and Economics Unit, National Infection Service, Public Health England, London, UK
[2]MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College School of Public Health, London, UK
[3]Center for Infectious Disease Dynamics, The Pennsylvania State University, University Park, USA
[4]Global Health Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL),Lausanne,Switzerland

*marcel.salathe@epfl.ch
+these authors contributed equally to this work

Abstract

There is increasing evidence that aerosol transmission is a major contributor to the spread of influenza. Despite this, virtually all studies assessing the dynamics and control of influenza assume that it is transmitted solely through direct contact and large droplets, requiring close physical proximity. Here, we use wireless sensors to measure simultaneously both the location and close proximity contacts in the population of a US high school. This dataset, highly resolved in space and time, allows us to model both droplet and aerosol transmission either in isolation or in combination. In particular, it allows us to computationally quantify the potential effectiveness of overlooked mitigation strategies such as improved ventilation that are available in the case of aerosol transmission. Our model suggests that recommendation-abiding ventilation could be as effective in mitigating outbreaks as vaccinating approximately half of the population. In simulations using empirical transmission levels observed in households, we find that bringing ventilation to recommended levels had the same mitigating effect as a vaccination coverage of 50% to 60%. Ventilation is an easy-to-implement strategy that has the potential to support vaccination efforts for effective control of influenza spread.

Keywords: Influenza, Disease Dynamics, Wireless Sensor Networks, Control, Ventilation

Introduction

Despite extensive clinical experience and decades of research on influenza, it is still not fully understood how influenza is transmitted among humans. Traditionally, influenza transmission has been assumed to occur through the air, physical contact between humans, and by touching contaminated surfaces (i.e., fomites) (Brankston et al., 2007; Killingley and Nguyen-Van-Tam, 2013) . Airborne transmission can occur in two ways: either through relatively large particles of respiratory fluid (droplets; 10^1-10^2 μm) or through smaller such particles that can remain aerosolized (droplet nuclei; «10^1 μm) (Fabian et al., 2008; Gralton et al., 2011; Stilianakis and Drossinos, 2010; Weber and Stilianakis, 2008; Tellier, 2006). As larger droplets are pulled to the ground by gravity quickly, droplet transmission requires close physical proximity between infected and susceptible individuals, whereas aerosolized transmission can occur over larger distances and does not necessarily require that infected and susceptible individuals are at the same location at the same time (Tellier, 2006).

Until recently, close contact transmission was considered to be the dominant transmission pathway, largely because the evidence to support the importance of transmission through aerosols was mixed (Brankston et al., 2007; Tellier, 2006). However, the question of the importance of the various transmission routes has received renewed attention recently, and multiple studies have in the past few years provided evidence for the importance of aerosol transmission (Atkinson and Wein, 2008; Tellier, 2009; Mubareka et al., 2009; Wong et al., 2010; Noti et al., 2012; Cowling et al., 2013; Lau et al., 2015). There is increasing evidence from experiments with mammalian hosts that airborne transmission is much more efficient than fomite transmission (Mubareka et al., 2009; Xiao et al., 2018). Data from randomized controlled trials of hand hygiene and surgical face masks in households provided evidence that aerosol transmission accounts for half of all transmission events (Cowling et al., 2013; Lau et al., 2015). A nosocomial influenza outbreak with subsequent airflow analysis provided further evidence for the important role of aerosol transmission

(Wong et al., 2010). An experimental laboratory study using a patient examination room containing a coughing manikin provided further support for aerosol transmission (Noti et al., 2012). A recent study with outpatients who tested positive for influenza A virus demonstrated that 53% and 42% produced aerosol particles containing viable influenza A virus during coughing and exhalation, respectively (Lindsley et al., 2016). A mathematical model of influenza transmission within a household has suggested that the aerosol transmission route may not only be important, but indeed dominant (Atkinson and Wein, 2008). Another mathematical model suggests that aerosol transmission is the dominant mode of transmission in long-term epidemics, whereas larger droplets could play a dominant role for short-term epidemics with high attack rates (Stilianakis and Drossinos, 2010).

Given the increasing evidence supporting an important role of aerosol transmission, it is prudent to revisit our expectations on disease dynamics of influenza outbreaks, and the best measures to control the spread of influenza. To address this issue, we use a high-resolution dataset of a medium-sized US high school, where both individual's indoor positions and close proximity contacts to others were measured using wireless sensor network technology and perform computation simulations on it.

Results

We first investigate the dynamics of influenza spread in three different transmission models, namely a droplet-based, an aerosol-based, and a combined droplet-aerosol-based model (for a schematic explanation of these transmission models, please see Figure 3.1). These three models were chosen to compare the two extreme situations (droplet-only, and aerosol-only) as well as an intermediate situation where the two transmission modes are equally relevant (Cowling et al., 2013; Lau et al., 2015). Simulations of influenza outbreaks were based on an SEIR model run on a high-resolution contact network collected at a US high school (Salathé et al., 2010) using wireless sensor network technology. In addition, we used the location information of each individual obtained using the same technology. In order to simulate partial or full aerosol transmission, we combined this data with building data from the school in order to compute the relevant infection probabilities due to aerosol transmission as per eq. 3.3. The entire model, and the data sources, are described in full detail in the Methods. Figure 3.2 shows an example of quanta concentrations in multiple classrooms during a day.

As illustrated in Figure 3.1, both droplet and aerosol transmission can be represented as weighted networks. It has been shown previously that strength (i.e., the weighted degree) is a key network metric for understanding disease spread in networks: hosts with a high strength generate on average more secondary cases, hosts with a low strength less (Smieszek and Salathé, 2013; Bell and Atkinson, 1999; Christley et al., 2005). Figure 3.3 shows distribution of and correlation between measures of strength for both modes of transmission, aerosol (quanta) and droplet (contact duration). As can be seen, distributions are similar and both strength measures are correlated across modes of transmission.

For all three transmission models, we measure three epidemiological quantities: the final size of an outbreak (number of recovered $r(t)$ individuals after a simulation run), the total duration of an outbreak (time until there are no more exposed $e(t)$ or infected $i(t)$ individuals, respectively), and the time to reach the peak of the outbreak, i.e. to reach the maximal prevalence. In addition, in order to be able to put these quantities in context, we also measure R_0 for the three transmission models. In an individual-based model like the one used here, measuring R_0 is straightforward using the droplet-based transmission model, because one can directly track which individual infected which. However, in the case of partial or full aerosol-mediated transmission where infection is mediated by the air in a room, this tracking is harder because one would need to computationally keep track of the individual sources of infectious aerosol particles. We thus measure an alternative but similar quantity, namely the number of all cases infected during the initial time period until the index case recovers. We call this quantity R_0'. Theoretically, there is a chance that this overestimates the true R_0 by including third generation cases that were not infected by the index case, but given the values for the incubation period and the recovery rate, this is rather rare.

In fig. 3.4, we can observe that increasing the relative importance of aerosol-based transmission (i.e., shifting from pure close-contact transmission via combined to pure aerosol-based transmission) has no major effect overall on disease dynamics. We note that outbreak sizes in the pure droplet model are slightly increased, and the time to outbreak peak is slightly increased. However, the small increase in outbreak size is also reflected in a small increase of R_0' (see fig. 3.4-D). In summary, our results indicate that one should not necessarily expect influenza disease dynamics to be very different when taking aerosol-based transmission into account.

For a better understanding of the behavior of the transmission models, we performed a sensitivity analysis on the core infectivity parameters (fig 3.5). In particular, we altered the droplet infectivity in the close-contact component (baseline set to 0.003 – see eq. 3.1) and the shedding rate (q – see eq. 3.2) in the aerosol component within the model that captures both modes of transmission at 50%. We changed both parameters, relative to the baselines, by -20%, -10%, 0%, 10% and 20%. As can be seen in figure 3.5, changes on median outbreak size appear to be well-behaved and proportionate, regardless whether parameters for close-contact or aerosol transmission are changed.

In pure droplet-based transmission models, vaccination is a powerful strategy to mitigate the spread of an infectious disease. When transmission can also be aerosol-based, increasing ventilation is an additional way to curb the spread of disease.

We therefore compared the effect of ventilation to traditional vaccination strategies. According to the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE) (American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2016), a good ventilation in classrooms corresponds to 3 air changes per hour. Most classrooms, however, have poor ventilation at rates around 0.5 air changes per hour (see Methods for more details). In the pure aerosol-based model, bringing all rooms to the recommended ventilation rate would almost completely eliminate the chance of an outbreak. This corresponds to the same effect of complete vaccination coverage in the case of poor ventilation rates, as shown in fig. 3.6-B. In the combined droplet-aerosol scenario, improvements of ventilation still results in a significant decrease of outbreak sizes. In particular, fig. 3.6-A shows that in the combined droplet-aerosol model, a good ventilation would have a similar effect to a 50-60 % vaccination coverage in the poor ventilation scenario. This finding proved to be robust to changes in transmission parameters in the sensitivity analysis (data not shown).

In practice, upgrading the ventilation system of an entire school campus to the rates proposed by ASHRAE will often be challenging due to limited resources. We therefore asked how strong the mitigating effect would be of upgrading the ventilation of only a fraction of all rooms. We also asked how one would identify the optimal set of rooms for mitigation purposes. The room selection strategies we explore are *optimal*, *schedule-based*, and *size-corrected*, which are all described in the methods. As expected, applying good ventilation to less rooms instead of the entire school leads to less pronounced improvements. However, fig. 3.6-C shows that selecting only a fraction of rooms for improved ventilation, according to the criteria explained above, still results in median outbreak sizes comparable with those obtained in a setting with 30-40% vaccination coverage. In particular, the *size-corrected* strategy, which requires only information readily available to each school (i.e. school rosters and room size) can result in median outbreak sizes that are comparable to those obtained with vaccination rates above 40%, even when only applied to 25% of all rooms.

Comparisons between the effects of improved ventilation and vaccination also depend on vaccine efficacy. The baseline efficacy was assumed to be 60%. If the ASHRAE recommended air change rate were implemented school-wide, this would result in a protective effect in the combined model that would correspond to a vaccination coverage of 60-70% for an efficacy of 40%, 50-60% for 60% efficacy, and 40-50% for 80% efficacy. In the aerosol model, consistently improved ventilation beats vaccination even with full coverage if efficacies are low.

Discussion

There is mounting evidence that aerosol-transmission is an important factor in the spread of influenza (Atkinson and Wein, 2008; Tellier, 2009; Mubareka et al., 2009;

Wong et al., 2010; Noti et al., 2012; Cowling et al., 2013; Lau et al., 2015). Despite this, virtually all infectious disease dynamics models on influenza have thus far ignored aerosol-transmission. Here, we parameterized a model with empirically obtained values to investigate the dynamics and control of influenza under the assumptions of no, partial, or full aerosol transmission. In order to create a realistic model, we used contact network and location data that was previously obtained at a US high school using wireless sensor network technology (Salathé et al., 2010). This dataset is well-suited for the objective of the study: modeling droplet-based transmission requires data on close proximity contacts, and modeling aerosol-transmission requires data on location in rooms and information about the rooms, such as the room size. The dataset used here, even though limited in scope, and particularly also in duration (data were of only one day and had to be used repeatedly, thus exaggerating correlation between school days), contains all of this information.

Using empirical estimates of various influenza-related parameters, we found that the overall disease dynamics does not differ substantially between the models using no, partial, or full aerosol transmission. This isn't entirely surprising, given the fact that aerosol transmission parameters were estimated from the same influenza outbreak that was used to parameterize previous (pure droplet-based) influenza models. However, aerosol transmission does change the underlying transmission network (see fig. 3.1), which in turn could nevertheless have a substantial impact on disease dynamics, depending on the specific co-location patterns of individuals. Our finding that the dynamics do not change substantially in a model parameterized by empirical co-location data may simply be a reflection of the fact that schools are high density environments with comparatively limited movement, and - in line with this - we found that the underlying transmission-dependent network structures to be very similar with respect to the key network property strength. It should also be noted that we did not assume any virus inactivation in our model, largely because it is generally known to be slow for low relative humidity values (typical for indoor air during influenza season) (Weber and Stilianakis, 2008; Tellier, 2006; Yang and Marr, 2011). Nevertheless, this assumption may overestimate the relative effectiveness of ventilation somewhat.

While vaccination is at the heart of influenza prevention efforts, aerosol-based transmission of influenza opens up additional possibilities to control the spread of the disease. In particular, when infectious agents can remain airborne, air ventilation is a well known method to mitigate disease spread (Li et al., 2007). In this study, we assessed potential effects of bringing the air change rates up to the recommended levels by the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE) (American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2016), which defines an acceptable ventilation in classrooms to be 3 air changes per hour. We found that by doing so, we were able to generate reductions in expected outbreak sizes that would normally only be possible with a substantial vaccination coverage of 50-60%, which is within the range of observed vaccination rates in school settings (Barclay et al., 2014). Moreover, even when bringing only a quarter of the rooms to the recommended air change rates, using easy-to-obtain data in order to select the best rooms, we were still able to obtain outbreak size reductions that would require 30-40% vaccination coverage when air change rates are at levels commonly reported at US schools. The concrete percentages, obviously, depend on modelling assumptions as well as transmission parameters. Sensitivity analyses on the transmission parameters (not shown) revealed that both the close-contact and the aerosol transmission model component reacted similarly to changes of the transmission parameters.

Our results suggest that improvements of ventilation in high density public spaces could be an important and relatively easy-to-implement strategy supplementing vaccination efforts for effective control of influenza spread. This observation rests on the assumption that at a substantial part of influenza spread is due to aerosol-based transmission, for which there is mounting evidence. Given that increased air ventilation rates are not known to have any negative side effects, and that there are numerous infectious diseases that are entirely or partially transmitted via aerosol (e.g. tuberculosis), the findings here thus provide an additional argument corroborating the public health recommendations for good air ventilation. It should be noted that influenza vaccine effectiveness is often less than the here assumed 60% (Belongia et al., 2016; Osterholm et al., 2012a), whereas good ventilation would provide increased protection, further underlining its importance.

Methods

Data

The data used in this paper were collected at a US high school during one school day using wireless sensor technology. In total, 789 individuals (94% of the school population) participated. They wore small sensors that detect and record radio signals broadcast by other nearby sensors. Further, stationary devices broadcasting signals were attached to fixed locations (at least one per room) throughout the school campus to keep track of the participants' locations.

Consequently, the data include two types of records. Close proximity interactions (CPIs) are records that indicate two participating individuals standing face-to-face with a distance of less than three meters at a certain point in time. Location records are records that indicate the presence of an individual nearby a stationary device (location information is at the level of rooms).

A detailed description on how information and noise were separated in the data is provided elsewhere (Smieszek and Salathé, 2013). Data were collected at time intervals of 20 seconds.

Model of influenza spread

We used an individual-based model with a susceptible, exposed, infectious, recovered (SEIR)-type structure. We assumed that influenza is introduced into the school population by one index case at the beginning of a simulation run and that no further introductions from outside occur. The duration of a simulation time step was half a day (i.e., contact information was aggregated at this level, which was shown to be a reasonable approximation for full-resolution networks (Stehlé et al., 2011)); also note that the temporal resolution of the model of virus particle air concentrations was kept at 20 seconds and only exposure levels were aggregated at the this level, see also below. Infection transmission could only occur during the half-day including school, not during the half-day including the night. Individual j's probability P_j to switch from the susceptible to the exposed state depends on the mode of transmission. We defined one function $P_{a,j}$ for aerosol transmission and one function $P_{cc,j}$ for close-contact transmission, as laid out below. We ran simulations for an all-aerosol scenario, for an all-close-contact scenario, and one for a scenario where both aerosol and close-contact transmission occur as $0.5P_{a,j} + 0.5P_{cc,j}$. The duration of the exposed state follows a Weibull distribution with an offset of half a day; the power parameter is 2.21, the scale parameter is 1.10 (Ferguson et al., 2005; Salathé et al., 2010)). After that period in the exposed state, every individual will be in the infectious state for exactly one time step before turning into home confinement and, finally, recovering. To allow for the fact that the onset of influenza symptoms is typically sudden and that affected individuals will be dismissed quickly, we reduced P_j by 75%, as described in Salathé et al. (Salathé et al., 2010), which also contains more details on all parameter choices other than those specific to the aerosol transmission probability.

Close-contact transmission probability

Assumptions, parameters, and structure of the close-contact transmission model are described in detail elsewhere (Salathé et al., 2010), and therefore only described briefly here.

Close-contact transmission requires social interaction between an infectious and a susceptible individual, and it includes transmission via large droplets that do not travel far and do not stay suspended in the indoor air as well as transmission via direct, physical contact (Smieszek et al., 2014). Risk of transmission is usually operationalized as a function of contact duration (Smieszek, 2009). Based on data from an outbreak on a commercial airliner (Moser et al., 1979), the probability of transmission was estimated as

$$P_{cc,j} = 1 - (1 - 0.003)^{T} \tag{3.1}$$

with T being the contact duration between two individuals in number of sensor recordings (every 20s)(Salathé et al., 2010).

Aerosol transmission probability

In our model, we quantify amounts of aerosolized virus particles in 'quanta', following Wells' (Wells, 1955) quantum theory of disease transmission. A quantum is defined as the amount of infectious droplet nuclei required to infect the fraction $1 - 1/e$ of a susceptible population exposed to it.

We assume that every room of the school is a well-mixed airspace that is only connected to the outside, but does not exchange air with other rooms. We further assume that removal of aerosolized virus particles by ventilation is the dominant removal process and that, e.g., neither inactivation nor settling play an important role at low levels of relative humidity typical for influenza season (which is a standard assumption in the literature, cf., e.g., (Azimi and Stephens, 2013; Yang and Marr, 2011)). Under these assumptions, we model the concentration of virus particles in a particular room r as

$$\frac{\Delta C_{r,t}}{\Delta t} = \frac{\sum_{i \in I_{r,t}} q_{i,t}}{V_r} - C_{r,t}\frac{Q_{r,t}}{V_r} \tag{3.2}$$

where $C_{r,t}$ is the quanta concentration in room r at time t, $I_{r,t}$ is the set of all infectious individuals that are in room r at time t, $q_{i,t}$ is the quanta shedding rate of infector i at time t, V_r is the volume of room r, and $Q_{r,t}$ is the fresh air supply rate of room r at time t. The quotient $Q_{r,t}/V_r$ is also known as the air change rate (ACR).

The instantaneous dose of infectious material, $D_{j,t}$, inhaled by individual j at time step t is given by

$$D_{j,t} = C_{r,t} p_j \Delta t$$

where $C_{r,t}$ is the quanta concentration in room r - the room in which individual j is located at time t - at time t, p_j is the breathing rate of individual j, and Δt is the duration of a simulation time step, here 20s.

Individual j's total exposure, D_j during an entire school day is given by

$$D_j = \sum_{t=t_0}^{t_x} D_{j,t}.$$

Combining the total daily exposure with Wells' definition of quanta allows to model the probability $P_{a,j}$ of a fully susceptible individual to become infected during one simulation school day as

$$P_{a,j} = 1 - \exp(-D_j) \tag{3.3}$$

where the total exposure D_j is the only parameter required.

Shedding rate: Both bottom-up (Fabian et al., 2008; Chen and Liao, 2010) and top-down approaches (Rudnick and Milton, 2003) have been used by others to estimate quanta-based shedding rates for influenza. Bottom-up studies (mechanistic approaches starting from basic measurements and processes) suffered from huge uncertainties and differed by three orders of magnitude. We used data from Rudnick and Milton's (Rudnick and Milton, 2003) top-down study that back-calculated quanta shedding rates from the same outbreak data (Moser et al., 1979) that was also used to parameterize the close-contact model (Salathé et al., 2010). They estimated shedding rates of between 79 quanta/h and 128 quanta/h, depending on model assumptions. We chose a shedding rate of 100 quanta/h for our aerosol transmission model.

Ventilation rate: We assumed different scenarios for the ventilation rate. According to the ventilation recommendations for schools by the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE) (American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2016), the ventilation rate for classrooms should be at least 8 l/s per person. Daisey et al. (Daisey, Angell, and Apte, 2003) estimate that for a typical classroom situation, this corresponds to an air change rate (ACR) of 3.0 air changes per hour. This estimate served as our good-ventilation scenario. Various studies found substantially lower ACR in US schools, (Daisey, Angell, and Apte, 2003; Shendell et al., 2004; Mullen et al., 2011) and CO_2 concentrations at the high school we collaborated with (unpublished data) indicated very poor ventilation conditions, too. In line with reported ACR in other US schools, we assumed 0.5 air changes per hour for a poor ventilation scenario. Additionally, we used 1.5 air changes per hour as a middle scenario.

Breathing rate: The breathing rate of humans depends mainly on their age, gender, and activity levels (Adams, 1993). In line with Adams' (Adams, 1993) measurements and in accordance with other work in the field (Rudnick and Milton, 2003), we assumed a constant breathing rate of 8 l/min for every individual.

Interventions

We compared ventilation-based interventions (effect only on aerosol transmission) with vaccinating individuals (effect independent on transmission pathway).

Ventilation rate: Baseline scenario to which all intervention scenarios were compared with was the poor ventilation scenario (ACR 0.5 h^{-1}). Basic interventions were improving the ventilation to ASHRAE standards (ACR 3.0 h^{-1}) and to achieve an intermediate improvement (ACR 1.5 h^{-1}), respectively.

We further analyzed how ventilation improvement only in some rooms would affect infection spread. We defined three different methods to identify rooms for which the ACR was increased from 0.5 h^{-1} to 3.0 h^{-1}: (i) *optimal*, using all available information from simulation runs, we identified rooms with the highest cumulative exposure, i.e., where most susceptibles will be exposed or where doses are highest in a typical simulation run; (ii) *schedule-based*, identifying rooms with the highest cumulative occupancy throughout a school day according to the school's official roster; (iii) *size-corrected*, which is similar to the schedule-based approach, but the total occupancy was divided by the volume of the room to give priority to small rooms with a high occupancy, as quanta concentration builds up faster in small rooms.

All three methods were used to identify rooms that represent 5%, 10%, 15%, 20% and 25% of the total indoor space. Methods (ii) and (iii) could realistically be applied in a school setting. Comparing interventions based on them with optimal ones allows assessing their relative performance to the theoretical optimum.

Vaccination: We assumed a vaccination effectiveness (protection against transmission) of 60% (Osterholm et al., 2012b) as a standard scenario (In sensitivity analyses we also assumed 40% and 80%) and a random distribution of vaccination status among the school population. We simulated the impact of vaccination for vaccination coverage values between 0% (baseline) and 100% with increments of 10%.

Ethical approval

All measurements involving human subjects were conducted according to the relevant regulations and involved an informed consent obtained from all subjects. The whole study was previously approved by the Stanford IRB (Institutional Review Board).

Acknowledgements

This research was supported by a fellowship from the German Academic Exchange Service DAAD to T.S. (Grant D/10/52328); T.S. also thanks the UK National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Modelling Methodology at Imperial College London in partnership with Public Health England (PHE) for funding (grant HPRU-2012-10080). The views expressed are those of the authors and not necessarily those of the MRC, the NHS, the NIHR, the Department of Health, Public Health England or any other organization the authors' are affiliated with or received funding from.
We also thank the High-Performance Computing group (SCITAS) of EPFL for his technical support.

Author contributions statement

T.S. and M.S. conceived and designed the study. M.S. collected the data. G.L. and T.S. performed the simulations and analyses. All authors wrote, read and approved the final manuscript.

Additional information

Accession codes

The data code to reproduce results and figures are available on a GitHub repository at `https://github.com/salathegroup/aerosol`. The input files needed to reproduce the computer simulations can be found in a dedicated folder at `https:`

//github.com/salathegroup/aerosol/tree/master/input_files

Competing interests

The authors declare no competing interests.

Figures

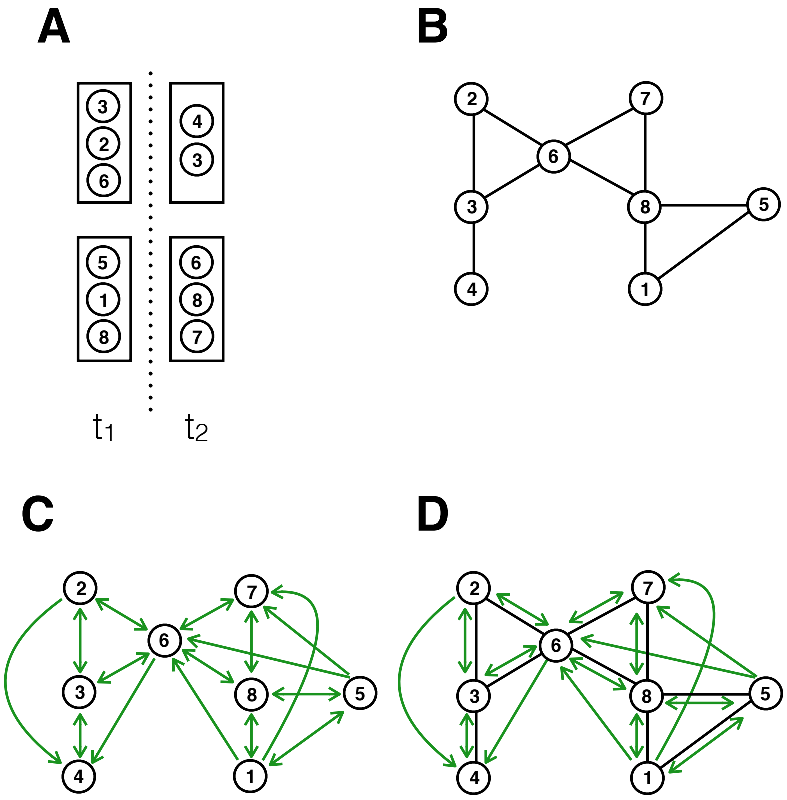


FIGURE 3.1: Simplified scheme of transmission routes. Different sets of individuals in the school can occupy the same rooms, at different times t1 and t2 (A). In these rooms, individuals may be in close proximity to one another. For visual simplicity, we assume in this figure that all individuals in the same room at the same time are in close proximity (but note that in the model, proximity is given by the sensor measurements). In the aerosol model, infected individuals can shed infectious material while in the room, which in turn may infect those individuals in the room concurrently or later on. The network of possible transmission pathways will therefore look different depending on the transmission routes. Based on the spatio-temporal pattern shown in panel (A), panel (B) shows the network of pure droplet transmission; panel (C) shows the network of pure aerosol transmission; and panel (D) shows the network of droplet and aerosol transmission combined. The edges for droplet transmission are always bidirectional, hence no arrows are shown. The edges for aerosol transmission may be unidirectional due to the temporal delay of virus shedding and virus uptake, hence arrows are shown.

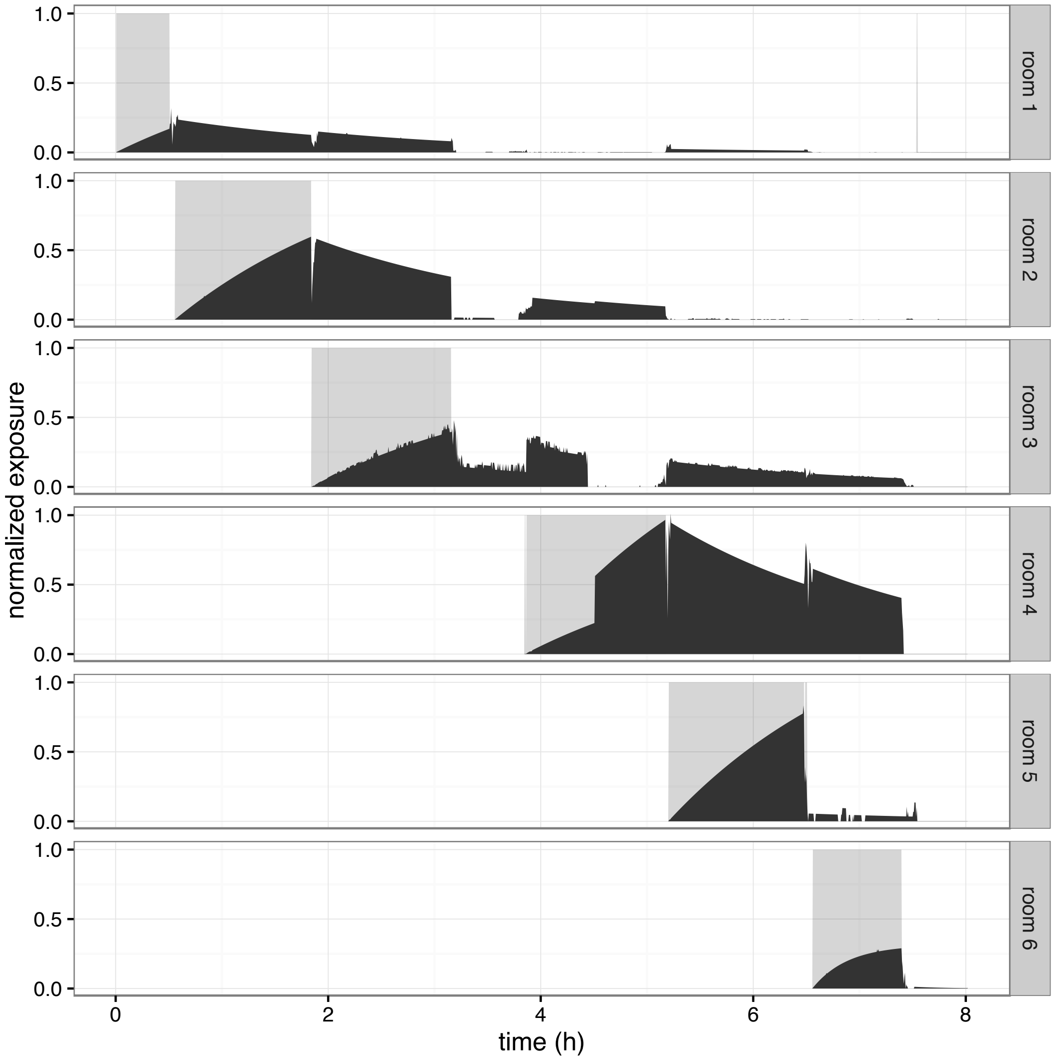


FIGURE 3.2: Illustration of relationship between presence of infected individual and exposure to others in the aerosol model. Presence of one (selected) infected individual in different rooms (gray area) and other people's exposure to infectious material shed by the individual in the respective rooms (black bars) across 8 hours; the scale of the x-axis is hours, the scale of the y-axis is exposure in [0 - 0.02] quanta. Exposure levels depend on the number of exposed individuals in the room following deposition of infectious material, as well as air change rates (here 0.5).

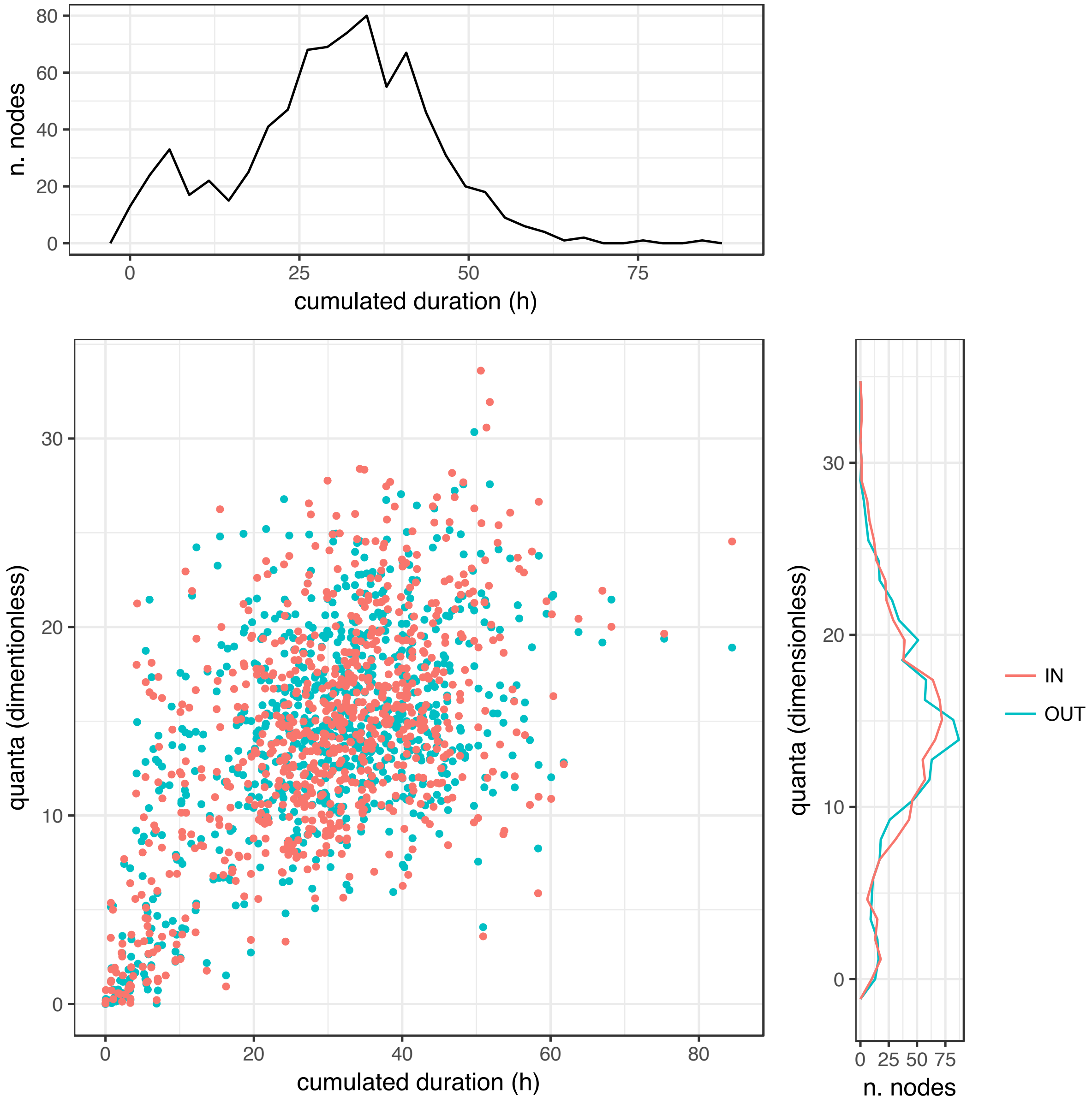


FIGURE 3.3: Weighted network representations of droplet transmission (cumulated contact duration) and aerosol transmission (cumulated quanta exposure). Quanta exposure caused in others represents outdegree, quanta exposure to the individual of interest represents indegree. Contact duration in droplet transmission is symmetic, therefore there is no difference between in- and outdegrees. Top figure and right figure illustrate weighted degree distributions; the scatter plot relates network metrics for droplet and aerosol transmission to each other.

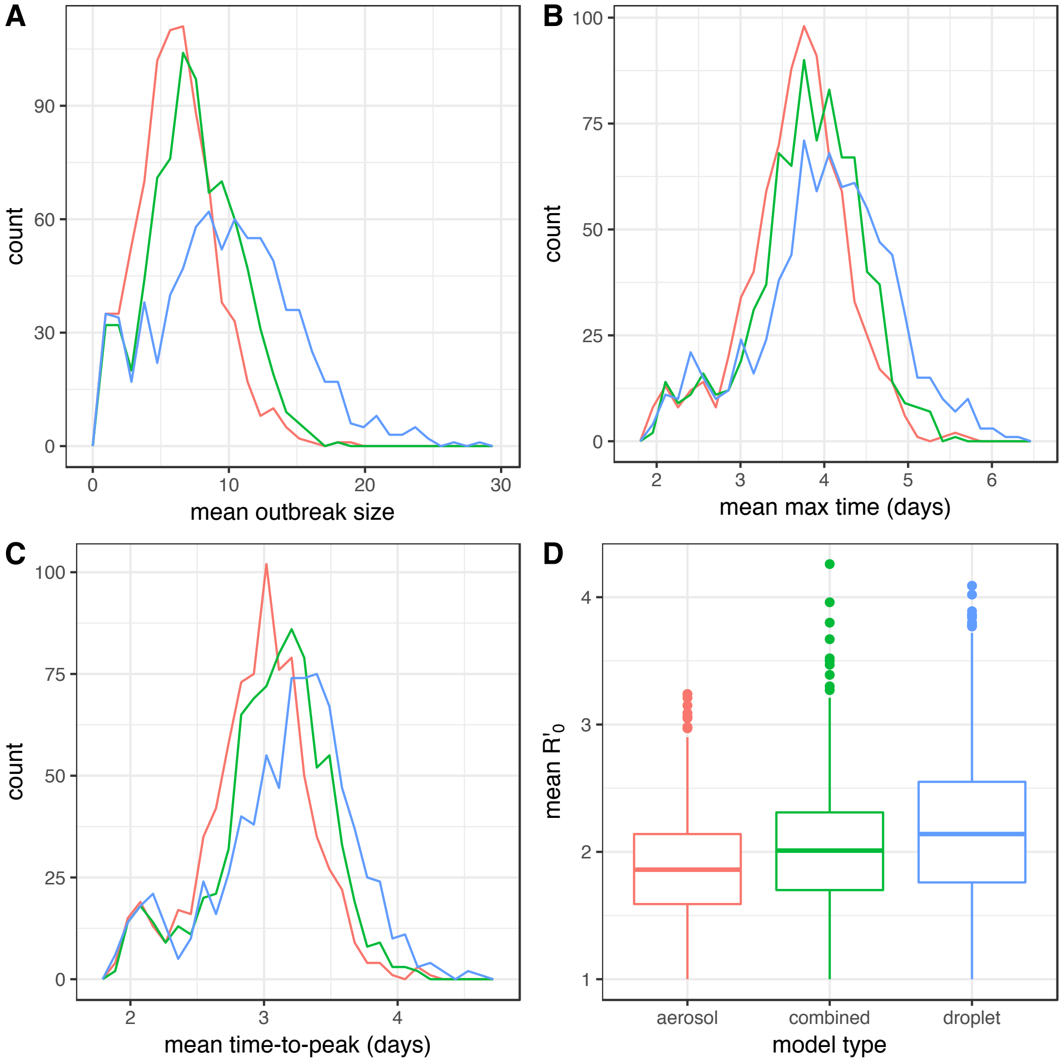


FIGURE 3.4: Frequency distributions of mean final size (A), mean duration of outbreak (B) mean time to reach the maximum prevalence (C), and mean R_0' (D), for the three different transmission models. Color-codes in panels A, B and C follow the labels on the x-axis in panel D (red for aerosol-based, green for combined, and blue for droplet-based transmission). For each transmission model, the values are based on 78,900 simulations where each individual served as index case in 100 independent simulation runs.

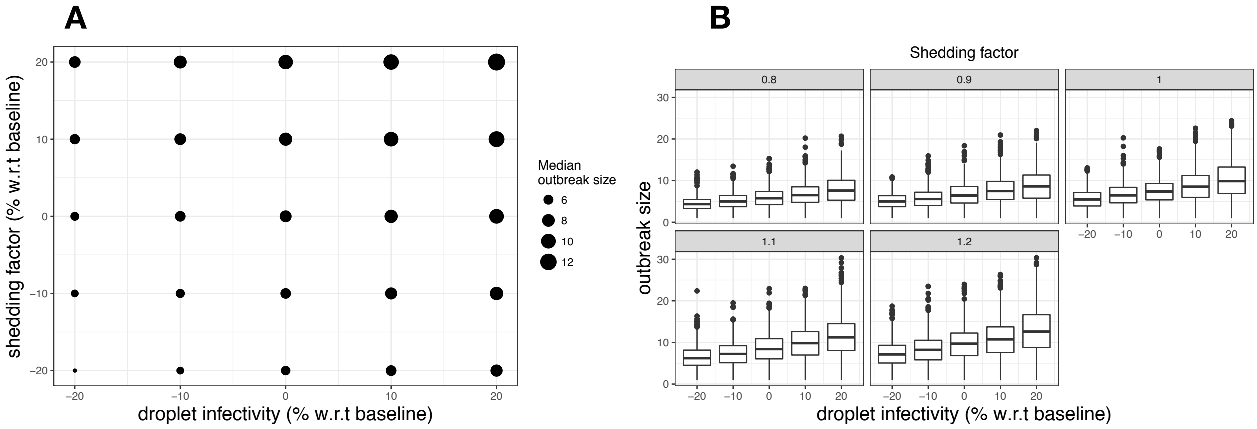


FIGURE 3.5: Sensitivity plots showing the effect of changes in relevant transmission parameters on outbreak size. In particular, both shedding rate and droplet infectivity were shifted from -20% up to +20% with respect to baseline values, as reported in the literature. (A) Scatter plot summarizing the effect of both parameters changes on median outbreak size (proportional to circles size, as shown in legend). (B) Boxplots plot presenting the details on the outbreak size distributions, depending on changes in droplet infectivity (x axis) and shedding rate (horizontal facets).

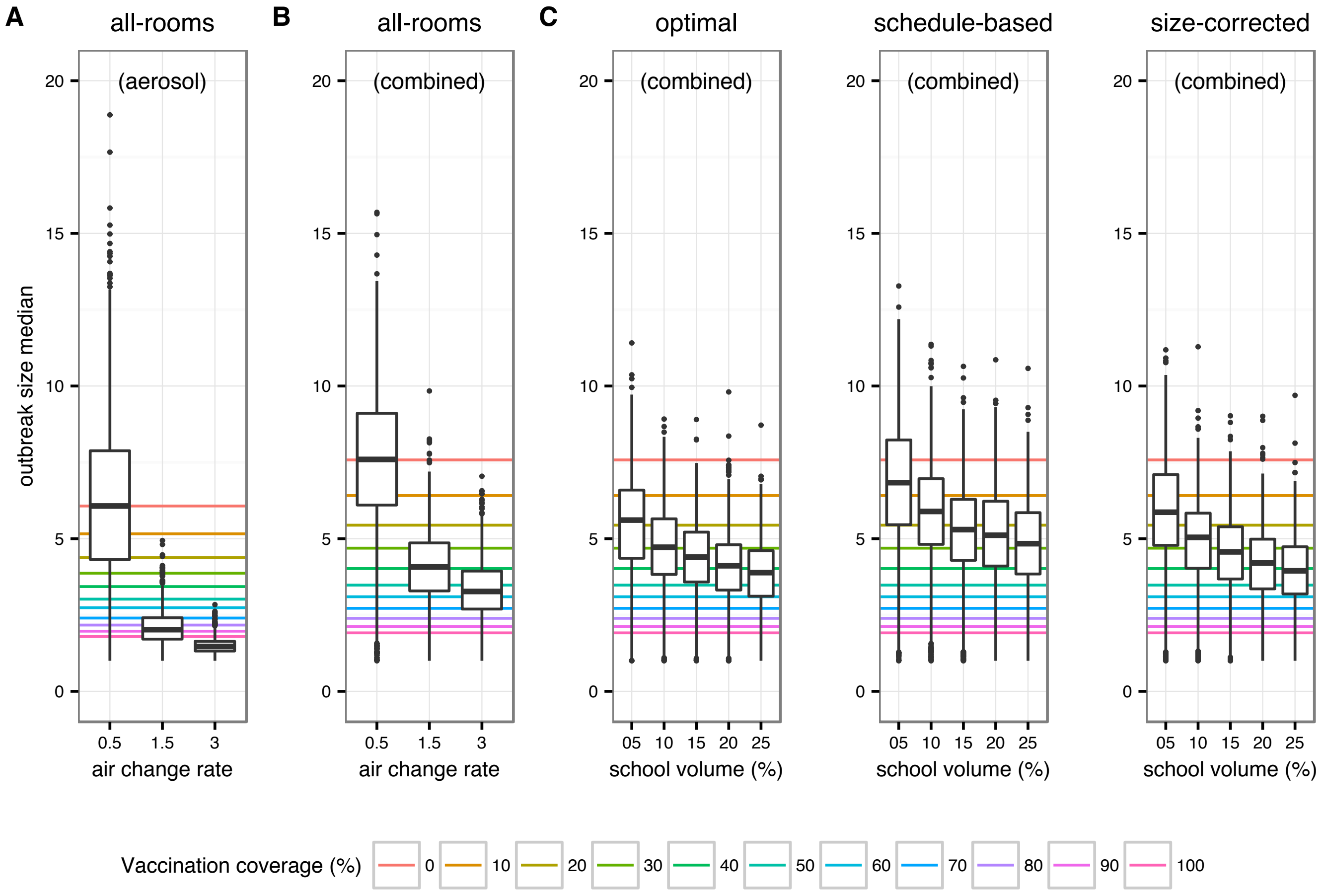


FIGURE 3.6: Comparison of the effect of ventilation (boxplots) and vaccination coverage (horizontal lines) on outbreak size. Colors refer to different vaccination coverages. Results are reported for aerosol-based (A) and combined transmission models (B, C). For (A) and (B), air change rate varies from 0.5 to 3.0 changes/h, while (C) assumes an air change rate of 0.5 changes/h. (C) Comparison of partially improved ventilation strategies (boxplots) in the school. Classrooms for improved ventilation were selected by ranking them according to the amount of inhaled infectious particles (optimal), their occupancy according to the school roster (schedule-based), or occupancy corrected by room size (size-corrected). Effect of vaccination coverage is also reported (median outbreak sizes under corresponding vaccination coverages, reported as horizontal lines).

Chapter 4

Breaking Apart Contact Networks with Vaccination

Gianrocco Lazzari[1], Marcel Salathé[1,*]

[1]Global Health Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[*]marcel.salathe@epfl.ch

Abstract

Infectious diseases can cause large disease outbreaks due to their transmission potential from one individual to the next. Vaccination is an effective way of cutting off possible chains of transmission, thereby mitigating the outbreak potential of a disease in a population. From a contact network perspective, vaccination effectively removes nodes from the network, thereby breaking apart the contact network into a much smaller network of susceptible individuals on which the disease can spread. Here, we look at the continuum of small world networks to random networks, and find that vaccination breaks apart networks in ways that can dramatically influence the maximum outbreak size. In particular, after the removal of a constant number of nodes (representing vaccination coverage), the more clustered small world networks more readily fall apart into many disjoint and small susceptible sub-networks, thus preventing large outbreaks, while more random networks remain largely connected even after node removal through vaccination. We further develop a model of social mixing that moves small world networks closer to the random regime, thereby facilitating larger disease outbreaks after vaccination. Our results show that even when vaccination is entirely random, social mixing can lead to contact network structures that strongly influence outbreak sizes. We find the largest effects to be in the regime of relatively high vaccination coverages of around 80%, where despite vaccination being random, outbreak sizes can vary by a factor of 20.

Introduction

The spread of infectious diseases remains a central public health issue in the 21st century. On the one hand, emerging or re-emerging diseases with no known vaccines pose a fundamental threat, and pandemics of such diseases remain on the list of potentially catastrophic events for humanity (*Ten threats to global health in 2019* 2019; *Bill Gates: deadly flu epidemic is one of biggest threats to humanity - Insider* 2018). On the other hand, even vaccine-preventable diseases continue to cause substantial morbidity and mortality, for two main reasons: vaccines are generally not perfectly protective (Ward et al., 2005; La Torre et al., 2007; Osterholm et al., 2012a), and a vaccination coverage of 100% is rarely achieved (*WHO | Data, statistics and graphics* 2019). Immunological issues such as limited vaccine efficacy, vaccine effectiveness, extent and duration of vaccine immunogenicity contribute to an imperfect protection, and remain under active investigation for improvements. Societal issues such as limited access to vaccines, as well as medical and personal reasons that prevent individuals from getting vaccinated contribute to an incomplete coverage (Hill et al., 2016).

Despite these issues, vaccination has substantially reduced the burden of many diseases in general, and childhood diseases in particular (Rappuoli et al., 2014; Peter, 1992). However, some vaccine preventable diseases have been making worrying comebacks in recent years. The case of measles is particularly concerning, for numerous reasons. First, measles is one of the most infectious agents known to humans, with a basic reproductive number R_0 anywhere between 12 and 18 (Guerra et al., 2017). Second, measles does not only cause substantial morbidity and mortality, but has recently also been shown to diminish previously acquired immune memory of other pathogens (Petrova et al., 2019; Mina et al., 2019). Third, the measles vaccine is one of the most efficacious and affordable vaccines, providing life-long immunity in 97% of people who have received two doses (Rosenthal and Clements, 1993; Demicheli et al., 2013). Because of these factors, the WHO and other health organizations recommend (*WHO position paper on measles vaccine* 2017) a vaccination coverage of 90-95% for two routine doses of measles-containing vaccines, and most WHO member states have committed to achieving these goals. However, by 2015, the global two-dose vaccine coverage was only 61%, with high variance between countries (*WHO position paper on measles vaccine* 2017). Even in high-income countries such as those in Europe, only a few countries have achieved the coverage goal. Concerningly, the number of countries who have achieved the target has declined recently, from 14 countries in 2007 to 4 countries in 2017 (*ECDC: Insufficient vaccination coverage in EU/EEA fuels continued measles circulation* 2019). In early 2019, the WHO declared vaccine hesitancy to be one of the top global health issues (*Ten threats to global health in 2019* 2019).

Interestingly, however, countries with similar vaccination coverages show markedly different patterns with respect to the number of measles cases experienced (fig. 4.1).

For example, Canada and Switzerland have almost identical vaccine coverages, but the yearly number of measles cases per capita differ by an order of magnitude. Similarly, Germany reports an almost identical coverage than the US, but has almost an order of magnitude more measles cases per capita than the US. For what reasons could similar vaccine coverages lead to large differences in relative outbreak sizes? Some hypotheses have been put forward. For example, even with similar vaccination coverage, the risk of large outbreaks can vary if unvaccinated individuals are clustered (Salathé and Bonhoeffer, 2008). If, for example, 10% of the population is not vaccinated, and those 10% live close to each other (geographically and socially), outbreaks will likely be larger than if those 10% are more randomly distributed in the population. In the former case, the protective effect of herd immunity is larger than in the later case of clusters of unvaccinated individuals. Such a clustering phenomenon has been argued to be a likely contributor to recent outbreaks (Salathé and Bonhoeffer, 2008). Another hypothesis is that the speed emerging outbreaks are being tackled can vary greatly from one country to the next. In the US, measles outbreaks are treated with extreme urgency and even relatively small outbreaks receive substantial media coverage, something that is not observed in other countries.

Here, we report on another phenomenon that can lead to substantially different outbreak sizes in populations with identical vaccination coverages. When large parts of a population gets vaccinated, the vast majority of possible chains of transmission is broken, thereby hampering the spread of a disease. As we will show below, the structure of the underlying contact network can greatly influence the magnitude of that effect on outbreak dynamics, and in particular on outbreak size. To do this, we will use a well-established contact network approach, where the nodes of the network represent individuals, and the edges between the nodes represent contacts along which a disease can spread. Vaccinating a node with a very effective vaccine can be thought of as removing that nodes and all its edges from the network, as no disease transmission can go through this node. When removing nodes in such a way, we are left with a much smaller and sparser network of unvaccinated nodes, on which the disease can spread. The structure of the original complete network will affect the structure of the remaining susceptible network. Indeed, with high vaccination coverages, the susceptible network will often fall apart into multiple disconnected subnetworks. This will substantial lower outbreak sizes, as the spread of the disease is confined to its network of origin, and outbreaks in a given subnetwork are limited to the size of the subnetwork. The maximal magnitude of this effect is shown to be dependent on the vaccination coverage, but given such a constant coverage, the outbreak size can differ by more than a factor of 20.

Results

The results reported are based on network simulations, where nodes can be in one of two states, vaccinated and unvaccinated. Using measles as our infectious disease

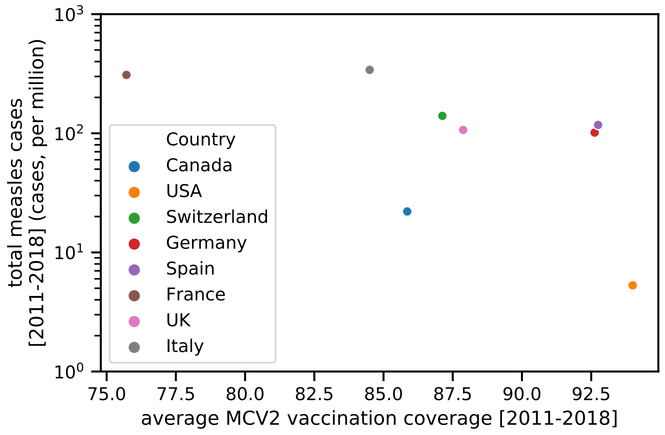


FIGURE 4.1: Scatter plot showing total measles cases (per million inhabitants) against average second-dose-vaccination coverage in the 2011-2018 time window, for North-American and the largest European countries. Data taken from (*WHO | Data, statistics and graphics* 2019; *WHO | Measles and Rubella Surveillance Data* 2019)

of interest, we make the simplifying assumptions that a vaccinated note is fully protected from getting infected, and that given a contact (edge) between an infected and a susceptible node, an infection is guaranteed to happen. We further ignore any timing issues with respect to incubation period and recovery times. While these assumptions do not reflect reality accurately, they are sufficiently representative to understand the worst case situation for measles, given that MCV2 status confers 97% protection, and the extremely contagious nature of the measles virus. These assumptions make stochastic disease simulations unnecessary, as the first infected node will go on to saturate the entire network of connected susceptible nodes with the disease. Thus, given multiple sub-networks (connected components) of susceptible nodes following vaccination in the complete network, and removal of all vaccinated nodes, the expected outbreak size $\bar{F}$ is given by $\bar{F} = \sum_i \frac{c_i}{\sum_i c_i} * c_i = \frac{1}{N_s} \sum_i c_i^2$, where c_i is the size of the i-th sub-network, and N_s the number of unvaccinated nodes in the networks (i.e. the sum of the size of all sub-networks). In other words, $\bar{F}$ is simply the weighted sum of the sub-networks' sizes.

In order to understand the effect of the structure of the original network on the expected outbreak size $\bar{F}$, we begin with a small-world network (Watts and Strogatz, 1998) of size $N = 1000$, and an initial rewiring probability of p. We then vaccinate a fraction V of the nodes, meaning that V represents the MCV2 vaccination coverage in our model. The vaccinated nodes are subsequently removed from the network, and the remaining $N_s = (1 - V)N$ susceptible nodes may then form multiple disconnected sub-networks, whose sizes determines the expected outbreak size $\bar{F}$ as indicated above. In order to understand the effect of the network structure on the expected outbreak size, we are calculating the outbreak size for different rewiring probabilities p, and different vaccination coverages V. Beginning from a rewiring probability $p = 0.001$, we explore increasing p values up to 0.8. Thus, starting from highly modular small-worlds network structures, we move increasingly towards random networks by increasing p, thereby lowering the modularity of the networks. Importantly, rewiring keeps the number of nodes and edges in the networks constant, making comparisons more meaningful. Figure 4.2 shows the effect of increasing rewiring on the size of the largest connected component of unvaccinated sub-networks (which dominates the expected outbreak size $\bar{F}$ given its calculation above). Overall, less modular networks are likelier to retain a large connected component after node removal than more modular networks.

Even though the difference in connectedness of the unvaccinated networks may appear visually subtle, as in Figure 4.2, its effect can nevertheless be quite consequential in terms of expected outbreak size $\bar{F}$. Figure 4.3a shows the effect of increasing rewiring on the expected outbreak size, for vaccination coverages $V = 0.5$, 0.6, 0.7, 0.8, and 0.9. While rewiring has initially little effect, we start to see noticeable effects at around p =0.01, initially for lower vaccination rates only, and later for higher vaccination rates as well. For each of the vaccination coverages, we can observe a transition from outbreak sizes that are far below the maximum possible outbreak sizes (as indicated by the horizontal lines in Figure 4.3a), approaching the maximum value with increasing rewiring. This transition spans at least an order of magnitude under all vaccination coverages, highlighting the magnitude of the effect. Overall, this demonstrates that rewiring changes the original network structure in such a way that the breaking apart of the network through vaccination-driven removal of nodes strongly influences the expected outbreak size.

We next explore a social model that may drive the rewiring process. Social contacts may change over time for a number of reasons, and while previous infectious disease models with vaccination have focused on social dynamics due to vaccination opinions (Mbah et al., 2012), we focus here on social dynamics that are entirely independent of vaccination. To begin, we assign a random social status s between 0 and 1 to each node, and then rewire edges assortatively, i.e. in such a way as to implement a similarity-seeking behavior of the nodes (see Methods for detail). We measure the strength of the similarity-seeking rewiring with τ, which captures the threshold of dissimilarity, above which nodes seek to change their contacts to more similar nodes (with respect to social status s). Once the network reaches a stable equilibrium, nodes are vaccinated at random, given vaccination coverage V. Thus, the social dynamics in this model are independent of vaccination, and vaccination is completely random. Figure 4.3b shows the effect of the dissimilarity threshold τ on the expected outbreak size with varying vaccination coverages. We observe that the dynamics are similar to the ones described in Figure 4.3a. We further quantify the difference τ can make, given a vaccination coverage V, by calculating the ratio between the expected outbreak size $\bar{F}$ at $\tau = 0.001$ (the minimal value), and the value of τ where $\bar{F}$ is maximal for the given vaccination coverage. Notably, at the minimal value $\tau = 0.001$, there are barely any rewirings, because the desire for similarity (or rather the dislike of dissimilarity) is so great that nodes cannot find suitable similar nodes. This value thus represents largely unmodified small-worlds network. Therefore, the calculated ratio quantifies the maximum strength of the effect of social dynamics. As can be seen in Figure 4.3c, this ratio can reach values of up to around 20, especially at vaccination coverage around $V \sim 0.8$. In other words, depending on the structure of the network due to social dynamics, outbreak sizes can differ by a factor of 20, even though the vaccination coverages are the same, and vaccination is at random. Importantly, these effect do not appear to be captured well by modularity - outbreak sizes can vary considerably in the range $\tau < 0.03$ even though the modularity of the networks is roughly the same (see Figure 4.4, right panel).

Finally, we explore the structural dynamics of social changes depending on the dissimilarity threshold τ. Low values of τ mean that nodes are generally seeking to connect to other, more similar nodes, but finding other nodes is challenging, given the very low dissimilarity threshold. Thus, the number of overall rewirings is low, as seen in Figure 4.4 (left panel). As τ is increasing, nodes are less likely to seek new connections, but when they do, they are more likely to find them due to the higher dissimilarity threshold. Thus, increasing τ leads to more rewirings. At a certain level of τ, the dynamics reverses, and rewirings become more rare: with increasing dissimilarity thresholds, nodes have little desire to seek out new connections. These overall dynamics of rewiring have a direct impact on the assortativity with respect to the social status s, and on the modularity of the network. As the rewirings are increasing, assortativity is increasing (as nodes are seeking, and finding, more similar nodes to connect to), and modularity is decreasing due to the random structural nature of the rewire (note that while the rewiring process itself is not random, but based on the value of s, the structural effect is nevertheless random in nature, because the values of s have initially been assigned randomly to nodes). This effect eventually weakens again, when τ becomes so high as to prevent most nodes from seeking to rewire in the first place.

Discussion

Vaccination is a powerful tool to curb the spread of infectious diseases in human contact networks because of its ability to break apart potential transmission chains. Given sufficiently high vaccination coverage, vaccination does not only break apart transmission chains, but has the potential to break apart a large contact network into many sub-networks, therefore substantially lowering the maximum possible size of an outbreak. We showed here that the original network structure influences the subnetwork structure in ways that can have very strong effects on expected outbreak sizes. In some cases, we observed a 20-fold difference of expected outbreak size, despite identical vaccination coverages.

We started from the observation that the number of measles cases per capita can differ substantially among countries even if they have very similar vaccination coverages. In particular, as can be seen in Figure 4.1, the number of measles cases per

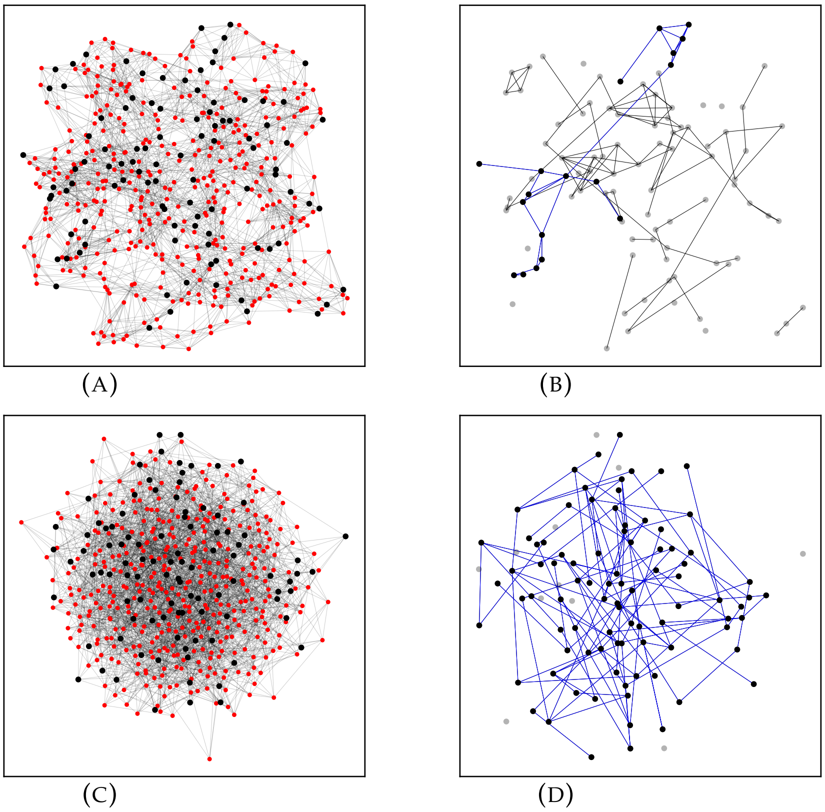


FIGURE 4.2: Graphs of equal size and similar structure break apart differently after random removal of 80% of all nodes. Two cases are shown, starting from the same Watts-Strogatz network ($N = 500$ and $k = 10$), but with different rewiring values: $p = 0.1$ in panel (a) with degree coefficient of variation $CV = 0.096$, and $p = 0.8$ in panel (c) with degree coefficient of variation $CV = 0.214$. In the right column, the largest connected component of the resulting graph after node removal (vaccination) is highlighted with blue edges. For top row, the expected outbreak size $\bar{F} = 8.72$, for bottom row $\bar{F} = 79.32$.

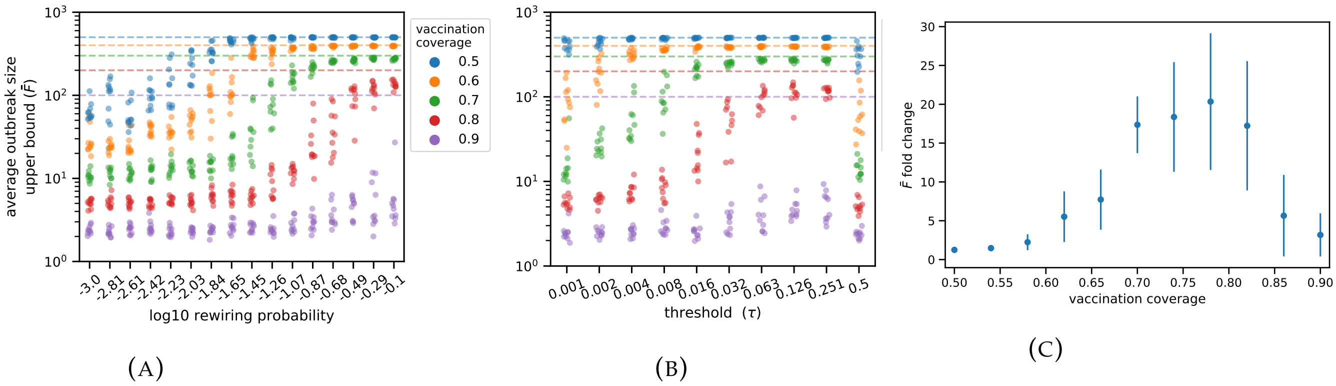


FIGURE 4.3: (a) Upper bound of outbreak sizes ($\bar{F}$), as a function of the rewiring probability p in the Watts-Strogatz (WS) model. (b) $\bar{F}$ measured after running the social dynamics algorithm, as a function of the social distance threshold τ (see `Methods`). Each dot corresponds to a single simulations run, with 10 runs for each value of τ and vaccination coverage. Dashed lines in panels (a) and (b) represent the fraction of unvaccinated nodes, i.e. the theoretical maximum outbreak size. (c) Fold change of $\bar{F}$, defined as the ratio between its highest value $\max_\tau(\langle\bar{F}\rangle)$ and its value at lowest threshold τ, $\langle\bar{F}\rangle(\tau = min(\tau))$, after taking mean over 10 simulation runs, for each value of τ. In all simulations WS graphs with $N = 1000$ and $k = 10$ were used; $p = 0.01$ in (b) and (c).

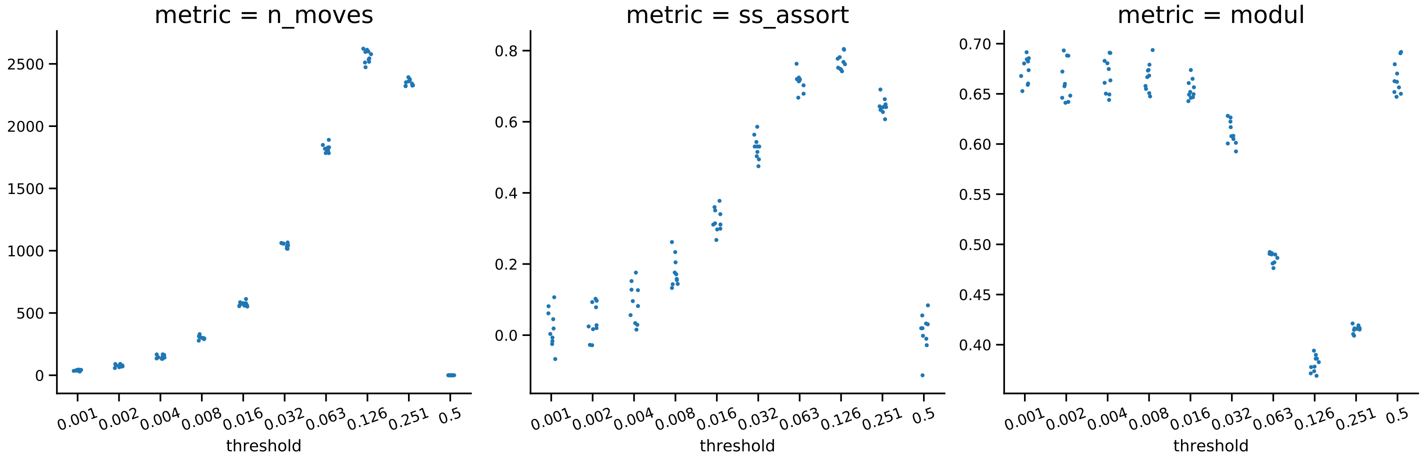


FIGURE 4.4: Effect of social dynamics algorithm on network properties, as a function of threshold τ (log scale). Number of actual edge rewirings performed (left panel), assortativity of nodes' 'social status' (central panel), and network modularity (right panel). As in figure 3, dots corresponds to a single simulations run (10 simulation runs for each value of τ).

capita on the North American continent are roughly an order of magnitude lower than in European countries with similar vaccination coverages. While there may be multiple reasons for this, we suggest that social dynamics influencing contact network structures, as shown here, can also play a role. For example, some evidence points to higher social segregation in the US compared to Europe (DiPrete et al., 2011; Mossong et al., 2008; Hens et al., 2009). It is plausible that higher social segregation will manifest itself in higher network modularity. The lowering of network modularity is the main topological reason why vaccination would fail to break apart the network into many disjoint subnetworks. In other words, in a network with weakly inter-connected communities (and higher modularity), the same level of vaccination will more likely disconnect communities from each other than in a network with strongly inter-connected communities (and lower modularity), thereby reducing the expected outbreak size in the former.

Previous models have associated social clustering with higher probabilities of vaccine-preventable disease outbreaks (Salathé and Bonhoeffer, 2008). Such models are generally based on the assumption of a vaccine decision-making process (Mbah et al., 2012), whereby vaccination is clustered in the network due to individuals' beliefs about vaccination, or other personal views that are correlated with vaccine decision-making. In contrast, our model strictly assumes a random distribution of vaccination, and thus describes a different phenomenon. Given that both effects are likely to be in play in reality, it will be interesting to see how these two phenomena interact in future work.

Methods

We generated and manipulated the networks using the `networkx` python library. In particular, we used the `community.greedy_modularity_communities` (Clauset, Newman, and Moore, 2004) and `community.modularity` functions to compute the graphs' modularity (Newman, 2006). The Watts-Strogatz networks used for the social dynamics model (fig. 4.4, 4.3b and 4.3c) were generated with rewiring probability $p = 0.01$ and number of initial nearest neighbors $k = 10$. Each node was assigned a random variable (s, its 'social status'), uniformly distributed in $[0, 1]$. Then we introduced a circular distance in the social-status space between node n_1 and node n_2, defined as: $d(s_1, s_2) \equiv min(|s_1 - s_2|, 1 - |s_1 - s_2|)$. The distance takes therefore values in the range $[0, 1/2]$. We then run our social dynamics algorithm, summarized hereby:

For each edge in the graph (say $(n_1, n_2) \in E(G)$):

1. decide if the 'social connection' between the two nodes n_1 and n_2 is too weak, based on a global threshold τ: $d(s_1, s_2) \geq \tau$

2. if yes, pick at random one of the two nodes (n_{old}) linked by the edge (e.g. $n_{old} = n_1$)

3. pick at random another node of the graph (n_{new}), outside of the neighborhood of n_{old} ($n_{new} \notin N_G(n_{old})$)

4. if a new link is possible (i.e. $d(s_{new}, s_{old}) < \tau$), rewire the old edge to the new contact (remove (n_{old}, n_2) and add (n_{old}, n_{new}))

were τ is a free-parameter of the model. The algorithm was stopped after the actual rewiring slows dramatically; in our case we set the max number of iterations equal to 4 times the number of edges $4 * E = 20000$, much bigger than the highest number of moves actually observed (see fig. 4.4, left panel). Note the algorithm preserves the total number of edges, as well as the mean degree, while it does not necessary keep the graph connected.

Chapter 5

FoodRepo: An Open Food Repository of Barcoded Food Products

Gianrocco Lazzari[1], Yannis Jaquet[1], Djilani Kebaili[1], Laura Symul[1], Marcel Salathé[1,*]

[1]Global Health Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

*marcel.salathe@epfl.ch

Abstract

In the past decade, digital technologies have started to profoundly influence health systems. Digital self-tracking has facilitated more precise epidemiological studies, and in the field of nutritional epidemiology, mobile apps have the potential to alleviate a significant part of the journaling burden, for example by allowing users to record their food intake via a simple scan of packaged products' barcodes. Such studies thus rely on databases of available products, their barcodes, ingredients, and nutritional values, which are not yet openly available with sufficient geographical and product coverage. In this paper, we present FoodRepo (https://www.foodrepo.org), an open food repository of barcoded food items, whose database is programmatically accessible through an Application Programming Interface (API). With currently more than 24,000 items available on the Swiss market, our database represents a solid starting point for large-scale studies in the field of digital nutrition, with the aim to lead to a better understanding of the intricate connections between diets, health in general, and metabolic disorders in particular. We are outlining the workflow of growing and maintaining the database, and discuss future plans for a broader geographic expansion.

Keywords: Open data, digital health, nutrition, API, digital epidemiology

Introduction

Metabolic disorders, such as diabetes or obesity, have become a major public health concern, with increasingly large parts of the global population affected (*WHO | Diabetes*; *WHO | Obesity and overweight*). Nutritional epidemiologists hope to better understand the underlying causes, the potential treatments and prevention strategies by analyzing population and individual patterns through studies that generally rely on surveying dietary habits. Traditional food-intake survey methods are based on questionnaires filled by participants at a given frequency. The frequency of diet records is an important factor contributing to the accuracy of the study (Satija et al., 2015). Multiple-day diet records might provide good accuracy when not based on memory, but require strong motivation and time commitment by the participants. Approaches like multiple / single 24-h recalls – involving a specialized interviewer performing surveys in person or on the phone with the participants – require less engagement, but pose issues with missing data as they rely on short-term memory. Finally, so-called Food Frequency Questionnaires, where participants are asked to indicate the frequency of intake of certain foods over long periods of time (typically 1 year), demand minimal participants' commitment, therefore allowing for large cohort studies on long-term dietary habits. However, the likelihood of missing or incorrect data increases as they count on participants' long-term memory. Overall, self-reported dietary data present biases which limit their applications, especially when they heavily rely on participants' memory (Archer, Pavela, and Lavie, 2015). Such limitations, which should be properly addressed in further epidemiological studies, may be overcome with more advanced recording methodologies such as dietary biomarkers and digital technologies (Subar et al., 2015).

Recent technological advances, and in particular the emergence and almost complete market penetration of smartphones, have offered interesting surveying alternatives. In particular, mobile phones have been successfully deployed in several food-related studies (Sharp and Allman-Farinelli, 2014), for example using food photography (Chae et al., 2011; Kong and Tan, 2012; Lee et al., 2012; Dibiano, Gunturk, and Martin, 2013; Zhu et al., 2011; Zhu et al., 2010). Other research has also explored the possibility of recording dietary habits by asking participants to scan the barcodes of their consumed food (Siek et al., 2006; Eyles, Jiang, and Mhurchu, 2010). Although further investigations are required to assess self-reporting biases, these advances in nutritional research have triggered the release of mobile apps oriented mainly towards diabetes and weight-loss self-management (Pagoto et al., 2013; Dunford et al., 2014; Stephens, Allen, and Himmelfarb, 2011; Tsai et al., 2007; Azar et al., 2013), showing the willingness and interest of users to monitor their food intake if it provides potential health benefits.

The further expansion of self-monitoring for research and medical purposes relies on comprehensive and continuously updated food databases. A few databases of barcoded products already exist, for example *Open Food Facts* or the *USDA Food*

Composition Database. While they each have their strength, not all of them are openly accessible or, they often have a limited product coverage, and are often not regularly updated. For Switzerland, we did not find any database whose product coverage was sufficiently high, where the data was completely open, and easily accessible through an Application Programming Interface (API). The last point was particularly important to us, as APIs are necessary for third parties to dynamically use the data in their products and services. Our approach was therefore to build an openly accessible database of barcoded food products with sufficiently high coverage, accessible through a stable API. Rather than focusing on a wide geographic range, we focused on a small country (Switzerland) in order to obtain the necessary coverage. The focus on the Swiss market further benefits from the need to support multiple languages from the beginning, thus making the system readily expandable to other countries, which we are now planning to do.

Here, we present this system, which we call FoodRepo (https://www.foodrepo.org), an openly accessible database of barcoded food products, and we describe the data-acquisition framework, its quality control and maintenance. Here, the word repository is meant to be understood as a data repository, where the community can deposit an increasing number of datapoints on food products. The growing community around FoodRepo and the validation of new products make our database robust, scalable and self-sustainable in the long run. Currently, the FoodRepo database mostly holds products sold in Switzerland, from the main grocery stores in the country. Its international expansion is under development.

Any item in the database is accessible through the FoodRepo website (for an example of products contained in the FoodRepo database, please see fig. 5.1-a) or via our API, described in section Usage Notes. The CC-BY-4 license under which our database is released will allow its exploitation by different type of users, from academic researchers to commercial partners. For instance, a Swiss consumers association is using FoodRepo data in their NutriScan mobile app (*Application NutriScan*) to make the food package information more accessible, and to provide their users with an overall nutritional score.

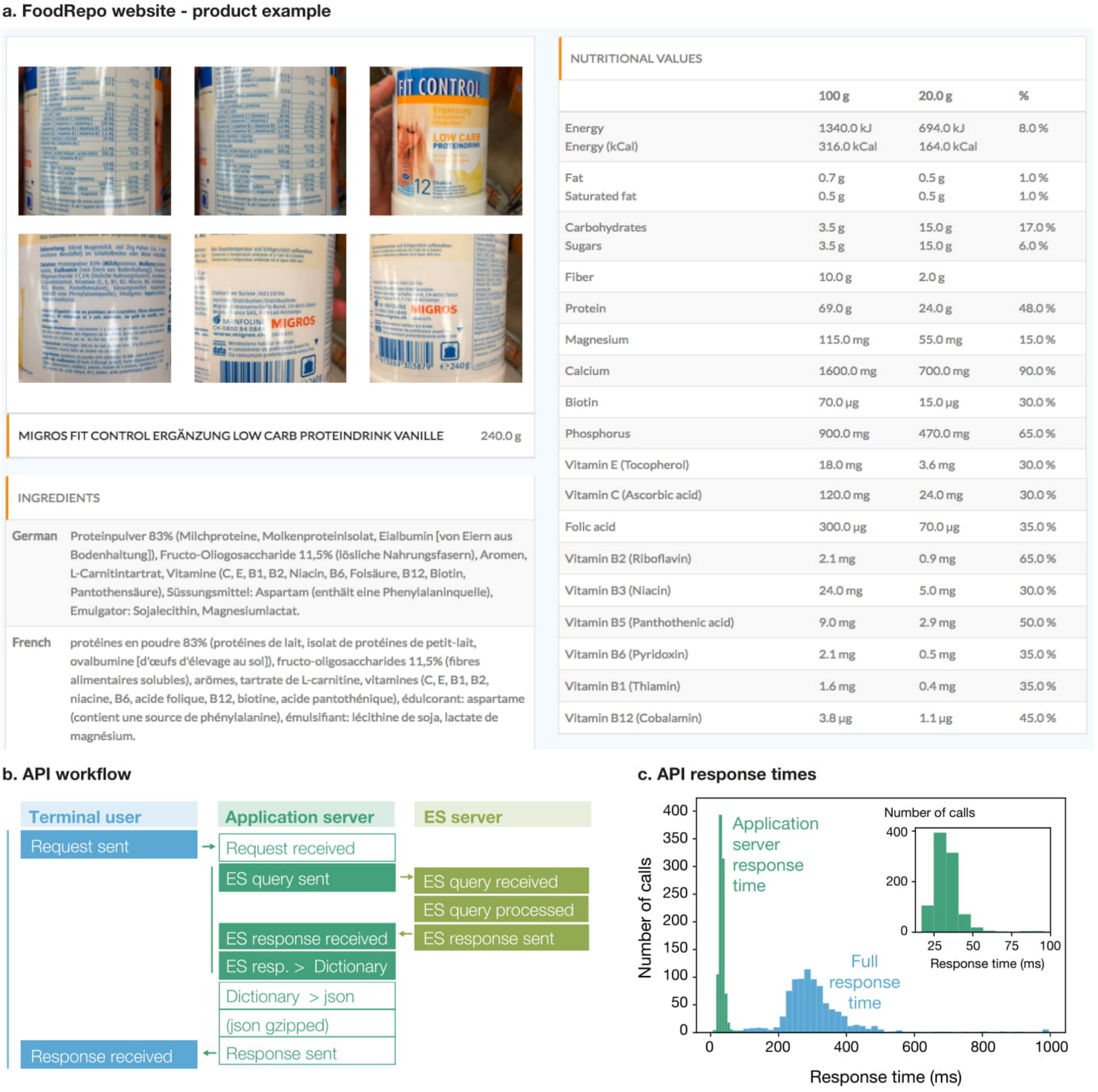
Beyond this specific example, the FoodRepo database opens the way for promising research opportunities in the field of digital epidemiology and personalized nutrition. Notably, we foresee that, through dietary live-tracking, this database can support studies which combine other recent technological developments and new findings in our understanding of the human metabolism. For example, phone-connected devices for continuous monitoring of blood glucose levels have recently been made available to diabetic patients (Pfeiffer, 1989; Aljasem et al., 2001), as well as numerous direct-to-consumer devices to estimate glucose levels have appeared on the market. A plethora of other wireless sensors are now also available to record various physiological parameters such as heart rate or blood pressure, marking a new era of 'high-throughput human phenotyping' (Elenko, Underwood, and Zohar, 2015). Studies that would simultaneously track participants' parameters, food intake, glycemic response and physical activity might provide detailed insights on the variability of individual metabolic responses. Interestingly, one of the factors which has recently been found to account for a large part of this variability is microbiota (Griffin et al., 2017; Turnbaugh et al., 2006; Le Chatelier et al., 2013; Zeevi et al., 2015; Pedersen et al., 2016). Large-scale testing of these hypotheses through self-tracking could contribute to the assessment of the complex metabolic response of the human body to different energy sources. This requires detailed records of food intake that includes nutritional information as well as eating times (Scheer et al., 2009) and food portion sizes (Ello-Martin, Ledikwe, and Rolls, 2005; Ledikwe, Ello-Martin, and Rolls, 2005; Young and Nestle, 2002), all challenges that FoodRepo may help to overcome.

However, we highlight an important limitation of all food databases. Generally, the curators of such repositories cannot ensure the validity of the data reported by the producers on the nutrition facts labels. It is indeed well known in the literature that there might be large discrepancies between the reported nutrients and the actual food content, due to different factors, such as food pre-processing or the different industry standards (Ng and Popkin, 2012; Ng and Dunford, 2013; Ahuja et al., 2009; Merchant and Dehghan, 2006; Phillips et al., 2006; Deharveng et al., 1999). Therefore, all studies using databases such as the one presented here would do well to assess the validity of such data and ideally quantify the reporting errors, especially when using the reported data on nutritional values.

Analyses of the database evolution will give interesting indication on the dietary trends and on the overall modification of the nutritive quality of packaged food. Although the database itself does not inform on the buying frequency, the continuous introduction of specific products in the market and thus in the database can potentially indicate how retailers react to customer demands and changing dietary habits.

Methods

The database building and maintenance process relies on the following steps: i) collection of product pictures from local retailers, ii) data extraction from the pictures, iii) validation of the extracted data, and iv) permanent storage in the database (Fig. 5.2). For the initial build of the database, we designed a specific pipeline (bootstrap workflow, Fig. 5.2-a, which allowed us to validate the first 20,000 food products in a few months. Given the dynamic nature of our data and the cost of the bootstrap workflow, we designed a second pipeline (currently under development) which relies on the growing FoodRepo community. This workflow (community-based, Fig. 5.2-b) allows us to keep up with the new and seasonal products introduced to the



a. FoodRepo website - product example

MIGROS FIT CONTROL ERGÄNZUNG LOW CARB PROTEINDRINK VANILLE 240.0 g

INGREDIENTS

German	Proteinpulver 83% (Milchproteine, Molkenproteinlsolat, Eialbumin [von Eiern aus Bodenhaltung]), Fructo-Oliogosaccharide 11,5% (lösliche Nahrungsfasern), Aromen, L-Carnitintartrat, Vitamine (C, E, B1, B2, Niacin, B6, Folsäure, B12, Biotin, Pantothensäure), Süssungsmittel: Aspartam (enthält eine Phenylalaninquelle), Emulgator: Sojalecithin, Magnesiumlactat.
French	protéines en poudre 83% (protéines de lait, isolat de protéines de petit-lait, ovalbumine [d'œufs d'élevage au sol]), fructo-oligosaccharides 11,5% (fibres alimentaires solubles), arömes, tartrate de L-carnitine, vitamines (C, E, B1, B2, niacine, B6, acide folique, B12, biotine, acide pantothénique), édulcorant: aspartame (contient une source de phénylalanine), émulsifiant: lécithine de soja, lactate de magnésium.

NUTRITIONAL VALUES

	100 g	20.0 g	%
Energy Energy (kCal)	1340.0 kJ 316.0 kCal	694.0 kJ 164.0 kCal	8.0 %
Fat Saturated fat	0.7 g 0.5 g	0.5 g 0.5 g	1.0 % 1.0 %
Carbohydrates Sugars	3.5 g 3.5 g	15.0 g 15.0 g	17.0 % 6.0 %
Fiber	10.0 g	2.0 g	
Protein	69.0 g	24.0 g	48.0 %
Magnesium	115.0 mg	55.0 mg	15.0 %
Calcium	1600.0 mg	700.0 mg	90.0 %
Biotin	70.0 µg	15.0 µg	30.0 %
Phosphorus	900.0 mg	470.0 mg	65.0 %
Vitamin E (Tocopherol)	18.0 mg	3.6 mg	30.0 %
Vitamin C (Ascorbic acid)	120.0 mg	24.0 mg	30.0 %
Folic acid	300.0 µg	70.0 µg	35.0 %
Vitamin B2 (Riboflavin)	2.1 mg	0.9 mg	65.0 %
Vitamin B3 (Niacin)	24.0 mg	5.0 mg	30.0 %
Vitamin B5 (Panthothenic acid)	9.0 mg	2.9 mg	50.0 %
Vitamin B6 (Pyridoxin)	2.1 mg	0.5 mg	35.0 %
Vitamin B1 (Thiamin)	1.6 mg	0.4 mg	35.0 %
Vitamin B12 (Cobalamin)	3.8 µg	1.1 µg	45.0 %

b. API workflow

c. API response times

FIGURE 5.1: (a) Screenshot from the webpage of a product on the FoodRepo website. (b) Schematic representation of the pipeline behind our API. When a user or an application (left column) sends a call to the API, the request is handled by the server that hosts the API (middle column). This sends then a query to the server which hosts the FoodRepo database (right column), where the query is handled by the Elastic Search engine. The data is returned to the API server which performs final formatting before giving it back to the user or the application. (c) Distribution of API response times, color-coded according to different sections of the back-end pipeline, as shown in panel b. In green (main plot and inset) the response-times of the Elastic Search server to the application server; in blue the full time needed for a user to have the data after a call to our API.

market by the retail shops, as well as to ensure the scalability and self-sustainability of FoodRepo in the long run.

The bootstrap workflow (Fig.5.2-a) consists of 3 main steps. The first step entailed a massive manual data collection from three large groceries stores in Switzerland upon approval from the shops (specifically Migros, Coop, and Lidl). We hired students to take pictures of all barcoded food items in retail shops located in the Lausanne area. To facilitate the data collection, we specifically designed a simple phone app with which students could scan the products' barcode and take pictures of the front and back of the package, the product's name, ingredients list, and nutrition facts. These pictures were then automatically uploaded to the database. At the end of this step, students had collected on average 4.4 pictures per item.
The second step focused on the extraction of information contained in the pictures. Due to the presence of multi-language ingredients and the often wrinkled surfaces of item packaging, Optical Character Recognition (OCR) systems could not achieve a reliable accuracy. We therefore opted for a crowd-sourced solution and in particular we decided to recruit workers on *Amazon Mechanical Turk* (AMT). AMT is a platform connecting *requesters* to *workers*, the latter being financially compensated to achieve tasks requiring human intelligence (HITs - Human Intelligence Tasks). Here, we designed a graphical user interface (GUI) allowing workers to transcribe the text they could read from product pictures. Specifically, the GUI presented text boxes where AMT workers provided the product name, nutritional values (in a table format) and ingredients, in every language present on the label (German and/or French for almost all items; Italian and/or English in addition for some products). Three different HITs were set up: one for nutrients, one for product name and one for ingredients. For the last two, we set up qualification rounds for AMT workers as their transcription involved some language skills. AMT workers could choose to either enter from scratch the information they saw on the pictures, or to approve / modify the suggestions given by an OCR system (*Text Recognition API Overview | Google Developers*). At the end of the second step, all annotated products were uploaded into the database, flagged as ready for validation.
The third step was thus dedicated to data validation, which was based on extensive manual checking by the FoodRepo team, and was additionally informed by manual reports from visitors to the FoodRepo website and with error-detection analyses of nutritional values. Such online reports are encouraged by the presence of a `'report an issue'` button on each product web-page, which prompts a visitor to file an issue when spotting a potential error. Details about the error-detection analyses are given in the Technical Validation section. Before the final validation of the data, the FoodRepo team as well as students manually checked all products thoroughly.

The community-based workflow (fig. 5.2-b) is similar to the bootstrap workflow, but instead of counting on AMT workers, it relies on the growing FoodRepo community. As new products become available in retail shops, FoodRepo users can submit them by uploading the corresponding package pictures, using the FoodRepo smartphone app. Currently, the information extraction is still performed by the FoodRepo team, but additional features are being implemented in the app, which will allow users to directly type the product details contained on the package. Before user-provided information is permanently stored in the FoodRepo database, consistent entries will need to be submitted by at least three different FoodRepo users. If such consensus will not be reached after seven independent submissions (i.e. there are still less than three consistent entries), the item will be manually analyzed by the FoodRepo team for definitive validation and inclusion into the database.

This procedure will ensure minimal intervention from our team, while still guaranteeing the reliability of the data. The FoodRepo team is currently fostering the development of an active community through which the continuity of FoodRepo is assured, and which will likely accelerate the birth of independent exploitations of the database, from both public and private partners.

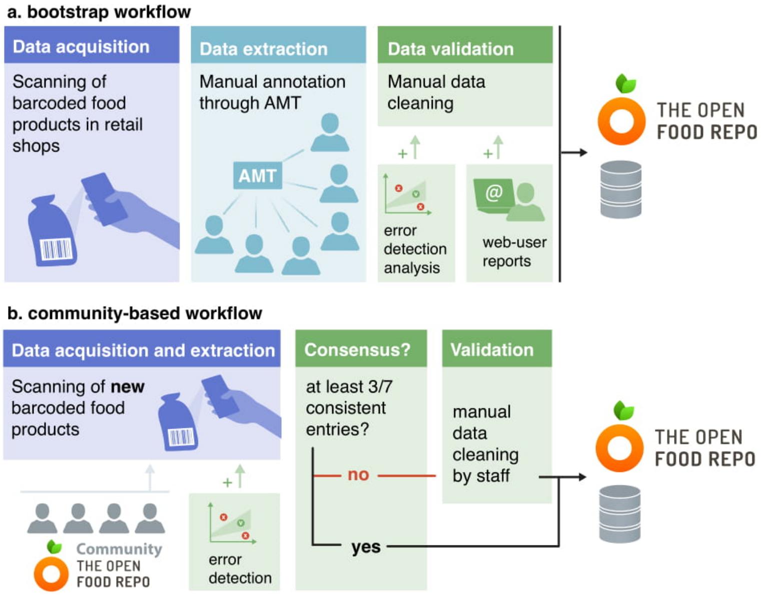

a. bootstrap workflow
Data acquisition
Scanning of barcoded food products in retail shops
Data extraction
Manual annotation through AMT
AMT
Data validation
Manual data cleaning
error detection analysis
web-user reports
THE OPEN FOOD REPO
b. community-based workflow
Data acquisition and extraction
Scanning of new barcoded food products
Community
THE OPEN FOOD REPO
error detection
Consensus?
at least 3/7 consistent entries?
no
yes
Validation
manual data cleaning by staff
THE OPEN FOOD REPO

FIGURE 5.2: Schematic overview of FoodRepo data collection and validation processes. The two workflows are illustrated here. The bootstrap workflow (a) was based on the joint work of the FoodRepo team and crowd-sourced workers collecting and validating the data. This allowed the storage of the first 14,000 or so products in the database. The community-based workflow (b) allows for long-term sustainability of the database thanks to customers uploading new products through FoodRepo mobile app and the continuous support of the FoodRepo team.

Data Records

All FoodRepo data are stored in a (*PostgreSQL: The world's most advanced open source database*) database, physically hosted on a server in Ireland. For a quick overview of the dataset, a database dump can be downloaded from the dedicated folder in our

API repository (*FoodRepo database dumps*). However, these dumps are not generated regularly, and we strongly encourage the use of the API which delivers up-to-date information. For each product, which comes with a unique numerical identifier, the database contains pictures of the item as found in the shop (usually between three to seven `.jpg` files), together with the main information presented on the package, i.e. the product name, nutritional values, ingredients list, barcode, and country of origin. The database holds as well the dates of the creation and last modification of the related item in the database (see Table 5.1). The programmatic access to the database is allowed by an API, described in the section Usage Notes

Technical Validation

As described in the Methods section, during the bootstrap stage (Fig. 5.2-a) the final validation was performed manually by the FoodRepo team, while in the community workflow (Fig. 5.2-b), the accuracy of the data is ensured by the consensus test (the FoodRepo team intervenes only if fewer than three matches are achieved after the uploads of the same product by seven different users). We highlight here that FoodRepo strictly reflects the information printed on products packages, even when suspicious values are present on the labels. All validation processes have thus been set-up to detect transcription errors.

Within this rationale, computational analyses were implemented for the detection of outliers, in particular regarding the nutritional values. These tests reflect basic constraints, such as the mass upper-limit:

$$p + f + c \leq 100 \tag{5.1}$$

where p, f, c are respectively the product's protein, fat and carbohydrates concentrations expressed in grams per 100 grams of product. From equation 5.1, one can also derive other linear inequalities for a single nutrient or couples of nutrients, namely $p + f \leq 100$, $p + c \leq 100$, and $c + f \leq 100$. These simple tests allowed us to detect transcription errors in earlier versions of the database, as illustrated by the outliers in fig. 5.3-a which shows the distribution of products in the fat-carbohydrates space with the joint mass boundary.

Similarly, other typos could be spotted by checking that the concentration of a sub-class of nutrient is smaller than the one of the parent-class. This is the case for instance of sugars VS carbohydrates, or saturated-fat VS fat, shown in fig. 5.3-b.

Another simple relation that helps check products' nutrition facts can be derived from the standard approximation of energy density based on nutrients composition (*COUNCIL DIRECTIVE of 24 September 1990 on nutrition labelling for foodstuffs*):

$$E \sim 4p + 9f + 4c, \tag{5.2}$$

where the product's energy content E is expressed in $kCal/100\,g$. Combining expressions 5.1 and 5.2 provides upper and lower boundaries for the energy content (for example fig. 5.3-c). In this case however, not all dots that fall outside the boundaries were due to typos in transcription. Indeed, the approximation in equation 5.2 does not take into account the different contribution to energy of complex carbohydrates such as polyols, which account for less than 4 $kCal/g$. This is why products such as candies and chewing gums would fall below the energy boundaries.

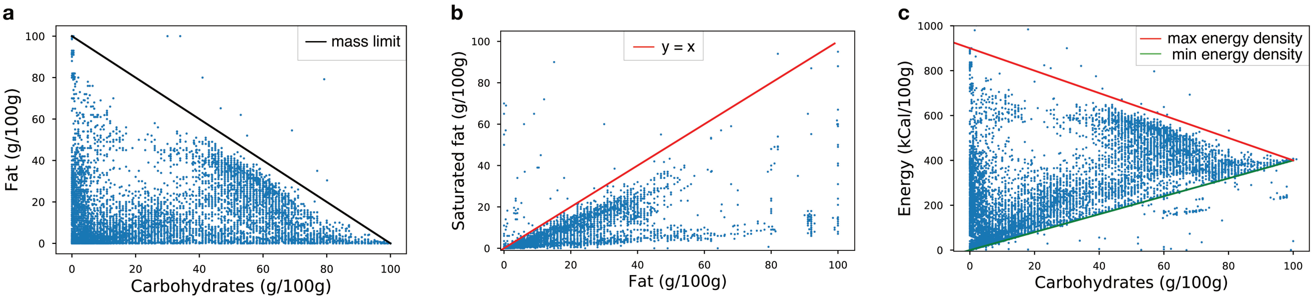


FIGURE 5.3: Examples of tests implemented with linear boundaries on nutritional values. Dots outside the boundaries have been inspected and corrected whenever data were different from the products packages.

Usage Notes

In order to facilitate the access to the database, we built an openly accessible application programming interface (API). Any terminal user, including third party apps or services, can send API requests to retrieve specific data. The API pipeline is illustrated in fig. 5.1-b. User's requests are handled on an application server, where an Elastic Search (ES) application handles the queries on another cloud computing service, based in Ireland. The ES response is then returned to the user after JSON formatting and compression (on demand). We checked that handling the request between the two servers does not critically compromise the total user-response time. We run series of single-page API calls, every 6 hours, over a week, in order to measure the full response-time and the application server response-time. We observed that the latter was consistently fast across all experiments (in the range of 20-50 ms) and that the bottleneck was rather the transmission between the terminal user and the application server (the average full response time was about 250 ms - see Fig. 5.1-c).

For a quick introduction to the API endpoints, users are welcome to try them out on the *OpenFood API Documentation*. Furthermore, on the project's GitHub repository, one can also find usage cases (*OpenFood API GitHub repository*) in Python, Ruby, Curl and JavaScript, as well as examples of complex queries which include fuzzy searches

(*Elasticsearch queries example*). When fetching a large amount of data, we suggest using the option of compressed data[1] and the possibility to include/exclude specific fields of each product (see for details the *OpenFood API Documentation*). In this way, one could reduce the response payload size by up to a factor of 10.
We remind readers that all contents (other than computer software) made available by FoodRepo on its websites, apps or services are licensed under the Creative Commons Attribution 4.0 International License. We however would like to highlight the fact that product images may contain copyrighted data such as brand logos.

Nomenclature

- API: Application Programming Interface - a set of tools and methods that allow to types of software to communicate. The FoodRepo API allows other applications to get and use the data.
- CC-BY-4: Creative-Commons public license, with the 'Attribution' term. It implies that anyone is free to share and transform the content of FoodRepo, even for commercial purposes, with the obligation to properly give credit to FoodRepo, and to display any modification without claiming direct endorsement from FoodRepo. For a detailed description, see the license text at https://creativecommons.org/licenses/by/4.0/
- OCR: Optical Character Recognition - tools that allow for automatic conversion of text contained in images to machine-readable formats.
- AMT: Amazon Mechanical Turk - web platform providing a marketplace, where workers perform tasks set up by requesters, usually in exchange of money.
- HIT: Human Intelligence Task - task related performed by workers in crowdsourcing platform, such as AMT.
- PostgreSQL: A popular and freely available relational database.
- JSON: a JavaScript-based file format commonly used for browser-server data exchange.
- Elastic Search: a very popular open-source search-engine.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

[1] This can be done by simply setting in the request header: `Accept-Encoding: gzip`

Fields	**Sample**
Product ID	972
Barcode	7611654884033
Name	Chocolat au lait aux noisettes
Quantity	150
Units	g
Portion Quantity	30
Portion Unit	g
Alcohol by Volume	0
Origin	Switzerland
Ingredients	(FR) sucre de canne brut* (Paraguay), cacao en pâte (Pérou), pâte de noisette 4.5% (Turquie), gousses de vanille*. Teneur en cacao du chocolat: 32% minimum. * Ingrédients conformes aux standards du commerce équitable Fairtrade. 58.6% du poids total. Dont sucre et produits à base de cacao avec bilan de masse. Tous les ingrédients agricoles sont issus de l'agriculture biologique (DE) Rohrohrzucker* (Paraguay), karamellisierte geröstete Haselnüsse 22% (Haselnüsse [Türkei], Rohrohrzucker [Paraguay], Wasser), Vollmilchpulver (Schweiz), Kakaobutter* (Dominikanische Republik), Kakaomasse* (Peru), Haselnusspaste 4.5% (Türkei), Vanilleschoten*, Kakaobestandteil in der Schokoladenmasse: mindestens 32%, * Nach Fairtrade-Standards gehandelte Zutaten. Gesamtanteil 58.6%. Davon Kakaoerzeugnisse und Zuckerarten mit Mengenausgleich. Alle landwirtschaftlichen Zutaten stammen aus biologischem Anbau. Allergie: Enthält Haselnuss, Milch. Kann spuren von Mandeln, Soja enthalten.
Nutrients (per 100g)	Energy 2410.0 kJ; Energy (kCal) 577.0 kCal; Fat 40 g; Saturated fat 16 g; Carbohydrates 43 g; Sugars 42 g; Fiber 4 g; Protein 10 g; Salt 0.2 g
Created at	2016-05-31, 17:54:07
Updated at	2017-11-16, 10:13:31
Pictures	Url to the front picture of the sample product: e.g. https://goo.gl/PyjjNa

TABLE 5.1: Sample product from the FoodRepo database with its values for the most relevant fields. While here we only provide the link to the front image of the product, an API call would provide the links to all pictures available for the requested products. A complete description of the fields provided by the API is available in the API documentation, on the project's GitHub repository.

Author Contributions

G.L. performed the descriptive and validation analysis of the dataset. Y.J. built the FoodRepo database, website, API and AMT HITs. J.D.K. maintained the API, coordinated the manual data validation and built the framework for the FoodRepo community. G.L., L.S. and M.S. wrote the manuscript. M.S. initiated and supervised the project.

Funding

This project was supported by a grant from the *Kristian Gerhard Jebsen Foundation*.

Acknowledgments

We are grateful to Migros, Coop, and Lidl for access to their retail shops.

Chapter 6

Conclusions and perspectives

The first contribution of our project on the 1630 Venice plague epidemic (Lazzari et al., 2019) surely lies in the dataset. Although similar datasets have been already recorded on other past outbreaks, the records contained in the Patriarchal Archive of Venice clearly excel for their space-time resolution, as well as for the details of deaths, considering the scientific knowledge and technology available at that time. Such dataset, first of all allowed us to discover the presence of a long tail of infections after the main outbreak (in late 1630), as never reported before in the literature (to the best of our knowledge). This forced us to look for models that could indeed reproduce such secondary outbreak, in 1631. We eventually proposed two approaches, not necessarily mutually exclusive. The first, based on a deterministic model, manage to well reproduce the entire epidemic curve, but requires a quick change in SIR model parameters (β, γ), with a resulting small increase in the basic reproduction number (R_0). The second one manages to show the secondary outbreak as a result of a delayed adaptation of transmission rate to the number of infected. While the former approach leaves the possibility to changes in both human behavior or biological factors in the disease dynamics, the latter focuses on hosts' behavioral changes and therefore does not require a change in the pathogen's transmission route. This would be the case for instance in a scenario in which bubonic plague evolved into pneumonic plague.

This project would definitely benefit in the future from further data collection. In particular detailed records of deaths from other parishes will help clarify whether or not another outbreak was co-occurring in the late 1630 in the city, and was still reported as plague (due for instance to the lack of medical knowledge of the public health officers, or/and the necessity to speed up sanitization procedures during the weeks of high disease incidence). Further resources will be required, as such data collection it is quite time consuming and requires the deployment of ancient language experts, that can read and transcribe information contained in the old parish books.

The work on influenza spread (Smieszek, Lazzari, and Salathé, 2019) leverages the increasing evidence of the importance of aerosol on disease transmission, in addition to the droplet-mediated one (Cowling et al., 2013). We were able to

establish a range for the relative impact of improved ventilation on epidemic control, with respect to standard immunization procedures. Furthermore, we performed sensitivity analysis on few parameters (shedding rate, droplet infectivity and air change rate – whose values were chosen anyways from the literature) in order to show that the model does not behave abruptly for small variation of such parameters.
An improved ventilation, closer to the recommended levels would surely have a positive impact on diseases control. However, the quantification of such benefits provided in our work still depends on few assumptions, in particular the relative importance of aerosol- versus droplet-mediated transmission, which we set to be equal. As in the year of the data collection (Salathé et al., 2010) no influenza outbreak was detected in that school, it was impossible to compare our simulated outbreaks with any real data. Apart from the possibility to repeat the study in the future waiting for an outbreak to happen, a more cost-effective way would be to run a retrospective validation. Namely, one would have to recover data from past outbreaks in similar settings such as high schools, where ventilation levels were recorded, in order to test the model hypothesis, assuming that the underlying contact networks are similar (down to the structural details relevant to disease transmission).

Our third work on infectious disease dynamics (Lazzari and Salathé, 2019) proposes a simple explanation of differences in the incidence of highly infectious diseases such as measles, across countries with similar vaccination coverages. Our model uses as main ingredient a coarse-grained representation of social features, implemented as one-dimensional nodes' attribute. Although it was not fitted to any specific outbreaks, the model shows that significant differences in outbreak sizes, up to 20-fold, are possible across different levels of segregation (for a given vaccination coverage). Furthermore, such effect it is more pronounced for vaccination coverages closed to $\sim$ 80%. Further work could try to first identify the main socio-economical factors leading to social distancing, and then embed them in a simple network model for epidemic spreading. An additional significant expansion of this work would include the combination of another important social phenomenon, relevant to diseases spread, namely the diffusion of anti-vax sentiment. It has been shown indeed that clustering of such opinions can lead to higher outbreak size (Salathé and Bonhoeffer, 2008). Therefore, it would be interesting to show how the effects of social segregation (unrelated to diseases spread) and clustering of vaccine hesitancy combine in real-world scenarios.

On a different research line, our FoodRepo database (Lazzari et al., 2018) aims to fill the need of digital technology in nutritional epidemiology, namely tools which can provide higher coverage and time-resolution in intake records, without an increase of the researchers' and participants' work-load. Furthermore, the digital approach has the potential to considerably reduce the error estimations for instance in calories and single-nutrients intake, which is quite hard to assess, especially with previous methodologies.

Developers form both public and academic sector have been using the FoodRepo API[1] in the last three years, showing the usefulness of an *open* database, which tries to keep up with the constantly evolving market. One of the challenge of such type of databases is indeed to keep including and transcribing the new food products constantly appearing in retail shops. The FoodRepo staff is trying to establish a community around the database, leveraging the use of our end-user mobile app, that allows the uploading of new products, as well as the transcription of the information printed on the package. Furthermore, next to quantity we want to ensure the correctness of the digitized data. For this, both algorithmic and manual checks are being set up, in order to account for transcription errors, as well as edge cases, such as typos or other sort of errors present actually on the food package (in which case FoodRepo choses to report the original package content).

As already mentioned, database like FoodRepo are meant to hopefully help collect food intake data with minimal burden from both participants and researchers, through the use for instance of mobile apps, that constantly fetch data from such repositories. In particular in our lab, the rest of the FoodRepo team has already developed *myFoodRepo*[2], a mobile app that allows indeed tracking of food intake, by pictures or barcodes. Tools like *myFoodRepo* will hopefully allow to quickly scale up cohort studies on nutritional epidemiology, such as the one ongoing in the lab, *Food&You*[3], devoted to assess the variability of the glycemic response across healthy subjects.

Overall, I am glad I had the possibility to touch different research areas, although the works on infectious diseases dynamics clearly played a larger role than the one in nutritional epidemiology. Thanks to the diversity of the projects I have been involved, I could work with different types of dataset, together with collaborators coming from diverse scientific communities and backgrounds. I believe that having such interdisciplinary exposure and portfolio is nowadays an important asset towards which, both academic and industrial research has received a considerable shift, in the last decade or so. As upcoming fields like personalized medicine will increasingly require such approach, I hope that this will be more and more reflected in the way also *under*graduate programs are designed, especially for those 'traditional' disciplines, like biology or physics, that I had the pleasure to cross in these last 6 years.

[1] Application Programming Interface

[2] https://www.myfoodrepo.org

[3] https://www.foodandyou.ch/en

Chapter 7

Curriculum vitae

Gianrocco Lazzari
gianrocco.lazzari@epfl.ch

Personal

Born on June 12, 1988. Italian Citizen. Swiss work permit B.

Summary

PhD student at EPFL in computational biology with broad interdisciplinary interests and computational skills. Main focus in infectious diseases modeling and nutritional epidemiology. Experience in: stochastic and statistical modeling; complex network science; data scraping, analysis and visualization applied to different data types, from nutritional databases to bio-luminescence single-cell video data.

Education

PhD student, prof. Marcel Salathé's lab, EPFL (2015 – 2020)
EPFL courses as student:
– Applied Data Analysis; dr. Michele Catasta (Stanford)
– Pattern Classification and Machine Learning; dr. Emtiyaz Khan (Riken)
– Responsible Conduct in Biomedical Research; dr. Hirosue Sachiko (former EPFL)
Trainee student, prof. Felix Naef's lab, EPFL (2014 – 2015)
Project topic: Transcription regulation in cell-cycle.
Signal extraction from single-cell bio-luminescence video data (ImageJ, Matlab) for cell genealogy reconstruction; unsupervised learning and statistical modeling of signal data (R).
PhD student, prof. E. van Nimwegen's lab, University of Basel (2013 – 2014)
Project topic: Stochastic models for genome evolution (Python)
Master of Science, Theoretical Physics, Utrecht University (2011 – 2013)
Main topics: Quantum field theory, Cosmology, Plasma Physics, Computational Physics, String theory, General Relativity
Bachelor of Science, Physics, University of Catania (Italy) (2007 – 2010)

Student at *Scuola Superiore di Catania* (2007 – 2010)
General topics in classical and quantum physics

Main PhD projects

- Modeling of influenza spread in public buildings, aiming to assess the relative effect of ventilation in outbreak reduction, w.r.t. vaccination strategies. The diffusion model was based on the contact network of an high-school, measured in a previous study [1].
- Analysis of FoodRepo data, our open database of swiss barcoded food products, for data quality assessment, data sharing and visualization [2].
- Modeling spread of ancient plague, based on historical data from the 1630 plague outbreak in Venice. The work includes the use of stochastic and deterministic modeling of diseases spread, as well unsupervised spatial analysis of time-series data [3].
- Implementation of a network stochastic model that can explain differences in outbreak sizes for highly infectious diseases, such as measles, in countries with similar vaccination coverages [4].

Other projects

- `https://hopsuisse.github.io` – project realized for the Applied Data Analysis course at EFPL. The project involved web data scraping, analysis and visualization (Python) of running events results in Switzerland.
- `https://www.foodrepo.org/ch/nutri-score-visualization-intro` – interactive visualization of food products of FoodRepo database, in the macronutrients space

Teaching

Teaching assistant for EPFL M.Sc./B.Sc. courses:

- Unsupervised and Reinforcement Learning in Neural Networks; fall 2016; dr. M.-O. Gewaltig
- Physical Biology of the Cell; spring 2016, fall 2017; prof. P. De Los Rios

Languages

Italian (mother tongue); English, French (fluent); Spanish, Portuguese (beginner)

Computer skills

Python / R (current daily use – modeling, data analysis and visualization, machine learning)
Interactive data visualization (e.g.: Plotly, Leaflet)
Matlab / ImageJ (previous daily use – image analysis, at EPFL)
Mathematica (previous daily use – master thesis work)

Awards

Winner of the "National Physics and Mathematics Certamen – Fabiana D'Arpa", 6th ed., 2007, L. da Vinci high school
Winner of the "Giovanni Raciti" prize, 2010 ed. – University of Catania (best B.Sc. career and thesis of the year)

Other activities

Current PhD students representative for the Computational Biology Doctoral School (EDCB) at EPFL. Member of doctoral school's working group on improvements of phd students' life conditions in satellite campuses.
Previously involved in different EPFL students associations, such as the ones for latin dance, mountaineering and sustainable development.

Publications[1]

1 Smieszek, T., Lazzari, G., & Salathé, M. (2019). Assessing the dynamics and control of droplet-and aerosol-transmitted influenza using an indoor positioning system. Scientific reports, 9(1), 2185.

2 Lazzari, G., Jaquet, Y., Kebaili, D. J., Symul, L., & Salathé, M. (2018). FoodRepo: An Open Food Repository of Barcoded Food Products. Frontiers in Nutrition, 5, 57.

3 Lazzari, G., Colavizza, G., Drago, D., Zugno, F., Bortoluzzi, F., Erboso, A., Kaplan, F., & Salathé, M., (2019) Death in Venice: Two-stage plague outbreak during the Second Pandemic (1630-31) (in preparation).

4 Lazzari, G. & Salathé, M., (2020) Breaking Apart Contact Networks with Vaccination (under review).

5 Lazzari, G., & Prokopec, T. (2013). Symmetry breaking in de Sitter: a stochastic effective theory approach. arXiv preprint arXiv:1304.0404.

[1] Google Scholar profile

Bibliography

Abbott, Alison (2001). “Digital history”. In: *Nature* 409, p. 556.

Abrate, Mario (1972). *Popolazione e peste del 1630 [ie un mille seiciento e trenta] a Carmagnola*. Vol. 1. Centro studi piemontesi.

Adams, WC (1993). “Measurement of breathing rate and volume in routinely performed daily activities, Final report”. In: *Human Performance Laboratory, Physical Education Department, University of California, Davis. Prepared for the California Air Resources Board, Contract* A033-205.

Ahuja, Jaspreet KC et al. (2009). “The impact of revising fats and oils data in the US Food and Nutrient Database for Dietary Studies”. In: *Journal of Food Composition and Analysis* 22, S63–S67.

Alfani, Guido and Samuel K Cohn Jr (2007). “Nonantola 1630. Anatomia di una pestilenza e meccanismi del contagio (con riflessioni a partire dalle epidemie milanesi della prima Età moderna)”. In: *Popolazione e storia* 8.2, pp. 99–138.

Alfani, Guido and Tommy E. Murphy (2017). “Plague and Lethal Epidemics in the Pre-Industrial World”. en. In: *The Journal of Economic History* 77.01, pp. 314–343. ISSN: 0022-0507, 1471-6372. DOI: 10.1017/S0022050717000092. URL: https://www.cambridge.org/core/product/identifier/S0022050717000092/type/journal_article (visited on 11/06/2017).

Aljasem, Layla I et al. (2001). “The impact of barriers and self-efficacy on self-care behaviors in type 2 diabetes”. In: *The Diabetes Educator* 27.3, pp. 393–404.

Amazon Mechanical Turk. https://www.mturk.com/.

American Society of Heating, Refrigerating and Air-Conditioning Engineers (2016). *ANSI/ASHRAE Standard 62.1-2016: Ventilation for Acceptable Indoor Air Quality*. American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Anderson, Roy M, B Anderson, and Robert M May (1992). *Infectious diseases of humans: dynamics and control*. Oxford university press.

Application NutriScan. https://www.bonasavoir.ch/nutriscan.

Archer, Edward, Gregory Pavela, and Carl J. Lavie (2015). “The Inadmissibility of What We Eat in America and NHANES Dietary Data in Nutrition and Obesity Research and the Scientific Formulation of National Dietary Guidelines”. In: *Mayo Clinic Proceedings* 90.7, pp. 911–926.

Atkinson, Michael P and Lawrence M Wein (2008). “Quantifying the routes of transmission for pandemic influenza”. In: *Bulletin of mathematical biology* 70.3, pp. 820–867.

Azar, Kristen MJ et al. (2013). “Mobile applications for weight management: theory-based content analysis”. In: *American journal of preventive medicine* 45.5, pp. 583–589.

Azimi, Parham and Brent Stephens (2013). “HVAC filtration for controlling infectious airborne disease transmission in indoor environments: Predicting risk reductions and operational costs”. In: *Building and Environment* 70, pp. 150–160.

Bamji, Alexandra (2016). “Medical Care in Early Modern Venice”. en. In: *Journal of Social History* 49.3, pp. 483–509. ISSN: 0022-4529, 1527-1897. DOI: 10.1093/jsh/shv060. URL: http://jsh.oxfordjournals.org/lookup/doi/10.1093/jsh/shv060 (visited on 06/05/2016).

Barclay, Victoria C et al. (2014). “Positive network assortativity of influenza vaccination at a high school: implications for outbreak risk and herd immunity”. In: *PloS one* 9.2, e87042.

Bell, D and J Atkinson (1999). “Centrality measures for disease transmission networks”. In: *Social Networks* 21, pp. 1–21.

Belongia, Edward A et al. (2016). “Variable influenza vaccine effectiveness by subtype: a systematic review and meta-analysis of test-negative design studies”. In: *The Lancet Infectious Diseases* 16.8, pp. 942–951.

Beltrami, Daniele (1954). *Storia della popolazione di Venezia dalla fine del secolo XVI alla caduta della Repubblica*.

Bill Gates: deadly flu epidemic is one of biggest threats to humanity - Insider (2018). https://www.insider.com/deadly-flu-epidemic-biggest-threat-bill-gates-2018-learnings-2018-12.

Bradley, Leslie (1977). “The most famous of all English plagues: a detailed analysis of the plague at Eyam 1665-6”. In: *Local Population Studies* Supplement 4, pp. 63–94.

Brankston, Gabrielle et al. (2007). “Transmission of influenza A in human beings”. In: *The Lancet infectious diseases* 7.4, pp. 257–265.

Bubonic plague confirmed in China (2019). https://www.washingtonpost.com/world/2019/11/18/bubonic-plague-inner-mongolia/.

Cassell, Gail H and John Mekalanos (2001). “Development of antimicrobial agents in the era of new and reemerging infectious diseases and increasing antibiotic resistance”. In: *Jama* 285.5, pp. 601–605.

Cattuto, Ciro et al. (2010). “Dynamics of person-to-person interactions from distributed RFID sensor networks”. In: *PloS one* 5.7, e11596.

Chae, Junghoon et al. (2011). “Volume estimation using food specific shape templates in mobile image-based dietary assessment”. In: *Proceedings of SPIE*. Vol. 7873. NIH Public Access, 78730K.

Chen, Szu-Chieh and Chung-Min Liao (2010). “Probabilistic indoor transmission modeling for influenza (sub) type viruses”. In: *Journal of Infection* 60.1, pp. 26–35.

Christakos, G., R. A. Olea, and H. L. Yu (2007). "Recent results on the spatiotemporal modelling and comparative analysis of Black Death and bubonic plague epidemics". In: *Public Health* 121.9, pp. 700–720. ISSN: 0033-3506. DOI: 10.1016/j.puhe.2006.12.011. URL: http://www.sciencedirect.com/science/article/pii/S0033350607000145 (visited on 12/17/2018).

Christley, R et al. (2005). "Infection in social networks: using network analysis to identify high-risk individuals". In: *American Journal of Epidemiology* 162, pp. 1024–1031.

Clauset, Aaron, Mark EJ Newman, and Cristopher Moore (2004). "Finding community structure in very large networks". In: *Physical review E* 70.6, p. 066111.

COUNCIL DIRECTIVE of 24 September 1990 on nutrition labelling for foodstuffs. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1990L0496:20081211:EN:PDF.

Cowling, Benjamin J et al. (2013). "Aerosol transmission is an important mode of influenza A virus spread". In: *Nature communications* 4.

Curtis, Daniel R. and Joris Roosen (2017). "The sex-selective impact of the Black Death and recurring plagues in the Southern Netherlands, 1349-1450". en. In: *American Journal of Physical Anthropology* 164.2, pp. 246–259. ISSN: 00029483. DOI: 10.1002/ajpa.23266. URL: http://doi.wiley.com/10.1002/ajpa.23266 (visited on 11/06/2017).

Daisey, Joan M, William J Angell, and Michael G Apte (2003). "Indoor air quality, ventilation and health symptoms in schools: an analysis of existing information". In: *Indoor Air* 13.1, pp. 53–64.

Dean, Katharine R. et al. (2018). "Human ectoparasites and the spread of plague in Europe during the Second Pandemic". en. In: *Proceedings of the National Academy of Sciences*, p. 201715640. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1715640115. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1715640115 (visited on 01/16/2018).

Deharveng, G et al. (1999). "Comparison of nutrients in the food composition tables available in the nine European countries participating in EPIC". In: *European Journal of Clinical Nutrition* 53.1, p. 60.

Demicheli, Vittorio et al. (2013). "Vaccines for measles, mumps and rubella in children". In: *Evidence-Based Child Health: A Cochrane Review Journal* 8.6, pp. 2076–2238.

DeWitte, Sharon N (2009). "The effect of sex on risk of mortality during the Black Death in London, AD 1349–1350". In: *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* 139.2, pp. 222–234.

DeWitte, Sharon N. (2010). "Age patterns of mortality during the Black Death in London, A.D. 1349–1350". en. In: *Journal of Archaeological Science* 37.12, pp. 3394–3400. ISSN: 03054403. DOI: 10.1016/j.jas.2010.08.006. URL: http://linkinghub.elsevier.com/retrieve/pii/S0305440310002803 (visited on 11/19/2017).

Dibiano, Robert, Bahadir K Gunturk, and Corby K Martin (2013). "Food image analysis for measuring food intake in free living conditions." In: *Medical Imaging: Image Processing*, 86693N.

DiPrete, Thomas A et al. (2011). "Segregation in social networks based on acquaintanceship and trust". In: *American journal of sociology* 116.4, pp. 1234–83.

Drancourt, Michel, Linda Houhamdi, and Didier Raoult (2006). "Yersinia pestis as a telluric, human ectoparasite-borne organism". In: *The Lancet infectious diseases* 6.4, pp. 234–241.

Dunford, Elizabeth et al. (2014). "FoodSwitch: a mobile phone app to enable consumers to make healthier food choices and crowdsourcing of national food composition data". In: *JMIR mHealth and uHealth* 2.3, e37.

ECDC: Insufficient vaccination coverage in EU/EEA fuels continued measles circulation (2019). https://www.ecdc.europa.eu/en/news-events/ecdc-insufficient-vaccination-coverage-eueea-fuels-continued-measles-circulation.

Elasticsearch queries example. https://github.com/salathegroup/foodrepo_api/blob/master/v3/code/meta/es_sample_queries_product.md.

Elenko, Eric, Lindsay Underwood, and Daphne Zohar (2015). "Defining digital medicine". In: *Nature biotechnology* 33.5, pp. 456–461.

Ell, Stephen R. (1989). "Three days in October of 1630: detailed examination of mortality during an early modern plague epidemic in Venice". In: *Review of Infectious Diseases* 11.1, pp. 128–139. URL: http://cid.oxfordjournals.org/content/11/1/128.short (visited on 04/06/2016).

Ello-Martin, Julia A, Jenny H Ledikwe, and Barbara J Rolls (2005). "The influence of food portion size and energy density on energy intake: implications for weight management". In: *The American journal of clinical nutrition* 82.1, 236S–241S.

Eyles, Helen, Yannan Jiang, and Cliona Ni Mhurchu (2010). "Use of household supermarket sales data to estimate nutrient intakes: a comparison with repeat 24-hour dietary recalls". In: *Journal of the American Dietetic Association* 110.1, pp. 106–110.

Fabian, Patricia et al. (2008). "Influenza virus in human exhaled breath: an observational study". In: *PloS one* 3.7, e2691.

Favero, Giovanni et al. (1991). "Le anime dei demografi. Fondi per la rilevazione della popolazione di Venezia nei secoli XVI e XVII". In: *Bollettino di demografia storica* 15, pp. 23–110.

Ferguson, Neil M et al. (2005). "Strategies for containing an emerging influenza pandemic in Southeast Asia". In: *Nature* 437.7056, pp. 209–214.

FoodRepo database dumps. https://github.com/salathegroup/foodrepo_api/tree/master/data.

Gage, Kenneth L and Michael Y Kosoy (2005a). "Natural history of plague: perspectives from more than a century of research". In: *Annu. Rev. Entomol.* 50, pp. 505–528.

Gage, Kenneth L. and Michael Y. Kosoy (2005b). “Natural History of the Plague: Perspectives from More than a Century of Research”. en. In: *Annual Review of Entomology* 50.1, pp. 505–528. ISSN: 0066-4170, 1545-4487. DOI: 10.1146/annurev.ento.50.071803.130337. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev.ento.50.071803.130337 (visited on 03/22/2016).

Gómez, José M. and Miguel Verdú (2017). “Network theory may explain the vulnerability of medieval human settlements to the Black Death pandemic”. en. In: *Scientific Reports* 7, p. 43467. ISSN: 2045-2322. DOI: 10.1038/srep43467. URL: https://www.nature.com/articles/srep43467 (visited on 12/17/2018).

Gordon M., Weiner (1970). “The Demographic Effects of the Venetian Plagues of 1575-77 and 1630-31”. In: *Genus* 26.1/2, pp. 41–57.

Gralton, Jan et al. (2011). “The role of particle size in aerosolised pathogen transmission: a review”. In: *Journal of Infection* 62.1, pp. 1–13.

Griffin, Nicholas W. et al. (2017). “Prior Dietary Practices and Connections to a Human Gut Microbial Metacommunity Alter Responses to Diet Interventions”. In: *Cell Host & Microbe* 21.1, pp. 84–96. ISSN: 19313128. DOI: 10.1016/j.chom.2016.12.006. URL: http://linkinghub.elsevier.com/retrieve/pii/S1931312816305170.

Guerra, Fiona M et al. (2017). “The basic reproduction number (R0) of measles: a systematic review”. In: *The Lancet Infectious Diseases* 17.12, e420–e428.

Hagberg, Aric A., Daniel A. Schult, and Pieter J. Swart (2008). “Exploring Network Structure, Dynamics, and Function using NetworkX”. In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, pp. 11 –15.

Hand, Eric (2011). “Culturomics: word play”. In: *Nature News* 474.7352, pp. 436–440.

Hens, Niel et al. (2009). “Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium”. In: *BMC infectious diseases* 9.1, p. 5.

Hill, Andrew B et al. (2016). “Improving global vaccine accessibility”. In: *Current opinion in biotechnology* 42, pp. 67–73.

Hufthammer, Anne Karin and Lars Walløe (2013). “Rats cannot have been intermediate hosts for Yersinia pestis during medieval plague epidemics in Northern Europe”. en. In: *Journal of Archaeological Science* 40.4, pp. 1752–1759. ISSN: 03054403. DOI: 10.1016/j.jas.2012.12.007. URL: http://linkinghub.elsevier.com/retrieve/pii/S0305440312005286 (visited on 12/17/2017).

Jefferson, Tom, Vittorio Demicheli, and Mark Pratt (1998). “Vaccines for preventing plague”. In: *Cochrane database of systematic reviews* 1.

Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2001). *SciPy: Open source scientific tools for Python*. https://www.scipy.org/.

Keeling, M. J. and C. A. Gilligan (2000a). “Metapopulation dynamics of bubonic plague”. In: *Nature* 407.6806, pp. 903–906. DOI: 10.1038/35038073.

Keeling, MJ and CA Gilligan (2000b). “Bubonic plague: a metapopulation model of a zoonosis”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 267.1458, pp. 2219–2230.

Kermack, William Ogilvy and Anderson G McKendrick (1927). “A contribution to the mathematical theory of epidemics”. In: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115.772, pp. 700–721.

Killingley, Ben and Jonathan Nguyen-Van-Tam (2013). “Routes of influenza transmission”. In: *Influenza and other respiratory viruses* 7.s2, pp. 42–51.

Kim, Jayoung et al. (2019). “Wearable biosensors for healthcare monitoring”. In: *Nature biotechnology*, p. 1.

Kong, Fanyu and Jindong Tan (2012). “DietCam: Automatic dietary assessment with mobile camera phones”. In: *Pervasive and Mobile Computing* 8.1, pp. 147–163.

Kristian Gerhard Jebsen Foundation. http://www.kgjf.org/.

La Torre, Giuseppe et al. (2007). “HPV vaccine efficacy in preventing persistent cervical HPV infection: a systematic review and meta-analysis”. In: *Vaccine* 25.50, pp. 8352–8358.

Lau, Max SY et al. (2015). “Inferring influenza dynamics and control in households”. In: *Proceedings of the National Academy of Sciences* 112.29, pp. 9094–9099.

Lazzari, Gianrocco and Marcel Salathé (2019). “Effects of assortative social mixing on infectious diseases dynamics”. In: *(under peer-review)*.

Lazzari, Gianrocco et al. (2018). “FoodRepo: An Open Food Repository of Barcoded Food Products”. In: *Frontiers in Nutrition* 5, p. 57.

Lazzari, Gianrocco et al. (2019). “Death in Venice: Two-stage plague outbreak during the Second Pandemic (1630-31)”. In: *(under peer-review)*.

Le Chatelier, Emmanuelle et al. (2013). “Richness of human gut microbiome correlates with metabolic markers”. In: *Nature* 500.7464, pp. 541–546.

Ledikwe, Jenny H, Julia A Ello-Martin, and Barbara J Rolls (2005). “Portion sizes and the obesity epidemic”. In: *The Journal of nutrition* 135.4, pp. 905–909.

Lee, Christina D et al. (2012). “Comparison of known food weights with image-based portion-size automated estimation and adolescents’ self-reported portion size”. In: *Journal of diabetes science and technology* 6.2, pp. 428–434.

Li, Yiping et al. (2007). “Role of ventilation in airborne transmission of infectious agents in the built environment–a multidisciplinary systematic review”. In: *Indoor air* 17.1, pp. 2–18.

Lindsley, William G et al. (2016). “Viable influenza A virus in airborne particles expelled during coughs versus exhalations”. In: *Influenza and other respiratory viruses*.

Manfredini, Matteo, Sergio De Iasio, and Enzo Lucchetti (2002). “The plague of 1630 in the territory of Parma: Outbreak and effects of a crisis”. In: *International Journal of Anthropology* 17.1, pp. 41–57.

Mbah, Martial L Ndeffo et al. (2012). "The impact of imitation on vaccination behavior in social contact networks". In: *PLoS computational biology* 8.4.

Merchant, Anwar T and Mahshid Dehghan (2006). "Food composition database development for between country comparisons". In: *Nutrition Journal* 5.1, p. 2.

Mina, Michael J et al. (2019). "Measles virus infection diminishes preexisting antibodies that offer protection from other pathogens". In: *Science* 366.6465, pp. 599–606.

Monecke, Stefan, Hannelore Monecke, and Jochen Monecke (2009). "Modelling the black death. A historical case study and implications for the epidemiology of bubonic plague". In: *International Journal of Medical Microbiology* 299.8, pp. 582–593.

Mongolian couple die of bubonic plague after eating marmot (2019). https://www.theguardian.com/world/2019/may/06/mongolian-couple-die-of-bubonic-plague-after-eating-marmot.

Monthly measles and rubella monitoring report, March 2019. https://www.ecdc.europa.eu/en/publications-data/monthly-measles-and-rubella-monitoring-report-march-2019.

Moser, Michael R et al. (1979). "An outbreak of influenza aboard a commercial airliner". In: *American Journal of Epidemiology* 110.1, pp. 1–6.

Mossong, Joël et al. (2008). "Social contacts and mixing patterns relevant to the spread of infectious diseases". In: *PLoS medicine* 5.3.

Mubareka, Samira et al. (2009). "Transmission of influenza virus via aerosols and fomites in the guinea pig model". In: *Journal of Infectious Diseases* 199.6, pp. 858–865.

Mullen, NA et al. (2011). "Ultrafine particle concentrations and exposures in six elementary school classrooms in northern California". In: *Indoor Air* 21.1, pp. 77–87.

Newman, Mark EJ (2006). "Modularity and community structure in networks". In: *Proceedings of the national academy of sciences* 103.23, pp. 8577–8582.

Ng, Shu Wen and Barry M Popkin (2012). "Monitoring foods and nutrients sold and consumed in the United States: dynamics and challenges". In: *Journal of the Academy of Nutrition and Dietetics* 112.1, pp. 41–45.

Ng, SW and E Dunford (2013). "Complexities and opportunities in monitoring and evaluating US and global changes by the food industry". In: *obesity reviews* 14.S2, pp. 29–41.

Noti, John D et al. (2012). "Detection of infectious influenza virus in cough aerosols generated in a simulated patient examination room". In: *Clinical Infectious Diseases*, cis237.

Open Food Facts. https://world.openfoodfacts.org/.

OpenFood API Documentation. https://www.foodrepo.org/api-docs/swaggers/v3.

OpenFood API GitHub repository. https://github.com/salathegroup/foodrepo_api/tree/master/v3/code.

Osterholm, Michael T et al. (2012a). “Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis”. In: *The Lancet infectious diseases* 12.1, pp. 36–44.

— (2012b). “Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis”. In: *The Lancet Infectious Diseases* 12, pp. 36–44.

Pagoto, Sherry et al. (2013). “Evidence-based strategies in weight-loss mobile apps”. In: *American journal of preventive medicine* 45.5, pp. 576–582.

Pedersen, Helle Krogh et al. (2016). “Human gut microbes impact host serum metabolome and insulin sensitivity”. In: *Nature*.

Peter, Georges (1992). “Childhood immunizations”. In: *New England Journal of Medicine* 327.25, pp. 1794–1800.

Petrova, Velislava N et al. (2019). “Incomplete genetic reconstitution of B cell pools contributes to prolonged immunosuppression after measles”. In: *Science immunology* 4.41.

Pfeiffer, EF (1989). “The glucose sensor: the missing link in diabetes therapy.” In: *Hormone and metabolic research. Supplement series* 24, pp. 154–164.

Phillips, Katherine M et al. (2006). “Quality-control materials in the USDA national food and nutrient analysis program (NFNAP)”. In: *Analytical and Bioanalytical Chemistry* 384.6, pp. 1341–1355.

Plague outbreak situation reports | WHO | Regional Office for Africa (01/21/2020). https://www.afro.who.int/health-topics/plague/plague-outbreak-situation-reports.

Pormann, Peter E (2015). “Interdisciplinarity: Inside Manchester’s’ arts lab’”. In: *Nature* 525.7569, p. 318.

PostgreSQL: The world’s most advanced open source database. https://www.postgresql.org/.

Rappuoli, Rino et al. (2014). “Vaccines, new opportunities for a new society”. In: *Proceedings of the National Academy of Sciences* 111.34, pp. 12288–12293.

Rosenthal, Sol Roy and C John Clements (1993). “Two-dose measles vaccination schedules.” In: *Bulletin of the World Health Organization* 71.3-4, p. 421.

Rossetti, Giulio et al. (2018). “NDlib: a python library to model and analyze diffusion processes over complex networks”. In: *International Journal of Data Science and Analytics* 5.1, pp. 61–79. ISSN: 2364-4168. DOI: 10.1007/s41060-017-0086-6. URL: https://doi.org/10.1007/s41060-017-0086-6.

Rudnick, SN and DK Milton (2003). “Risk of indoor airborne infection transmission estimated from carbon dioxide concentration”. In: *Indoor air* 13.3, pp. 237–245.

Salathé, Marcel and Sebastian Bonhoeffer (2008). “The effect of opinion clustering on disease outbreaks”. In: *Journal of The Royal Society Interface* 5.29, pp. 1505–1508.

Salathé, Marcel and James H Jones (2010). “Dynamics and control of diseases in networks with community structure”. In: *PLoS computational biology* 6.4, e1000736.

Salathé, Marcel et al. (2010). "A high-resolution human contact network for infectious disease transmission". In: *Proceedings of the National Academy of Sciences* 107.51, pp. 22020–22025.

Satija, Ambika et al. (2015). "Understanding nutritional epidemiology and its role in policy". In: *Advances in Nutrition: An International Review Journal* 6.1, pp. 5–18.

Scheer, Frank AJL et al. (2009). "Adverse metabolic and cardiovascular consequences of circadian misalignment". In: *Proceedings of the National Academy of Sciences* 106.11, pp. 4453–4458.

Schelling, Thomas C (1971). "Dynamic models of segregation". In: *Journal of mathematical sociology* 1.2, pp. 143–186.

Schmid, Boris V. et al. (2015). "Climate-driven introduction of the Black Death and successive plague reintroductions into Europe". en. In: *Proceedings of the National Academy of Sciences* 112.10, pp. 3020–3025. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1412887112. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1412887112 (visited on 12/17/2017).

Schnettler, Sebastian (2009). "A structured overview of 50 years of small-world research". In: *Social networks* 31.3, pp. 165–178.

Schofield, Roger (1977). "An anatomy of an epidemic: Colyton November 1645 to November 1646". In: *Local Population Studies* Supplement 4, pp. 95–126.

— (2016). "The last visitation of the plague in Sweden: the case of Bräkne-H oby in 1710–11". In: *The Economic History Review* 69.2, pp. 600–626.

Scott, Susan and Christopher J Duncan (2001). *Biology of plagues: evidence from historical populations*. Cambridge University Press.

Sharp, Darren B and Margaret Allman-Farinelli (2014). "Feasibility and validity of mobile phones to assess dietary intake". In: *Nutrition* 30.11, pp. 1257–1266.

Shendell, Derek G et al. (2004). "Air concentrations of VOCs in portable and traditional classrooms: results of a pilot study in Los Angeles County". In: *Journal of Exposure Science and Environmental Epidemiology* 14.1, pp. 44–59.

Siek, Katie A et al. (2006). "When do we eat? An evaluation of food items input into an electronic food monitoring application". In: *Pervasive Health Conference and Workshops, 2006*. IEEE, pp. 1–10.

Smieszek, Timo (2009). "A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread". In: *Theoretical Biology & Medical Modelling* 6, p. 25.

Smieszek, Timo, Gianrocco Lazzari, and Marcel Salathé (2019). "Assessing the dynamics and control of droplet-and aerosol-transmitted influenza using an indoor positioning system". In: *Scientific reports* 9.1, p. 2185.

Smieszek, Timo and Marcel Salathé (2013). "A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks". In: *BMC Medicine* 11.1, p. 35.

Smieszek, Timo et al. (2014). “How should social mixing be measured: comparing web-based survey and sensor-based methods”. In: *BMC infectious diseases* 14.1, p. 136.

Stehlé, Juliette et al. (2011). “Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees”. In: *BMC Medicine* 9, p. 87.

Stehlé, Juliette et al. (2011). “High-resolution measurements of face-to-face contact patterns in a primary school”. In: *PloS one* 6.8.

Stephens, Janna, Jerilyn K Allen, and Cheryl R Dennison Himmelfarb (2011). ““Smart” coaching to promote physical activity, diet change, and cardiovascular health”. In: *The Journal of cardiovascular nursing* 26.4, p. 282.

Stilianakis, Nikolaos I and Yannis Drossinos (2010). “Dynamics of infectious disease transmission by inhalable respiratory droplets”. In: *Journal of the Royal Society Interface*, rsif20100026.

Subar, Amy F et al. (2015). “Addressing Current Criticism Regarding the Value of Self-Report Dietary Data, 2”. In: *The Journal of nutrition* 145.12, pp. 2639–2645.

Tellier, Raymond (2006). “Review of aerosol transmission of influenza A virus”. In: *Emerg Infect Dis* 12.11, pp. 1657–1662.

— (2009). “Aerosol transmission of influenza A virus: a review of new studies”. In: *Journal of the Royal Society Interface*, rsif20090302.

Ten threats to global health in 2019 (2019). https://www.who.int/news-room/feature-stories/ten-threats-to-global-health-in-2019.

Text Recognition API Overview | Google Developers. https://developers.google.com/vision/text-overview.

Tran, Thi-Nguyen-Ny et al. (2011). “High Throughput, Multiplexed Pathogen Detection Authenticates Plague Waves in Medieval Venice, Italy”. en. In: *PLoS ONE* 6.3. Ed. by Tara Smith, e16735. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0016735. URL: http://dx.plos.org/10.1371/journal.pone.0016735 (visited on 12/17/2017).

Tsai, Christopher C et al. (2007). “Usability and feasibility of PmEB: a mobile phone application for monitoring real time caloric balance”. In: *Mobile networks and applications* 12.2-3, pp. 173–184.

Turnbaugh, Peter J et al. (2006). “An obesity-associated gut microbiome with increased capacity for energy harvest.” In: *Nature* 444.7122, pp. 1027–31. ISSN: 1476-4687. DOI: 10.1038/nature05414. URL: http://www.ncbi.nlm.nih.gov/pubmed/17183312.

Ulvioni, Paolo (1989). *Il gran castigo di Dio. Carestia ed epidemie a Venezia e nella Terraferma 1628-1632*. Milano.

USDA Food Composition Database. https://ndb.nal.usda.gov/ndb/.

Voirin, Nicolas et al. (2015). “Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care hospital”. In: *Infection Control & Hospital Epidemiology* 36.3, pp. 254–260.

Ward, Joel I et al. (2005). "Efficacy of an acellular pertussis vaccine among adolescents and adults". In: *New England Journal of Medicine* 353.15, pp. 1555–1563.

Watts, Duncan J and Steven H Strogatz (1998). "Collective dynamics of 'small-world'networks". In: *Nature* 393.6684, p. 440.

Weber, Thomas P and Nikolaos I Stilianakis (2008). "Inactivation of influenza A viruses in the environment and modes of transmission: a critical review". In: *Journal of infection* 57.5, pp. 361–373.

Welford, Mark R. and Brian H. Bossak (2009). "Validation of Inverse Seasonal Peak Mortality in Medieval Plagues, Including the Black Death, in Comparison to Modern Yersinia pestis-Variant Diseases". In: *PLoS ONE* 4.12, e8401. DOI: 10.1371/journal.pone.0008401.

Wells, William Firth (1955). *Airborne contagion and air hygiene. An ecological study of droplet infections*. Cambridge, MA: Harvard University Press.

Whittles, Lilith K. and Xavier Didelot (2016). "Epidemiological analysis of the Eyam plague outbreak of 1665–1666". en. In: *Proceedings of the Royal Society B: Biological Sciences* 283.1830, p. 20160618. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2016.0618. URL: http://rspb.royalsocietypublishing.org/lookup/doi/10.1098/rspb.2016.0618 (visited on 01/20/2017).

WHO | Data, statistics and graphics (2019). https://www.who.int/immunization/monitoring_surveillance/data/en/.

WHO | Diabetes. http://www.who.int/mediacentre/factsheets/fs312/en/.

WHO | Measles and Rubella Surveillance Data (2019). https://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/active/measles_monthlydata/en/.

WHO | News | Plague (2020). https://www.who.int/news-room/fact-sheets/detail/plague.

WHO | Obesity and overweight. http://www.who.int/mediacentre/factsheets/fs311/en/.

WHO | Plague (2020). https://www.who.int/health-topics/plague.

WHO position paper on measles vaccine (2017). https://www.who.int/immunization/policy/position_papers/WHO_PP_measles_vaccine_summary_2017.pdf?ua=1.

Wilson-Aggarwal, Jared K et al. (2019). "High-resolution contact networks of free-ranging domestic dogs Canis familiaris and implications for transmission of infection". In: *PLoS neglected tropical diseases* 13.7, e0007565.

Wong, Bonnie CK et al. (2010). "Possible role of aerosol transmission in a hospital outbreak of influenza". In: *Clinical infectious diseases* 51.10, pp. 1176–1183.

Xiao, S et al. (2018). "Probable transmission routes of the influenza virus in a nosocomial outbreak". In: *Epidemiology & Infection*, pp. 1–9.

Yang, Wan and Linsey C Marr (2011). "Dynamics of airborne influenza A viruses indoors and dependence on humidity". In: *PLOS One* 6.6, e21481.

Young, Lisa R and Marion Nestle (2002). “The contribution of expanding portion sizes to the US obesity epidemic”. In: *American journal of public health* 92.2, pp. 246–249.

Yue, Ricci P. H., Harry F. Lee, and Connor Y. H. Wu (2017). “Trade routes and plague transmission in pre-industrial Europe”. en. In: *Scientific Reports* 7.1, p. 12973. ISSN: 2045-2322. DOI: 10.1038/s41598-017-13481-2. URL: http://www.nature.com/articles/s41598-017-13481-2 (visited on 01/31/2020).

Zeevi, David et al. (2015). “Personalized Nutrition by Prediction of Glycemic Responses”. English. In: *Cell* 163.5, pp. 1079–1094. ISSN: 00928674. DOI: 10.1016/j.cell.2015.11.001. URL: http://www.cell.com/article/S0092867415014816/fulltext.

Zhu, Fengqing et al. (2010). “The use of mobile devices in aiding dietary assessment and evaluation”. In: *IEEE journal of selected topics in signal processing* 4.4, pp. 756–766.

Zhu, Fengqing et al. (2011). “Multilevel segmentation for food classification in dietary assessment”. In: *Image and Signal Processing and Analysis (ISPA), 2011 7th International Symposium on*. IEEE, pp. 337–342.