COVID-19 IN IRISH WORKPLACES AND COMMUNITIES - MODELLING OUTBREAKS FROM INFECTION DATA

COVID19 outbreaks in workplace settings: understanding and preventing superspreading events - Work Package 1 Report

A Science Foundation Ireland COVID19 Rapid Research Scheme Funded Project

Rita Howe, Fintan Costello, Carolyn Ingram, Conor Buggy and Carla Perrotta

Covid-19 in Irish Workplaces and Communities - Modelling Outbreaks from Infection data

Rita Howe

Fintan Costello

Carolyn Ingram

Conor Buggy

Carla Perrotta

University College Dublin

A Science Foundation Ireland Covid-19 Rapid Research Scheme funded project

Authors: Rita Howe, Fintan Costello, Carolyn Ingram, Conor Buggy and Carla Perrotta

Publisher: UCD School of Public Health Physiotherapy and Sports Science (cc) 2022

Funding Body: Science Foundation Ireland (SFI)

ISBN: 978-1-910963-64-7

Recommended Citation: Howe, Rita; Costello, Fintan; Ingram, Carolyn; Buggy, Conor; and Perrotta, Carla (2022). COVID19 in Irish Workplaces and Communities - Modelling Outbreaks from Infection data (COVID19 outbreaks in workplace settings: understanding and preventing superspreading events - Work Package 1 Report). University College Dublin, School of Public Health, Physiotherapy and Sports Science, Dublin, Ireland.

Contact the Author(s): rita.howe@ucd.ie

Executive Summary

During 2020-2021, the government of Ireland in line with international recommendations imposed the closure of non-essential trades, services, and commerce. Food plant factories, meat processing plants among others were deemed essential and remained open. During that time, many workers were exposed to outbreaks in their workplaces. Some of the questions arising included if workers will adapt to new safety measures, if those measures could prevent and mitigate workplace outbreaks and , if an outbreak occur in a closed facility, if it will impact community transmission. The most vulnerable workplaces were typically front-line industries, with healthcare and food processing facilities among the hardest hit by Covid-19 infections.

To complete the core aims, statistical models were developed for WP1. These models could accurately predict the scale of an outbreak in a meat processing plant based on the infection transmission in the community in the weeks preceding the outbreak and account for patterns in infection spread in both Ireland and worldwide using a 'behavioural response' mechanism. In addition to this, vaccine effectiveness was calculated using a method that made use of surveillance data. This demonstrated the strength and limitations of surveillance data.

One clear aspect of behaviour in the COVID-19 pandemic has been people's focus on, and response to, reported or observed infection numbers in their community. WP1 developed a simple model of infectious disease spread in a pandemic situation where people's behaviour is influenced by the current risk of infection and where this behavioural response acts homeostatically to return infection risk to a certain preferred level. Analysis of worldwide COVID-19 data confirmed the model predictions at both an overall and an individual country level.

Building on the findings of the infectious disease spread model, the research team aimed to

investigate how individuals adapted their behaviours throughout the pandemic at an individual level, using the number of community cases and the number of contacts reported by cases to the contact-tracing program as a proxy for behavioural response. This work is ongoing at this time.

In addition to this, estimations on vaccine effectiveness were calculated using a method that made use of surveillance data. This demonstrated the strength and limitations of surveillance data.

There were significant challenges in completing WP1, primarily caused by a difficulty in accessing the required data, however, the primary aims and goals of the work package were achieved and a meaningful body of research was produced on disease spread in specific, controlled environments and among the general population.Our work will certainly inform future pandemics. The main messages are 1) that community transmission can predict the occurrence of outbreaks -suggesting that managers and Public Health officials should work together to reinforce surveillance during peaks of community transmission and 2) high risk settings -like meat factories- can reduce or mitigate outbreaks if they introduce timely protective measures.

Contents

List of Figures

List of Tables

Acknowledgements

This research was funded by the Science Foundation Ireland (SFI). Without their support, this research would not have been possible.

We extend our thanks to our research advisory panel which provided invaluable guidance and knowledge: Professor Anne Drummond (UCD), Professor Francis Butler (UCD), Dr Penpatra Sripaiboonkij (UCD), Dr Elizabeth Alvarez (McMaster University) and Dr Claire Buckley (HSE).

The research team also gratefully acknowledges the support and advice of Professor Grace Mulcahy (UCD), Dr Nicola Walshe (UCD), and Ms Charlene Grice (UCD), the Department of Agriculture Veterinary Inspectors, Professor Patricia Kearney (HSE), the Covid Care Tracker team (HSE), Dr Paul Watts (Maynooth), and Dr Patricia Garvey (HPSC).

Glossary

Acronym	Full name
ARR	Absolute Risk Reduction
CCT	Covid Care Tracker
CI	Confidence Interval
CIDR	Computerised Infectious Disease Reporting
CMP	Contact Management Programme
CSO	Central Statistics Office
HSE	Health Service Executive
HSPC	Health Protection Surveillance Centre
DAFM	Department of Agriculture, Food and the Marine
ICU	Intensive Care Unit
LEA	Local Electoral Area
MPP	Meat-processing plant
NPHET	National Public Health Emergency Team
NNV	Numbers Needed to Vaccinate
OWID	Our World in Data
PCR	Polymerase chain reaction
PPE	Personal Protective Equipment
R/R_0	Basic reproductive number
RDP	Researcher Data Portal
RRR	Relative Risk Reduction
SEIR	Susceptible Exposed Infectious Recovered
VE	Vaccine Effectiveness
VI	Veterinary Inspector
WHO	World Health Organisation
WP	Work package

Part I.

Overview of Work Package 1

1. Project introduction

1.1. Background to the project

The advent of the SARS-CoV-2 pandemic has given rise to unique challenges worldwide in all aspects of life. These challenges were observed at both macro and micro levels of society with governments, organisations, religious and social groups, and individuals have made significant and sometimes difficult changes to prevent disease spread. For individuals, one of the most significant changes was the curtailing of social interaction with other people. In particular, the spread of the disease in workplaces has led to radical changes in the ways in which people work, how they work, and how they behave in their workplaces.

Initially, there was considerable uncertainty about the spread of the disease. In response to this, many workplaces/organisations implemented stringent personal protection measures but were still beset with Covid-19 outbreaks of significant size. With many businesses, institutions, and industries moving to remote working at their own or their governments' behest, the majority of these outbreaks occurred in sectors with essential workers who were unable to work from home. Some seemed particularly vulnerable to large and devastating outbreaks including nursing homes and other health facilities, and meat factories (Thompson et al., 2020; Hashan et al., 2021; Burton et al., 2020; Illingworth et al., 2021; Dyal, Grant, Broadwater, & et al., 2020; Waltenburg et al., 2020; Pokora et al., 2021; Mallet et al., 2021;

Herstein et al., 2021; Di Leone et al., 2020; Günther et al., 2020).

Typically, the spread and expected size of an epidemic (or pandemic) can be predicted by the basic (or initial) reproductive number (R_0). This value indicates the expected number of secondary infections produced by an infectious individual and predict the expected magnitude of the pandemic in absence of any interventions. Where $R_0 > 1$, it signals that the rate of disease is increasing in the community. Early research gave an unclear picture on the rate at which the disease was spread from person-to-person, with meta-analysis of studies suggesting that the R_0 of SARS-CoV-2 was approximately 3.38 ± 1.40 (Alimohamadi, Taghdir, & Sepandi, 2020). Later studies found an R_0 of approximately 2.2 for Western Europe, and 2.69 worldwide (Ahammed et al., 2021; Locatelli, Trächsel, & Rousson, 2021). Considerable variation in this value was observed between countries (Shaw & Kennedy, 2021). With this difficulty in predicting and controlling the spread of Covid-19 into vulnerable environments, it is important to consider how the transmission occurs into these environments initially, the role of the community infections in this, and the conditions under which the disease enters the workplace environment.

1.2. Role and Objectives

The primary aim of this package was to understand Covid-19 outbreaks (and super-spreading events) in workplace settings, workers behaviour adaptation in those settings, with a view to anticipating and preventing future outbreak events including super-spreading events. The secondary aim of this package was to understand how contact patterns influence and modulate the community transmission and spread of SARS-CoV-2 among a given population.

The role of work package 1 was to develop and test statistical models that could respond to the question if outbreaks in workplaces were responsible for increasing number of community transmission. To this end, a number of steps were needed. Firstly, data on pre-existing outbreaks in workplaces was required. Meat processing facilities were chosen as the target workplaces to investigate as they are a relatively homogeneous environment, they have been subject to outbreaks in striking scale and number since the beginning of the pandemic, and data on these outbreaks were less sensitive that those on, for instance, nursing home outbreaks. Once this information and information on community rates of infection was obtained, then the outbreaks inside MPPs could be modelled.

Secondly, WP1 aimed to develop and test a theory on the patterns of Covid-19 transmission throughout the pandemic. To do this, data on Covid-19 cases, including demographic information on the infected persons and vaccination status was required. Using this and the publicly available infection data, it would be possible to model contact patterns as a function of age, gender, vaccination, and health status.

Additionally, WP1 hoped to use the available surveillance data to gauge how effective Ireland's vaccination programme was at a given time. This is useful for a number of reasons; no study of vaccine effectiveness was available for Ireland at the time of this research, the results could be compared to empirical studies to examine whether surveillance data could approximate empirical results and finally, knowing if there was an observable difference in disease transmission for vaccinated and non-vaccinated individuals was useful to inform the conclusions for other areas of this work package.

1.3. Methodology

1.3.1. MPP model

For this model, the aim was to develop a statistical model of infectious disease spread in the workplace, by comparing community rates of infection to workplace rates. Following this, the model should be able to identify anomalous reported incidents such as uncontrolled spread of disease in the workplace (outbreak) and super-spreading events. To successfully model these events, information about the community prior to and during the outbreak and the workplace prior to and during the outbreak was required.

Community location (LEA/County) Using the community location at county or Local Electoral Area level, accurate Covid-19 rates could be determined in that area at the time of outbreak and the preceding weeks. This would be an essential model test of whether the outbreak is a function of the community Covid-19 rates or whether this was an exceptional outbreak event

Workplace demographics To model the workplace outbreak, it was necessary to obtain information about the workplace including number of employees and number of susceptible employees.

Outbreak information To model the scale of the outbreak information on the number of infections and type of testing (e.g., mass testing) done in the workplace was required.

Timeline To determine the course and spread of the outbreak, data was required on the timing of infection rise (or fall) prior to the outbreak event in the community, the timing of the infection contact events (for superspreading events), and the timing of the identified infections in the workplaces.

1.3.2. Behavioural Response model

This model described infectious disease spread in a pandemic situation where people's behaviour is influenced by the current risk of infection and where this behavioural response acts homeostatically to return infection risk to a certain preferred level. To model this, data on global infection numbers over time was obtained from OWID.

1.3.3. Contact model

The aim of the contact model was to develop a statistical model of community contact rate characteristics and change in relation to reported or observed infection numbers in their community. Perception of risk was expected to be a key predictor of people's contact behaviours during the pandemic. People would accept a certain number of contacts based on their perception of acceptable risk and this would change as infection rates change. People in high-risk situations such as those with severe illnesses would have smaller tolerances for risk than those low-risk situations and this would be reflected in their behaviours as the infection rates change. To successfully model this, information about the community, demographic information about the population, information about infection rates and outcomes, information about infection timelines, and information about contacts was required.

Contact information Detailed contact information was required to model contact rates over the pandemic in relation to infection rates. As people's perception of risk changes, their contact rates should show a comparable change.

Demographic information A key way of identifying individual perceptions of risk is through demographic information. Individual risk factors were determined by age, health status, and likelihood of contact with infectious individuals.

Infection rates Infection rates and outcomes would affect individual perceptions of risk in

the community. Communities with higher rates of disease, hospitalisation, poor disease outcomes, long-term illness and higher rates of death would be perceived as higher risk than communities with good disease outcomes.

1.3.4. Vaccine effectiveness

Standard measures of vaccine effectiveness were applied to Irish surveillance data to calculate the vaccination programme's performance. These measures include relative risk reduction (RRR), absolute risk reduction (ARR) and numbers needed to vaccinate (NNV). To calculate these measures information on population size, numbers of vaccinated individuals in the population, numbers of vaccinated and non-vaccinated individuals admitted to hospital, admitted to ICU, and died due to Covid-19 was required.

1.4. Data

By necessity, the data required for WP 1 was data collected by various government bodies during the pandemic. For the MPP model, the data required was collected during Covid-19 outbreaks in meat plants and surveillance data on community rates of infection. The outbreak data was held by a number of bodies and collated in the CSO research data portal. Within it, outbreak data was held within the CIDR database. The information on community rates of infection was publicly available with the HSE COVID-in-Ireland database. Data was also obtained from DAFM veterinary inspectors who worked in MPPs during outbreaks.

For the contact model, the data required was HSE data collected on persons infected with Covid-19, including their contacts, vaccination status, age, and health status. Data on rates of infection in the population at the time of a given person's infection was also required. This data was held by various bodies and accessed through the CMP, where the contact data

is containing inside the CCT database. The data on infection rates in the population was publicly available and could be accessed through OWID or the HSE COVID database.

Data for the behavioural response model was accessed through OWID, while the data for the vaccine effectiveness analyses came from CSO surveillance data on outcomes of vaccinated and unvaccinated individuals.

1.5. Structure of Chapters

WP 1 investigated the patterns and rate of transmission of Covid-19 in both specific workplace settings and among the general public. In the process of this, new statistical models for the purposes of understanding the spread of this disease were developed and tested. The contributions of the research team and their collaborators are acknowledged in each section.

This report encompasses two main strands of investigation: how disease spreads in specific, controlled environments and how disease spreads among the general population. Both of these will be discussed in terms of statistical modelling but for clarity, they will be split into separate sections.

Chapter 1 provides a brief overview of the background, aims, and methodology of work package 1.In chapter 2, the available literature on MPP outbreaks is reviewed and a novel model is developed to account for these outbreaks, and how they relate to community infections. The model builds on SEIR-type models but includes an alternative method to account for community-to-workplace transmission. The model's predictions are tested with data collected from MPP outbreaks that occurred between March 2020 - January 2022 in

Irish MPPs. Chapter 3 and 4 provides the theoretical background to role of risk judgements play in the transmission of Covid-19 infections. We propose and develop a model for the spread of Covid-19 in terms of risk-based decision-making among the general population. This chapter discusses the progress made with this model. Chapter 5 provides some contexts for vaccine effectiveness in Ireland. The successes and failures of analysis using surveillance data as a proxy for empirical data is discussed. Finally, in chapters $6 - 8$, we discuss our attempts to acquire the necessary data, the limitations thereof, our recommendations for future data collection and access.

Part II.

Workplace outbreaks

2. Meat plant outbreaks model

Rita Howe
Charlene Grice
Fintan Costello
Nicola Walshe
Vicky Downey
Carla Perrotta
Grace Mulcahy

2.1. Introduction

Large-scale Covid-19 outbreaks in workplaces have been a feature since the beginning of the SARS-CoV-2 pandemic (Middleton, Reintjes, & Lopes, 2020). The food-processing industry was particularly vulnerable to mass outbreaks with large numbers of cases reported in these facilities worldwide (Dyal et al., 2020; Waltenburg et al., 2020; Pokora et al., 2021; Mallet et al., 2021; Herstein et al., 2021; Di Leone et al., 2020; Günther et al., 2020). These outbreaks have led to significant illness and deaths among the workers and have had severe impacts on the food industry with factory closures (Waltenburg et al., 2020; Karodia et al., 2020).

The unique environmental factors and operational practises in meat-processing facilities make Covid-19 infections easy to spread and difficult to control when the virus begins to circulate within the facility (Walshe et al., 2021). In particular, areas of these facilities where ventilation is sub-optimal (such as the animal processing areas) have been linked to the highest number of employee infections (Walshe et al., 2021; Pokora et al., 2021).

In the initial stages of the pandemic, considerable effort was made to document the environmental role of the Covid-19 spread within MPPs, however, less well documented is how the initial infections enter the plant. The outbreaks must originate, at some level, in the community and spread to the plants but the conditions under this happens haven't been studied in detail. It is difficult to trace the entry point of the initial infection as many investigations are conducted after the outbreak has taken place or during the late stages of the outbreak when COVID infections are widespread within the facility.

Initially, an outbreak in a plant may be hard to detect-, in US meat plants asymptomatic or pre-symptomatic cases accounted for $12\% - 14\%$ of total cases (Waltenburg et al., 2020). Investigations of outbreaks in Ireland found that less than 50% of the cases were symptomatic in some plants (Department of Health, 2020a, 2020b). Other workplaces were found to have asymptomatic rates of $19\% - 88\%$, while a meta-anaylsis suggests that asymptomatic cases have a prevalence of approximately 30% (Oran & Topol, 2020; Payne et al., 2020; Sah et al., 2021). Where there are high numbers of asymptomatic cases, individual symptomatic cases may seem like cases in isolation and outbreaks may be harder, and take longer, to detect when they are circulating in plants.

There is some evidence that the initial infection stemming from contact in the community. The largest documented outbreak in Germany was traced to a social event (Günther et al., 2020). In another case in South Dakota, the cleaning shift employees who were socially distanced, not in directed contact with the other shift teams, and routinely wore PPE had the same attack rates of Covid-19 as the other shift employees who were engaged in high-density, non-distances meat processing (Steinberg et al., 2020). This suggests that the infections, in

this case, arose from an outside source.

It is non-trivial to assess the role that community infections play in MPP outbreaks and vice versa. Meat plant outbreaks have been implicated in ongoing community transmission and increased community infection rates, but with little evidence to prove this (Waltenburg et al., 2020; Dyal et al., 2020).

2.2. Theory

In our model, we argue that workplace outbreaks are a function of Covid-19 infections in the community. Where community incidence is high, an outbreak is more likely to occur and where community incidence rate low, an outbreak is less likely to occur. As initial infections may be difficult to detect, an infection may enter the plant weeks prior to the outbreak being detected, and as such, the precise time at which the infection entered the plant may vary from plant to plant.

We assume that new, secondary infections ('offspring' infections) are produced by a generative process (in this case, the disease) that controls the probability of a new infection being produced, given contact between an infected (I) and a susceptible (S) person. Our analysis considers the number of infections observed in a workplace for a given interval, t. The precise interval is not described here, but it is expected to be the interval from the first observed infection to the end of the outbreak (i.e., the time when no new infections are reported and the outbreak is considered resolved by the observing authorities).

In developing our model of workplace infections, we assume accurate knowledge of the

number of people in the workplace (N) and the number of susceptible (S) and infectious (I) people, in a given interval. We take a workplace to be defined by the following two properties. First: there is relatively small population N in the workplace, which is known and is subject to controlled change over time. Second: patterns of contact between members of this population are also controlled by the structure of the workplace.

Where an individual worker is susceptible, they may become infected in two situations (a) they had contact with an infected person in the community (for the purpose of this model, community contact includes household contact between infectious and susceptible persons) or (b) they had contact with an infected person in their workplace.

2.3. Data

The model's predictions were tested using data collected from the by UCD School of Veterinary Science UPCOM project from March 2020 - January 2022 (inclusive) in the Republic of Ireland. Information on the outbreak was gathered from the Veterinary inspectors (VI) in meat processing plants (MPP) that had reported outbreaks during the study period. This data was used in conjunction with the Central Statistics Office (CSO) Computerised Infectious Disease Reporting (CIDR) database which records investigated outbreaks countrywide and the Health Service Executive (HSE) Covid-19 database, which records information on Covid-19 infections at national and local levels.

MPP Data

The VI reports recorded information on the location (at county level), the size of the outbreak, the dates of the outbreak, number of outbreaks, and the number of employees at the facility. In total, data from 53 outbreaks across 35 meat-processing facilities was obtained

from the VI reports. A number of these facilities (18 in total) had two outbreaks, but no facility reported more than two outbreaks during the study period.

No information on staff turnover, reinfection rates (and possible subsequent immunity), or vaccination uptake rates among staff[1] were available from either the VI reports or the CIDR database for the 2nd outbreak, for this reason, the N of total staff was not adjusted for the second outbreak.

2.3.1. Data validation

The data provided by the VIs on the 35 MPPs was cross verified against records of outbreaks held in the CSO CIDR database, where possible. Data from the CIDR database could not be used to test the model as it does not capture some key information required by the model, such as number of employees working in the facility at the time of the outbreak.

Data provided by the VIs was matched to the CIDR outbreaks by outbreak type, location, and date in the database. In some cases, the outbreaks in the VI reports did not correspond perfectly to the CIDR reports, this was particularly true for outbreaks that had occurred at the beginning of the pandemic in spring 2020. Small discrepancies in dates between the VI reports and the CIDR database are expected, particularly early in the pandemic and during Covid-19 surges, where infrastructure capacity issues may delay reporting. Where such discrepancies did exist, the earlier of the two dates was chosen as the "start" of the outbreak, to avoid including any of the plant infections in the community incidence rates.

Where discrepancies in the number of infections existed, the VI reported number of in-

[1] Ireland's vaccination program began on 28/12/2020 with health care workers and high-risk individuals prioritised. Vaccination of the general public under the age of 70 did not begin until March 2021 so it is unlikely that a significant proportion of the MPP employees where vaccinated until autumn 2021

fections was chosen as the "correct" estimate. This was for two reasons (1) CIDR infection numbers may only include infections reported by mass testing in the plant once an outbreak had been identified and (2) VI reports may include infection numbers reported to the plant by employees who self-tested or had PCR-tests done at HSE testing centres and thus may not have been reported with the main body of the outbreak.

We excluded plants that reported two outbreaks less than two months apart from the sample, as these may be a single ongoing outbreak. In total there were two plants where this occurred. Additionally, we excluded one outbreak which we could not verify through the CIDR database or HSE reported infections for that location and date.

Community incidence rates

For this study, the "community" was defined as the local administrative unit (in this case; the county) where the facility was located. Information on the proportion of the employees that resided and worked in the same county was not available for all outbreaks, however, NHPET report that $> 95\%$ of employees were resident in the county of their workplace (Department of Health, 2020a).

The community incidence rates were calculated on the daily reported infections in the county where the facility is located. These case numbers were found using Ireland's Covid-19 Data Hub[2] and the approximate population of the county was derived from the 2016 census[3], the latest official population count available.

The model used 7-day incidence rates from the day prior to the first identified plant infection up to 100 days before the outbreak to make predictions. A 7-day incidence rate is a

[2] https://covid19ireland-geohive.hub.arcgis.com/
[3] https://www.cso.ie/en/statistics/population/

Table 2.1.: The relationship between the community incidence rate and outbreak size at an MPP, March 2020 - January 2022

	Outbreak 1		Outbreak 2	
Week	r	p	r	p
1	0.34	0.05	0.22	0.4
2	0.43	0.01	0.56	0.02
3	0.34	0.05	0.55	0.02
4	0.06	0.7	0.50	0.04
5	-	-	0.48	0.05
6	-	-	0.39	0.1

The correlations above show that the for the first outbreak, the size of the outbreak is more related to the recent rate of infection spread in the community, while the 2nd outbreak is more dependent on the sequence of infections over a number of weeks.

more appropriate measure of community infection than a 14-day incidence rate in this case, as it represents (approximately) one cycle of infection and so, is more sensitive to sudden changes in infection spread and approximates disease spread more accurately.

The county incidence rate was calculated using the standard method of determining the rate of infection spread in a population. It is then given by:

$$\frac{\sum(Cases_{d1}...Cases_{dn})}{\text{County population}} * 100,000$$

To establish that there was a relationship between the community incidence and the size of the outbreak at the facility, we calculated the weekly incidence rate for the community for 100 days prior to the outbreak. These values were then correlated with the outbreak size for the first and second (if applicable) outbreak. The correlations between the weekly infection rates and outbreaks are shown in table 2.1. Here, we observed that for the first outbreak the community rate is significantly correlated with the outbreak size for weeks 1 - 3 prior to the outbreak and for the 2nd outbreak, the community incidence rate is significantly correlated with the outbreak size up to 5 weeks prior to the outbreak.

2.4. Modelling

In this section, we describe a model based on a standard $SEIR$ model for infection evolution and spread within the workplace and using the binomial distribution to account community-to-workplace transmission. We assume a workplace with N employees, take β, α and γ to be the transmission, incubation and recovery rates for the disease in this workplace (so that the reproductive number is $R = \beta\gamma$) and take S_t, E_t, and I_t the number of susceptible, exposed, and infectious individuals in workplace at time t, and we have the standard SEIR expressions for disease spread

$$S_{t+1} = S_t - \beta S_t I_t / N$$

$$E_{t+1} = E_t + \beta S_t I_t / N - \alpha E_t$$

$$I_{t+1} = I_t + \alpha E_t - \gamma I_t$$

We assume that, as well as infections being communicated in the workplace as described by these expressions, infections periodically enter the workplace from the external community; we take p_t to represent the probability of infection in the community (the incidence rate) on a given day t, and take C to represent a list of such probabilities from some initial time 1. We assume that the periodic nature of workplace-to-community contact is represented in this list by the insertion of $p_t = 0$ values on days when such contact does not take place, and $p_t > 0$ values on days where such contact does take place. Given a list of p_t community infection probabilities up to some time $end = length(C)$ we wish to estimate the number of exposed or infected workers in the workplace at time end, assuming no initial workplace infections.

To estimate this number we define a function $SEIR_W$ which takes as input some current

time t, some list of community incidence values C, and some estimated S_t,E_t and I_t values for the workplace on that day. This $SEIR_W$ function returns a discrete probability distribution for the estimated number of infected and exposed individuals at the final time $t = end$ (that is, for the value $E_{end} + I_{end}$). This discrete probability distribution consists of a list of $N+1$ probabilities, with the first entry in this list representing the estimated probability that $E_{end} + I_{end} = 0$ will hold given the inputs t,C,S_t,E_t and I_t; the second entry the estimated probability that $E_{end} + I_{end} = 1$ will hold, given those inputs, and so on.

We define this function recursively. As a base case, we note that if this function is given values S,E, I and t is greater than or equal to the length of the list C ($t \geq |C|$), then the probability distribution returned must be one where $E + I = 1$ and all other values in the distribution are 0. We also note that, if the function is given values S, E, I and $t < |C|$ where $C[t] = 0$ (where there is no contact with the community on day t), then susceptible, exposed and infectious numbers in the workplace will be updated as in the standard SEIR model (representing spread within the workplace).

Finally, we note that if this SEIR updating has taken place with time moving on to $t+1$, but $p_t = C[t] > 0$ (so there is was some probability of community infection on the previous day), then there will now be S, E and I susceptible, exposed and infectious individuals produced solely through workplace spread, plus some additional number of newly exposed individuals produced by community contact on day t. If the number of such new community infections is k, then the probability distribution of cases at time end is obtained by calling the $SEIR_W$ function with arguments t, $S-k, E+k$ and I (those additional community exposures moving from the S to the E compartment). This number k can take on any value from 0 up to S (from no community exposures up to all susceptible individuals becoming exposed), and we

can estimate the probability of k exposures through community contact using the standard binomial expression

$$P_k = \binom{n}{S} p_t^k (1-p_t)^{S-k}$$

so that the probability of the distribution at time end being given by $SEIR_W(t, S-1, E+1, I)$ is P_1, the probability of that distribution being given by $SEIR_W(t, S-2, E+2, I)$ is P_2, and so on. The overall probability distribution at time end is thus given by the sum of all those possible distributions, each weighted by its probability of occurrence:

$$\sum_{k=0}^{S} P_k \times SEIR_W(t, S-k, E+k, I)$$

(where multiplying a discrete distribution by P_k means multiplying each individual term in that distribution by P_k, and where summing two discrete distributions means adding the corresponding terms in those two distributions). Figure 1 gives pseudocode for this recursive estimation process.

We used this model to predict outbreak numbers in 53 meat plant outbreaks in Ireland, involving 35 different meat plants with 18 plants subject to two separate outbreaks. We remove from this dataset 2 plants which had 2 outbreaks less than 2 months apart (since these cases likely reflect a single extended outbreak rather than two separate cases) leaving 49 outbreaks for analysis from 33 plants. For each outbreak we recorded the outbreak date, the number of reported infections in the outbreak, the number of employees in the plant, and the daily community incidence rates for the county in which the outbreak was located, for up to 100 days before the outbreak date.

For initial analysis we assume a reproductive number of $R_0 = 3$, incubation time of 6 and recovery time of 6, values consistent with the ranges for estimates found in, for example

Algorithm 1 $SEIR_W$. Assumes list of community incidence rates (infection probabilities) C, and parameters β (transmission rate) α (incubation rate) and γ (recovery rate).

function $SEIR_W(t, S, E, I)$
  repeat
    $S \leftarrow S - \beta SI/N$ // SEIR updates for within-workplace spread.
    $E \leftarrow E + \beta SI/N - \alpha E$
    $I \leftarrow I + \alpha E - \gamma I$
    $t \leftarrow t + 1$
  until $t \geq |C|$ **or** $C[t-1] > 0$ // If community contact at $t-1$, add contacts.
  length(D) $\leftarrow$ N+1 // discrete probability distribution (initialised to 0).
  if $t \geq |C|$ **then** // If end of time reached, point distribution is returned.
    $D[E+I] \leftarrow 1$ // Entry $E+I$ has probability 1; all others are 0.
  else
    $p \leftarrow C[t-1]$ // Probability of community infection at current time.
    for $k \leftarrow 0$ to S **do**
      $P_k \leftarrow \binom{S}{k} p^k (1-p)^{S-k}$ // binomial probability of k community infections.
      D $\leftarrow$ D + $P_k \times$ $SEIR_W(t, S - k, E + k, I)$
// sum of distributions returned by recursive calls
    end for // with each term in each distribution multiplied by P_k.
  end if
  return D // Return distribution.
end function

Overall Estimated Distribution $\leftarrow$ $SEIR_W(1, 0, 0, 0)$
// Initial call to function

(Y. Liu, Gayle, Wilder-Smith, & Rocklöv, 2020; Park, Cook, Lim, Sun, & Dickens, 2020), giving $\alpha = 1/6, \gamma = 1/6$ and $\beta = R_0\gamma = 1/2$.

To construct a list C of community infection probabilities for a given outbreak in week i, we calculated the weekly incident rate in that outbreak location for the week before the outbreak (W_i), the week before that (W_{i-1}), and so on. We take these incident rates to be estimates of the probability of being infected through community transmission, with some lag L,so that with lag L, the incident rate W_i is an estimate of the probability of community infection in week W_{i-L}. We assume that community incidence rates from some week $i - x$ up to the week prior to the outbreak act as causes of that outbreak; with lag L we assume that incident rate W_i estimates the actual probability of community infection at prior week W_{i-L}. For weeks from $i - L + 1$ up to week i (for which, due to the assumed lag L, we have no estimates), we assume a constant community incidence rate equal to W_i.

We assume that community contact takes place weekly, and so for a given starting week x prior to outbreak i, we construct a list of the form

$$C = [W_{i-x}, 0, 0, 0, 0, 0, 0, W_{i-x-L+1}, \ldots W_i, 0, 0, 0, 0, 0, 0, \ldots W_i, 0, 0, 0, 0, 0, 0]$$

(where the W_i sequence occurs L times).

In constructing these lists C for each outbreak, we thus have as input data the reported community incidence rates in the county where the outbreak was located up to the week prior to the outbreak report, and two free parameters: the lookback parameter x and the lag parameter L. We assume that these parameter values will be different for first and second outbreaks (that plant management will be more aware of and responsive to the risk of com-

munity infection for second outbreaks, since they have experienced an outbreak already); in fitting this model to data on outbreaks, we assume the same parameter values x_{first} and L_{first} for all first outbreaks in the dataset, and the same parameter values x_{second} and L_{second} for all second outbreaks.

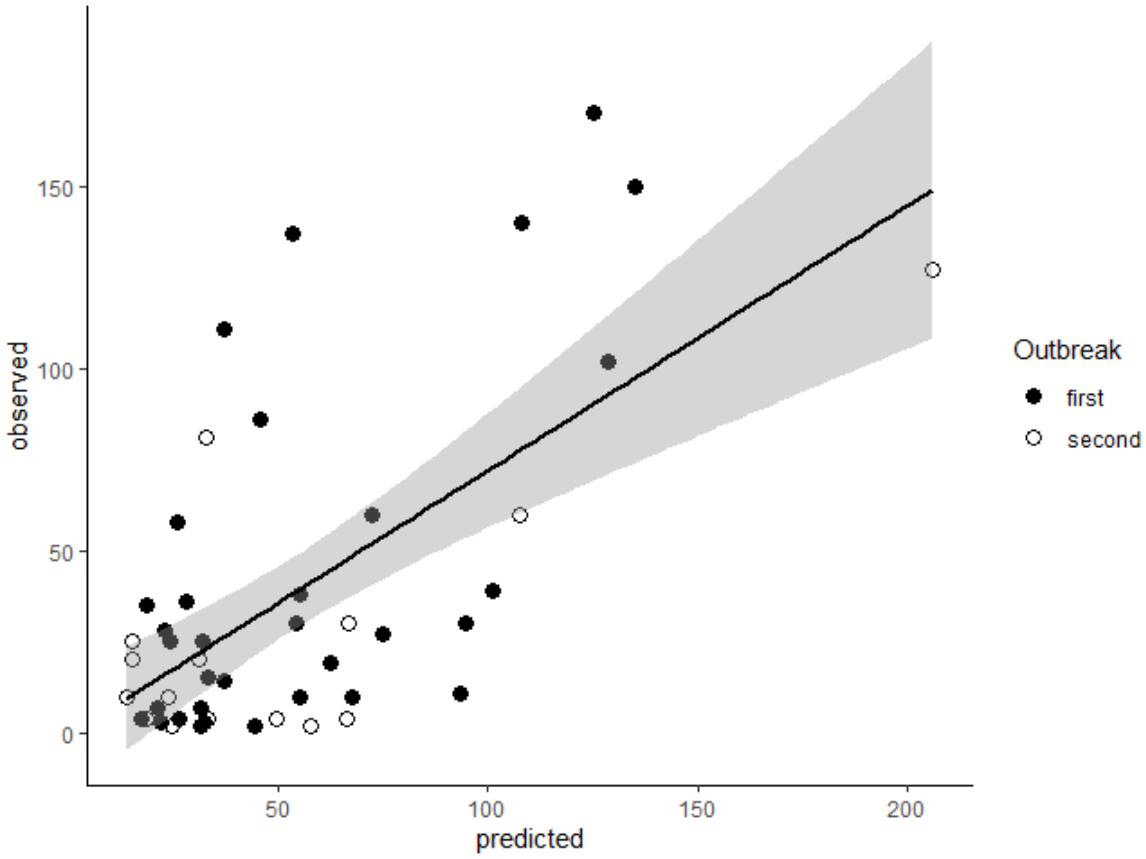


Figure 2.1.: Scatterplot of observed and predicted outbreak size for 49 meat plants in Ireland across the Covid-19 pandemic

The first outbreak in a plant are represented by the filled circles; the second outbreak in a plant are represented by the hollow circles. The diagonal line shows the linear regression line of best fit between observed and predicted values, the shaded area shows the standard error in that fit.

We run this model with a range of values of x and L when fitting this model to data, and assessed degree of fit in terms of the correlation between observed and predicted out-

break size, and in terms of the average difference (Root Mean Squared Deviation) between observed and predicted size. The best fit arises with values $x_{first} = 1, L_{first} = 3$ and $x_{second} = 3, L_{second} = 0$; with these parameter values the correlation between observed and predicted outbreak numbers overall was $r = 0.62$ ($p < 0.0001, RMSD = 5.6$; see Figure 2.1). These results suggest that community infection rates in the period before a Covid-19 outbreak in a meat plant can be used to predict outbreak size to at least some degree: over the 49 outbreaks considered here, the difference between reported outbreak size and the size predicted using the community incidence rate and the plant size (and this model of infection spread) was less than 6 cases, at least on average.

Analysing further we found that prediction accuracy was notably higher for second outbreaks ($r = 0.77, RMSD = 7.6$) than for first outbreaks ($r = 0.61, RMSD = 11.7$). It seems likely that this difference in accuracy arises because reported community incidence rates were weaker reflections of the true incidence rate in the community early in the course of the Covid-19 pandemic. First outbreaks were more likely to occur earlier in the pandemic, and so outbreak numbers (which are based on these reported incidence rates) are likely to have weaker predictive power as a consequence. The $SEIR_W$ algorithm doesn't address the uncertainty associated with community infection numbers (the binomial probability distribution assumes that the exact probability of community infection is known).

Where the above analysis assumes a constant reproduction number R, we also developed a modified version of this algorithm where first outbreaks were subject to some value R_{first} and second outbreaks some different value R_{second}. Since second outbreaks are very likely to arise in the context of protective workplace measures (that is, lower workplace R numbers), we would expect model fit to improve with with higher values for R_{first} than R_{second}.

Running the model just as above but $R_{first} > R_{second}$ produced a higher overall correlation between predicted and observed outbreak numbers: with $R_{first} = 3.5$ and $R_{second} = 2.5$, for example, the overall correlation between predicted and observed numbers was $r = 0.66$ ($RMSD = 5.1$). Cases where $R_{first} < R_{second}$ gave the opposite result: with $R_{first} = 2.5$ and $R_{second} = 3.5$ the overall correlation between predicted and observed numbers was $r = 0.49$ ($RMSD = 12.2$). The fact that model fit is improved by giving more realistic values for R supports the $SEIR_W$ model as an account of workplace infection spread.

Uncertainty In the SEIR model, the probability of an infection (a positive test) is a fixed value based on the incidence rate of infection in the community, however, the incidence rate in the community is an uncertain value and will vary around the "true" value of infection in the community. The random variance in infection rates could be approximated using a beta- binomial model. While the chance of success in a binomial model is a fixed value, in a beta-binomial, this free value. However, this is a very computational expensive process and is unlikely to produce values significantly different to those observed in the SEIR model above.

2.5. Community spread

A significant concern during the pandemic is that large outbreaks such as those in meat plants, hospitals, and nursing homes will seriously impact the community in which the outbreak occurs, through either infection being *introduced* to the community because of the outbreak or the outbreak leading to higher rates of infection in the community overall. However, in this model, we argue that outbreaks are, by-and-large, a consequence of community infection spread rather than the inverse.

To investigate whether community rates are a consequence of individual outbreaks rather than infections in the plants being a consequence of the *community infection rates*- that is, the infection originated and spread from the community to the meat plant, rather than the infection originating and spreading from the plant to the community, we examined whether there was a statistical difference between the national incidence rates and the county incidence rates, pre-, during-, and post-outbreak. If the infection originated or spread from the meat plant to the community, (the outbreak causes spread to the community), then it is likely that the characteristics of the community infections will differ significantly to the characteristics of national infections, and a difference between the incidence rates will be observed. In the case where the outbreak causes community rates to rise consistently, then higher rates of infection should be observed even after the outbreak has been halted in the plant. However, if the situation is consistent with the national transmission and spread, then the community rates should be consistent with national rates of infection.

The county incidence rate was calculated for 60 days prior to the outbreak and 90 days after "day 0," the day which the outbreak was notified. As the exact duration of each outbreak was unknown, the post-outbreak period was taken to be day 31 - day 90 after the outbreak had been notified. The outbreak period itself was taken to be day 0 - day 30, or approximately 4 cycles of infection *after* the outbreak had been detected in the plant. The county incidence rate was calculated as the sum of reported cases over a 7-day period divided by the population of the county and normalised to a population of 100,000 people.

The county incidence rate was compared to an adjusted national incidence rate. This adjusted national incidence rate was calculated as the sum of the daily cases over a 7-day period less the sum of the daily cases for 7-day period outbreak affected county over the

national population less the county population, standardised to a population of 100,000 people.

$$\text{National incidence rate} = \frac{\sum(Nationalcases_{d1...dn} - Countycases_{d1...dn})}{\text{National population } - \text{County Population}} * 100,000$$

Again, all incidence rates were calculates as a 7 day incidence rate.

T-tests were used to determine if there was any significant difference between the national incidence rate and the county incidence rate before the outbreak occurred, $t = 1.0973$, $p = 0.2726$ with a mean of 102 cases per $100,000$ for the national rate and 96 cases per $100,000$ for the county rate prior to the outbreak. Correspondingly, no significant difference was observed between the national rate and county rate post-outbreak, $t = 0.39838$, $p = 0.6904$, here the mean national incidence rates was 94 cases per $100,000$ and 93 cases per $100,000$ for the outbreak county.

No significant difference was observed for county incidence rates pre- and post-outbreak, $t = 0.72787, p = 0.4667$. This suggests that the outbreaks are unlikely to result in higher community rates in the long-term as the community rates are typically consistent with each other pre- and post-outbreak and consistent with the national rate of infection post-outbreak.

The period while the outbreak is ongoing has statistically higher community rates than the pre- and post-outbreak periods, however, these periods where there are high community rates coincide with national high periods, suggesting that the outbreaks are likely to occur when there are periods of high infections in the community and nationally. Typically, the community rate is rising and falling in-line with the national rate.

That the infections don't appear to cause continued spread outside the plant in the long-term is most likely a consequence of the mitigation measures in place at the time of the outbreak. The period where the analysis was done included stringent measures such as lockdowns, systematic mass-testing, and isolation of confirmed and suspected cases. It seems that these public health measures are successful at preventing sustained infection growth in the community.

2.6. Conclusions

In this section, we have presented a simple extension of the standard *SEIR* compartmental model of infectious disease spread. We've shown that this model can effectively predict the size of an outbreak in a workplace given the community infection spread in the weeks preceding the outbreak. The predictions of the model are consistent with outbreaks observed in Irish MPPs.

There are a number of ways in which the model could be extended in future work. This model could be applied without alteration to other workplaces such as nursing homes, which have seen large outbreaks. The current model could also be refined by using data on reinfection, immunity, and vaccination. Additionally, a large proportion of infections appear to occur in particular areas of these plants such as the boning hall, so future investigations of outbreaks should take this into consideration.

There are a number of limitations to these results. Firstly, we must assume that the community infection rates are a true reflection of the disease circulation in the community. This is an unrealistic assumption as reporting may be less accurate at certain times, such as

when infection surges overwhelm reporting capacity and at the beginning of the pandemic when the infrastructure for testing and reporting is not fully operational. Secondly, no information was available on re-infection rates or staff turnover for the plants which had two outbreaks. In some cases, previous infection may lead to a degree of immunity to re-infection and in plants where the first outbreak affected a large number of the employees, there may be a so-called "herd immunity effect" among the employees, slowing and containing the spread during the second outbreak. This could potentially effect the predictive power of the model.

Part III.

Disease reproduction

3. Behavioural Response Model

Fintan Costello
Paul Watts
Rita Howe

3.1. Introduction

In simple epidemiological models of disease spread, infection numbers at time t are a function of disease transmissibility p, incubation rate α and recovery rate γ (properties of the disease), of the proportion of infectious and susceptible individuals in the population at time t, and of behaviour: in particular, of the average number of contacts individuals make with others at that time. In some models (Bertozzi, Franco, Mohler, Short, & Sledge, 2020) this contact number is taken as to be constant; in others is treated as a free parameter, varying with time in a way that is not described within the epidemiological model but instead is estimated via fitting the model to data (Ndaïrou, Area, Nieto, & Torres, 2020; IHME COVID-19 forecasting team, 2020; Gleeson et al., 2022) by using mobility or contact tracing datasets(Nouvellet et al., 2021; Badr et al., 2020; Russo et al., 2020) or by using assumed seasonal changes in behaviour (X. Liu et al., 2021; Bukhari, Jameel, Massaro, D'Agostino, & Khan, 2020).

It is clear, however, that contact rates between individuals in a population will tend to vary as a function of infection risk, with people reducing contacts and changing behaviour when risk is high in what has been termed a 'behavioural immune response' (Schaller, 2011;

Verelst, Willem, & Beutels, 2016). Capturing this relationship between human behaviour and infectious diseases is seen as 'the hard problem of epidemiology' (Perra, 2021) and a wide variety of behavioural response models have been proposed which link infection numbers and behavioural response in different and often complex ways (Ajbar, Alqahtani, & Boumaza, 2021; Weitz, Park, Eksin, & Dushoff, 2020; Tkachenko et al., 2021; Manrubia & Zanette, 2021; Steinegger, Arola-Fernández, Granell, Gómez-Gardeñes, & Arenas, 2022; Steinegger, Arenas, Gómez-Gardeñes, & Granell, 2020; Avery, 2021) .

A critical problem for research in this area is that of validation: many models are not tested (they give purely theoretical presentations), and when testing is done, it is almost exclusively carried out by fitting the model to existing data; that is, by varying model parameters until model and data agree to some extent (Verelst et al., 2016; Funk, Salathé, & Jansen, 2010; Perra, 2021). Such model fits do not act as confirmatory evidence in favour of a model, for at least three reasons. First, a good model fit may arise, not because the model is a useful description of the underlying process, but because parameter variation gives the model flexibility to fit any data. Second, quite different models can often give good fits to the same data; because of this, a good model fit leaves the underlying process unclear. Third, because model fit is specific to both the parameters and the data used, the fact that a model gives a good fit to one specific set of data with a particular choice of parameter values does not imply that this fit will generalise.

Our aim here is to address this problem of validation by presenting a simple and generic behavioural response model (an extension of the standard SEIR compartmental model) and by showing that this model leads to three parameter-free numeric predictions about infection numbers; predictions that, if the model describes the underlying process well, should hold

across all sets of data. The first prediction is that the effective reproduction number R prior to herd immunity will have a median of 1; the second is that proportional changes in infection numbers will follow the standard Cauchy distribution $C(0, 1)$; the third is that the frequency of high infection numbers will follow a power-law distribution x^{-k} with exponent $k = 2$. We show that these predictions do, in fact, hold in a large Covid-19 dataset covering 190 countries: the mean estimated R value across all countries is statistically indistinguishable from 1, relative changes in new infection numbers follow a standard $C(0, 1)$ distribution very closely, and fitting a power law to the frequency distribution for infection numbers for each country, the estimated exponent is statistically indistinguishable from 2.

3.2. ASEIR Model of behavioural response to infection risk

Models of behavioural response assume that when people are aware of infection risk, they will change their behaviour (their level of risky contact) with the aim of balancing the risk of infection associated with contact against the various (economic, social, and psychological) gains associated with contact. Our model assumes that each person has a certain constant risk or probability of infection per day, X, which they are willing to accept (a level which balances gain from contact with risks from contact), and when they become aware of increased infection risk, they will reduce their number of contacts per day until their overall estimated risk that day is at that level. In a pandemic situation, we expect that awareness will spread as the infection itself spreads, rapidly reaching some saturation level where a large proportion of the population are responding to infection risk. Once this point is reached the probability of infection, and so the overall number of new infections arising in the population, will tend vary around some constant value or set-point depending on X (being pushed away from that point by changes in the infection itself or in various other factors, and being returned to that

point by behavioural response to those changes). Assuming that infection confers immunity, this pattern of behavioural response will continue until the total number infected reaches some 'herd immunity' level; after this point the number of new infections will necessarily decline irrespective of behavioural changes.

To adjust their behaviour in response to infection risk, people must have some way of estimating risk; as in most behavioural change models, we assume that people estimate the risk of infection at time t based on (some approximation of) the number of infections in the population at a previous time $t - L$, where L is the lag between an infection occurring and it being known or reported. We express these ideas of homeostatic behavioural response to risk, spreading awareness of risk, and risk estimation with lag, in an extension of a standard SEIR compartmental model where S_t represents the number of susceptible individuals, E_t the number of exposed individuals (who are incubating infection but not yet infectious), and I_t the number of infectious individuals in a population of size N at time t, and where i_t represents the number of individuals who were newly infected at that time. Assuming that infection confers lasting immunity, we also have a recovered or removed compartment containing $N - S_t - E_t - I_t$ immune individuals: to avoid confusion with R_t, the effective reproduction number at time t, we do not refer to this removed compartment here. In this model we have

$$S_{t+1} = S_t - i_t$$

$$E_{t+1} = E_t + i_t - \alpha E_t$$

$$I_{t+1} = I_t + \alpha E_t - \gamma I_t$$

(newly infected individuals at time t move from the S to the E compartment, αE individuals move from the E to the I compartment at time t, and γI individuals recover at time t). In

a standard SEIR model new infection numbers are given by

$$i_t = p(KI_t/N)S_t$$

where the (KI_t/N) term approximates the probability of contact with an infected individual given K contacts at time t (assuming contacts take place at random) and so $p(KI_t/N)$ gives the probability of a susceptible individual becoming infected given that they make K random contacts, and the expected number of new infections is this probability times S_t. In our extension of this approach we assume that at time $t = 0$ (before the introduction of a new infectious disease), this contact number K is set relative to the risk of infection from pre-existing diseases at that time: letting e represent the risk of infection from any of those diseases, we then have $eK = X$ (the number of contacts is set so that the risk of infection is approximately X) and so $e = X/K$.

Susceptible individuals who are not aware of a new infection risk at time $t > 0$ maintain this level of contact K, so their risk of infection from the new disease remains at pKI_t/N as in the standard SEIR model. Individuals who are aware of and actively responding to infection risk will adjust their number of contacts K_t so that their estimated probability of infection is, on average, X. Let $I_{est}(t)$ represent the *estimated* number of infectious individuals in the population at time t. Then for these aware individuals, their total estimated probability of infection from a single contact is

$$e + \frac{pI_{est}(t)}{N} = \frac{X}{K} + \frac{pI_{est}(t)}{N}$$

and so, these individuals will adjust their contacts so that

$$\left[\frac{X}{K}+\frac{pI_{est}(t)}{N}\right]K_t = X \rightarrow K_t = \frac{X\,K\,N}{XN+pKI_{est}(t)}$$

Letting A_t represent the number of individuals who are aware of and actively responding to infection risk at time t, then since awareness rises when individuals hear about infections among people they know but falls where there are no such infections, we have

$$A_{t+1} = A_t + \left(\frac{bI_{est}(t)}{N}\right)(N-A_t) - f\left(1-\frac{bI_{est}(t)}{N}\right)A_t$$

where b represents the average number of people an individual knows and f represents the rate at which low risk of infection causes individuals to cease responding. Given this, the average number of contacts at time t for the population as a whole is

$$\left(1-\frac{A_t}{N}\right)K + \left(\frac{A_t}{N}\right)\frac{X\,K\,N}{XN+pKI_{est}(t)}$$

The evolution of this ASEIR model depends on the estimated number of infections $I_{est}(t)$. Individuals can only observe or find out about an infection after a certain observation lag L (necessarily greater than the incubation time for the infection), and so we simply take $I_{est}(t) = I_{t-L}$ (the estimated number of infections at time t is equal to the actual number of infections at time $t-L$), where L is within that range. Given this the expected number of new infections at time t is

$$\begin{aligned} i_t &= \left[\left(1-\frac{A_t}{N}\right)K + \left(\frac{A_t}{N}\right)\frac{X\,K\,N}{XN+pKI_{t-L}}\right]p\,S_tI_t/N \\ &= \left[1-\left(\frac{A_t}{N}\right)\left(\frac{pKI_{t-L}}{XN+pKI_{t-L}}\right)\right]p\,K\,S_tI_t/N \end{aligned}$$

We assume that there is initially no awareness ($A_0 = 0$). This means that if $b = 0$ then

$A_t = 0$ for all t, and so $i_t = pKS_tI_t/N$ and this model includes the standard SEIR approach as a special case. This model depends on 4 parameters from the standard SEIR model (p, K, α and γ) and 4 behavioural awareness parameters (b, f, X and L). Here we treat b as a switching parameter taking on values 0 or K, where 0 gives a standard SEIR model while $b = K$ gives an ASEIR model with behavioural response to infection (and where the number of people an individual knows is, on average, equal to their number of pre-pandemic contacts).

Assuming that p, K and γ are such that the initial reproduction number $R_0 = pK/\gamma > 1$, this model shows a characteristic pattern of evolution in which infection numbers initially rise until awareness A reaches a certain level, at which point infection numbers return towards some relatively stable value, with this stability continuing until a point at which herd immunity is reached (after which infection numbers necessarily decline). Figures 1(A) and 1(B) compare the evolution of new infection numbers in a standard SEIR model against their evolution in an ASEIR model (with parameter values selected for demonstrative purposes). Where in the standard SEIR model new infection numbers rise to a high level and then decline to 0, in the ASEIR model infection numbers rise to a much lower level and then tend to oscillate around a constant number of new infections.

3.2.1. Modelling disease variants

This stabilisation of new infection numbers assumes no other perturbations or shocks affecting infection numbers. We can extend this simple model to account for such perturbations by considering the emergence of new disease variants, each with different transmission rates p. Assuming variants $j \in \{1 \dots m\}$ each with transmission rate p_j and each entering the population at time T_j, we define $E_{j,t}$ and $I_{j,t}$ to be the number of exposed/infectious individuals

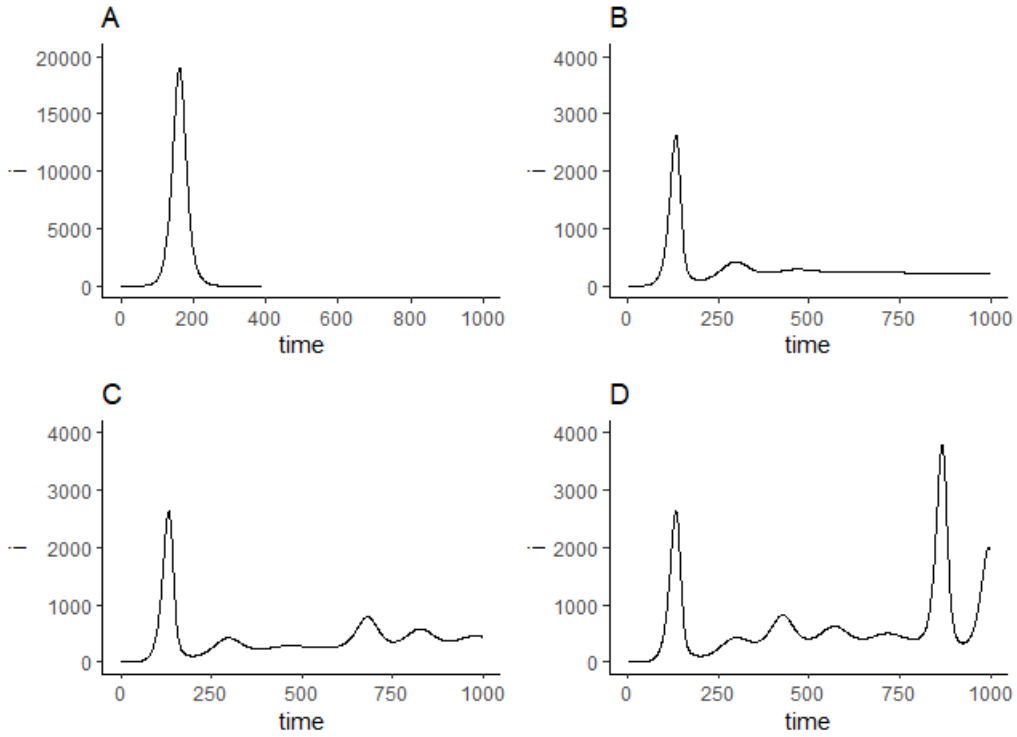


Figure 3.1.: Number of new infections i_t over time generated in simulations runs of the standard SEIR model ($b = 0$) and the ASEIR behavioural response model ($b = K$).

All simulations with population size $N = 10^6$, incubation rate $\alpha = 1/10$, recovery rate $\gamma = 1/10$, fixed contact number $K = 10$ and with behavioural response parameters $x = 1/5000$, $b = K$, $f = 1/100$, and lag $L = 1/\alpha + 1/\gamma = 20$; initial values are $S_0 = N$ and $E_0 = I_0 = A_0 = 0$. (A) SEIR model with a single disease variant with transmission rate and entry time $(p_1 = 3/100, T_1 = 1)$. Identical graphs are produced for the SEIR model with 2 or 3 disease variants entering the population (not shown). (B) ASEIR model with a single disease variant $(p_1 = 3/100, T_1 = 1)$; (C) ASEIR model with two variants $(p_1 = 3/100, T_1 = 1)$ and $(p_2 = 7/100, T_2 = 500)$; (D) ASEIR model with three disease variants $(p_1 = 3/100, T_1 = 1)$, $(p_2 = 7/100, T_2 = 250)$ and $(p_3 = 25/100, T_2 = 750)$. In these graphs $t = 1$ is the first time-step on which $i \geq 1$ (there is at least one infectious individual in the population), and new infection numbers $i_t < 1$ are not shown.

with variant j at time t and let

$$\overline{p}_t = \frac{1}{I_t} \sum_j p_j I_{j,t}$$

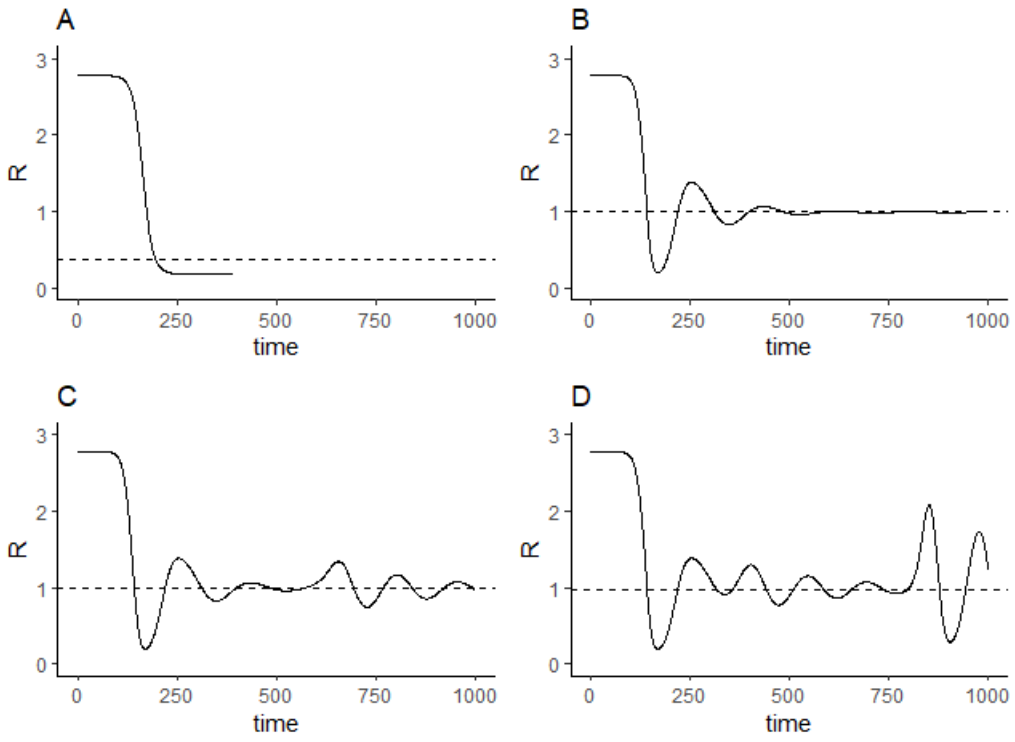


Figure 3.2.: Effective reproduction numbers R_t for the SEIR and ASEIR simulations in Figure 2.

Effective reproductive number is calculated from simulated data as $R_t = i_t/(\gamma I_t)$. The dashed line shows the median of the r values calculated from i_t and I_t values in each simulation. For the ASEIR simulations, the median R_t was approximately 1; for the SEIR model, the median R_t was 0.18.

be the weighted average transmission probability at that time. Then we have

$$i_{j,t} = \left[1 - \left(\frac{A_t}{N}\right)\left(\frac{\overline{p}_{t-L} K I_{t-L}}{XN + \overline{p}_{t-L} K I_{t-L}}\right)\right] p_j K S_t I_{j,t}/N$$

as the number of new cases of variant j at time t, and so

$$S_{t+1} = S_t - \sum_j i_{j,t}$$

$$E_{j,t+1} \begin{cases} 1 & \text{if } t = T_j \\ E_{j,t} + i_{j,t} - \alpha E_{j,t} & \text{if } t > T_j \end{cases}$$

$$E_{t+1} = \sum_j E_{j,t+1}$$

$$I_{j,t+1} = I_{j,t} + \alpha E_{j,t} - \gamma I_{j,t}$$

$$I_{t+1} = \sum_j I_{j,t+1}$$

(and A_t as before).

Figures 1(C) and 1(D) illustrate ASEIR model simulations with the same parameters as before but with 2 or 3 disease variants with different transmission probabilities entering the population at various times. We also ran the SEIR model with these variants: however, the introduction of these variants had no effect on SEIR infection numbers (since herd immunity has been reached by the time these variants entered the population). For the ASEIR model, by contrast, these new variants has a substantial effect on new infection numbers, with 'plateaus' in new infection numbers between variant arrival (in Figure $2(C)$ there is a plateau of around 300 new infections per time-step between times $t = 300$ and $t = 600$, for example).

It is useful to consider the effective reproduction number, R_t, produced in these simulations. R_t represents the number of new infections generated by existing infections at time t, and can be calculated from simulated data as $R_t = i_t/(\gamma I_t)$; Figure 2 shows the R_t values calculated at each time-step from the i_t and I_t values generated for each simulation in Figure 1 (with 3 disease variants for the SEIR model simulation, and 1, 2 and 3 disease variants for the ASEIR model). The median R_t value for the SEIR model is low (around 0.18) while the

median R_t value for the three ASEIR simulations are all almost exactly 1 (even when there is significant variability in R_t due to the arrival of disease variants in the population).

A median R value of 1 is clearly predicted in the ASEIR model with a single infection (Graph B in Figures 1 and 2) because in that situation behavioural response acts to maintain infection numbers at a relatively constant 'plateau' level after the initial wave, necessarily maintaining $R = 1$. Similar predictions of $R \sim 1$ arising from such 'plateaus' have been made in a number of other models. However, even with multiple infection waves and no plateaus (Graphs C and D), the ASEIR model still predicts a median R value of 1. This more general prediction arises because each individual infection variant, in this model, will be returned to a relatively stable level after its initial wave, and so the total number of new infections (made up of a 'superposition' of these individual infection variants) will similarly return to a relatively stable level (until herd immunity is reached).

As these figures illustrate, infection and effective reproduction numbers R_t produced by the ASEIR model have a number of general characteristics: infection numbers do not immediately rise to herd immunity levels and then decline to 0; reproduction numbers, similarly, do not rise and then decline monotonically, but instead vary over time around a median of 1; infection numbers can show various 'plateaus' of relatively constant numbers of new infections over long time periods; and both infection and reproduction numbers show noticeable effects of new variants and other stochastic shocks. These characteristics are evident in reported infection and reproduction numbers for the Covid-19 pandemic (Manrubia & Zanette, 2021; Arroyo-Marioli, Bullano, Kucinskas, & Rondón-Moreno, 2021; Koyama, Horie, & Shinomoto, 2021; Tkachenko et al., 2021).

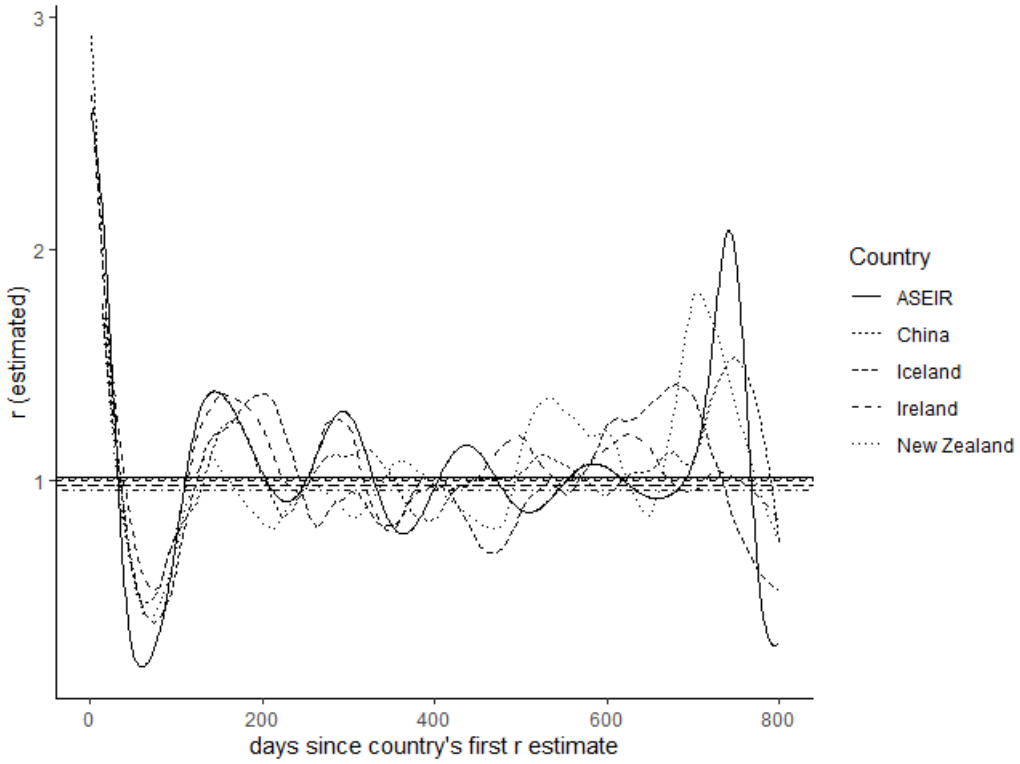


Figure 3.3.: Reported effective reproduction numbers R_t per day for China, Ireland, New Zealand, and Iceland

Taken from the OWID COVID dataset and aligned on initial reproduction number (see section 3.4) and reproductive numbers R_t generated by the ASEIR model (from graph D in Figure 2, aligned by taking day 1 to be $t = 110$ in that graph). Horizontal lines show the median R_t value for each country/the model calculated across the the entire period (all are almost exactly 1). Model and country R_t values show a common pattern of decline and rebound over the first ≈ 150 days, and agree closely: in the period up to day 146 (the first ASEIR peak) the Pearson product-moment correlations between ASEIR R_t values and country R_t values were $r = 0.81$ (China), $r = 0.96$ (Ireland) $r = 0.9$ (New Zealand) and $r = 0.94$ (Iceland), with all correlations significant at $p < 10^{-15}$.

3.2.2. Comparison with observed reproduction numbers

The model described above assumes a single initial exposed individual and random homogenous spread in a single population. As such, the most natural points of comparison are to patterns of infection spread at the origin of a new disease, and to patterns of spread when that new disease enters a comparatively isolated and relatively small population: in these cases the ASEIR model predicts that the same trends in infection and R numbers will be

seen, at least in the initial period of infection.

To test this prediction we compare estimated R numbers for Covid-19 in 4 countries (the origin country China and 3 island countries with relatively small populations: Ireland, Iceland, and New Zealand) against each other and against R numbers produced by the ASEIR model in Figure 2(D). Estimated R numbers were taken from the Our World in Data Covid-19 Hub(Ritchie et al., 2020) accessed June 30, 2022 (see 'Availability of Data and Materials'). Since Covid-19 arrived at different dates in China, Ireland, New Zealand and Iceland (first R estimates for these countries were on 23/01, 25/03, and 27/03 and 05/04/2020, respectively) for comparison purposes we aligned the first R number estimate for Ireland, Iceland and New Zealand with the closest reported R number for China in the initial phase of the pandemic. R_t numbers produced by the ASEIR model in Figure 2(D) were aligned with these numbers by taking day 1 for the ASEIR model to be time $t = 110$ in figure 2(D) (the point at which R numbers generated by ASEIR begin declining rapidly). Figure 3 shows the aligned R values: at the initial stage of infection (up to around 150 days after the first reported R_t number for each country), R values for these countries show very similar patterns of steep decline in R followed by 'rebound' at around $R_t \approx 0.5$, followed by oscillation around approximately 1 in all cases. The ASEIR model follows this pattern closely: over the initial ≈ 150 day period the correlations between ASEIR R_t values and country R_t values were strong (all $r > 0.8$, all significant at $p < 10^{-15}$).

3.3. Predictions

This fit between predicted and observed R_t values are suggestive, demonstrating as it does that the ASEIR model with the selected parameter values can match the observed evolution of R_t for these countries to at least some degree. This fit does not, however, give support

for the behavioural response approach to infectious disease modelling in general; evidential support for a given model is not obtained by fitting a parametrised model to specific data. Instead, evidential support is obtained by testing hypotheses derived from that model which are independent of the model parameters and which should apply to all observed data, not just to specific fitted data. Because the behavioural response approach assumes that infection and reproduction numbers will return to a given acceptable level and because, like all compartmental models, these are 'mean field' predictions (based on and describing expected means, with observed values expected to vary around these means following some error distribution) this approach leads to various predictions about the distributions of R_t, i_t, and related values.

The ASEIR model's predictions about the distribution of these values typically hold only in the long run, when infection numbers have reached approximate stability after stochastic shocks (clearly, the model does not predict $R \sim 1$ will hold during an initial infection wave). These predictions also do not hold when herd immunity has been reached (because at that point $R < 1$ necessarily holds independent of any behavioural response). To test these predictions, we must specify the domain where they apply, which we refer to as the 'oscillatory' domain.

Assuming that recovery from infection confers lasting immunity, herd immunity is reached at some time t where $S_t \approx 1/R_0$, and infection numbers necessarily decrease after that time. In the standard SEIR model, this point is reached relatively quickly, because infection spreads exponentially through the population. In the ASEIR model, this point is reached more slowly: with a single infection and taking X to be the acceptable probability of infection, approximately XN individuals will be infected per day, and herd immunity will be reached

at time $t = h$ where

$$N - hXN \approx 1/R_0 \rightarrow h \approx \frac{1}{X}\left(1 - \frac{1}{R_0 N}\right) \approx \frac{1}{X}$$

We distinguish between the herd immunity and the oscillatory domains by noting a quantitative difference between these two domains: in the herd immunity domain R_t must necessarily remain below 1 (because infection numbers must decrease in this domain), while in the oscillatory domain R_t can go from below 1 to above 1. Taking h to be the highest value for which $R_{h-1} \leq 1 \leq R_h$ holds, we see that the region $t \leq h$ must be in the oscillatory domain. Similarly, since R_t can go from below 1 to above 1 only after the initial wave of a new variant (or at the start of that initial wave), we see that all variants should have reached approximate stability by time h. These ASEIR model predictions are thus expected to hold only in the oscillatory domain $t \leq h$ (that is, in the period of time from the start of pandemic infection up to the most recent date at which R_t moved from below to above 1).

3.3.1. The median value of R_t is 1

We can state the ASEIR model's specific prediction for R as follows: For a given country c we take $R_{t,c}$ to be the reproduction number in that country on day t. Defining h_c for that country as the most recent day on which $R_{t-1,c} \leq 1 \leq R_{t,c}$, this model predicts that $R_{t,c}$ will vary around 1 in the oscillatory domain $1 \leq t \leq h_c$. This prediction holds both in situations where there are clear plateaus in the number of new infections (when these numbers are flat for a long period of time) and also in cases where no such plateaus are observed: in both cases this homeostatic return is active. Letting $R_{t \leq h_c}$ be the median value of R in the region $t \leq h_c$ for country c, random between-country variability means that country medians $R_{t \leq h_c}$ will themselves vary across countries around an overall expected mean of 1. More formally, letting M_1 represent the mean value of $R_{t \leq h_c}$ across a set of different countries,

our hypothesis is that the 95% confidence interval for M_1 will contain the predicted value 1.

Note that various forms of this general prediction $R_t \sim 1$ have been derived in various behavioural response models and supporting results have been seen in various countries. (Manrubia & Zanette, 2021; Tkachenko et al., 2021) The main novelty in our proposal is a formal statement of the domain in which this prediction is expected to hold, a formal statistical test of the hypothesis, and a general application of this test to data from all countries worldwide.

3.3.2. Proportional change in i_t follows Cauchy distribution $C(0,1)$

In the oscillatory domain infection numbers i_t will tend to vary, in the ASEIR model, in a way that depends on the lag L between an infection occurring (at time $t-L$) and that infection being observed by others and causing a behavioural response (at time t). This lag is necessarily greater than the incubation period for the infection (an infection becoming observable only after incubation) and means that the observed rate of new infections at time t is equal to the actual rate of new infections at time $t-L$. If $i_{t-L} > X$, the acceptable risk level, then the overall behavioural response at time t will reduce contact numbers, pushing i_t downwards, while if $i_{t-L} < X$ then the overall behavioural response at time t will increase contact numbers, pushing i_t upwards, and so the difference $i_t - i_{t-L}$ varies around 0. Since this overall behavioural response is the sum of all individual responses in the population, from the Central Limit theorem this difference $i_t - i_{t-L}$ will follow a Normal distribution $i_t - i_{t-L} \sim \mathcal{N}(0, \sigma_t^2)$ with some variance σ_t^2 (which may change over time). The difference $i_{t-2L} - i_{t-L}$ will follow the same distribution (albeit with variance σ_{t-L}^2). Defining a measure

of proportional change in new infection numbers from time $t - 2L$ to time t,

$$D_L(t) = \frac{i_t - i_{t-2L}}{i_t + i_{t-2L} - 2i_{t-L}} = \frac{(i_t - i_{t+L}) - (i_{t-2L} - i_{t-L})}{(i_t - i_{t-L}) + (i_{t-2L} - i_{t-L})}$$

we see that D_L is the ratio of two standard Normal variables (sums of common standard deviations cancelling), and so follows the standard Cauchy distribution $C(0, 1)$ (the Cauchy distribution with location parameter 0 and scale parmeter 1). The ASEIR model thus predicts that in the oscillatory domain this measure D_L will follow $C(0, 1)$ for values of L in a region greater than the incubation period of the infection.

We can assess this prediction informally via measures of goodness-of-fit, by asking to what extent the distribution $C(0, 1)$ gives a close fit to the distribution of of D_L values in the $t \leq h_c$ domain. More formally, we note that, if a set of numbers is drawn from some Cauchy distribution C, the median of those numbers is an unbiased estimate for the location parameter of C, and the median of the absolute values of those numbers is an unbiased estimate for the scale parameter of C. Defining d_c to be the median value of D_L for country c in the domain $t \leq h_c$, and $|d|_c$ to be the median of the absolute values of D_L in that domain, we thus expect that d_c will be distributed around 0 and $|d|_c$ around 1. Letting M_2 be average value of d_c across a set of different countries and M_3 be the average value of $|d|_c$ across those countries, our specific hypotheses are that the 95% confidence interval for M_2 will contain the predicted location parameter value 0, and that the 95% confidence interval for M_3 will contain the predicted scale parameter value 1, for values of L in a region greater than the incubation period of the infection.

3.3.3. Frequency distribution of i_t follows a power law with $k = 2$

In the ASEIR model the degree of response to infection risk depends on the degree to which current estimated risk is above the acceptable level X: the higher the current value of $I_{est}(t)$, the greater the behavioural response to risk. Here we consider the distribution of values i_t in this model when $I_{est}(t)$ is high: specifically, where

$$X \ll pKI_{est}(t)/N$$

(where the probability of infection given K contacts and the estimated number of infections in the population is much greater than the acceptable level of risk, X).

In this situation we assume that $A_t = A_{t+1} \approx N$ (because infection numbers are high, almost everyone is aware of infection risk); this gives

$$\begin{aligned} i_t &= \left[1 - \left(\frac{pKI_{est}(t)}{XN + pKI_{est}(t)}\right)\right] p\,K\,S_tI_t/N \\ &= \left(\frac{1}{1 + \frac{pKI_{est}(t)/N}{X}}\right) p\,K\,S_tI_t/N \end{aligned}$$

Similarly, in this situation we can assume

$$1 + \frac{pKI_{est}(t)/N}{X} \approx \frac{pKI_{est}(t)/N}{X}$$

giving

$$i_t \approx \left[\frac{X}{pKI_{est}(t)/N}\right] p\,K\,S_tI_t/N = XS_t\left(\frac{I_t}{I_{est}(t)}\right)$$

Finally, assuming that estimated infection numbers are to some degree realistic (that the

ratio $I_t/I_{est}(t)$ varies around 1) we can approximate the change in values of i at time t as

$$i_{t+1} - i_t \approx X\left[S_{t+1} - S_t\right] = -Xi_t$$

and the rate of change in i at time t is proportional to the value of i at that time. In the continuous case this this corresponds to

$$\frac{dt}{di} = -\frac{1}{Xi}$$

and given that some number of infections i has occurred, the amount of time infection numbers will remain in some region Δ around i will be proportional to Δ/i. Assuming that infection numbers are 'measured' at some constant rate and the true infection number has reached a value i some time since the last measurement, this means that the probability of obtaining a measured infection number in the region Δ around i will also be proportional to Δ/i; in other words, the conditional probability of recording a new infection count in the region Δ around i (given that there were i new infections at some time since the last measurement) is expected to follow a power law $p(i) \sim i^{-\lambda}$ with exponent $\lambda = 1$.

Given some fixed bin size Δ, let n_j be the number of recorded new infection counts i that fall into bin j (that is, where $\Delta j < i \leq \Delta(j+1)$). Similarly, let ϕ_Δ be the frequency distribution for values $n_j > 0$, so that $\phi_\Delta(n)$ gives the number of bins for which $n_j = n$. Since infection numbers in each bin in this distribution have occurred at least once, the probability of observing an infection i that falls into any one of these bins is approximated by the conditional probability $p(i) \sim i^{-\lambda}$ with $\lambda = 1$. If a variable's probability distribution follows a power law with exponent λ, then the associated frequency distribution ϕ will follow a power law with exponent $k = 1 + 1/\lambda$ (Adamic & Huberman, 2002; Hanel, Corominas-Murtra, Liu,

& Thurner, 2017); since probabilities $p(i)$ for these bins follow a power law with exponent $\lambda = 1$, our prediction is the frequency distribution ϕ_Δ will follow a power law with exponent $k = 1 + 1/\lambda = 2$. Note that, unlike our predictions about R and D_L (both of which describe long-run oscillatory behaviour and so are limited to the oscillatory domain), this power-law prediction is focused on the tail of high infection numbers (primarily caused by the arrival of new variants in the population), and thus holds generally, and not just in the oscillatory domain.

As before, this prediction can be assessed informally via measures of goodness-of-fit: by asking to what extent the frequency distribution for i (given a certain value of Δ) is fit by a power-law distribution with exponent 2. More formally, fitting a general power law to the frequency distribution of i for a given country c and letting k_c be the best-fitting exponent value obtained for that country and then taking M_4 to be mean value of k_c across a set of different countries, our hypothesis is that the 95% confidence interval for M_4 will contain the predicted value 2.

3.4. Methods

We tested these predictions about M_1, M_2, M_3 and M_4 using publicly available data from the Our World in Data COVID hub (Ritchie et al., 2020) for the period from the start of the pandemic up to June 30, 2022 (see 'Availability of Data and Materials'). This dataset gives the number of new Covid-19 infections reported each day for 231 countries under the variable name *new cases*, from the Johns Hopkins University Covid-19 Data Repository (Dong, Du, & Gardner, 2020), and a smoothed version of this measure under the variable name *new cases smoothed* (alongside a population-normalised measure *new cases smoothed*

per million). This dataset also gives the estimated reproduction number each day under the variable name *reproduction rate*, with estimation carried out using a Kalman Filter approach. (Arroyo-Marioli et al., 2021) Some countries in the dataset had no values associated with one or more of these variables on any day: we cleaned the dataset by removing all such countries, leaving data from 190 countries for analysis.

One problem with the OWID Covid-19 data arises because countries frequently reported 0 new case numbers on certain days: around 25% of new case numbers reported in the OWID dataset were 0, with many countries having reliable patterns of 0 new case numbers on weekend days only. These 0 values clearly do not reflect a change in infection numbers; instead, they simply indicate gaps in reporting. Derived measures such as reproduction rate and smoothed new case numbers are calculated from these reported values and so are similarly affected by these gaps, but to a lesser degree. In an attempt to avoid these gaps in our analysis we further clean the dataset by excluding, for each country, any day with 0 new case numbers reported for that country. Our analysis thus considers the evolution of infection numbers over consecutive reporting days, with reporting gaps removed.

For a given country c and day t we take $R_{t,c}$ to represent the value of the OWID *reproduction rate* variable for that country on that day. Taking h_c to be the highest value (the latest day) for which $R_{h-1,c} \leq 1 \leq R_{h,c}$ holds for country c, the oscillatory domain for that country is $t \leq h_c$ and our predictions concern the value and confidence intervals for M_1, M_2 and M_3 calculated from the cleaned dataset in that domain. There were 6 countries where $R_{t,c} < 1$ held for all reported days: these countries were excluded from analysis of oscillatory domain results, leaving 184 countries giving oscillatory domain data.

In assessing predictions about M_2 (the mean proportional change in new infection numbers, d_c) and M_3 (the mean of the absolute value of that change, $|d|_c$) we take i_t for a given country c to represent the OWID variable *new cases* for that country, and calculate d_c and $|d|_c$ from values of this variable in the oscillatory domain. Note that, while the OWID variable *new cases smoothed* gives a more accurate estimate of new infection numbers at a given time (because the smoothing process reduces the variability caused by reporting gaps), we cannot use this smoothed variable to assess the distribution of proportional changes in new infection numbers over time (because the smoothing process itself removes some proportional changes from the data, and so systematically alters this distribution).

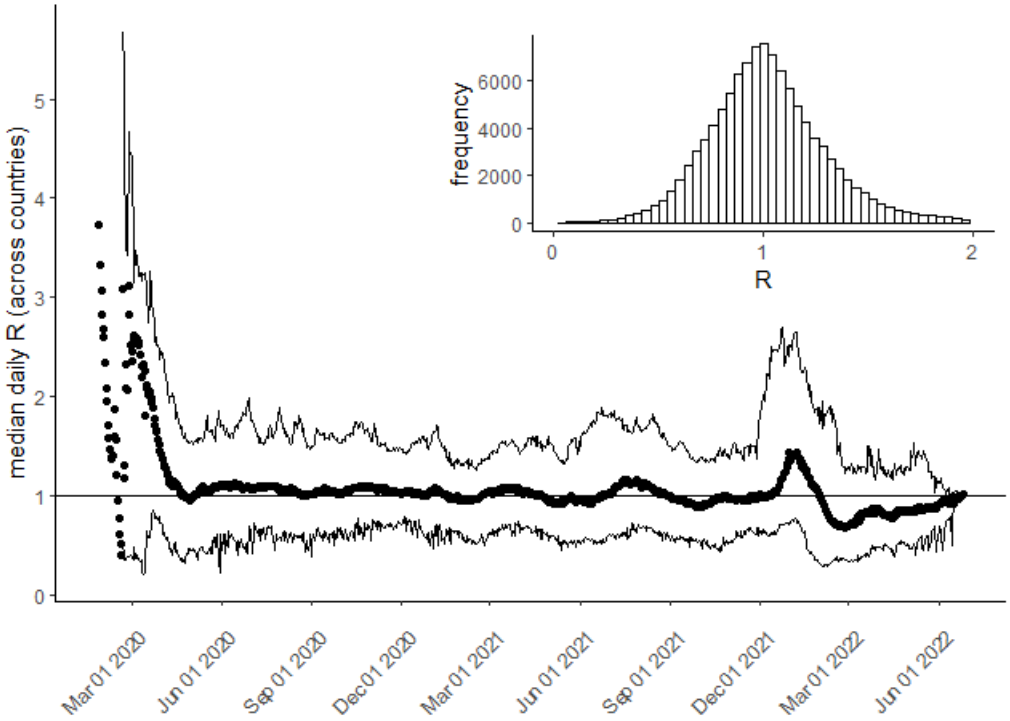


Figure 3.4.: Median of R values on each day (points) with 2.5% and 97.5% quantiles for country R values on each day (lines).

The inset shows a histogram of R values (bin size 0.1). There are no quantiles before 21 February 2020, because only R values for China are reported before that date.

In assessing our prediction about M_4 (the mean estimated power-law exponent k_c), we take i_t for a given country c to represent the OWID variable *smoothed new cases* for that country, and calculate k_c from the frequency distribution of this variable in the entire dataset (not just the oscillatory domain). We use the *new cases smoothed* variable for this analysis because that variable gives a more accurate estimate of new infection numbers, and because the smoothing process has no particular effect on exponent estimation.

Sine we do not know the statistical distributions for values of interest M_1,M_2,M_3 and M_4, we estimate confidence intervals for these values using both the assumption of normally distributed error (via a t-test) and using a standard non-parametric bootstrapping method. Our analysis of the power-law prediction M_4 applies to frequency data, which is produced by placing numbers i_t into bins of a certain size Δ. A central problem for such frequency analysis arises with the choice of bin size Δ: different choices for Δ will produce different numbers of bins for a given set of infection numbers i_t, making the resulting frequency data easier or harder to fit (depending on whether the number of bins obtained is small or large). We deal with this problem by automatically setting a bin size Δ_c for each country so that each country's infection number data falls into the same number of bins B, where B is the largest number such that every country's data can be placed into at least B distinct bins. This procedure ensures that power-law fits to frequency data for different countries are not affected by artefacts arising from the choice of bin size.

All statistical and modelling analysis was carried out `RStudio` using `R` version $4.0.5$ (R Core Team, 2021), using packages `data.table`, `ggplot2`, `lubridate` and `patchwork` for general analysis and graphing (Dowle & Srinivasan, 2021; Wickham, 2016; Grolemund & Wickham, 2011; Pedersen, 2020) and using packages `qqplotr`(Almeida, Loy, & Hofmann,

2018) for quantile-quantile plots, `nptest`(Helwig, 2021) for non-parametric confidence intervals; and `poweRlaw`(Gillespie, 2015a) for power-law fits. A complete `R` script that implements the ASEIR model, downloads the OWID Covid-19 data, carries out all statistical and data analysis, and generates all figures reported here is available online (see 'Availability of Data and Materials').

3.5. Results

Figure 4 shows the distribution of R values for countries in the OWID dataset on each day t in the oscillatory region, with a histogram showing the frequency of individual R values. Both show R centered around 1, consistent with our first prediction.

To test this $R \sim 1$ prediction formally, we calculated the median $R_{t \leq h_c}$ value for each country c in the OWID dataset across the Covid-19 pandemic. The overall mean of these values was $M_1 = 1.0$ with a 95% confidence interval for the mean of $0.99 \ldots 1.01$ both when calculated via a t-test ($t = -0.34, df = 183, p = 0.73$) and via the non-parametric bootstrap estimate, confirming the prediction.

To assess our D_L predictions informally we compare D_L values calculated from the cleaned OWID dataset against the theoretical distribution $C(0, 1)$. For each country, we calculated $D_L(t)$ for every day in the dataset using the OWID *new cases* variable in that country's oscillatory region, for values of L from 7 (assuming the incubation period for Covid-19 is approximately 6) to 30 (assuming that the reporting of case numbers will not take more than one month). For some days $D_L(t)$ could not be calculated because one of the component infection numbers was missing or resulted in division by 0; these values of $D_L(t)$ were dropped from analysis. Figure 5 (inset) shows a probability-probability plot comparing the cumulative probability of D_L for $L = 7$ against that of C. Correlation of cumulative probabilities is

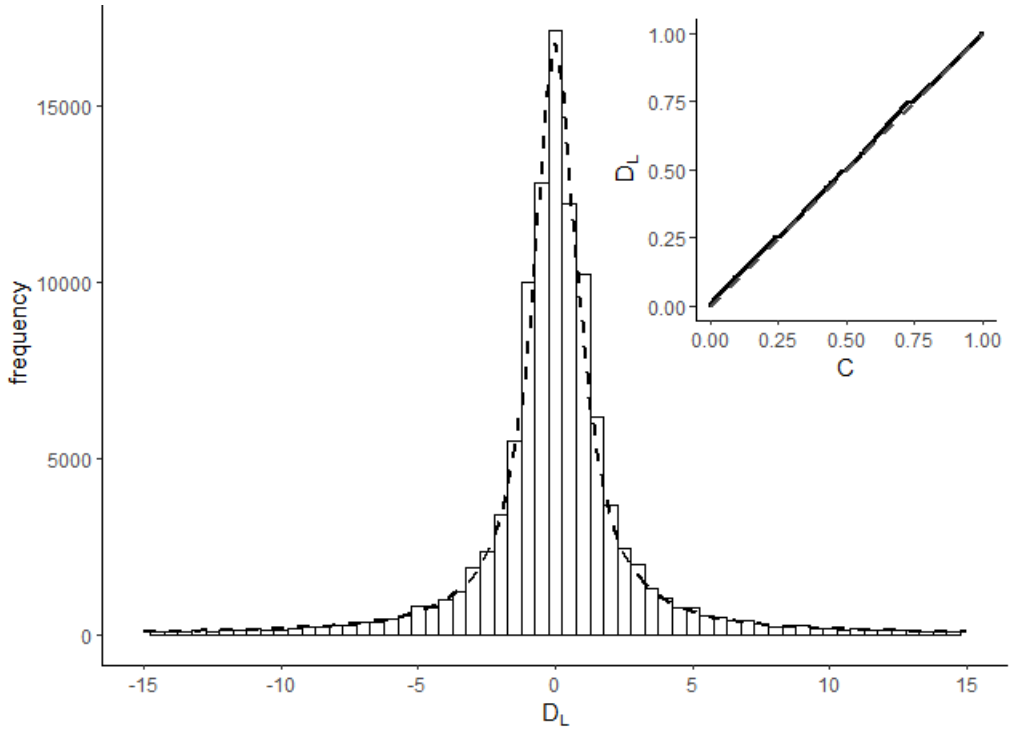


Figure 3.5.: Histogram of D_L for $L = 7$ calculated from the cleaned dataset in the central $-15 \ldots 15$ range with standard Cauchy distribution $C(0, 1)$.

The inset shows a probability-probability plot comparing theoretical and observed cumulative probabilities across the entire range: the solid line in that plot is actually made up of over $50,000$ points, one for each D_L value calculated in the dataset for $L = 7$: the dashed diagonal line (mostly hidden by these points) is the line of identity between theoretical and observed cumulative probabilities (bin size 0.5 with standard Cauchy distribution $C(0, 1)$ shown in dashed line, C distribution scaled by bin size and total histogram frequency for comparison).

a measure of goodness of fit between observed and theoretical values; here the correlation was high ($r = 0.999$). Since probability-probability plots overweight extreme values, we also analysed the relationship between $C(0, 1)$ and D_L for values near the midpoint of the range, by selecting the subset of D_L values between -15 and 15 (93% of the total sample). Figure 5 (main) shows a histogram of these values. The correlation between D_L and C values for this central-region histogram was $r = 0.99$. Similar results held for other values of L.

To test prediction M_2 formally we obtained, for each country c, location estimates d_c for values of L from 7 to 30 by calculating the median value of D_L for that country for each

value of L, and setting $M_2(c)$ to be the mean of these location estimates for that country across all values L. We took M_2 to be the average of these $M_2(c)$ values across all countries. The overall mean of these values was $M_2 = 0.01$ with a 95% confidence interval for the mean of $-0.01 \ldots 0.02$ when calculated via a t-test ($t = 1.15, df = 183, p = 0.25$) and the same confidence interval when calculated via the non-parametric bootstrap estimate. This confirms prediction 2.

We similarly obtained, for each country c, scale estimates $|d|_c$ for values of L from 7 to 30, by calculating the median of the absolute value of D_L for each value of L, and setting $M_3(c)$ to be the average of these scale estimates. The overall mean of these values was $M_3 = 1.08$ with a 95% confidence interval for the mean of $1.04 \ldots 1.1$ in both t-test and non-parametric analysis. While this is very close to the predicted scale estimate of $M_3 = 1.0$, the predicted value falls outside the calculated confidence interval, and so prediction M_3 is not confirmed.

The fact that the estimated scale parameter here is marginally higher than the predicted value (1.08 versus 1) could arise as a consequence of overextension of the oscillatory region for some countries: if the identified oscillatory region bound h_c for country c in fact included the initial rising section of an infection wave, values D_L in that region will be biased upwards by that wave, producing an increase in the scale estimate. As a post-hoc test of this proposal, we calculated for each country the number of days in the dataset outside the oscillatory region (days where $t > h_c$) and re-ran our analysis excluding any countries where this number of days was small. Excluding all countries where $|t > h_c| \leq 10$ give $M_3 = 1.04$ with a 95% confidence interval for the mean of $0.99 \ldots 1.08$, supporting prediction M_3.

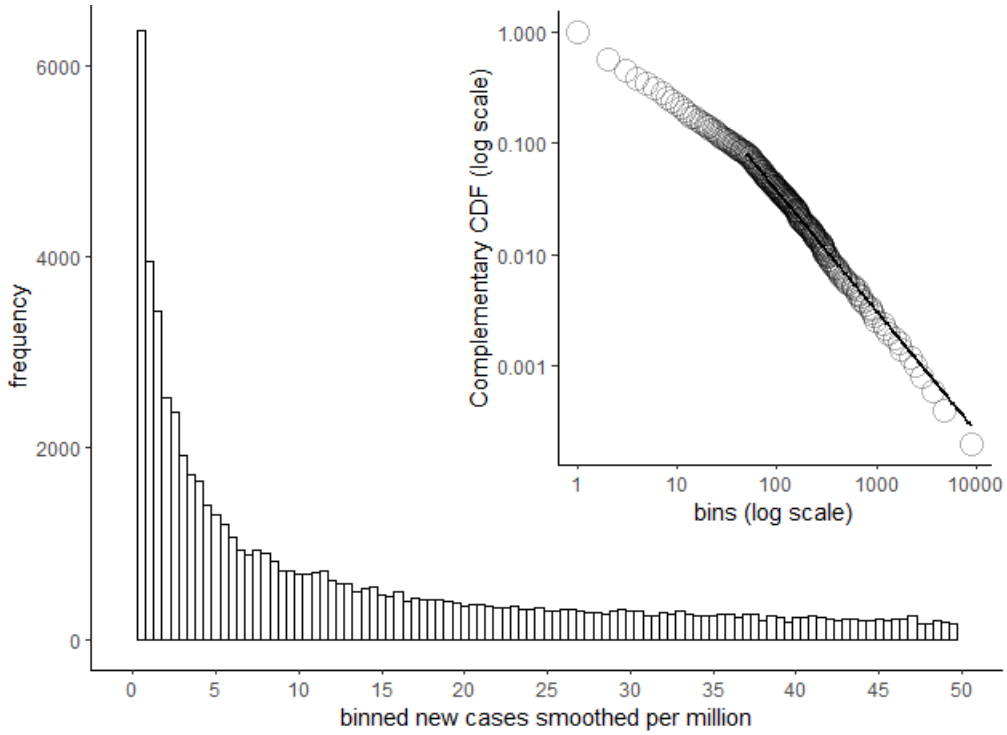


Figure 3.6.: Histogram showing the frequency of smoothed new cases per million (bin size 0.5) across all countries in the OWID dataset.

For illustrative purposes only the first 50 frequency bins are shown. The inset shows a plot of complementary cumulative probability (CCDF) across all bins: the solid line shows the theoretical CCDF value for the best fitting power law for this frequency data, with $k = 2.08$ and $x_{min} = 49$. The agreement between the solid line and the CCDF datapoints gives an informal illustration of the power-law fit.

To test our power-law prediction M_4 informally, we produced a frequency table of smoothed new cases per million across all countries in the OWID dataset, with a bin size of 0.5 (Figure 6). We found the best-fitting power law for this frequency data using the R powerLaw package. (Gillespie, 2015b) When fitting a power law to data, it is usually argued that only the tails of the distribution (greater than some value x_{min}) follow a power law; this assumption is explicit in the behavioural response account, where a power law is assumed to hold only for high new infection numbers. The powerLaw package returns the best-fitting x_{min} and k values for the given data; for the OWID data, the best fit was obtained with $k = 2.08$ and $x_{min} = 49$ (new infection numbers greater than 49 per million are best fit by a power law with $k \approx 2$). Figure 6 plots the frequency of smoothed new cases per million across all

countries in the OWID dataset in the first 50 of these bins. A standard way to assess power-law fits informally is via comparison of observed and theoretical 'complementary cumulative distribution functions' or CCDFs (Clauset, Shalizi, & Newman, 2009); the inset in Figure 6 plots the observed CCDF versus the theoretical CCDF predicted for this value of k. Note that the theoretical CCDF (solid line) starts at $x_{min} = 49$, and that there is a noticable 'turn' in the observed CCDF at that point. In the context of the ASEIR model, this point represents a transition to the 'high infection numbers' domain.

To test prediction M_4 formally, we first obtained, each country in the cleaned OWID dataset, the number B_c equal to the largest integer such that the full set of i data for country c can be placed into B_c equal-sized bins. We then set B equal to the minimum value of B_c across all countries, so that B is the largest number such that every country's data can be placed into at least B distinct bins. Given this B we then obtained, for each country c, the largest bin size Δ_c such that country's data will be placed into B bins, and using that bin size Δ_c produced a frequency table of smoothed new case numbers for that country. For each country we used the powerLaw package to find the best-fitting power law for that country's frequency table. Letting k_c be the best-fitting power law exponent for country c, we took M_4 to be the mean value of k_c across all countries. The overall mean of these values was $M_4 = 2.06$ with a 95% confidence interval for the mean of $1.97 \ldots 2.15$ when calculated via a t-test $(t = 1.33, df = 189, p = 0.18)$ and via the non-parametric bootstrap estimate. This confirms prediction 4.

3.6. Discussion and Conclusions

In this paper we've presented an extension of the standard SEIR compartmental model of infection to include spreading awareness of and behavioural response to infection risk. We've

shown that this model can naturally account for the effect of various disease variants arriving in a population over time and matches initial patterns of rapid decline and rebound in reproduction numbers for the Covid-19 pandemic for selected countries. To validate this model we derive various parameter-free numeric predictions from this approach; analysis of Covid-19 data at both aggregate (world) and individual country levels gives explicit confirmation for these predictions, validating the behavioural response approach to modelling infection spread, and demonstrating some striking statistical regularities in the distribution of infection numbers.

It is useful to specify the situations in which we expect these statistical regularities to hold. First, these results assume that a large proportion of the population will become aware of and respond to the risk of infection, and so apply to epidemic or pandemic situations only: we do not expect this model to describe infection spread in narrower outbreak situations. Second, this model depends on the assumption that people's estimates of infection risk at time t will reflect the number of new infections at some recent time $t - L$. This assumption holds for infections with short incubation and recovery periods; for infections where these periods are longer, this assumption doesn't hold. Third: this model makes the simplifying assumption that people are free to limit their number of contacts to match their acceptable level of risk. For some demographics this is not the case: people in poverty, for example, may be economically unable to limit their contacts in this way, and so will have an estimated risk of infection systematically above their acceptable risk level. Assuming that people's acceptable risk levels are well-calibrated, this predicts increased infections in such demographics relative to the population as a whole. (J. Patel et al., 2020; Little et al., 2021) Letting K_{min} represent the lowest possible average contact rate for the population as a whole given these constraints on contact numbers, then $R_{min} = pK_{min}/\gamma$ is

the minimum possible reproduction number, and if $R_{min} > 1$ then the disease will spread exponentially through the population irrespective of behavioural response; while if $R_{min} < 1$ then behavioural response will act to maintain $R \sim 1$ in the oscillatory period of the infection.

The model makes a number of other simplifying assumptions: no vaccination, perfect and lasting immunity after infection, no quarantining or reduction of contact numbers among infected individuals. More realistic (and so more complex) versions of the model can be constructed to include vaccination, waning immunity, and quarantine responses. However, the statistical regularities described above will necessarily hold in these more complex models just as in the simple model described above. This is because while vaccination, waning immunity and quarantine all have clear effects on infection risk, in the ASEIR model behavioural response to this risk will continue to act to maintain $R \sim 1$ (with increased vaccination numbers, for example, causing a reduction in both perceived infection risk and in infection numbers, and this reduction in risk causing a corresponding increase in contact numbers and so a subsequent rise in infections, thus maintaining R around 1). The effect of vaccination, in these more complex models, is to shorten the oscillatory period and increase progress towards herd immunity, while the effect of waning immunity is to lengthen the oscillatory period and postpone herd immunity. An important aim for future research is to test these predictions about the effects of vaccination programs and of reinfection rates against data on Covid-19.

Availability of Data and Materials

All data used in this analysis is publicly available online from the Our World In Data COVID hub `https://ourworldindata.org/coronavirus` in the combined data file `https://raw`

.githubusercontent.com/owid/Covid-19-data/master/public/data/owid-covid-data.csv. R code implementing the ASEIR model, downloading this data file and running all analyses is publicly available online from the Open Science Foundation repository https://osf.io/29ayn/.

4. Contact patterns model

Rita Howe
Jwenish Kumawat
Michelle Gilbert
Patricia Kearney
Ciara Carroll
Carla Perrotta
Claire Buckley
Fintan Costello

4.1. Background

Epidemiological models of behavioural response to infection risk assume that people respond to perceived increases in the risk of infection by reducing their number of contacts, and respond to perceived decreases in risk by increasing their number of contacts (Ajbar et al., 2021; Bukhari et al., 2020; Manrubia & Zanette, 2022; Steinegger et al., 2020, 2022; Tkachenko et al., 2021; Weitz et al., 2020; Costello, Watts, & Howe, Pre print). Capturing this 'behavioural immune response' (Funk et al., 2010; Verelst et al., 2016) has been described as the hard problem of epidemiology (Perra, 2021); understanding the feedback processes linking infection risk and behaviour would allow us to more accurately predict patterns of infection spread in a population, and to identify effective interventions. This study is intended to test various predictions of models of behavioural response to infection risk during the Covid-19 pandemic, by assessing the relationship between reported COVID infection numbers in Ireland over time and the number of close contacts between individuals over

time, as recorded in data gathered by the Irish Contact Management Programme (CMP) within the HSE as part of the pandemic response. The software used for this purpose was the CovidCare Tracker (CCT).

4.2. Variables

The national Irish Contact Tracing dataset collected by the CMP consists of a row for each case whose contacts are being traced; each row contains responses to questions about that individual. The Data Dictionary for this dataset is shown in Table C.1 in the appendices. Also available for this analysis is public data on Covid-19 infection in Ireland: e.g., new case numbers at each date, number of Covid-19 tests carried out, and so on. We take these from the Our World In Data COVID Hub dataset. The Data Dictionary for the variables used in this dataset shown in Table 4.2. Code to carry out all analysis will be written in R and will initially involve loading these two datasets and joining them on the TestDate==Date variables, so that for each Contact Tracing record we also have new case numbers, positivity rate, R number etc on the date of test, from the OWID dataset.

4.3. Research questions

This study is intended to test various general predictions of models of behavioural response to infection risk during the Covid-19 pandemic. In these models' people respond to perceived increases in the risk of infection by reducing their contact numbers and respond to perceived decreases in risk by increasing their contact numbers. Since decreases in contacts decrease the rate of infection, while increases in contact increase the rate of infection, these models predict homeostatic or 'set point' patterns in infection numbers (long periods where infection numbers 'plateau' or are approximately constant) and thus predict that estimated reproductive numbers will vary around R=1 (corresponding to constant infection numbers).

4.4. Predictions

Our general hypothesis is that the value of the “number of close contacts“ field in the Contact Tracing dataset (numberContacts) will vary inversely with the OWID "Infection numbers" field (newCases): when risk is high, number of reported close contacts will be lower, when risk is lower, number of reported close contacts will be higher. We also hypothesise that the level of perceived risk (and so response to infection numbers) will vary by age, by health status, and vaccination status. We expect that reported close contacts will vary over the course of the pandemic primarily as a function of infection risk (number of reported infections) at that time, with other factors (e.g., pandemic fatigue) having a secondary influence. These predictions apply to the general population. Since the risk and contact situation for health-care workers and for patients in the health-care system is different from that of the general population because (health-care workers and patients have less choice over their number of contacts and different levels of risk), we expect that these hypotheses will hold only in data excluding health-care staff and those who contracted infection in a health-care setting . More specifically, propose the following hypotheses (refer to Tables for variable descriptions):

H1 Letting AvgContacts(date) be the average number of close contacts reported on "date", we hypothesise that AvgContacts and newCases will be significantly negatively correlated across dates.

H2 Letting AvgContacts(date|group) be the average contact numbers for members of some subgroup on “date”, we hypothesise that AvgContacts(date| hasUnderlyingCondition=TRUE) $<$ AvgContacts(date| hasUnderlyingCondition=FALSE) will hold across dates and overall.

H3 We hypothesise that AvgContacts(date| age > 60) $<$ AvgContacts(date| age < 60) will hold across dates and overall (assuming that age > 60 is itself a risk factor).

H4 The correlation between AvgContacts(date|group) and newCases(date) will be stronger and more reliable for cases where hasUnderlyingCondition=TRUE or Age > 60 than for other cases.

H5 Contact numbers for members of high-risk groups will vary more than those for members of low-risk groups (measured in terms of range of variation for the same dates): Var(AvgContacts(date| hasUnderlyingCondition=TRUE)) $>$ Var(AvgContacts(date| hasUnderlyingCondition=FALSE)), with variance measured across dates.

H6 For a given "case boundary number" X, identifying dates Date1 and Date2 where newCases(Date1) $< X$ and newCases(Date2)$> 2X$ (new case numbers at least doubled between Date1 and Date2), AvgContacts(date1|group) $-$ AvgContacts(date2|group) (change in contact numbers between those dates for a given group) will be higher for groups hasUnderlyingCondition=TRUE and Age > 60 than for the complementary groups. The smaller the difference between Date1 and Date2, and the larger the value of the case boundary X, the greater the difference between groups.

H7 Matching by risk group we predict that contact rates will be higher among vaccinated individuals than non-vaccinated individuals: for example, that AvgContacts(date| Vaccinated=TRUE, age > 60) $>$ AvgContacts(date| Vaccinated=FALSE, age > 60). AvgContacts(date| Vaccinated=TRUE, hasUnderlyingCondition=TRUE) $>$ AvgContacts(date| Vaccinated=FALSE, hasUnderlyingCondition=TRUE). And so on for other groups.

4.4.1. Time periods with approximately constant infection numbers

There are various periods where the infection numbers in Ireland were approximately the same: from 2020/11/06 to 2020/12/20 (period A1) and from 2021/04/05 to 2021/06/29 (period A2), infection rates were constant at around 350 infections per day, while from

2020/10/09 to 2020/11/05 (period B1) 2021/01/29 to 2021/03/07 (period B2) and 2021/7/10 to 2021/8/9 (period B3) infection rates were between 500 and 1500 per day (average around 1000).

H8 : In these of time periods (with approximately constant risk due to contact) we predict that contact rates will be higher among vaccinated individuals than non-vaccinated individuals AvgContacts(date| Vaccinated=TRUE) $>$ AvgContacts(date| Vaccinated=FALSE) for dates in A1, A2, B1,B2,B3.

H9 Dividing the CT dataset into two periods where period1 is before vaccination roll-out and period2 is after widespread vaccination, we predict that the proportion of covid cases among older and higher risk individuals will be lower in period 2 than in period 1, but the proportion of covid cases among younger and lower risk individuals will be higher. Letting ProportionCases(period, group) be the proportion of contact tracing cases in "period" who are in a given group, we thus predict that ProportionCases(period2, hasUnderlyingCondition=FALSE) $>$ ProportionCases(period1, hasUnderlyingCondition=FALSE) will hold (and similarly for age < 60).

Sampling Plan

- Existing Data: This project will use existing CT and OWID datasets. Registration will take place before datasets are combined and analysed.
- Sample size: Over one million CT records.
- Stopping rule: All CT and OWID data available at the start date of the project will be used. Only this data will be used.

Design Plan

- Study type: Observational Study - Data is collected from study subjects that are not randomly assigned to a treatment.
- Blinding: No blinding is involved in this study.
- Study design: This study involves secondary analysis of existing health-care data.
- Randomization: no randomization is involved.

Analysis Plan Correlational analyses will use Pearson's r, with variables paired by date. Group comparisons involve comparisons of average contact numbers in groups, paired by date. Since we assume that the number of individual contact tracing records on each date will be relatively large, we assume normally distributed error (central limit theorem) and so use paired t-tests for comparisons. Dates on which contact numbers in a given group are less than N=100 will be excluded from analysis. Hypothesis tests will be one-sided, in the predicted direction. For comparisons for which no groups with N>100 are available, non-parametric tests will be carried out. We will conduct our confirmatory analysis strictly according to this analysis plan, using scripts written in R version 4.0.2.

4.4.2. Exploratory analysis

Exploratory analyses may also be carried out: in reporting these will be explicitly labelled as exploratory and distinguished from the confirmatory tests of the hypotheses listed above. Possible analyses include: To what extent do these patterns hold for health-care workers? How do they change their contacts with changing risk in the population? Do their contact numbers change at all? Similarly, do these patterns hold for people under 20? Is there any difference in close contact numbers between periods of remote schooling versus periods of face-to-face school? (Note that no tests of behavioural response model predictions will be carried out in these exploratory analyses).

Table 4.1.: Variables selected for the analysis from the Covid Care Tracker data

Variable		
Description	Name	Type
Covid ID	ID	integer
Recorded Number of contacts	NumberContacts	integer
Date of Test	TestDate	date
Did patient have symptoms	hasSymptoms	boolean
Date of Symptom onset	symptomOnsetDate	date
Date of Contact Tracing completion	TracingCompletionDate	date
Age	Age	integer
Underlying condition	hasUnderlyingCondition	boolean
Did you receive vaccine?	Vaccinated	boolean
Number of doses?	vaccineDoses	integer

4.4.3. Data sources

This project will use existing Covid Care Tracker (CCT) and Our World in Data (OWID) datasets. Registration will take place before datasets are combined and analysed.

Sample size: Over one million CT records.

Stopping rule: All CT and OWID data available at the prior to January 2022 will be used. Only this data will be used.

Variables The CT dataset consists of a row for each infected individual whose contacts being traced; each row contains responses to a number of questions about that individual. That data available for this analysis contains the following variables for each row (each contact tracing case):

Also available for this analysis is public data on Covid-19 infection in Ireland: e.g., new case numbers at each date, number of Covid-19 tests carried out, and so on (from the Our World In Data COVID Hub dataset downloaded from URL 'https://raw.githubusercontent.com/owid/Covid-19-data/master/public/data/owid-covid-data.csv').

Analysis code will be written in R and will initially involve loading these two datasets and

Table 4.2.: OWID variables from the Ireland dataset

Variable		
Description	Name	Type
Date	Date	date
Ireland: Positivity rate on this date	positivityRate	real
Ireland: number of tests on this date	NumberTests	integer
Ireland: number of new cases on this date	newCases	integer
Ireland: smoothed number of new cases	smoothedNewCases	real
Ireland: Reproductive number	R	real

joining them on the TestDate==Date variables, so that for each CT record we also have new case numbers, positivity rate, R number etc on the date of test.

4.4.4. Progress to date

The following progress has been made with this portion of the research to date:
The required variables needed to test the hypotheses and exploratory analyses described in sections 4.4 and 4.4.2 have been identified within the CMP database and agreed upon with the CMP. Variables were minimized as much as possible to protect anonymity of subjects.

An ethics application was made to the HSE so that the data could be used for research purposes. This application was submitted in May 2022 and provisionally approved in July 2022, pending clarification on minor points.

Pre-registration of the hypotheses in section 4.4 was completed and data analysis is expected to being in Autumn 2022.

Part IV.

Vaccine Effectiveness

5. Vaccine effectiveness

Rita Howe
Mark Roe
Carolyn Ingram
Carla Perrotta

5.1. Introduction

Ireland initiated a phased vaccination campaign on the 28^{th} December 2020. Among the first cohorts vaccinated were health care workers, high-risk individuals (i.e., those with underlying medical conditions), and individuals ≥ 85 years of age. As the vaccination campaign progressed, different age cohorts were vaccinated in descending order. Vaccines used in the campaign were BNT162b2 (Pfizer–BioNTech) (71% of issued vaccines by September 1st), ChAdOx1 nCoV-19 (AstraZeneca) vaccine (17% issued vaccines), the mRNA-1273 (Moderna) vaccine (8%), and the Ad26.COV2.S (Johnson & Johnson–Janssen)(3%) vaccine (Health Service Executive, 2022). With this high rate of vaccination coverage, we propose that vaccine effectiveness (VE) rates can be investigated using population data on acute hospitalizations, ICU admissions and death due to Covid-19 and severe Covid-19 to (1) help policy makers determine national strategy and (2) provide evidence on VE to enhance vaccine confidence and public trust (M. K. Patel et al., 2021; Zheng et al., 2022).

Clinical trials have demonstrated vaccine efficacy of 95% (ARR: 0.84%, NNV: 119) for

Pfizer, 94% (ARR: 1.2%, NNV: 81) for Moderna, 67% (ARR: 1.3%, NNV: 78) for AstraZeneca, and 67% (ARR: 1.2%, NNV: 84) for Johnson & Johnson-Janssen (Polack et al., 2020; Baden et al., 2021; Voysey et al., 2021; Sadoff et al., 2021; US Food and Drug Administration, 2020; Olliaro, Torreele, & Vaillant, 2021). Studies on real world effectiveness have demonstrated a range of VE rates from $77\% - 95\%$, depending on a number of conditions including vaccine administered and time since vaccination, and a decline in effectiveness over time, facilitating the need for booster vaccines (Poukka et al., 2021; Glatman-Freedman, Bromberg, Dichtiar, Hershkovitz, & Keinan-Boker, 2021; Vitek et al., 2022). Evaluating how the vaccines perform at a nationwide level is an important step in a comprehensive analysis of vaccine effectiveness. To date, no analysis of Covid-19 VE in Ireland has been completed.

Global Covid-19 vaccine deployment has faced high uncertainty and complexity relating to effectiveness, risks for various age groups, duration of immunity, and repeated vaccinations; challenges which are aggravated by an infodemic and the rise of misinformation stemming from conditions of uncertainty (Mills, Rahal, Brazel, Yan, & Gieysztor, 2020). To overcome these obstacles and minimize Covid-19 morbidity and mortality, governments and public health officials need to understand the most cost-effective and equitable approaches to mass vaccination rollouts. Interim analyses from real-life mass vaccination campaigns are thus crucial and urgently needed to reinforce or improve current policies on widespread immunization (Flacco et al., 2021).

While vaccine coverage is high in Ireland, an estimated 9% of the eligible adult population remains resistant to a Covid-19 vaccine, with statistically significant increases in resistance reported over the course of the pandemic (Murphy et al., 2021; Hyland et al., 2021). Vaccine resistant individuals are more likely to harbour conspiracy beliefs, to have lower levels of

trust in scientists, healthcare professionals, and the state, and to consume more information from social media than formal information sources (Murphy et al., 2021). Addressing these barriers to vaccination willingness and uptake requires public health messaging that is clear, direct, repeated, positively oriented, and emphasises the personal benefits of vaccination against Covid-19 (Murphy et al., 2021). Thus, by reporting on and communicating Covid-19 vaccine effectiveness to the general public in Ireland, this study can help improve trust in the Covid-19 vaccine and its providers.

5.2. Methods

5.2.1. Study Aim and Data sources

This study aims to examine vaccine effectiveness for hospital admission, ICU admission, and mortality due to Covid-19 across time and age groups in Ireland using "real world" publicly available data. The study period covered five months from August 1st, 2021 (month 08 since the start of the vaccination campaign) to November 30th (month 11 of the vaccination campaign) and focused on the population over 12 years of age.

In Ireland, no data is currently available that can trace vaccination status, vaccine type, and subsequent outcomes, therefore vaccine effectiveness has to be estimated using population data. Vaccine effectiveness was determined using three measures: *Relative Risk Reduction*, using the formula RRR $= 1 - RR$, where RR is the proportion of vaccinated individuals hospitalised (or admitted to ICU or recorded as a death) due to Covid-19 to the proportion of unvaccinated individuals hospitalised (or admitted to ICU or recorded as a death) due to Covid-19. To avoid reporting bias, *Absolute risk reduction* (ARR) was calculated using the ratio of unvaccinated individuals that are hospitalised (or admitted to ICU or died) due to Covid-19 minus the ratio of vaccinated individuals hospitalised (or admitted to ICU or died)

due to Covid-19. Finally, *Numbers needed to vaccinate* (NNV) was calculated using $1/ARR$. Average VE, ARR, and NNV rates were calculated used weighted mean values,

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

ARR and NNV are important measures for real-world settings, as they demonstrate the practical consequences of vaccination and are important measures for public health policy (Olliaro et al., 2021). Where sample size was very small and error rates high, such as in the weekly estimates for some of the age cohorts, estimates are not presented.

The data used to determine VE estimates for the eligible population for August 1st to November 30th was obtained from the Health Protection Surveillance Centre (HSPC). The HSPC provide estimates of total number of fully vaccinated, partially vaccinated, or unvaccinated individuals that are admitted to hospital, ICU or die due to a Covid-19 infection for a given month. The HPSC calculate the number of vaccinated or unvaccinated individuals using the cumulative number of vaccinated individuals in the eligible population (> 12 years of age) by mid-point of each month (Health Protection Surveillance Centre, 2022).

The data used to estimate VE for different age cohorts was obtained from the Central Statistics Office (CSO) and the Health Service Executive (HSE). The CSO provides data on rates of hospitalisation and ICU admission for different age cohorts by vaccine status from September 1st - October 31st (Central Statistics Office, 2022). However, unlike the HPSC, the CSO data does not distinguish between fully and partially vaccinated individuals. The estimates of population size per age cohort was obtained from the CSO's population database (Central Statistics Office, 2022). Finally, an estimate of the total number of unvaccinated

and vaccinated individual in each age cohort was obtained from the HSE (Health Service Executive, 2022). Vaccine effectiveness (VE) was determined for five age cohorts (0 – 24, 25 – 44, 45 – 64, 65 - 79, and $\geq$ 80 years old). These cohorts were determined by the data available through the CSO, HPSC, and HSE at the time of analysis.

5.2.2. Case definition

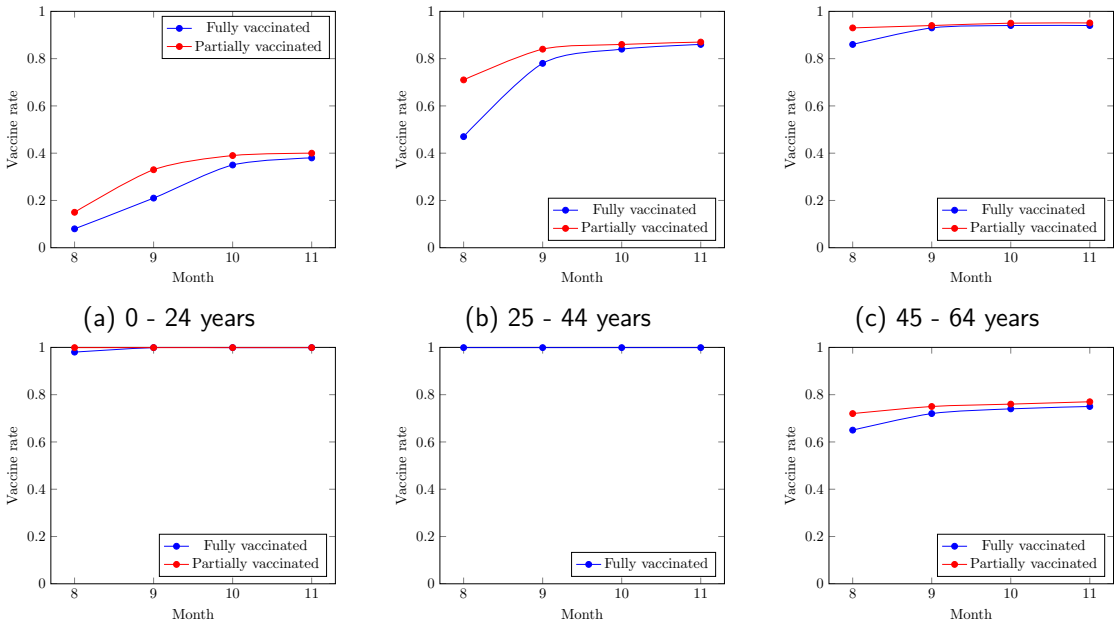
The criteria used by the HPSC and HSE to define a case was a laboratory confirmed PCR SARS-CoV-2 tests collected from oropharyngeal and nasopharyngeal swabs during the study period (August - November). Hospitalised individuals were new admissions to an acute hospital for a given week/month with a PCR-confirmed test. This criterion was the same for ICU admissions (Central Statistics Office, 2022).

5.2.3. Vaccination rates for age cohorts

The proportion of population vaccinated at a given date was found using vaccine rates $\geq$ 14 days before date of admission to hospital or date of admission to ICU. Individuals that had received *at least* one dose of any of the four available vaccines were considered vaccinated (see figure 5.1). The proportion of the population that was unvaccinated for a given date was calculated by determining the difference between the vaccination rates 14 days prior to that date and the CSO population estimate for that age cohort.

Approximately 100% of the population over 70 was fully vaccinated (two doses of the Pfizer or Moderna vaccine) in the period studied. As such, the denominators for the non-vaccinated groups in this age cohort were calculated in the following way: the CSO provides a year-on-year estimate for the current population by age in April of each year. The reasonable assumption that the population had increased since that estimate was made following examination of current population trends. On average, the population >70 had increased by

approximately 3% each year for the previous 10 years (Central Statistics Office, 2022). It is expected then, that the > 70 population will see a small, nominal increase of approximately 3% by April 2022. This was calculated and used for the estimate of population for the over 70s. The denominator for these cohorts was then calculated by finding the difference between the estimated number of fully vaccinated individuals > 70 and the 2022 estimated population.



(a) 0 - 24 years (b) 25 - 44 years (c) 45 - 64 years

(d) 65 - 79 years (e) 80+ years (f) Total Population

Figure 5.1.: Vaccine rate per cohort and for the total population (incl < 12) for August - November 2021

5.3. Results

5.3.1. Vaccine effectiveness over time

Over the course of August - November, VE against hospitalization for the fully vaccinated eligible population[1] fluctuated, rising from 73% (95% CI: $72\% - 79\%$) in August and peaking to 80% (95% CI: $77\% - 82\%$) in September. After this, VE declined again to 73% (95% CI: $68\% - 76\%$) in November. This may be a consequence of declining vaccination protection in

[1]VE estimates for the partially vaccinated population are available in table SD.1

Table 5.1.: Overall Vaccine Effectiveness against hospitalisation, severe illness and death, August - November 2021, Ireland

Month	VE	$(1-rr)$	95% CI	NNV	95% CI NNV	ARR
August	Hospital	0.76	0.72 - 0.79	2198	1,917 - 2,577	0.045%
	ICU	0.90	0.88 - 0.93	8712	6,887 - 11,857	0.011%
	Mortality	-	-	-	-	-
September	Hospital	0.80	0.77 - 0.82	1,342	1,189 - 1,540	0.075%
	ICU	0.91	0.91 - 0.95	5,010	4,079 - 6,493	0.019%
	Mortality	0.53	0.37 - 0.67	19,589	12,167 - 50,224	0.005%
October	Hospital	0.71	0.66 - 0.74	1,825	1,563 - 2,194	0.055%
	ICU	0.98	0.88 - 0.93	6,228	4,899 - 8,544	0.016%
	Mortality	0.51	0.27 - 0.67	25,248	14,504 - 97,418	0.004%
November	Hospital	0.73	0.68 - 0.76	1,508	1,296 - 1,802	0.066%
	ICU	0.91	0.88 - 0.93	4,749	3,779 - 6,388	0.021%
	Mortality	0.59	0.36 - 0.74	23,457	13,682 - 82,140	0.0043%
Weighted Mean per month	Hospital	0.74	0.70 - 0.77	1,718	1,655 - 1,786	0.058%
	ICU	0.91	0.87 - 0.94	6,123	5,756 - 6,541	0.016%
	Mortality†	0.46	0.18 - 0.65	32,849	27,095 - 41,706	0.003%

† Calculations were based on values up for September 1st - November 30th, 2021.

members of the older and at-risk cohorts, many of whom had been fully vaccinated for over 6 months at the time of study. ARR and NNV rates against hospitalisation also fluctuated during this period, with ARR rising from 0.05% in August to 0.07% in November (Table 5.1).

VE against critical illness (ICU admissions) was consistently high, with effectiveness of 87% or greater during this time. ARR rates remained stable during September - November (ARR: $0.01\% - 0.02\%$). NNV rates were highest in August (NNV: $8,713$, $95\% CI : 6,887 - 11,857$) but were consistently lower after this (NNV: $4,749 - 6,228$).

Significant relative risk reduction against death was observed in the vaccinated cohort. In September, the relative risk reduction was 53% ($95\% CI : 37\% - 67\%$), in October, it was 51% ($95\% CI : 27\% - 67\%$), and in November, a reduction of 59% ($95\% CI : 36\% - 74\%$) was observed.

5.3.2. Vaccine effectiveness by age

The VE rates against hospitalisation were highest for 25 - 79-year-olds, with a VE of 88% - 90% observed for these cohorts (Table 5.2). Marginally lower estimates of VE were observed for the ≥ 80 group, here the VE was 73% ($95\%CI : 59\% - 82\%$). A lower relative risk reduction of 47% ($95\%CI : 27\% - 62\%$) was found for the $0 - 24$ cohort. ARR was lowest and NNV was highest for the $\leq$ 45 cohorts, while similar ARR and NNV scores were observed for the $65-79$ (ARR: 0.85% NNV: 117) and ≥ 80 cohorts (ARR: 0.62%, NNV: 161).

The relative risk reduction for critical illness (ICU admission) was high for all cohorts[2], with a VE of $\geq 95\%$ observed in each case (Table 5.2). ARR scores were highest for the $65 - 79$ cohort (ARR: 0.4%) and declined for the younger cohorts (ARR $25 - 44$: 0.2%, ARR $45 - 64$: 0.07%). Equally, the NNV rates were lowest for the $65 - 79$ cohort (NNV: 241, $95\%CI : 181 - 359$) and increased for the $25 - 44$ cohort (NNV: 4,872, $95\%CI : 3,728 - 7,029$).

In figure 5.2, it is observed that the VE rates remain highly stable across the 9-week period, even as incidence rates fluctuate (Supplementary figure SF.1). This is particularly true for the $45 - 79$ cohorts, where a maximum change of 2% was observed. The $0 - 24$ cohort saw a rise in VE over the period from 19% to 55% at the end of October, consistent with the increased vaccination rates among that group. The weekly VE estimate for the ≥ 80 cohort observed declined over the same period, from 77% to 64% protection against hospitalisation.

Weekly estimates of VE against critical illness (ICU) were not available for the $0 - 24$ and ≥ 80 cohorts due to small sample size and considerations of anonymity. For the cohorts that

[2]ICU data was not available for the $0 - 24$ cohort due to anonymity and material sensitivity

were available, the VE was both high and stable for that period (Figures 5.2f, 5.2g, 5.2h). Each of the cohorts, $25-44$, $45-64$, $65-79$, had high VE values of $\geq 95\%$.

Table 5.2.: Overall Vaccine Effectiveness against severe illness by age cohort, September - October 2021, Ireland

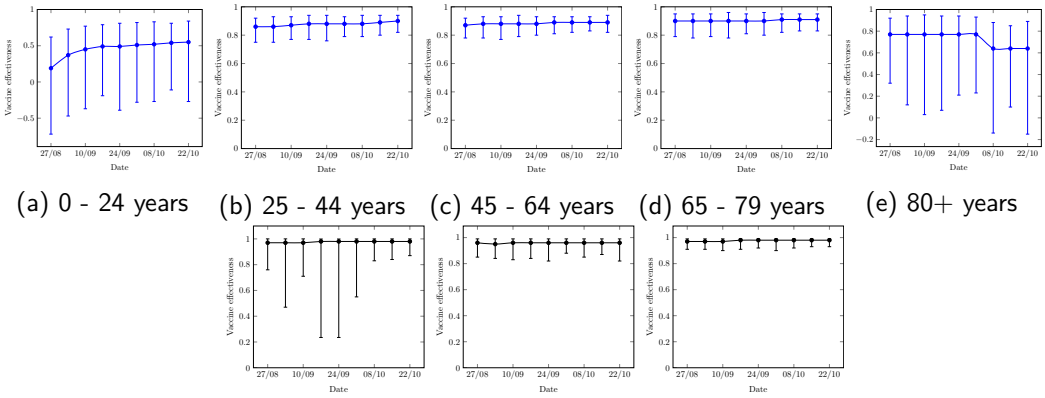
Age group	VE	$(1-rr)$	95% CI	ARR	NNV	95% CI NNV
0 - 24	Hospital	0.47	0.27 - 0.62	0.0075%	13,400	9,234 - 24,414
	ICU	-	-	-	-	-
25 - 44	Hospital	0.88	0.85 - 0.90	0.096%	1,042	906 - 1,228
	ICU	0.98	0.94 - 0.99	0.02%	4,872	3,728 - 7,029
45 - 64	Hospital	0.88	0.86 - 0.90	0.22%	447	383 - 537
	ICU	0.96	0.93 - 0.97	0.07%	1,482	1,146 - 2,095
65 - 79	Hospital	0.90	0.88 - 0.92	0.85%	117	95 - 154
	ICU	0.98	0.96 - 0.98	0.4%	241	181 - 359
80+	Hospital	0.73	0.59 - 0.82	0.62%	162	104 - 358
	ICU	0.95[†]	0.76 - 0.99	-	-	-

[†] Due to the relatively small sample size (< 10) for this cohort, the ICU results should be interpreted with caution.

5.3.3. Sensitivity Analyses

Across the age cohorts, the vaccine status of approximately 13% of hospitalised cases and 1% of ICU cases was unknown. The largest proportion of non-specified vaccination status was among the youngest age cohort. It is not expected then that this proportion of unknown cases will have a considerable impact on overall VE values for the other age cohorts.

Sensitivity analysis $\#1$ When the unspecified cases were classified as unvaccinated, no changes were observed to the ICU rates for any of the age cohorts, except the ≥ 80 cohort. Here, the VE rose from 95% to 97%. The hospitalisation VE rates increased for both the $0-24$ (47% to 60%) and ≥ 80 (73% to 90%) cohorts. The ARR and NNV values changed in-line with this increase in VE. For the $0-24$ cohort, the ARR rates rose from 0.0075% to 0.01% (NNV: $13,400$ to $7,968$) and for the ≥ 80 cohort, the ARR rates rose from 0.62%



(a) 0 - 24 years (b) 25 - 44 years (c) 45 - 64 years (d) 65 - 79 years (e) 80+ years

(f) 25 - 44 years (g) 45 - 64 years (h) 65 - 79 years

Figure 5.2.: Vaccine effectiveness against hospitalisation and ICU admission by age during September and October 2021.

Vaccine effectiveness against hospitalisation is shown in figures 5.2a - 5.2e. Vaccine effectiveness against ICU admission by age cohort is shown in figures 5.2f - 5.2h. Weekly ICU data was unavailable for the $0-24$ and $80+$ cohort.

to 2%[3] (NNV: 162 to 47). Marginal increases were observed in the $25-44$, $45-64$, and $65-79$ cohorts, with VE increasing between 2 - 5 percentage points for these cohorts. Small increases in ARR rates and decreases in NNV rates were observed for these cohorts (Appendix table E.1).

Sensitivity analysis $\#2$ When the unspecified cases were classified as vaccinated, the only cohort which saw a change to ICU VE was ≥ 80 which fell from 95% to 92%. For VE against hospitalisation, the greatest change was observed for the $0-24$ cohort, which fell from 47% to -11%. In the other cases, changes to the hospitalisation rates were minimal. VE rates decreased approximately from 1 to 7 percentage points depending on the cohort, and no significant change was observed to ARR and NNV rates (Appendix table E.2).

For both sensitivity analyses, the cohorts where change in rates was most likely to be observed were the $0-24$ and ≥ 80 age cohorts. This can likely be attributed to two reasons:

[3] Note: this value is higher than those observed in clinical trials and may be an artefact of uncertainty over the number on unvaccinated individuals ≥ 80 in the population at the time of the study.

1. In both cases, and particularly for the ICU rates, the sample sizes were smaller than the other cohorts, making these two groups more sensitive to numerator changes.

2. At the time of the study, vaccines had not been approved for individuals < 12 years of age. These individuals constitute a large proportion of the $0-24$ cohort (approximately 50%). However, these could not be excluded from the analyses as the information to do so was unavailable at that time of this study.

It is unlikely that all the non-specified cases would be either vaccinated or unvaccinated, however, age differences in the distribution of unspecified cases are expected to be observed. The $0 - 24$ cohort had much lower rates of vaccination than the other cohorts at the time of the study, so it is unrealistic that all or the majority of unspecified cases in the group were vaccinated. In contrast, approximately 100% of the ≥ 80 cohort were fully vaccinated, so it seems equally unrealistic that all the unspecified patients were unvaccinated. For both sensitivity analyses, the VE change was approximately proportional from the primary analysis. For analysis $\#1$, average VE change for these cohorts was $+4$ percentage points. For analysis $\#2$, average VE change for these cohorts was -3.3 percentage points. As such, the primary analysis is appropriate to estimate the rates of VE for illness in the vaccinated and unvaccinated groups.

5.4. Comparisons with empirical studies

One of the primary aims of this study was to explore whether surveillance data could be used as an effective proxy for an empirical study under a certain set of conditions. To this end, the results of a number of other vaccine effectiveness studies (all empirical) were compared to the results found in this study in two areas: comparison for reported effectiveness against disease severity and comparison of reported effectiveness by age.

5.4.1. Comparison by outcome

By outcome, the vaccine effectiveness against hospital admissions was lower than the *vaccine efficacy* clinical trials for Pfizer and Moderna but higher than that for AstraZeneca and Johnson & Johnson-Jansen vaccines. As all four of those vaccines were used for the population of Ireland, but primarily the Pfizer and Moderna vaccines, this is not a surprising result. For severe or critical illness, the VE values were in-line with those of Pfizer and Moderna (see table 5.3.

The VE values we observed are consistent with those seen in other empirical studies for both hospitalisation and severe illness, as shown in table 5.5.

None of the empirical studies calculated NNV or ARR values, so direct comparison is not available to us. Against the clinical trials, we not that the NNV rates are much higher and ARR rates are much lower than expected.

5.4.2. Comparison by age

A clearer picture of vaccine performance is shown when comparing by age group. All of the cohorts except the $0-24$ cohort had VE values consistent with the FDA clinical results (see table 5.4). This was also the case when comparing our study to the empirical studies. Again, the VE values reported were consistent with what we observed, aside from the $0-24$ cohort (see table 5.6).

For the age cohorts greater than 45, the NNV and ARR values are consistent with the FDA clinical trials, but again, the NNV and ARR values from the empirical studies were not available for a full comparison.

Table 5.3.: Comparison with FDA vaccine efficacy results (Overall VE)

Outcome	Study	Vaccine type	RRR	NNV	ARR
Hospital admission	FDA	Pfizer	95%	119	0.84%
		Moderna	94%	81	1.2%
		AstraZeneca	67%	78	1.3%
		J & J	67%	84	1.2%
	Current study	-	74%	1,718	0.058%
Severe or Critical illness	Current study	-	91%	6,123	0.016%

Table 5.4.: Comparison with FDA vaccine efficacy results (Age)

Study	Vaccine	Study age	RRR	NNV	ARR
FDA	Pfizer	–	95%	119	0.84%
	Moderna	–	94%	81	1.2%
	AstraZeneca	–	67%	78	1.3%
	J & J	–	67%	84	1.2%
Current Study	–	$0-24$	47%	13,400	0.0075%
	–	$25-44$	88%	1,042	0.096%
	–	$45-64$	88%	447	0.22%
	–	$65-79$	90%	117	0.85%
	–	≥ 80	73%	162	0.62%

The table above compares the VE observed in our study to that of the values that the FDA report for the for vaccines used in Ireland.

Table 5.5.: Comparison with clinical trials (Hospitalisation)

Outcome	Study	Vaccine type	RRR	NNV	ARR
-	FDA	Pfizer	95%	119	0.84%
		Moderna	94%	81	1.2%
		AstraZeneca	67%	78	1.3%
		J & J	67%	84	1.2%
Hospital	Current study	-	74%	1,718	0.058%
	Poukka et al	-	88%	-	-
	Rosenberg et al	Pfizer	$91\% - 72\%$	-	-
	Rosenberg et al	Moderna	$97\% - 78\%$	-	-
	Rosenberg et al	J & J	$87\% - 69\%$	-	-
Severe or Critical illness	Current study	-	91%	6,123	0.016%
	Bernal et al	Pfizer	88%	-	-
	Bernal et al	AstraZeneca	67%	-	-
	Glatman-Freeman et al	Pfizer	$99\% - 77\%$	-	-

Table 5.6.: Comparison with clinical trials (Age)

Age range	Study	Vaccine	Study age	RRR
≤ 49	Current study	–	$0-24$	47%
	Current study	–	$25-44$	88%
	Rosenberg et al	Pfizer	$18-49$	$93\%-67\%$
	Rosenberg et al	Moderna	$18-49$	$96\%-77\%$
	Rosenberg et al	J & J	$18-49$	$84\%-69\%$
$50-64$	Current study	–	$45-64$	88%
	Rosenberg et al	Pfizer	$50-64$	$95\%-75\%$
	Rosenberg et al	Moderna	$50-64$	$97\%-82\%$
	Rosenberg et al	J & J	$50-64$	$86\%-75\%$
≥ 65	Current study	–	$65-79$	90%
	Current study	–	≥ 80	73%
	Nunez et al	–	$65-79$	94%
	Nunez et al	–	≥ 80	82%
	Rosenberg et al	Pfizer	≥ 65	$91\%-76\%$
	Rosenberg et al	Moderna	≥ 65	$96\%-83\%$
	Rosenberg et al	J & J	≥ 65	$81\%-69\%$

5.5. Discussion

In this study, we analysed the effect of vaccination in preventing infection and severe illness due to Covid-19 over time and for different age cohorts. When considering the total population, the results showed that vaccinations provide significant protection against both hospitalisation and critical illness outcomes. During the 4 months studied the vaccinated individuals had lower rates of severe illness, critical illness, or death than unvaccinated individuals over time.

VE rates against hospitalisation were, in general, high (with mild fluctuations) throughout the study period, with an average VE of 74% observed. Protection against serious illness was consistently high during the same period, an average VE of 91% was observed against ICU admissions. Additionally, fully vaccinated individuals were significantly less likely to be admitted to ICU or die due to Covid-19 than unvaccinated individuals (Average VE: 46%).

The age cohort with the highest VE and lowest number needed to treat was the 65 - 79 cohort, the majority of this group (approximately 70%) had received their 2nd vaccination three - four months prior to September thus this demonstrated a higher protection in the three months posterior -for the delta variant. The lowest VE was observed for the youngest cohort, which had the lowest vaccination rates. VE improved substantially as vaccination coverage increased in this group. The three cohorts comprised of individuals between $25-79$ years, overall, had high and steady levels of VE. The oldest cohort had declining levels of VE for hospitalisations. This group was approaching 6 months vaccination at the time of the study, and waning antibodies may have contributed to the decline in VE. Strikingly, the decline that is apparent for hospitalisations, is not observed in the ICU data, where VE for the oldest cohort remained high and stable, though this sample is small.

ARR and NNV calculated rates were generally lower than those observed for clinical trials, however, in some age cohorts, ARR and NNV rates were close to the clinical rates. For instance, the estimated vaccine effectiveness for $65-79$ cohort during September and October, was 90% (ARR: 0.85%, NNV: 117), this is consistent with the clinical efficacy observed for the pfizer vaccine (VE: 95%, ARR: 0.84%, NNV: 119). In general, the rates were variable between the different cohorts, with the highest ARR and lowest NNV rates among the older cohorts. The youngest cohort, which had the lowest vaccine coverage rates, consistently had the lowest ARR and highest NNV rates. From a cost-effective perspective, the youngest cohort appeared to have the least effectiveness with the applied vaccination strategy. The ARR and NNV rates for hospitalisation, ICU admission, and death had small degrees of fluctuation over time.

As this study is reliant on surveillance data, there are a number of limits to highlight.

Firstly, the data on vaccination status does not include information required to determine which vaccine an infected individual received. Due to this, VE could not be calculated for vaccine type. This is particularly important to know for policy decisions. Secondly, due to issues surrounding the granularity of the data, the analysis of the $0-24$ cohort included case numbers from the < 12 population. At the time of the study, this cohort was not eligible for vaccination but could not be excluded. All values from this cohort must be interpreted with extreme caution.

Finally, more extensive data on the fully vaccinated cohorts prior to September was not available at the time of the study. More extensive data coverage on age cohorts and their outcomes, data on previous infections among vaccinated individuals, and further information on vaccination status would allow us to calculate adjusted-VE values and provide more in-depth analysis on the effectiveness of the vaccines used over time and by age. While these limitations exist, the method applied here is a novel way to measure vaccine effectiveness which population coverage is high and is an appropriate method when data is restricted or minimal.

5.6. Conclusions

Vaccines are effective at reducing both rates of Covid-19 illness and at reducing complications arising from that illness. Vaccinated cohorts were less likely to become ill than the unvaccinated cohorts for all age groups, with the greatest effectiveness in the older age cohorts. Vaccinated cohorts were also significantly less likely to suffer from severe illness following a Covid-19 infection. Public data is too limited to allow in-depth analysis of VE according to vaccine type, additional age-groups, and exact time periods following vaccination. Future surveillance efforts and research projects should address these limitations to better inform

public health policies.

Part V.

Summations

6. Data Access

6.1. Timeline of data requests

Channel 1: Health Protection Surveillance Centre

- December 2020: Initial CIDR data access form submitted to HPSC
- February 5, 2021: Received email that decision from the National CIDR Peer Review expected the week of 15th February
- April 10, 2021: Request for revisions received from the CIDR Peer Review group
- April - August 2021: We proceeded with our request for data through the Research Data Governance Board process for accessing CIDR within the CSO Covid-19 Research Data Hub (this process had stalled as of July 2021, and we recommenced with our CIDR data access request submitted directly to the HPSC).
- August 31, 2021: Revised CIDR data access form submitted to HPSC
- October 2021: No response received regarding the submitted revisions. Decision to pause this process as access to the CSO Data Hub had been granted.
- January 2022: Some required data is not available through RDP, recommence HSPC process

- April 2022: Request for CIDR data on specific outbreaks to cover data required for modelling submitted to HSPC

Channel 2: Research Data Governance Board

- April 12, 2021: Application submitted
- April 15, 2021: Preliminary review of application returned with queries
- May 06, 2021: Conditional approval granted dependent upon response to second round of queries
- May 27, 2021: Response to second round of queries submitted
- November 17, 2021: Access to requested RDGB datasets granted
- December 10, 2021: Researcher accounts added to CSO RDP

Channel 3: UCD School of Veterinary Science UPCOM project

- February 2022: Contact UCD School of Veterinary Science UPCOM team who have collected outbreak data from MPPs about possibility of collaboration
- February 2022: Collaboration agreed upon
- March - May 2022: Data received from VIs on-site at MPPs countrywide

Channel 4: Contact Management Programme

- February 2022: Contacted the CMP directly to request contact tracing data
- March 2022: CMP data request process is too lengthy for the duration of this project, to shorten this, the UCD and CMP teams agree that the analyses will be done "in house" by the CMP team using *R* code provided by the UCD team. The UCD team will guide and advise on the appropriate analyses

- March - April 2022: Data request is finalised
- May 2022: Ethics application is submitted to HSE so data can be used for research purposes
- July 2022: Ethics application is provisionally approved, with minor clarifications required, and re-submitted

The following permissions were required to gain access to the CSO Datasets accessed through the RDGB: Health Research Consent Declaration Committee

- March 15, 2021: Application submitted
- June 2, 2021: Preliminary review of application returned with queries
- June 22, 2021: HRCDC application approved

UCD Ethics Committee

- April 01, 2021: Application to the Human Research Ethics Committee – Sciences (HREC – LS) submitted
- April 16, 2021: Preliminary review of application returned with queries
- May 21, 2021: Response to queries submitted
- May 25, 2021: Conditional approval granted dependent upon response to second round of queries
- May 25, 2021: Response to second round of queries submitted
- May 28, 2021: HREC – LS final approval granted

Data Protection Impacts Assessment:

- March 16, 2021: DPIA submitted to the University College Dublin Data Protection Officer for review
- March 16, 2021: DPIA approved by UCD DPO

7. Limitations

The primary limitation of work package 1 (and its outputs) was its reliance on outside sources of data such as those collected by government bodies during the pandemic. At the time of this study, it was not feasible for the researchers to collect this sensitive data on the scope and scale required for the models. This issue with outside data can be broadly characterised at two limitations: access and quality.

Access Two types of data were used during this study: publicly available anonymous surveillance data and confidential health data. The available surveillance data is extensive and easy to access. However, obtaining access to the required health data was a lengthy process as noted in section 6.1. For the MPP portion of the investigation, we required information on MPP outbreaks to test our model performance. For the contact model portion of the investigation, we required information on the number of contacts for a given positive Covid-19 case. To obtain this data, we applied for access to the CSO RDP databases so we could access the CIDR data which contained records of investigated outbreaks in Ireland and the CCT data on contact tracking information gather from individual cases. Unfortunately, the data in both databases was insufficient to complete our research and additional data had to be applied for through the HSPC and VIs (MPP study) and the CMP (Contact study).

Quality Due to the sensitive nature of the data, in some cases it was unclear exactly what

data had been collected and how complete each data variable was prior to gaining access to the data. While the CIDR records have an extensive amount of information available to researchers, the variables available are contingent on the data that has already been collected by the HSPC and other government bodies. Here, the problem lies in which data is chosen for collection and the CIDR data lacked key information required for use in the modelling process. For the CCT data, the issues arose around how variables were defined. For instance, in the case of outbreaks, contacts were recorded as $T - C$ where T is the number of total contacts and C is the number of contacts that have already completed contact tracing. In some cases, the recorded contacts in the database will not be consistent with the total contacts an individual had. Someone with many contacts could be recorded as having 0 contacts if all their contacts had already completed contact tracing and identifying these cases was not always possible with the information available in the CCT database.

8. Recommendations

The major challenge in completing this work package stemmed from data access. As described in chapters 7 and 6, the data access process was lengthy in nature and the data difficult to obtain. With this in mind, the bulk of our recommendations are regarding this issue.

In the process of this research, our researchers noted that much of the private data collected on this health emergency is fragmented between different government bodies. The HPSE collects much of the outbreak data, while the contact data is collected by the CMP and the vaccination data is collected by the HSE. All of this data should be available through the CSO databases. However, in reality, the databases were difficult, if not impossible, to link together. Given this, we propose the creation of a system to link all these disparate pieces of data together. This would facilitate:

- Creation of a system with greater utility to researchers, with greater flexibility to test and explore hypotheses
- Clearer procedures for data access, as all data would be stored together
- Increase potential research outputs, which researchers being able to trace the course and outcome of infections, and its relationship to outbreaks, vaccinations programmes and the success of such programmes

Another issue noted by our researchers was related to *which* data was chosen for collection by the different government bodies. In particular, some crucial information on the scale and severity of Covid-19 infections was not captured. In the case of outbreaks, data collected did not include information on potential numbers of exposed individuals. In workplaces, no record was made of the total number of employees. We recommend the creation of a "best practise" protocol for data collection for future health emergencies. This would achieve the following:

- Strengthen the monitoring of disease spread
- Establish the scale of the outbreak/infection spread
- Improve the identification of vulnerable facilities and workplaces
- Faster mobilisation of investigatory teams in future outbreak situations

Finally, publicly available surveillance data was also fragmented between the government bodies, and while easier to access, it was not consistently published or sometimes published in forms that made it difficult to extract and use for analysis. For instance, data on positivity rates for the individual counties (a useful metric to determine the variability in infection reporting) was published in NHPET reports. However, it was not consistently reported, and when it was, it was embedded in PDFs. In future, we propose that raw data used to create these reports are published alongside the reports in a format accessible to researchers, such as CSV files. This would:

- Allow researchers to access this data without having to request if from the government bodies
- Increase reporting transparency around reported figures, allowing the public a greater degree of trust in health measures undertaken by the government and health service

Appendices

A. Dummy data for MPP model

The table below displays an example of the type of data used for the MPP model in chapter 2. Due to issues surrounding data anonymity, only dummy data which mirrors but does not correspond to the real-world outbreak data is presented here. Community incidence rates were calculated as a 7-day incidence rate for up to 100 days before the outbreak, contingent on the length of time between the outbreak and the start of the pandemic in Ireland. The date corresponds to the first day in which a case of Covid-19 was confirmed in the MPP.

Table A.1.: Dummy data for MPP model

MPP ID	N staff	N infected	Date	County	Incidence rate
A	250	38	07/07/2021	County A	130.89
B	300	2	14/12/2020	County G	45.23
C	82	50	22/12/2020	County H	2002.21
D	100	17	26/10/2021	County B	273.13
E	200	39	11/05/2021	County D	66.70
F	400	20	02/06/2020	County C	23.82
G	250	70	01/11/2021	County E	400.56
H	150	100	04/05/2020	County F	3700.78

B. Code for MPP analysis

The *R* code below was used to test the model described in chapter 2, using data collected from MPP outbreaks and community infection rates.

```
library(data.table)
library(lubridate)
library(ggplot2)
library(readxl)
```

```
## model

# assumes that the list of community infection probabilities is
    in order from
# day outbreak_day - length(community) (first entry) up to
   outbreak_day (last entry)
```

```
RSEIR <- function(beta, alpha ,gamma,N, community){
  SEIR_W <- function(t,S,E,I){
    S_t <- S
    E_t <- E
    I_t <- I
    t_val <- t

    repeat({
      i <- beta*S_t*I_t/N
      S_t <- S_t-i
      E_t <- E_t+i-alpha*E_t
      I_t <- I_t + alpha*E_t -gamma*I_t
      t_val <- t_val + 1
      if( t_val > length(community) || community[t_val-1] > 0){
          break }
    })
    if(t_val >  length(community) ){
      dist <- rep(0,N+1)
      dist[E_t+I_t] <- 1
      return(dist)
    }
    prob <-community[t_val-1]
    n <- round(S_t,0)
    dist_weighted <- function(k){
      binom_density <- dbinom(k,n,prob)
```

```
        if (k <= n & binom_density> sqrt(.Machine$double.eps) ) {
          i
          return(   SEIR_W(t_val,S_t-k,E_t+k,I_t)*binom_density  )
        } else {
          return(rep(0,N+1))
        }
      }
      dist <- Reduce("+",mapply(dist_weighted,seq(0,n),SIMPLIFY=
         FALSE))
      return(dist/sum(dist))
    }
    final_dist <- SEIR_W(1,N,0,0)
    data.table(infections=seq(0,length(final_dist)-1),probability
       =final_dist)
}



first_outbreaks_file <- 'MPcol4.csv'
first_outbreaks_data  <- read.csv(first_outbreaks_file,header=T
    ,fill=TRUE)
setDT(first_outbreaks_data)
colnames(first_outbreaks_data) <- c("Day","X7.day.rate","
    outbreak_id","Outbreak.date","Infections","Employees","
    Dailytests","positive","prate")
```

```
first_outbreaks_first_day<- first_outbreaks_data[ Day==0,]


mpp_file <- 'MPfullset.csv'
new_data  <- read.csv(mpp_file ,header=T, fill=TRUE)
setDT(new_data)


colnames(new_data) <- c("day","X7.day.rate","outbreak_id","
    Outbreak.date","Infections","Employees")
new_data$secondOutbreak <- !new_data$outbreak_id %in% unique(
    first_outbreaks_data$outbreak_id)




new_data$plant_id <- new_data$outbreak_id
new_data[outbreak_id > 1000,]$plant_id <- (new_data[outbreak_id
    > 1000,]$outbreak_id-100)/1000


outbreak_data <- data.table(new_data[day == 0,])
outbreak_data$Outbreak.date <- as.Date(dmy(outbreak_data$
    Outbreak.date))
outbreak_data$day <- NULL
```

```
outbreak_data$time_diff <- as.numeric(mapply(function(id)
    difftime(outbreak_data[plant_id == id & secondOutbreak ==
    TRUE]$Outbreak.date,
          outbreak_data[plant_id == id & secondOutbreak == FALSE
             ]$Outbreak.date,
          units = "days"), outbreak_data$plant_id))

outbreak_data[is.na(time_diff),]$time_diff <- Inf




# This assumes that new_data$X7.day.rate is the daily rate.
    What does the X7 mean here?

incident_rates_by_outbreak <- data.table(outbreak_id=new_data$
    outbreak_id,day=new_data$day,incidence=new_data$X7.day.rate)



summed_incident_rate_by_week <- function(week){
  day_numbers <- as.character(seq(0,6,1) + (week-1)*7)
  incident_rate <- aggregate(.~outbreak_id,incident_rates_by_
      outbreak[day %in% day_numbers], sum, na.action=na.pass)
  data.table(outbreak_id=incident_rate$outbreak_id,week=rep(
      week,length(incident_rate$day)),incidence = incident_rate$
```

```
    incidence)
}

# call that function for every week from 1 to 22 and then join
    the results together
weekly_community_incidence <- do.call(rbind,mapply(summed_
    incident_rate_by_week, seq(1,22,1),SIMPLIFY = FALSE))

# make a table with named week entries
weekly_community_incidence_named <- weekly_community_incidence
weekly_community_incidence_named$week <- paste("incidence_for_
    week",weekly_community_incidence_named$week,sep="")

# cast weekly incidence to columns
wide_weekly_community_incidence <- dcast(na.omit(weekly_
    community_incidence_named),outbreak_id ~ week,value.var="
    incidence")
# and add to outbreak data
outbreak_data <- outbreak_data[wide_weekly_community_incidence,
    on = .(outbreak_id)]


# exclude outbreaks where first and second outbreak are within
    2 months of each other
outbreak_data <- outbreak_data[time_diff > 60,]
```

```
# test correlation between infection numbers for outbreaks and
    weekly incidence rates for 1 week before outbreak,
# two weeks before outbreak etc (weeks are numbered "back" from
     outbreaks: incidence_for_week2 is the incidence rate
# two weeks prior to the outbreak
cor.test(outbreak_data$Infections,outbreak_data$incidence_for_
    week1)
cor.test(outbreak_data$Infections,outbreak_data$incidence_for_
    week2)
cor.test(outbreak_data$Infections,outbreak_data$incidence_for_
    week3)
cor.test(outbreak_data$Infections,outbreak_data$incidence_for_
    week4)

# do the same test for first outbreaks only and second
    outbreaks only
cor.test(outbreak_data[secondOutbreak==FALSE]$Infections,
    outbreak_data[secondOutbreak==FALSE]$incidence_for_week1)
cor.test(outbreak_data[secondOutbreak==FALSE]$Infections,
    outbreak_data[secondOutbreak==FALSE]$incidence_for_week2)
cor.test(outbreak_data[secondOutbreak==FALSE]$Infections,
    outbreak_data[secondOutbreak==FALSE]$incidence_for_week3)
cor.test(outbreak_data[secondOutbreak==FALSE]$Infections,
    outbreak_data[secondOutbreak==FALSE]$incidence_for_week4)
```

```
cor.test(outbreak_data[secondOutbreak==TRUE]$Infections,
    outbreak_data[secondOutbreak==TRUE]$incidence_for_week1)
cor.test(outbreak_data[secondOutbreak==TRUE]$Infections,
    outbreak_data[secondOutbreak==TRUE]$incidence_for_week2)
cor.test(outbreak_data[secondOutbreak==TRUE]$Infections,
    outbreak_data[secondOutbreak==TRUE]$incidence_for_week3)
cor.test(outbreak_data[secondOutbreak==TRUE]$Infections,
    outbreak_data[secondOutbreak==TRUE]$incidence_for_week4)
cor.test(outbreak_data[secondOutbreak==TRUE]$Infections,
    outbreak_data[secondOutbreak==TRUE]$incidence_for_week5)
cor.test(outbreak_data[secondOutbreak==TRUE]$Infections,
    outbreak_data[secondOutbreak==TRUE]$incidence_for_week6)
```

```
 periodic_incident_rate <- function(weeks){
  outbreak_incident_rate <- function(outbreak){
    get_week_incidence <- function(week_number){
      incident_rate <-weekly_community_incidence[outbreak_id==
          outbreak & week==week_number]$incidence
      probs <- c(rep(0,6),incident_rate)
      day_numbers <- seq(0,6,1) + (week_number-1)*7
```

```
        data.table(outbreak_id=rep(outbreak,7),day=day_numbers,
            incidence = probs)
    }
    do.call(rbind,mapply(get_week_incidence,weeks,SIMPLIFY =
        FALSE))
  }
}


periodic_community_incidence <- do.call(rbind,mapply(periodic_
    incident_rate(seq(1,22,1)),outbreak_data$outbreak_id,
    SIMPLIFY = FALSE))


# predict outbreak size


predicted_outbreak_size <- function(outbreak,beta,alpha,gamma,
    lag,end_day){
  cat("outbreak",outbreak,"\n")
    is_second_outbreak <- outbreak_data[outbreak_id==outbreak,]$
        secondOutbreak
    plant_size <- outbreak_data[outbreak_id==outbreak]$Employees
    community_for_plant = periodic_community_incidence[outbreak_
        id==outbreak & day < end_day,]$incidence
    community_for_plant[is.na(community_for_plant)] <- 0
```

```
    most_recent_week <- community_for_plant[1:7]
    community_for_plant <- c(rep(most_recent_week,lag),community
        _for_plant)
    com <-rev(community_for_plant)/100000
    plant_dist <- RSEIR(beta,alpha,gamma,plant_size,com)
    dist_size <- length(plant_dist$probability)
    data.table(outbreak_id=rep(outbreak,dist_size),
        secondOutbreak=rep(is_second_outbreak,dist_size),
        infections=plant_dist$infections,probability=plant_dist$
        probability)
}


# calculate means etc.

expected_incidence <- function(a,b,predictions){
  function(outbreak){
    probs <- predictions[outbreak_id==outbreak,]$probability
    cumulative <- data.frame(cumsum(probs))
    indx1 <- which(cumulative >=a )
    indx2 <- which(rev(cumulative) >= b )
    low <- indx1[1]
    if(is.na(low)) {low <- 1 }
    high <-indx2[1]
    if(is.na(high)) {high <- length(probs)-1 }
```

```
    is_second_outbreak <- unique(predictions[outbreak_id==
        outbreak,]$secondOutbreak)
    weighted_vals <-predictions[outbreak_id==outbreak,]$
        probability*predictions[outbreak_id==outbreak,]$
        infections
    predicted <- sum(weighted_vals)
    data.table(outbreak_id=outbreak,secondOutbreak=is_second_
        outbreak,predicted=predicted,low=low,high=high)
  }
}



# exclude outbreaks where the difference between the first and
# second outbreak was less than 2 months



# set default alpha, gamma and R values
alpha <- 1/6
gamma <- 1/6
R_first <- 3
R_second <- 3

# best fitting values for community infection list
end_day_first <- 7
```

```
lag_first <- 3


end_day_second <- 21
lag_second <- 0



# get predictions for outbreaks


get_predictions <- function(outbreak){
  if (outbreak_data[outbreak_id==outbreak]$secondOutbreak){
    predicted_outbreak_size(outbreak,R_second*gamma ,alpha,
        gamma,lag_second,end_day_second)
  } else {
    predicted_outbreak_size(outbreak,R_first*gamma ,alpha,gamma
        ,lag_first,end_day_first)
  }
}


# get predictions for all outbreaks


predictions <- do.call(rbind,mapply(get_predictions,outbreak_
    data$outbreak_id,SIMPLIFY=FALSE))


# get means for those predictions
```

```
CI <- 0.99
prediction_analysed <- do.call(rbind,mapply(expected_incidence(
    (1-CI)/2 ,1-(1-CI)/2    ,predictions),outbreak_data$outbreak
   _id,SIMPLIFY=FALSE))

length(prediction_analysed$outbreak_id)

prediction_results_full <- prediction_analysed[outbreak_data,on
   =.(outbreak_id)]

prediction_results <- prediction_results_full # use for all
   outbreaks
prediction_results <- prediction_results_full[Outbreak.date > "
   2020-05-01"]



# first outbreak results
cor(prediction_results[secondOutbreak==FALSE,]$predicted,
   prediction_results[secondOutbreak==FALSE,]$Infections)
sqrt(mean(prediction_results[secondOutbreak==FALSE,]$predicted-
   prediction_results[secondOutbreak==FALSE,]$Infections)^2)

# second outbreak results
cor(prediction_results[secondOutbreak==TRUE,]$predicted,
   prediction_results[secondOutbreak==TRUE,]$Infections)
```

```
sqrt(mean(prediction_results[secondOutbreak==TRUE,]$predicted-
    prediction_results[secondOutbreak==TRUE,]$Infections)^2)



# overall results
cor(prediction_results$predicted,prediction_results$Infections)
sqrt(mean(prediction_results$predicted-prediction_results$
    Infections)^2)

# draw figure

prediction_results$outbreak_label <- "first"
prediction_results[secondOutbreak==TRUE,]$outbreak_label <- "
    second"

ggplot(prediction_results,aes(x=predicted,y=Infections))+
  geom_point(size=3,aes(group=outbreak_label,shape=outbreak_
      label))+
  scale_shape_manual(values=c(16,1))+
  geom_smooth(method="lm",color="black")+
  guides(shape=guide_legend(title="Outbreak type"))+
  theme_bw()+theme_classic()
```

C. Data Dictionary for the Contact Tracing dataset

Table C.1.: Data Dictionary for the Contact Tracing dataset

Column	Variable Name	Variable description	Type	Options
A	ID	Unique Identifier	String	NA
B	NumberContacts	Number of Close contacts	Integer	NA
C	DateofTest	Test date	DateTime	NA
D	hasSymptoms	Symptomatic Status	Categorical	1 = Symptomatic 2 = Asymptomatic 3 = Unknown (Default: NULL)
E	symptomOnsetDate	Date of Symptom onset	Date	NA
F	Resolution	Contact tracing Resolution Status	Categorical	1 = Complete 2 = not required 3 = Unable to inform 4 = Patient Advised to self-isolate 5 = Self administered antigen
G	TracingCompletionDate	Date of contact	DateTime	NA

		tracing resolution		
H	Portal	TYC Portal Submitted	Boolean	
I	Age	Age	Integer	NA
				1 = Male
J	Gender	Gender	Categorical	2 = Female
				3 = Other
K	isHealthcareWorker	Health Care Worker	Boolean	NA
				1 = Yes
				2 = No
L	hasUnderlyingCondition	Underlying condition opt	Categorical	3 = Unknown
				4 = NULL
				1 = Yes
				2 = No
M	Vaccinated	vaccination history	Categorical	3 = Unknown
				4 = NULL

				1 = One
				2 = Two
N	VaccineDoses	vaccination doses	Categorical	3 = First Round
				4 = First Round + Booster
				5 = Unknown
				1 = Close Contact of a confirmed Case
				2 = Close contact with person who has symptoms suggestive of COVID
				3 = Healthcare setting acquired: patient
				4 = Healthcare setting acquired: staff
				5 = Travel related
				6 = Close contact of a known confirmed travel case
O	SourceTransmission	Most likely source	Categorical	7 = Associated with an outbreak
		of transmission		8 = Undergoing source investigation
				9 = Community transmission (only if none of above apply)
				10 = Travel related/close contact with a known confirmed travel case

11 = NA

12 = NULL

D. Vaccine effectiveness over time for partially vaccinated population >12

The data used to determine VE estimates for the eligible population for August 1st to November 30th was obtained from the Health Protection Surveillance Centre (HSPC). The HSPC provide estimates of total number of fully vaccinated, partially vaccinated, or unvaccinated individuals that are admitted to hospital, ICU or die due to a Covid-19 infection for a given month. Vaccine status was classified as partially vaccinated if (1) an individual had received one dose of a two-dose regimen and the epidemiological date is ≥ 14 days after receipt of this first dose or (2) an individual had completed the vaccine schedule and the epidemiological date is ≤ 14 days after receipt of the second (or final) dose.

Vaccine effectiveness was calculated using the Relative Risk Ratio $(1 - RR)$ of partially vaccinated individuals to unvaccinated individuals for hospital and ICU admission. ARR and NNV values were calculated as for the fully vaccinated individuals.

For partially vaccinated individuals, mortality rates were not calculated due to small sample size. The partially vaccinated values may also be subject to a degree of selection bias: by Autumn 2021, the majority of the eligible population had been fully vaccinated, and partially vaccinated individuals were likely to be from younger and low-risk cohorts, which had the lowest priority during the vaccination program. In contrast, the unvaccinated individuals may be distributed among any section of the population.

Table D.1.: Vaccine Effectiveness against hospitalisation and critical illness for partially vaccinated individuals, August - November 2021, Ireland

						Absolute risk reduction	
Month	VE	$(1-rr)$	95% CI	NNV	95% CI NNV	ARR	Case reduction per 100,000
August	Hospital	0.58	0.48 - 0.67	2,861	2,311 - 3,755	0.03%	35
	ICU	0.83	0.64 - 0.92	9,526	7,203 - 14,060	0.01%	11
September	Hospital	0.50	0.35 - 0.62	2,155	1,631 - 3,175	0.046%	46
	ICU	0.86	0.62 - 0.95	5,432	4,202 - 7,681	0.018%	18
October	Hospital	0.25	-0.05 - 0.48	-	-	-	
	ICU	0.63	-0.02 - 0.86	9,002	5,296 - 29,968	0.01%	11
November	Hospital	0.40	0.15 - 0.58	2,702	1,723 - 6,251	0.037%	37
	ICU	0.53	-0.03 - 0.78	8,179	4,557 - 39,828	0.012%	12
Weighted Mean	Hospital	0.40	0.23 - 0.54	3,127	2,182 - 5,516	0.032%	32
	ICU	0.76	0.47 - 0.89	7,304	5,274 - 11,875	0.014%	14

E. Sensitivity analysis for Vaccine Effectiveness

The primary analysis considered the vaccine effectiveness using reported rates of vaccination for hospitalisation and ICU admission for 5 age cohorts ($0-24$, $25-44$, $45-64$, $65-79$, $80+$). Cases were classified as "vaccinated", "unvaccinated", or "unknown". Approximately 13% of hospital admissions and 1% of ICU admissions were classified as "unknown". In the primary analysis, these "unknown" cases were excluded from the analysis.

To consider the impact of these cases on the analysis, two sensitivity analyses were performed. Sensitivity analysis $\#1$ considered the impact of all the "unknown" cases being unvaccinated, while sensitivity analyses $\#2$ considered the impact of all the "unknown" cases being vaccinated.

In sensitivity analysis $\#1$, all the unknown cases were reclassified as unvaccinated and the VE, ARR, and NNV values were recalculated. The results are shown in tables E.1.

In sensitivity analysis $\#2$, all the unknown cases were reclassified as vaccinated and the VE, ARR, and NNV values were recalculated. The results are shown in tables E.2 below.

Table E.1.: Vaccine Effectiveness against severe illness by age cohort, adjusted for vaccination status uncertainty (sensitivity analysis #1)

Age group	VE	$(1-rr)$	95% CI	ARR	NNV	95% CI NNV
0 - 24	Hospital	0.60	0.46 - 0.71	0.013%	7,968	6,179 - 11,215
	ICU	-	-	-	-	-
25 - 44	Hospital	0.91	0.89 - 0.93	0.14%	730	650 - 833
	ICU	0.98	0.94 - 0.99	0.0002	4,872	3,728 - 7,029
45 - 64	Hospital	0.92	0.90 - 0.93	0.3%	316	278 - 367
	ICU	0.96	0.94 - 0.97	0.0007	1,365	1,066 - 1,898
65 - 79	Hospital	0.95	0.94 - 0.96	1.7%	58	50 - 70
	ICU	0.98	0.96 - 0.98	0.4%	241	181 - 359
80+	Hospital	0.90	0.87 - 0.92	2.1%	47	37 - 63
	ICU	0.97	0.87 - 0.99	-	-	-

Table E.2.: Vaccine Effectiveness against severe illness by age cohort, adjusted for vaccination status uncertainty (sensitivity analysis #2)

Age group	VE	$(1-rr)$	95% CI	ARR	NNV	95% CI NNV
0 - 24	Hospital	-0.11	-0.42 - 0.13	-0.00017%	-56473	-16799 - 41473
	ICU	-	-	-	-	-
25 - 44	Hospital	0.81	0.77 - 0.84	0.09%	1126	968 - 1348
	ICU	0.98	0.94 - 0.99	0.02%	4,872	3,728 - 7,029
45 - 64	Hospital	0.86	0.84 - 0.89	0.2%	458	391 - 553
	ICU	0.96	0.93 - 0.97	0.07%	1,423	1,107- 1,999
65 - 79	Hospital	0.89	0.86 - 0.91	0.8%	119	96 - 157
	ICU	0.98	0.96 - 0.98	0.4%	241	181 - 359
80+	Hospital	0.70	0.55 - 0.80	0.6%	168	107 - 392
	ICU	0.92	0.54 - 0.99	-	-	-

F. Incidence rates for Vaccine Effectiveness

The National 14-day incidence rates are shown in Figure F.1. This period covered is part of Wave 4 (week 26 2021 – week 50 2021) of the SARS-CoV-2 pandemic in Ireland, where the Delta variant was predominant (Health Protection Surveillance Centre, 2022)

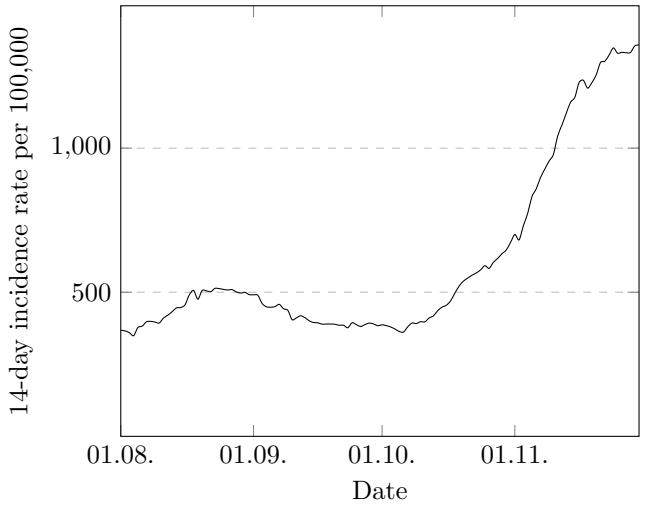


Figure F.1.: National 14-day incidence rates, August - November, 2021

G. CIDR data requested from CSO

Variables requested and approved

1. Core Dataset
    a) ID Variables (required to identify the data)
        i. COVID19 IDPIK
        ii. DOBSURNAME PIK
        iii. Event ID
        iv. Outbreak Identifier
        v. PATIENT IDPIK
    b) Demographic variables
        i. Age Years at time of event
        ii. Country
        iii. Country of Birth
        iv. Date of Birth
        v. Ethnicity
        vi. Gender Name
        vii. Patient Age at time of event

viii. Patient Type

c) Community rate variables

i. CCA Name

ii. CHO Area

iii. County

iv. Health Board Name

v. Hospital of Admission current

d) Infection timeline variables

i. Date of Death

ii. Date of Diagnosis

iii. Date of first admission

iv. Event Creation Date

v. Lab Reported Date

vi. Lab Specimen Collected Date

vii. Onset Date

e) Infection outcome variables

i. Interpreted Overall Lab Result

ii. Case Classification

iii. Outcome

f) Workplace variables

i. Occupation

g) Contact variables

i. Country of Infection

2. ENHANCED Dataset

a) Risk variables

i. BMI 40

ii. Cancer malignancy

iii. Chronic heart disease

iv. Chronic kidney disease

v. Chronic liver disease

vi. Chronic neurological disease

vii. Chronic respiratory disease

viii. Diabetes

ix. Hypertension

x. Other co-morbidity

xi. Is the case pregnant

xii. Smoking status

xiii. Underlying clinical conditions

b) Infection timeline variables

i. Date of admission to ICU

ii. Symptomatic

iii. Was the case admitted to ICU

iv. Date case placed in isolation

c) Workplace variables

i. Is the Case currently employed as a HCW

ii. If Yes HCW Role

iii. If other allied HCW please specify

d) Infection outcome variables

i. If recovered date of recovery

e) Contact variables

i. Most likely transmission source

f) ID variables

i. Event ID

3. HIU Dataset

a) ID variables

i. Event ID

b) Community rate variables

i. CHO

ii. CSO ED

iii. CSO ED Name

iv. LEA ID

v. LEA Name

4. ICU Dataset

a) ID variables

i. Event ID

b) Infection timeline

i. Date of ICU admission

ii. Length of stay in ICU days

5. OUTBREAKS Dataset

a) ID variables

i. Outbreak Identifier

b) Infection timeline variables

i. First-reported-date

c) Community rate variables

i. Outbreak CCA

ii. Outbreak CHO

iii. Outbreak Health Board

iv. Outbreak county

v. Outbreak location

d) Infection timeline variable

i. Outbreak created date

e) Infection rate variables

i. Outbreak extent

ii. Outbreak status

f) Infection outcome variables

i. Total dead

ii. Total hospitalised

iii. Total Ill

iv. Total Lab Investigated

H. CCT Data requested from CSO

Variables requested and approved

1. ADMISSIONSDISCHARGES dataset
    a) ID Variables (required to identify the data)
        i. COVID19 IDPIK
        ii. DOBSURNAME PIK
        iii. SWIFTQUEUE IDPIK
        iv. PATIENT IDPIK
    b) Demographic variables
        i. PatientDateOfBirth
        ii. GenderPatient
    c) Community rate variables
        i. Hospital
    d) Infection timeline variables
        i. AdmissionEpisodeStartDate
        ii. DischargeEpisodeEndDateT
    e) Infection outcome variables

i. PatientDateOfDeath

2. AMBULATORYASSESSMENTS

a) ID Variables

i. COVID19 IDPIK

ii. DOBSURNAME PIK

iii. SWIFTQUEUE IDPIK

iv. PATIENT IDPIK

b) Demographic variables

i. PatientDateOfBirth

ii. GenderPatient

c) Community rate variables

i. AddressCounty

ii. Eircode RoutingKey

d) Infection timeline variables

i. CreatedOn

e) Infection rate variables

i. Covid19TestUndertaken

f) Workplace variables

i. PatientIsACarer

g) Infection outcome variables

i. Ambul Assessment Outcome Plan

h) Risk variables

i. AreThereSignificantRiskFacto

ii. BMI

iii. PatientIsSociallyVulnerable

3. ASSESSMENTS Dataset

a) ID Variables

i. COVID19 IDPIK

ii. DOBSURNAME PIK

iii. SWIFTQUEUE IDPIK

iv. PATIENT IDPIK

b) Demographic variables

i. PatientDateOfBirth

ii. GenderPatient

iii. Gender

c) Infection timeline variables

i. Covid19VirologyResultDate

ii. CreatedOn

iii. DateOfFirstSymptoms

iv. DateOfHospitalAdmission

v. DateOfIcuAdmission

vi. PatientAdmittedToHospital

vii. PatientAdmittedToIcu

d) Infection rate variables

i. AssessmentOutcome

ii. Covid19VirologyResultFinding

e) Workplace variables

i. HealthCareWorker

ii. PatientIsACarer

f) Infection outcome variables

i. DateOfDeath

ii. DateOfRecovery

g) Risk variables

i. AreThereSignificantRiskFacto

ii. Cancer

iii. ChronicHeartDisease

iv. ChronicKidneyDisease

v. ChronicLiverDisease

vi. ChronicNeurologicalDisease

vii. ChronicRespiratoryDisease

viii. CurrentlyPregnant

ix. Diabetes

x. PatientHasUnderlyingCondition

4. CASES Dataset

a) ID Variables

i. COVID19 IDPIK

ii. DOBSURNAME PIK

iii. SWIFTQUEUE IDPIK

iv. PATIENT IDPIK

v. CaseNumber PIK

vi. RecordId

b) Demographic variables

i. PatientDateOfBirth

ii. GenderPatient

c) Infection timeline variables

i. InformedDate

ii. LastActivityDate

iii. ContactTracingResolutionDate

d) Infection rate variables

i. Covid19Positive

e) Infection outcome variables

i. Outcome

f) Contact variables

i. ContactTracingMethod

ii. ContactType

iii. HaveYouCompletedContactTraci

5. CONTACTS Dataset

a) ID Variables

i. COVID19 IDPIK

ii. DOBSURNAME PIK

iii. PATIENT IDPIK

b) Demographic variables

i. ContactDateOfBirth

ii. ContactGender

iii. InterpreterLanguageRequired

iv. InterpreterRequired

c) Infection timeline variables

i. CreatedOn

ii. RecordCreatedOn

d) Infection rate variables

i. ContactOutcomeDate

ii. Symptomatic

e) Community rate variables

i. ContactEircode RoutingKey

ii. County

f) Infection outcome variables

i. Outcome

g) Contact variables

i. CircumstancesOfContact

ii. ContactType

iii. DateOfLastContactOccurred

iv. IdentifiedAsContactByApp

v. IdentifiedAsContactByPatient

vi. CongregatedResidentialSetting

vii. ReasonForBeingComplex

viii. ReasonForBeingExceptional

6. NEGPATIENTTESTS Dataset

a) ID Variables

i. COVID19 IDPIK

ii. DOBSURNAME PIK

iii. PATIENT IDPIK

iv. SWIFTQUEUE IDPIK

b) Demographic variables

i. PatientDateOfBirth

c) Infection timeline variables

i. CreatedOn RegardingLabResult

ii. DateCreated

iii. DateReported RegardingLabResult

iv. DateofTest

v. ModifiedOn RegardingLabResult

7. PATIENT Dataset

a) ID Variables

i. COVID19 IDPIK

ii. DOBSURNAME PIK

iii. PATIENT IDPIK

iv. RecordIdSourceContact

b) Demographic variables

i. Gender

ii. InterpreterRequired

iii. Nationality

iv. PatientDateOfBirth

c) Infection timeline variables

i. DateofTest

d) Community rate variables

i. Eircode RoutingKey

e) Infection Rate Variables

i. Covid19Result

f) Infection Outcome Variables

i. PatientStatus

g) Contact variables

i. ContactType

h) Workplace Variables

i. HealthCareWorker

8. POSPATIENTASSESSMENTS Dataset

a) ID Variables
    i. COVID19 IDPIK
    ii. DOBSURNAME PIK
    iii. PATIENT IDPIK
    iv. ResidentialIdSqAppointment

b) Demographic variables
    i. EthnicityPpa
    ii. OtherEthnicityDetailsPpa
    iii. PatientDateOfBirth
    iv. GenderPatient

c) Infection timeline variables
    i. CreatedOn
    ii. DateOfFirstSymptoms
    iii. DateOfHospitalAdmission
    iv. DateOfICUAdmission
    v. PatientAdmittedToHospital
    vi. PatientAdmittedToIcu

d) Infection rate variables
    i. Covid19Result
    ii. DidPatientHaveSymptoms
    iii. Symptomatic

e) Community rate variables

i. WhichHospital

ii. Eircode RoutingKey

f) Infection outcome variables

i. DateOfDeath

ii. PositiveAssessmentOutcome

g) Contact variables

i. CongregatedResidentialSetting

ii. CongregatedResidentialSetting1

iii. ContactTracingMethodPatient

iv. InCongregatedResidentialSett1

v. InCongregatedResidentialSett

vi. MostLikelyTransmissionSource

vii. TypeOfResidencePpa

h) Workplace variables

i. CurrentOccupationSectorOther

ii. HcwWhoHasDirectContactPpa

iii. HealthCareWorker

iv. HealthCareWorkerRole

v. HealthCareWorkerRolePpa

vi. OtherCurrentOccupationRole

vii. OtherHealthCareWorkerRoleDe

viii. OtherTypeOfHcwPlaceOfWork

ix. PatientIsACarer

x. WhatTypeOfFacilityDoYouWor

References

Adamic, L. A., & Huberman, B. A. (2002). Zipf's Law and the Internet. *Glottometrics*, *3*(1), 143–150.

Ahammed, T., Anjum, A., Rahman, M. M., Haider, N., Kock, R., & Uddin, M. J. (2021). Estimation of novel coronavirus (COVID-19) reproduction number and case fatality rate: A systematic review and meta-analysis. *Health science reports*, *4*(2), e274.

Ajbar, A., Alqahtani, R. T., & Boumaza, M. (2021). Dynamics of an SIR-based COVID-19 Model With Linear Incidence Rate, Nonlinear Removal Rate, and Public Awareness. *Frontiers in Physics*, 13–13.

Alimohamadi, Y., Taghdir, M., & Sepandi, M. (2020). Estimate of the basic reproduction number for COVID-19: a systematic review and meta-analysis. *Journal of Preventive Medicine and Public Health*, *53*(3), 151.

Almeida, A., Loy, A., & Hofmann, H. (2018). Ggplot2 Compatible Quantile-Quantile Plots in R (Vol. 10) (Computer software manual No. 2). Retrieved from `https://doi.org/10.32614/RJ-2018-051`

Arroyo-Marioli, F., Bullano, F., Kucinskas, S., & Rondón-Moreno, C. (2021). Tracking R of COVID-19: A New Real-Time Estimation Using the Kalman Filter. *PloS One*, *16*(1), e0244474.

Avery, C. (2021). *A Simple Model of Social Distancing and Vaccination* (Tech. Rep.). National Bureau of Economic Research.

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., ... Creech, C. B. (2021). Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine*, *384*(5), 403–416.

Badr, H. S., Du, H., Marshall, M., Dong, E., Squire, M. M., & Gardner, L. M. (2020). Association Between Mobility Patterns and COVID-19 Transmission in the USA: A Mathematical Modelling Study. *The Lancet Infectious Diseases*, *20*(11), 1247–1254.

Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B., & Sledge, D. (2020). The Challenges of Modeling and Forecasting the Spread of COVID-19. *Proceedings of the National Academy of Sciences*, *117*(29), 16732–16738.

Bukhari, Q., Jameel, Y., Massaro, J. M., D'Agostino, R. B., & Khan, S. (2020). Periodic Oscillations in Daily Reported Infections and Deaths for Coronavirus Disease 2019. *JAMA Network Open*, *3*(8), e2017521–e2017521.

Burton, J. K., Bayne, G., Evans, C., Garbe, F., Gorman, D., Honhold, N., ... Swietlik, S. (2020). Evolution and effects of COVID-19 outbreaks in care homes: a population analysis in 189 care homes in one geographical region of the UK. *The Lancet Healthy Longevity*, *1*(1), e21–e31.

Central Statistics Office. (2022). *Annual Population Estimates*. Retrieved from `https://data.cso.ie/` (Accessed 10 January 2022)

Central Statistics Office. (2022). *Covid-19 Insight Bulletins: Deaths and Cases*. Retrieved from `https://www.cso.ie/en/releasesandpublications/ep/p-covid19/covid-19informationhub/health/covid-19deathsandcasesstatistics/` (Accessed 10 January 2022)

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, *51*(4), 661–703.

Costello, F., Watts, P., & Howe, R. (Pre print). A model of behavioural response to risk

accurately predicts the statistical distribution of COVID-19 infection and reproduction numbers.

Department of Health. (2020a). *National Public Health Emergency Team – COVID-19 Meeting Agenda* (Vol. 50).

Department of Health. (2020b). *National Public Health Emergency Team – COVID-19 Meeting Agenda* (Vol. 52).

Di Leone, G., Drago, P., Troiano, M., Mascoli, F., Dahbaoui, N., Scorrano, D., ... Iurilli, M. (2020). Integrated management method in the prevention department of a COVID-19 epidemic outbreak in a large meat processing plant in Bari province. *Epidemiologia e prevenzione*, *44*(5-6 Suppl 2), 334–339.

Dong, E., Du, H., & Gardner, L. (2020). An Interactive Web-Based Dashboard to Track COVID-19 in Real Time. *The Lancet Infectious Diseases*, *20*(5), 533–534.

Dowle, M., & Srinivasan, A. (2021). data.table: Extension of 'data.frame' [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=data.table` (R package version 1.14.2)

Dyal, J., Grant, M., Broadwater, K., & et al. (2020). COVID-19 among workers in meat and poultry processing facilities—19 states, April 2020. *Morbidity and Mortality Weekly Report*, *69*.

Flacco, M. E., Soldato, G., Acuti Martellucci, C., Carota, R., Di Luzio, R., Caponetti, A., & Manzoli, L. (2021). Interim Estimates of COVID-19 Vaccine Effectiveness in a Mass Vaccination Setting: Data from an Italian Province. *Vaccines*, *9*(6). Retrieved from `https://www.mdpi.com/2076-393X/9/6/628` doi: 10.3390/vaccines9060628

Funk, S., Salathé, M., & Jansen, V. A. (2010). Modelling the Influence of Human Behaviour on the Spread of Infectious Diseases: A Review. *Journal of the Royal Society Interface*, *7*(50), 1247–1256.

Gillespie, C. S. (2015a). Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software*, *64*(2), 1–16. Retrieved from `http://www.jstatsoft.org/v64/i02/`

Gillespie, C. S. (2015b). Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software*, *64*(i02).

Glatman-Freedman, A., Bromberg, M., Dichtiar, R., Hershkovitz, Y., & Keinan-Boker, L. (2021). The BNT162b2 vaccine effectiveness against new COVID-19 cases and complications of breakthrough cases: A nation-wide retrospective longitudinal multiple cohort analysis using individualised data. *EBioMedicine*, *72*, 103574.

Gleeson, J. P., Brendan Murphy, T., O?Brien, J. D., Friel, N., Bargary, N., & O'Sullivan, D. J. (2022). Calibrating COVID-19 Susceptible-Exposed-Infected-Removed Models With Time-Varying Effective Contact Rates. *Philosophical Transactions of the Royal Society A*, *380*(2214), 20210120.

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy With Lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from `https://www.jstatsoft.org/v40/i03/`

Günther, T., Czech-Sioli, M., Indenbirken, D., Robitaille, A., Tenhaken, P., Exner, M., ... Brinkmann, M. M. (2020). SARS-CoV-2 outbreak investigation in a German meat processing plant. *EMBO molecular medicine*, *12*(12), e13296.

Hanel, R., Corominas-Murtra, B., Liu, B., & Thurner, S. (2017). Fitting Power-Laws in Empirical Data With Estimators That Work for All Exponents. *PloS One*, *12*(2), e0170920.

Hashan, M. R., Smoll, N., King, C., Ockenden-Muldoon, H., Walker, J., Wattiaux, A., ... Khandaker, G. (2021). Epidemiology and clinical features of COVID-19 outbreaks in aged care facilities: A systematic review and meta-analysis. *EClinicalMedicine*, *33*,

100771.

Health Protection Surveillance Centre. (2022). *Surveillance for Covid-19.* Retrieved from https://www.hpsc.ie/a-z/respiratory/coronavirus/novelcoronavirus/surveillance/ (Accessed 16 January 2022)

Health Service Executive. (2022). *Ireland's COVID-19 Data Hub: Data and services.* Retrieved from https://covid-19.geohive.ie/ (Accessed 14 January 2022)

Helwig, N. E. (2021). Nptest: Nonparametric Bootstrap and Permutation Tests [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=nptest (R package version 1.0-3)

Herstein, J. J., Degarege, A., Stover, D., Austin, C., Schwedhelm, M. M., Lawler, J. V., ... Donahue, M. (2021). Characteristics of SARS-CoV-2 transmission among meat processing workers in Nebraska, USA, and effectiveness of risk mitigation measures. *Emerging infectious diseases*, *27*(4), 1032.

Hyland, P., Vallières, F., Shevlin, M., Bentall, R. P., McKay, R., Hartman, T. K., ... Murphy, J. (2021). Resistance to COVID-19 vaccination has increased in Ireland and the United Kingdom during the pandemic. *Public Health*, *195*, 54–56.

IHME COVID-19 forecasting team. (2020). Modeling COVID-19 Scenarios for the United States. *Nature Medicine*.

Illingworth, C. J., Hamilton, W. L., Warne, B., Routledge, M., Popay, A., Jackson, C., ... Hosmillo, M. (2021). Superspreaders drive the largest outbreaks of hospital onset COVID-19 infections. *elife*, *10*.

Karodia, A., Hargovan, M., Fadal, R., Parker, W., Singh, R., Hoosen, A., ... Mahomed, S. (2020). COVID-19 and its effects on the food production industry of South Africa. *Occupational Health Southern Africa*, *26*(4), 158–161.

Koyama, S., Horie, T., & Shinomoto, S. (2021). Estimating the Time-Varying Reproduction

Number of COVID-19 With a State-Space Method. *PLoS Computational Biology*, *17*(1), e1008679.

Little, C., Alsen, M., Barlow, J., Naymagon, L., Tremblay, D., Genden, E., ... van Gerwen, M. (2021). The Impact of Socioeconomic Status on the Clinical Outcomes of COVID-19; A Retrospective Cohort Study. *Journal of Community Health*, *46*(4), 794–802.

Liu, X., Huang, J., Li, C., Zhao, Y., Wang, D., Huang, Z., & Yang, K. (2021). The Role of Seasonality in the Spread of COVID-19 Pandemic. *Environmental Research*, *195*, 110874.

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The Reproductive Number of COVID-19 Is Higher Compared to SARS Coronavirus. *Journal of Travel Medicine*.

Locatelli, I., Trächsel, B., & Rousson, V. (2021). Estimating the basic reproduction number for COVID-19 in Western Europe. *Plos one*, *16*(3), e0248731.

Mallet, Y., Pivette, M., Revest, M., Angot, E., Valence, M., Dupin, C., ... Ballet, S. (2021). Identification of Workers at Increased Risk of Infection During a COVID-19 Outbreak in a Meat Processing Plant, France, May 2020. *Food and Environmental Virology*, *13*(4), 535–543.

Manrubia, S., & Zanette, D. H. (2021). Individual Risk-Aversion Responses Tune Epidemics to Critical Transmissibility ($R = 1$). *arXiv Preprint arXiv:2105.10572*.

Manrubia, S., & Zanette, D. H. (2022). Individual risk-aversion responses tune epidemics to critical transmissibility (r$=$ 1). *Royal Society Open Science*, *9*(4), 211667.

Middleton, J., Reintjes, R., & Lopes, H. (2020). *Meat plants—a new front line in the covid-19 pandemic* (Vol. 370). British Medical Journal Publishing Group.

Mills, M., Rahal, C., Brazel, D., Yan, J., & Gieysztor, S. (2020). COVID-19 vaccine deployment: Behaviour, ethics, misinformation and policy strategies. *London: The Royal Society & The British Academy*.

Murphy, J., Vallières, F., Bentall, R. P., Shevlin, M., McBride, O., Hartman, T. K., ... Gibson-Miller, J. (2021). Psychological characteristics associated with COVID-19 vaccine hesitancy and resistance in Ireland and the United Kingdom. *Nature communications*, *12*(1), 1–15.

Ndaïrou, F., Area, I., Nieto, J. J., & Torres, D. F. (2020). Mathematical Modeling of COVID-19 Transmission Dynamics with a Case Study of Wuhan. *Chaos, Solitons & Fractals*, *135*, 109846.

Nouvellet, P., Bhatia, S., Cori, A., Ainslie, K. E., Baguelin, M., Bhatt, S., ... Cooper, L. V. (2021). Reduction in Mobility and COVID-19 Transmission. *Nature Communications*, *12*(1), 1–9.

Olliaro, P., Torreele, E., & Vaillant, M. (2021). Covid-19 vaccine efficacy and effectiveness—the elephant (not) in the room. *The Lancet Microbe*.

Oran, D. P., & Topol, E. J. (2020). Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. *Annals of internal medicine*, *173*(5), 362–367.

Park, M., Cook, A. R., Lim, J. T., Sun, Y., & Dickens, B. L. (2020). A Systematic Review of COVID-19 Epidemiology Based on Current Evidence. *Journal of Clinical Medicine*, *9*(4), 967.

Patel, J., Nielsen, F., Badiani, A., Assi, S., Unadkat, V., Patel, B., ... Wardle, H. (2020). Poverty, Inequality and COVID-19: The Forgotten Vulnerable. *Public Health*, *183*, 110.

Patel, M. K., Bergeri, I., Bresee, J. S., Cowling, B. J., Crowcroft, N. S., Fahmy, K., ... Lanata, C. F. (2021). Evaluation of post-introduction COVID-19 vaccine effectiveness: Summary of interim guidance of the World Health Organization. *Vaccine*, *39*(30), 4013–4024.

Payne, D. C., Smith-Jeffcoat, S. E., Nowak, G., Chukwuma, U., Geibe, J. R., Hawkins, R. J.,

... Weiner, Z. (2020). SARS-CoV-2 infections and serologic responses from a sample of US Navy service members—USS Theodore Roosevelt, April 2020. *Morbidity and Mortality Weekly Report*, *69*(23), 714.

Pedersen, T. L. (2020). Patchwork: The Composer of Plots [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=patchwork` (R package version 1.1.1)

Perra, N. (2021). Non-Pharmaceutical Interventions During the COVID-19 Pandemic: A Review. *Physics Reports*, *913*, 1–52.

Pokora, R., Kutschbach, S., Weigl, M., Braun, D., Epple, A., Lorenz, E., ... Rietschel, P. (2021). Investigation of superspreading COVID-19 outbreak events in meat and poultry processing plants in Germany: A cross-sectional study. *PloS one*, *16*(6), e0242456.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., ... Zerbini, C. (2020). Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *New England Journal of Medicine*, *385*, 1761–1773.

Poukka, E., Baum, U., Palmu, A. A., Lehtonen, T. O., Salo, H., Nohynek, H., & Leino, T. (2021). Cohort study of Covid-19 vaccine effectiveness among healthcare workers in Finland, December 2020 - October 2021. *Vaccine*.

R Core Team. (2021). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ritchie, H., Mathieu, E., Rodes-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., ... Roser, M. (2020). Coronavirus Pandemic (COVID-19). *Our World in Data*. (https://ourworldindata.org/coronavirus)

Russo, L., Anastassopoulou, C., Tsakris, A., Bifulco, G. N., Campana, E. F., Toraldo, G., & Siettos, C. (2020). Tracing Day-Zero and Forecasting the COVID-19 Outbreak in Lombardy, Italy: A Compartmental Modelling and Numerical Optimization Approach.

PloS One, *15*(10), 240–649.

Sadoff, J., Gray, G., Vandebosch, A., Cárdenas, V., Shukarev, G., Grinsztejn, B., ... Spiessens, B. (2021). Safety and efficacy of single-dose Ad26. COV2. S vaccine against Covid-19. *New England Journal of Medicine*, *384*(23), 2187–2201.

Sah, P., Fitzpatrick, M. C., Zimmer, C. F., Abdollahi, E., Juden-Kelly, L., Moghadas, S. M., ... Galvani, A. P. (2021). Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. *Proceedings of the National Academy of Sciences*, *118*(34).

Schaller, M. (2011). The Behavioural Immune System and the Psychology of Human Sociality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1583), 3418–3426.

Shaw, C. L., & Kennedy, D. A. (2021). What the reproductive number R0 can and cannot tell us about COVID-19 dynamics. *Theoretical Population Biology*, *137*, 2-9. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0040580921000010` doi: https://doi.org/10.1016/j.tpb.2020.12.003

Steinberg, J., Kennedy, E. D., Basler, C., Grant, M. P., Jacobs, J. R., Ortbahn, D., ... Clayton, J. L. (2020). Covid-19 outbreak among employees at a meat processing facility—South Dakota, March–April 2020. *Morbidity and Mortality Weekly Report*, *69*(31), 1015.

Steinegger, B., Arenas, A., Gómez-Gardeñes, J., & Granell, C. (2020). Pulsating Campaigns of Human Prophylaxis Driven by Risk Perception Palliate Oscillations of Direct Contact Transmitted Diseases. *Physical Review Research*, *2*(2), 023181.

Steinegger, B., Arola-Fernández, L., Granell, C., Gómez-Gardeñes, J., & Arenas, A. (2022). Behavioural Response to Heterogeneous Severity of COVID-19 Explains Temporal Variation of Cases Among Different Age Groups. *Philosophical Transactions of the Royal Society A*, *380*(2214), 20210119.

Thompson, D.-C., Barbu, M.-G., Beiu, C., Popa, L. G., Mihai, M. M., Berteanu, M., & Popescu, M. N. (2020). The impact of COVID-19 pandemic on long-term care facilities worldwide: an overview on international issues. *BioMed research international*, *2020*.

Tkachenko, A. V., Maslov, S., Wang, T., Elbana, A., Wong, G. N., & Goldenfeld, N. (2021). Stochastic Social Behavior Coupled to COVID-19 Dynamics Leads to Waves, Plateaus, and an Endemic State. *Elife*, *10*, e68341.

US Food and Drug Administration. (2020). Vaccines and Related Biological Products Advisory Committee Meeting. FDA Briefing Document Moderna COVID-19 Vaccine. , *5*, 2021. Retrieved from `https://www.fda.gov/advisory-committees/advisory-committee-calendar/vaccines-and-related-biological-products-advisory-committee-february-26-2021-meeting-announcement`

Verelst, F., Willem, L., & Beutels, P. (2016). Behavioural Change Models for Infectious Disease Transmission: A Systematic Review (2010–2015). *Journal of the Royal Society Interface*, *13*(125), 20160820.

Vitek, M. G., Klavs, I., Učakar, V., Serdt, M., Mrzel, M., Vrh, M., & Fafangel, M. (2022). Vaccine effectiveness against severe acute respiratory infections (SARI) COVID-19 hospitalisations estimated from real-world surveillance data, Slovenia, October 2021. *Eurosurveillance*, *27*(1), 2101110.

Voysey, M., Clemens, S. A. C., Madhi, S. A., Weckx, L. Y., Folegatti, P. M., Aley, P. K., ... Bhorat, Q. E. (2021). Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet*, *397*(10269), 99–111.

Walshe, N., Fennelly, M., Hellebust, S., Wenger, J., Sodeau, J., Prentice, M., ... Downey, V. (2021). Assessment of Environmental and Occupational Risk Factors for the Mitigation

and Containment of a COVID-19 Outbreak in a Meat Processing Plant. *Frontiers in public health*, 1544.

Waltenburg, M. A., Victoroff, T., Rose, C. E., Butterfield, M., Jervis, R. H., Fedak, K. M., ... Austin, C. (2020). Update: COVID-19 among workers in meat and poultry processing facilities—United States, April–May 2020. *Morbidity and Mortality Weekly Report*, *69*(27), 887.

Weitz, J. S., Park, S. W., Eksin, C., & Dushoff, J. (2020). Awareness-Driven Behavior Changes Can Shift the Shape of Epidemics Away From Peaks and Toward Plateaus, Shoulders, and Oscillations. *Proceedings of the National Academy of Sciences*, *117*(51), 32764–32771.

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from `https://ggplot2.tidyverse.org`

Zheng, C., Shao, W., Chen, X., Zhang, B., Wang, G., & Zhang, W. (2022). Real-world effectiveness of COVID-19 vaccines: a literature review and meta-analysis. *International Journal of Infectious Diseases*, *114*, 252–260.