# Data-driven analytics of disease spread under close contact for optimal testing and mitigation

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Biology, Medicine and Health

2023

Hugo J Lewkowicz
School of Health Sciences

# Contents

Word count: 71706

# List of figures

# List of tables

# Terms and abbreviations

| Acronym | Meaning |
|---|---|
| ABM | Agent-Based Model |
| AIDS | Acquired Immuno-Deficiency Syndrome |
| API | Application Programming Interface |
| ART | Anti-retroviral therapy |
| CD4 | Cluster of Differentiation 4 |
| CDC | The Centre for Disease Control |
| CDF | Cummulative Density Function |
| COG-UK-HOCI | COVID-19 Genomics UK Hospital Onset COVID-19 Infection |
| COPE | COVID in Older People |
| COVID | Coronavirus |
| EECA | Eastern Europe and Central Asia |
| ELISA | Enzyme-linked immunosorbent assay |
| EMOD | Epidemiological MODelling |
| FAIR | Findable Accessible Interoperable Reusable |
| FOI | Force of Infection |
| HCV | Hepatitis-C Virus |
| HIV | Human Immunodeficiency Virus |
| IPC | Infection Prevention and Control |
| LFT | Lateral flow test |
| LTBI | Latent tuberculosis infection |
| MERS | Middle Eastern Respiratory Syndrome |
| MLE | Maximum Likelihood Estimate |
| MPLE | Maximum Pseudo-Likelihood Estimator |
| NHS | National Health Service |
| NPSs | Novel Psychoacitve Substances |
| ODE | Ordinary Differential Equation |
| OST | Opioid Substitution Therapy |
| PCR | Polymerase Chain Reaction |
| PDF | Probability Density Function |
| PEPI | Programs for Epidemiologists |
| PHE | Public Health England |
| PMF | Probability Mass Function |
| PPE | Personal Protective Equipment |
| PWID | People who inject drugs |
| QALY | Quality-adjusted life year |
| REACT-2 | Real-time Assessment of Community Transmission 2 |
| RNA | Ribo-Nucleic Acid |
| SARS-CoV-2 | Severe Acute Respiratory Symdrome - Coronavirus 2 |
| SEIR | Susceptible→Exposed→Infectious→Recovered |
| SIR | Susceptible→Infectious→Removed |
| SIS | Susceptible→Infectious→Susceptible |
| SQL | Structured Query Language |
| St.d. | Standard deviation |
| STEC | Shiga-toxin producing Escherichia coli |
| STEM | Spaciotemporal Epidemiological Modeler |
| SVD | Swine Vesicular Disease |
| TB | Tuberculosis |
| TITAN | Treatment of Infection and Transmission ABNăNetwork |
| Titan | TMC114/r In Treatment-experienced pAtients Naïve to lopinavir |
| UK | United Kingdom |
| USA | The United States of America |

# Abstract

Models emulating the spread of infectious diseases in close-contact environments present a set of unique challenges. This field of research has exploded over the past three years due in part to the SARS-CoV-2 pandemic. In this thesis, we present and explore five different models of infection in close-contact environments which aim to fulfill five different needs.

The first model (Chapter 2) is used to study the ability to estimate the outbreak size, i.e. the total number of individuals in a group who have been infected following an exposure event, based on the number of observed symptomatic individuals by a certain time after the event. Of the three quantities we investigated, the proportion of individuals who have been observed to be symptomatic, the outbreak size and the group size, the first one is shown to have the greatest influence on the entropy of the resulting predicted distribution for the outbreak size.

The second model (Chapter 3) explores the effect of rota patterns on workplace infection rates by calculating a central estimate for the length of time an individual is at work whilst infectious. We first explore this model numerically and then approach an analytical solution by representing rota patterns as a Fourier series. In both cases, we find that longer shifts reduce in-work infectiousness and, for a parameter set emulating SARS-CoV-2, a rota length of approximately 10–11 days is optimal.

Nosoco (introduced in Chapter 4) is a tool we have generated for approximating the total number of in-hospital infections based on the timing of positive swabs. We explore how efficient it is as a tool compared to declaring as nosocomial infections all individuals diagnosed after a fixed number of days since their admission, explore how the proportions of total cases attributed to infection within and outside the hospital changes over time and estimate each individual's daily rate of infection.

The fourth model (Chapter 5) is used to estimate the incidence of Hepatitis C from results of a cross-sectional survey when assuming a constant rate of infection. We use an example study of a cross-sectional survey of Scottish prisons to show no evidence, among people who inject drugs, of a lower incidence in prison compared to the external incidence .

The final model (Chapter 6) incorporates a distance element into transmission trees as we investigate an outbreak aboard a cruise-liner, using cabin location as a proxy for infector location.

# Declaration of originality

I hereby confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

# Copyright statement

# Acknowledgements

These past few years have been challenging for a multitude of reasons and I know that I am incredibly grateful for the support I have received along the way. To name a few (but certainly not all) of the people who more than deserve my gratitude:

Thank you to my PhD supervisors, Ian Hall, Lorenzo Pellis, Thomas House and Andy Ustianowski. Thank you for your patience and guidance as I embarked in learning areas of statistics and modelling I had never heard of prior to this project. And thank you Andy for being my medical island in this stormy seas of mathematics.

Thank you Chris Overton, Shaz Ahmad and Ryan George. Whilst neither of you were officially my supervisors, you certainly helped and advised me along the way.

Thank you to all the members of my office, who accepted me despite my complete lack of knowledge regarding anything in the Applied Mathematics field. Thank you for your patience when I asked for the forty-billionth time some obvious question regarding proper mathematical notation.

Thank you to my family, for always believing in me and supporting me in this mad endeavour. In particular, thank you George for always listening to my ramblings about some statistical model that I am sure were far beneath your work, and thank you Josh. Very early on you told me that whatever I did, you knew I'd ace it, and that thought has helped me through a lot of things that needed acing.

Thank you Luke Chaplin, God of Chicken Sex. Thank you for paving a way for me, and showing that medicine is not the only option. Also, as with George, thank you for listening to coding queries that were no doubt way beneath you.

Finally, thank you Elaine McFarlane, my wife. Thank you for supporting me emotionally, financially and, after one fateful Christmas party, physically. Thank you for marrying me, which we did during this whole project! Thank you for listening to me and being patient with me when I could not stop thinking about this project. Thank you for not throttling me at any point during the 80 years of lock down. Also, it is probably worth saying: Thank you for all your work during that whole pandemic thing we had! You are my hero, my rock star and my inspiration. Thank you for our future. I love you. Thank you.

# The author

Hugo Lewkowicz is a PhD student from the University of Manchester. He has a joint Bachelor's degree in Medicine and Surgery (2014) and an MSc in Bioinformatics and Systems Biology (2019), both from the University of Manchester. Prior to starting his PhD, he had five years of experience working as a junior doctor across a number of different hospitals and specialities. During this time, he did exactly one shift in an Infectious Diseases department.

# Chapter 1

# Introduction

This chapter is an introduction to “Data-driven analytics of disease spread under close contact for optimal testing and mitigation”. We will start in Section 1.1 by discussing literature relevant to this area of research in the form of a literature review. This helped inform us of gaps in the field and possible areas of future research. Section 1.2 explores our understanding of the thesis title and how it informed our goals in this PhD project. Finally, in Section 1.3 we introduce the five core research projects of this PhD.

## 1.1 Literature Review

### 1.1.1 Search protocol

The following section was originally written in 2020, based on a literature search performed on 21/10/2019. We performed a PubMed search for the phrase “(((epidemic OR infect*)) AND (model* OR simulat*)) AND (household OR close-contact OR prison OR ship)”. This produced 14749 articles. We refined the search to include only articles produced in the past five years pertaining to humans. We reviewed the titles and abstracts of the resulting 5074 articles in order of “Best match” according to the PubMed search. We ruled articles in if they were focused on an infectious disease, if they considered a mathematical model for the transmission of said disease and if they involved a close-contact or closed environment. We stopped once we had identified 105 articles.

In these 105 articles, the were two systematic reviews, looking at the spread of infectious diseases related to incarceration[1] and the mathematical modelling of the spread of Hepatitis C[2]. We followed the citations used in these articles. Additionally, we identified five modelling programmes from the previously mentioned articles. These were each chosen to highlight specific common aspects of available epidemiological modelling tools. We searched for these tools on PubMed for their use in literature over the past 5 years.

The breadth of research in modelling infectious diseases has expanded dramatically over the course of this 3 and a half year research project. Repeating the same search criteria in March 2023, we found a further 10,044 papers published since 2019, of which approximately 50% fulfill our more selective criteria of being focused on an infectious disease, considering the mathematical modeling of transmission and if they involve a close-contact or closed environment, and 29% fulfill this criteria and are focused on SARS-CoV-2. These numbers are based on the first 200 publications appearing in this search. Put in context, this is 4.23 relevant papers per day from January 2020 to March 2023 and 2.45 papers per day focused on SARS-CoV-2 over the same time frame.

To the best of our ability, we have updated this literature review to include recent papers that are similarly relevant. This is most evident in Section 1.1.6 where we have included an update on additional popular outbreak modelling tools.

There are many approaches to analysing the spread of infectious diseases. This thesis aims to show a few mathematical models tailor-made for understanding their spread in close-contact environments, such as prisons or hospitals. In order to do so, we need to establish what tools are already available. This chapter will highlight the wide range of analytical approaches that have been taken in the past to help understand real-life and so-called "toy" model outbreaks. We give examples of key literature on the subject, as well as, where relevant, briefly discuss the mathematics involved. We also identify what we feel is missing from the tools currently available, as this is the motivation behind the work in our thesis.

### 1.1.2 Historical models of disease spread

#### Compartmental models

Compartmental models represent an important method for understanding the spread of infectious diseases. They do not appear in the body of this thesis as, for reasons we will discuss later, they can be an inappropriate way to model certain close-contact environments. However, they are so foundational to the field of infectious disease epidemiology that it would be wrong of us to exclude them from our literature review.

Compartmental models are used to approximate how different populations change in size over time. Each population is represented by a parameter, whose size indicates the size of the population. We then write a series of equations, called Ordinary Differential Equations (ODEs), that show how the size of these populations change over time. Rather than considering individuals, we can consider the overall flow of all individuals between compartments. This model style can be useful when the rate at which the population sizes change is dependent on the size of the population itself. In a predator-prey scenario, the rate at which the predator population changes is dependent on both the predator population size (for breeding) and the prey size (for availability of food)[3].

In infectious disease modelling, we can divide the population up by their infection status. A classic example of this is the $SIR$ model, in which a community is divided into groups $S$, $I$ and $R$. The population in the $S$ (susceptible) group is still susceptible to the disease. They flow into the infectious $I$ group at a rate dependent both on the total number of available susceptible individuals and the number of infectious individuals. Over time, infected individuals recover from the disease and move into the $R$ (recovered) group where they are immune to the disease, at a rate dependent on the size of the infectious group. The movement between these three groups over time can be expressed through the following ODEs:

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta I S\\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta I S - \sigma I\\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \sigma I\end{aligned}$$

In these equations, $\beta$ represents the infection rate from an infected individual to a susceptible individual and $\sigma$ denotes the recovery rate from the disease. We can analyse these ODEs to draw conclusions about disease spread. For example, the number of people one person can infect in an entirely susceptible population ($R_0$) is $\frac{\beta}{\sigma}$. If this number is less than 1 then an outbreak cannot occur, as the rate of growth of the $I$ group would always be negative. We could therefore show a vaccination plan (where vaccination converts a susceptible person to a recovered person) needs to reach $1-\frac{1}{R_0}$ of the population to prevent an outbreak.

Looking at this model, we can see that any epidemic will eventually die out. The number of susceptible people will always reduce, meaning there will always come a point where the infectious population recover faster than they infect new people (see Figure 1.2 A). However, if birth and death are introduced to the population this is not necessarily true. With each birth, new susceptible people are introduced to the population, meaning that given the right $R_0$ the disease will not die out and instead enter an endemic state (see Figure 1.2 B)[4].

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= \gamma(I+R) - \beta I S\\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta I S - I(\sigma+\gamma)\\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \sigma I - \gamma I\end{aligned}$$

With this change to the model, individuals die and are born at the same rate, $\gamma$. They are born susceptible, meaning new susceptible individuals are constantly being introduced. The infection is no longer guaranteed to die out, as $\frac{\mathrm{d}I}{\mathrm{d}t}$ does not necessarily turn negative over time, meaning the disease can become endemic.

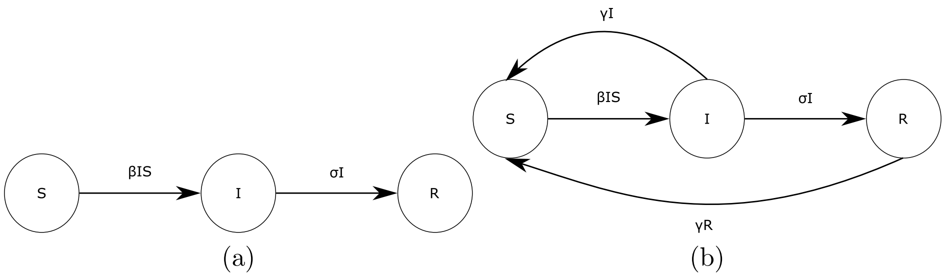


Figure 1.1. Two SIR compartmental models show the flow of individuals between compartments. In the left hand model, the Susceptible group can become infected and move to the Infectious group, who over time end their infectious period and move to the Recovered group. This movement corresponds to the change in group sizes seen in Figure 1.2 A. The same movement occurs in the right hand model, except individuals die and are born at a rate $\gamma$, corresponding with Figure 1.2 B.

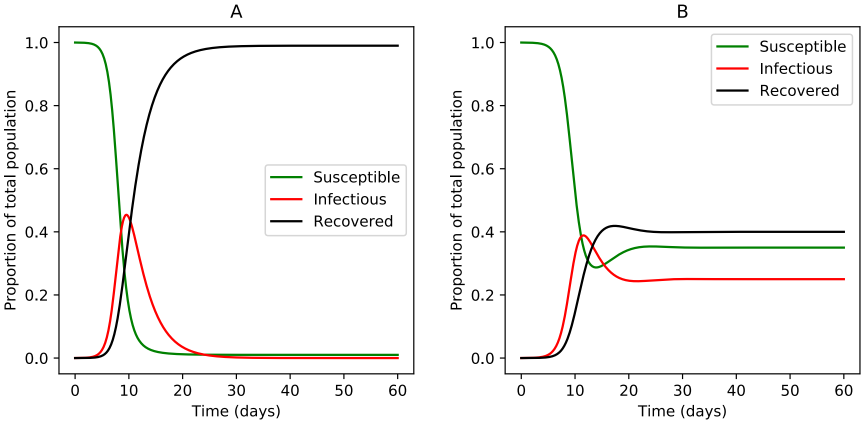


Figure 1.2. The changes of population proportions an SIR model over time. Both figures demonstrate models of a disease with a transmission rate ($\beta$) of 1.42 transmission per day and a recovery rate ($\sigma$) of 0.31 recoveries per day. This results in an $R_0$ of 4.64. 1.2A shows a model that does not include deaths or births. A peak can be observed at approximately time 12. After this point, with a low susceptible population, the infectious population runs out of people to infect and decreases in size. 1.2B shows a model that accounts for births and deaths. With the introduction of a new susceptible population, the infectious population now tend towards an endemic equilibrium.

This simple yet effective model, first proposed by Kermack and McKendrick[5], can be used as the basis to investigate multiple disease scenarios. A search of PubMed for SIR model[Title/Abstract] produced a total of 355 articles prior to 2020 and a further 534 after this time[6]. Part of the reason for this popularity is the models adaptability. For example, having recovering individuals move from the Infected group to the Susceptible group and removing the Recovered group creates a model where people do not develop immunity following exposure to a disease, such as seen in many sexually transmitted diseases[7] (see Figure 1.3). Introducing a new group E (Exposed) between being the Susceptible and Infectious individuals creates a delay in infectiousness as seen in Influenza or SARS-CoV-2[8] (see Figure 1.4).

The adaptability of compartmental models and the fact that they look at populations rather than individuals mean that epidemiological data for a disease can be fitted to an appropriate model to parameterise factors such as the disease recovery time and transmission rate. An example of this fitting was seen in 2014 when Lewnard et al. published a paper using a deterministic compartmental model to simulate the

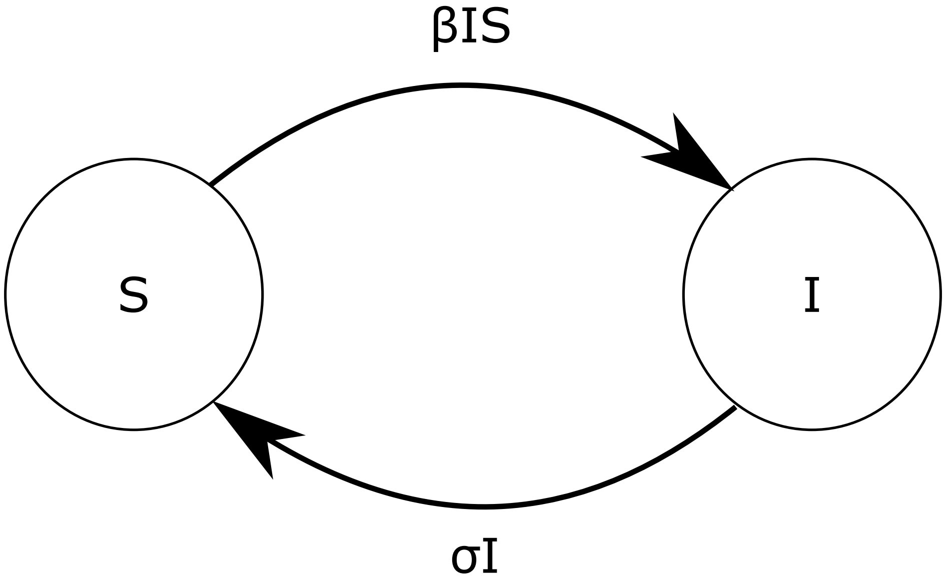


Figure 1.3. An SIS compartmental model, showing a scenario where infected individuals do not gain any immunity after recovering from an infection. This is in keeping with many sexually transmitted diseases.

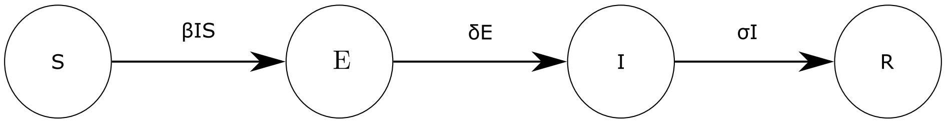


Figure 1.4. An SEIR compartmental model, where, before becoming infectious, infected individuals spend a period of time in an exposed group, creating a delay in the model.

spread of Ebola in Montserrado, Liberia based on cumulative incidence reports[9]. By adapting the above model to include such factors as hospitalisation, caring for household members and the infection risk of incorrectly disposed bodies, they investigated the effect of additional treatment centre beds on preventing an outbreak of the disease. They showed that extra hospital beds, which at the time had been promised by the USA, were inadequate to control the outbreak and that the time window to act was rapidly closing. They used a least squares fit to fit their data and assumed an independence in variation of their data points. As a result they demonstrated an apparent narrow error margin in their estimations.

King et al. investigated the same data using similar models[10]. They had two concerns. Firstly, the cumulative data used in the Lewnard study meant that any variation was dependent rather than independent, resulting in a wider error margin in their estimated prediction. Secondly, the use of a deterministic model raises issues when compared to equivalent stochastic models. Deterministic models are ones where the result is necessarily determined by the starting conditions. For example, compartmental ODE models are deterministic: given the same set of starting conditions, they will always follow the same path. Stochastic models allow for some random variation between iterations of the model. We will discuss them in more depth later. Deterministic models can overestimate values. If the rate of departure from a compartment is modelled as an exponential function, as was the case in their ODEs, the length of time in each compartment has an infinitely long positive tail, resulting in overestimation and broader error windows. Stochastic models conversely do not and allow for random variance in a population. King generated an equivalent stochastic model to the one used by Lewnard and fitted it to the same data. In doing so they showed to a greater accuracy that the spread of Ebola would be far less rapid than predicted. In raising these concerns, King showed the important role stochastic modelling will have in the prediction of the spread of infectious diseases, in that it was necessary for modelling small populations, early outbreak dynamics and scenarios where $R_0$ is close to 1.

With regards to SARS-CoV-2, an SEIR model that allows for some presymptomatic period tends to be more popular, as it reflects an important delay between infection and symptom onset in the SARS-CoV-2 career. Many research groups have produced papers fitting an SEIR model to SARS-CoV-2 outbreak data, For example, Al-Khani et al. fitted an SEIR model to outbreak data across Saudi Arabia to estimate the effect of social distancing and predict the shape of the pandemic by Hajj season[11], Nguyen et al. used an SEIR model with parameterisation dependent on 8 age groups in New Zealand to predict the outcome of multiple different age dependent vaccination programmes[12] and Song et al. used a similar SEIR model to examine the effect of age-dependent immune responses to vaccination against SARS-CoV-2 to on the overall outbreak size[13].

One basic assumption that is shared by each deterministic ODE model of this variety is that every infectious individual has the same ability to infect every susceptible individual. By thinking about the dynamics of most infectious diseases (the need for proximity, for example) we can conclude that this assumption is not accurate. We look at work that has focused on correcting this assumption in the next section.

#### Heterogenous mixing

One of the most important factors in human infectious disease modelling is the unpredictable nature of social interactions making a stochastic system of disease transmission. Multiple researchers have investigated at environments of heterogeneous mixing. Heterogeneous mixing describes an environment where the probability that individual A will interact with individual B is not necessarily the same as the probability that they will interaction with individual C. This is different from homogeneous mixing, where every individual is equally likely to interact with every other individual.

One way to simulate heterogeneous mixing is to further compartmentalise populations into households. Ball has published multiple papers looking at parameterising stochastic spread of diseases given that populations live in households. In these models, each member of a population belongs to a smaller household. A disease is more likely to transmit between members of the same household than between unrelated members of the population. His team found that in early stages, where a household has only been infected by one incident, the attack rate of completed epidemics in households will follow a distribution dependent on the within-household spread rate. Therefore, the final size distribution in early epidemics can be used to estimate the within-household infection rate[14].

From household models, Ball et al. went on to include schools in their models [15]. They defined schools as larger additional groups that household members belong to, with increased probability of intra-school transmission, but less so than household transmission. They analysed household and household/school models, looking at how school allocation based on household allocation would affect the spread of disease through a population. They compared two extreme scenarios, one where every household member belonged to the same school and one where schools were assigned independently of household. They found $R_0$ to be inadequate in these scenarios, and so explored multiple alternative $R$ values. Additionally, they discovered that with a more structured social network the principle of vaccinating $1 - \frac{1}{R_0}$ of the population would be inadequate to prevent a large epidemic.

Keeling et al. used household models to further investigate vaccination policies[16]. They were looking at an allocation system in which members of a household are assigned vaccines until there is the same number of unvaccinated individuals in each household. Whilst this is an optimum strategy for a density-dependent spread of disease (dependent on the proportion of the population infected), they found that it was no longer as effective when the disease spread was frequency dependent (dependent on the total number of infected individuals).

Zhang et al. used household models to investigate the efficacy of household isolation for scenarios where vaccination may not be readily available[17]. They used a stochastic model to simulate household self-isolation once one member presented with influenza symptoms and found that this reduced the reproductive value of each house. Their aim was to show the importance of household isolation in the reduction of between household transmissions, or more importantly the effect of imperfect compliance with household isolation. This prescient paper (it was written in 2015) showed that household isolation can be very effective, and that even poor compliance can have a considerable reduction in the number of households one infectious household infects.

An alternative representation of heterogeneous mixing to household models is an interaction network. Bioglio et al. showed that parameters found based on homogeneous mixing models can be adjusted linearly to be used with network-based contact models[18]. Ball et al. also looked at the application of contact networks on household models[19]. They focused on different distribution methods of limited vaccines in order to prevent epidemics. They found that prioritising people on contact networks with a high degree of connectivity may be more effective than targeting certain household sizes. However, they criticized this theoretical outcome, noting that it may be both unethical and impractical to enact in real life.

Certain diseases are often the focus of household-transmission investigations. Otero et al showed Tuberculosis risk in Lima, Peru to be 10 times higher when living with a person with a positive diagnosis[20]. Zeiner et al. corroborated this increased risk and further showed that in cases of TB, testing and treating households with a positive diagnosis is more cost effective than testing the whole community[21]. Similar investigations have been performed looking at real world household data for Neisseria meningitidis[22], HIV[23] and influenza, including a meta-analysis of 56 published influenza by Tsang et al. that showed a 38% increase in risk of a secondary infection for household members of a person with a positive diagnosis[24].

Of note are the works of Walker et al. and Geard et al.. Walker inferred parameters for an SIR household model using a Markov Chain Monte Carlo analysis (MCMC) to represent transitions between states of high and low transmission, fitted to FF100 data[25]. FF100 are data from the first 100 households affected by an influenza outbreak so predictions made accurately from these data could make a contemporaneous estimate of the overall size of an ongoing outbreak. This process was both computationally efficient and presented good approximations for the real world parameters.

Geard created an extensive SIR household model tracking over multiple decades of Australian census data[26]. They parameterised factors such as decreasing fertility, increasing life expectancy and increasing rates of household disruption. Geards team made their Python code for this model publicly available, raising the possibility of incorporating it into more complicated household models.

Household structures were influential in the spread of SARS-CoV-2 during the pandemic. Madewell et al. performed a meta-analysis of studies looking at secondary infections in households where SARS-CoV-2 had been introduced. The found 54 such studies and showed an elevated rate of infection once SARS-CoV-2 was introduced to a household. They also broke down these interaction, showing there was further heterogeneity within households[27].

These corroborations of the household principle show the importance of its further investigation. Ball et al. identified 7 possible areas of further mathematical investigation of household models or metapopulations:

1. Clarifying the utility and limitations of weakly coupled large sub-populations

2. Developing a theory of endemic models within the household structure

3. Generating a generalised framework for including household models in more complex socio-economic structures

4. Developing meta-population models that reflect spatial populations

5. Developing inferential methods for emerging epidemics

6. Developing inferential methods for emerging epidemics with computational efficiency
7. Using metapopulations to model the individual as a habitat of the disease

[28]

### Age based models

Household models are an effective explanation for some heterogeneity in the transmission of diseases. Another key factor is age. Multiple studies monitor contact rates between individuals. For example, using RFID chips, Ozella et al. showed that, in households, 70% of interactions include an infant[29]. The POLYMOD study used contact diaries as an alternative self-monitoring method of monitoring contacts[30].

There is a consistent link between young age and high contact rate, both with other younger individuals and people of other ages. This information can be used to predict multiple different phenomena in epidemiology. Rohani et al. showed that an observed change in the average age of pertussis infections following vaccination was not due to new vaccination resistance but was expected due to the age-related heterogeneity in contact rates[31].

There are conflicting conclusions from studies into which age-ranges are at risk of secondary infections in households with diarrhoeal outbreaks. Miura et al. created multiple different permutations of households and generated a complex deterministic model looking at toilet use, washing habits and eating habits to investigate the introduction then spread of Norovirus into households in Japan[32]. They found that mothers and infants were susceptible to the highest rates of secondary infections. Tokunda et al., however, found from questionnaires of households with Shiga-toxin producing Escherichia coli that it was 6-9 year old children, especially males, that are most at risk[33]. They also found that early face-to-face hand-hygiene interventions were most effective in reducing secondary infections, particularly if the household still had a STEC negative child. Whilst these two studies do slightly contradict each other, they show the importance of age when considering transmission of infectious disease.

With outbreaks modelled over short time-frames, the change in ages of a population does not need to be considered. Instead, as with the above models, they can be considered static. However, chronic infectious diseases like HIV and HCV need to be considered on a longer scale and therefore an age-structured model needs to be used.

Age-structured models consider the change of distributions of ages in each compartment as time passes. There are three factors to consider that change as age distributions change. As mentioned earlier, the POLYMOD trial demonstrated how interactions vary between age ranges. The changes in age distributions in the infectious and susceptible compartments will therefore have a direct influence on the transmission rates over time. The birth rate in a population is dependent on the proportion of the population that are of a fertile age. Similarly, the death rate increases as the age of the population increases. Including these factors by representing each population compartment as a size and distribution can therefore be used to show the long-term effects of a disease[34].

Depending on the disease, higher physical contacts among younger populations may not be the age-related risk factor of transmission. Sexually transmitted diseases have transmission rates that are dependent on age, but target a different age group. An example of an age-structured model used to represent disease dynamics can be seen in the work by Bershteyn et al. investigating age-dependent partnering among people who have HIV[35]. They showed that including age into their model greatly adjusted the efficacy of HIV interventions. Early introduction of ART (within one year) saw a higher drop in incidence than predicted from a non-age structured model, but also increased the interventions cost. Additionally, by considering the change of age they showed that this intervention would increase the expected age of incidence.

### Agent-based models

Once age changes are included in a disease model, the logical conclusion becomes to include other factors that may influence infection and death rates. Before intervention, the rate of transmission of HIV between men who have sex with men is higher than any other gender couplings[36], so sexual practices would be an important factor to include when modelling the spread of HIV. Meena et al. showed that smokers had significantly worse outcomes the non-smokers when diagnosed with tuberculosis[37], so if we wanted to create a model to approximate expected outcomes from TB infections, we should include individual's smoking statuses. Agent-based models (ABMs) are models that create individual agents with multiple descriptive factors and consider how they would interact given rules based on these factors[38]. This can be applied to infectious disease models by considering populations not just in terms of compartments but in terms of individuals in these compartments. Each individual is assigned a number of different factors representing their demographics. The chosen demographics are relevant to the modelled disease and dictate how the disease is transmitted to and from the individual, alongside how the disease will affect the individual as time progresses.

This step away from purely equation-based modeling comes with the opportunity for additional complexities. It is now possible to directly track individuals' locations. Venkatramanan et al. used an agent-based model to track the movements of individuals exposed to Ebola, which proved vital in predicting the diseases rate of spread across sparsely populated regions[39]. Hunter et al. used a similar method to model the spread of measels in small towns and villages in Ireland, focusing on students due to their unusual movement patterns and lower probability of vaccination[38]. Lum et al. modelled incarceration as an infectious disease with an SIS process of transmission. Using ABMs to single out race as a factor they explained a racial bias in incarceration frequency in African American populations in California and make a sociological argument against longer incarceration times[40].

The relative flexibility of agent-based models made them incredibly popular when modelling the wide range of different interests surrounding SARS-CoV-2. A PubMed search reveals 139 papers that use ABMs to model some aspect of the SARS-CoV-2 pandemic. These range from including genetic[41] and geospatial data[42], and investigating outcomes such as economic implications of the SARS-CoV-2 pandemic[43] and even the effect of lockdown on lower back pain[44]. One common agent-based model developed specifically written for SARS-CoV-2 modelling is Covasim[45]. Covasim is a open-source Python based tool specifically designed to be both flexible and easy to use. It appears in multiple studies and has an online app, although as of writing we have been unable to access it.

There are two common issues with ABMs. The most important is speed. ABMs are computationally complex, with decisions made for each agent in the model for each time step. This complication increases exponentially when considering each possible interaction between individual agents. This speed will be discussed later when comparing different publicly available modelling programs.

The second noted issue with ABMs is also their strength. They can account for a large amount of variation by accumulating and utilizing multiple pieces of data from real world populations. This requires accurate real world data in order to produce accurate results. This means, as Hunter et al. noted, that a large amount of information must be collected from a population before an accurate conclusion can be made through an ABM[38].

### 1.1.3 Methods of Transmission

In a standard SIR model, we assume transmission occurs when a susceptible individual contacts an infectious individual. As a result, the rate of transmission is assumed to be proportional to the size of the infectious population multiplied by the size of the susceptible population. Whilst this method of contraction is acceptable for air-borne diseases, more complicated methods of transmission often require different means of modelling.

#### Blood-borne diseases

HIV HIV is a blood-borne disease, meaning that it can only be transmitted through the transferal of bodily fluids. For this reason, HIV models have often focused on couples rather than households. Chemaitelly et al. used the Demographic and Health survey data from sub-Sahara African countries to show that couples make up a significant proportion of HIV incidence, with HIV most frequently introduced to a couple through sexual contact outside the couple[46].

Zhang et al. studied the seroconversion of HIV-negative partners in serodiscordant couples in Liuzhou, China using HIV epidemiology databases from 1996-2013. They identified a total of 125 conversion over 4963.5 person-years, giving an incidence of 2.52/100 person-years. In particular, those whose partners had a high CD4 count ($> 350$) had lower risks than those with a low CD4 count ($< 200$) and men with female HIV positive partners, people whos partner did not receive ART and intermittent condom use were linked with an increased seroconversion risk[47]. Additionally, Oldenburg et al. published a study looking at the effect of offering ART to household members of those newly diagnosed with HIV, and found a significant drop in household HIV as a result[23].

Partners should not be the only consideration when modelling HIV, with particular individual factors being shown to increase HIV risk. Fagbamigbe et al. used the 2012 Nigerian population-based HIV/AIDS and reproductive health survey to demonstrate that HIV amongst women in Nigeria was linked with transactional sex, sexual debut before 15 years of age and in women who have been married before. They suggested that these populations could be the target of interventions[48].

The high dependence of partners and specific contacts in HIV transferal show that a household or household/school model would not be enough to describe the character of its spread in communities. This is why specialist HIV modelling tools like TITAN[49] (see later) use networks to map its transmission.

Finally, HIV transmission can also be considered within models for other diseases. For example, Martinez et al. showed a possible decrease in infectiousness among HIV seropositive patients who have a positive diagnosis of tuberculosis. Households in Kampala, Uganda where the index case of tuberculosis was HIV seronegative had a higher rate of latent TB than those that were seropositive. However, this could be due to an increased speed of progress of HIV positive cases leaving less time to infect their households[50]. Velen et al. analysed a program in South Africa that offered HIV counselling and testing to all household members with an index case of tuberculosis alongside all other TB contact tracing. 8.6% of individuals who agreed to testing were new diagnoses thanks to the study[51]. It may be sensible therefore to include the spread of TB into a population model of HIV spread. These examples all demonstrate why the method of transmission must be considered carefully when modelling HIV spread.

Hepatitis C In 2016 the World Health Organisation proposed a target of eliminating Hepatitis C as a public threat by 2030[52]. Pitcher et al. performed a systematic review in 2019 looking at modelling efforts towards understanding the spread of Hepatitis C. They collated papers looking at incarceration and co-infection of HIV. At the time, the models they found did not necessarily show a positive outlook for elimination in a decades time. However, one opportunity they did observe was incarceration, which takes a higher proportion of people who inject drugs (PWID) and in turn Hepatitis C positive patients and therefore represented an opportunity for targeted treatment[2].

Martin et al. performed an extensive review of treatment of PWID with Hepatitis C over 2014 in 7 different sites. They found a range of treatment levels (maximum 26/1000 PWID/year) and used this to support a model evaluating upcoming treatment changes. They found that the maximum treatment level (26/1000) is needed to effect any change over the next 10 years, but the more the better[53].

He et al. also created a model looking specifically at increased screening rates of Hepatitis C in prisons. They found that increasing the screening rates in USA prisons, and screening indiscriminately with regards to age increased diagnosis and treatment in a cost-effective manner (more cost-effective than targeting those born between 1945-1965)[54]. This went against current recommended practice, as in 2012 a policy was created in the USA to offer a one-time screen for Hepatitis C only to anyone born between 1945-1965[55]. It was believed that they made up three quarters of all individuals with chronic Hepatitis C in America, although to the best of this authors knowledge this statistic has only been presented once at The Liver Meeting in San Francisco in 1999[56] and is based on a study that failed to include incarcerated or homeless individuals in their calculations[57].

#### Vector borne diseases

Malaria and dengue fever are not primarily transmitted directly through human contact but are instead transmitted via mosquitoes [58], although vertical transmission can be considered in the case of dengue fever[59]. As a result, rather than modelling transmission through direct contact, researchers have created spacio-temporal models to investigate shared areas of mosquito populations.

Stresman et al. attempted to investigate the presence of hotspot households for malaria transmission among 12 villages in Gambia. They used a geospatial model to show there was no evidence of a link between high prevalence households and high rate of transmission to nearby households. They considered factors that would increase mosquito presence, such as proximity to water and rainfall[60]. Pinchoff et al. evaluated these factors further in Nchelenge District, Zambia. They found a significant increase in malaria risk for each 250m closer a household was to a category 1 stream (origin stream) and a seasonal effect for each 250m closer to a category 2 stream a household was (formed when two category 1 streams join)[61]. Kabaghe et al. also showed the temporal-spatial nature of malaria hotspots, and that they were linked with both a high mosquito population and a high infected mosquito population[62].

Dengue fever has also been shown to follow geographically-weighted mosquito-vector-transmission models. Using incidence in municipalities of Brazil, Rodrigues et al. showed the change in dengue levels in the country over the past decade. They found a link with socio-economically deprived areas and increased dengue risk, as well as coastal regions, urban regions, borders, subtropical climates, poor bin collection and lack of sewer supplies[63]. A novel control method for dengue fever in Guerrero, Mexico is to put larvivorous fish in the household water supply, with the hopes that they would eat mosquito larvae and in doing so reduce the local mosquito population. Morales-Perez et al. showed a reduction in the rate of dengue fever in these households, along with living in a rural setting[64]. They showed that this reduction in the geographical risk was effective, rather that individual risk factors, further strengthening the evidence for use of spacio-temporal models when looking at these vector-borne diseases.

#### Diarrhoeal diseases

Household and age-based transmission rates are to be important when modelling diseases with faecal-oral transmission. For example, Tsang et al. investigated in infectiousness of Norovirus in different household sizes in urban and rural environments. They found that households of smaller sizes (both in space and number of people) experienced a higher secondary attack rate, and urban houses had a higher secondary attack rate than rural settings[65].

Transmission of these diseases is more complicated that purely contact between two individuals. Emont et al. performed an interesting study looking at an epidemic of diarrhoeal diseases found in Tuvalu in 2013 during a drought (due to a lack of laboratory equipment, the exact diagnosis of the disease was not possible). They managed to link the epidemic to a decline in hand-washing techniques. In turn they linked the decline in hand-washing techniques to a lack of water. When a push for hygiene was made, the epidemic died off, long before more water was supplied to the people of Tuvalu. They proposed that this epidemic may be due to the islands inhabitants being more likely to try less clean water supplied in a drought or for lower water supplies to be more easily concentrated with contaminates[66]. Friedrick et al. went further and quantified the risks of different hand-washing techniques in the spread of communicable diseases by evaluating their techniques and seeing the bacterial biome on their hands before and after washing[67].

Rather than modeling transmission as direct contact, Miura et al. produced a deterministic model for household spread of Norovirus across a year. They parameterised interactions between four different person types (mother, father, child and nappy wearing infant). They only looked at diarrhoea based transfer, with factors such as meal times, times to open bowels and which members of a household would help with nappy changes, and modeled a continual external risk from oysters during the fresh oyster season in Japan (November to March). They developed different household designs based on number of different types of humans. They found that mothers and infants were susceptible to a high level of secondary Norovirus infections[32]. The model created shows the extra detail that can be included when considering transmission methods.

### 1.1.4 Closed Environments

Closed environments pose an interesting challenge for modelling the spread of infectious diseases. It is easier to track the relatively rare flow of people in and out of a closed environment, but the small numbers involved means that transmission rates are strongly influenced by stochastic changes. A historical investigation into measles epidemics on ships travelling to Australia in the 1800s found a far lower $R_0$ than is typically quoted for the disease[68]. Paterson et al. studied logbooks written by ships' surgeons (the Garrow, the Roslin Castle, the Trevelyan, the Duntrune and the America). They observed generations of infections in the doctors reports. Using the data they estimated the $R_0$ from parameters that gave the maximum likelihood of each outbreak. The values they found, 7-10, were lower than their expected 12-18[69], a discrepancy that they explained through stochastic factors such as non-homogenous mixing, varying levels of immunity due to largely adult populations and recorded isolation of infectious individuals.

The easier it is to closely follow the spread of closed-environment outbreaks, the better we can note risk factors for transmission. Distasio et al. studied an outbreak of tuberculosis aboard a naval ship in 1987, involving 180 crew members[70]. As this was a naval ship, everyone's berths and workstations were known in the investigation, meaning correlations could be drawn around where people worked and slept. They identified an index case and Distasio et al. showed through simple regression that people working in the index case's department were at higher risk of contracting the disease, while working and berthing in different compartments were protective factors. They also noted that ventilation may have been a factor, studying how the air flowed from the index case's workstation into communal corridors. Additionally, they observed that the index case's workstation was an area of high foot-traffic as it had useful amenities such as a photocopier. Their study showed how powerful information can be extracted from a closed environment outbreak without the need for a full epidemic model. However, with each individual's location being fairly predictable throughout the outbreak, it does seem like a spatial model could be fitted to such a data set to help understand the effect of limited ventilation.

Understanding the physical environment an outbreak takes place in can be critical to understanding the outbreak itself. For example, Lister ei al. investigated an outbreak of colonisation of vanB vancomycin-resistant Enterococcus faecium in a multi-site neonatal unit in Australia[71]. Across the sites, they found 44 colonised babies (31%) followed by a rapid drop in point-prevalence once containment measures were taken. Of note, they also found colonisation sites on baby scales, a baby bath and a pharmacy cupboard, each of which was linked to the outbreak through genetics. In a larger environment an infected area may not contribute significantly to the spread of an infectious disease as much as the movement of people, but where the population size is so small it is likely to have contributed in an unpredictable manner. Holmes and Simmons analused details of an outbreak of viral gastroenteritis on a long-haul flight between Los Angeles and Aukland[72]. They included seat number in their data analysis, and showed the contribution good ventilation had on decreased disease spread, as well as posit the effect that toilet and exit location in relation to the primary case had.

Closed environments also give opportunity to understand features of a disease free from contamination from other infections. For example, the 1950s saw the advent of blood tests for enzymes released by the liver to indicate viral hepatitis infection. At the time, viral hepatitis (individual types of viral hepatitides were not yet recognised), was a diagnosis made either through biopsy, on post-mortem analysis or through clinical observations of jaundice, hepatomegaly and clinical symptoms of a viral illness. However, there was no guarantee of jaundice or hepatomegaly occurring, and other symptoms were easy to mistake for viral gastroenteritis. De Ritis er al. took advantage of an outbreak of viral hepatitis in a institute in Rome to show that amino-transferase levels can be used to diagnose viral hepatitis without jaundice. It was the closed environment, resulting in a relatively high prevalence and no cross-contamination, that enabled them to come to this conclusion[73].

With this lack of cross contamination, we can use the genetic structure of a disease to create its phylogenetic tree and better track its transmission between people. We can then show if an infection is endemic, or as the result of multiple introductory events into the closed environment. Falchi et al. performed a longitudinal study sequencing the HA gene of influenza A on Corsica Island between 2006-2010 and found multiple strains on the island[74]. Using the prevalence of each strain over time and the relative genetic biodiversity of the sequences they identified fixed mutations in each seasons dominant lineage, in keeping with a closed environment. Sequencing Hepatitis D virus RNA from participants in a study on Miyako Island revealed the source of the infection for Arakawa et al.[75]. The homology between the Miyako sequences and the three identified sequences in Japan suggesting a shared lineage. Additionally, the Miyako sequences each had thirteen shared differences compared to the Japanese sequences, indicative that the infection was introduced from mainland Japan to Miyako Island and not the other way around. Finally, Sasaki et al. showed a high level of genetic variation in a Norovirus outbreak on a ship surveying Tokyo Bay[76]. This made it less likely that the outbreak had come from a singular human source and instead was from contaminated seafood[77].

Through a literature review of influenza epidemics in small societies across 120 years[78], Finnie et al. investigated factors that increased risks of larger epidemics for small closed societies[79]. In particular, they were assessing the attack ratio, the proportion of the overall environment that end up having been infected once the epidemic is over. They found that there is a steep drop in the attack ratio as the community sizes increase and that, once again, children represent an important risk factor for disease spread (alongside military personnel). This first observation is in keeping with findings from a prospective study performed by Viboud et al.[80]. They observed that in France during the 1999-2000 winter season, there was a decreasing risk of secondary influenza infections as a household size increased. These studeis both show the importance of introducing smaller heterogeneous mixing groups in infectious disease models, as with the household models discussed earlier in this chapter.

### 1.1.5 Prisons

Prisons represent an environment between completely closed environments and (such as ships) and completely open environments (such as cities). Careful consideration must be made when modelling the spread of diseases in prisons.

Prisons and incarceration are important factors when considering the spread of infectious diseases. A meta-analysis performed by Dolan et al. looked at published papers on the epidemiology of HIV, Hepatitis B, Hepatitis C and tuberculosis in prisons globally. They identified 299 publications covering 196 countries from 2005 to 2016. They demonstrated an elevated level of each of these diseases in prisons almost universally, when compared to the general public, with multiple individual outbreaks demonstrated of HIV, Hepatitis B and tuberculosis. They therefore showed the importance of these institutions when considering infectious disease modelling. Additionally they approximated the infection rate of HIV in prisons, and showed that reducing incarceration rates could contribute considerably to reducing the spread of this disease, a conclusion that could be applied to the other diseases as well[1].

Taylor et al. performed a national sero-behavioural study looking at each closed prison in Scotland[81]. 5187 prisoners volunteered to be tested for Hepatitis C and fill in a behavioural questionnaire. 32% reported a drug injecting history. 53% of PWID were found to be Hepatitis C positive, and only 4 new infections were found (identified as being anti-HCV negative, HCV-RNA positive  a 51-75-day from exposure window [82]). The resulting incidence of Hepatitis C among incarcerated PWID was calculated at 3-4.3%. However, in order to exclude false positives resulting from HCV exposure soon before incarceration, prisoners were excluded from the study if they were under 75 days into their stay. This, combined with the narrow window in which a positive diagnosis of new HCV infection could occur, results in the possibility of the incidence in prisons being higher than calculated.

Using this lowered incidence combined with the shown increased risk of injection related deaths following release from incarceration[83], [84], Stone et al. fitted known PWID HCV incidence data to a deterministic model that mapped both a persons progress through prison systems as well as HCV contraction and treatment[85]. They proposed that if a person had an increased risk of injection related deaths immediately after leaving prison, they would also have an increased risk of contracting HCV. Their model suggested a 45% decrease of Hepatitis C incidence in Scotland should this risk be reduced compared to a 22% reduction in risk if injecting drugs was legalised.

Rather than a lower incidence, Altice et al. observed a higher incidence of HIV and tuberculosis in Eastern European and Central Asia prisons among PWID in a similar study to Taylor et al[86]. They found that incarceration may responsible for up 75% of all TB incidence for people who inject drugs, and 28-55% of all new cases of HIV in the EECA over 15 years from 2016, a difference from the previous studies which needs to be explored further.

PWID are not the only group who are more likely to be incarcerated. Adams et al. used TITAN, the network driven HIV spread modelling tool discussed earlier[49], to explore the effect of increased incarceration on the African American community in Philadelphia[87]. They found that an increased rate of incarceration in the male Africa American population had a knock on effect of an increased HIV risk for the female African American community when they left prison, showing the importance of the external effects of closed environments.

This “spill-over” effect was investigated further by Mabud et al. when looking at tuberculosis in prisons in Brazil[88]. They collected data looking at incidence of TB given time in prison, including incidence at time of release. They then parameterised a compartmental model looking at the effect of people moving in and out of prison. Annual mass TB screening in prisons reduced this models in-prison TB incidence by 47.4% and out-of-prison incidence by 19.4%.

A systematic review of treatment studies for tuberculosis in correctional facilities by Al-Darraji et al. revealed a low level of completion of treatment for TB for those with a diagnosis of latent TB infection (LTBI)[89]. This review revealed a median of 44% completion rate between the studies, with a prevalence in the USA between 6.8-105/100,000 people. They noted that jails and short-term incarceration facilities may in part be responsible for these low rates of completion, as there often may not be enough time to complete the treatment course. Treatment with Isoniazid Preventative Treatment for 6 months if HIV negative and 12 months if positive has been shown to be effective in preventing LTBI developing into active TB[90], [91], which may not be possible to complete for shorter sentences.

There is also a link between incarceration and worse outcomes for HIV. Cohen et al. showed increased odds of mortality and poor HIV outcomes for women who are incarcerated through a longitudinal study[92] and Erickson et al. showed consistent evidence throughout literature that women are then less likely to achieve adequate HIV treatment following incarceration[93]. However, this is not an effect exclusive to women, as any short-stay incarceration is strongly associated with virologic failure in individuals undergoing ART treatment for HIV[94].

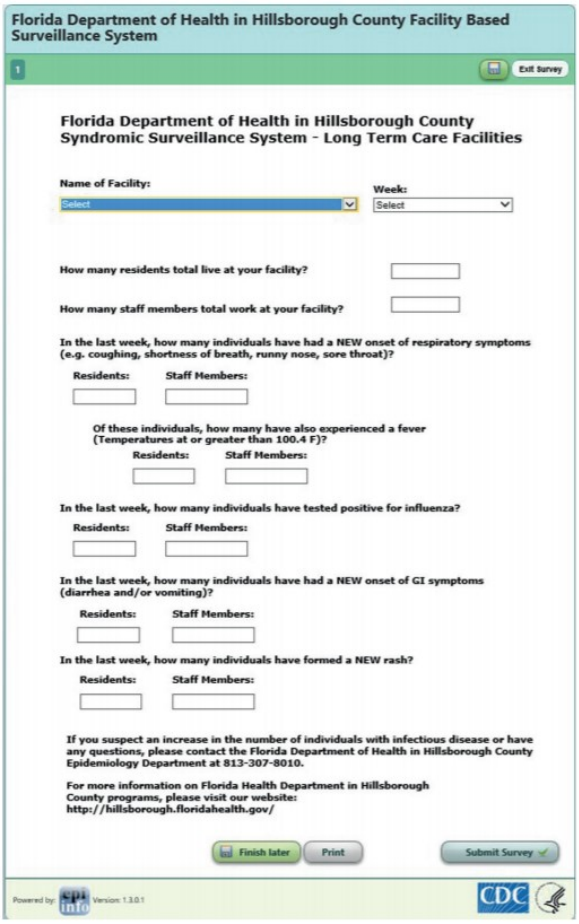
Given the shown importance of incarceration in the spread of infectious diseases, Ndeffo-Mbah et al. performed a systematic review looking for papers published between 1970 and 2017 that model the spread of one or more diseases and include incarceration as part of their model[95]. In total they found 34 models published over this time period, some of which have already been mentioned in this chapter. The overview of these models showed the impact of incarceration across infectious diseases. For example, in communities of people who inject drugs, HIV prevalence greater than 5% resulted in incarcerations being linked to 12-55% of HIV incidence[1], [86]. They also noted that parameter uncertainty was only accounted for in 14 of the 34 models, of which only 8 performed a sensitivity analysis and model fitting and only 1 showed model validation. This gives a good indication as to what is required for future models of infectious disease spread that include incarceration.

### 1.1.6 Software

Multiple tools are available for generic epidemiological analysis. These programs enable users to perform regression analyses and draw links from epidemiological data. With recent advancements, more sophisticated tools have been created that compare epidemiological data to theoretical models. In this section we are going to examine the advantages and disadvantages of a number of available software and review their active use in epidemiological literature.

#### Epi Info

Epi Info was created by the Centers for Disease Control and Prevention in 1985[96]. It is a free-to-access suite of software tools designed for rapid epidemiological data collection, analysis and presentation, with mapping, graphs and report creation. It has been used in studies monitoring multiple diseases, including the spread of haemorrhagic fever in West Africa[97], the range of symptom presentations from nursing homes and assisted living facilities in Florida[98] and the health outcomes of tuberculosis in Kazakhstan[99]. In 2002, 300,000 downloads of the software occurred. It has multiple advantages. Its free access and online and mobile accessibility options for data collection enable a user to gain more information than may have been manageable from in-person surveys. As it is run by the CDC they can use it to focus infor-



Florida Department of Health in Hillsborough County Facility Based Surveillance System

**Florida Department of Health in Hillsborough County Syndromic Surveillance System - Long Term Care Facilities**

Name of Facility: Select

Week: Select

How many residents total live at your facility?

How many staff members total work at your facility?

In the last week, how many individuals have had a NEW onset of respiratory symptoms (e.g. coughing, shortness of breath, runny nose, sore throat)?

Residents: Staff Members:

Of these individuals, how many have also experienced a fever (Temperatures at or greater than 100.4 F)?

Residents: Staff Members:

In the last week, how many individuals have tested positive for influenza?

Residents: Staff Members:

In the last week, how many individuals have had a NEW onset of GI symptoms (diarrhea and/or vomiting)?

Residents: Staff Members:

In the last week, how many individuals have formed a NEW rash?

Residents: Staff Members:

If you suspect an increase in the number of individuals with infectious disease or have any questions, please contact the Florida Department of Health in Hillsborough County Epidemiology Department at 813-307-8010.

For more information on Florida Health Department in Hillsborough County programs, please visit our website: http://hillsborough.floridahealth.gov/

Figure 1.5. An example questionnaire provided to nursing homes and assisted living facilities in Florida to gather data on symptomology of residents[98].

mation gathering on particular epidemics of a global concern, as seen in the haemorrhagic fever study. The multiple language options available and clear questionnaire layout (see Figure 1.5) reduce the difficulties of on-site data gathering. Once the data is collected, Epi Info can package it in a manner that is interpretable to those unfamiliar with computing software or statistical analysis. It is designed with accessibility and ease of use as a priority. This results in accurate data management vital to understanding disease epidemiology.

This comes at a loss of certain complexities. Specific model creation and validation must be done with separate software once the data has been collected. Epi Info is not designed for modelling and instead is to be used for summarising epidemiological data and showing trends. One of its major issues is that it is currently officially only licensed for Windows. This lack of flexibility can cause difficulties with data collection from multiple sources. Camp et al. created a cross-platform package that creates a usable version of Epi-Info on Linux, Mac and Windows, greatly increasing its accessibility[100].

### WINPEPI

PEPI (Programs for Epidemiologists) is a suite of tools for epidemiologists to use both in the field and to train with that was originally designed for calculators in 1983[101]. The original intent was for it to "make life easier for investigators, extend the use of appropriate analytic methods, and enable researchers to concentrate on substantive issues rather than on procedural technicalities". WINPEPI in turn is the implementation of PEPI for Windows[102].

WINPEPI currently consists for 7 programs:

1. COMPARE2  A tool for comparing two independent groups or samples
2. DESCRIBE  A tool for basic descriptive epidemiology
3. LOGISTIC  A tool for multiple logistic regression
4. PAIRSetc  A tool for appraising similarities and differences between matched samples
5. POISSON  A tool for Poisson regression
6. WHATIS  A tool for evaluating expressions, calculating p-values, confidence intervals and time spans
7. ETCETERA  A collection of miscellaneous epidemiological tools not contained in other WINPEPI programs[103]

Like Epi Info, WINPEPI was designed to be a teaching tool as well as a tool for epidemiological investigation. It is designed with ease of use in mind, although its writers do note that an early user may be overwhelmed by the options available. Due to its many options, it has been used to evaluate a range of epidemiological situations, such as the mortality of traumatic chest injuries[104], the changing epidemiology of oral cancers over recent decades[105] and the prevalence of malnutrition among veterans[106]. However, it is not designed for tracking the spread of infectious diseases from person to person. Deterministic or stochastic models must be validated separately.

### TITAN Model

TITAN Model (Treatment of Infection and Transmission in Agent-based Networks), created by the Marshall Research Group at Brown University and not to be confused with the TITAN trial (TMC114/r In Treatment-experienced pAtients Naïve to lopinavir) for HIV treatment[107], is a network modelling tool for the spread of HIV[49]. The team define agent-based network modelling first by defining an agent as an individual person and assigning each individual in a theoretical population with different attributes such as age, gender and socioeconomic status. They then arrange these agents in a network representing interactions between individuals in a population. Finally, they track each agent as events happen, such as contracting HIV from another agent or deciding to get tested for HIV, with each event happening randomly according to probabilities calculated from the demographics of the agent.

The 8,000 line Python 2.7 code that creates this model is adjusted specifically to simulate HIV based interactions by including factors such as a latent period between HIV contraction and detectability. It has been used in two studies by Adams et al., first to show the effect of mass incarceration of African-American men in America on the HIV status of African-American women[87], and then to show how targeted interventions for African-American men post-incarceration may decrease the transmission rate of HIV for African-American women[108]. In this way, Adams has shown how TITAN can be adapted to incorporate incarceration and targeted interventions. Due to the direct nature of transmission of HIV and other blood-borne infections, it can be argued that network models like TITAN are necessary for accurate representation of its spread. Additionally, the flexibility of the model enables the examination of multiple interventions at the same time and the stochastic nature of agent-based network modelling gives the user clear understanding of the variability in the outcome of their model.

The Marshall Research Group discuss drawbacks of the model. The most prominent of these is the speed of the program. As noted earlier, agent-based models are slow compared to compartmental models, with each agent requiring a random decision to be made at each time step rather than treating them all as a whole. The model design is sensitive to inaccuracies, so inaccurate information will result in false results. This means the input demographics from the user must be as accurate as possible and biases may be amplified by the models output. They also note that, due to the stochastic design, TITAN is not well suited for detailed predictions, but rather trends in HIV transmission and treatment rates.

An additional negative of TITAN is that it is written in Python 2.7. This language is considerably different to the up to date Python 3.8.16 which can lead to compatibility issues[109]. Also, this makes TITAN difficult to understand, use and adapt for epidemiologists unfamiliar with Python or coding.

Finally, TITAN is only designed to model the spread of HIV. With knowledge and of the source code its application could be extended to include other infectious diseases, but this is not currently part of its design.

#### EMOD

EMOD (Epidemiological MODelling) is an alternative stochastic agent-based model[110] to TITAN. Rather than modelling transmission of infectious diseases between agents as interactions on a network, EMOD creates the probability of an individual becoming infected based on proportion of individuals already infected in the population. In this way it is like traditional homogeneous mixing models of disease spread.

The key difference between it and deterministic models is, rather than using ODEs to determine the rate at which a population goes from one compartment to another, EMOD uses equations to calculate the probability that an agent will become infected or recover and then, for each agent at each time step, decides if this has occurred. As with TITAN, each agent is assigned specific demographics, with each demographic assigned parameters that will affect the agents journey through the model. With the network removed, EMOD assumes homogeneous mixing between all members of the population.

The advantage of EMOD compared to TITAN is its flexibility. Rather than focusing on one disease, EMOD starts with multiple disease transmission models. There are options for specific diseases, such as HIV, malaria and measles, for which EMOD provides the parameterisation and the probability equations, but these can be adjusted by the user. There is the capability for more experienced users to write their own infectious diseases models for EMOD to use in its calculations. These alterations occur as part of the executable program, not by altering the code itself, making it more accessible for users.

It has featured in multiple papers modelling interventions for different infectious diseases. McCarthy et al. used it to look at the effect varying the point in Plasmodium falciparum's in-human lifecycle at which a vaccine targets has on the spread of malaria[111], Bershteyn et al. investigated how age-dependent partnering changed the spread of HIV[35] and McCarthy et al. showed the risks of type 2 polio vaccination in response to a theoretical outbreak[112].

As observed by Kerr[113], speed, much like with TITAN, is EMODs greatest draw back. As more understanding is gained for the transmission of each disease, the computation time and code complexity required by EMOD to reflect our understanding increases. This makes researchers more likely to avoid ABMs and instead focus on simpler deterministic compartmental models.

#### STEM

STEM (Spatiotemporal Epidemiological Modeler) is an open source tool written in Java which is available on Microsoft, Apple and Linux operating systems[114]. Rather than the ABM of TITAN and EMOD, STEMs simulations focus on compartmental systems, tracking the flow of whole populations from one compartment to the next. This loses some of the specific focus of ABMs but gains computational efficiency.

STEMs largest strength lies in its spatial data. With the tool comes geography, transportation systems and population for 244 countries (this is the number of countries quoted by STEM, although this number is disputed by the United Nations[115]). This rich supply of information has enabled users to model the spread of a disease on national and global scales, including economic impacts of outbreaks and the roles of industry and commerce in disease spread. For example, Edlund et al. combined the geographical data with climate data to predict the changing location of Anopheles, and therefore the global change in incidence of malaria[116]. Its thorough parameterisation has meant that it has also been of interest when modelling theoretical bioterrorist attacks[117].

In general, although it enables intricate modelling over large areas, it is less effective at modelling smaller closed environments where global transport links are not as important. Additionally, whilst it is able to stochastically model process and account for variability, its loss of demographics for agents in its model means its models will not be as representative of individual humans in the real world.

#### PyGOM

A relatively new tool designed for epidemiological modelling is PyGOM[118]. Designed by Public Health England, PyGOM is a Python module for handling ODEs at the code level. It enables the user to find the algebraic expression, Jacobian, gradient and forward sensitivity of an ODE. It can also create a stochastic equivalent of a deterministic ODE model for further analysis and generates a graphical representation of models of easy visualisation and error checking (see Figure 1.6).

A search for PyGOM on PubMed does not reveal any results[6], suggesting that it is not widely used in epidemiological modelling. This is likely in part because it is relatively new. The fact that it must be run as part of the end users Python code may be reducing its accessibility. Its lack of user interface means that it is a versatile tool for someone well versed in Python coding but may be off-putting for users without said familiarity. Additionally, rather than presenting the user with pre-set models, the user is presented with a set of tools with which they can create their own ODE models. Again, this versatility may be a draw for a user with the right knowledge base, but may preventing other people from using PyGOM.

#### Outbreaker2

Outbreaker2 is a package which is designed to aid in reconstructing outbreak transmission trees. Through a Bayesian inference framework it can be used to model a wide variety of different outbreak types, and incorporate details such as phylogeny and spatial data to generate likely transmission trees[119].

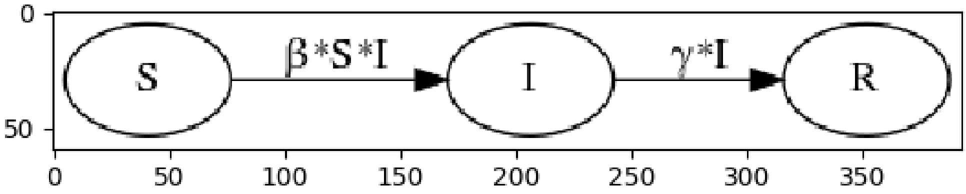


Figure 1.6. Visualisation of an SIR model through PyGOM [118]. This function is useful both for understanding the model and error checking the encoding of an ODE.

The fact that it is so adaptable has meant that it has been used in a variety of different projects. In more recent years, these projects have focused on outbreaks of SARS-CoV-2, such as reconstructing outbreaks in hospital wards in a trust in Switzerland[120], during Iceland's third wave of the SARS-CoV-2 pandemic[121] and from introductions of the Delta variant into Provincetown, Massachusets[122]. However, it is not limited to one disease, and has also been used by Pezzoni et al. to reconstruct transmissions of Swine Vesicular Disease (SVD) between farms in northern Italy. In each case, by iterating through possible transmission trees, the users showed consistent informative relationships in transmissions, such as the importance of healthcare workers in hospital outbreaks and the likelihood of undiagnosed farms in the 2006-2007 SVD outbreak.

Unfortunately, as it is a package in R, Outbreaker2 is not the most accessible tool for infectious disease modelling for individuals who are not familiar with the language or are uncomfortable with coding. Whilst it is exceedingly adaptable, this requires a knowledge and confidence to implement effectively.

### EpiEstim

A more popular R-package designed to investigate outbreaks is EpiEstim. Specifically, EpiEstim uses the rate of new diagnoses to estimate the current reproductive number, otherwise known as the $R_t$ value[123], [124]. This is the ever-changing average number of individuals an infectious person would infect during their entire infectious career. This number is incredibly useful, as a value of $R_t$ greater than 1 implies exponential growth, while a value less than 1 indicates a declining outbreak. However, it is not easy to infer, as it can be obfuscated by factors such as variable delays between infection and diagnosis (and therefore appearing in data-sets). An accurate, accessible tool that makes it easy for the user to estimate the $R_t$ from their data-set would clearly be popular.

EpiEstim has been used on a wide variety of epidemiological modelling projects. A PubMed search reveals 18 separate projects that used EpiEstim since its creation in 2018[125], including projects outside SARS-CoV-2. Whilst in specific circumstances researchers have shown more accurate ways of estimating $R_t$[126], it remains a popular tool, most likely due to its accessibility. It exists both as an R-package and a stand-alone app.

#### Summary of available modelling software

A spectrum can be used to describe the differences between TITAN, EMOD and STEM. The TITAN model is highly specific, in both the choice of disease and its description of individuals in the model. Not only are each of its agents assigned specific demographics, but they are also placed on a network of interactions along which HIV can be spread. This focus provides detailed and sensitive feedback on the effect on specific interventions on HIV epidemiology. This level of focused modelling is computationally expensive and so the program can run comparatively slowly.

STEM, at the other end of the spectrum, removes the information on individual agents and the networks that they interact on and instead divides the population into homogeneous compartments. In doing so it saves on computational complexity allowing it to model larger environments and include external data into its models, such as geography and transport. It is effective when looking at national and global impacts, but the loss of focused detail means it is less accurate with smaller local population sizes.

In between these two on the spectrum is EMOD, which maintains the agent-based modelling of TITAN whilst removing the network-based interactions and allowing agents to transmit diseases between each other homogeneously. Its flexibility enables it to become the more adaptable out of the three.

Finally, PyGOM represents an opportunity for the user to direct exactly at what specificity the model they want to create should be. With this versatility comes a requirement for the user the understand ODE modelling, design the model and utilise it at code level in Python.

### 1.1.7 Aims identified from reviewed literature

We have discussed the importance of accurate models of transmission when considering interventions for infectious diseases. We have considered the relative merits and drawbacks of deterministic compartmental models when compared to stochastic agent-based models. We noted that deterministic models, whilst easier to evaluate mathematically, often cause issues when attempting to parameterise from real world events as they do not allow for stochastic variation among a population. However, we also found that a greater accuracy, such as is seen with stochastic agent-based models, comes at the expense of speed as the models become computationally expensive.

We have also noted an interest in closed environments. Due to the limited movements in and out of these environments, epidemics are easier to track, model and ultimately intervene in. They also do not represent these epidemics entirely in isolation. Prisons, for example, have been shown to contribute to the spread of these diseases outside the prison environment. Thus they are an important opportunity for intervention. For these reasons, future studies could focus particularly on modelling closed environments.

From the modelling programs reviewed (there are others available, such as NetLogo[127], but they have similar strengths and weaknesses) we can draw aims for developing future transmission modelling tools for closed environments:

Accessibility Not every person whom these tools will be aimed at will have a familiarity with either coding or the mathematics behind epidemiological models. For this reason, they must be designed with a user interface with options that can model an environment without the user needing to manipulate the code or enter equations. The language of choice for the tools will be Python due to its high familiarity among biologists. This also opens up the possibility of producing both a tool with a user interface and a Python suite for more experienced coders to manipulate.

Accuracy For multiple reasons stated above, a standard deterministic SIR model should not be used to parameterise epidemic outbreaks. Modelling compartment transitions through exponential decay causes over-estimations of times spent in compartments. Any tools will have to correct for errors like these, for example by finding an alternative mathematical representation of time transitions. It is possible to subdivide compartments and then allow individuals to transit through each of these sub-compartments before transitioning to the next whole compartment. A series of exponential decays stacked in this manner produces an Erlang distribution of decay, preventing a long tail and the over-estimation of time spent in compartments. This technique has been used in infectious disease models in the past[128] and represents one of multiple ways a model could represent changing compartment sizes over time without only using exponential functions.

Speed Another alternative to Erlang distributions is including stochasticity into modelling. Multiple modelling tools calculate the probability of a disease spread at any one time, and then investigate the outcome depending on that probability. This stochastic modelling, repeated over multiple iterations, creates a more realistic picture of disease spread over time, and the multiple iterations give insight into the potential variation in outcomes. It is computationally expensive and as a result these models can take a long time to produce results. Similarly, agent-based models, where each individual in a model population is accounted for, are slow. This will delay output from the tools. Whilst these are currently our most accurate methods for modelling, easily accessible alternatives should be available in the tool should quicker estimates be needed.

The FAIR4RS guidelines advise mainly on the first of these three goals. Based on the FAIR guidelines for data accessibility[129], these criteria are designed to help direct users to making their research tools more available to all. They focus on four main goals:

1. Findable - The research software needs to be stored in a public manner so that it can be found relatively easily.
2. Accessible - The research software, and associated metadata can be retrieved using standard protocols.
3. Interoperable - Information can be transferred from the software to other softwares using standardised methods such as Application Programming Interfaces (APIs).
4. Reusable - Whilst the software can be used for its primary use, it can be readily adapted to suit another user's needs.

[130]

Looking at EpiEstim as an example, we can see the importance of accessibility as a vital part of successful research software. We can use these guidelines to help increase the accessibility of tools we create.

The current aim is to create a tool that models the spread of infectious diseases in closed environments. From there, the tool needs to be applied to real world data, and tested for Accessibility, Accuracy and Speed in order to compare it to current modelling tools.

Ultimately these tools should be able to be extrapolated to alternative settings with alternative diseases. The aim of our future investigation will be to achieve this and thus create an adaptable infectious disease modelling tool for closed environments.

## 1.2 Defining the subject area

“Data-driven analytics of disease spread under close contact for optimal testing and mitigation” covers a broad range of themes in the epidemiology of infectious diseases. We should break down this title to help explain our goals.

Available computational power has grown exponentially since the start of the 21st century and continues to grow. Statistical analysis of data-sets that would have been too large and too noisy to learn from has become not just a possibility but a reality. This has led to multiple fields of research such as artificial intelligence, bioinformatics and beyond. Data-driven analytics simply refers to an examination of a system centred around the data available to the user. The data available affects the way we analyse the system. It informs the conclusions we can draw. It is an important source of the biases influencing our understanding. To understand a system, we need to understand the data that informs us about it.

We specifically want to investigate analyses of disease spread. When referring to disease spread, it would be natural to assume that we are referring to infectious diseases such as SARS-CoV-2. Indeed through much of the thesis we do. However, sometimes it can be useful to think of infectious diseases not as something that is passed directly from an infectious individual to a susceptible individual but as an environmental exposure that everyone experiences. In Chapter 2, for example, we examine the analysis of individuals who have all been exposed to something which may cause symptoms to develop over time. We speculate that whilst this “something” could be an infectious disease, it could just as easily be a poison or something esoteric such as an idea. However, in general it is safe to assume that for most of this thesis we will be considering infectious diseases.

In this thesis, we will be focusing on the spread of infectious diseases in close-contact environments. This covers a broad range of possible locations or outbreak types, but inherently links spread to proximity. When considering spread through proximity, we have a choice. We could consider a scenario where each individual makes connections with each other member of the group with the same rate (or the same probability) as each other. There is no difference between each person and if one person is infectious, every susceptible individual is equally likely to be infected by them. This is called homogeneous mixing and can be true for a small enough scenario. Alternatively, we can consider a scenario where this assumption is not true. An example of this would be a larger, more populous group. In this scenario, proximity becomes more important, because individuals are more likely to make contacts with people near them. This is known as heterogeneous mixing. Infectious individuals in this scenario are more likely to infect one person than another, based on their location.

It is under these circumstances that we want to best understand how best to intervene to prevent further spread of an infectious disease. Interventions can come in a variety of different forms, such as isolation, vaccination and testing. In Chapter 3 we investigate how even a work schedule could be adjusted to limit the length of time an individual is at work whilst infectious.

In summary, this thesis will be looking at many possible ways data around infectious disease outbreaks in small or heterogeneous environments can be used to help inform interventions that can prevent further spread.

The study of the spread of infectious diseases has a few unique properties that makes it unlike any other subset of epidemiology. The most notable of these is the nature of exponential growth. In most cases of infectious diseases, the disease spreads from an infectious host to a susceptible individual. Then this newly infected individual becomes an infectious host. The more infectious hosts there are, the faster they can infect susceptible individuals and therefore the faster the growth of in the infectious group becomes (this is a gross over-simplification of the outbreak process for illustrative purposes).

This growth, where rate at which the infectious population grows is linearly proportional to the size of the infectious population, is know as exponential growth, which is not usually to be seen in other fields of epidemiology. With most other diseases, any intervention that reduces the total number of individuals affected by a proportion does exactly that. However, any intervention that reduces the number of individuals affected by an infectious disease also reduces the number of individuals they go on to infect and in turn the number of individuals those not infected individuals would have gone on to infect and so on. Declan et al. generated a compelling analogous illustration of a row of lit matches demonstrating how preventing one match from becoming lit prevented the remaining matches from lighting as well[131]. Similarly, a negative intervention (or an intervention that results in more individuals becoming infected) will have a knock-on effect that those individuals will go on to infect more individuals.

For clarity's sake, we should say that whilst the illustration by Declan et al. is narratively compelling, it is an imprecise representation of a true outbreak. Most notably, it assumes that all infections occur along a 1-Dimensional line linking us all, rather than the complicated web of connections that make up human interactions. With a disease spreading along a network, one individual removing themselves from the network will not necessarily stop an outbreak completely. However, the effect that they are trying to show, that preventing one infection will reduce the chance of further infections, still applies.

The underpinning mathematics is more complicated than just exponential growth and interventions that fail to fully account for them can lead to disastrous results. Consider temporary interventions of endemic diseases. An endemic disease is an infectious diseases whose rate of infection has reached a equilibrium as there are not enough susceptible individuals to maintain an exponential growth. A temporary intervention may reduce this endemic rate. So long as this intervention remains in place, the level of endemic infection will remain reduced. Over this time, the number of susceptible individuals will accumulate. Hollingsworth et al. showed that if this temporary intervention is relaxed (e.g. due to lack of funding) the accumulated susceptible population can lead to an explosion in the rate of infection to an extent that may be unmanageable[132].

There is another trap when trying to understand the spread of infectious diseases which is more noticeable when considered on a smaller, close-contact scale. On larger scales, we can analytically observe exponential growth and generate relatively simple models that emulate it and can accurately predict how the rate of infection will change over time. However, over smaller scales, the growth of an infectious disease is much more reliant on random stochastic events as an infectious individual interacts with a susceptible individual. This means that the growth over these smaller scales will vary from outbreak to outbreak. When an outbreak is large enough, whilst each of these interaction are happening at random times, enough of them are occurring that we can say that it is analogous to a constant rate, and the variability is reduced.

As a result of this variability, trying to fit a simple exponential curve to a smaller outbreak without considering the range of possible results can lead to wide margins of error. Indeed, smaller outbreaks can be so variable as to raise a concern that information gained from them would not be applicable to any other outbreaks. The one observed outbreak represents just one of a range of many possible outcomes given any parameterisation, and teasing out the exact parameterisation can be analytically complicated and computationally demanding.

We have not even considered the fact that pure exponential growth may not actually occur in a outbreak. The rate of growth is dependent on both the size of the infectious population and the size of the susceptible population. As more individuals as infected, the susceptible population is exhausted and the rate of growth slows. A model that fails to account for the exhausting of the susceptible population and assumes only exponential growth can grossly overestimate the ultimate infectious population size.

The ultimate aim of this thesis is to closer investigate specific scenarios involving outbreaks in close-contact environments in order to aid future researchers‘ analysis. By researchers, we mean both academics and health care providers who would benefit from a better understanding of such outbreaks. We present a suite of possible tools that can be used to analyse multiple aspects of an outbreak and help advise ways to better prevent further spread of an infectious disease.

## 1.3 Research outline

In Chapter 2 we consider a theoretical circumstance where are small group of individuals are exposed to a disease. As time passes and we monitor the number that develop symptoms, we use Bayesian statistics to estimate the true number that have been infected. The simple model grows in complexity as we introduce asymptomatic individuals, a limitless number of exposed individuals and a prolonged exposure window.

As mentioned previously, Chapter 3 investigates the role rota patterns can play in limiting the spread of a disease in the workplace. We assume that an individual has a varying risk of being infected at home compared to work and use this to predict when in a rota pattern they would most likely be infectious. We then investigate how changing a rota pattern changes the length of time they are likely to be infectious whilst at work, as well as when best to test to reduce this value.

Chapter 4 introduces the Nosoco. This a monitoring tool we have developed for estimating the total number of in-hospital "nosocomial" transmissions occurring on a particular ward. This tool extracts hospital ward admissions and pairs them with swab results for SARS-CoV-2 to estimate the total number of transmissions over a time period, the probability that an outbreak has occurred, and the rate at which transmissions are occurring. We first introduces some of the key mathematics underpinning Nosoco and assesses it for speed and how its results compare to standard definitions of nosocomial transmissions. Then we give an example of an analysis that can be performed by Nosoco, comparing the total number of nosocomial transmissions to the total number of community transmissions, and seeing how this relationship changes over time. Finally, we investigate the rate of nosocomial transmissions of SARS-CoV-2. We see how this has changed over time, how it varies between age groups and how best to represent it in an accurate and communicable manner.

Chapter 5 attempts to correct a repeated mistake that has occurred when estimating the incidence of Hepatitis C in prisons from a cross-sectional survey. Previously, this has been reduced to a fairly simple fraction. We demonstrate why and to what extent this calculation is incorrect, propose an alternative approach and show the real need to better understand the length of stay of the prison population in epidemic modelling.

Chapter 6 shows a brief investigation into spacial elements to an outbreak on a cruise-liner, the M.S. Braemar. Outbreaks of infectious diseases on cruise-liners are not uncommon, and understanding how spatial aspects can be involved in an outbreak (in this chapter we look specifically at cabin location) can help advise on screening and isolation when they inevitably occur again.

The final chapter, Chapter 7 reviews each of these tools and explores possible areas of development.

The SARS-CoV-2 pandemic has seen soaring demand for better understanding the spread of infectious diseases. This understanding, whilst useful in the minds of academics, is better placed shared and understood by healthcare workers and policy makers. The tools discussed in this thesis vary between analytical complexity and relative simplicity. It is out hope that we present them each in a way that is accessible to all.

Access to the GitHub repository, including Python code for Chapter 2 and the software behind Nosoco is available on request.

# Chapter 2

# Estimating the total size of an outbreak event from the speed at which individuals develop symptoms

## 2.1 Introduction

Part of the analysis in this chapter has been published in a peer reviewed group paper[133]. Section 2.3 appears in an abridged format in said paper as "Section 2.4: Estimating the size of the first generation from the observed number of symptomatic individuals". This was my contribution to a group paper looking at multiple alternative analytical methods available early in a pandemic. In this chapter we expand far beyond the concepts raised in the original paper. Early on we will highlight when this analysis moves away from the analysis in the paper.

Early analysis of an outbreak event, when we are between the first exposure event and all of the first generation of individuals developing symptoms, is highly variable. The data available is heavily censored by time. It is difficult to know if an individual has not yet developed symptoms or become identifiable as infected because they were not in fact infected or because not enough time has passed. This censoring becomes more important the longer a disease's expected incubation period is.

Analyses of small outbreaks in closed environments can bring interesting insights into aspects of the disease process, such as incubation periods[134]–[138] and modes of transmission[139]–[141]. However, these studies are retrospective. During the outbreak itself, we may only be able to answer simple questions, such as "How many people do we think are currently infected?".

This question was asked early on in the SARS-CoV-2 global pandemic, before the virus was fully established in the UK. Sporadic stochastic introductory events occurred as travellers unwittingly brought the disease from other countries. Once these people were identified, the inevitable question becomes "How many people did they infect?". When the disease's prevalence was minimal, it was theoretically practical to trace all contacts and identify people who had been exposed (although in practice there were multiple issues with effective contact tracing[142], [143]). However, with a delay between transmission and symptom onset, and at a time where testing was not readily available, how could we tell if asymptomatic individuals were not infected or were just yet to develop symptoms? Asymptomatic infection, where an infected individual will never develop symptoms, is another wrinkle to this problem which we will discuss later.

To demonstrate the logic behind answering these questions, I will present two intentionally extreme scenarios. Firstly, we can imagine that it has been 99 days since an individual was exposed to a disease like SARS-CoV-2 and they still have not developed symptoms. In a world where it was impossible to be infected and not eventually develop symptoms, we would conclude that this individual was not originally infected. After a certain length of time, our intuition tells us that the incubation period of SARS-CoV-2 cannot be that long, so we conclude it is exceedingly unlikely that the individual was infected (although more insidious diseases, such as HIV or Hepatitis-C, may result in incubation periods this long, or longer).

Our second scenario has 100 people receiving the same "level" of exposure. It is difficult to imagine 100 people all being exposed to one infectious individual to the same level, unless they are an exceedingly potent viral-shedder, so I find it easier to imagine that they were all in the same room with radioactive material or all ate the same poisoned meal. The results are the effectively identical. We still have a group of exposed individuals, it is unclear who has been affected and there is a delay between being affected and developing symptoms. If, in under 5 minutes, 99 out of the 100 exposed individuals start developing symptoms and we were asked to guess if the 100th individual would develop symptoms, our intuition tells us that yes, they probably will (in the case of the poisoning scenario, perhaps a more sceptical mind may suspect the 100th individual of being the poisoner, but that is a complication that we are not trying to demonstrate here). The more people who develop symptoms, the more certain we get that those that have not developed symptoms will go on to do so. Of course it may be that there is some variability in the level of exposure, as has been explored in models by Pratt et al[144]. The reason why this possibility is not relevant to our model is explored in Section 2.3.6.

So what happens if we combine those two scenarios? 100 individuals were exposed, 99 of whom developed symptoms in under 5 minutes, but the 100th person has remained symptom-free for 99 days. At some point over those 99 days our assumption that the 100th person was also infected will switch and we will instead be convinced that they avoided infection. When does this switch occur?

Before we define our scenario, it is worth being specific about how we define individuals. Given an exposure, individuals can be in one of four states:

1. Not infected - Despite exposure, this individual was not infected.

2. Pre-symptomatic - As a result of exposure, this individual has been infected. They have not yet developed symptoms. At some time in the future they will develop symptoms. From an observation stand-point, this individual is temporarily indistinguishable from a Not infected individual.
3. Symptomatic - As a result of exposure, this individual has been infected. Their incubation period has finished and they can now be observed to be displaying symptoms. They are known to be infected.
4. Asymptomatic - Despite being infected, this individual will never develop symptoms. They will permanently be indistinguishable from Not infected individuals without testing. To simplify our initial analysis, we will assume that this option is not possible.

This work was started at a time in the SARS-CoV-2 pandemic when testing was not readily available. We will refer to individuals as being "Pre-symptomatic or Symptomatic, but note that this could now be extended to Pre-detectable or Detectable. Figure 2.1 demonstrates each of these possibilities.

We start with the simplest scenario. A group of individuals of size $n$ were exposed exactly at time $t = 0$. They each have the same unknown probability $\rho$ of being infected as a result of this exposure. If they were infected, this is the only time it could have occurred. A total of $E_0$ individuals are infected at this time, although this number is unknown. They cannot infect each other - As such this method can also apply to non-infectious exposures, such as radiation poisoning or, to some extend, advertising. We do not know how many were infected from this exposure, nor do we know how infectious the disease normally is. $F(\tau)$ is a cumulative density function that describes the length of an infected individual's incubation period. This is also equal to the probability that an individual infected at time 0 will have developed symptoms by time $\tau$. This function is the same for all infected individuals and is not dependent on the probability that they were infected. In other words, an increased or decreased exposure will not increase or decrease the time it takes for any one infected individual to develop symptoms. Similarly, if the group were particularly intrinsically vulnerable (e.g. frailty, immunocompromised), this would not affect the time it takes for any one infected individual to develop symptoms. As $\tau$ increases, $F(\tau)$ will tend towards 1. The probability that an infected person will have developed symptoms before time $\tau$ always tends to 1 as $\tau$ increase, so long as it is impossible for individuals to remain asymptomatic indefinitely.

This is very similar to the work done by Egan & Hall[138] in order to approximate the incubation period distribution of a disease from a closed outbreak incident. Their theoretical set-up is the same, except that the function $F(\tau)$ is unknown and $E_0$ is known. By seeing at what time infected individuals become symptomatic they demonstrate how to fit a distribution for the incubation period.

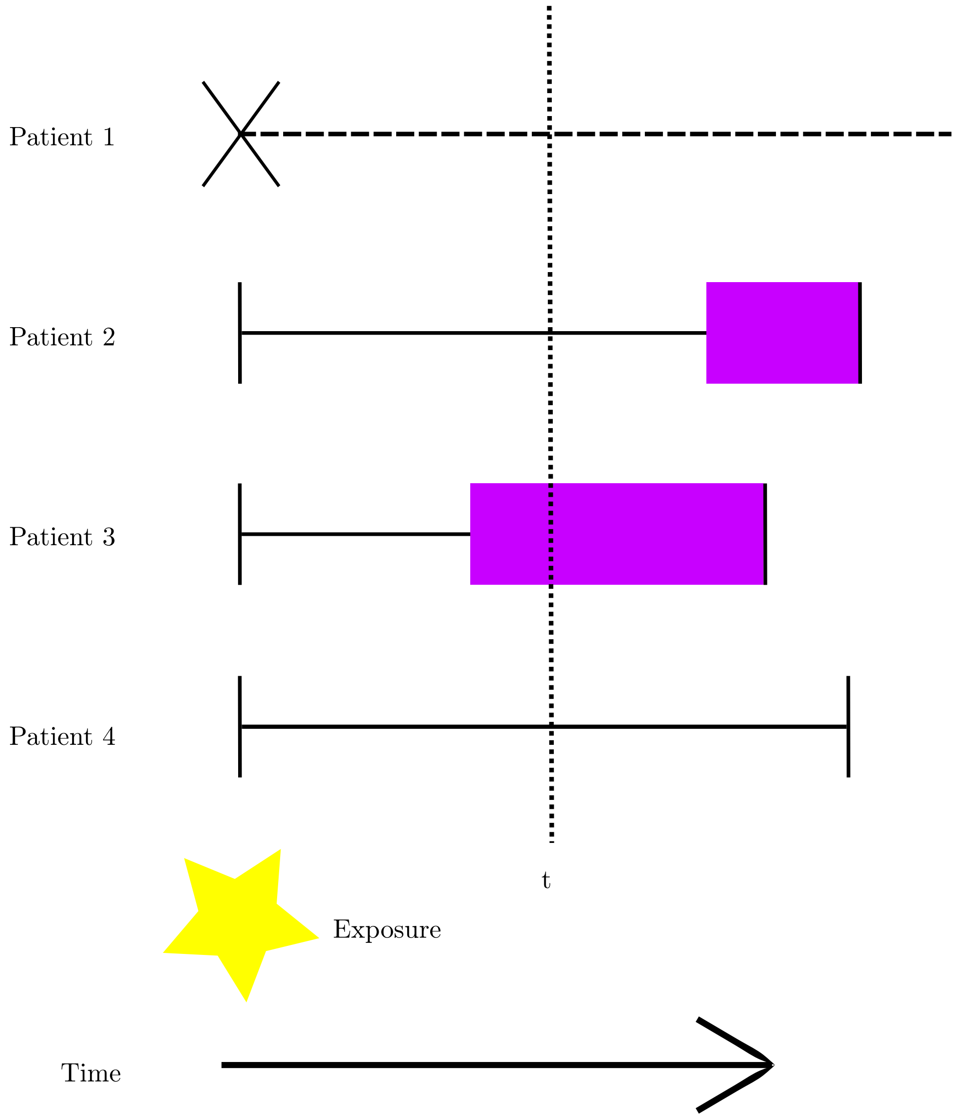


Figure 2.1. Possible observed scenario following a single exposure event. Each of the four patients are exposed at the same time. Patients 2-4 are infected. Patient 1 will never develop symptoms because they were not infected - Not infected. Patient 2 will eventually develop symptoms (represented by the purple block) but not before time $t$ - Pre-symptomatic. Patient 3 has developed symptoms by time $t$ - Symptomatic. Patient 4 was infected but will never go on to develop symptoms - Asymptomatic. At time $t$, Patients 1, 2 and 4 are indistinguishable.

As time passes, we record the number of individuals out of the original $n$ that have developed symptoms. $I_t$ is the total number of individuals that have developed symptoms at some point prior to time $t$. Every infected individual starts as pre-symptomatic and will eventually develop symptoms (i.e. it is currently impossible to be an asymptomatic individual). We could record the exact time each infected individual becomes symptomatic. By time $t$, $\mathbf{J}_t$ is a set of size $I_t$, which contains all times at which individuals were observed to become symptomatic prior to time $t$. However, with this current model this is not necessary in order to make the most accurate inference of $\rho$, as will be demonstrated later.

Using $n$, $I_t$ and $F(t)$, we aim to show the probability that the outbreak size is any value between $I_t$ and $n$ as $t$ increases, and how outbreak size, group size and proportion of individuals that are symptomatic by time $t$ affects this estimate. We show two special cases: the probability that the outbreak size is equal to $I_t$ (i.e. there are no remaining Pre-symptomatic or Asymptomatic individuals) and the probability that the outbreak size equals 0 (this probability is only greater than 0 if at time $t$, $I_t = 0$). We then consider the confounding scenarios where

1. individuals who are infected may never develop symptoms (Asymptomatic individuals are possible)
2. the total number of exposed individuals is unknown
3. the exposure is ongoing.

## 2.2 Parameter and Function description

Table 2.1 is a list of relevant parameter and function definitions for this chapter. The reader may find it useful to refer back to this table as and when required.

## 2.3 Model 1: Point Exposure with 100% symptomatic rate

### 2.3.1 Calculating a posterior distribution for the value of the probability that an individual was infected given a known distribution for the incubation period

In the case where we have no prior understanding of the distribution of possible values of $\rho$, aside from the fact that they fall somewhere between 0 and 1, we start by choosing an uninformative Jeffreys prior distribution for the value of $\rho$ to minimise our prior's effect on our posterior distribution[145]. A Jeffreys prior distribution for a parameter in a probability distribution can be calculated as the square root of the determinant of the distribution's Fisher's information matrix. In turn, where mathematically possible, the determinant of the Fisher's information matrix can be calculated as the negative expected value for the second derivative of the log of the probability density function (or mass function) with respect to the parameter given the value of the parameter[146]. All prior distributions affect our posterior distribution in some way. By choosing a Jeffreys prior we have chosen a prior that minimises the effect our prior has on our posterior calculations. In a simple Bernoulli trial of probability $\theta$, the Jeffreys prior for $\theta$ is a Beta distribution $\theta \sim Beta(0.5, 0.5)$[147]. If $x$ is the observed result of a Bernoulli trial with a probability of $\theta$ of success, where $x = 1$ if the trial is a success and $x = 0$ otherwise, then we can write the probability mass function for $x$, $g(x; \theta)$:

$$g(x;\theta) = \begin{cases} \theta & x = 1 \\ 1 - \theta & x = 0 \end{cases}$$

| Parameter/Function | Description |
|---|---|
| ${}_2\mathcal{F}_1(a,b,c,z)$ | The Hypergeometric 2F1 function |
| $a$ | The probability of developing symptoms given infection |
| $B(a,b)$ | The Beta function |
| $B_x(a,b)$ | The lower incomplete Beta function |
| $E_0$ | The number of individuals that were infected during an exposure at time $t=0$ |
| $E_t$ | The total number of infected individuals by time $t$ |
| $f(\tau)$ | The probability density function for the incubation period distribution |
| $F(\tau)$ | The cumulative density function for the incubation period distribution |
| $g(x;\theta)$ | A generic probability mass function describing the probability of outcome $x$ given parameter set $\theta$ |
| $\mathbb{H}$ | The Shannon entropy |
| $H(t)$ | The cumulative hazard function for infection given an ongoing exposure since time $t=0$ |
| $\mathbb{H}_{norm}$ | The normalised Shannon entropy |
| $I_t$ | The total number of symptomatic individuals by time $t$ |
| $\mathbf{J}_t$ | A set of all symptom onset timings for individuals that developed symptoms before time $t$ |
| $k(t)$ | The probability density function describing the distribution of time from start of exposure to developing symptoms |
| $K(t)$ | The cumulative density function describing the distribution of time from start of exposure to developing symptoms |
| $\mathbb{L}(\theta;x)$ | A generic likelihood function giving the proportional likelihood of input parameter set $\theta$ given observation $x$ |
| $\Lambda$ | The rate parameter for the Poisson distribution that describes the total number of infectious connections an individuals makes |
| $\mu$ | The mean probability of infection given variation between individuals |
| $n$ | The total size of the exposed group |
| $od$ | The overdispersion of the probability of infection, defined as the variance divided by the mean |
| $\omega$ | The rate parameter when the incubation period distribution is described as an exponential distribution |
| $\pi(\theta)$ | A prior distribution describing the believed likelihood of parameter $\theta$ prior to any observation |
| $\rho$ | The probability that an exposed individual has been infected |
| $\rho_t$ | An individual's probability of being infected before time $t$ |
| $\sigma$ | The standard deviation of the probability of infection given variation between individuals |
| $t$ | The amount of time that has passed since the start of the observation period |
| $\tau$ | The amount of time that has passed since an individual was infected. When an individual was infected at the start of the observation period, $\tau = t$ |
| $\mathbb{W}_p$ | The Wasserstein-$p$ distance |
| $\mathbb{W}_{prop}$ | The proportional Wasserstein-1 distance |
| $x_t$ | The observed outcome for an individual at time $t$, which equals 1 if they are already symptomatic and 0 otherwise |
| $y(t)$ | The probability deinsity function for the time of infection given an ongoing exposure starting at time $t=0$ |
| $z$ | The timing of the end of an exposure period |

Table 2.1. A complete description of parameters and functions used in this chapter

It is this function that is then used to calculate the Jeffreys prior. As stated earlier, the Jeffreys prior is defined as the square root of the determinant of the Fisher's information matrix of a probability mass function, which in turn has been show to be the second derivative of the natural Log of the probability mass function with respect to its parameterisation (we have previously defined it in terms of the likelihood function $\mathbb{L}(\theta; x)$. As this function is proportional to the probability mass function for $x$ and we normalise our prior distribution across all possible values between 0 and 1, we cam use the two terms interchangeably in this specific circumstance).

$$\begin{aligned}
\pi(\theta) &= \sqrt{-\mathbb{E}\left(\frac{\mathrm{d}^2 \ln\left[g\left(x;\theta\right)\right]}{\mathrm{d}\theta^2}\right)} \\
&= \sqrt{-\begin{cases} \theta \times \frac{\mathrm{d}^2 \ln[\theta]}{\mathrm{d}\theta^2}, & x = 1 \\ + & \\ (1-\theta) \times \frac{\mathrm{d}^2 \ln[1-\theta]}{\mathrm{d}\theta^2}, & x = 0 \end{cases}} \\
&= \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}
\end{aligned}$$

However, in our model, where we can only observe infected individuals if they have already developed symptoms, the prior distribution needs to be adjusted. $x_t$ is the observation of a symptomatic individual such that $x_t = 1$ if the individual develops symptoms some time before time $t$ and $x_t = 0$ otherwise. The probability that $x_t = 1$ is equivalent to the probability that they were infected ($\rho$) and that they developed symptoms prior to time $t$. As in this scenario they can only be infected at time 0, the probability that they develop symptoms prior to time $t$ given that they were infected is equivalent to the probability that their incubation period is less than $t$. The cumulative density function that describes the probability that an individual's incubation period is less than $\tau$ is $F(\tau)$. Therefore, the probability that an individual is symptomatic prior to time $t$ is $\rho F(t)$. This now gives us a new probability mass function for $x_t$ given $\rho$:

$$g\left(x_t;\rho\right) = \begin{cases} \rho F(t), & x_t = 1 \\ 1 - \rho F(t), & x_t = 0 \end{cases}$$

The term $1 - \rho F(t)$ encapsulates both the probability that the individual was not infected and that they were infected but are yet to develop symptoms. Using this probability mass function gives us a Jeffreys prior that is dependent on $t$, or more specifically $F(t)$.

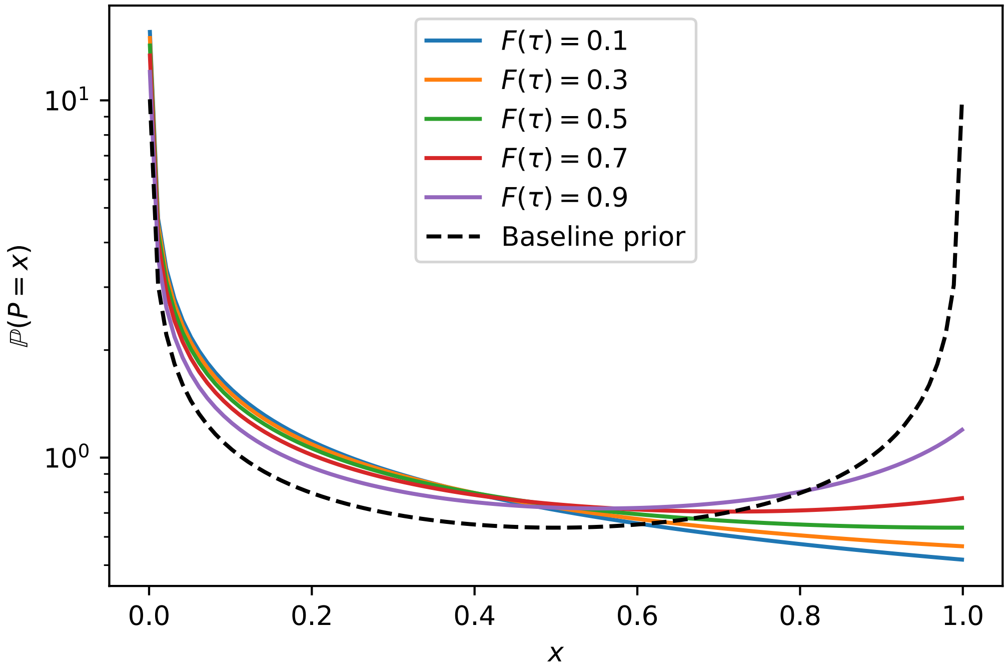


Figure 2.2. Demonstration of Jeffreys priors for a censored Bernoulli trial. The probability of observing a success is censored by $F(t)$, the probability of observing a positive result. In our model this is equivalent to the probability of an incubation period of a length less than $t$ (e.g. The line for $F(t) = 0.1$ demonstrates the Jeffreys prior for $\rho$ when observing an exposed group at time $t$ such that the probability of an infected individual developing symptoms prior to time $t$ is 0.1) Each prior has been normalised by dividing by the constant $2\sin^{-1}\left[\sqrt{F(t)}\right]$. A logarithmic y-scale has been used for clarity.

$$
\begin{aligned}
\pi(\rho|F(t)) &= \sqrt{-\mathbb{E}\left(\frac{\mathrm{d}^2 \ln\left[g\left(x_t;\rho\right)\right]}{\mathrm{d}\rho^2}\right)} \\
&= \sqrt{-\begin{cases} \rho F(t) \times \frac{\mathrm{d}^2 \ln[\rho F(t)]}{\mathrm{d}\rho^2}, & x_t = 1 \\ + & \\ (1-\rho F(t)) \times \frac{\mathrm{d}^2 \ln[1-\rho F(t)]}{\mathrm{d}\rho^2}, & x_t = 0 \end{cases}} \\
&= \rho^{-\frac{1}{2}} F(t)^{\frac{1}{2}} (1-\rho F(t))^{-\frac{1}{2}}
\end{aligned}
$$

Figure 2.2 shows the Jeffreys prior for multiple values of $F(t)$ alongside the baseline Jeffreys prior for an uncensored Bernoulli trial, equivalent to $F(t) = 1$. The uncensored prior distribution is symmetrical around line $x = 0.5$. For values of $F(t)$ less than 1 this symmetry is lost, favouring lower values of $\rho$. Whilst similar in structure, these prior distributions are not strictly Beta distributions.

We can find the posterior distribution for $\rho$ given the total number of symptomatic individual by time $t$, $I_t = i_t$, by stating:

$$\mathbb{P}\left(\rho|I_t = i_t, F(t)\right) \propto \mathbb{P}\left(I_t = i_t|\rho, F(t)\right) \times \pi\left(\rho|F(t)\right)$$

$$= \binom{n}{i_t} (\rho F(t))^{i_t} (1-\rho F(t))^{n-i_t} \times \rho^{-\frac{1}{2}} F(t)^{\frac{1}{2}} (1-\rho F(t))^{-\frac{1}{2}}$$

Calculating a normalising constant gives this posterior distribution:

$$\mathbb{P}(\rho | I_t = i_t, F(t)) = \frac{\rho^{i_t - \frac{1}{2}} F(t)^{i_t + \frac{1}{2}} (1-\rho F(t))^{n-i-\frac{1}{2}}}{B_{F(t)}(i_t + \frac{1}{2}, n - i_t + \frac{1}{2})} \tag{2.1}$$

where $B_x(a,b)$ is the lower incomplete Beta function $\int_0^x t^{a-1}(1-t)^{b-1}$[148]. This differs from the posterior distribution calculated in Overton et al.[133], the original paper this work was published in, which uses a Uniform distribution between 0 and 1, equivalent to the Beta distribution $\beta(1,1)$, for their prior distribution to give:

$$\mathbb{P}(\rho | I_t = i_t, F(t)) = \frac{\rho^{i_t} (i_t + 1) (1 - \rho F(t))^{n-i_t}}{{}_2\mathcal{F}_1 (i_t + 1, i_t - n, i_t + 2, F(t))}$$

where ${}_2\mathcal{F}_1(a,b,c,z)$ is the Hypergeometric 2F1 function[148]. We will continue to use the Jeffreys prior, as it is both numerically easier and has less of an influence on our posterior calculations, so there may be some differences between here and the paper.

But what if we wanted to include the exact timings that each infected individual developed symptoms prior to time $t$? This may seem like it would give us a more accurate posterior distribution, but in fact it does not. This is because the exact timing of symptom onset for any one infected individual is independent of the size of their exposure or the probability that they have been infected. If multiple people become symptomatic early in our model, this indicates a larger value of $\rho$ only because it is indicative of a larger total number of samples of the incubation period, not because of some intrinsic link between the incubation period and the probability of infection. This means that, in terms of this model, two scenarios for the same group, one where $I_t$ individuals immediately became symptomatic after exposure and one where $I_t$ individuals became symptomatic just before time $t$, will have the same predictions by our model after time $t$ (see Figure 2.3).

We can show this analytically. The possible observation $x_t$ now represents two terms, $\delta$, which is 1 if the individual becomes symptomatic before time $t$ and 0 otherwise, and $j$, which is the time at which the individual became symptomatic if they became symptomatic before time $t$, and 0 otherwise. With this observation, the probability density function for $x_t$ is changed:

$$g(x_t; \rho) = \begin{cases} \rho f(j) & \delta = 1; 0 \leq j < t \\ 1 - \rho F(t) & \delta = 0 \end{cases}$$

where $f(\tau)$ describes the probability density function of the disease's incubation period. For the sake of analysis we choose to use the same prior distribution. For an

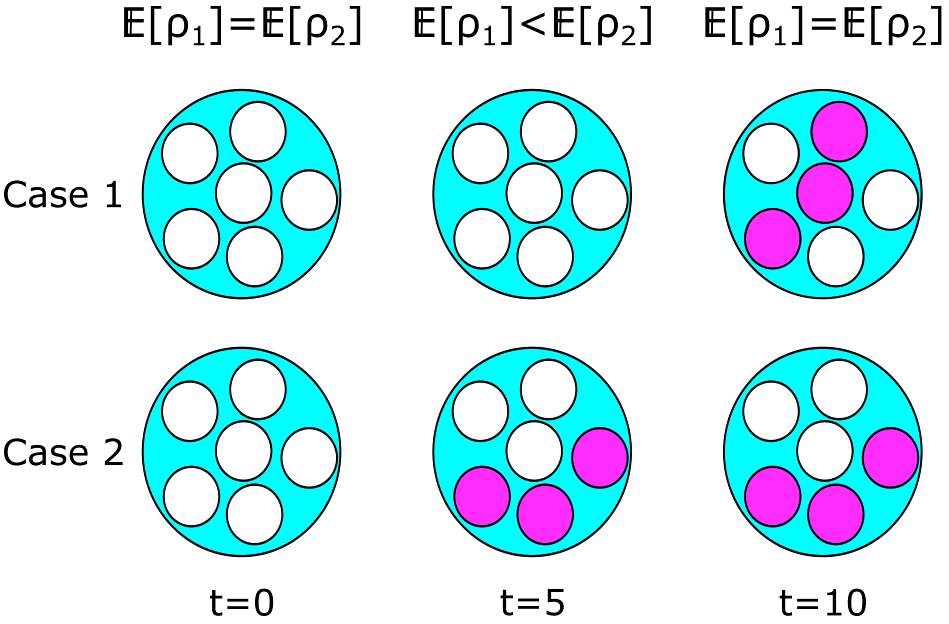


Figure 2.3. Estimation of infection probability $\rho$ at time $t$ is independent of the exact timing of symptom onset, and instead relies on the total number of symptomatic individuals by time $t$. Two groups of the same size ($n = 6$) are both exposed at time 0. Each individual is represented by a circle, which is purple if they have developed symptoms and white otherwise. By time $t = 5$, three individuals have developed symptoms in Case 2, but none have developed symptoms in Case 1. We would estimate the probability of infection to be higher in Case 2 ($\mathbb{E}[\rho_1] < \mathbb{E}[\rho_2]$). However, by time $t = 10$, the two cases have the same number of symptomatic individuals so we would generate the same posterior estimate for $\rho$ in both cases, regardless of when each individual became symptomatic.

observed set of $x_t$, the exact timings at which individuals developed symptoms prior to $t$, we will appear to need to adjust our calculation for the posterior distribution of $\rho$:

$$
\begin{aligned}
\mathbb{P}(\rho|x_t, t) &\propto \mathbb{P}(x_t|\rho, t) \times \pi(\rho|F(t)) \\
&= \left( \prod_{q=1}^{n} \left(\rho f\left(j_q\right)\right)^{\delta_q} \times \left(1 - \rho F(t)\right)^{1-\delta_q} \right) \times \rho^{-\frac{1}{2}} F(t)^{\frac{1}{2}} (1 - \rho F(t))^{-\frac{1}{2}} \\
&= \left( \prod_{q=1}^{n} f\left(j_q\right)^{\delta_q} \right) \rho^{I_t - \frac{1}{2}} F(t)^{\frac{1}{2}} (1 - \rho F(t))^{n - I_t - \frac{1}{2}}
\end{aligned}
$$

With this new product term $\prod_{q=1}^{n} f(j_q)^{\delta_q}$ it may appear that we have a new, more accurate posterior distribution. However, when a normalising constant is found such that the integral of $\rho$ between 0 and 1 is equal to 1, this product term integrates out leaving us with the same posterior distribution seen in Equation 2.1. The posterior distribution for $\rho$ is independent of the exact timings that individuals become symptomatic, and is instead dependent on the total number that are symptomatic prior to the observed time $t$.

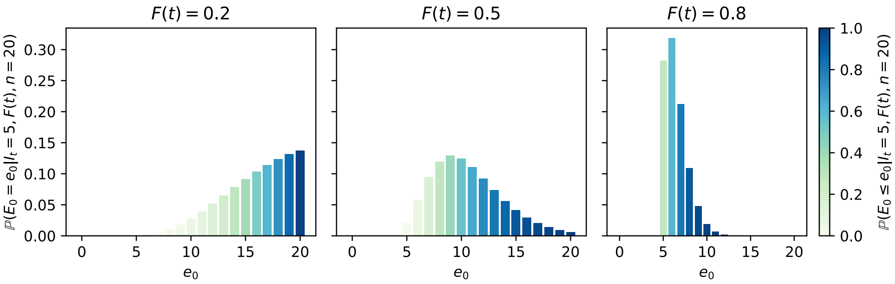


Figure 2.4. The probability that $e_0$ individuals were infected given $I_t$ are symptomatic out of a group of $n$. All three graphs show the results of calculations for a group of $n = 20$ individuals, of which $I_t = 5$ are symptomatic. We can see how this calculation changes over time, with the left plot representing a time when $F(t) = 0.2$, the middle plot representing when $F(t) = 0.5$ and the right plot representing when $F(t) = 0.8$.

### 2.3.2 Calculating a probability distribution for the total number of individuals infected given a known probability of being symptomatic before the observed time

In the case that we know the value of our censoring probability $F(t)$, we can be specific regarding the exact probability of each eventuality.

$$
\begin{aligned}
\mathbb{P}\left(E_0 = e_0 | I_t = i_t, F(t), n\right) &= \int_0^1 \frac{\mathbb{P}\left(E_0 = e_0 \;\; \& \;\; I_t = i_t | \rho, F(t)\right)}{\mathbb{P}\left(I_t i_t | \rho, F(t)\right)} \\
&\quad \times \mathbb{P}\left(\rho | I_t = i_t, F(t)\right) \mathrm{d}\rho \\
&= \int_0^1 \frac{\binom{n}{e_0} \rho^{e_0} \left(1-\rho\right)^{n-e_0} \binom{e_0}{i_t} F\left(t\right)^{i_t} \left(1 - F\left(t\right)\right)^{e_0 - i_t}}{\binom{n}{i_t} \left(\rho F\left(t\right)\right)^{i_t} \left(1 - \rho F\left(t\right)\right)^{n-i_t}} \\
&\quad \times \frac{\rho^{i_t - \frac{1}{2}} F(t)^{i_t + \frac{1}{2}} \left(1 - \rho F\left(t\right)\right)^{n - i_t - \frac{1}{2}}}{B_{F(t)} i_t + \frac{1}{2}, n - i_t + \frac{1}{2})} \mathrm{d}\rho \\
&= \binom{n - i_t}{n - e_0} \frac{F(t)^{i_t + \frac{1}{2}} (1 - F(t))^{e_0 - i_t}}{B_{F(t)}(i_t + \frac{1}{2}, n - i_t + \frac{1}{2})} \\
&\quad \times \int_0^1 \rho^{e_0 - \frac{1}{2}} (1-\rho)^{n - e_0} (1 - \rho F(t))^{-\frac{1}{2}} \mathrm{d}\rho \\
&= \frac{B\left(n - e_0 + 1, e_0 + \frac{1}{2}\right) F(t)^{i_t + \frac{1}{2}} \left(1 - F(t)\right)^{e_0 - i_t}}{(n - i_t + 1) B\left(n - e_0 + 1, e_0 - i_t + 1\right)} \\
&\quad \times \frac{{}_2\mathcal{F}_1\left(\frac{1}{2}, e_0 + \frac{1}{2}, n + \frac{3}{2}, F(t)\right)}{B_{F(t)}\left(i_t + \frac{1}{2}, n - i_t + \frac{1}{2}\right)} \qquad (2.2)
\end{aligned}
$$

where $E_0$ is the total number of individuals that were infected, $n$ is the total number of individuals that were exposed, $I_t$ is the total number of individuals that are symptomatic by time $t$ and $F(t)$ is the probability that an individual infected at time 0 is symptomatic by time $t$.

We could show the effect of including the exact timings of symptom onset, but this turns out to be unnecessary information of the same reasons as discussed previously.

### 2.3.3 Performance analysis

We have chosen two metrics that we feel best evaluate our estimates: The Wasserstein-1 metric[149] and the Shannon entropy[150], [151].

#### Wasserstein-1 metric

The Wasserstein-1 can be used to compare how well two distributions match. It is a measure for the amount of change required to transform one distribution into another. Let us consider two independent random variables, $\mathbf{X}$ and $\mathbf{Y}$. These random variables are sampled from two different probability distributions, who have cumulative density functions of $G_{\mathbf{X}}(x)$ and $G_{\mathbf{Y}}(y)$ respectively. Their quantile functions, $G_{\mathbf{X}}^{-1}(q)$ and $G_{\mathbf{Y}}^{-1}(q)$ are the inverse of their cumulative density functions. We calculate the Wasserstein-$p$ metric as:

$$\mathbb{W}_p = \left( \int_0^1 \left| G_{\mathbf{X}}^{-1}(q) - G_{\mathbf{Y}}^{-1}(q) \right|^p \mathrm{d}q \right)^{\frac{1}{p}}$$

In the case of the Wasserstein-1 metric, this integral captures the absolute area between the two quantile curves. As demonstrated in Figure 2.5, this is equal to the absolute area between the two cumulative density function curves. This mean, in the special case of the Wasserstein-1 metric, we can calculate this distance as:

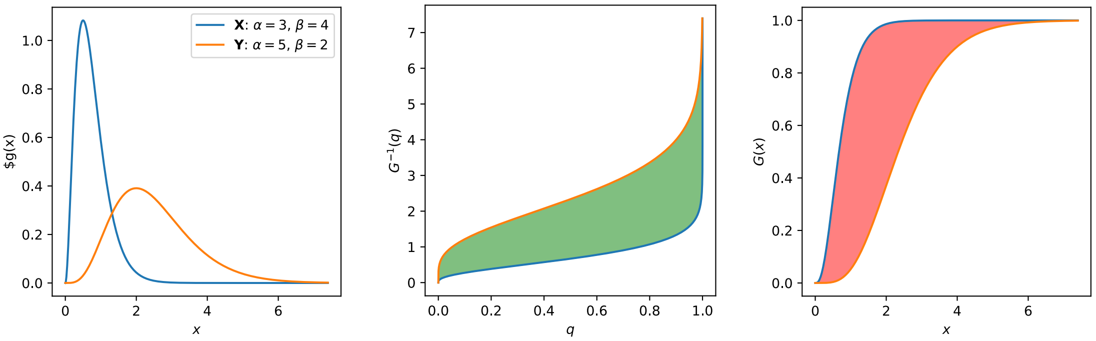
$$\mathbb{W}_1 = \int_{\mathbb{R}} |\mathbb{P}(\mathbf{X} < k) - \mathbb{P}(\mathbf{Y} < k)| \, \mathrm{d}k$$

In the case of two discrete probability distributions, this can be written as:

$$\mathbb{W}_1 = \sum_{k=-\infty}^{+\infty} |\mathbb{P}(\mathbf{X} \leq k) - \mathbb{P}(\mathbf{Y} \leq k)|$$

The closer $\mathbf{X}$ resembles $\mathbf{Y}$, the smaller this metric is. If the two distributions are identical, then the metric will be at a minimum of 0. We chose to calculate the Wasserstein-1 metric, rather than any other Wasserstein-$p$ metric, for mathematical simplicity. We want to see how well our estimated distribution for $E_0$ matches values for $E_0$ occurring through simulation.

We simulated scenarios with groups of size $n$ from 10 to 50 individuals. For each value of $n$ we generate 100,000 random values of $P$, each member of the group's probability of infection, by sampling from a Beta distribution with parameters $(\frac{1}{2}, \frac{1}{2})$. For each value of $P$ we randomly select a value for $E_0$, the total number of infected individuals, by sampling from the binomial distribution $Binomial(n, P)$ For each infected individual we assign a normalised incubation period by sampling a Uniform



(a) (b) (c)

Figure 2.5. Calculating the Wasserstein-1 metric for two Gamma distributions. The left hand plot shows the PDFs of two Gamma distributions, $Gamma(\alpha = 3, \beta = 4)$ and $Gamma(\alpha = 5, \beta = 2)$, in blue and orange respectively. The middle plot shows the quantile functions of these distributions. The Wasserstein-1 metric of these two distributions is equal to the absolute area between these two functions, shown in green. The right hand plot shows the cumulative density functions of these distributions, which is the inverse of the quantile functions. The absolute area between these curves, in red, is equal to the green area in the middle plot and therefore also equal to the Wasserstein-1 metric.

distribution between 0 and 1. With an incubation period distribution of $U(0, 1)$, $F(\tau) = \tau$. This means we can present the passage of time in our results in terms of $F(t)$, which would therefore be agnostic to our distribution choice.

At any time during any simulation, we can observe $I_t$, the number of individuals who are symptomatic by time $t$, $n$, the total number of individuals in the simulation, and $F(t)$, the probability that an infected individual would be infected prior to this time. From this we can generate a probability mass function for the true value of $E_0$, the total number of infected individuals using Equation 2.2. From our simulated results we can also calculate the “true” distribution for $E_0$ given any observed scenario. By finding the difference of the cumulative mass functions of both distributions, we estimate a value for $\mathbb{W}_1$ given $n, I_t, F(t)$.

$$\mathbb{W}_1(n, I_t, F(t)) = \sum_{e_0=I_t}^{n} |\mathbb{P}(E_{calc} \leq e_0) - \mathbb{P}(E_{sim} \leq e_0)|$$

For any given value of $n$, $E_0$ and $I_t$, the maximum potential difference between our calculated, estimated distribution and the simulated, observed distribution would be if one distribution predicted $E_0$ definitely equalled $I_t$ (i.e. all infected individuals have already developed symptoms) and if the other predicted $E_0$ definitely equalled $n$ (i.e. all exposed individuals were infected). In this case, the Wasserstein-1 metric would be:

$$Max(\mathbb{W}_1) = \sum_{e=I_t}^{n-1} 1 = n - I_t$$

The larger the difference between $n$ and $I_t$, the larger the Wasserstein-1 distance can become. To compare meaningfully between two scenarios where the difference between $n$ and $I_t$ is different, we calculate $\mathbb{W}_{prop}$, a normalised version of the Wasserstein-1 metric:

$$\mathbb{W}_{prop}(n, I_t, F(t)) = \begin{cases} \frac{\mathbb{W}_1(n, I_t, F(t)}{n - I_t} & I_t < n \\ 0 & I_t = n \end{cases}$$

In the case where $I_t = n$, we say the proportional Wasserstein-1 distance is equal to 0. The Wasserstein-1 metric will show us the total distance between our estimated distribution and the simulated value. The proportional Wasserstein-1 metric will show that metric proportional to the total possible error.

We find the overall distribution for these values for fixed values of $F(t)$, and then examine the mean values across simulations at fixed time intervals and fixed:

1. Group sizes $(n)$
2. Proportions of individuals who are symptomatic $(\frac{I_t}{n})$
3. Proportions of individuals who are infected $(\frac{E_0}{n})$

As the randomised model is based on the same assumptions as our calculated distributions, we expect the simulated results and the calculated distribution to be similar. As a result, we should see a fairly low Wasserstein-1 and proportional Wasserstein metric. They should only differ due to the random outcomes of our simulation and so the areas where the difference (and the Wasserstein metrics) is higher may indicate areas of elevated variation in this model.

Shannon Entropy

The Wasserstein-1 metric shows how well our calculated distribution matches the observed distributions from our simulations. It does not tell us how informative our calculated distribution is. We use the Shannon entropy $\mathbb{H}(n, I_t, F(t))$ to calculate the information provided by our calculated distribution:

$$\mathbb{H}(n, I_t, t) = \sum_{e=I_t}^{n} -\mathbb{P}(E_0 = e|I_t, t) \times \log_2 \left[\mathbb{P}(E_0 = e|I_t, t)\right]$$

In this case, a fully informative distribution (one where, for one value of $e$, $\mathbb{P}(E_0 = e) = 1$) will have a Shannon metric of 0. Conversely, we can show that a minimally informative distribution, where there is a uniform probability of any possible value of $E_0$, will have a Shannon metric of $\log_2 [n - I_t + 1]$.

We state that the maximum Shannon entropy of a discrete probability distribution is $-\log_2[n]$ where $n$ is the total number of possible values if the distribution. This can be shown by looking at the local information provided by two probabilities in a discrete distribution.

$\mathbf{x}$ is a range of possible discrete values for which the probability of selecting each value is fixed. We focus on two values in this possible set, $a$ and $v$. The probability of selecting $u$ or $v$ is $w$, a fixed value between 0 and 1. $\mathbb{H}_{u,v}$ is the local entropy of just the probabilities of $u$ and $v$. $U$ is the unknown probability of $u$, which falls between 0 and $W$. We can therefore calculate $\mathbb{H}_{u,v}$:

$$\begin{aligned}\mathbb{H}_{u,v} &= -\mathbb{P}(\mathbf{x}=u)\log_2[\mathbb{P}(\mathbf{x}=u)] - \mathbb{P}(\mathbf{x}=v)\log_2[\mathbb{P}(\mathbf{x}=v)] \\ &= -U\log_2[U] - (w-U)\log_2[w-U]\end{aligned}$$

With a Shannon entropy, we are specifically looking at $\log_2$ terms, but as converting between different logarithms involves dividing by a fixed value, we are going to simply write $\log$ as a shorthand from now on. We solve the derivative of this equation in terms of $U$ to find a value of $U$ that gives a maximum entropy.

$$\begin{aligned}\frac{\mathrm{d}\mathbb{H}_{u,v}}{\mathrm{d}U} = \log[w-U] - \log[U] &= 0 \\ U &= \frac{w}{2}\end{aligned}$$

The second derivative is always negative when $U = \frac{w}{2}$, meaning this a maximum.

We now know value of $U$, the probability of $u$, that maximises the local entropy $\mathbb{H}_{u,v}$ is $\frac{p}{2}$. As the probability of $b$ is $p - Y$, the value of $\mathbb{P}(\mathbf{x} = v)$ at this maximum is also $\frac{w}{2}$. In other words, the localised entropy is at its maximum when the probability of $u$ is equal to the probability of $v$. Any change will result in a decrease in entropy. Taking this observation further we can say that the maximum total entropy for a discrete distribution would be when the probability of each individual value is identical. For a distribution with $n$ possible values, this would be when the probability of any one value is equal to $\frac{1}{n}$.

This seems in keeping with our understanding of entropy as a measure of information provided. A maximum entropy means the distribution provides the minimum information it can. The case of all discrete probabilities being equal seems to be an extremely uninformative distribution, as aside from declaring all the possible values, it gives no indication which one the resulting value will be.

With a discrete probability distribution of size $n$ we can therefore calculate the maximum possible entropy:

$$\mathbb{H}_{max} = n \times -\frac{1}{n} \log \left[\frac{1}{n}\right] = \log[n]$$

We can state $0 \leq \mathbb{H}(n, I_t, F(t)) \leq \log_2[n - I_t + 1]$, with a lower Shannon metric representing a more informative distribution. Once again, this means the possible range of values for our metric changes dependent on $n$ and $I_t$. We also calculate a normalised Shannon metric $\mathbb{H}_{norm}$ by dividing by this maximum value.

$$\mathbb{H}_{norm}(n, I_t, F(t)) = \sum_{e=I_t}^{n} -\frac{\mathbb{P}(E_0 = e|I_t, t) \times \log_2[\mathbb{P}(E_0 = e|I_t, t)]}{\log_2[n - I_t + 1]}$$

We calculate these Shannon metrics for each occurrence in our simulation and demonstrate the entropy range based on how frequently each observation occurs in our simulation.

All simulations were performed in Python with a NumPy random seed of 1111.

Figure 2.6 demonstrates the results from an example simulation. By and large, as time progresses, the estimate for the value of $E_0$ gets narrower and closer to its true value. However, we see multiple skips coinciding with when new individuals become symptomatic. These jumps do not necessarily make our estimates more accurate. Clearly there is a more complicated relationship at work than simply "more symptomatic individuals result in a more accurate model", which we hope to get a better understanding of by averaging these metrics across multiple iterations.

### 2.3.4 Wasserstein-1 results

We can see from the simulated results in Figure 2.7 that our calculated estimates do not perfectly match the simulated data, although they are closely related, a relationship that improves over time. If our model perfectly predicted the simulated distribution of the total number of individuals that were infected, the Wasserstein-1 distances would always be 0. The higher the Wasserstein-1 distances, the worse our model predicted the simulated results.

The total Wasserstein-1 distance tends to be below 1 ($10^0$ - note the Log scale on the y-axis) by the time of $F(t) = 0.5$ (i.e. the time of the median estimate for the incubation period), as seen in Figure 2.7a. In this time, the median proportional Wasserstein-1 distance is below 0.05, indicating a good fit (Figure 2.7b).

Of the three factors we investigated, group size appears to have the smallest effect when it comes to both Wasserstein-1 and proportional Wasserstein-1 metrics in terms of orders of magnitude. As predicted, larger groups will result in a larger error as there is a wider range of possible mistakes to make. A larger group will, up until

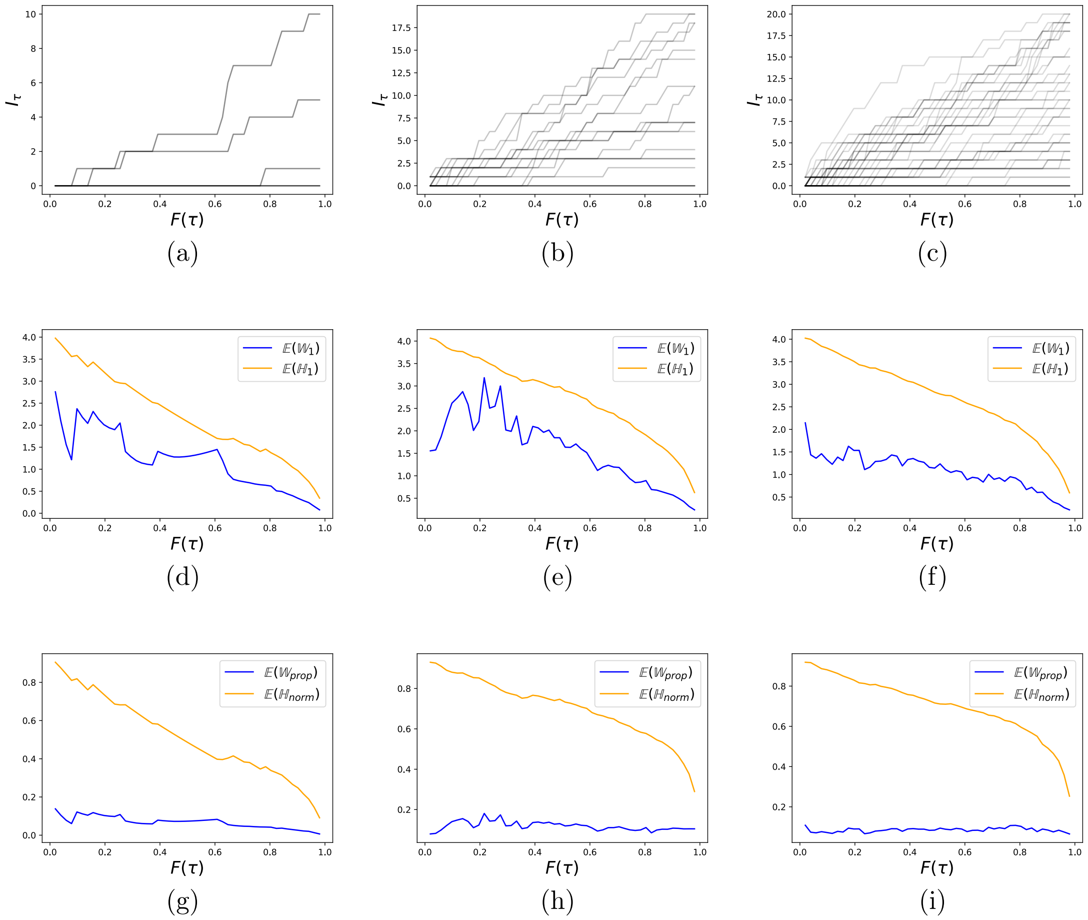
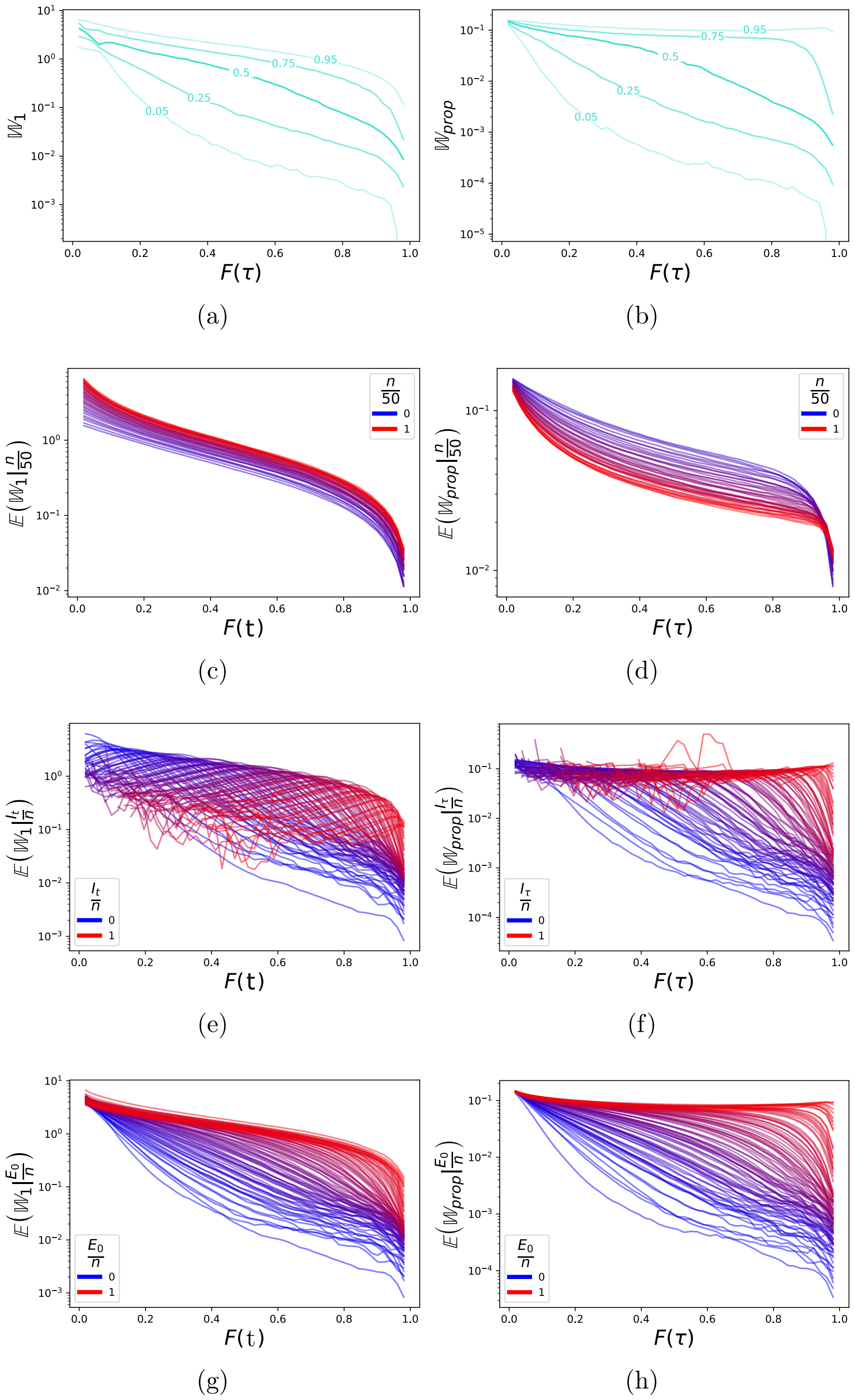


Figure 2.6. Example simulation and metric results of a simple exposure event. At time $t = 0$, a group of individuals of size $n = 20$ are all exposed to a disease with a $P$ chance of being infected, where $P$ is drawn from a Beta distribution with input values of $(\frac{1}{2}, \frac{1}{2})$. From left to right, the results of 5, 20 and 50 simulations are shown. The top graphs are spaghetti plots of the total number of symptomatic individuals over time, the middle graphs estimate the mean Wasserstein-1 metrics and Shannon entropies from these simulations and the bottom graphs show their proportional equivalents. In this case, $n = 20$, but our true simulations cover values of $n$ between 10 and 50 and perform 100,000 simulations per value of $n$.

late in the incubation period distribution, result in a lower proportional Wasserstein-1 metric, suggesting that more individuals results in more information and therefore a proportionally more accurate estimation. However, the range of central estimates in both the proportional and real Wasserstein-1 metric barely spread across one order of magnitude and the effect of time is far more prominent.

Both proportion of the group that are symptomatic and proportion of the group that are infected have a greater effect on the accuracy of our system. They are, of course, inherently linked, as you cannot have symptomatic individuals without having infected individuals, and this link will strengthen over time as the proportion of individuals that are symptomatic tends towards the proportion of people who are infected.

In real terms, for any proportion of infected individuals, the expected Wasserstein-1 metric of our model compared to the simulated data decreases over time. A higher proportion of infected individuals results in a higher level of error, the disparity of



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 2.7. Plot of the changing Wasserstein distances when comparing our calculated distributions to simulated results. Plots 2.7a and 2.7b show the 5th, 25th, 50th, 75th and 95th centiles of errors across the simulations. Each strand on the spaghetti plots Figures 2.7c - 2.7h show the changing average Wasserstein-1 distances (left) and proportional Wasserstein-1 distances (right) given one fixed parameter, coloured according to said parameter. Figures 2.7c and 2.7d show the effect of initial population size on error during the same simulations, with red strands showing large population sizes. Figures 2.7e and 2.7f show the effect of the proportion of individuals who are symptomatic by time on accuracy, with red strands representing a more symptomatic population. Finally, Figures 2.7g and 2.7h show the effect total outbreak size has on the accuracy of our predictive model, with red strands showing a larger outbreak. As the analysis is agnostic to the type of distribution used to represent the incubation period, the passage of time is represented in terms of $F(t)$.

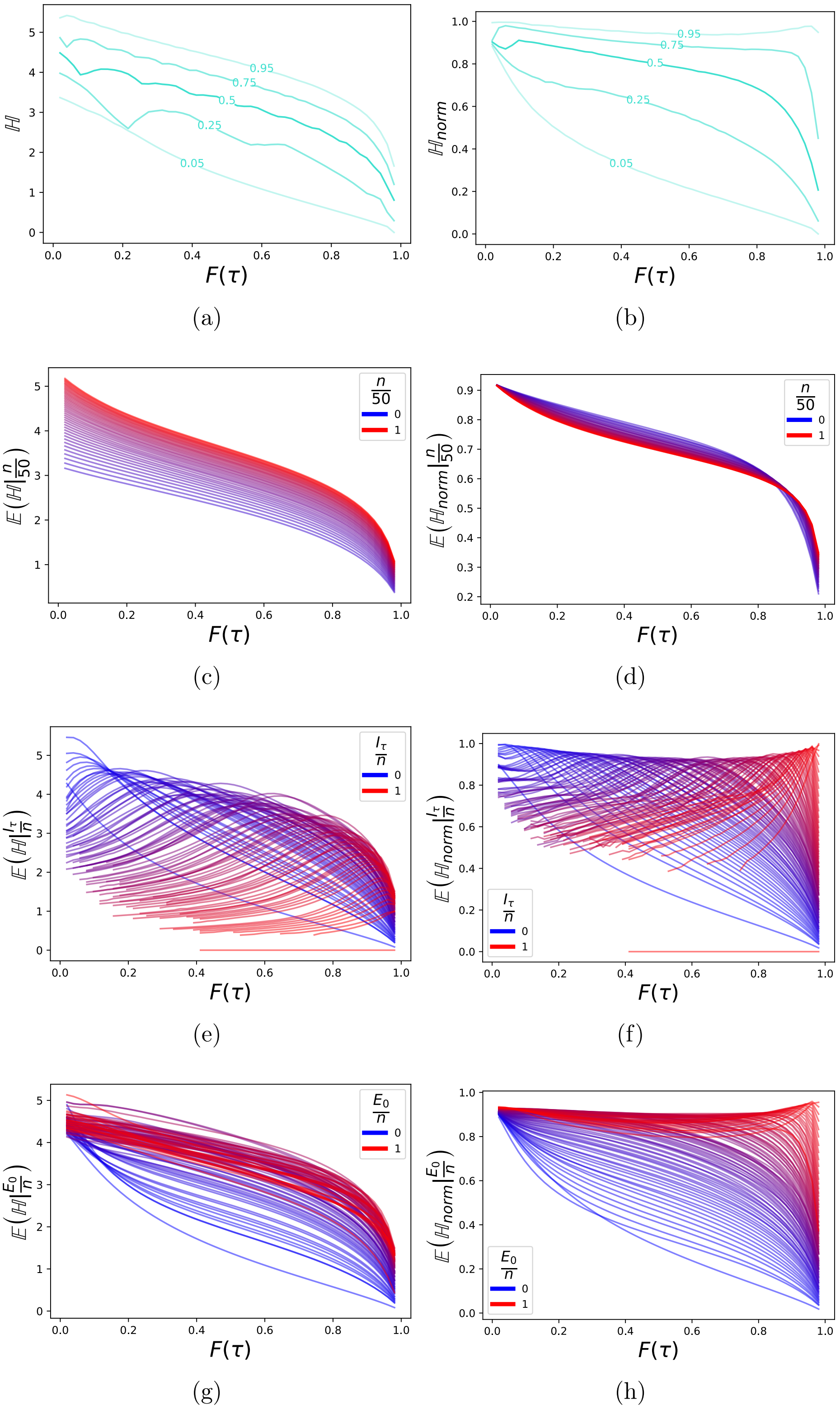
which increases over time. For high enough values of $\frac{E_0}{n}$ this can result in some increase in the porportional Wasserstein-1 metric.

The proportion of individuals that are symptomatic has a far more dramatic effect on the estimated Wasserstein-1 metric. Firstly, it is worth noting the erratic estimates for high proportions of symptomatic individuals early in our model. This is because of a low count in the simulations, as it we are unlikely to see high proportions of symptomatic individuals early in the model. The progression of the estimated Wasserstein-1 metric is not monotonic, unlike when looking at other fixed parameters. In real terms, we can clearly see an increase in the Wasserstein-1 metric over time which hits an apex and then decreases. This apex occurs later for higher values of $\frac{I_t}{n}$. By and large the increase results in a constant proportional error that drops off when the apex is reached. As a result of this apex, initially lower proportions of symptomatic individuals results in higher total error, but over time the higher the proportion of individuals that are symptomatic, the greater the error. This shift is in part explained when we look at the Shannon entropy.

### 2.3.5 Shannon results

Figure 2.8 shows the changing Shannon metric of our distribution over multiple different scenarios. In effect, it shows the amount of information our calculation can be expected to provide at any one time. The higher the line, the less information the distribution provides. In the case of the normalised Shannon metrics (right-hand side figures), the closer the y-value is to 1, the closer our calculated distribution is to an uninformative Uniform distribution. That being the case, how is it that as time increases from $F(t) = 0$, our distribution becomes less informative? How can observing for longer result in more uncertainty?

The key to answering this question lies in the three scenarios presented at the start of this chapter. In the first, one person was observed for 100 days and we were certain they were not infected. In the second 99 out of 100 people were symptomatic within 5 minutes, so we were certain that the 100th person would become symptomatic eventually. In the third scenario, that 100th person was observed for 100 days and at some point during that observation period we went from being certain they were infected to being certain they were not. At some point between these two times, we hit a peak of maximum uncertainty where, based on how much time has passed, there is a 50% chance the last remaining individual was infected and a 50% chance that they were not. Before that point, certainty that the individual was infected was steadily decreasing as each day our prediction that they would develop symptoms did not come true. After that point, our uncertainty that they were infected became out-weighed by our increasing certainty that the individual had not been infected, up until time $F(t) = 1$, at which point we could be certain that an asymptomatic individual was not infected as it would be impossible for them to have



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 2.8. Plot of the expected changes in Shannon entropy of our calculated distribution given our simulated observations. Left-hand plots reveal the true Shannon metric, while right-hand plots show a normalised Shannon metric by dividing by the maximum possible value in each scenario. The top graphs show the spread of Shannon entropy with the 5th, 25th, 50th, 75th and 95th centiles, and the graphs below show the expected Shannon entropy for fixed values of group size, proportion symptomatic and proportion infected in our simulations.

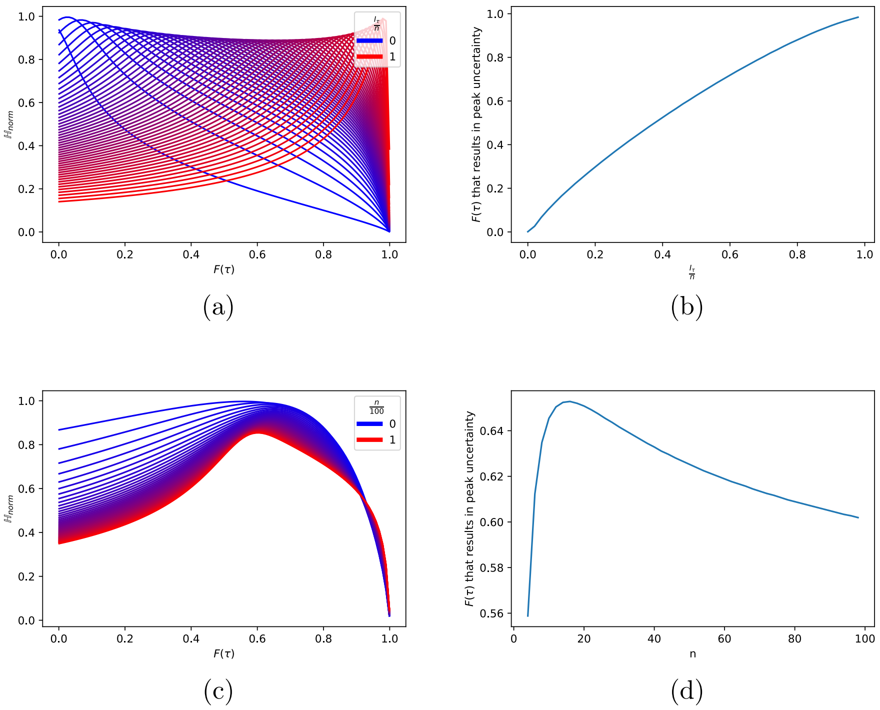
an incubation period so long. At this point, the Shannon metric equals 0. This initial certainty, followed by an increasing uncertainty to a peak and then an increasing certainty is worth keeping in mind when consider the shape of the Shannon curves we see in Figure 2.8.

First, at the start of our model we can see that the proportional Shannon entropy tends to be close to 1, indicating a distribution close to uniform. At this time we have little information regarding the outbreak so we would expect this value to be close to 1. In general, we can see that as time increases, the range of possible values for the Shannon entropy decreases. However, a certain proportion still remain close to a normalised Shannon entropy of 1. For the correct parameter set is it possible to estimate a near-uniform distribution for any time (except a time at which $F(t) = 1$).

Figures 2.8c and 2.8d show us that, as with the Wasserstein-1 metric, group size has little effect on the expected Shannon entropy, and even less effect on the normalised Shannon entropy. In general, a larger group will result in more uncertainty, but time has a far bigger effect.

The proportion of individuals that are symptomatic has a far more dramatic effect on the timing of the peak of uncertainty (Figures 2.8e and 2.8f). The more of the group that are symptomatic, the later the peak, with the timing of these peaks ranging from $F(t) \approx 0$ to $F(t) \approx 1$. In the proportional Shannon entropy, we can see that the expected values at the peak closely approach 1, indicating a near uniform-distribution. We believe this peak represents the time at which, for a set value of $\frac{I_t}{n}$, our model starts to shift from being certain that there are more infected individuals to being uncertain how many more individuals are infected, to being certain that there are no more infected individuals left. We believe that it is this peak in uncertainty that least matches our simulated data, resulting in a peak in proportional Wasserstein-1 metric as well. The timing of this peak is highly dependent on the proportion of individuals that are symptomatic, with a higher proportion of symptomatic individuals resulting in a later shift in uncertainty. This is in keeping with our intuition: the more symptomatic individuals there are, the later we will be certain that we have seen all the infected individuals.

This possibility is perhaps better demonstrated in Figure 2.9, which investigates the changing timing of peak uncertainty for a fixed value of $n = 50$ with a varying proportion of symptomatic individuals (Figures 2.9a and 2.9b) and a fixed proportion of symptomatic individuals $\frac{I_t}{n} = 0.5$ with a varying group size (Figures 2.9c and 2.9d). In these Figures we can see that the proportion of symptomatic individuals is far more important than the size of the exposed group when it comes to the timing of peak uncertainty. In terms of $F(t)$, this timing in approaching 1:1 with the proportion in the group that are symptomatic. Its relationship with group size is not monotonic, initially increasing and then slowly decreasing. We allowed $n$ to vary between 2 and 100 and the timing of peak uncertainty only ranged between



(a)

(b)

(c)

(d)

Figure 2.9. Plot demonstrating timing of peak uncertainty with either constant group size or constant proportion to symptomatic individuals. Figures 2.9a and 2.9b demonstrate the changing timing of peak uncertainty when group size is constant at $n = 50$. Figures 2.9c and 2.9d demonstrate the comparatively constant timing of peak uncertainty when group size is allowed to vary, but the proportion of individuals who are symptomatic, $\frac{I_t}{n}$ is constant at 0.5.

$0.55 < F(t) < 0.66$.

The effect the proportion of infected individuals has on the Shannon entropy needs to be thought about carefully. Unlike the group size and the proportion of individuals that are symptomatic, the proportion that are infected does not directly affect the calculations of the Shannon entropy. Instead it dictates the environment which may result in a different proportion of symptomatic individuals. In general we see that for higher values of $\frac{E_0}{n}$, our expected proportional Shannon entropy remains closer to 1 for longer. The larger the outbreak, the longer we are likely to be uncertain about the exact size of the outbreak. With each new individual becoming symptomatic, the apex of uncertainty based on the proportion of symptomatic individuals becomes later and later in our model. With a larger outbreak, a higher proportion of symptomatic individuals will be achieved, resulting in a later apex and therefore a longer period of time where the normalised Shannon entropy is closer to 1.

### 2.3.6 Discussion

In general, we can see that there is a good fit between the simulated data and our calculated distribution. This only improves with time as we observe the exposed group for longer. By the time $F(t) = 1$, both metrics we have observed equal 0. We would expect this to be true. At this time, all infected individuals should be symptomatic and therefore we should know the exact number of infected individuals. This logic will change if it is possible for infected individuals to be permanently asymptomatic, or if the individuals continue to be exposed after the start of the observation period. Both scenarios will be investigated later.

Not only is the fit fairly in keeping with the simulated data, dependent on various parameters, but also our calculations can be fairly informative. In nearly all scenarios (no observed symptomatic individuals is the exception), the uncertainty of our model will increase as time continues, until it hits a peak largely dependent on the proportion of individuals that are symptomatic. It would be tempting to hope that we can avoid passing through this peak uncertainty by chance as new symptomatic individuals will change the proportion of symptomatic individuals and therefore also the timing of the peak. However, there can only ever be an increase in symptomatic individuals, never a decrease, which in turn delays the timing of the peak, meaning it would still be in our future. In fact, in a very unfortunate scenario, it would not be impossible for individuals to become symptomatic in such timing that the model would pass through multiple peaks of uncertainty.

In general, a higher proportion of symptomatic individuals results in a more informative model early on, as symptomatic individuals, who are therefore definitely infected, tell us more about how many individuals are infected than asymptomatic individuals do, who may not be infected or may be infected but are yet to develop symptoms. As time increases and the probability that an asymptomatic individual is in fact infected decreases, this relationship switches and the lower the proportion that are symptomatic, the more informative our model will be.

Comparing the Wasserstein-1 and Shannon entropy results, there appears to be a clear link between these two outcome measures. Indeed, the same set of parameters that will result in a less informative distribution also seem to result in a greater Wasserstein-1 distance. If we consider what it means to be less informative, we can see why this observation would make sense. A less informative distribution allows for greater range of options when selecting from it in simulation. If our analytical model perfectly matched the system we simulated, we would still expect some difference between the observed, simulated data and our model distribution owing to the stochasticity in the simulated data. In turn, we would be more likely to see this difference in low-information parameter sets, as there would be a great possibility of random selection less in keeping with the calculated distribution.

Figure 2.10 demonstrates this changing relationship for a fixed group size of $n = 10$. In this visualisation we can see that both $\mathbb{W}_{prop}$ and $\mathbb{H}_{norm}$ initially increase to an apex and then decrease together as $F(t)$ increases. There is some variation to this pattern at the very beginning and end of the model but this is likely due to a decreased number of observations and rounding errors respectively. In general we can conclude that a considerable proportion of the decrease in perceived accuracy in our model can be explained by a decrease in information the model provides.

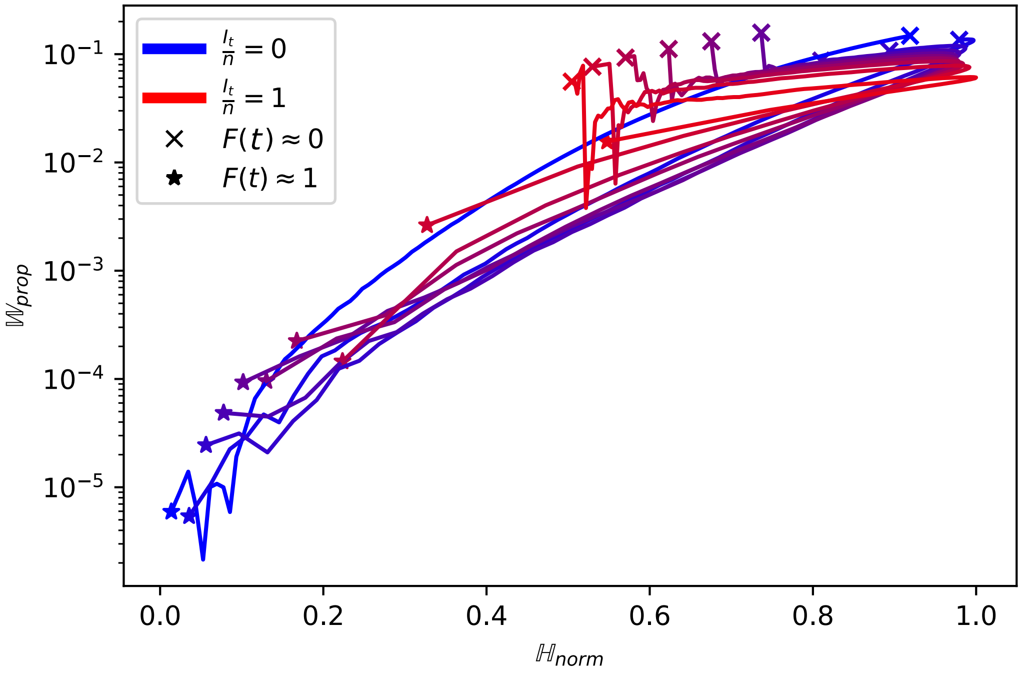


Figure 2.10. Demonstration of the relationship between proportional Wasserstein-1 distance and normalised Shannon entropy for our model when compared to simulated data for different parameter sets. Each line represents a different proportion of symptomatic individuals for a group size of $n = 10$, and tracks how $\mathbb{H}_{norm}$ and $\mathbb{W}_{prop}$ change for a range of values of $F(\tau)$ starting at $F(t) \approx 0$ ($\times$) and going up to $F(t) \approx 1$ ($\star$). These results are based on 10,000,000 simulations, an impractical number of simulations for the larger investigation.

It is, however, a fairly basic model. It assumes that all infected individuals will become symptomatic, that we know the total number of individuals that were exposed and that all individuals were exposed at the same time. We explore these assumptions in Sections 2.4, 2.5 and 2.6. Additionally we assume that every exposed individual has the same probability of infection. Fortunately, Section 2.3.6 reveals that breaking this assumption does not actually change the accuracy of our model.

The censoring problem we observe in our model is not a problem unique to epidemiology. One important example is the count of species in ecology. When attempting to parameterise how many of a particular species may be found in a particular region, it is not unusual to find a high occurrence of zero-counts. This happens for multiple reasons, but can result in underestimating or overestimating the total distribution of counts if not handled carefully.

One well recognised approach is a so called zero-inflation model. In this circumstance, the modeller recognises that there are effectively two separate and often independent decisions occurring when the species count is decided:

1. Is the species count greater than zero?
2. Given that the species count is greater than 0, what is the species count?

The first decision may be effected by factors that are completely separate to the second decision. To take an extreme example to demonstrate my point, I may want to sample average crocodile congregation size. Over 50 days I take two samples, one from two different sites, counting the number of crocodiles I see in each site. I then perform a Poisson regression on these counts to calculate a model for the total congregation size across my sites. Except, what I have not said is that one of these sites is in the swamp-lands of Florida, and the other is in my office in Manchester. My Manchester-based crocodile hunt is likely to artificially inflate the number of zeros I observe (hopefully). However, even if I did separate out my data by site, I may come to the incorrect conclusion that a crocodile congregation in my office would be small. It may not be. I just have not seen one yet. Obviously, this is a slightly ridiculous example that hopefully serves as a demonstration of how zero-inflation can occur. Martin et al. have written an informative summary of a range of approaches that can be taken when handling zero-inflated data. They each come down to treating these two decisions separately (is my datum greater than 0 Vs given that my datum is greater than zero what is it) and performing some form of regression on each of them separately[152].

In a true outbreak model, we should expect to see some element of zero-inflation. Some outbreaks will never grow past a certain point, while others will cross a threshold into exponential growth. This is in part why, whilst on average SARS-CoV-2 outbreaks on cruise ships involved only one person[153], there was a point where transmissions on the Diamond Princess made up over half of all SARS-CoV-2 transmissions outside of China[154]. The difference between our model and a zero-inflation model is that we expect the observed number of infected individuals to be reduced in a uniform manner (by a factor of $F(t)$), rather than reduced to 0, meaning a zero-inflation model would not be appropriate in our case.

Later in this chapter, we will look how changing the method of censoring affects our observations. First though, we will discuss one final question to be answered with this initial scenario: What is the probability that all infected individuals have been observed?

#### Calculating the probability that the number of infected individuals is equal to the number of symptomatic individuals

As a reminder, Equation 2.2 shows us that we can write the probability that $E_0$ takes any value between $I_t$ (the number of symptomatic individuals observed) and $n$ (the total number of exposed individuals):

$$\begin{aligned}\mathbb{P}(E_0 = e|I_t = i_t) = &\frac{B\left(n - e_0 + 1, e_0 + \frac{1}{2}\right) F(t)^{i_\tau + \frac{1}{2}} \left(1 - F(t)\right)^{e_0 - i_t}}{(n - i_t + 1)B\left(n - e_0 + 1, e_0 - i_t + 1\right) B_{F(t)}\left(i_t + \frac{1}{2}, n - i_t + \frac{1}{2}\right)}\\ &\times {}_2\mathcal{F}_1\left(\frac{1}{2}, e_0 + \frac{1}{2}, n + \frac{3}{2}, F(t)\right)\end{aligned}$$

When $E_0 = I_t$, this equation simplifies:

$$\mathbb{P}\left(E_0 = I_t\right) = \frac{B\left(n - I_t + 1, I_t + \frac{1}{2}\right) F(t)^{I_t + \frac{1}{2}} {}_2\mathcal{F}_1\left(\frac{1}{2}, I_t + \frac{1}{2}, n + \frac{3}{2}, F(t)\right)}{B_{F(t)}(I_t + \frac{1}{2}, n - I_t + \frac{1}{2})} \tag{2.3}$$

This is important because it effectively calculates the probability that all the infected individuals have developed symptoms and that there are no more undetected symptomatic individuals.

Figure 2.11 demonstrates how this probability changes with respect to $F(t)$ for a group of size $n = 10$ ($I_t = 10$ was not included in this analysis for obvious reasons). As would be expected, this probability increases as $F(t)$ increases. Aside from $I_t = n$, the first value for $I_t$ where we can be 95% confident that there are no further infected individuals is $I_t = 0$. This stands up to common sense. If you were looking at two groups who had been exposed and asked to guess which one had undiagnosed infected individuals in it, you would go for the one where there is at least some evidence of transmission ($I_t > 0$).

Interestingly, for low enough values of $F(t)$, the probability that there are no more infected individuals is next highest (although still low) for $I_t = n - 1 = 9$. As the number of remaining individuals who could be infected dwindles, the probability that they are also infected decreases, resulting in a higher probability that $E_0 = I_t$.

However, in general, for a 95% confidence that $E_0 = I_t$, the higher the value for $I_t$, the higher the corresponding value of $F(t)$ needs to be. The more symptomatic individuals observed, the longer the wait before we can be confident that there are no remaining undiagnosed infected individuals.

#### Calculating the probability that no one was infected

Given a scenario where no symptomatic individuals have been observed, we can simplify our calculations even further to answer the question "What's the probability that no individuals were infected?" or perhaps more importantly "How long do we have to wait until we can be confident that no individuals were ever infected from the exposure?".

If $I_t = 0$, even our equation for the posterior distribution for $\rho$ is simplified:

$$\mathbb{P}(\rho | I_t = 0) \propto \rho^{-\frac{1}{2}} \left(1 - \rho F\left(t\right)\right)^{n - \frac{1}{2}}$$

By finding a normalising constant, we can write the posterior distribution:

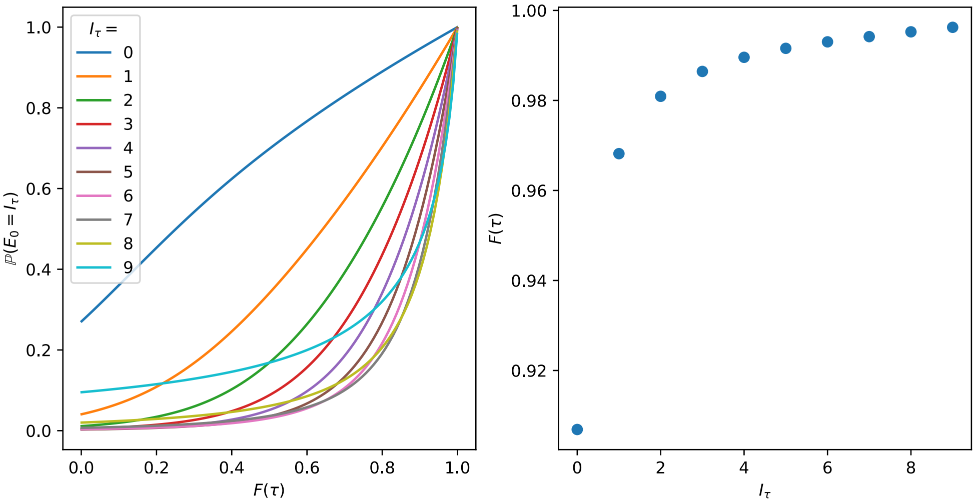


Figure 2.11. Plots of the probability that all infected individuals have been observed in a simplistic model of an exposure event given that $I_t$ out of 10 individuals have developed symptoms. The left-hand graph demonstrates how the probability changes as $F(t)$ increases. The right-hand graph shows the value of $F(t)$ such that we can say with a 95% confidence that no further individuals were infected.

$$\mathbb{P}(\rho|I_t = 0) = \frac{\rho^{-\frac{1}{2}} F(t)^{\frac{1}{2}} \left(1 - \rho F\left(t\right)\right)^{n-\frac{1}{2}}}{B_{F(t)}\left(\frac{1}{2}, n + \frac{1}{2}\right)}$$

In turn we can feed this back in to our formula for the probability that $E_0 = e_0$ given $I_t = 0$, remembering that we are only interested in the case where $E_0, I_t = 0$:

$$\begin{aligned}\mathbb{P}(E_0 = 0|I_t = 0) &= \int_0^1 \frac{(1-\rho)^n}{(1-\rho F(t))^n} \times \frac{\rho^{-\frac{1}{2}} F(t)^{\frac{1}{2}} (1-\rho F(t))^{n-\frac{1}{2}}}{B_{F(t)}\left(\frac{1}{2}, n+\frac{1}{2}\right)} \\ &= \frac{B\left(\frac{1}{2}, n+1\right)}{B_{F(t)}\left(\frac{1}{2}, n+\frac{1}{2}\right)} \times {}_2\mathcal{F}_1\left(\frac{1}{2}, \frac{1}{2}, n + \frac{3}{2}, F(t)\right)\end{aligned}$$

Figure 2.12 demonstrates how this probability varies with respect to $n$, the size of the population exposed. Initially, the smaller the population size, the more likely it is that no individuals were infected. However, as time (and $F(t)$) increases, this flips. The larger the exposed population, the sooner we are going to reach a 95% certainty that no one was infected.

### Calculating the effect of a variable probability of infection

Each of our models is primarily tied to approximating the value of $\rho$, a fixed and unknown probability that an exposed individual would be infected (in Section 2.6 the definition of $\rho$ changes slightly depending on the nature of the prolonged exposure, and cast aside entirely in favour of an infector-focussed model in Section 2.5). This

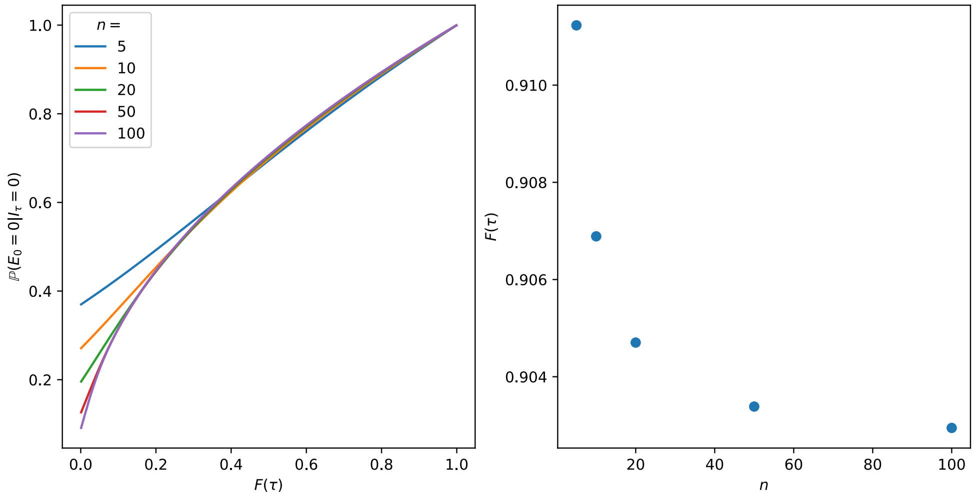


Figure 2.12. Plot of probability of no infected individuals given no observed symptomatic individuals. The left-hand plot demonstrates how this probability changes with respect to $F(t)$. The right-hand plot shows the value of $F(t)$ required to by 95% certain that no individuals were infected during the exposure.

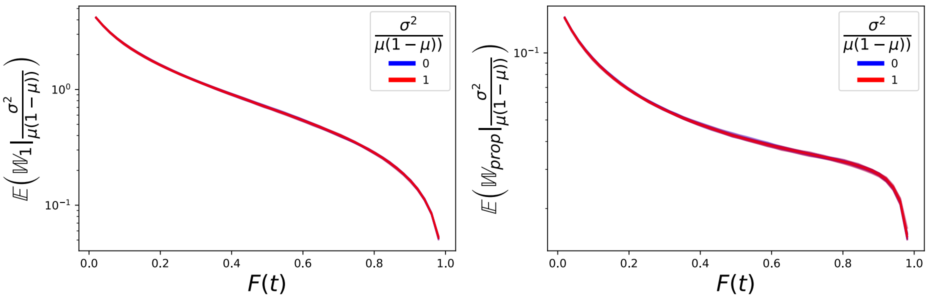
value is assumed to be constant for all exposed individuals and is the basis on which we go on to estimate the total number of infected individuals. But what if it was not constant?

For multiple reasons, it would be unreasonable to assume that each individual's probability of exposure would be exactly the same. It could be that some exposed individuals are more frail, more vulnerable or had a higher level of interaction with the exposing agent, increasing their probability of infection. Conversely, some individuals may have a level of immunity, either acquired or innate, that may reduce their probability of infection. These both seem like fairly reasonable situations in the real world, but it is unclear what effect they would have on our model.

To investigate this, we repeat the Wasserstein-1 metric investigation of our Point Exposure 100% symptomatic rate model (Section 2.3). With each iteration of the model we assign a random average probability of infection $\mu$. However, each individual's probability of infection is taken from a Beta distribution with a mean of $\mu$ and a variation of $\sigma^2$ such that there is a proportional fixed level of dispersion (defined in this analysis as the variation divided by the mean). We want to see how an increase in variance with respect to the mean probability of infection affects the accuracy in our model with regards to predicting the total number of infected individuals.

With a Beta distribution with input parameters $\alpha$ and $\beta$, the mean and variance are calculated as $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ respectively. With a small amount of rearrangement, we can find the variance in terms of $\alpha$ and the mean $\mu$:

$$\sigma^2 = \frac{\mu^2(1-\mu)}{\alpha+\mu}$$



(a) (b)

Figure 2.13. Mean Wasserstien-1 and proportional Wasserstien-1 metric for fixed levels of overdispersion in the probability of infection.

As $\alpha$ approaches 0, $\sigma^2$ approaches $\mu(1-\mu)$ and as $\alpha$ approaches $+\infty$, $\sigma^2$ approaches 0. This gives us our range for $\sigma^2$ for any value of $\mu$. Additionally, it gives us our range for over-dispersion, $od$:

$$0 < od < 1 - \mu$$

We sample $\mu$ from the beta distribution $B(0.5, 0.5)$ and then calculate the corresponding Beta distribution such that over-dispersion of the resulting distribution is a fixed distance along its possible range (i.e. $\frac{\sigma^2}{\mu(1-\mu)}$ is fixed). It is this Beta distribution that we then use to sample each exposed individual's probability of infection. Experimenting with group sizes between 10 and 50 inclusive, we estimate the mean Wasserstein-1 and proportional Wasserstein-1 metrics for different values of this fixed proportion, to see if increasing over-dispersion decreases our model's accuracy.

Figure 2.13 shows the effect variance in the probability of infection has on the accuracy of our model in terms of expected Wasserstein metrics. Perhaps, more accurately, it shows the lack of effect. It could be expected that varying the over-dispersion in the probability of infection would result in a wider variety of outcomes, thereby decreasing our model's accuracy as it fails to account for this change. However, we can see that this change does little if nothing to alter the expected Wasserstein metrics. So what is happening?

This lack of change revolves around a hidden variable in our model, $\rho_{true}$. Where $\rho$ is the probability that an exposed individual is infected, $\rho_{true}$ is the probability that an individual from the exposed group chosen at random is infected once assignment has already occurred. Both probabilities have the same likelihood functions and could be argued to have the same posterior function based on observation of symptomatic individuals (there is an argument to say that $P_{true}$ has a discrete distribution with probabilities $\frac{\mathbf{j}}{n}$ where $\mathbf{j}$ is the set of all integers between $I_\tau$ and $n$ inclusive). So when we are finding a posterior distribution for $\rho$, we are also finding a posterior distribution for $\rho_{true}$.

When we replace $\rho$ with a Beta distribution, $\rho_{true}$ remains a fixed constant. A set number of individuals are assigned as infected, these dictate the probability of selecting an infected individual from the group and likelihood function for $\rho_{true}$ remains the same. In reality, it is this posterior distribution that we then use to estimate the posterior distribution for the total number of infected individuals. The accuracy of this posterior distribution is completely independent to the variance on the value of $\rho$, so introducing variance in the probability of infection does not affect the accuracy of our model.

## 2.4 Model 2: Introducing asymptomatic infected individuals

Asymptomatic transmission has become an important part of understanding SARS-CoV-2 outbreaks. An early narrative study by Oran et al. collating 16 case sources estimated that 40-45% of all cases will remain asymptomatic[155]. A cohort study in the Republic of Korea found no significant difference in the viral load of symptomatic and asymptomatic infected individual, although the viral load in symptomatic individuals appeared to take longer to go down[156]. This has led some to believe that asymptomatic individuals can be just as infectious as those with symptoms. This theory has been corroborated by Almadhi et al., who, through a retrospective study looking at contact tracing databases in Bahrain, found no significant difference between the infectiousness of symptomatic contacts and asymptomatic contacts[157], and Emery et al., whose models of the SARS-CoV-19 outbreak on the Diamond Princess Cruise-Liner suggests that asymptomatic transmission must have had to play an important role in the spread of the virus[158].

In the early stages of the pandemic, asymptomatic individuals posed a particular problem, as tests were not readily available and were reserved for people with clinical suspicion of SARS-CoV-2. In our initial scenario, asymptomatic individuals would go completely unobserved. We relied on knowing the total number of symptomatic individuals by time $t$ to infer the probability that the remainder of their cohort are or are not infected. The possibility that an infected individual could never develop symptoms could throw a spanner into our model. It could be assumed that, in order to accommodate for this, the model needs rewriting from the very beginning, but actually ends up being a very simple fix.

We aim to adjust our model to analyse scenarios when an infected individual will remain forever asymptomatic with a known probability. We will then investigate how how informative and accurate the resulting model becomes.

### 2.4.1 Calculating a posterior distribution for the total number of infected individuals when the rate of asymptomatic transmission is known

We start by assuming that $a$ is the known probability that an infected individual will eventually become symptomatic. In the case of SARS-CoV-2 this value may not be fixed from person to person and likely depends on demographic features like age[159]–[161]. However, for our model we will assume that it remains fixed for all exposed individuals. For every time we have had to calculate the probability that an individual is both infected and symptomatic, we now need to replace this with being infected, able to become symptomatic and currently symptomatic. Whilst this sounds difficult, it is simply a matter of replacing each instance of $F(t)$ with $aF(t)$ in our formulae. Our prior distribution for $\rho$ becomes:

$$\pi(\rho|F(t), a) = \rho^{-\frac{1}{2}}(aF(t))^{\frac{1}{2}}(1 - \rho aF(t))^{-\frac{1}{2}}$$

In turn, our posterior distribution for $\rho$ given $I_t$ becomes:

$$\mathbb{P}(\rho|I_t - i_t, a, F(t)) = \frac{\rho^{i_t - \frac{1}{2}}(aF(t))^{i_t + \frac{1}{2}}(1 - \rho aF(t))^{n - i_t - \frac{1}{2}}}{B_{aF(t)}\left(i_t + \frac{1}{2}, n - i_t + \frac{1}{2}\right)}$$

Finally, our probability for the value of $E_0$ becomes:

$$\begin{aligned}\mathbb{P}(E_0 = e_0|I_t = i_t, a) = & \frac{B\left(n - e_0 + 1, e_0 + \frac{1}{2}\right)(aF(t))^{i_t + \frac{1}{2}}(1 - aF(t))^{e_0 - i_t}}{(n - i_t + 1)B\left(n - e_0 + 1, e_0 - i_t + 1\right)} \\ & \times \frac{{}_2\mathcal{F}_1\left(\frac{1}{2}, e_0 + \frac{1}{2}, n + \frac{3}{2}, aF(t)\right)}{B_{aF(t)}\left(i_\tau + \frac{1}{2}, n - i_t + \frac{1}{2}\right)}\end{aligned} \tag{2.4}$$

This bears a striking resemblance to Equation 2.2, except now all "$F(t)$" terms have been replaced with "$aF(t)$". Essentially, where in the previous model this term would tend towards 1, it now tends towards $a$. This will have some important effects if we want to be certain that no further individuals will be infected (for low enough values of $a$, we will never reach a 95% certainty that all infected individuals have been observed).

### 2.4.2 Performance analysis

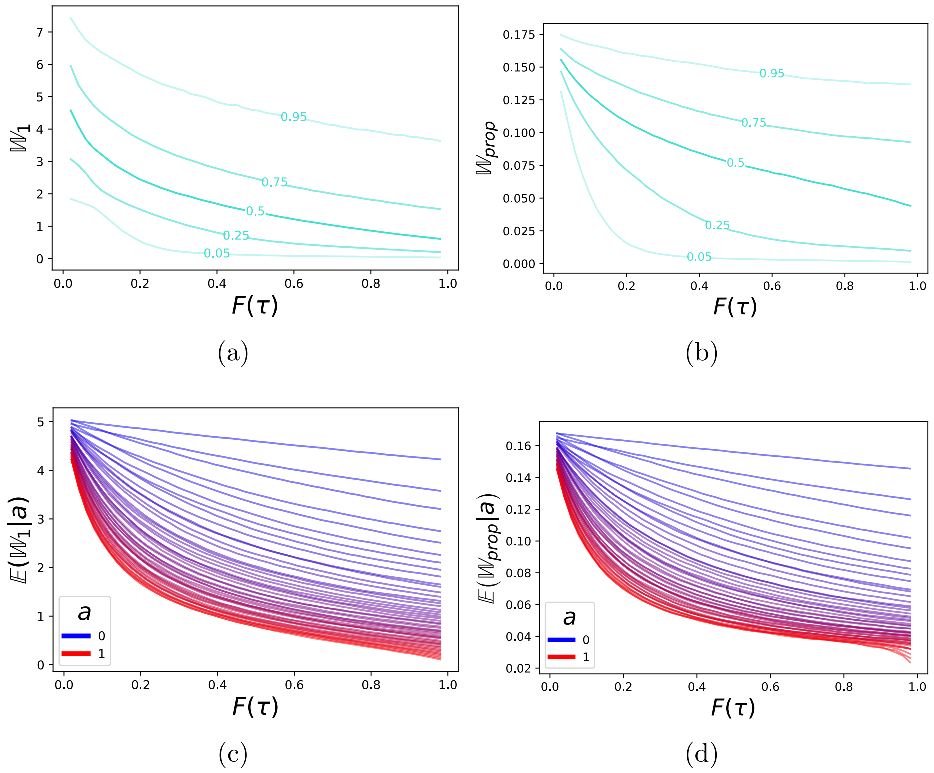
We repeated the simulations in Section 2.3.3. However, we now included a known variable $a$, the probability that an infected individual will ever be symptomatic. This was fixed for each simulation, and was sampled in a range between 0.001 and 0.999. We calculated the Wasserstein-1 and proportional Wasserstein-1 metrics of our expected distributions from our simulated results. In order to see the effect of previously examined parameters (group size, proportion infected, proportion symptomatic), we also repeat the simulations for values $a = 0.1, 0.3, 0.5, 0.7, 0.9$ to show possible outcomes across a range of set values of $a$. We also compare these outcomes to when the value of $a$ is unknown. Additionally, we calculate the the Shannon metrics for each possible scenario.

### 2.4.3 Wasserstein-1 results

The Wasserstein-1 metrics from our simulations can be seen in Figure 2.14. As time passes, our calculated distribution more closely resembles the simulated results. Both the Wasserstien-1 metric and normalised Wasserstein-1 metric decreases (Figures 2.14a and 2.14b). The fit by time $F(t) = 0.5$ is fairly good, with an expected normalised Wasserstein-1 distance under 0.1. However, by introducing uncertainty in the form of possible asymptomatic individuals, we can no longer be certain of the total number of infected individuals at time $F(t) = 1$. This can be seen by the fact that the Wasserstein-1 distance no longer tends towards 0 in our simulations.

The rate of asymptomatic infection greatly influences the accuracy of our model. For a low enough symptomatic rate ($a \approx 0$) there is barely any improvement in the average Wasserstein metric from our model. This is to be expected, seeing as with a low enough probability of symptoms onset, we are unlikely to ever see anyone who is symptomatic, meaning we are unlikely to gain any information as time progresses. If we consider how introducing $a$ affected the equations behind our model, we can effectively say that its accuracy is pushed back in time by a factor of $a$. For any time $t_1$ when $a \neq 1$ our model is equivalent to the same model ($n = n$, $I_t = I_t$) where symptom onset is certain ($a = 1$) at time $t_2$ where $F(t_2) = aF(t_1)$. This is easier to see in our demonstrations where the passage of time is put in terms of $F(t)$. It is perhaps best seen when we compare the effects of group size, proportion symptomatic and proportion infected for fixed values of $a$.

Figures 2.15 and 2.16 demonstrate how the value of $a$ interacts with group size, outbreak size and proportion symptomatic when it comes to the accuracy of our model. These effects remain the same as when it was guaranteed that all infected individuals would eventually become symptomatic. However, if we consider the effect the value of $a$ had on our model, we can realise that our simulated outcomes are elongated by a factor of $a$. Each value of $F(t)$ in the original model was multiplied by $a$, resulting in this elongation. This means that for low values of $a$ we will only see the equivalent to very early on in simulations when all individuals can become symptomatic and we will be barely able to see a difference when the value of $a$ is close to 1. This is not as important for group size and outbreak size where their relationships with the Wasserstein-1 and proportional Wasserstein-1 metrics are largely monotonic (for large enough values of $aF(t)$, the relationship between population size and pro-



(a) (b)

(c) (d)

Figure 2.14. Plot of the changing error in estimation of total outbreak size compared to exposed $F(t)$ if asymptomatic individuals are included in the model. Plots 2.14a and 2.14b show the 5th, 25th, 50th, 75th and 95th centiles of errors across the simulations. Figures 2.14c and 2.14d demonstrate the effect of $a$, the probability that an infected individual will eventually become symptomatic.

portional Wasserstein-1 metric switches from negative to positive). However, with the proportion of symptomatic individuals the change in accuracy over time is no longer homogeneous, starting with a decrease in accuracy to a peak, followed by an increase. The value of $a$ changes when this peak occurs, elongating it to the point where for low enough symptomatic rates an apex may never be reached.

In general, knowing the value of $a$ will make us more likely to make an accurate assessment of the situation. This is increasingly true for larger outbreaks and higher proportions of symptomatic individuals. In a circumstance where underestimating a larger outbreak would be worse than overestimating a smaller outbreak it would always be advisable to find out the rate of symptomatic infections if possible rather than assuming $a = 1$.

### 2.4.4 Shannon results

The effect of the value of $a$ on the Shannon entropy is similar to its effect on the Wasserstein-1 metric. By multiplying all $F(t)$ by $a$, we are elongating the passage of time. Figures 2.18 and 2.19 clearly demonstrate this elongation in action. We can see that, as $a$ increases, it appears as though the graphs squeeze in on the x-axis towards 0, revealing more and more. As $a$ tends towards 1, the amount of information our model is expected to provide at any one time looks more and more like our ini-

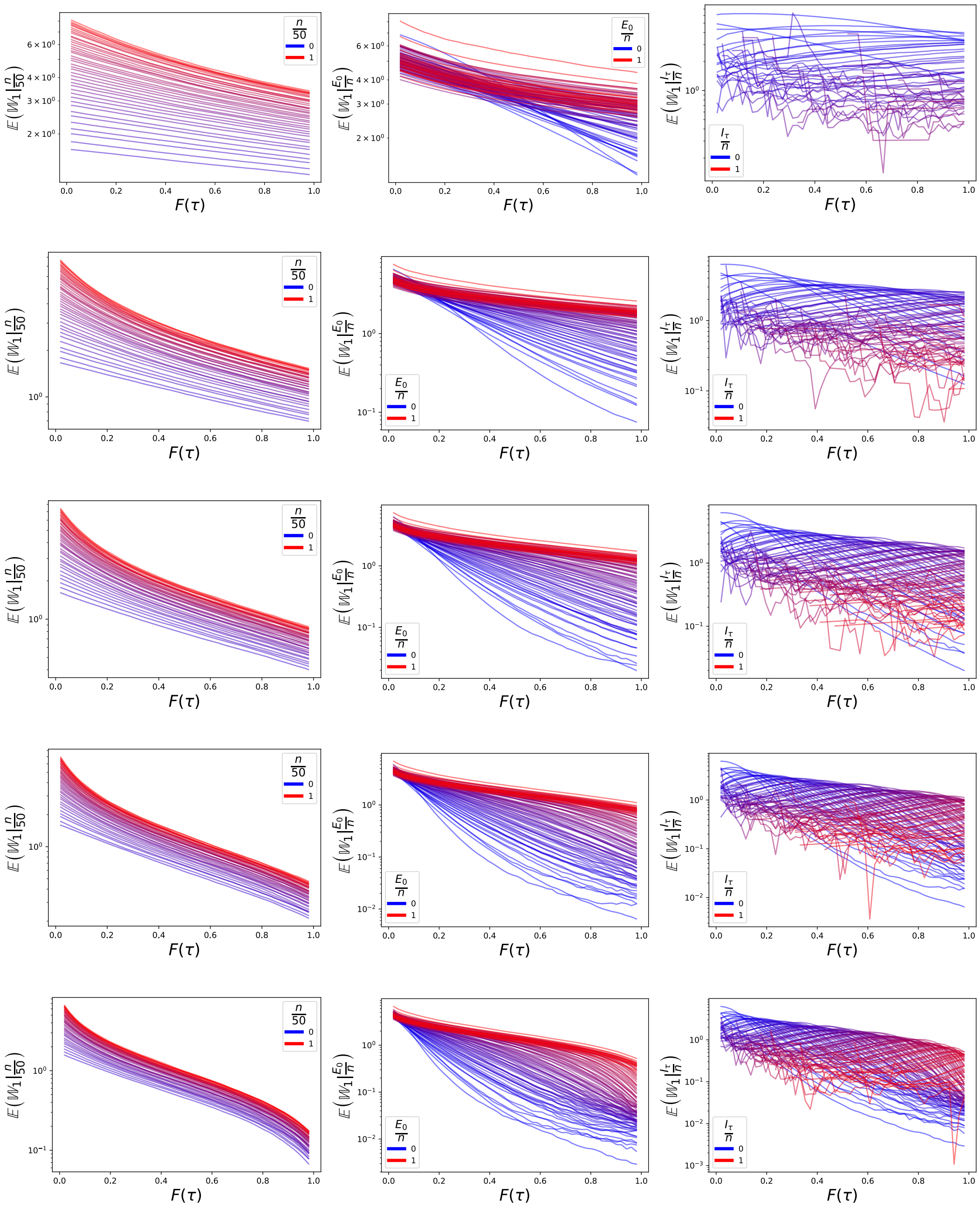


Figure 2.15. Plot demonstrating the effect asymptomatic rate has on the influence other variables have on the accuracy of our model in simulation. Each row has a fixed symptomatic probability of $a = 0.1, 0.3, 0.5, 0.7, 0.9$ from top to bottom. From left to right we see the effect of group size, proportion infected and proportion symptomatic have on the total Wasserstein-1 metric $\mathbb{W}_1$

tial calculations as in Figures 2.7a and 2.7b. This is, of course, to be expected.

What is more stark is the expected Shannon entropy when $a \approx 0$. In Figure 2.19 we can see that for low enough values of $a$, the expected proportional Shannon entropy barely changes over time. If the value of $a$ is low enough, the passage of time is expected to provide us with minimal to no information. As this is an expected value based on our simulations, for a low value of $a$ it is unlikely that any symptomatic individuals appear over time, providing us with minimal information. When $a = 1$, in the early stages of a simulation, the lower the proportion of symptomatic individuals, the higher the Shannon entropy. The same is true for low enough values of $aF(\tau)$ when $a < 1$. The difference is that if $a = 1$, as time increases, the Shannon entropy

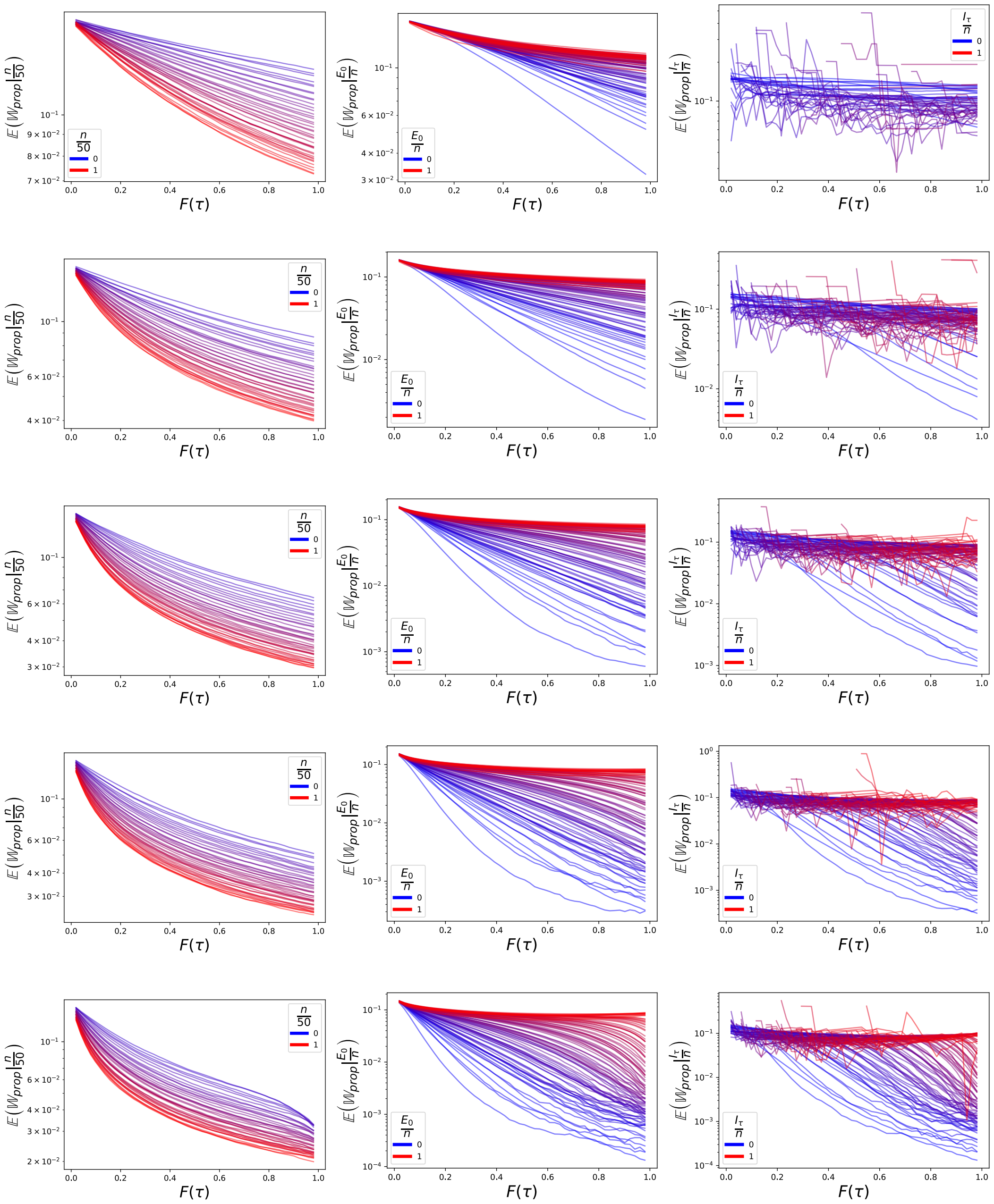
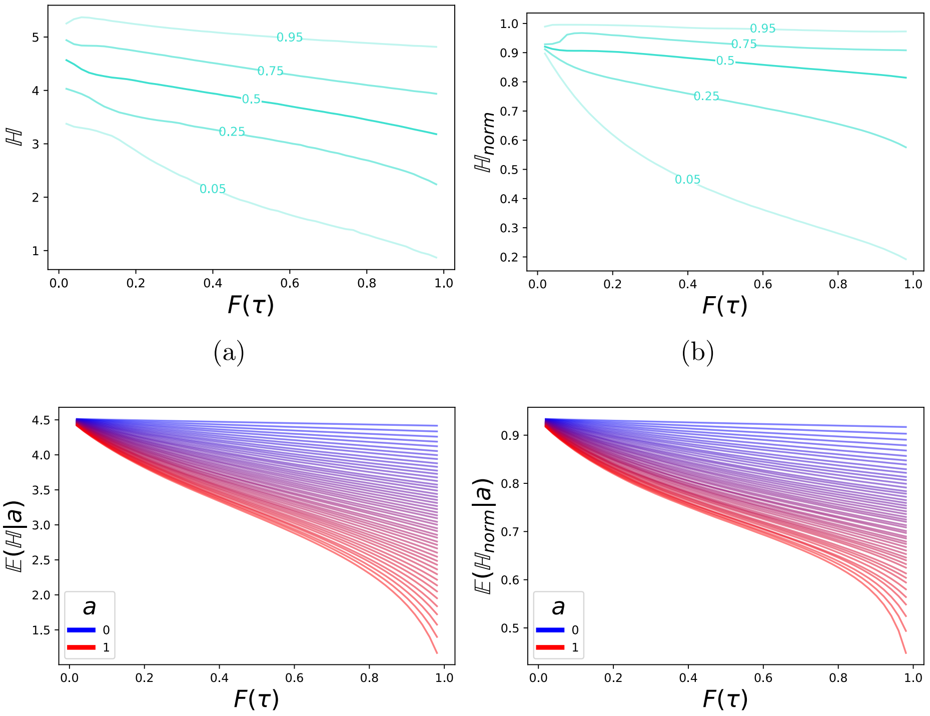


Figure 2.16. Plot demonstrating the effect asymptomatic rate has on the influence other variables have on the proportional accuracy of our model in simulation. The layout is the same as Figure 2.15, but now we demonstrate $\mathbb{W}_{prop}$ rather than $\mathbb{W}_1$.

will eventually decrease, something that will happen sooner rather than later for lower proportions of symptomatic individuals. However, as time is effectively elongated by a factor of $a$, for low enough values of $a$ the same fall in entropy will never occur. If there were more symptomatic individuals, for low values of $a$, the Shannon entropy would be increased. However, this is unlikely to occur in simulation as, with low values of $a$, by very definition infected individuals would be unlikely to ever develop symptoms.

If we compare Figure 2.17 to Figure 2.8 (Page 69), we could initially conclude that varying the probability that an infected individual will eventually develop symptoms, $a$, has little effect on the overall progression of information in our model. There is an initial increase in uncertainty to an apex, followed by a decline. If anything, there



(a)

(b)

(c)

(d)

Figure 2.17. Plot of the changing Shannon metrics of our calculated distribution when it is possible for infected individuals to never develop symptoms. Left-hand plots reveal the true Shannon metric, while right-hand plots show a normalised Shannon metric by dividing by the maximum possible value in each scenario. The top plots reveal the range of Shannon entropies with the 5th, 25th, 50th, 75th and 95th centiles and the bottom plots show the expected entropy during our simulation for fixed values of $a$, the probability an infected indivdual will ever develop symptoms.

may be an improvement: the initial increase in uncertainty is slower and the apeces are lower. Somehow, decreasing the probability that we would ever observe an infected individual has increased the amount of information our model provides at any one time.

The effect of $a$ in a non-simulated scenario is perhaps best demonstrated in Figure 2.20, where we can see the normalised Shannon metric for multiple iterations of the model when $n$ and $I_t$ are constant ($n = 50$, $I_t = 25$). Each iteration is the same curve elongated in the horizontal direction according to the value of $a$. The start points are all the same. The apeces (if they are reached) are all the same height. Importantly, though, if $a \neq 1$, the iteration does not reach 0. In fact, a lot of the iterations in Figure 2.20 do no reach their uncertainty threshold. This point is the time at which we go from being increasingly uncertain that there are undetected infected individuals to being certain that there are no remaining undetected individuals in our group. We will never be certain that we have observed all the infected individuals (unless $E_0 = n$).

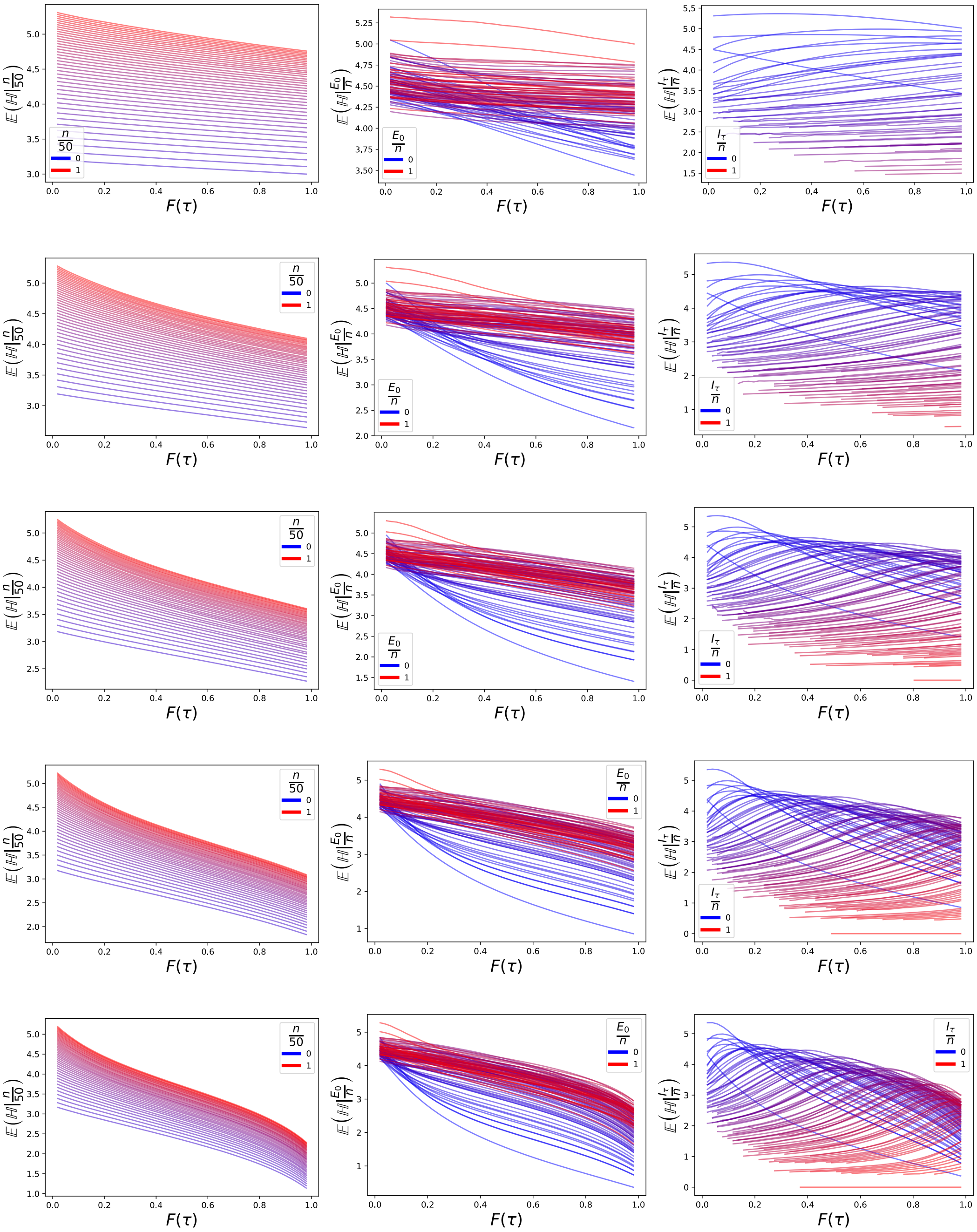


Figure 2.18. Plot demonstrating the effect asymptomatic rate has on the influence other variables have on the information gained from our model in simulation. Each row has a fixed symptomatic probability of $a = 0.1, 0.3, 0.5, 0.7, 0.9$ from top to bottom. From left to right we see the effect of group size, proportion infected and proportion symptomatic have on the expected Shannon Entropy $\mathbb{H}$

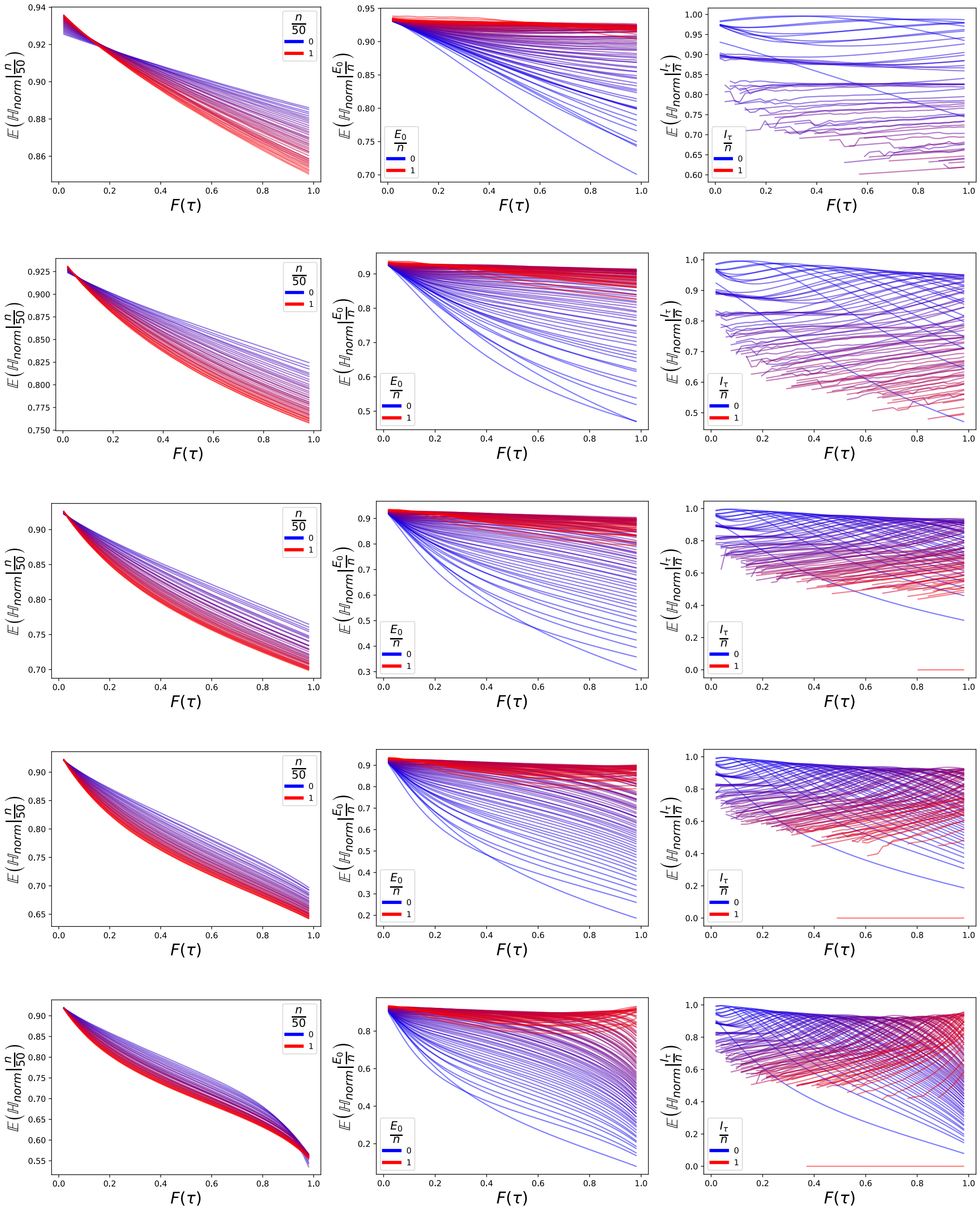


Figure 2.19. Plot demonstrating the effect asymptomatic rate has on the influence other variables have on the proportional expected information our model provides in simulation. The layout is the same as Figure 2.18

### 2.4.5 Discussion

Including the possibility that an infected individual may never develop symptoms can be expected to reduce the accuracy of our model. However, failure to include $a$ in our model can result in far more catastrophic errors in our estimates. Including $a$ is a fairly simple step (assuming symptomatic rates remain constant across the demographic of individuals in our model) that should be taken if the value of $a$ is known. Including $a$ as an unknown variable and integrating through all possible values is a step that could be taken, but would likely result in too much uncertainty in the model.

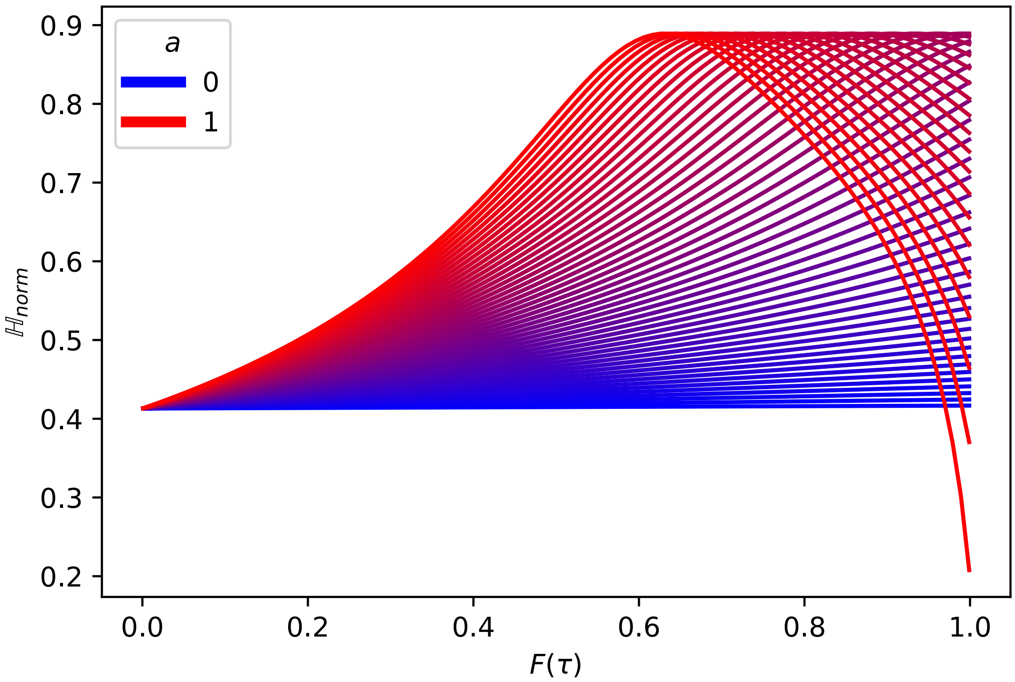


Figure 2.20. Plot of information gained from a model with a consistent group size and proportion of symptomatic individuals, but varying probability of ever developing symptoms. In this case, $n = 50$ and $I_t = 25$. As $a$ decreases from $a = 1$, the same line is stretched in the horizontal direction, time. However, as $F(t) = 1$ represents the end of observable time, for all cases other than $a = 1$, our model never ends on 100% certainty as to how many infected individuals there are, as the curves can never reach $\mathbb{H}_{norm} = 0$.

The question that we really want to answer may not be "How many people have been infected?" but "Have we seen all the infected individuals?". The uncertainty that $a < 1$ introduces means we may never be certain of the answer to this question. The next couple of sections seek to explore under what circumstances we can be certain.

Calculating the probability all infected individuals have been seen

In Equation 2.4, we replaced $F(t)$ with $aF(t)$ from Equation 2.2 to account for the probability that an infected individual was ever going to develop symptoms. We can perform a similar conversion to Equation 2.3 to give us the probability that we have observed all the infectious individuals in the group.

$$\mathbb{P}\left(E_0 = I_t | a\right) = \frac{B\left(n - i_\tau + 1, i_t + \frac{1}{2}\right)\left(aF(t)\right)^{i_\tau + \frac{1}{2}}\,{}_2\mathcal{F}_1\left(\frac{1}{2}, i_t + \frac{1}{2}, n + \frac{3}{2}, aF(t)\right)}{B_{aF(t)}\left(i_t + \frac{1}{2}, n - i_t + \frac{1}{2}\right)} \quad (2.5)$$

As with our model, this calculation elongates the curve of Equation 2.3. As $F(t)$ tends to 1, Equation 2.3 also tends towards 1. The probability that we have observed all the infected individuals in our model when $a < 1$ will tend towards the value of Equation 2.3 when $F(t) = a$.

This means that for certain values of $n$ and $I_t$, we may never be $> 95\%$ certain that we have observed all the infected individuals in our model. This stands up to fairly logical reasoning, as there may well be infected individuals we will never detect because they will never become symptomatic. It also means that, for any value for $n$ and $I_t$, we can calculate a threshold minimum value of $a$. If the symptomatic rate $a$ is below this value we will never be 95% we have observed all the symptomatic individuals.

We know that when $F(t) = 1$, $\mathbb{P}(E_0 = I_t|n, I_t, a = \theta, F(t) = 1)) = \mathbb{P}(E_0 = I_t|n, I_t, a = 1, F(t) = \theta)$. If we find the value of $F(t)$ such that $\mathbb{P}(E_0 = I_t|n, I_t, a = 1, F(t)) = 0.95$, we have in effect found the value for $a$ where the probability that all the infected individuals are symptomatic at time $F(t) = 1$ is equal to 0.95. For any value of $a$ less than this value, a 95% confidence can never be achieved, meaning this becomes a threshold value.

As we can see from Figure 2.21, a high symptomatic rate is required to be 95% certain all infected individuals have been observed. Even in cases where there are no symptomatic individuals, $a$ needs to be at least greater than 0.91. This value decreases for larger group sizes, but this decrease appears to be limited. The threshold increases with $I_t$. This means that with each newly symptomatic individual, the threshold increases. Once $a$ is below the threshold, the model can never change in a way that will bring it back above the threshold - $I_t$ can only increase with time, never decrease. Once this happens, the only way we could be more than 95% certain that all infected individuals have been observed is if all of the group developed symptoms.

Calculating the probability no individuals were infected given asymptomatic infection is possible

Again, this is a fairly trivial conversion, replacing $F(t)$ with $aF(t)$, giving us

$$\mathbb{P}(E_0 = 0|a) = \frac{B\left(\frac{1}{2}, n+1\right)}{B_{aF(t)}\left(\frac{1}{2}, n+1\right)} \times {}_2\mathcal{F}_1\left(\frac{1}{2}, \frac{1}{2}, n+\frac{3}{2}, aF(t)\right)$$

Once again though, as seen in Figure 2.22, any sort of uncertainty in if an individual will go on to develop symptoms undermines the probability that no individual has been infected based on prolonged observation (i.e. you can observe the group for as long as you like, there's still no guarantee none of them have been infected). This demonstrates the importance of being able to identify infected individuals in this scenario.

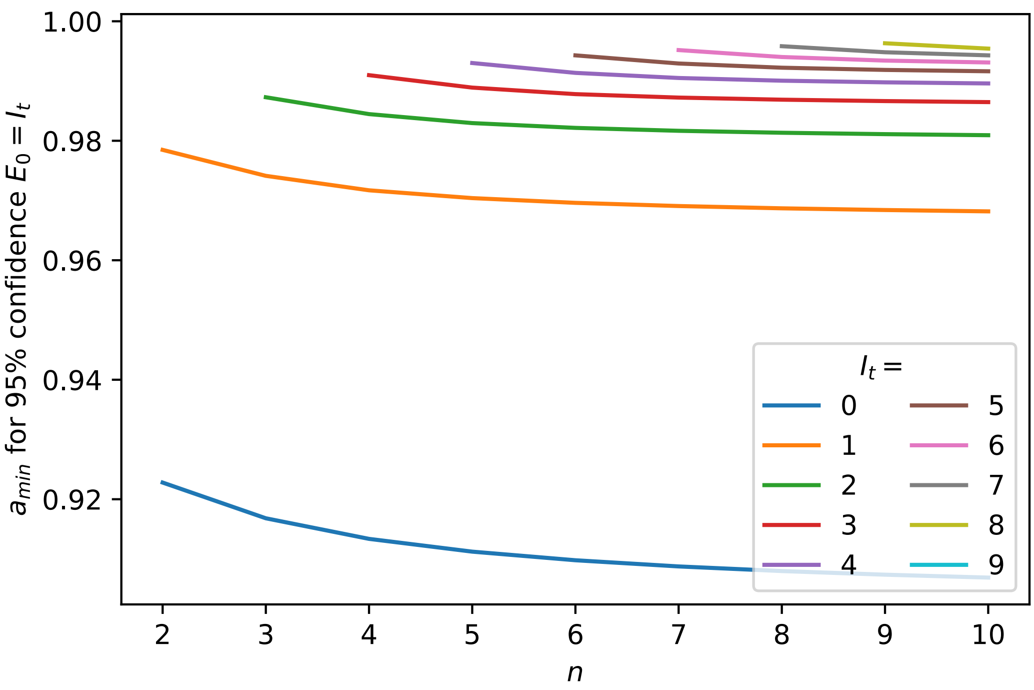
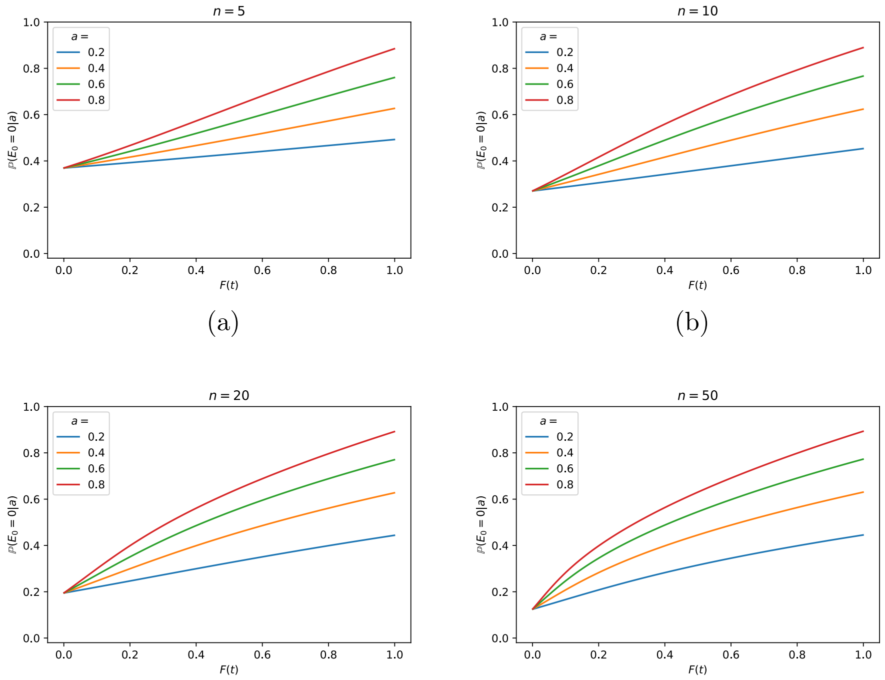


Figure 2.21. Plot of minimum symptomatic rates required to be 95% certain that all infected individuals have been observed. Each line represents a value of $I_t$, the final number of symptomatic individual, the x-axis demonstrates $n$, the group size, and the threshold values of $a$ are shown on the y-axis.



(a)

(b)

(c)

(d)

Figure 2.22. Plot of the probability of no individuals having been infected in a group given a possibility for asymptomatic infections.

## 2.5 Model 3: Estimating outbreak size when the total number of exposed individuals is unknown

In the previous scenarios one element that has been taken for granted is that we know the exact number of individuals that have been exposed, $n$. However, this may not be realistic. If the exposure happened in a public setting, for example, we may not be able to track down all the exposed individuals for monitoring. One way to allow for this is to decide on an upper limit of individuals who possibly could have been exposed and treat this as our new value of $n$. However, this assumption may introduce an error in our model, either by grossly overestimating the number of individuals exposed, stopping monitoring earlier than would be appropriate (larger groups result in sooner conclusions that no further individuals were infected - see Figure 2.12), or by underestimating the total number of exposed individuals and therefore the number of individuals that were infected.

Instead, we propose an alternate model relying on a Poisson distribution. Suppose at the moment of exposure a series of infectious contacts are made. We do not know how many were made, but their number falls on some Poisson distribution. Even if each connection has a fixed probability of success, the number of successes would still fall on a Poisson distribution. The number of successes is equal to the number of infected individuals resulting from the exposure, so $E_0 \sim Poisson(\Lambda)$ where $\Lambda$ is the unknown input parameter for our Poisson distribution. We will find an estimate for $E_0$ by first finding and estimate for $\Lambda$ based on the number of symptomatic individuals by time $t$ and the distribution of incubation periods, much in the same way that when $n$ was known we first found a posterior distribution for $\rho$, the probability of infection given exposure then found an estimate for the true value of $E_0$.

### 2.5.1 Finding a posterior distribution for the Poisson rate of infectious contacts

Although our model is different to our initial model, our approach remains very much the same. For a given value of $\Lambda$, we can write a probability mass function of seeing $x_t$ symptomatic individuals by time $t$:

$$g\left(x_t; \Lambda, F(t)\right) = \frac{(\Lambda F(t))^{x_t} \exp\left[-\Lambda F(t)\right]}{x_t!}$$

This is a Poisson distribution with rate $\Lambda F(t)$. As with before, we can use this PMF to calculate to Jeffreys prior for $\Lambda$:

$$\pi(\Lambda | F(t)) = \sqrt{-\mathbb{E}\left[\frac{\mathrm{d}^2 \ln\left[g(x_t; \Lambda, F(t))\right]}{\mathrm{d}\Lambda^2}\right]}$$

$$= \sqrt{-\sum_{i=0}^{\infty} g(i;\Lambda,F(t))\frac{\mathrm{d}^2 \ln\left[g(i;\Lambda,F(t))\right]}{\mathrm{d}\Lambda^2}}$$
$$= \sqrt{\frac{F(t)}{\Lambda}}$$

Repeating the same steps we took to find a posterior distribution for $\rho$, we now remember that the probability that $\Lambda$ is any value given that $I_t$ individuals are symptomatic by time $t$ is proportional to the probability that $I_t$ individuals would be symptomatic by time $t$ multiplied by our prior distribution for $\Lambda$:

$$\begin{aligned}\mathbb{P}(\Lambda|I_t,F(t)) &\propto g(I_t;\Lambda,F(t)) \times \pi(\Lambda|F(t)) \\ &= \frac{\Lambda^{I_t-\frac{1}{2}}F(t)^{I_t+\frac{1}{2}}\exp\left[-\Lambda F(\tau)\right]}{I_t!}\end{aligned}$$

By finding the normalising constant such that $\int_0^\infty \mathbb{P}(\Lambda|I_t,F(t))\mathrm{d}\Lambda = 1$ shows us that the posterior distribution is a Gamma distribution:

$$\Lambda \sim \mathit{Gamma}\left(I_t + \frac{1}{2}, F(t)\right)$$

### 2.5.2 Calculating a distribution for the total number of infected individuals when an upper limit is unknown

Again, we can repeat our methods for estimating a distribution for the value of $E_0$ when $n$ was known, except now our estimates fall on a Poisson distribution of an unknown rate $\Lambda$:

$$\begin{aligned}\mathbb{P}(E_0 = e_0|I_t = i_t) &= \int_0^\infty \frac{\mathbb{P}(E_0 = e_0 \ \ \& \ \ I_t = i_t|\Lambda,F(t))}{\mathbb{P}(I_t = i_t|\Lambda,F(t))} \times \mathbb{P}(\Lambda|I_t = i_t)\mathrm{d}\Lambda \\ &= \int_0^\infty \frac{\frac{\Lambda^{e_0}\exp[-\Lambda]}{e_0!}\binom{e_0}{i_t}F(t)^{i_t}(1-F(t))^{e_0-i_t}}{\frac{(\Lambda F(t))^{i_t}\exp[-\Lambda F(t)]}{i_t!}} \\ &\quad \times \frac{F(t)^{i_t+\frac{1}{2}}\Lambda^{i_t-\frac{1}{2}}\exp\left[-\Lambda F(t)\right]}{\Gamma\left(i_t+\frac{1}{2}\right)}\mathrm{d}\Lambda \\ &= \frac{F(t)^{i_t+\frac{1}{2}}(1-F(t))^{e_0-i_t}}{\left(e_0+\frac{1}{2}\right)B\left(e_0-i_t+1,i_t+\frac{1}{2}\right)}\end{aligned}$$

Compared to Equation 2.2, this calculation seems relatively simple. However, without an upper limit for $E_0$, it may be unclear to what value of $E_0$ we need to calculate this distribution to in order to understand it fully. This will cause us problems later

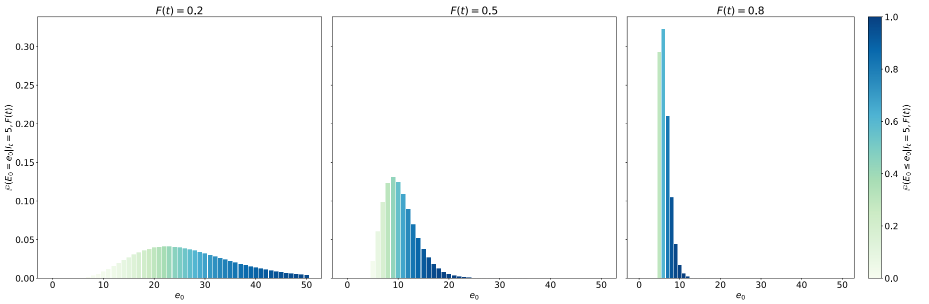


Figure 2.23. Estimating the total number of infected individuals when no upper limit is know. The number of symptomatic individuals observed is $I_t = 5$, but we no longer know how many this is out of. The left hand plot shows this scenario at time $F(t) = 0.2$, the middle plot shows it at time $F(t) = 0.5$ and the right hand plot shows it at time $F(t) = 0.8$. This is analogous to the scenario depicted in Figure 2.4.

when attempting to calculate the information this distribution careers and how well it matches up with simulated data.

First, though, because of its analytical simplicity, we can find the first two moments of this distribution, the mean $\mu$ and the variance $\sigma^2$:

$$
\begin{aligned}
\mu &= \sum_{e_0=I_t}^{\infty} e_0 \times \frac{F(t)^{I_t+\frac{1}{2}}(1-F(t))^{e_0-I_t}}{\left(e_0+\frac{1}{2}\right) B\left(e_0-I_t+1, I_t+\frac{1}{2}\right)} \\
&= \frac{1-F(t)+2I_t}{2F(t)} \\
\sigma^2 &= \sum_{e_0=I_t}^{\infty} (\mu-e_0)^2 \times \frac{F(t)^{I_t+\frac{1}{2}}(1-F(t))^{e_0-I_t}}{\left(e_0+\frac{1}{2}\right) B\left(e_0-I_t+1, I_t+\frac{1}{2}\right)} \\
&= (1-F(t))\frac{2I_t+1}{2F(t)^2}
\end{aligned}
$$

There is a linear relationship between the number of symptomatic individuals $I_t$ and the estimated total number of infected individuals, with each additional symptomatic individual increasing this estimate by a factor of $\frac{1}{F(t)}$. As time increases, $F(t)$ will also increase and our mean estimate will tend towards $I_t$. We can get an approximate idea for the certainty of this distribution by calculating the dispersion, in this chapter calculated as $\frac{\sigma^2}{\mu}$:

$$
\begin{aligned}
\frac{\sigma^2}{\mu} &= \frac{(1-F(t))(2I_t+1)}{F(t)(1-F(t)+2I_t)} \\
\frac{\mathrm{d}\sigma^2/\mu}{\mathrm{d}F(t)} &= -\frac{(2I_t+1)((F(t)-1)^2+2I_t)}{F(t)^2(F(t)-2I_t-1)^2} \\
\frac{\mathrm{d}\sigma^2/\mu}{\mathrm{d}I_t} &= -\frac{2(1-F(t))}{(F(t)-2I_t-1)^2}
\end{aligned}
$$

Whilst looking at the initial dispersion formula, it can be difficult to interpret how

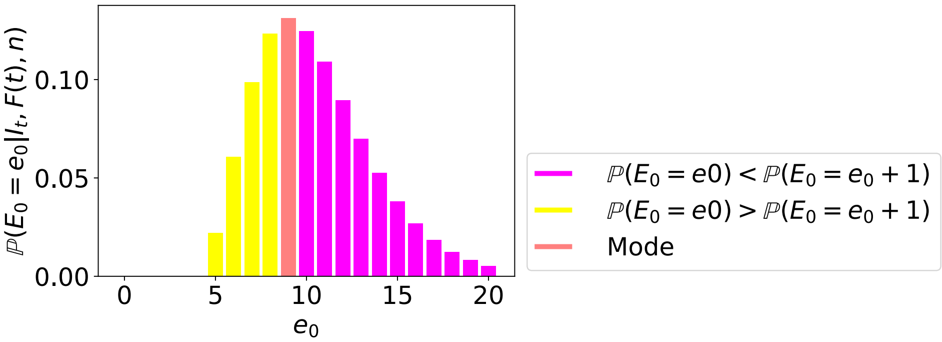


Figure 2.24. The mode value can be found at the first value in a series where the subsequent probability is less than the probability o the current value

time and the total number of symptomatic individuals affects the uncertainty in our system. However, its derivative with respect to both $I_t$ and $F(t)$ is negative. As time progresses and/or we seem more symptomatic individuals, the dispersion decreases. This gives us an interesting clue as to a difference between this model and when a total number of exposed individuals is known. In that case, we saw that the uncertainty in our system may initially increase to an apex before decreasing again, the timing of this apex dependent predominantly on the proportion of individuals who are symptomatic. While these two measures are different, we may expect to see a similar result when we look at the entropy of our new system in the next section.

The final piece of analysis we can perform is to search for the mode value of $E_0$. We may no be certain from looking if there is only one mode in out distribution. However, we can define the mode value as the first in any series of values where $\mathbb{P}(E_0 = e_0|I_t, F(t)) > \mathbb{P}(E_0 = e_0 + 1|I_t, F(t))$. We can use this inequality to find conditions where modes occur.

$$
\begin{aligned}
\mathbb{P}(E_0 = e_0|I_t, F(t)) &> \mathbb{P}(E_0 = e_0 + 1|I_t, F(t)) \\
\frac{F(t)^{I_t+\frac{1}{2}}(1-F(t))^{e_0-I_t}}{\left(e_0+\frac{1}{2}\right) B\left(e_0 - I_t + 1, I_t + \frac{1}{2}\right)} &> \frac{F(t)^{I_t+\frac{1}{2}}(1-F(t))^{e_0-I_t+1}}{\left(e_0+\frac{3}{2}\right) B\left(e_0 + I_t + 2, I_t + \frac{1}{2}\right)} \\
\frac{\Gamma\left(e_0+\frac{1}{2}\right)}{\Gamma\left(e_0 - i_t + 1\right)\Gamma\left(I_t+\frac{1}{2}\right)} &> \frac{(1-F(t))\Gamma\left(e_0+\frac{3}{2}\right)}{\Gamma\left(e_0 - I_t + 2\right)\Gamma\left(I_t+\frac{1}{2}\right)} \\
e_0 - I_t + 1 &> \left(e_1 + \frac{1}{2}\right)(1-F(t)) \\
e_0 &> \frac{2I_t - F(t) - 1}{2F(t)}
\end{aligned}
$$

This distribution is uni-modal, with that mode occurring at the smallest value of $E_0$ such that $E_0 > \frac{2I_t - F(t) - 1}{2F(t)}$. Any value higher than this value will always be less likely than the one before, as will any value prior to it.

### 2.5.3 Performance analysis

In previous scenarios we have based our assessment of the performance of our model on variations of the Wasserstein-1 metric ($\mathbb{W}_1$) and the Shannon entropy ($\mathbb{H}$) of the output distributions. As a reminder, for given values of $E_0, I_t$ and $n$, we calculated these using the following finite sums:

$$\begin{aligned}\mathbb{W}_1 &= \sum_{e=I_t}^{n} |\mathbb{P}(E_{calc} \leq e) - \mathbb{P}(E_{sim} \leq e|I_t)| \\ \mathbb{H} &= \sum_{e=I_t}^{n} -\mathbb{P}(E_{calc} = e|I_t) \times \mathbb{P}(E_{sim} = e|I_t) \times \log_2 \left[\mathbb{P}(E_0 = e|I_t)\right]\end{aligned}$$

In these circumstances we knew $n$ to be the upper limit value of $E_0$ (i.e. we could not have more infected individuals than exposed individuals). However, in our current model, we do not have an upper limit for the number of possible infected individuals, meaning in order find the Wasserstein-1 metric and Shannon entropy we need to solve the following infinite sums:

$$\begin{aligned}\mathbb{W}_1 &= \sum_{e=I_t}^{\infty} |\mathbb{P}(E_{calc} \leq e|I_t) - \mathbb{P}(E_{sim} < e|I_t)| \\ \mathbb{H} &= \sum_{e=I_t}^{\infty} -\mathbb{P}(E_0 = e|I_t) \times \log_2 \left[\mathbb{P}(E_0 = e|I_t)\right]\end{aligned}$$

Fortunately the $\mathbb{W}_1$ sum can be found analytically. We start by calculating the cumulative mass function for our model:

$$\begin{aligned}\mathbb{P}(E_0 \leq e|I_t) &= \sum_{k=I_t}^{e} \frac{F(t)^{I_t+\frac{1}{2}}(1-F(\tau))^{k-I_t}}{\left(k+\frac{1}{2}\right) B\left(k-I_t+1, I_t+\frac{1}{2}\right)} \\ &= \begin{cases} 1 - \frac{B_{1-F(t)}\left(e-I_t+1, I_t+\frac{1}{2}\right)}{B\left(e-I_t+1, I_t+\frac{1}{2}\right)} & e \geq I_t \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

Gratifyingly, this probability tends towards 1 as $e$ tends towards $\infty$, in keeping with a cumulative mass function for an infinite series. Also, as $F(t)$ tends to 1, this value also tends to 1, in keeping with the fact that when $F(t) = 1$, $I_t = E_0$. Following simulation, we would now be able to calculate part of the Wasserstein-1 metric numerically upto all observed values of $E_{sim}$ given a value of $I_t$ at time t:

$$\mathbb{W}_1(I_t, F(t)) = \left(\sum_{e=I_t}^{E_{max}} |\mathbb{P}(E_{calc} \leq e|I_t, F(t)) - \mathbb{P}(E_{sim} < e|I_t, F(t))|\right) + \phi$$

where $E_{max}$ is the highest number of infected individuals observed in our simulations for input values of $F(t)$ and $I_t$ and $\phi$ is the remaining as of yet uncalculated part of the Wasserstein-1 distance. As $\mathbb{P}(E_{sim} > E_{max}) = 1$, the remainder of this Wasserstein-1 distance simplifies:

$$\mathbb{W}_1 = \left[\sum_{e=I_t}^{E_{max}} |1 - \frac{B_{1-F(t)}\left(e - I_t + 1, I_t + \frac{1}{2}\right)}{B\left(e - I_t + 1, I_t + \frac{1}{2}\right)} - \mathbb{P}(E_{sim} \leq e)|\right] + \sum_{e=E_{max}+1}^{\infty} \frac{B_{1-F(t)}\left(e - I_t + 1, I_t + \frac{1}{2}\right)}{B\left(e - I_t + 1, I_t + \frac{1}{2}\right)}$$

We estimate the second part of this series by calculating it for each value of $e$ until its addition to the Wasserstein-1 metric is less than 0.01% of the current calculated total.

The infinite sum for the Shannon entropy cannot be calculated analytically. Instead, again, we use the cumulative mass function for our model. If the infinite sum for the Shannon entropy does converge, then calculating

$$\sum_{k=I_\tau}^{K} \mathbb{P}(E_0 = k|I_t) \times \log_2 [\mathbb{P}(E_0 = k|I_t)]$$

for increasing values of $K$ will generate increasingly accurate estimates of $\mathbb{H}$. We label the estimate for $\mathbb{H}$ generated from the sum between $I_\tau$ and $k$ as $\mathbb{H}_{k \text{ est}}$. Figure 2.25 demonstrates how closely related the increase in $\mathbb{H}_{k \text{ est}}$ is to $\mathbb{P}(E_0 = k|I_t)$. We therefore estimate $\mathbb{H}$ as $\mathbb{H}_{k \text{ est}}$ for the minimum value of $k$ such that $\mathbb{P}(E_0 \leq k|I_t) \geq 0.99999$.

We now have methods for calculating both the Wasserstein-1 metric and the Shannon entropy. As there is no upper limit, a normalised version for either of these metrics is not necessary. We simulate 1000000 outbreak events where the upper limit of exposed individuals is unknown. For each simulation we randomly select a rate for the Poisson distributed infectious connections from an exponential distribution of rate 20. Using this rate, we then assign a random number of infectious connections made, and for each infected individual, an incubation period, again selected from a Uniform distribution between 0 and 1 so that $t = F(t)$ for $0 \leq t \leq 1$. We calculate the range of the Wasserstein-1 metric and Shannon Entropy given our simulation and their expected values given a fixed outbreak size and the proportion of the outbreak that are already symptomatic.

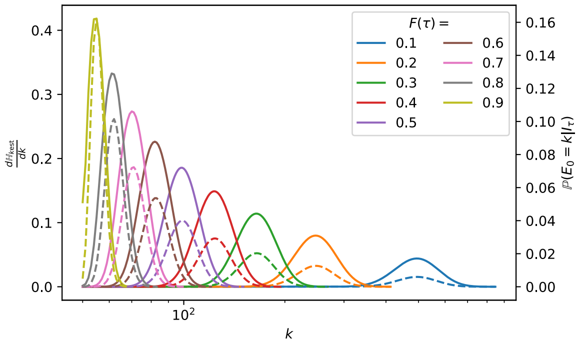


Figure 2.25. Demonstration of the close relationship between the probability density function and the increase in $\mathbb{H}_{k\text{ est}}$ for different values of $F(t)$ when $I_t = 50$. The complete lines show the probability mass at that value of $e_0$, while the dotted line shows the corresponding change in the entropy estimate $\mathbb{H}_{k=e_0\text{ est}}$.

### 2.5.4 Wasserstein-1 results

As expected, some aspects of the Wasserstein-1 metrics resulting from our simulation remain the same regardless of if $n$, the number of exposed individuals, is known or not, as can be seen in Figure 2.26. For example, as time progresses, the Wasserstein-1 distance decreases and by time $F(t) = 1$, $\mathbb{W}_1 = 0$. However, there are some notable changes in our results, as well as how we can display them.

Our fixed parameters are no longer the same. As we no longer have a value for $n$, we can no longer work out the proportion of the exposed group that was infected, nor the proportion of the exposed group that is symptomatic by time $t$. Instead, we have shown the effect of fixed outbreak size (compared to a fixed value of 363 which was the largest outbreak seen in our simulation) and the proportion of infected individuals that are symptomatic by time $t$ compared to the true outbreak size $E_0$. In a real-world scenario, the true value of $E_0$ would be expected to be unknown.

The relationship with outbreak size is fairly clear: The larger the outbreak, the worse our calculated distribution matches the simulated distribution. There are orders of magnitude difference between the larger and smaller outbreaks, which broadens (in terms of orders of magnitude) as time progresses. The expected Wasserstein-1 metrics for the largest outbreaks are still approximately 10 even late into the model. This may seem large, but as it represents an outbreak of a size greater than 300, an error of $\pm 10$ may be acceptable. We are not able to demonstrate a meaningful proportional error as there is no upper limit to our distribution.

The relationship between expected Wasserstein-1 metric and proportion of infected individuals that are symptomatic is less clear. There definitely seems to be some effect where a higher proportion of symptomatic individuals results in a less accurate model. However, this is not monotonic, with some fixed proportions crossing over over time. This in part may be due to the fact that certain values of $\frac{I_t}{E_0}$ were rare to see in simulation (we were less likely to see a high proportion of symptomatic in-

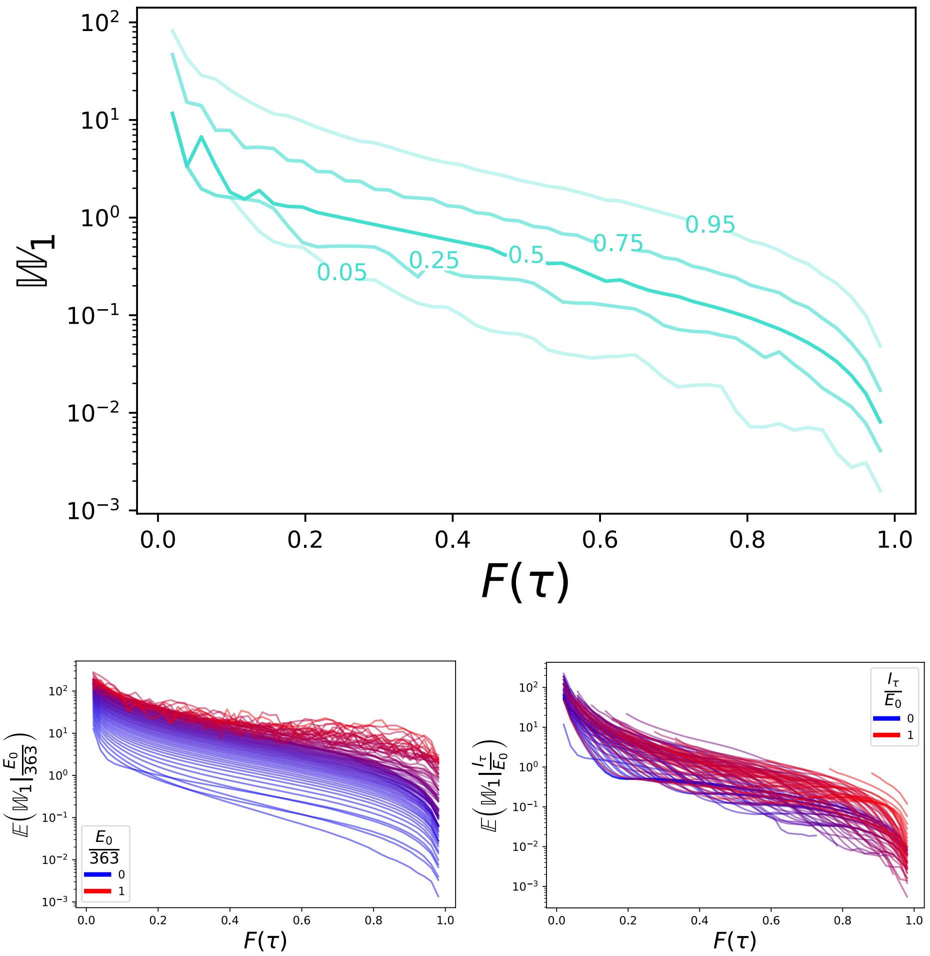


Figure 2.26. Estimated Wasserstein-1 metrics for simulated outbreaks when the exposure size is unknown. The top result shows the centile range of results, bottom left demonstrates the expected effect of group size and bottom right demonstrates the expected effect of the proportion of the infected group that are symptomatic. Unlike when the exposed group size was known and fixed, we rely on Poisson distribution to dictate the outbreak size and therefore number of individuals symptomatic at a particular time. The result is a less smooth looking graph, with erratic values at under-represented parameter sets (for example, large outbreak sizes) and narrow centile gaps at over-represented points (for example the 0.05 and 0.25 centile appear to touch when $F(t) \approx 0$, likely due to the Wasserstein-1 metric bein over-represented when $I_t = 0$.

dividuals early of in the simulation, for example) making early values of an estimate given a fixed value of $\frac{I_t}{E_0}$ less reliable. Investigating the Shannon entropy with these fixed terms will give us more insight into the nature of this relationship.

### 2.5.5 Shannon entropy results

The Shannon entropy of our model has similar behaviours to when the size of the exposed group is known. Over time we gain more knowledge, our model becomes more informative and the Shannon entropy decreases. A larger outbreak size results in a less informative model. What is interesting and new is the effect of the total number of symptomatic individuals.

By the nature of the simulation, we are no longer able to demonstrate the number of infected individuals in terms of $\frac{I_t}{n}$ and instead calculate it in terms of the size of the outbreak. As this is a simulation, we saw few instances of high proportions of symptomatic individuals until late on in the simulations, meaning estimates for the expected Shannon entropy for high values of $\frac{I_t}{E_0}$ early on are erratic if not non-existent. From our simulations it did appear that we can expect less information from higher proportion of symptomatic individuals, but this needed investigating outside the simulation.

Figure 2.28 demonstrates the calculated Shannon entropy for fixed values of symptomatic individuals in the range $0 \leq I_t \leq 100$. From this we can see far more clearly that more symptomatic individuals results in a less informative model. Earlier we noted that increasing the number of symptomatic individuals in this system will decrease the overall dispersion of the probability mass function. It is interesting that this decrease in dispersion also results in an increase in entropy.

This plot also reveals a very interesting difference when compared to a system where we know the true number of exposed individuals, $n$. In that scenario, our model would initially become less informative to an apex, only for the Shannon entropy to decrease again, reaching 0 when $F(t) = 1$. The proportion of symptomatic individuals caused said apex to happen later in our model. We concluded that initially we were certain that more individuals were infected than symptomatic, but as time increases, we became less certain until eventually we became more certain that no more individuals are infected.

The same shift in estimation is happening in our new model. Initially, we believe there are more infected individuals than there are symptomatic individuals. Then as time progresses, our estimate for the value of $E_0$ decreases until eventually we believe that only the symptomatic individuals were infected. The difference is that at the start, when $n$ is known, the number of infected individuals is constrained by the value of $n$. When we think there are many more infected individuals, our posterior estimates are forced to be less than or equal to $n$, resulting in a very informative distribution. When we do not know the value of $n$, our posterior distribution is no longer constrained. It is able to spread wider, resulting in a less informative distribution, or higher Shannon entropy. Figure 2.29 demonstrates how, for the same total number of symptomatic individuals, having a constraint on the total number of exposed individuals increases the information our model provides, but that over time this effect decreases.

### 2.5.6 Discussion

Without the constraint of a known value of $n$, we can calculate a posterior distribution for the total number of infected individuals from an exposure event by treating the total number of infectious connections made as falling on a Poisson distribution. This alters how our model behaves, but in some regards simplifies it dramatically. Now the more individuals infected and/or the more individuals that are symptomatic, the less informative it will be and the less in keeping it will be with simu-

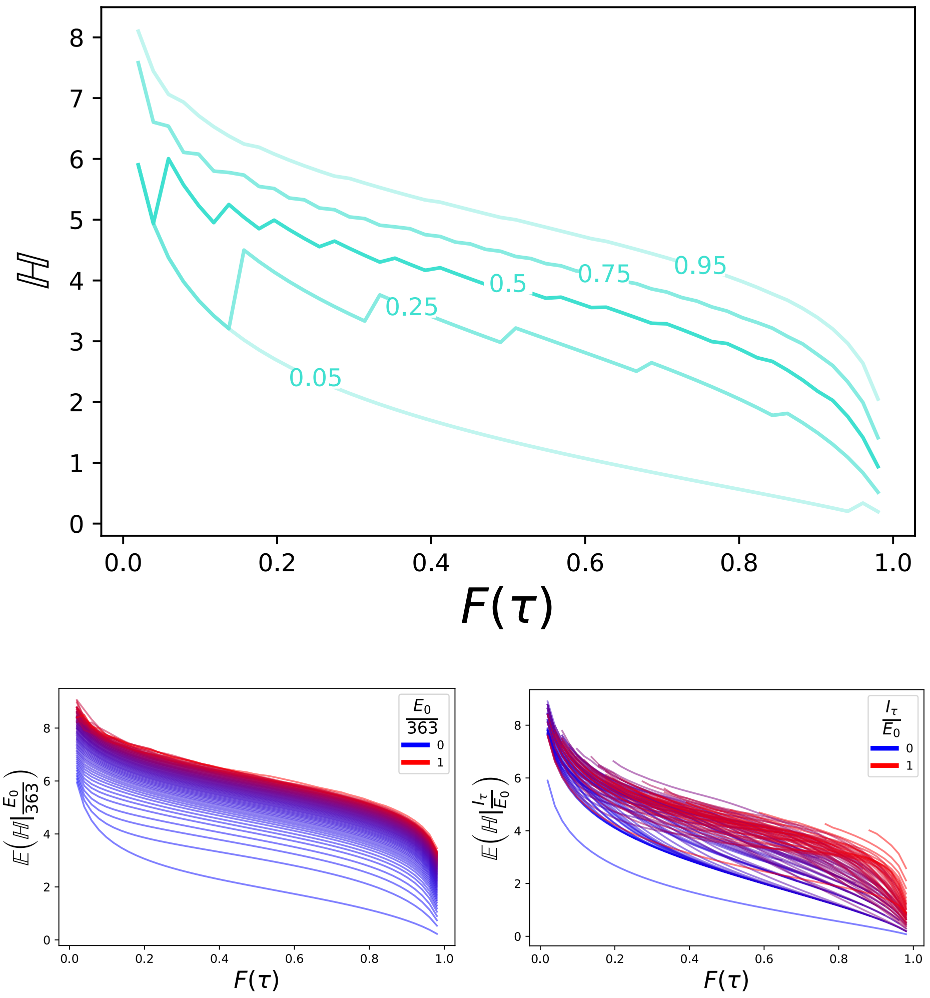


Figure 2.27. Expected Shannon entropy for distributions calculated during an outbreak simulation where the total number of exposed individuals is unknown. The top graph shows the range of possible values, the left graph shows the effect of outbreak size on the mean estimate for the Shannon entropy and the right side shows the effect of the proportion of an outbreak that has developed symptoms. We see a similar erratic nature for similar reasons as discussed in Figure 2.26.

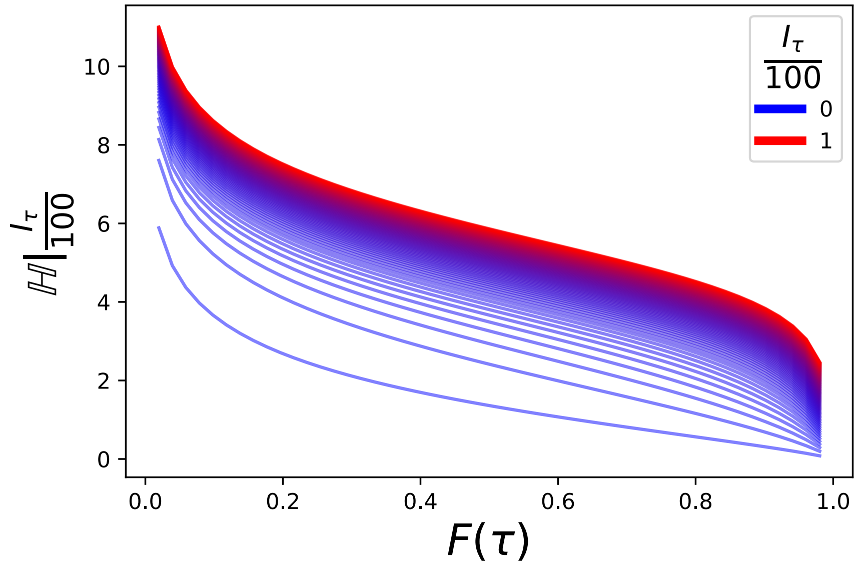


Figure 2.28. Changing Shannon entropy for a fixed number of observed symptomatic individuals

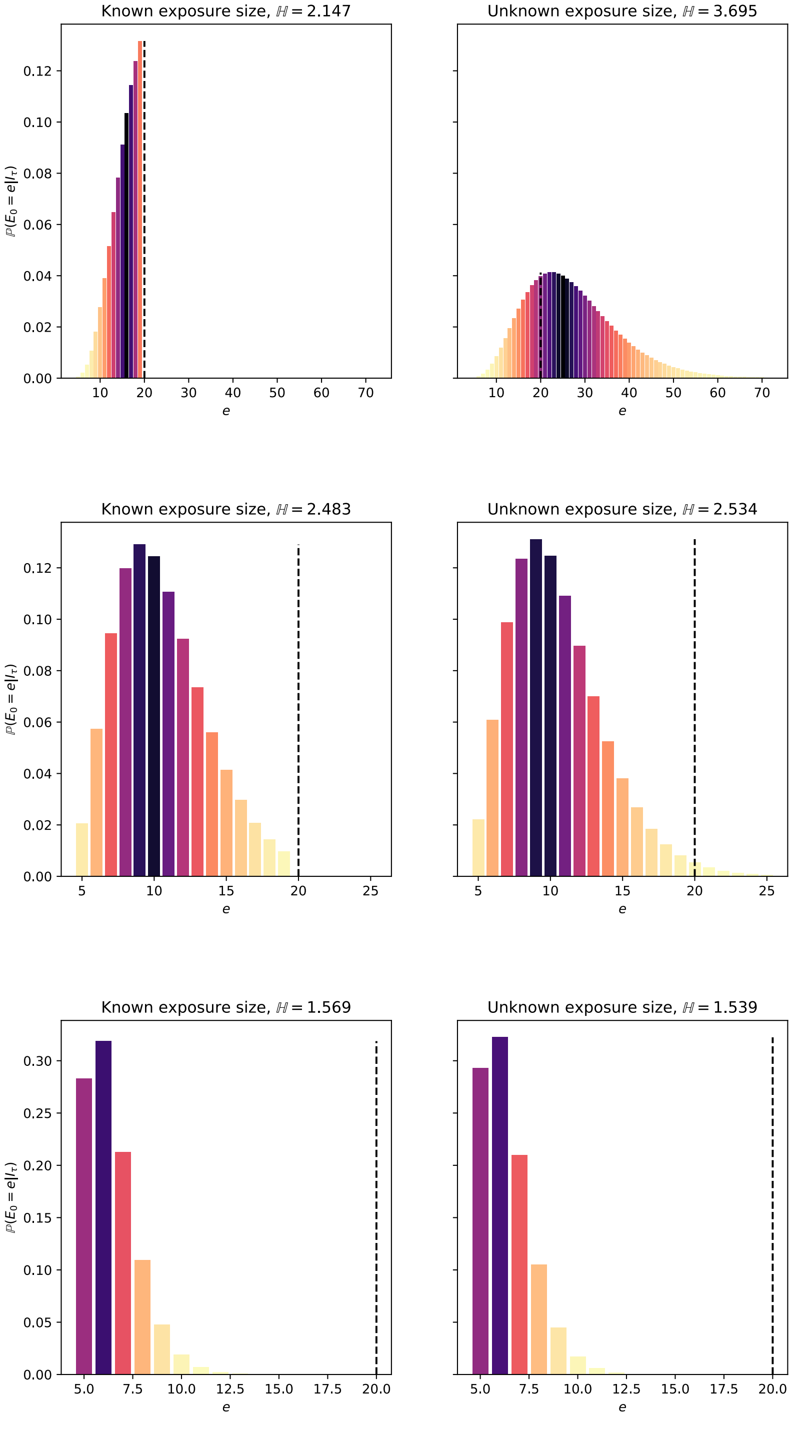


Figure 2.29. Comparison of posterior estimates for the total number of infected individuals when the exposure size is known and unknown. Left-hand graphs show the posterior when the exposure group is known to be of size $n = 20$ (as indicated by the dashed black line), while the right-hand group shows the effect of an unknown exposure group size. The effect changes over time, from $F(t) = 0.2$ (top) to $F(t) = 0.5$ (middle) to $F(t) = 0.8$ (bottom).

lated results. These errors are, of course, relative, as the model still remains fairly accurate.

Once again, our final thoughts go to answering to questions: “What is the probability that we have seen all infected individuals?” and “What is the probability that no individuals were infected in the first place?” We can solve both these questions with the same manipulation of our posterior function as before.

Estimating the probability that all infected individuals are symptomatic given no upper limit to the size of the exposure group

As with previous examples, the probability that we have seen all the symptomatic individuals is equivalent to the probability that the number of symptomatic individuals and the number of infected individuals is the same. This results in a surprising simplification:

$$\mathbb{P}(E_0 = I_t | n \text{ is unknown}, F(t)) = F(t)^{I_t + \frac{1}{2}} \tag{2.6}$$

Without a constraint on the total number of exposed individuals, the probability that we have seen all the infected individuals (that they are all already symptomatic) relies only on the total number of symptomatic individuals and the time that has passed (in terms of $F(t)$). As $F(t)$ increases, the probability that all infected individuals are already symptomatic also increases. However, whenever $I_t$ increases (whenever a new individual becomes symptomatic), the probability that we have seen all the infected individuals decreases, as $0 \leq F(t) \leq 1$, by a factor of $F(\tau)$. Late on in the incubation period this is unlikely to make a grand effect, but early symptomatic individuals reduce this probability at that time dramatically. Figure 2.30 demonstrates this changing probability for values of $I_t$ between 0 and 9.

Calculating the probability that no individuals were infected without an upper limit to the number of individuals exposed

As has been discussed previously, the probability that no individuals were infected is equivalent to the probability that all the infected individuals are symptomatic ($E_0 = I_t$) when there are no symptomatic individuals ($I_t = 0$). We assume this question is not being asked when there are symptomatic individuals, as it results in a rather trivial answer: If there are symptomatic individuals, the probability that no one was infected is 0 (investigating the effect of false positive diagnosis is beyond the scope of this work).

Equation 2.6 shows us that when $I_t = 0$, the probability that nobody was infected is

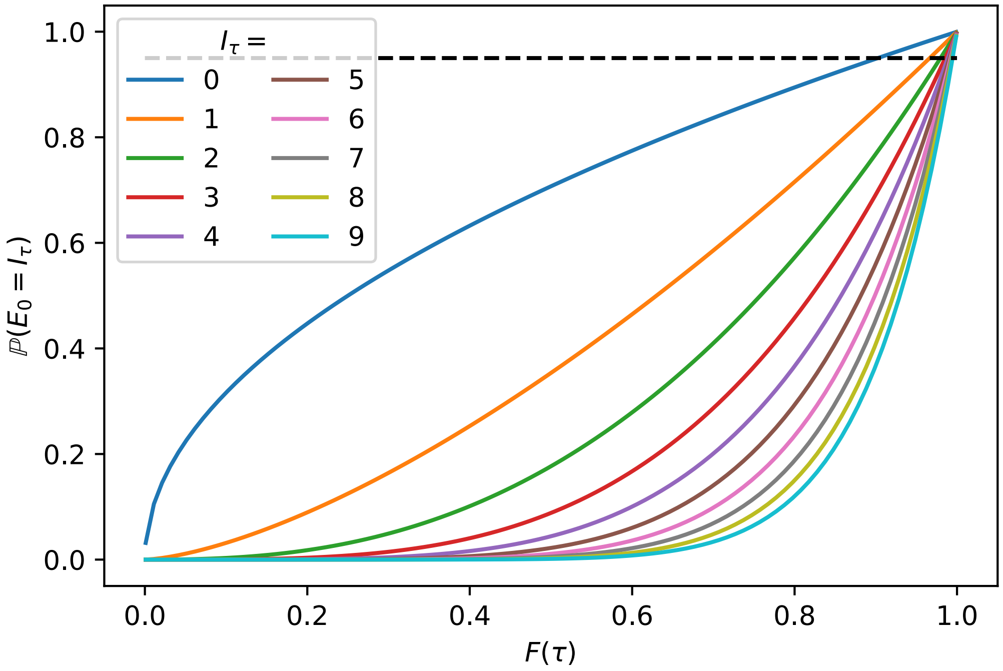


Figure 2.30. Changing probability that all symptomatic individuals have been observed. The black dashed line demonstrates the point at which this probability equals 0.95. The special case calculating the probability that no individuals were infected is equivalent to the line $I_t = 0$.

$$\mathbb{P}(E_0 = 0 | I_t = 0, n \text{ is unknown}, F(t)) = F(t)^{\frac{1}{2}}$$

We can see how that changes over time in Figure 2.30 by looking at the $I_t = 0$ line. For any scenario, we can say with a 95% certainty that no one was infected at the time $F(\tau) = 0.95^2 = 0.9025$.

## 2.6 Model 4: Estimating outbreak size for ongoing exposure

The previous sections have created models based on a scenario of instantaneous exposure at time $t = 0$. In each case we have assumed that all the individuals were exposed at once, and that that exposure immediately stopped. Their incubation periods could only start at time $t = 0$. In a lot of scenarios, this, of course, is unrealistic. It would be useful to extend this work to include some distribution for a transmission time.

If individuals in a group were exposed between 0 and $z$, we may be able to write the probability density function that describes when an individual is infected between times 0 and $z$ given that they are infected between times 0. We will call this function $y(t)$, which is equal to 0 outside of 0 and $z$, and integrates to 1 between these values. We start with the trivial case where the structure of $y(t)$ is independent to the probability of infection during this time, $\rho_z$. We want to find the probability that an individual would become symptomatic before some time $t$. This would require the "relatively" simple task of convolving $y(t)$ with $f(t)$ (the distribution for the incubation period) to generate a new function $k(t)$ that describes the distribution of time from start of exposure to developing symptoms. The cumulative density function $K(t)$ would replace $F(t)$ in our previous calculations, which would then pretty much stay the same.

However, what if $y(t)$, the distribution for time from start of exposure to infection, was not independent of the probability of infection during the exposure period, $\rho_z$?

Let's consider an ongoing exposure, starting at time $t = 0$, that provides a cumulative force of infection on a group. This cumulative force of infection is described by the cumulative hazard function $H(t)$. It is related to the probability of an individual being infected, in that the probability that an individual survives infection past time $t$, $1 - \rho_t$, is $\exp\left[-H(t)\right]$. This can be rearranged:

$$H(t) = -\ln\left[1 - \rho_t\right]$$

If $\exp\left[-H(t)\right]$ is the probability of surviving past time $t$ without infection, we can write the probability density function for the timing of infection, $y(t)$:

$$\begin{aligned} y(t) &= \frac{\mathrm{d}(1 - \exp\left[-H(t)\right])}{\mathrm{d}t} \\ &= \exp\left[-H(t)\right] \frac{\mathrm{d}H(t)}{\mathrm{d}t} \end{aligned}$$

If we can parameterise $H(t)$, the cumulative hazard function for any time $t$, in terms of $\rho_z$, the probability of being infected before a specific time $z$, we can find $y(t)$ in terms of $\rho_z$. In turn, $K(t)$, the probability that an individual is symptomatic by time $t$, would be calculated as the convolution of $y(t)$ and $F(t)$ and as a result would also be dependent on the probability of infection prior to time $t$. Dependent of the shape of $H(t)$, a change in the probability of infection will result in a change in the distribution for symptom onset given infection. The only exception to this is if the distribution for timing of infection was uniform between 0 and $z$. In this case, the cumulative hazard function is structured in such a way that the probability of being infected at any one time is equal throughout this time period.

The effect $\rho_z$ has on the shape of $K(t)$ puts a very important spanner in our previous model. Before, we showed that we did not have to keep track of the timing of symptom onset in order to get the maximum amount of information out of our model. The logic behind this was that the timing of an individual's symptom onset in itself was independent of the probability of infection. if an individual in an exposed group had a quick onset of symptoms, it simply indicated that they were one of many samples from the incubation period distribution and that therefore many other people were infected. We also showed this mathematically: the exact timing of symptom onsets integrated out of our posterior equations. However, this is now not true. We would need to keep track of the exact timings of symptom onset as $K(t)$ is now dependent on the value of $\rho_z$ (except in the Uniform case). This would result in the investigation of very specific relationships and simulations that we would no longer be able to generalise. It is theoretically possibly, but beyond the scope of this project.

There is an exception, though. If no individuals become symptomatic, then there would be no timing of symptom onset to keep track of. We aim to investigate how this model could be used to approximate the probability that no individuals were infected during the exposure period.

### 2.6.1 Calculating a posterior function for the probability of having been infected before a certain time

As with previous models, we start by calculating a prior distribution for the overall probability of infection, $\rho_z$. One of the advantages of a Jeffreys prior distribution is that it is in-variate on reparameterisation[147]. We could choose $z$ to represent any particular time frame in our model and our Jeffreys prior would still have the same influence on our posterior calculations. Similarly, we could calculate the probability of infection in terms of the cumulative hazard function or the constant rate of infection and we would have the same effect. As before, $x_t$ is the observation that someone is symptomatic prior to time $t$, meaning they were both infected and developed symptoms under time $t$. It equals 1 if they are symptomatic, and 0 otherwise. We are only using this method to analyse an environment where as of yet no one has developed symptoms, so we are not interested in observing the exact time of symptom onset. $K(t, \rho_z)$ is the cumulative density function for the time from start of exposure to developing symptoms. If $g(x_t, \rho_z)$ gives the probability of observation $x_t$ given the probability of being infected by time $z$ is $\rho_z$, we can write $g(x_t, \rho_z)$:

$$g(x_t; \rho_z) = \begin{cases} K(t, \rho_z) & x_t = 1 \\ 1 - K(t, \rho_z) & x_t = 0 \end{cases}$$

We can write a generalisation of the Jeffreys prior based on this probability mass function:

$$\pi(\rho_z) = \sqrt{-\mathbb{E}\frac{\mathbf{d}^2 \ln\left[g(x_t; |\rho_z)\right]}{\mathbf{d}\rho_z^2}}$$

$$= \sqrt{-\begin{cases} K(t,\rho_z)\frac{d^2 \ln[K(t,\rho_z)]}{\mathrm{d}\rho_z^2} & x_t = 1 \\ + & \\ (1-K(t,\rho_z))\frac{\mathrm{d}^2 \ln[1-K(t,\rho_z)]}{\mathrm{d}\rho_z^2} & x_t = 0 \end{cases}}$$
$$= K(t,\rho_z)^{-\frac{1}{2}}(1-K(t,\rho_z))^{-\frac{1}{2}}\frac{\mathrm{d}K(t,\rho_z)}{\mathrm{d}\rho_z}$$

Given that we have not seen any symptomatic individuals by time $t$, the posterior distribution for $\rho_z$ becomes:

$$\mathbb{P}(\rho_z|I_t = 0) = \begin{cases} K(t,\rho_z)^{-\frac{1}{2}}(1-K(t,\rho_z))^{n-\frac{1}{2}}\frac{\mathrm{d}K(t,\rho_z)}{\mathrm{d}\rho_z} \times c & 0 \leq \rho_z < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

where $c$ is a normalising constant. Through integration by parts we can find $c$:

$$\begin{aligned} c &= \left(\int_0^1 K(t,p)^{-\frac{1}{2}}(1-K(t,p))^{n-\frac{1}{2}}\frac{\mathrm{d}K(t,p)}{\mathrm{d}p}\mathrm{d}p\right)^{-1} \\ u &= K(t,p) \\ c &= \left(\int_{K(t,0)}^{K(t,1)} u^{-\frac{1}{2}}(1-u)^{n-\frac{1}{2}}\mathrm{d}u\right)^{-1} \\ &= \left[B_{K(t,1)}\left(\frac{1}{2}, n+\frac{1}{2}\right) - B_{K(t,0)}\left(\frac{1}{2}, n+\frac{1}{2}\right)\right]^{-1} \\ B_{K(t,0)}(a,b) &= 0 \\ c &= B_{K(t,1)}\left(\frac{1}{2}, n+\frac{1}{2}\right)^{-1} \end{aligned}$$

This exact value of $c$ will be necessary later when calculating the probability that no individuals have been infected. Interestingly, it is in keeping with our calculations for $c$ when all exposure occurs at time $t = 0$ in Equation 2.1, where $K(t,\rho_z) = \rho F(t)$.

### 2.6.2 Calculating the probability no individuals have been infected in an ongoing exposure

As with previous calculations, finding the posterior distribution for the total number of infected individuals by time $t$ (which we will write as $E_t$) is a case of integrating across all possible values of $\rho_z$. This task is made a lot easier when looking only for the probability that no individuals were infected:

$$
\begin{aligned}
\mathbb{P}\left(E_t = 0 | I_t = 0\right) = & \int_0^1 \frac{\mathbb{P}\left(E_t = 0 \;\;\&\;\; I_t = 0 | \rho_z = p\right)}{\mathbb{P}\left(I_t = 0 | \rho_z = p\right)} \\
& \times \mathbb{P}\left(\rho_z = p | I_t = 0\right) \mathrm{d}p \\
= & \int_0^1 \frac{(1-p)^n}{(1-K(t,p))^n} \\
& \times K(t,p)^{-\frac{1}{2}}(1-K(t,p))^{n-\frac{1}{2}} \frac{\mathrm{d}K(t,p)}{\mathrm{d}p} \times c\mathrm{d}p \\
= & \, c \times \int_0^1 (1-p)^n K(t,p)^{-\frac{1}{2}}(1-K(t,p))^{-\frac{1}{2}} \frac{\mathrm{d}K(t,p)}{\mathrm{d}p} \mathrm{d}p \\
= & \, c \times \Bigg[ (1-\rho_z)^n \left(-2\mathrm{arccos}\left[\sqrt{K(t,\rho_z)}\right]\right) \\
& - \int -n(1-p)^{n-1}\left(-2\mathrm{arccos}\left[\sqrt{K(t,p)}\right]\right) \mathrm{d}p \Bigg]_{\rho_z=0}^{\rho_z=1}
\end{aligned}
$$

$$
\mathbb{P}(E_t = 0 | I_t = 0) = c \times \left(\pi - \int_0^1 n(1-p)^{n-1} 2\mathrm{arccos}\left[\sqrt{K(t,p)}\right] \mathrm{d}p\right) \tag{2.8}
$$

Calculating $c$ is fundamental to calculating the probability that no individuals were infected during an ongoing exposure.

These terms are dependent on the function $K(t, \rho_z)$, the probability that an individual is symptomatic prior to time $t$ given a probability of $\rho_z$ of being infected prior to time $z$. In turn, $K(t, \rho_z)$ is dependent on the cumulative hazard function $H(t)$ and the cumulative density function for the incubation period $F(t)$. We are going to investigate two possibilities for the shape of the cumulative hazard function. In both cases, we are going to use an cumulative density function for the incubation period $(F(t))$ that is equivalent to an exponential distribution with the mean $\frac{1}{\omega}$.

### 2.6.3 Calculating the probability that no one has been infected during a constant force of infection

One of the simplest options for a cumulative hazard function could arguably be a constant force of exposure. This could be used to represent a constant environmental exposure, such as radiation from a material with a long half-life, or a constantly contaminated water supply (although in the short term the exposure would result from stochastic events where an individuals would drink from / wash with the water supply, over a long enough period of time these events could be represented as a constant exposure). With a constant force of exposure, the cumulative Hazard function can be written as:

$$H(t) = \lambda t$$

where $\lambda$ is the unknown force of exposure. We can calculate $\lambda$ in terms of $\rho_z$ by remembering the relationship between $\rho_z$ and $H(t)$:

$$\lambda = -\frac{\ln\left[1-\rho_z\right]}{z}$$

In turn, for any value of $\rho_z$, we can write the probability density function for the time from exposure start to infection, $y(t)$:

$$\begin{aligned}
y(t) &= \frac{\mathrm{d}H(t)}{\mathrm{d}t} \times \exp\left[-H(t)\right] \\
H(t) &= \lambda t \\
&= -\frac{\ln\left[1-\rho_z\right]}{z} t \\
y(t) &= -\frac{\ln\left[1-\rho_z\right]}{z} \times \exp\left[\frac{\ln\left[1-\rho_z\right]}{z} t\right] \\
&= -\left(1-\rho_z\right)^{\frac{t}{z}} \frac{\ln\left[1-\rho_z\right]}{z}
\end{aligned}$$

This is a reparameterisation of an exponential distribution with mean $\frac{-z}{\ln[1-\rho_z]}$. In order to find $K(t,\rho_z)$ we convolve our infection time distribution with $F(t)$, which in this case is $1-\exp[-\omega t]$:

$$K(t,\rho_z) = \int_0^t y(t-\tau|\rho_z) F(\tau) d\tau$$

In special cases $K(t,\rho_z)$ can be found analytically. However, its solution can be cumbersome, specific to $F(t)$ and does not at this stage provide any additional analytical insight and so will not be included here.

We can insert $K(t,\rho_z)$ into Equation 2.8 to calculate the probability that no individuals have been infected by time $t$ given that no one is symptomatic by time $t$ in an environment of some constant exposure. With this version of $K(t,\rho_z)$, $K(t,0) = 0$, meaning Equation 2.8 simplifies:

$$\mathbb{P}(E_t = 0|I_t = 0) = \frac{\Gamma(n+1)}{\sqrt{\pi}\Gamma\left(n+\frac{1}{2}\right)} \times \left(\pi - \int_0^1 n(1-p)^{n-1} 2\text{arccos}\left(\sqrt{K(t,p)}\right) dp\right)$$

Figure 2.31 shows how this probability changes with respect to time and size of exposure group. As time progresses with no individuals becoming symptomatic, we can become more and more convinced that no individuals have been infected. As time progresses and the incubation period becomes more and more insignificant when compared to the time that has passed, $K(t, \rho_z)$ will increasingly resemble $\rho_z$ (i.e. the probability that an individual will be symptomatic by time $t$ will be more dependent on the probability that they are infected before time $t$ than the probability that their incubation period is completed for large enough values of $t$). Replacing $K(t, p)$ with $p$ in Equation 2.8 and following the equation through proves this asymptotic relationship:

$$\begin{aligned}\mathbb{P}(E_t = 0 | I_t = 0, K(t, \rho_t) = \rho_t) &= \frac{1}{B\left(\frac{1}{2}, n + \frac{1}{2}\right)} \times \left(\pi - \int_0^1 n(1-p)^{n-1} \text{arccos}\left[\sqrt{p}\right] \mathrm{d}p\right) \\ &= 1\end{aligned}$$

However, it may take longer than expected to reach a significant level of certainty that no individuals have been infected. In a scenario where the incubation period falls on an exponential distribution of mean $\frac{1}{\omega}$, for even small group sizes we can expected to wait approximately 10x the average incubation period before we can be certain that no one has been infected. This value increases as our group size increases. This might go against our instincts that with more individuals we would gain information quicker and therefore be convinced earlier that no one has been infected. As we gain this extra information, we are also exposing more individuals, resulting in an increase in the probability that one of them may be infected without us knowing. This has important implications when drawing conclusions from large groups who have been exposed over a long period of time to a constant agent. While we could conclude that the individual risk to each individual is low over such a long period of time, we may still be missing one infection out of such a large group.

In this model, it is assumed that given enough time, everyone exposed will eventually be “infected”. This may not be accurate, as a subset of the exposed group may be immune to the exposure. Alternatively, the exposure may wain in strength. It is this latter option which we shall now investigate as our final model in this investigation.

### 2.6.4 Calculating the probability that no one was infected during an exposure to a waning force of infection

The previous section looked at the effect of a constant force of exposure resulting in the cumulative hazard function:

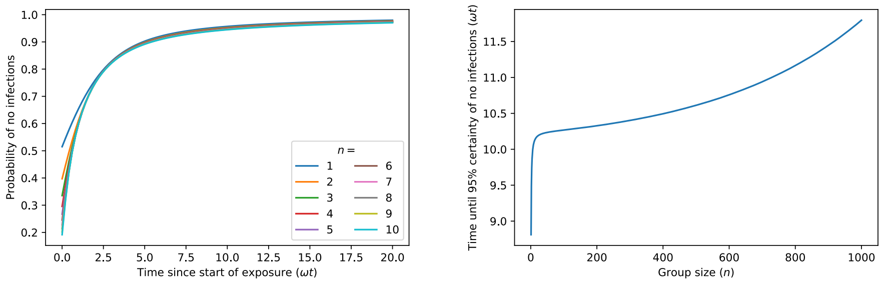


Figure 2.31. Demonstration of the probability that no individuals have been infected given a constant force of exposure. Time is given in terms of $\omega$, where $\frac{1}{\omega}$ is the mean incubation period. The left-hand plot shows how this probability changes over time for different group sizes ($n$) and the right-hand plot show the time at which we can be 95% certain that no one in the group has been infected.

$$H(t) = \lambda t$$

Now we consider a scenario where the force of infection is decreasing over time. Specifically, in a closed environment where an outbreak has been occurring, the most recent individual to have a positive diagnosis was diagnosed at time $t = 0$. There are $n$ remaining individuals who have not had a positive diagnosis. For sake of mathematical ease, we say that both the incubation period and the infectious period of the disease fall on exponential distributions. This being the case, if the incubation period for the disease falls on an exponential distribution of mean length $\frac{1}{\omega}$, then we can say that the mean of the infectious period is $\frac{1}{z\omega}$, where $z$ is a non-negative value. The force of infection caused by an infectious individual is constant throughout their infectious period and constant between people. Finally we assume that an individual's infectious period starts on or before they receive a positive diagnosis. Therefore we can workout the expected cumulative force of exposure generated by individuals who were infectious on or prior to time $t = 0$ from time 0 to time $t$ as a product of the integral of the probability that they will still be infectious at said time:

$$\begin{aligned} H(t) &= \lambda \int_0^t \exp\left[-z\omega\tau\right] \mathrm{d}\tau \\ &= \lambda \frac{1 - \exp\left[-z\omega t\right]}{z\omega} \end{aligned}$$

In the example with a constant force of exposure, $\int_0^\infty H(t)\mathrm{d}t = \infty$. As a result, given an infinite exposure period, every individual would be infected. However, as our new hazard function does have a finite integral between 0 and infinity, the probability of eventual infection, $\rho_\infty$ is less than 1. Specifically:

$$\begin{aligned}\rho_\infty &= 1 - \exp\left[-H(\infty)\right] \\ &= 1 - \exp\left[-\frac{\lambda}{z\omega}\right]\end{aligned}$$

Putting $H(t)$ in terms of the probability of ever being infected, we find:

$$H(t) = -\ln\left[1-\rho_\infty\right]\left(1-\exp\left[-z\omega t\right]\right)$$

In turn, we can calculate a new probability density function for the time of infection, $y(t, \rho_\infty)$

$$\begin{aligned}y(t,\rho_\infty) &= \exp\left[-H(t)\right]\frac{\mathrm{d}H(t)}{\mathrm{d}t} \\ &= -(1-\rho_\infty)^{1-\exp[-z\omega t]} z\omega \ln\left[1-\rho_\infty\right]\exp\left[-z\omega t\right]\end{aligned}$$

Unfortunately we cannot find an analytical solution for $K(t, \rho_\infty)$, the convolution between the infection time distribution $y(t)$ and the probability that an individual's incubation period is less than $\tau$, $F(\tau)$, for which we were using an exponential distribution with a mean of $\frac{1}{\omega}$. $K(t, \rho_\infty)$ can be found numerically fairly easily, as can, therefore, $\pi(\rho_\infty|z, \omega, F(t))$, the prior distribution for the value of $\rho_\infty$. Finally, using previously shown formulae we can calculate not just the probability that no exposed individuals have been infected by time $t$, but the probability that no individuals will ever be infected by individuals that were infectious prior to time $t = 0$ (this is why we put our calculations in terms of $\rho_\infty$ instead of $\rho_t$). In effect, this is the probability that the previous generation will fail to infect the next generation, or the probability that the outbreak has come to an end. Of course there is the possibility that an individual infected prior to time $t = 0$ will only become infectious after time $t = 0$, but we assume that if this is the case a) this individual would be identified meaning we would know not to declare the outbreak as over and b) that they would be identified prior to the time it would take for us to be 95% certain the outbreak is over.

Figure 2.32 shows how this probability changes over time for different values of $n$ and $z$. For group sizes greater than $n = 10$, the probabilities remain fairly constant. This pattern continues up to $n = 1000$. If we are observing a group exposed to a waning force of infection where the force of infection wanes at an exponential rate, so long as all involved remain asymptomatic, the time at which it will take for us to be convinced that no one will ever be infected is nearly independent of the total number of exposed individuals. This has important implications regarding observing the end of an outbreak. The size of the remaining susceptible population should by and large not influence how long it takes for us to declare an outbreak is over.

What is far more influential is the relationship between the expected incubation period and the expected infectious period. As the passage of time is given in terms of the expected incubation period, it is not hard to see that a longer incubation period will result in a longer time until we are certain the outbreak is over. What has a more dramatic and obvious effect, though, is the length of the infectious period compared to the incubation period $\left(\frac{1}{z}\right)$. A longer infectious period greatly delays the time it takes for us to be certain that no individuals in the next generation will be infected.

Although infectious periods and incubation periods may not be exactly exponentially distributed, this approximation serves as a good rule of thumb to approximate the end of an outbreak in a closed environment if the means of both distributions are known. An assumption that it is vulnerable to is that all infected individuals will eventually be symptomatic. If an individual manages to be infectious without ever developing symptoms, then a generation of the outbreak could go undetected resulting in a large apparent time between the two observable generations and a possible premature conclusion that the outbreak is over. One possible way of preventing this is through rigorous testing of asymptomatic individuals. Any positive case discovered would be counted as a new symptomatic individual, alongside any actually symptomatic individuals, and the wait for the outbreak to be over would start again, with a new value of $n$ exposed individuals.

A final point to be made regarding this model is the concept of homogeneous mixing. As soon as we introduced the concept of an infectious individual, we made the assumption that they would be mixing with every other individual equally at all times. This is implicit in our choice of cumulative hazard function. Further investigations could be made looking at alternative hazard function, as well as different distributions for the infectious period and incubation period, but this is beyond the scope of this study. For the time being, our model just calculates a useful ready-reckoner for approximating if an outbreak has drawn to a conclusion.

## 2.7 Conclusions

This chapter has focused on calculating an approximation for the size of a first generation during an outbreak event. We started with a simple point exposure to a group and showed how the number of individuals who had developed symptoms before a set time could be used to approximate the total number of individuals who had been infected in the first place. We showed how group size, outbreak size and proportion symptomatic affected the accuracy of this approximation, as well as the information it provides.

We showed the interesting relationship between our model and the probability of ever becoming symptomatic given an infection $(a)$. This probability stretches out

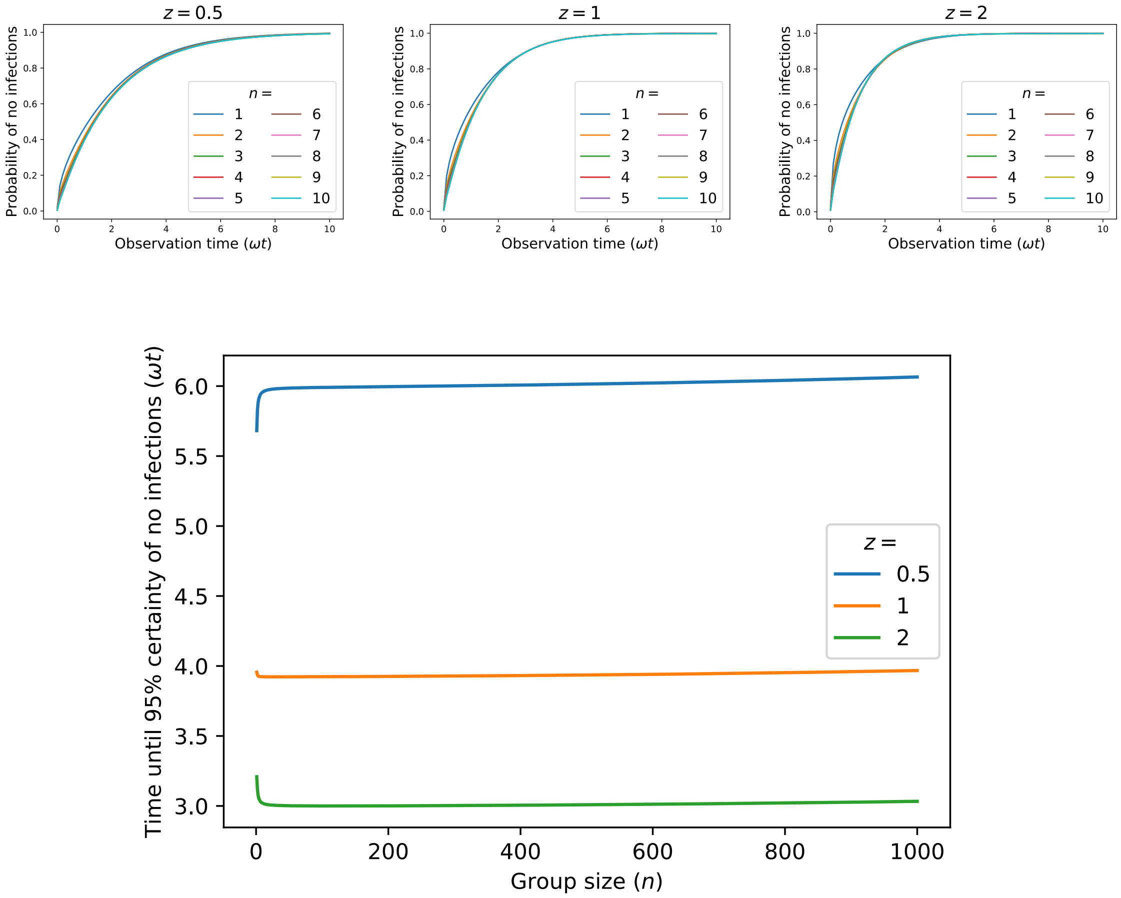


Figure 2.32. Demonstration of the approximate probability that we have reached the end of an outbreak. From left to right, the top graphs show how this probability changes over time (in terms of the expected length of the incubation period $\left(\frac{1}{\omega}\right)$ when the expected length of an infectious period is twice as long $\left(\frac{2}{\omega}\right)$, the same length $\left(\frac{1}{\omega}\right)$ and half as long $\left(\frac{1}{2\omega}\right)$ as the incubation period. The bottom figure shows us that whilst for $n \approx 10$, group size has an effect on the time until we are certain an outbreak has concluded, the relationship between the length of the incubation period and the infectious period is far more important, with a proportionally longer infectious period ($z = 0.5$) resulting in a far longer waiting time.

time in our model, and ultimately decreases our certainty.

By turning our approach around and considering an infector making a series of infectious contacts, we removed the necessity to know the size of the exposed group. This gave us useful insights into the actual effect knowing the group size had on our model, as well as showing us the astonishingly simple calculations for the probability that all infected individuals had been observed.

Finally, we have investigated the effect of allowing infections to continue to occur after the start of our model. Due to analytical complexities we stuck to the scenario where, whilst there has been some exposure, people are yet to develop symptoms. In the scenario of a constant force of infection, population size continues to delay the time at which we can be certain no one has been infected. However, in our waning force of infection, group size quickly becomes far less important than incubation period and infectious period length.

In each of our models, we have always chosen the most simplistic model or distribution possible. If possible, we have eschewed choosing particular distributions at all. By design we put the passage of time in terms of the cumulative function for the incubation period, as this would mean that it would remain true for any incubation period chosen. Further complexities could be brought to this model. For example, we assumed each exposed individual would have the same probability of infection given exposure, but what if this probability was different for each individual, falling on some unknown Beta distribution? Similarly, what if the probability that an infected individual would ever develop symptoms varied from person to person, or was unknown. For that matter, what if the distribution for the incubation period itself was unknown? These are possible avenues for exploration at a later date. It is our hope, that should someone want to answer these questions, the work has laid down a strong enough foundation for their project.

As a final thought, let us consider the function $K(t, \rho)$. This function gave us the probability that an individual would be observed by time $t$ given a probability of $\rho$ of being a positive case. In our initial examples, $\rho$ represented the probability of being infected immediately. Then we extended the model so $\rho$ represented the probability of being infected prior to time $t$. Finally, we concluded with $\rho$ being the probability of ever being infected.

In each case, $\rho$ could have taken any value between 0 and 1, while $K(t, \rho)$ had a maximum of $F(t)$ for any value of $t$. The maximum probability of observing an infected person before time $t$ was equal to the probability that their incubation period was shorter than $t$.

By looking back through our example models, we can see the profound effect the shape of $K(t, \rho)$ with respect to $\rho$ can have on our model. This perhaps is best demonstrated by looking how changing our model changed the influence group size has on the time it took to be 95% certain that no infections have occurred.

In the first model, where individuals were infected at time 0, larger groups decreased the time it took to be 95% certain of no infections (Figure 2.12). In this case $K(t, \rho)$ was equivalent to $\rho F(t)$. In the case of a constant force of infection, we can see that larger groups increase the time this takes (Figure 2.31). Finally, with an exponentially waning force of infection, we can barely observe the effect group size has on this measure (Figure 2.32). It is our understanding that it was the changing shape for $K(t, \rho)$ that resulted in these dramatic differences in conclusion.

Perhaps then, a standardised expression of $K(t, \rho)$ would be sensible. For future analysis, we propose the following possible structure for $K(t, \rho)$:

$$K(t, \rho) = \rho F(t) \frac{B_\rho(\alpha, \beta)}{B(\alpha, \beta)}$$

By varying $\alpha$ and $\beta$, the way in which the probability of infection affected the probability of being observed can be altered. The shape of $K(t, \rho)$ can be dramatically changed to reflect the indirect influence the probability of infection has on detection. In turn, we could observe the the effect other factors, such as group size, has on the overall distribution. Future research into this model structure, and others like it, could be very useful for understanding the nature of missing data.

Summary:

This chapter shows a quick method of analysing small outbreaks in their first generation, where it is not yet certain if all individuals have developed symptoms. This is a good ready-reckoner that indicates the eventual size of the first generation, as well as when we can be certain that we have observed all the infected individuals. Extending this to include an ongoing exposure is certainly more analytically complicated, but possible in a scenario where no individuals have actually been shown to be infected. This is a useful tool to give up-to-date analysis of small introductory events into closed environments such as care-homes and prisons.

In the case of an exposure event, we can be 95% certain that everyone from a first generation has been observed once:

$$\frac{B(n-I_t,I_t+\frac{1}{2})F(t)^{I_t+\frac{1}{2}}{}_2\mathcal{F}_1(\frac{1}{2},I_t+\frac{1}{2},n+\frac{3}{2},F(t))}{B_{F(t)}(I_t+\frac{1}{2},n-I_t+\frac{1}{2})} \geq 0.95$$

or, in the case where the total number of infected individuals is unknown, once:

$$F(t)^{I_t+\frac{1}{2}} \geq 0.95$$

This occurs later when there is a greater proportion of symptomatic individuals, and it is possible for us to never be certain if the rate of asymptomatic infection is high enough. By analysing this distribution in terms of its Shannon entropy, we learnt that when an upper limit of exposed individuals is known, the distribution initially counter-intuitively becomes less informative over time to a peak when the proportion of symptomatic individuals is approximately equal to the probability that in infected individual is symptomatic.

# Chapter 3

# The role of rotas in mitigating workplace outbreaks

## 3.1 Introduction

The following chapter features research written in collaboration with Carl Whitfield. In particular, Carl generated the numerical shorthand for describing a working week and started the initial investigation into how test timing was important in reducing time at work whilst infectious. This enabled us to observe how important the rota pattern length was, as well as the timing of the test result.

The first UK lock-down for the SARS-CoV-2 pandemic started on the evening of the 23rd of March, 2020. On the 10th of May, then Prime Minister Boris Johnson announced the "road map" for easing lock-down restrictions[162]. During this time period, key-workers were identified whose work must continue and must continue on location. Other workers were either furloughed or started working from home, reducing contact time with individuals outside their households. This could have had two effects:

1. Key-workers were likely to see their highest level of exposure at work. This was both because they were interacting with more people at work than at home and because the individuals they were interacting with, if other key-workers, were in turn likely to have interacted with more individuals than members of their household. This is not necessarily true for scenarios where there are two or more key-workers in a household.

2. Disease transmission occurring at work was more likely to result in more infections than disease transmission within households (see Figure 3.1a). Transmission at work had the capability of resulting in further work place infections as well as household infections (see Figure 3.1b). This could create a larger infection tree than a household transmission, which could only result in more infections in that one household. Again, this observation would not be true for households with more than one key-worker.

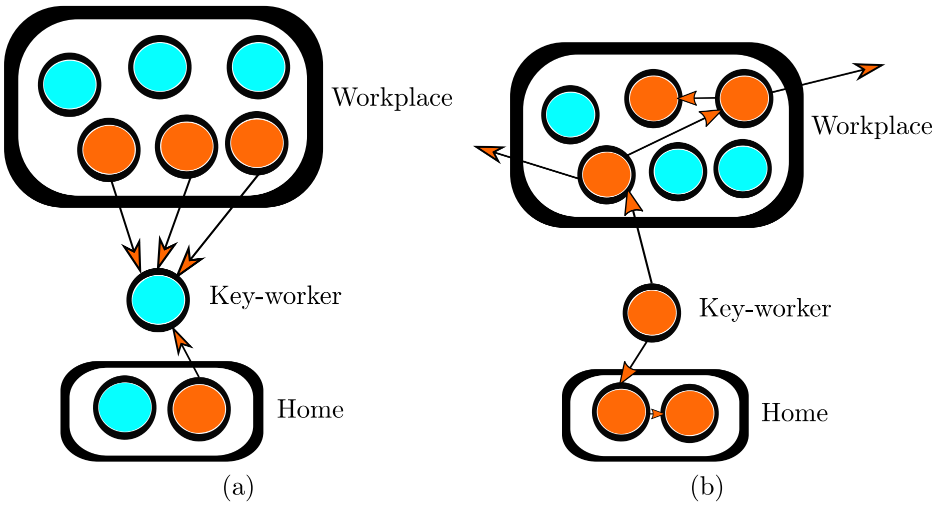


Figure 3.1. Visual demonstration of the effect lock-downs have on the infection pattern of key-workers. Blue circles represent uninfected individuals, orange circles represent infected individuals and orange lines show possible lines of transmission. In Figure 3.1a . we can see the higher number of people in the work place increases the probability of being infected there. The non-infected key-worker is more likely to be infected at work than at home. Additionally, if their household members are not key-workers, it would be less likely for them to get infected and therefore less likely for the the key-worker to pick up the infection at home than at work (this ignores the "strength" of an infectious connection, which may be stronger at home rather than work). In Figure 3.1b, the infected key-worker infects the same number of individuals in their home and their work place. The infection chain at home dies out once everyone is infected. However, at work, as their are more people, the infection chain spreads wider, including to outside work to other households.

Indeed, multiple studies demonstrated elevated risks of SARS-CoV-2 infection for key-workers. The initial phase of the REACT-2 study, a cohort study of previous and, at the time, current police officers, showed an elevated proportion of key-workers with associated antibodies following the first wave of the Coronavirus in the UK when compared to non-key-workers[163]. Similarly, four longitudinal studies through web-based surveillance showed an elevated proportion of infected individuals in the UK among key-workers in the first wave[164]. We should note that the elevated risk was not just associated with health-care environments. As stated earlier, the REACT-2 study looked at police officers and early genomic data from the COG-UK study showed strains of SARS-CoV-2 that spread between healthcare workers but never reached or came from patients[165].

Given the demonstrated increased risks from workplace transmission, reducing the rate of workplace transmission should be important in preventing a workplace outbreak. Some workplaces did institute workplace testing policies[166]. and it seems justified that in their paper outlining prioritise regarding vaccine administration, key-workers or individuals associated with key-workers (those that live in the same household, for example) made up 3 out of the 6 groups Kohns Vasconcelos et al. identified for targeted vaccination[167]. One area that has not been investigated, however, is the time that key-workers actually spend at work.

It is reasonable to assume that there is a relationship between the length of time an infectious person is at work and the probability that they transmit to a co-worker.

As part of the lock-down, if a key-worker developed symptoms they were to immediately stop working and return to their household for this reason. Hu et al. showed that that the presymptomatic, prodromal period plays an important role in the transmission of SARS-CoV-2[168]. Relying on self-reported symptoms may not be adequate for preventing work place transmission. Instead, we aim to investigate the relationship between rota schedules and the length of time an infectious person spends their prodromal period at work.

We are particularly interested in the length of time individuals spend in handover periods at the start and end of their shifts. A possible method of curtailing the spread of an infection through a workforce is scheduling key-workers to work in set cohorts, such that the same group of individuals work together. The aim with this intervention is to limit workplace transmission of an infection to only one cohort. Between shifts, a number of jobs require handovers between cohorts. These often need to be face-to-face, as they rely heavily on communication, and may require the transfer of a physical object, such as a paging device. Reducing the number of handovers an infectious person attends would be crucial to limiting the spread of a disease between cohorted groups. We aim to explore how changing shift patterns affects the distribution of the number of handovers attended whilst infectious (at the beginning and end of each shift) as well as the distribution of total time spent at work whilst infectious. The former will dictate how the disease will spread between cohorted groups, whilst the latter will dictate how the disease spreads within cohorted groups.

The following chapter will demonstrate two methods of approximating the length of time in infectious individual is at work given that they were infected at work, based on their rota pattern. Sections 3.3.1 and 3.3.2 demonstrate a numerical version of this analysis and its outcomes, and Section 3.4 will show how the same problem can be approached analytically through Fourier transforms. Finally, Section 3.5.1 explores a mathematical explanation for effect the rota pattern structure has on our results.

## 3.2 Parameter and Function description

Table 3.1 is a list of relevant parameter and function definitions for this chapter. The reader may find it useful to refer back to this table as and when required.

## 3.3 A numerical description of a rota pattern

### 3.3.1 Methods

We identify three key points in a person's infectious career for our model:

| Parameter/Function | Description |
|---|---|
| $a$ | Proportion of working day spent working |
| $\alpha_a$ | Alpha parameter for the Gamma distribution describing the latent period |
| $\alpha_{b1}$ | Alpha parameter for the Gamma distribution describing the infectious period |
| $\alpha_{b2}$ | Alpha parameter for the Gamma distribution describing the prodromal period |
| $b$ | Proportion of the working day spent resting |
| $\beta_a$ | Beta parameter for the Gamma distribution describing the latent period |
| $\beta_{b1}$ | Beta parameter for the Gamma distribution describing the infectious period |
| $\beta_{b2}$ | Beta parameter for the Gamma distribution describing the prodromal period |
| $c$ | Number of working days in a rota pattern |
| $c'$ | Number of full days worked in a rota pattern where $c$ is not an integer value |
| $C_{F,j}$ | The $j$th constant of the Fourier series that describes function $F$ |
| $d$ | Number of rest days in a rota pattern |
| $\delta(x)$ | The Dirac delta function |
| $\hat{F}(\omega)$ | The Fourier transform of the function $F(x)$ |
| $G\left(x, y, T_t, T_d, c+d\right)$ | The total number of tests that would be performed after $x$ that would come back before $y$ given regular testing at time $T_t$, and delay of $T_d$ and a rota length of $c+d$ |
| $\Gamma(a, x)$ | The upper incomplete Gamma function |
| $\Gamma(x)$ | The Gamma function |
| $H_{T_e}$ | The total number of handovers an individual infected at time $T_e$ attends whilst infectious |
| $\mathbf{h}_x$ | The set of all handovers after time $x$ given in terms of time $x$ |
| $\kappa(\tau)$ | The probability that an individual is infectious at time $\tau$ given they were infected at time 0 |
| $\kappa_1\left(t, T_e\right)$ | The probability that an individual is infectious at time $t$ given they were infected at time $T_e$ |
| $\kappa_2\left(t, T_i\right)$ | The probability that an individual is infectious at time $t$ given their infectious period started at time $T_i$ |
| $L(t)$ | The expected total length of time an individual will spend at work whilst infectious given they became infectious at time $t$ |
| $\Lambda(t)$ | The expected force of infection an infectious individual will generate at any point of a rota pattern $t$ |
| $\lambda(\tau)$ | The expected force of infection of an individual at time $\tau$ given they were infected at time 0 |
| $\mathbf{m}$ | The set of all starts of shifts in a rota pattern cycle, including the end of the rota cycle |
| $\mathbf{n}$ | The set of all ends of shifts in a rota pattern cycle |
| $\omega$ | The frequency of a wave |
| $\rho$ | The asymptomatic infection rate |
| $\varrho$ | The false negative rate for testing |
| $R(t)$ | The relative risk of infection at time $t$ in a rota pattern |
| $s(t)$ | An indicator function that equals 1 only if an individual is at work at time $t$ |
| $T$ | The length of a rota pattern |
| $\bar{t}$ | The amount of time worked in one rota pattern |
| $T_d$ | The delay between test and result |
| $T_e$ | The time at which an individual was infected |
| $T_i$ | The time at which an individual becomes infectious |
| $T_s$ | The time at which an individual stops being infectious |
| $T_t$ | The time in a rota pattern when a test is taken |
| $T_w$ | The total time at work whilst infectious |
| $u$ | The relative risk of infection at work |
| $v$ | The relative risk of infection outside of work |
| $w$ | The proportion of time worked in a rota pattern |
| $W(t)$ | The expected force of infection an individual will generate at time $t$ in the rota pattern whilst at work |
| $Y(j)$ | A function that simplifies the notation of certain Fourier transforms |

Table 3.1. A table of parameters and functions used in this chapter.

1. $T_e$ - The time at which the individual is infected
2. $T_i$ - The time at which an infected individual starts their infectious period
3. $T_s$ - The time at which an infectious individual stops being infectious, either by developing symptoms and isolating or completing their infectious period

An individual's shift pattern starts on the first day of their shift at the time $t = 0$, which is the start of their first shift. We assume we are looking at the rota cycle in which our individual is infected. Each shift is a proportion of 1 day $a$ with a proportion between shifts $b$ such that $a + b = 1$ day. The individual works $c$ shifts in a row before taking $d$ days off. The total shift pattern is $c + d$ days long.

We assume that individuals are infected with a Uniform probability while at work, but are never infected outside of work. The distribution for the time of infection during a workday is therefore Uniform:

$$\mathbb{P}\left(T_e = t_e | \text{Infected on day } x\right) = \begin{cases} \frac{1}{a}, & \text{if } x \leq t_e \leq x + a \\ 0, & \text{otherwise} \end{cases}$$

The probability of being infected on a particular day of a shift pattern is $\frac{1}{c}$ if it is a working day and otherwise 0. Therefore, the total distribution for chance of being infected at time $t_e$ given that the individual is infected during the first rota equals:

$$\mathbb{P}\left(T_e = t_e\right) = \begin{cases} \frac{1}{ac}, & \text{if } n \leq t_e \leq a + n \text{ for any integer } n \text{ in range } 0 \leq n < c \\ 0, & \text{otherwise} \end{cases}$$

We choose to use a Gamma distribution to represent the latent period for the disease we are representing. From the literature, it is difficult to estimate the true parameterisation of this distribution, as we can directly observe neither when an individual is infected nor when they become infectious. Instead, we subtract a 2 day prodromal period with a standard deviation of 1.5[159], [169], [170] from a incubation period with a mean length of 4.84 days and a standard deviation of 2.79. We approximate the result as a Gamma distribution with a mean of 2.84 days and a standard deviation of 2.34. This is not meant to be wholly accurate representation of the SARS-CoV-2 latent period and is for demonstration purposes only. A full list of parameterisations can be seen in Table 3.2. An alternative parameterisation could be chosen to emulate a different infection. We can use the mean and standard deviation to calculate input parameters for this distribution, $\alpha_a$ and $\beta_a$, such that a PDF for the latent period $(T_i - T_e)$ can be written as:

$$\mathbb{P}(T_i - T_e = t) \propto \frac{\beta_a^{\alpha_a}}{\Gamma\left(\alpha_a\right)} t^{\alpha_a - 1} \exp\left[-\beta_a t\right]$$

.

Similarly, we use a Gamma distribution to represent both the prodromal period (length of time from start of infectious period to development of symptoms if symptoms ever appear) and total infectious period, with means of 2 and 6 days and standard deviations of 1.5 and $\sqrt{12}$ respectively.

| Parameter | Value (s.d.) | Explanation/Sources |
|---|---|---|
| Prodromal period | 2 (1.5) days | Time from becoming infectious to developing symptoms (unless asymptomatic). [159], [169], [170] |
| Latent period ($T_i - T_e$) | 2.84 (2.34) days | Period from time of infection to becoming infectious. Estimated from the mean and standard deviation of the prodromal period (above) and incubation period [133], assuming independence of the prodromal and latent periods. |
| Asymptomatic infectious period | 6 ($\sqrt{12}$) days | The time an asymptomatic individual remains infectious, estimated from [168]. |
| Asymptomatic rate ($\rho$) | 0.3 | Probability of an individual being asymptomatic.* |
| False-negative rate ($\varrho$) | 0.3 | Probability that a test taken during the infectious period return negative.* |

Table 3.2. Parameters used in our working model. Due to high variability in literature, parameters marked with an * were chosen for illustrative purposes only.

If the infected individual does not develop symptoms, then they do not change their work pattern and the distribution for $T_s - T_i$ is taken from the total infectious period. If they do develop symptoms then it will be taken from the prodromal period, as once their prodromal period is completed they develop symptoms and remove themselves from the system, effectively ending their infectious period. If the probability of remaining asymptomatic is equal to $\rho$ and the total infectious period and prodromal period distributions take the parameters $\alpha_{b1}, \beta_{b1}$ and $\alpha_{b2}, \beta_{b2}$ respectively, then we can now write a distribution for $T_s - T_i$, the effective infectious period:

$$\begin{aligned}\mathbb{P}\left(T_s - T_i = t\right) = \rho &\left(\frac{\beta_{b1}^{\alpha_{b1}}}{\Gamma\left(\alpha_{b1}\right)} t^{\alpha_{b1}-1} \exp\left[-\beta_{b1} t\right]\right) \\ &\times (1-\rho)\left(\frac{\beta_{b2}^{\alpha_{b2}}}{\Gamma\left(\alpha_{b2}\right)} t^{\alpha_{b2}-1} \exp\left[-\beta_{b2} t\right]\right)\end{aligned}$$

With these two distributions, we can compute the probability that an individual is infectious at a particular time if we know when they are infected. Put another way, conditional on the individual being infected at time $T_e$, the probability that they are infectious at time $t \geq T_e$ (it is obviously 0 if $t < T_e$) equals the probability that the time the individual's infectious period starts, $T_i$, is less than $t$ and the end of their infectious period, $T_s$ is greater than $t$, i.e.:

$$\mathbb{P}\left(\text{Infectious at time } t | T_e, t \geq T_e\right) = \int_0^{t-T_e} \frac{\beta_a^{\alpha_a}}{\Gamma\left(\alpha_a\right)} x^{\alpha_a - 1} \exp\left[-\beta_a x\right]$$

$$\times \left( p\frac{\Gamma\left(\alpha_{b1}, \beta_{b1}\left(t - T_e - x\right)\right)}{\Gamma\left(\alpha_{b1}\right)} \right.$$
$$\left. + (1-p)\frac{\Gamma\left(\alpha_{b2}, \beta_{b2}\left(t - T_e - x\right)\right)}{\Gamma\left(\alpha_{b2}\right)} \right) \mathrm{d}x$$

where $\Gamma(a, y)$ is the upper incomplete Gamma function $\int_y^\infty t^{a-1} \exp\left[-t\right] \mathrm{d}t = \int_{by}^\infty b^a t^{a-1} \exp\left[-bt\right] \mathrm{d}t$

Unfortunately, a solution to the integral in this form does not exist (although we will consider alternatives in the following chapter). With a known parameter set we can calculate it numerically. We will use the function $\kappa_1\left(t, T_e\right)$ to describe this probability and $s(x)$ as an indicator function to show if an individual is at work at time $x$.

$$s(x) = \begin{cases} 1 & \text{at work at time } x \\ 0 & \text{otherwise} \end{cases}$$

With these functions we can now calculate the expected time at work whilst infectious, $\mathbb{E}\left[T_w\right]$ given that an individual is infected at time $T_e$:

$$\mathbb{E}\left[T_w|T_e\right] = \int_{T_e}^\infty \kappa_1\left(x, T_e\right) \times s(x)\mathrm{d}x \tag{3.1}$$

Similarly, if we define $\mathbf{h}_x$ as the infinite set of all starts and ends of shifts after time $x$ (i.e. the timing of all handovers after time $x$) we can calculate the mean of $H_{T_e}$, the total number of handovers an infectious individual will attend given that they were infected at time $T_e$:

$$\mathbb{E}\left[H_{T_e}\right] = \sum \kappa_1\left(\mathbf{h}_{T_e}, T_e\right) \tag{3.2}$$

Both Equation 3.1 and 3.2 involve either an infinite integral or infinite sum that we cannot solve in their current form. However, as $x$ increases in $\kappa_1\left(x, T_e\right)$ the resulting calculated probability will tend towards 0, because the further we are from the time at which an individual is infected, the less likely they are to be infectious. This means that whilst we cannot find exact solutions to Equations 3.1 or 3.2 we can approach their solutions numerically to any desired degree of accuracy.

In both cases, to find an overall estimate for the value of either $T_w$ or $H$, we need to integrate across all possible values for $T_e$:

$$\mathbb{E}\left[T_w\right] = \int_0^{c+d} \mathbb{P}\left(T_e = t\right) \mathbb{E}\left[T_w|T_e = t\right] \mathrm{d}t$$

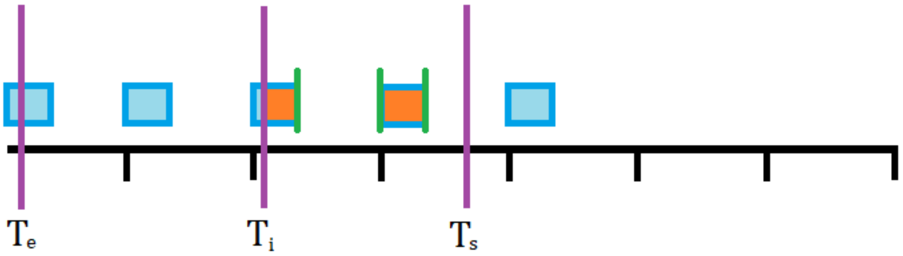


Figure 3.2. An example demonstration of an individual's at-work infectious profile. In this case, the individual works a 9-5 rota pattern, with at-work times demonstrated by the blue boxes. They are infected during their first shift ($T_e$), start becoming infectious part way through their third shift ($T_i$) and end their infectious period between their four and fifth shift ($T_s$). We are interested in the total length of time they are at work whilst infectious (sum of the widths of the orange boxes) and the number of beginning and end of shift handovers they attend (green lines).

$$\mathbb{E}\left[H\right] = \int_0^{c+d} \mathbb{P}\left(T_e = t\right) \mathbb{E}\left[H_t | T_e = t\right] \mathrm{d}t$$

remembering that $c$ is the number of days at work, $d$ is the number of days of rest and $c + d$ is the total length of a rota pattern. We are assuming that we start our observation of the infected individual at the start of the rota pattern they were infected in and in doing so ensure that we have integrated across every possible time that they could have been infected.

#### Including a test pattern

In many workplace environments it would be impractical to assign an exact time in a rota for individuals to receive a test. There are many systemic factors that may prevent such a rigid regime, such as availability of tests, the requirement for testing outside of work time and, in a 9-to-5 rota at least, the fact that a testing centre may not be able to cope with a sudden influx of tests once a week. However, we will continue to search for an optimum time for testing regardless, as it will give insight into priorities when designing a more flexible, real-world appropriate work place testing regime.

In practical terms, let's define $T_t$ as the time in a rota pattern that a test is taken and $T_d$ as the delay between test and result. This means that for any integer value of $n$, a test occurs at time $T_t + n(c + d)$, whose results come back at $T_t + T_d + n(c + d)$. When a positive test comes back, an infected individual isolates for the entire remainder of their infectious period (we are choosing to ignore individuals coming back from work early because of a misjudged infectious period or a subsequent false negative result). We also denote by $\varrho$ the fixed probability of a false negative given that an individual is infectious at the time of testing, and assume the probability of a positive test when an individual is not infectious is 0 (in reality, the relationship between test result and time of infection is more complicated than this, but will not be explored in this model). Therefore, after each test result, an individual's probability of still being infectious is reduced by a factor of $\varrho$.

The probability of an infected individual being detected is directly dependent on if they are infectious before or after the test is taken. This means that we can no longer directly combine the distributions for an individual's start of infectious phase ($T_i$) and end of infectious phase ($T_s$) to simply calculate the probability that they are infectious at a particular time, as the value now depends if they were detectable for each prior test. Instead we need to think of these distributions separately.

The distribution for the start of the infectious period (and therefore the detectable period) is a convolution of the distribution for the time the individual was infected and their latent period. We start by assuming that we know the exact shift that the individual was infected in, shift $z$, which starts at time $z$ and ends at time $z + a$. As stated earlier, the distribution for the time of infection of the individual would then be Uniform between $z$ and $z + a$:

$$\mathbb{P}\left(T_e = t | z \leq T_e \leq z + a\right) \propto \begin{cases} \frac{1}{a} & z \leq t \leq z + a \\ 0 & \text{otherwise} \end{cases}$$

In turn, the distribution for their time of infectious period onset would be a convolution of this Uniform distribution and the Gamma distribution that describes their latent period, $\mathit{Gamma}\left(\alpha_a, \beta_a\right)$, as a convolution of two probability density functions represents the probability density function of their sum. The PDF of a Uniform distribution is constant in the range of its possible values ($\frac{1}{a}$ in the case of a Uniform distribution between $z$ and $z + a$). Therefore, with careful consideration, this convolution is relatively straightforward. Theoretically, in order for an individual's infectious period to start at $t$ given that they were infected on day $z$, they would have to have a latent period of a length somewhere between $t - z - a$ and $t - z$. However, for certain values of $t$, this would result in a negative latent period. This is not possible and is indeed a constraint of the Gamma distribution, whose values in the form we are using must be greater or equal to 0. For this reason, we must constrain our convolution to only allow for lengths of the latent period greater than 0:

$$\begin{aligned}\mathbb{P}\left(T_i = t | z \leq T_e \leq z + a\right) &= \int_{\text{Max}[0,t-z-a]}^{\text{Max}[0,t-z]} \frac{1}{a} \times \frac{\beta_a^{\alpha_a}}{\Gamma\left(\alpha_a\right)} x^{\alpha_a - 1} \exp\left[-\beta_a x\right] \mathrm{d}x \\ &= \frac{\Gamma\left(\alpha_a, \beta_a \text{Max}\left[0, t - z - a\right]\right) - \Gamma\left(\alpha_a, \beta_a \text{Max}\left[0, t - z\right]\right)}{a\Gamma\left(\alpha_a\right)}\end{aligned}$$

If $t$ is less than $z$ then this probability is equal to 0.

An individual has a $\frac{1}{c}$ chance of being infected on any one day in a single rota cycle,

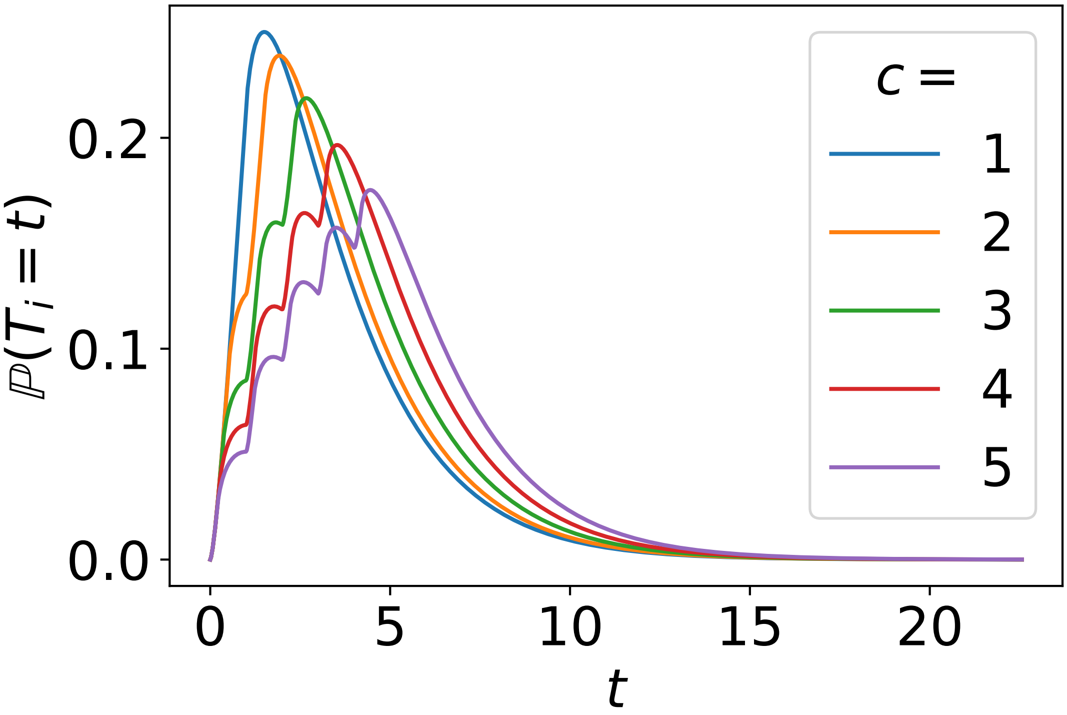


Figure 3.3. Probability density functions for the onset of an individual's infectious period who is infected in a rota pattern starting at time $t = 0$. Each line represents week long rota where individuals work a total of 24 hours over $c$ consecutive days.

meaning the complete pdf for the distribution of $T_i$ is the sum of the probability for every work day in the rota multiplied by $\frac{1}{c}$:

$$\mathbb{P}\left(T_e = t\right) = \sum_{z=0}^{c-1} \frac{1}{c} \times \mathbb{P}\left(T_e = t | z \leq T_e \leq z + a\right)$$

We have already discussed our distribution for $T_s - T_i$, the length of time an individual is infectious for, but as a reminder, this is dictated by a combination of two Gamma distributions, $Gamma\left(\alpha_{b1}, \beta_{b1}\right)$ and $Gamma\left(\alpha_{b2}, \beta_{b2}\right)$ which describe distributions for an individual's total infectious and prodromal periods respectively, and $\rho$, the probability that the individual will remain asymptomatic and never remove themselves from work. Without testing, given a known time of infectious period onset, we can write the probability that an individual is still infectious at time $t$:

$$\begin{aligned}\kappa_2\left(t, T_i\right) =& \rho \frac{\Gamma\left(\alpha_{b1}, \beta_{b1}\left(t - T_i\right)\right)}{\Gamma\left(\alpha_{b1}\right)} \\ &+ \left(1 - \rho\right) \frac{\Gamma\left(\alpha_{b2}, \beta_{b2}\left(t - T_i\right)\right)}{\Gamma\left(\alpha_{b2}\right)}\end{aligned}$$

$\kappa_2\left(t, T_i\right)$ is slightly different from $\kappa_1\left(t, T_e\right)$ from earlier in that it calculates the probability of still being infectious at time $t$ given a known time of infectious period onset, rather than a known time of infection. We can then include a function

$G(x, y, T_t, T_d, c+d)$ which returns the total number of tests that would occur after time $x$ whose results would be back before time $y$ given that testing occurs at time $T_t$ on the rota and the rota is of length $c + d$. The probability of still being infected by time $y$ given the individual became infectious and detectable at time $x$ would be reduced by the false negative rate $\varrho$ to the power of this value:

$$\mathbb{P}\left(\text{Infectious at time } t | T_i, T_t, T_d, c+d, \varrho\right) = \kappa_2(t, T_i) \times \varrho^{G(T_i, t, T_t, T_d, c+d)}$$

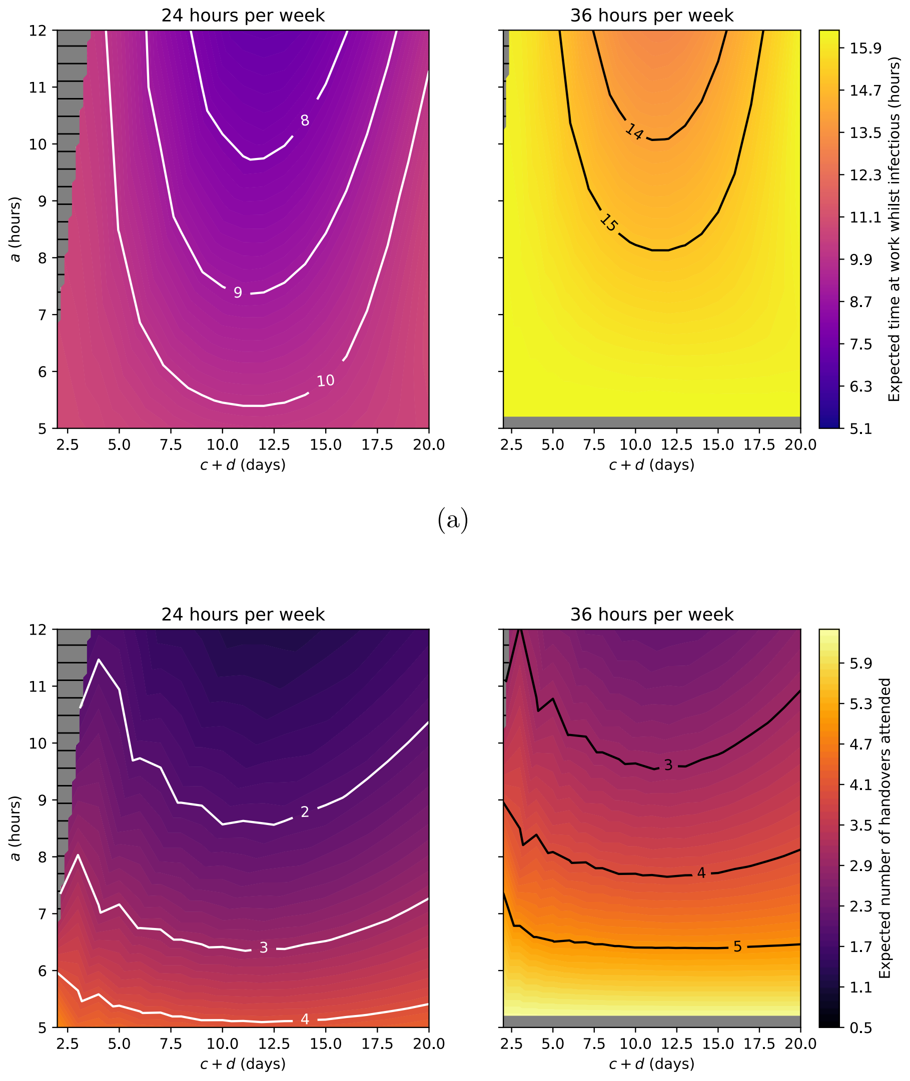
As we now have a distribution for the time of infectious period onset and a function for the probability that an individual is still infectious at a particular time given that we know their time of infectious period onset, we are now ready to numerically integrate the total probability that an infected individual is infectious at a particular time:

$$\begin{aligned}
\mathbb{P}\left(\text{Infectious at time } t\right) &= \int_0^t \mathbb{P}\left(T_i = x\right) \times \kappa_2(t, x) q^{G(x, t, T_t, T_d, c+d)} \mathrm{d}x \\
\mathbb{E}\left[T_w | T_t, T_d, \varrho\right] &= \int_0^\infty s(x) \times \mathbb{P}\left(\text{Infectious at time } x\right) \mathrm{d}x \\
\mathbb{E}\left[H | T_t, T_d, \varrho\right] &= \sum \mathbb{P}\left(\text{Infectious at time } \mathbf{h}_0\right)
\end{aligned}$$

where $h_0$ is the set of all handovers after time $t = 0$ inclusive.

### 3.3.2 Results

The gross effects changing rota patterns can have on the length of time an individual is infectious at work and the number of handovers they attend can be seen in Figures 3.4a and 3.4b. In general, for both metrics, increasing the length of individual shifts ($a$) whilst keeping the rota pattern the same length sees a positive outcome (in both cases, a lower metric, be it less time at work whilst infectious or fewer handovers attended whilst infectious, is seen as a positive outcome). There appears to be an optimum rota pattern length at approximately 11 days during which both metrics are minimised. With regards to handovers attended, this optimum is far more noticeable with longer individual shifts. A final comment on the effect of rota patterns is to note that with regards to the handovers metric, unlike the equivalent $T_w$ plots, the contours are not smooth, rising and falling as we increase the length of the rota pattern $c + d$. It would be unclear from this analysis alone if this is a true signal from the data set, a side-effect of numerical integration introducing errors into our calculations, or a result of us attempting to interpolate continuous data despite only being able to test discrete models ($c$ and $d$ must both be integers in out model). This will be explored further in the following chapter, when we investigate a way to generate a model with continuous rota parameters.

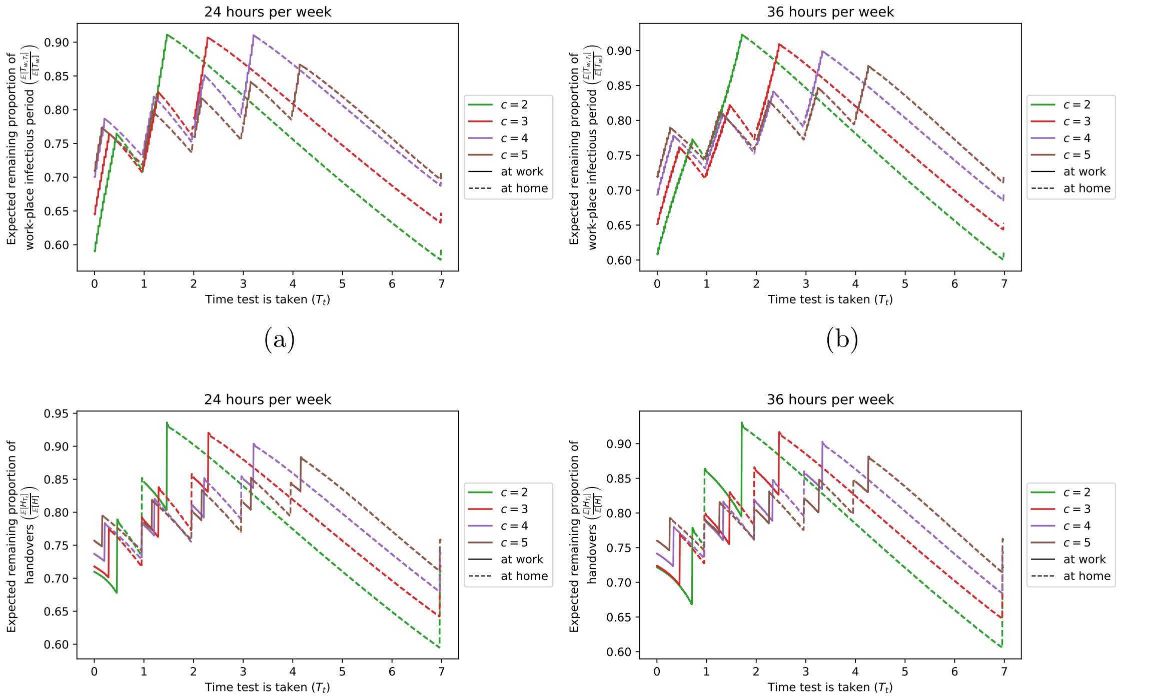


(a)

(b)

Figure 3.4. Contour plots of the expected in-work infectious time $T_w$ (Figure 3.4a) and number of handovers attended whilst infectious $H$ (Figure 3.4b) given a particular rota pattern. In each case, the x-axis gives the length of the rota pattern cycle $(c + d)$ and the y-axis gives the length of an individual shift. The left-hand plots represent rotas that result in an individual working an average of 24 hours a week and the right-hand plots come from rota patterns that result in an average of 36 hours per week. As an individual could only work an integer number of days which must be less that or equal to the total length of the rota pattern, for certain rota lengths there were upper and lower limits of $a$ outside of which an appropriate rota pattern was not possible. These regions have been greyed out in the contour plots.

Figure 3.5 demonstrates the effect of regularly timed testing on the total length of time at work whilst infectious and total number of handovers attended whilst infectious. In the case of in-work infectious time, there is a clear optimum timing of tests such that their results come in at the start of the rota pattern. Similarly, the worst time to arrange testing would be such that results come back at the end of the last shift in the rota, with a less than 10% reduction in time at work in some cases. Testing to reduce the number of handovers attended whilst infectious has similar optimal and worst timings with two differences. Firstly, as this is a discrete count rather than a smooth continuous measure of time, testing at a time such that results will come back immediately after a handover results in a sharp spike in the number of han-



(a) (b)

(c) (d)

Figure 3.5. The effect the timing of regular tests can have on the expected total length of time whilst at work (Figures 3.5a and 3.5b) and the expected number of handovers attended whilst infectious (Figures 3.5c and 3.5d). In this case we have investigated a 7 day rota where an individual works $c$ days in a row and then has $7 - c$ days off. We show the times periods when at work as solid lines and the time periods between shifts or in the rest days. The y-axis shows the expected measure proportional to the expected measure if no interventions were performed. So the top graphs show the expected length of time at work whilst infectious given a certain rota pattern (determined by the value of $c$) if regular testing was performed, divided by the expected value of $T_w$ for the same rota pattern if no testing was performed. In this case, the probability of a false negative is 0.5 and the test takes one hour to return.

dovers attended while infectious. As a result, the second difference is that each shift results in two rises in the metric with declines between spikes rather than one continuous rise whilst at work and declines whilst not at work. Each spike represents a handover missed by not testing in time. Again, testing at the wrong time can result in only a 10% reduction in this metric.

Figure 3.6 demonstrates the effect of changing the delay between test and results, $T_d$ on the length of time an individual will be infectious whilst at work. In this case, whilst we keep the rota pattern fixed as a 9-5 working week, we vary the value to $T_d$. The x-axis shows the timing of test-result rather than test-taking ($T_t + T_d$ rather than simply $T_t$). Unsurprisingly, a longer delay results in a worse outcome, as it would take longer to identify a positive individual. Additionally, we can see more clearly that it is the timing of the test result rather that the timing of the test which dictates the optimum time to test, with the result coming back at the start of the week resulting consistently in the optimum test timing.

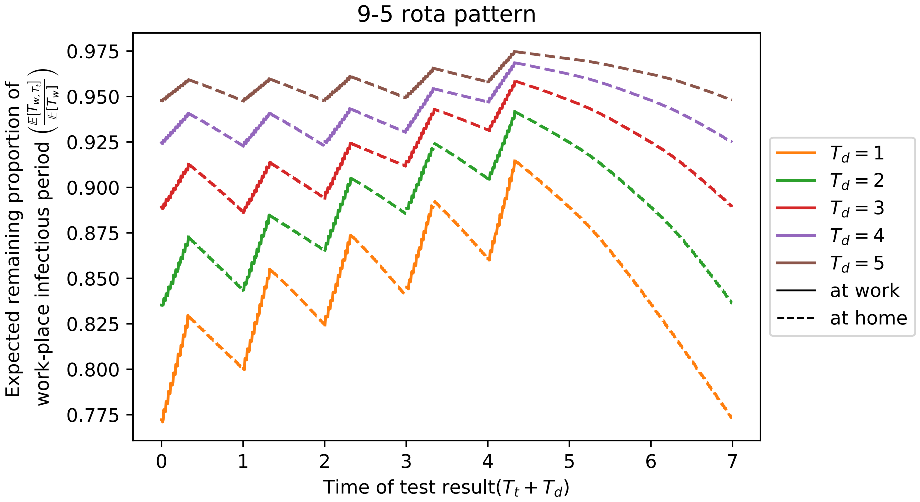


Figure 3.6. The effect the delay between test and result, $T_d$, has on the total length of time spent at work whilst infectious $T_w$. In this model, we look at a 9-5 rota and line up our testing patterns so that the x-axis shows the timing of test results rather that when the tests were taken.

## 3.4 Representing rota pattern effects through Fourier transforms

In the previous sections, we estimated the length of time whilst infectious at work through thenumerical convolution of multiple different functions, including a probability density function for the time of transmission, the cumulative density function of the latent period of the disease and the survival function for the infectious period of the disease. A numerical solution, whilst useful, leaves questions as to its accuracy. Additionally, the more complicated the rota pattern structure, the more computationally expensive the model would become and the greater level of accuracy that would be required. In this chapter, we aim to approach a semi-analytical solution for this problem through the use of Fourier transforms.

Fourier transforms are one of many ways in which a function can be converted. If we can express a function as a sum of multiple sine and cosine waves, then, for any given frequency, the function's Fourier transform gives the required waves amplitudes. A Fourier transform is often described as taking a function in the time-domain and finding its corresponding expression in the frequency-domain. They appear in multiple aspects and mathematics, physics and engineering. One of their advantages is that they can convert a digital signal into an analogue one, as a digital signal can be expressed as a sum of an infinite number of sine and cosine waves (see Figure 3.7).

This is useful for us as we can express rota patterns as digital signals: either an individual is at work or they are not. Expressing this mathematically without a Fourier transform can be tricky and resulted in inaccuracies in edge cases (approaching the time when an individual would switching between work and not).

Previously, the numerical convolution of multiple Uniform distributions (representing the times when an individual could get infected at work, based on the rota pattern)

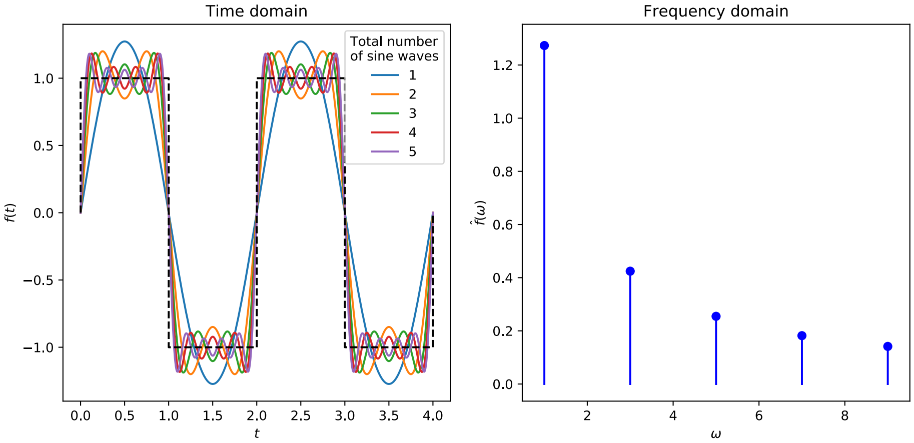


Figure 3.7. A demonstration of a square wave $f(t)$ and its equivalent Fourier transform $\hat{f}(\omega)$. A periodic function like $f(t)$ can be expressed as the sum of an infinite number of sine and cosine waves with discrete frequencies. In the case of a square wave with a frequency of $\frac{1}{2}$ and an amplitude of 1, $f(t) = \sum_{k=0}^{\infty} \frac{4}{(2k+1)\pi} \sin\left[(2k+1)\pi t\right]$.

with two Gamma distributions (representing the latent period and the time from starting being infectious to becoming symptomatic) became cumbersome and computationally expensive. We chose a relatively simple rota pattern to analyse, but a hospital rota may be as many as 12 weeks long with different shift patterns on each week. Including such a pattern with the current method of analysis would be impractical. If we consider the digital expression of rota patterns as a sum of an infinite number of waves rather than a number of discrete shifts, then we may be able to find a semi-analytical solution for our model for each discrete frequency and then sum over all frequencies.

Our model looks at the entire infectious career of one individual. We are going to define three functions and calculate their equivalent frequency domain function. The first is $R(t)$, a periodic function representing the relative risk of transmission at any one time in a rota pattern of period length $T$. Any $T$ length section of $R(t)$ represents the probability density function (pdf) for the time of transmission for an individual given that they are infected in this period. The function $\lambda(\tau)$ gives the infectiousness of the infected individual in our model at time $\tau$ since the time of transmission, i.e. given that they were infected at time $\tau = 0$. In the real world, this function would likely vary from person to person. It must be real and non-negative for $\tau > 0$, but otherwise be equal to 0. Finally, as with the numerical model, $s(t)$ is an indicator function showing if an individual is at work at time $t$, where they would be able to infect a colleague.

In this section, we aim to be careful to distinguish the terms transmission and infection. When discussing an individual's infectious career, their time of transmission is when they are initially infected, whereas their infectious window is when they can infect others.

We can see some similarities between $s(t)$ and $R(t)$, in that they both rely on the rota pattern for their structure. However, unlike $R(t)$ which varies dependent on the chosen risk of transmission inside and outside of work, $s(t)$ always returns 1 if the individual is at work at time $t$ and 0 otherwise. We will use this function to indicate when our infected individual is at work whilst infectious. Making use of the similarities between $R(t)$ and $s(t)$ will be key to the analytical solution for our model.

### 3.4.1 Model for time of infection with respect to rota pattern

We will need to calculate $\hat{R}(\omega)$, the Fourier transform for $R(t)$. However, first let us start by defining $R(t)$ in the time domain. Simply put, $R(t)$ can be one of two values:

$$R(t) = \begin{cases} \frac{u}{A} & \text{if at work} \\ \frac{v}{A} & \text{if not at work} \end{cases}$$

We define $u$ and $v$ as the relative risks of infections inside and outside of work respectively and $A = u\bar{t} + v(T - \bar{t})$ where $\bar{t}$ is the length of time worked in one rota period. The area under $R(t)$ for any period of length $T$ would therefore be 1, and the average value across a period would be $\frac{1}{T}$. Note that, unlike in the previous chapter, where individuals were assumed to be at risk of infection only at work, here we allow them to be infected also when not working (unless $v = 0$).

For numerical reasons, it is easier for us to design any expressions of our repeating rota pattern such that time 0 represents the start of an individual's first shift in the rota. However, it is worthwhile clarifying at this point that this is not related to the time 0 in the $\lambda(\tau)$ infectiousness function. Functions given in terms of $t$, such as $R(t)$ and $s(t)$ will be bound to time relative to the rota pattern (calendar time) such that a new rota pattern starts at time $t = 0$. Functions given in terms of $\tau$ will be bound to time relative to the time at which transmission occurs for the individual in our model (i.e. they are always infected at time $\tau = 0$, but this could be at any value of $t$).

We define two vectors, $\mathbf{m}$ and $\mathbf{n}$, which denote the beginning and end of each shift respectively. As the first shift starts at time 0, $m_1 = 0$. These two vectors are of length $c + 1$ and $c$ respectively. The reason why $\mathbf{m}$ is of length $c + 1$ is because its final value represents the start of a new rota pattern. This is a slightly more careful definition that $\mathbf{h}$, our method of describing shifts in the previous case. In that case, $\mathbf{h}$ represented both beginnings and ends of shifts, and it represented all possible shifts in an infinite time range.

For example, in the case of a 7 day, 9-to-5 shift, we define $\mathbf{m}$ and $\mathbf{n}$ as:

$$\mathbf{m} = [0, 1, 2, 3, 4, 7]$$
$$\mathbf{n} = \left[\frac{1}{3}, 1\frac{1}{3}, 2\frac{1}{3}, 3\frac{1}{3}, 4\frac{1}{3}\right]$$

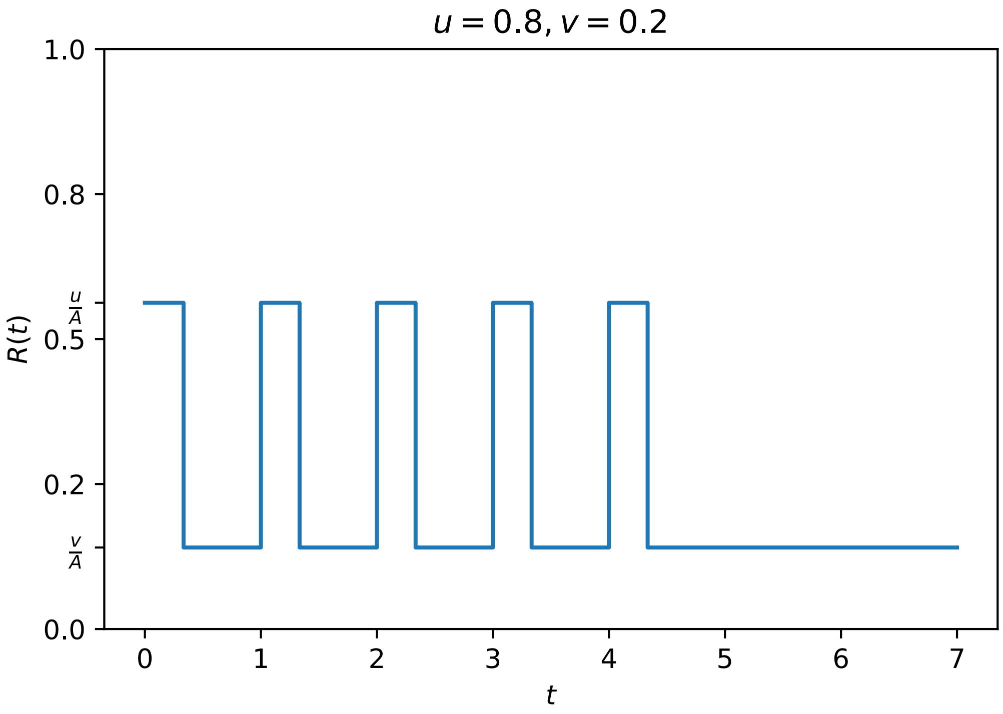


Figure 3.8. Demonstration of one period of $R(t)$, the repeating probability density function for an individual's time on infection. In this case the ratio of risks of being infected inside or outside of work is 0.8:0.2, with a weekly rota of 5 days of 9 to 5 shift pattern followed by 2 days of rest, starting from the beginning of the individual's first shift.

We have described this rota pattern such that it starts at the beginning of the first shift in the pattern, hence $m_1 = 0$, which in a normal 9-5 rota pattern would be equivalent to 9 O'clock on a Monday morning. As the pattern is 7 days long, and the final value of $\mathbf{m}$ represents the start of the new shift, $m_6 = 7$. Since a 9-5 rota pattern involves a shift length of 8 hours, which is a third of a day, each value of $\mathbf{n}$, representing the end of shifts, is $\frac{1}{3}$ greater than a start of a shift found in $\mathbf{m}$.

We now have enough parameters to fully describe $R(t)$. $T$ describes the total length of the rota pattern, $\mathbf{m}$ and $\mathbf{n}$ denote if an individual would be at work relative to a starting point $t = 0$ and $u$ and $v$ describe the relative risk of infection inside and outside of work. In turn, we can now explicitly calculate $\hat{R}(\omega)$, the Fourier transform of $R(t)$.

As a periodic function, the Fourier transform of $R(t)$ can be described as a series of Dirac delta functions at the harmonic frequencies of $R(t)$ (i.e. the frequencies $\frac{2\pi j}{T}$ where $j$ is a real integer). We can write $\hat{R}(\omega)$ as:

$$\hat{R}(\omega) = 2\pi \sum_{j=-\infty}^{\infty} C_{R,j} \delta \left( \omega - \frac{2\pi j}{T} \right)$$

where $\delta(x)$ is a Dirac delta function and $C_{R,j}$ is a constant specific to the function

$R(t)$ and the integer $n$. We calculate $C_{R,j}$ for all non-zero integer values of $j$:

$$\begin{aligned} C_{R,j} &= \frac{1}{T}\int_0^T R(t)\,\exp\left[-\frac{2\pi \mathrm{i} jt}{T}\right]\mathrm{d}t \\ &= \frac{1}{T}\sum_{k=1}^{c}\left(\frac{u}{A}\int_{m_k}^{n_k}\exp\left[-\frac{2\pi \mathrm{i} jt}{T}\right]\mathrm{d}t + \frac{v}{A}\int_{n_k}^{m_{k+1}}\exp\left[-\frac{2\pi \mathrm{i} jt}{T}\right]\mathrm{d}t\right) \end{aligned}$$

where the integral between $m_k$ and $n_k$ represents the time period during the $k$the day of work, and the integral between $n_k$ and $m_{k+1}$ represents the rest period between the $k$th day of work and the $k+1$th day of work. We are summing across every work day, so the final rest period will be between the end of the last shift, $n_c$, and the beginning of the next rota cycle, $m_{c+1} = T$. If we solve these integrals we find:

$$C_{R,j} = \frac{1}{T}\sum_{k=1}^{c}\left(\frac{uiT}{2\pi jA}\left(\exp\left[-\frac{2\pi ijn_k}{T}\right] - \exp\left[-\frac{2\pi ijm_k}{T}\right]\right) + \frac{viT}{2\pi jA}\left(\exp\left[-\frac{2\pi ijm_{k+1}}{T}\right] - \exp\right.\right.$$

The solutions to both integrals each involve some variant of the term $\exp\left[-\frac{2\pi ijn_k}{T}\right]$ so we can see how these parts can be combined:

$$\frac{uiT}{2\pi jA}\exp\left[-\frac{2\pi ijn_k}{T}\right] - \frac{viT}{2\pi jA}\exp\left[-\frac{2\pi ijn_k}{T}\right] = \frac{iT(u-v)}{2\pi jA}\exp\left[-\frac{2\pi ijn_k}{T}\right]$$

The solution to the "work-time" integral involves an $\exp\left[-\frac{2\pi ijm_{\mathbf{k}}}{T}\right]$, while the solution to the "rest-period" integral involves an $\exp\left[-\frac{2\pi ijm_{\mathbf{k+1}}}{T}\right]$ term (the bold font has been used to highlight their differences). It may seem like these two solutions cannot be easily combined. However, $m_{c+1} = T$, meaning $\exp\left[-\frac{2\pi ijm_{\mathbf{c+1}}}{T}\right] = 1 = \exp\left[-\frac{2\pi ijm_{\mathbf{1}}}{T}\right]$. This means that the sum $\sum_{k=1}^{c}\exp\left[-\frac{2\pi ijm_{k+1}}{T}\right]$ is equivalent to the sum $\sum_{k=1}^{c}\exp\left[-\frac{2\pi ijm_k}{T}\right]$ and we can combine the **m** values from the two integral solutions much in the same way we combined the **n** values in the same solution. This gives us:

$$\begin{aligned} C_{R,j} &= \frac{1}{T}\sum_{k=1}^{c}\frac{iT(u-v)}{2\pi jA}\exp\left[-\frac{2\pi ijn_k}{T}\right] + \frac{iT(v-u)}{2\pi jA}\exp\left[-\frac{2\pi ijm_k}{T}\right] \\ &= \frac{\mathrm{i}(u-v)}{2\pi jA}\sum\left(\exp\left[-\frac{2\pi \mathrm{i}j\mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi \mathrm{i}j\mathbf{m}}{T}\right]\right) \end{aligned}$$

For brevity, we use $\sum$ on its own to represent summing over every pair of shift starts and ends, **m** and **n**, ignoring $m_{c+1}$. $C_{R,0}$ is the average value of $R(t)$ over a period which we already know is $\frac{1}{T}$. Therefore, the complete expression for $\hat{R}(\omega)$ becomes:

$$\hat{R}(\omega) = \delta(\omega)\frac{1}{T} + \sum_{j=-\infty, j\neq 0}^{\infty} \delta\left(\omega - \frac{2\pi j}{T}\right) \\ \times \frac{\mathrm{i}(u-v)}{jA} \sum \left(\exp\left[-\frac{2\pi \mathrm{i} j\mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi \mathrm{i} j\mathbf{m}}{T}\right]\right) \tag{3.3}$$

### 3.4.2 Model for infectiousness at any point during a rota pattern

$\lambda(\tau)$, describing the infectiousness from an individual at time $\tau$ given that they were infected at time 0, is unique to the individual and as such it can never be known precisely. However, it must satisfy the following constraints:

1. $\lambda(\tau) \geq 0$ if $t \geq 0$ - The force of infection must be non-negative.
2. $\lambda(\tau) = 0$ if $t < 0$ - The individual cannot generate a force of infection prior to infection.
3. $0 \leq \int_0^\infty \lambda(\tau)\mathrm{d}t < \infty$ - The individual generates a finite force of infection over their infectious career.
4. $\Im\left[\lambda(\tau)\right] = 0$ - For all values of $t$, the force of infection must take a real value (including 0).

Given the first statement, the final statement must be true. However, it is worthwhile stressing that $\lambda(\tau)$ is real, as it will have implications when we consider its Fourier transform.

As $\lambda(\tau)$ has a finite integral, its Fourier transform, $\hat{\lambda}(\omega)$ must exist. As $\lambda(\tau)$ is real, $\Re\left[\hat{\lambda}(\omega)\right]$, the real part of $\hat{\lambda}(\omega)$, is an even function (meaning $\Re\left[\hat{\lambda}(\omega)\right] = \Re\left[\hat{\lambda}(-\omega)\right]$, or $\Re\left[\hat{\lambda}(\omega)\right]$ has reflective symmetry over the line $\omega = 0$. Similarly, $\Im\left[\hat{\lambda}(\omega)\right]$ is an odd function, in that it has rotational symmetry about the origin and $\Im\left[\hat{\lambda}(\omega)\right] = -\Im\left[\hat{\lambda}(-\omega)\right]$. This means that in total, $\hat{\lambda}(\omega)$ is the conjugate of $\hat{\lambda}(-\omega)$. This is not strictly useful information now, but will be very relevant later.

We want to convolve $R(t)$ and $\lambda(\tau)$ to get $\Lambda(t)$, a periodic function that describes the expected total force of infection that an infected individual will generate at each part of the rota pattern given that they have been infected. Depending on how long they are infectious for and how short the rota pattern is, they may provide a force of infection to a particular part of the rota more than once during their infectious career. This was why it was advantageous to describe $R(t)$ in the form of a periodic PDF, as we can calculate the force of infection given that the individual was infected multiple rota periods ago.

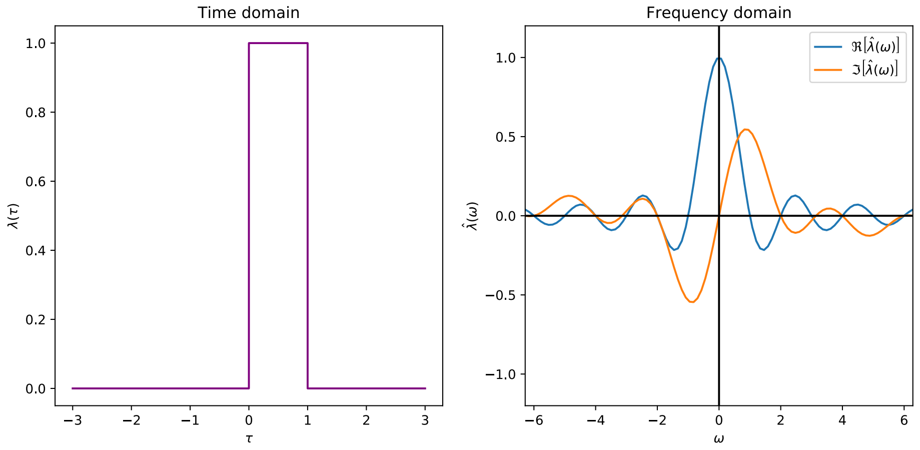


Figure 3.9. A function with only real values and its Fourier transform. In this case we are looking at a square pulse function between 0 and 1. We can see that the real part of its Fourier transform (blue) is reflected on the line $\omega = 0$ while its imaginary part (orange) has rotational symmetry about the origin.

This convolution is not necessarily obvious, especially as $\lambda(\tau)$ is currently an unknown function. Even if we did know $\lambda(\tau)$, its convolution with $R(t)$ may not be analytically possible. However, for functions with calculable Fourier transform, convolution in the time-domain is equivalent to multiplication in the frequency domain. That is to say, while we do not currently know the form $\hat{\lambda}(\omega)$ will take, we do know how it can be used to calculate $\hat{\Lambda}(\omega)$, the Fourier transform of $\lambda(\tau)$:

$$\hat{\Lambda}(\omega) = \hat{\lambda}(\omega) \times \left(\delta(\omega)\frac{1}{T} + \sum_{j=-\infty, j\neq 0}^{\infty} \delta\left(\omega - \frac{2\pi j}{T}\right) \times \frac{\mathrm{i}(u-v)}{jA} \sum \left(\exp\left[-\frac{2\pi \mathrm{i} j \mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi \mathrm{i} j \mathbf{m}}{T}\right]\right)\right) \tag{3.4}$$

### 3.4.3 Model of the expected force of infection while at work

Following the same rota pattern that defined $R(t)$, $s(t)$ is a repeating indicator function that equals 1 if $t$ is a point in the rota pattern where an individual would be at work, and 0 if not. By multiply $\Lambda(t)$ by $s(t)$ we would get a function that describes the expected force of infection an individual would generate at time $t$ at work (they obviously generate a force of infection also when not at work, but quantifying that element is not the focus of this project).

Setting $A = 1$, $u = 1$ and $v = 0$ in Equation (3.3) for the Fourier transform for $R(t)$ we could directly get the the Fourier transform for $s(t)$:

$$\hat{s}(\omega) = \delta(\omega)\frac{\bar{t}}{T} + \sum_{j=-\infty, j\neq 0}^{\infty} \delta\left(\omega - \frac{2\pi j}{T}\right) \times \frac{\mathrm{i}}{j} \sum \left(\exp\left[-\frac{2\pi \mathrm{i} j \mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi \mathrm{i} j \mathbf{m}}{T}\right]\right). \tag{3.5}$$

We want to set $A = 1$ because previously we used $A$ to normalise $R(t)$ so that it integrated to 1 over any given time period of length $T$, but in this case we do not want to normalise our indicator function.

We could define $W(t)$, the function that results from the multiplication of $\Lambda(t)$ and $s(t)$. As the period of the two is the same, this would be a periodic function showing the force of infection an infected individual would be expected to exert whilst at work. However, the convolution theorem states not only that convolution in the time-domain is equivalent to multiplication in the frequency-domain, but also that the opposite is true.

Therefore, instead of being interested in $W(t)$, we can focus on its Fourier transform $\hat{W}(\omega)$, which is a convolution of $\hat{\Lambda}(\omega)$ and $\hat{s}(\omega)$. More specifically, we can focus on $\hat{W}(0)$, which is total area under the curve of $W(t)$ in one period and therefore the total expected force of infection whilst at work.

As $\hat{\Lambda}(\omega)$ is a series of Dirac delta functions, its convolution with $\hat{s}(\omega)$ can be written in terms of an infinite sum:

$$\hat{W}(0) = \hat{\Lambda}(0)\hat{s}(0) + \frac{1}{2\pi} \sum_{j=-\infty, j\neq 0}^{\infty} \hat{\Lambda}\left(\frac{2\pi j}{T}\right) \times \hat{s}\left(-\frac{2\pi j}{T}\right)$$

The two Fourier transforms were each multiplied by $2\pi$ to normalise them. When we combine the to functions, we have to include the $\frac{1}{2\pi}$ term because otherwise we will have normalised this term twice.

If we then look at the similarities between $\hat{\Lambda}(\omega)$ (Equation 3.4.2) and $\hat{s}(\omega)$ (Equation 3.5), we notice that they both contain the term:

$$\sum \left(\exp\left[-\frac{2\pi \mathrm{i}j\mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi \mathrm{i}j\mathbf{m}}{T}\right]\right)$$

When we convolve $\hat{\Lambda}(\omega)$ with $\hat{s}(\omega)$ to find $\hat{W}(0)$, we will be performing the multiplication $\hat{\Lambda}(\omega) \times \hat{s}(-\omega)$ for all possible values of $\omega$. It would therefore be useful to define the function $Y(j)$:

$$\begin{aligned} Y(j) = &\left(\sum \exp\left[-\frac{2\pi \mathrm{i} j \mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi \mathrm{i} j \mathbf{m}}{T}\right]\right) \\ &\times \left(\sum \exp\left[\frac{2\pi \mathrm{i} j \mathbf{n}}{T}\right] - \exp\left[\frac{2\pi \mathrm{i} j \mathbf{m}}{T}\right]\right) \end{aligned}$$

$$\begin{aligned} &= \Re\left[\sum \exp\left[-\frac{2\pi \mathrm{i} j\mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi \mathrm{i} j\mathbf{m}}{T}\right]\right]^2 \\ &+ \Im\left[\sum \exp\left[-\frac{2\pi \mathrm{i} j\mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi \mathrm{i} j\mathbf{m}}{T}\right]\right]^2 // \end{aligned}$$

$\sum \exp\left[-\frac{2\pi \mathrm{i} j\mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi \mathrm{i} j\mathbf{m}}{T}\right]$ is the conjugate of $\sum \exp\left[\frac{2\pi \mathrm{i} j\mathbf{n}}{T}\right] - \exp\left[\frac{2\pi \mathrm{i} j\mathbf{m}}{T}\right]$, meaning multiplying these two terms together will give us a real value which is the square of their real part plus the square of their imaginary part. By remembering that the real part of $\exp[-i\theta]$, $\Re[\exp[-i\theta]]$, is equal to= $\cos[\theta]$ and that the imaginary part, $\Im[-i\theta]$ is equal to $-\sin[\theta]$ we can convert $Y(j)$ into a trigonometric function:

$$\begin{aligned} Y(j) = &\left(\sum \cos\left[\frac{2\pi j\mathbf{n}}{T}\right] - \cos\left[\frac{2\pi j\mathbf{m}}{T}\right]\right)^2 \\ &+ \left(\sum \sin\left[\frac{2\pi j\mathbf{m}}{T}\right] - \sin\left[\frac{2\pi j\mathbf{n}}{T}\right]\right)^2 \end{aligned}$$

Finally, we can adjust $Y(j)$ further by using the two trigonometric identities describing the difference between two Sine functions and two Cosine functions:

$$\begin{aligned} \sin[\alpha] - \sin[\beta] &= 2\cos\left[\frac{\alpha+\beta}{2}\right]\sin\left[\frac{\alpha-\beta}{2}\right] \\ \cos[\alpha] - \cos[\beta] &= -2\sin\left[\frac{\alpha+\beta}{2}\right]\sin\left[\frac{\alpha-\beta}{2}\right] \\ Y(j) &= 4\left(\left(\sum \sin\left[\frac{\pi j(\mathbf{n}+\mathbf{m})}{T}\right]\sin\left[\frac{\pi j(\mathbf{n}-\mathbf{m})}{T}\right]\right)^2 \right. \\ &\qquad \left. + \left(\sum \cos\left[\frac{\pi j(\mathbf{n}+\mathbf{m})}{T}\right]\sin\left[\frac{\pi j(\mathbf{n}-\mathbf{m})}{T}\right]\right)^2\right) \end{aligned}$$

When we insert $Y(j)$ into our formula for $\hat{W}(0)$, we are replacing the $\sum\left(\exp\left[-\frac{2\pi i j\mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi i j\mathbf{m}}{T}\right]\right)$ terms from both $\hat{\Lambda}\left(\frac{2\pi j}{T}\right)$ and $\hat{s}\left(-\frac{2\pi j}{T}\right)$. This gives us:

$$\begin{aligned} \hat{W}(0) &= \hat{\Lambda}(0)\hat{s}(0) + \frac{1}{2\pi}\sum_{j=-\infty, j\neq 0}^{\infty} \hat{\Lambda}\left(\frac{2\pi j}{T}\right) \times \hat{s}\left(-\frac{2\pi j}{T}\right) \\ &= \hat{\Lambda}(0)\hat{s}(0) \\ &\quad + \frac{1}{2\pi}\sum_{j=-\infty, j\neq 0}^{\infty} \hat{\lambda}\left(\frac{2\pi j}{T}\right)\frac{i(u-v)}{jA}\sum\left(\exp\left[-\frac{2\pi i j\mathbf{n}}{T}\right] - \exp\left[-\frac{2\pi i j\mathbf{m}}{T}\right]\right) \\ &\qquad\qquad \times -\frac{i}{j}\sum\left(\exp\left[\frac{2\pi i j\mathbf{n}}{T}\right] - \exp\left[\frac{2\pi i j\mathbf{m}}{T}\right]\right) \\ &= \hat{\Lambda}(0)\hat{s}(0) \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2\pi}\sum_{j=-\infty,j\neq 0}^{\infty}\hat{\lambda}\left(\frac{2\pi j}{T}\right)\frac{i(u-v)}{jA}\times -\frac{i}{j}\times Y(j)\\
&=\hat{\Lambda}(0)\hat{s}(0)\\
&+\frac{1}{2\pi}\sum_{j=-\infty,j\neq 0}^{\infty}\frac{(u-v)}{Aj^2}Y(j)\hat{\lambda}\left(\frac{2\pi j}{T}\right)
\end{aligned}
$$

The function $Y(j)$ is symmetrical, in that $Y(j) = Y(-j)$. Conversely, $\hat{\lambda}\left(\frac{2\pi j}{T}\right)$ is the conjugate of $\hat{\lambda}\left(-\frac{2\pi j}{T}\right)$, meaning:

$$
\hat{\lambda}\left(\frac{2\pi j}{T}\right)+\hat{\lambda}\left(-\frac{2\pi j}{T}\right)=2\Re\left[\hat{\lambda}\left(\frac{2\pi j}{T}\right)\right]
$$

In turn, we can calculate the sum of the $j$th term and $-j$th term in our infinite sum:

$$
\begin{aligned}
\frac{u-v}{Aj^2}Y(j)\hat{\lambda}\left(\frac{2\pi j}{T}\right)+\frac{u-v}{A(-j)^2}Y(-j)\hat{\lambda}\left(-\frac{2\pi j}{T}\right)&=\frac{u-v}{Aj^2}Y(j)\left(\hat{\lambda}\left(\frac{2\pi j}{T}\right)+\hat{\lambda}\left(-\frac{2\pi j}{T}\right)\right)\\
&=\frac{2(u-v)}{Aj^2}Y(j)\Re\left[\hat{\lambda}\left(\frac{2\pi j}{T}\right)\right]
\end{aligned}
$$

This means we can simplify the infinite sum in our term for $\hat{W}(0)$ to only require positive values of $j$:

$$
\hat{W}(0)=\hat{\Lambda}(0)\frac{\bar{t}}{T^2}+\sum_{j=1}^{\infty}\frac{u-v}{\pi Aj^2}Y(j)\Re\left[\hat{\lambda}\left(\frac{2\pi j}{T}\right)\right]
$$

Again, by multiplying $\hat{W}(0)$ value by $T$ we get an infected individual's total expected infectiousness whilst at work for their entire infectious career. The zeroth term in a Fourier series is its expected average value across an entire rotation. Multiplying this number by the length of the period will give the total area under the curve in one cycle. Since $W(t)$ gives the total infectiousness an infected person provides at time $t$ of the cycle, $T\hat{W}(0)$ will give the infectiousness they provide in total. It is this value that we would want to minimise to reduce the rate of workplace infections.

If we wanted to look at the complete distribution for $\hat{W}(\omega)$, the mathematics become more complicated as it would not be as easy to take shortcuts such as $Y(j)$ and summing conjugates. While not the subject of this chapter, such investigations do have merit, as they could be adjusted to indicate a new pdf for the time of infection, replacing $\hat{R}(\omega)$ for a second generation of infections. Repeating this process an infinite number of times would result in the pdf (and ultimately the in-work force of infection) when the disease in endemic in the work place. This investigation only seeks to estimate this value for the first generation of staff infections.

### 3.4.4 Including a more rigid structure for a rota pattern

So far we have described the rota pattern in terms of two vectors, **m** and **n**, which describe the starts and ends of $c$ shifts starting at time 0, and $T$, the total length of the rota pattern. This was to show how this work can be applied to any rota pattern.

However, in the numerical model, we used a far stricter rota pattern. An individual work shifts of length $a$ with breaks of length $b$, meaning $a+b=1$ day. They worked $c$ shifts before having $d$ days off, resulting in a total rota pattern length of $c+d$. Whilst keeping the average number of hours worked in a week fixed, we adjusted the shift length, $a$, and the rota length, $c+d$, to see if they affected the expected length of time while infectious at work. We found that a longer shift appeared to reduce the expected length of time while infectious at work, and there was a consistent optimum length of rota.

This nomenclature enabled us to treat rota structures as mathematical entities. We could directly investigate the effect of working longer days or having a longer rota pattern whilst maintaining the same average length of time worked. The current documentation for our Fourier transform model is a lot more loose (meaning in the future we can be more flexible as to the rota patterns we analyse). However, as we want to compare the two methods (and see if they show the same conclusions), we need to introduce the stricter original design for a rota pattern into our Fourier model.

To incorporate this stricter definition of a rota pattern into our calculations above, we primarily need to know how they affect the function $Y(j)$, but there are other values, such as $\bar{t}$ and $A$, that we also need to calculate.

Let $w = \bar{t}/T$ be the proportion of time that an individual spends at work in a full rota, as $\bar{t}$ is the length of time spent at work and $T$ is the length of the rota pattern. Conversely, $a$ is the proportion of time an individual spends at work on any working day. We can start by assuming that we choose a value of $a$ and $T$ such that integer values of $c$ and $d$ can be found. This assumption can be relaxed later. For the time being, however, we can make a number of definitions from the values $a$, $c$, and $w$:

$$\begin{aligned}
a &> w \\
c &= \frac{Tw}{a} \\
\bar{t} &= Tw \\
m_k &= k-1 \\
n_k &= k+a-1 \\
m_{c+1} &= T
\end{aligned}$$

$$A = uTw + vT(1-w)$$

From these statements we can calculate the $j$th constant terms of $\hat{R}(\omega)$ and $\hat{s}(\omega)$. We could skip this step and instead calculate $Y(j)$, but there is some understanding to be gained from looking at the rota pattern functions explicitly. Let us look at $C_{R,j}$. For simplicity sake, we will choose a scenario where it is impossible to be infected outside of work. In this case, $u = 1$, $v = 0$, $A = Tw$. This gives $C_{R,j}$ the following formula:

$$\begin{aligned} C_{R,j} &= \frac{\mathrm{i}}{2\pi jTw} \sum_{k=0}^{\frac{Tw}{a}-1} \left( \exp\left[-\frac{2\pi \mathrm{i} j(k+a)}{T}\right] - \exp\left[-\frac{2\pi \mathrm{i} jk}{T}\right] \right) \\ &= \frac{\mathrm{i}}{2\pi jTw} \left( \exp\left[-\frac{2\pi \mathrm{i} ja}{T}\right] - 1 \right) \sum_{k=0}^{\frac{Tw}{a}-1} \exp\left[-\frac{2\pi \mathrm{i} jk}{T}\right] \end{aligned}$$

The nature of this finite sum changes dependent on the value of $j$. Specifically, if $\frac{j}{T}$ is an integer, then the exponential term in the sum is equal to 1 and the sum is equal to $c = \frac{Tw}{a}$, the number of shift patterns worked. This is the equivalent to the the solution to a finite geometric sum when the common ratio between terms is equal to 1. If $\frac{j}{T}$ is not an integer, then the sum becomes the sum of a geometric sequence and takes a completely different form. The complete term of $C_{R,j}$ is therefore dependent on this fraction.

$$C_{R,j} = \frac{\mathrm{i}}{2\pi jTw} \left( \exp\left[-\frac{2\pi \mathrm{i} ja}{T}\right] - 1 \right) \times \begin{cases} \frac{Tw}{a} & \frac{j}{T} \in \mathbb{Z} \\ \frac{\exp\left[-\frac{2\pi \mathrm{i} jw}{a}\right]-1}{\exp\left[-\frac{2\pi \mathrm{i} j}{T}\right]-1} & \text{otherwise} \end{cases} \tag{3.6}$$

If we look at the solution when $\frac{j}{T}$ is not an integer, we can see we would need to divide by $\exp\left[-\frac{2\pi \mathrm{i} j}{T}\right] - 1$. In the case where $\frac{j}{T}$ is an integer, this denominator would be equal to 0, making this fraction not possible. We can use the value in Equation 3.6 to get a specific trigonometric expression for $R(t)$:

$$R(t) = \frac{1}{T} + \sum_{j=1}^{\infty} \frac{2}{\pi j} \sin\left[\frac{\pi aj}{T}\right] \times \begin{cases} \frac{1}{a} \cos\left[\frac{\pi j(a-2t)}{T}\right] & \frac{j}{T} \in \mathbb{Z} \\ \frac{1}{Tw} \frac{\sin\left[\frac{\pi jw}{a}\right]}{\sin\left[\frac{\pi j}{T}\right]} \cos\left[\pi j \left(\frac{w}{a} + \frac{a-2t-1}{T}\right)\right] & \text{otherwise} \end{cases}$$

This condition to check if $\frac{j}{T}$ is an integer also applied when we are calculating $Y(j)$. This is perhaps easiest to see when it is in its exponential form, although it be no means easy. We start by substituting our new values for **m** and **n** into our formula for $Y(j)$

$$\begin{aligned}
Y(j) &= \left(\sum \exp\left[\frac{2\mathrm{i}\pi j\mathbf{n}}{T}\right] - \exp\left[\frac{2\mathrm{i}\pi j\mathbf{m}}{T}\right]\right)\left(\sum \exp\left[-\frac{2\mathrm{i}\pi j\mathbf{n}}{T}\right] - \exp\left[-\frac{2\mathrm{i}\pi j\mathbf{m}}{T}\right]\right)\\
&= \left(\sum_{k=0}^{\frac{Tw}{a}-1} \exp\left[\frac{2\mathrm{i}\pi j(k+a)}{T}\right] - \exp\left[\frac{2\mathrm{i}\pi jk}{T}\right]\right)\\
&\quad\times\left(\sum_{k=0}^{\frac{Tw}{a}-1} \exp\left[-\frac{2\mathrm{i}\pi j(k+a)}{T}\right] - \exp\left[-\frac{2\mathrm{i}\pi jk}{T}\right]\right)
\end{aligned}$$

In both sums, we can factor out a value of either $\exp\left[\frac{2\pi ijk}{T}\right]$ or $\exp\left[-\frac{2\pi ijk}{T}\right]$:

$$\begin{aligned}
Y(j) &= \left(\sum_{k=0}^{\frac{Tw}{a}-1} \exp\left[\frac{2\mathrm{i}\pi jk}{T}\right]\left(\exp\left[\frac{2\pi\mathrm{i}ja}{T}\right] - 1\right)\right)\\
&\quad\times\left(\sum_{k=0}^{\frac{Tw}{a}-1} \exp\left[-\frac{2\mathrm{i}\pi jk}{T}\right]\left(\exp\left[-\frac{2\pi\mathrm{i}ja}{T}\right] - 1\right)\right)
\end{aligned}$$

Since neither $\exp\left[\frac{2\pi ija}{T}\right]-1$ nor $exp\left[-\frac{2\pi ija}{T}\right]-1$ depend on $k$, they can be extracted from the sum terms and multiplied together:

$$\begin{aligned}
Y(j) &= \left(2 - \exp\left[\frac{2\pi\mathrm{i}ja}{T}\right] - \exp\left[-\frac{2\pi\mathrm{i}ja}{T}\right]\right)\\
&\quad\times\left(\sum_{k=0}^{\frac{TW}{a}-1} \exp\left[\frac{2\pi\mathrm{i}jk}{T}\right]\right)\left(\sum_{k=0}^{\frac{TW}{a}-1} \exp\left[-\frac{2\pi\mathrm{i}jk}{T}\right]\right)
\end{aligned}$$

There are then two steps to simplifying this term outside of the two sums. First, we recognise the relationship between the Cosine function and the exponent of $i$:

$$\begin{aligned}
\cos[x] &= \frac{\exp[\mathrm{i}x] + \exp[-\mathrm{i}x]}{2}\\
Y(j) &= 2\left(1 - \cos\left[\frac{2\pi ja}{T}\right]\right)\left(\sum_{k=0}^{\frac{TW}{a}-1} \exp\left[\frac{2\pi\mathrm{i}jk}{T}\right]\right)\left(\sum_{k=0}^{\frac{TW}{a}-1} \exp\left[-\frac{2\pi\mathrm{i}jk}{T}\right]\right)
\end{aligned}$$

Finally, we must consider the relationship between $\sin^2[x]$ and $\cos[2x]$:

$$\begin{aligned}
2\sin^2[x] &= 1 - \cos[2x]\\
Y(j) &= 4\sin^2\left[\frac{\pi ja}{T}\right]\left(\sum_{k=0}^{\frac{Tw}{a}-1} \exp\left[\frac{2\pi\mathrm{i}jk}{T}\right]\right)\left(\sum_{k=0}^{\frac{Tw}{a}-1} \exp\left[-\frac{2\pi\mathrm{i}jk}{T}\right]\right)
\end{aligned}$$

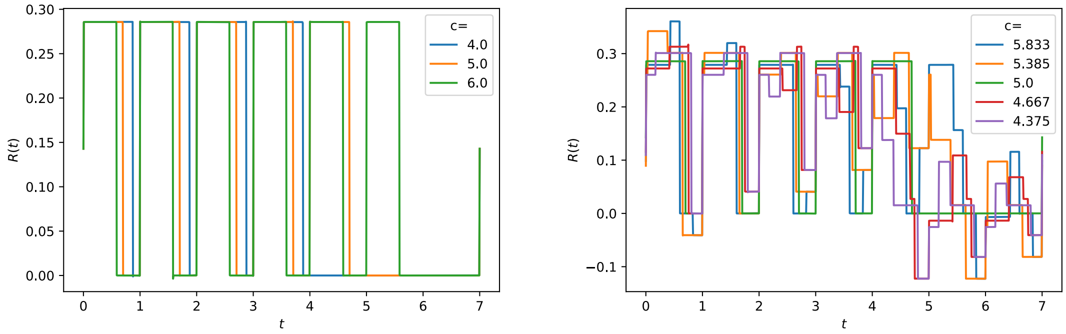
These two sum terms are geometric sums. As with before, when $\frac{j}{T}$ is an integer, the common ration in these sums will be 1. Whilst this is easy to compute, it does mean that we cannot write an expression for $Y(j)$ without a check for if $\frac{j}{T}$ is an integer.

$$Y(j) = 4\sin^2\left[\frac{\pi ja}{T}\right] \times \begin{cases} \left(\frac{Tw}{a}\right)^2 & \frac{j}{T} \in F, \\ \frac{\sin^2\left[\frac{\pi jw}{a}\right]}{\sin^2\left[\frac{\pi j}{T}\right]} & \text{otherwise.} \end{cases} \tag{3.7}$$

As with our expression for $C_{R,j}$, these two expressions are technically continuous, in that as $j$ approaches a value such that $\frac{j}{T}$ is an integer, $\frac{\sin\left[\frac{\pi jw}{a}\right]}{\sin\left[\frac{\pi j}{T}\right]}$ will approach $\frac{Tw}{a}$ according to L'Hopital's rule:

$$\begin{aligned}
\text{If } \lim_{x\to c}\frac{f(x)}{g(x)} &= \frac{0}{0} \\
\text{or } \lim_{x\to c}\frac{f(x)}{g(x)} &= \frac{\infty}{\infty}, \\
\text{then } \lim_{x\to c}\frac{f(x)}{g(x)} &= \frac{\frac{\mathrm{d}f(x)}{\mathrm{d}x}}{\frac{\mathrm{d}g(x)}{\mathrm{d}x}}. \\
f(j) &= \sin\left[\frac{\pi jw}{a}\right] \\
&= \sin\left[\frac{\pi jc}{T}\right] \\
g(j) &= \sin\left[\frac{\pi j}{T}\right] \\
\frac{x}{T} &\in \mathbb{Z} \\
f(x) &= 0 \\
g(x) &= 0 \\
\frac{\mathrm{d}f(j)}{\mathrm{d}j} &= \frac{c\pi}{T}\cos\left[\frac{\pi jc}{T}\right] \\
\frac{\mathrm{d}g(j)}{\mathrm{d}j} &= \frac{\pi}{T}\cos\left[\frac{\pi j}{T}\right] \\
\lim_{j\to x}\frac{\sin\left[\frac{\pi jw}{a}\right]}{\sin\left[\frac{\pi j}{T}\right]} &= c\frac{\cos\left[\frac{\pi xc}{T}\right]}{\cos\left[\frac{\pi x}{T}\right]} \\
\cos\left[\frac{\pi jc}{T}\right] &= 1 \\
\cos\left[\frac{\pi j}{T}\right] &= 1 \\
\lim_{j\to x}\frac{\sin\left[\frac{\pi jw}{a}\right]}{\sin\left[\frac{\pi j}{T}\right]} &= c = \frac{Tw}{a}
\end{aligned}$$

However, as we are likely using a computer to calculate this value, it would be better to write a condition for checking if $\frac{j}{T}$ is an integer as a precaution. We previously stated that we would assume that we have chosen values of $T$, the rota length, $a$, the proportion of time worked on a working day and $w$, the total proportion of time



(a)

(b)

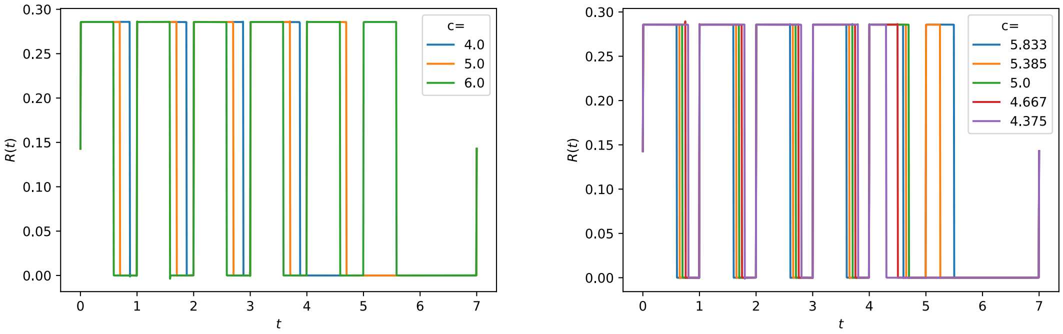
Figure 3.10. The effect of non-integer value for the number of days worked ($c$) on the infinite series representation of the rota pattern, $R(t)$. The left hand plot shows a stable rota pattern when the number of working days would be an integer, where the right hand plot shows similar values except $c$ is no longer an integer. In each case, $T = 7$, $w = 0.5$ and $a$ is allowed to vary.

worked, such that the value of $c$, the total umber of working days, that is $\frac{Tw}{a}$, would be an integer. In other words, we assume an individual works an integer number of shifts in a rota cycle. Our formula for $R(t)$ is no longer dependent on $c$ being an integer as we have calculated the resulting sums from working $c$ days. We therefore now have the opportunity to explore what happens when we relax this assumption.

Figure 3.10 shows what happens to our model if this assumption is relaxed. When $c$ is an integer value, as in the left hand plot, the model is well-behaved. The risk of infection inside work remains constant (aside from an occasional spike owing to a non-infinite approximation of an infinite summation, such as at the start of the second rest period on the green $c = 6$ plot) and shows the exact trend we were expecting. However, when $c$ is not an integer, our model falls apart. The risk of infection while at work is no longer constant. Additionally, it is possible for $R(t)$ to be negative. As $R(t)$ is supposed to be the cyclical probability density function for the time of infection of an individual, this is an inappropriate property. $R(t)$ should always be non-negative. If we do not address this it will have knock-on effects when calculating $Y(j)$ and $\hat{W}(0)$.

For this reason, we propose a slightly altered model for the rota pattern that allows us to realistically inspect parameters that result in a non-integer value of $c$. In this case, an individual works $c'$ days of length $a$, where $c'$ is an integer. They then work a further 1 shift with a length less than $a$ so that the total time they work is equal to $Tw$. This slightly changes our definitions for our parameters in our rota pattern:

$$\begin{aligned} a &> w \\ c' &= \lfloor \frac{Tw}{a} \rfloor \\ \bar{t} &= Tw \end{aligned}$$



(a) (b)

Figure 3.11. A robust model for handling rota parameters which result in a non-integer number of days. Now, an additional make-up shift is performed at the end of the rota so that each individual works the same proportion of time, regardless of the length of each individual shift. The same parameter set has been used as in Figure 3.10.

$$
\begin{aligned}
m_k &= k - 1 \\
m_{c'+2} &= T \\
n_{k,k\neq c'+1} &= k + a - 1 \\
n_{c'+1} &= Tw - ac' \\
A &= uTw + vT(1-w)
\end{aligned}
$$

Assuming an individual can only be infected at work, we now need to change our expression for $C_{R,j}$:

$$
\begin{aligned}
C_{R,j} = \frac{\mathrm{i}}{2\pi jTw}\Bigg( & T\left(\exp\left[-2\pi \mathrm{i}j\left(w+\frac{c'(1-a)}{T}\right)\right] - \exp\left[-\frac{2\pi \mathrm{i}jc'}{t}\right]\right) \\
& + \left(\exp\left[-\frac{2\pi \mathrm{i}ja}{T}\right]-1\right)\begin{cases} c' & \frac{j}{T}\in\mathbb{Z} \\ \frac{\exp\left[-\frac{2\pi \mathrm{i}c'}{T}\right]-1}{\exp\left[-\frac{2\pi \mathrm{i}}{T}\right]-1} & \text{otherwise} \end{cases}\Bigg)
\end{aligned}
$$

We can once again write $R(t)$ explicitly in the form of an infinite sum of trigonometric functions. This calculation is not included in this chapter for brevity, but its outcome is show in Figure 3.11. As we can see, now with this “make-up” day the model is stable and transitions between integer values of $c$ can be investigated. We could have chosen an infinite number of other options as to handle this problem, but investigating the issue further would be irrelevant to our analysis.

Similarly, knowing that we can calculate the limit when $\frac{j}{T}\in\mathbb{Z}$ using L’Hopital’s rule, we can calculate the trigonometric expression for $Y(j)$. We invite the reader to find this solution in their own time, but should they not we want to reassure them that it does not provide any further analytic insight.

### 3.4.5 Calculating the probability that an individual is infectious

So far in this section, we have used $\lambda(\tau)$ to represent an individual's infectiousness as a function of time since transmission, i.e. at time $t = \tau$ given that they were infected at time $t = 0$. Very early on we pointed out that the shape of this function would be different for each individual. A better understanding of the disease we are modelling would help parameterise how this curve could change between individuals. This is an avenue that could be explored at a later date.

In the previous chapter, we attempted to minimise the total time an individual is infectious at work, ignoring different levels of infectiousness and treating any time in their infectious period with the same priority. We convolved the probability density function of the latent period (the time from getting infected to becoming infectious) with the survival function for the infectious period to find a function that gives the probability that an individual is infectious at time $\tau$, where $\tau$ gives time relative to when the individual was infected:

$$\mathbb{P}(\text{Infectious at time } \tau) \int_0^\tau \mathbb{P}(\text{Latent period} = x) \times \mathbb{P}(\text{Infectious period} > \tau - x)\mathrm{d}x$$

Unfortunately, we could not find an analytical solution for this function when using Gamma distributions to represent the latent and infectious periods through non-Fourier methods. We will now show how we can calculate the Fourier transform for this function. We will start by replacing $\lambda(\tau)$ with $\kappa(\tau)$, a new function that gives the probability that an individual is infectious at time $t = \tau$ given that they were infected at time $t = 0$. If we think about this function carefully, we can see that it will have all the constraints $\lambda(\tau)$ has, as discussed in Section 3.4.2. $\hat{\kappa}(\omega)$ is the Fourier transform of this function, which we can calculate.

We start by saying that an infected individual has a fixed latent period of length $x$ and a fixed infectious period of length $y$. We are currently ignoring factors that might shorten an individual's effective infectious period such as developing symptoms and deciding not to go to work. The function $\kappa(\tau)$ can be written as follows:

$$\kappa(\tau) = \begin{cases} 1 & x \leq \tau \leq x + y \\ 0 & \text{Otherwise} \end{cases} \tag{3.8}$$

This is an indicator function. An individual will only be infectious between times $x$ and $x + y$. During this time period, their probability of being infectious is 1 ($\kappa(\tau) = 1$) and outside this time period the probability is 0 ($\kappa(\tau) = 0$). This function has a fairly simple real part to its Fourier transform:

$$
\begin{aligned}
\Re\left[\hat{\kappa}(\omega)\right] &= \Re\left[\int_{x}^{x+y} \exp\left[-\mathrm{i}\omega t\right] \mathrm{d}t\right] \\
&= \frac{\sin\left[\omega(x+y)\right]}{-} \sin\left[\omega x\right] \omega
\end{aligned}
$$

We can make use of the trigonometric identity describing the difference between two Sine functions to simplify this further:

$$
\begin{aligned}
\sin\left[\alpha\right] - \sin\left[\beta\right] &= 2\cos\left[\frac{\alpha+\beta}{2}\right] \sin\left[\frac{\alpha-\beta}{2}\right] \\
\Re\left[\hat{\kappa}(\omega)\right] &= \frac{2\cos\left[\frac{\omega(2x+y)}{2}\right] \sin\left[\frac{\omega y}{2}\right]}{\omega}
\end{aligned}
$$

The Sinc function, $\frac{\sin[x]}{x}$ can be incorporated into this calculation so that there is no longer an $\omega$ value in the denominator:

$$
\begin{aligned}
\sin\left[\frac{\omega y}{2}\right] &= \text{sinc}\left[\frac{\omega y}{2}\right] \times \frac{\omega y}{2} \\
\Re\left[\hat{\kappa}(\omega)\right] &= y\text{sinc}\left[\frac{\omega y}{2}\right] \cos\left[\omega\left(\frac{y}{2}+x\right)\right]
\end{aligned}
$$

However, in reality $x$ and $y$ should really be $X$ and $Y$, two unknown variables randomly sampled from distributions for the latent and infectious period respectively. We can calculate an analytical solution for $\hat{\kappa}(\omega)$ by integrating across all possible values of $X$ and $Y$:

$$
\Re\left[\hat{\kappa}(\omega)\right] = \int_0^\infty \mathbb{P}(X=x) \int_0^\infty \mathbb{P}(Y=y) \times y\text{sinc}\left[\frac{\omega y}{2}\right] \cos\left[\omega\left(\frac{y}{2}+x\right)\right] \mathrm{d}y\mathrm{d}x \qquad (3.9)
$$

We represent the distribution for $X$ with a Gamma distribution with input parameters $\alpha_a$ and $\beta_a$. We say that an infected individual has a probability $1-\rho$ that they will ever develop symptoms and therefore isolate themselves, the distribution for time from become infectious to developing symptoms is a Gamma distribution with parameters $\alpha_{b2}$ and $\beta_{b2}$, and the distribution for the infections period as a whole follows a Gamma distribution with input parameters $\alpha_{b1}$ and $\beta_{b1}$. We can therefore write PDFs for $X$ and $Y$:

$$
\begin{aligned}
\mathbb{P}(X=x) &\propto \frac{\beta_a^{\alpha_a}}{\Gamma\left[\alpha_a\right]} x^{\alpha_a-1} \exp\left[-\beta_a x\right] \\
\mathbb{P}(Y=y) &\propto \rho \frac{\beta_{b1}^{\alpha_{b1}}}{\Gamma\left[\alpha_{b1}\right]} y^{\alpha_{b1}-1} \exp\left[-\beta_{b1} y\right] + (1-\rho) \frac{\beta_{b2}^{\alpha_{b2}}}{\Gamma\left[\alpha_{b2}\right]} y^{\alpha_{b2}-1} \exp\left[-\beta_{b2} y\right]
\end{aligned}
$$

Gamma distributions have a very accommodating characteristic function (the Fourier transform of their probability density function), which means that in the case of our latent and infectious period distributions, we can find an analytical solution for $\hat{\kappa}(\omega)$.

Previously we showed the calculation required to generate an analytical solution for the real value part of $\hat{\kappa}(\omega)$, where $\hat{\kappa}(\omega)$ is the Fourier transform of a function $\kappa(\tau)$, which in turn gives the probability of an individual being infectious at time $\tau$ given that they were infected at time 0. $X$ and $Y$ are unknown Gamma distributed variables equal to the lengths of the individual's latent and effective infectious period respectively. When $\omega = 0$, $Re\left[\hat{\kappa}(\omega)\right]$ is easily calculated:

$$\Re\left[\hat{\kappa}(0)\right] = \mathbb{E}\left[Y\right] = \rho\frac{\alpha_{b1}}{\beta_{b1}} + (1-\rho)\frac{\alpha_{b2}}{\beta_{b2}}$$

By incorporating our sinc function, we can calculate $\Re\left[\hat{\kappa}(\omega)\right]$ for all values of $\omega$:

$$\begin{aligned}\Re\left[\hat{\kappa}(\omega)\right] = &\int_0^\infty\int_0^\infty \frac{\beta_a^{\alpha_a}}{\Gamma\left[\alpha_a\right]}x^{\alpha_a-1}\exp\left[-\beta_a x\right]\\ &\times\left(\rho\frac{\beta_{b1}^{\alpha_{b1}}}{\Gamma\left[\alpha_{b1}\right]}y^{\alpha_{b1}-1}\exp\left[-\beta_{b1}y\right] + (1-\rho)\frac{\beta_{b2}^{\alpha_{b2}}}{\Gamma\left[\alpha_{b2}\right]}y^{\alpha_{b2}-1}\exp\left[-\beta_{b2}y\right]\right)\\ &\times y\text{sinc}\left[\frac{\omega y}{2}\right]\cos\left[\omega\left(\frac{y}{2}+x\right)\right]\mathrm{d}x\mathrm{d}y\end{aligned}$$

We start by solving the internal, d$x$ integral:

$$\begin{aligned}\Re\left[\hat{\kappa}(\omega)\right] = &\int_0^\infty \frac{\beta_a^{\alpha_a}}{\Gamma\left[\alpha_a\right]}\left(\beta_a^2+\omega^2\right)^{-\frac{\alpha_a}{2}}\cos\left[\frac{\omega y}{2}+\arctan\left[\frac{\omega}{\beta_a}\right]\right]\\ &\times\left(\rho\frac{\beta_{b1}^{\alpha_{b1}}}{\Gamma\left[\alpha_{b1}\right]}y^{\alpha_{b1}-1}\exp\left[-\beta_{b1}y\right] + (1-\rho)\frac{\beta_{b2}^{\alpha_{b2}}}{\Gamma\left[\alpha_{b2}\right]}y^{\alpha_{b2}-1}\exp\left[-\beta_{b2}y\right]\right)\\ &\times y\text{sinc}\left[\frac{\omega y}{2}\right]\mathrm{d}y\end{aligned}$$

We then solve the external, d$y$ integral:

$$\begin{aligned}\Re\left[\hat{\kappa}(\omega)\right] = &\frac{\beta_a^{\alpha_a}}{\omega}\left(\beta_a^2+\omega^2\right)^{-\frac{\alpha_a}{2}}\\ &\times\Bigg(\rho\left(\beta_{b1}^{\alpha_{b1}}\left(\beta_{b1}^2+\omega^2\right)^{-\frac{\alpha_{b1}}{2}}\sin\left[\alpha_a\arctan\left[\frac{\omega}{\beta_a}\right]+\alpha_{b1}\arctan\left[\frac{\omega}{\beta_{b1}}\right]\right]\right)\\ &+(1-\rho)\left(\beta_{b2}^{\alpha_{b2}}\left(\beta_{b2}^2+\omega^2\right)^{-\frac{\alpha_{b2}}{2}}\sin\left[\alpha_a\arctan\left[\frac{\omega}{\beta_a}\right]+\alpha_{b2}\arctan\left[\frac{\omega}{\beta_{b2}}\right]\right]\right)\\ &-\sin\left[\alpha_a\arctan\left[\frac{\omega}{\beta_a}\right]\right]\Bigg)\end{aligned}$$

In order to simplify this further, we introduce three new terms, $\theta_a$, $\theta_{b1}$ and $\theta_{b2}$. Each is a corner of a right-angled triangle whose non-hypotenuse sides are $\omega$ and $\beta_a$, $\beta_{b1}$ and $\beta_{b2}$ respectively, such that $\tan\left[\theta_x\right] = \frac{\omega}{\beta_x}$. These triangles are demonstrated

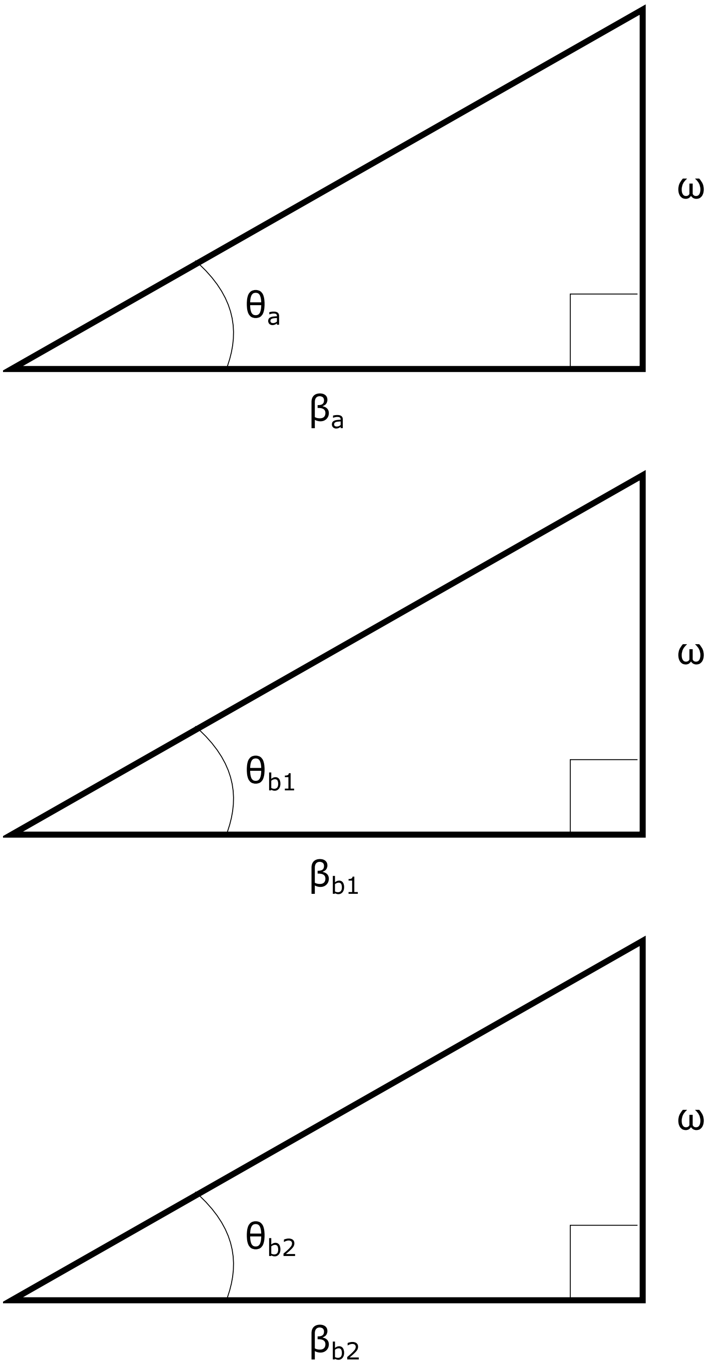


Figure 3.12. Graphical representation of the three triangles used to calculate values of $\theta$. In reality, $\omega$ can take negative values, but this is not depicted here.

in Figure 3.12. Due to the geometry of the triangles, we also find that $\cos\left[\theta_x\right] = \beta_x \left(\beta_x^2 + \omega^2\right)^{-\frac{1}{2}}$.

$$\begin{aligned}
\tan\left[\theta_a\right] &= \frac{\omega}{\beta_a} \\
\cos\left[\theta_a\right] &= \beta_a \left(\beta_a^2 + \omega^2\right)^{-\frac{1}{2}} \\
\tan\left[\theta_{b1}\right] &= \frac{\omega}{\beta_{b1}} \\
\tan\left[\theta_{b2}\right] &= \frac{\omega}{\beta_{b2}}
\end{aligned}$$

$$
\begin{aligned}
\Re\left[\hat{\kappa}(\omega)\right] = \frac{\cos^{\alpha_a}\left[\theta_a\right]}{\omega}& \\
\times \Bigg(&\rho\cos^{\alpha_{b1}}\left[\theta_{b1}\right]\sin\left[\alpha_a\theta_a + \alpha_{b1}\theta_{b1}\right] \\
&+ (1-\rho)\cos^{\alpha_{b2}}\left[\theta_{b2}\right]\sin\left[\alpha_a\theta_a + \alpha_{b2}\theta_{b2}\right] \\
&- \sin\left[\alpha_a\theta_a\right]\Bigg)
\end{aligned}
$$

For completeness, let us investigate $\Im\left[\hat{\kappa}(\omega)\right]$. As this is not strictly necessary for our future calculations and requires much the same logical leaps as when calculating $\Re\left[\hat{\kappa}(\omega)\right]$, this explanation shall be less verbose. We start, as with finding the real part, by finding the imaginary part of $\hat{\kappa}(\omega)$ if $x$ and $y$ are fixed:

$$
\begin{aligned}
\Im\left[\hat{\kappa}(\omega)\right] &= \Im\left[\int_x^{x+y}\exp\left[-\mathrm{i}\omega t\right]\mathrm{d}t\right] \\
&= \frac{\cos\left[\omega x\right] - \cos\left[\omega(x+y)\right]}{\omega} \\
\cos\left[\alpha\right] - \cos\left[\beta\right] &= -2\sin\left[\frac{\alpha+\beta}{2}\right]\sin\left[\frac{\alpha-\beta}{2}\right] \\
\Im\left[\hat{\kappa}(\omega)\right] &= -\frac{2\sin\left[\omega\left(\frac{y}{2}+x\right)\right]\sin\left[-\frac{\omega y}{2}\right]}{\omega} \\
&= y\text{sinc}\left[\frac{\omega y}{2}\right]\sin\left[\omega\left(\frac{y}{2}+x\right)\right]
\end{aligned}
$$

We can then incorporate this into a model where $x$ and $y$ are allowed to vary according to the previously established distributions:

$$
\begin{aligned}
\Im\left[\hat{\kappa}(\omega)\right] &= \int_0^\infty\int_0^\infty \frac{\beta_a^{\alpha_a}}{\Gamma\left[\alpha_a\right]}x^{\alpha_a-1}\exp\left[-\beta_a x\right] \\
&\times\left(\rho\frac{\beta_{b1}^{\alpha_{b1}}}{\Gamma\left[\alpha_{b1}\right]}y^{\alpha_{b1}-1}\exp\left[-\beta_{b1}y\right] + (1-\rho)\frac{\beta_{b2}^{\alpha_{b2}}}{\Gamma\left[\alpha_{b2}\right]}y^{\alpha_{b2}-1}\exp\left[-\beta_{b2}y\right]\right) \\
&\times y\text{sinc}\left[\frac{\omega y}{2}\right]\sin\left[\omega\left(\frac{y}{2}+x\right)\right]\mathrm{d}x\mathrm{d}y \\
&= \int_0^\infty \beta_a^{\alpha_a}\left(\beta_a^2+\omega^2\right)^{-\frac{\alpha_a}{2}}\sin\left[\frac{\omega y}{2} + \alpha_a\arctan\left[\frac{\omega}{\beta_a}\right]\right] \\
&\times\left(\rho\frac{\beta_{b1}^{\alpha_{b1}}}{\Gamma\left[\alpha_{b1}\right]}y^{\alpha_{b1}-1}\exp\left[-\beta_{b1}y\right] + (1-\rho)\frac{\beta_{b2}^{\alpha_{b2}}}{\Gamma\left[\alpha_{b2}\right]}y^{\alpha_{b2}-1}\exp\left[-\beta_{b2}y\right]\right) \\
&\times y\text{sinc}\left[\frac{\omega y}{2}\right]\mathrm{d}y \\
&= \frac{\beta_a^{\alpha_a}}{\omega}\left(\beta_a^2+\omega^2\right)^{-\frac{\alpha_a}{2}} \\
&\Bigg(\cos\left[\alpha_a\arctan\left[\frac{\omega}{\beta_a}\right]\right]
\end{aligned}
$$

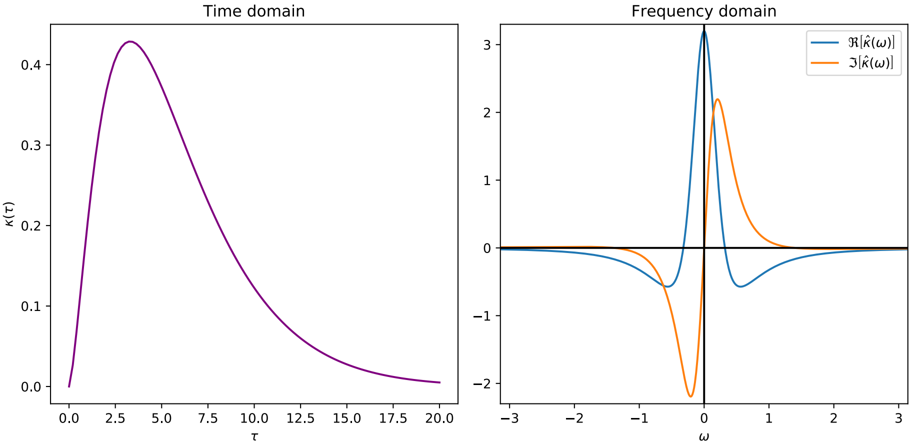


Figure 3.13. The time domain and frequency domain representation of the probability of being infectious at time $\tau$ after being infected.

$$
\begin{aligned}
&- \rho\beta_{b1}^{\alpha_{b1}} \left(\beta_{b1}^2 + \omega^2\right)^{-\frac{\alpha_{b1}}{2}} \cos\left[\alpha_a \arctan\left[\frac{\omega}{\beta_a}\right] + \alpha_{b1} \arctan\left[\frac{\omega}{\beta_{b1}}\right]\right] \\
&- (1-\rho)\beta_{b2}^{\alpha_{b2}} \left(\beta_{b2}^2 + \omega^2\right)^{-\frac{\alpha_{b2}}{2}} \cos\left[\alpha_a \arctan\left[\frac{\omega}{\beta_a}\right] + \alpha_{b2} \arctan\left[\frac{\omega}{\beta_{b2}}\right]\right]\Bigg) \\
= \frac{\cos^{\alpha_a}[\theta_a]}{\omega} \\
\Bigg(&\cos[\alpha_a\theta_a] \\
&- \rho\cos^{\alpha_{b1}}[\theta_{b1}]\cos[\alpha_a\theta_a + \alpha_{b1}\theta_{b1}] \\
&- (1-\rho)\cos^{\alpha_{b2}}[\theta_{b2}]\cos[\alpha_a\theta_a + \alpha_{b2}\theta_{b2}]\Bigg)
\end{aligned}
$$

Figure 3.13 shows both $\kappa(\tau)$ and $\hat{\kappa}(\omega)$ for the parameter set used in this chapter. Note that, as with Figure 3.9, $\Re[\hat{\kappa}(\omega)]$ is even and $\Im[\hat{\kappa}(\omega)]$ is odd.

We use the parameters from Table 3.2 to represent an individual's infectious career. In this case, we assume that infections only occur at work (i.e. $u = 1, v = 0$). The latent period has a mean of 2.84 days with a standard deviation of 2.34. There is a probability of asymptomatic infection of 0.3. If the individual remains asymptomatic then they are infectious for an average of 6 days with a standard deviation of $\sqrt{12}$. Otherwise they detect their symptoms an self-isolate, resulting in an infectious, prodromal period of 2 days with a standard deviation of 1.5. Each of these unknown values follow a Gamma distribution. We now have enough information to calculate a central estimate for the total length of time an individual in this scenario would be infectious whilst at work.

### 3.4.6 Incorporating regular testing into a Fourier model

At this stage we are unfortunately unable to fully integrate regular testing into our modelling approach based on Fourier transforms. With a fixed testing schedule, the force of infection function changes shape depending on when an individual becomes infectious or detectable. For example, in the model where an individual's force of infection is equal to 1 during their infectious period (this is the model that we have used to track how long an individual is infectious for), in the absence of testing the expected force of infection at time $t$ given that the individual was infected at time 0 is the probability that the individual is still infectious at time $t$ or, in other words, the survival function of the probability distribution that defines the length of the infectious period. Therefore, without testing we do not need to consider the exact time an individual becomes infectious and instead can integrate through all possible latent and infectious period lengths. With testing we need to pay attention specifically to when an individual becomes detectable, as the testing pattern will have no influence prior to this time (excluding the possibility of false positives).

As a result, once testing is introduced, if there is a delay between being infected and becoming detectable (as it would be reasonable to assume there is), we can no longer integrate through all possible values of the latent and infectious period length. We have to consider them separately. The survival function for an individual's effective infectious period (which can be stopped early if they are diagnosed before it is complete) is now dependent on if a test occurred before or after they became detectable, as anyone who is detected would finish their effective infectious period. However, just because we cannot use Fourier techniques to achieve the same simplifications as in a case where there are no tests does not mean that we cannot use some Fourier techniques to improve the accuracy of our analysis, as we shall now show.

Let us say that a disease's infectious period falls on a Gamma distribution with parameters $\alpha_b$ and $\beta_b$ (previously we included the possibility that an individual could become symptomatic and remove themselves from work, but as this does not greatly change the nature of this analysis we leave it to the reader to observe how this option could be included in our model). The probability that an individual would still be infectious at time $\tau$ given that they started their infectious period at time 0 is given by the upper incomplete Gamma function $\frac{\Gamma(\alpha_b,\beta_b\tau)}{\Gamma(\alpha_b)}$. We want to track the integral of this function as this will give us the total time an individual is infectious for. Fortunately, this can be found analytically:

$$\int \frac{\Gamma\left(\alpha_b, \beta_b t\right)}{\Gamma\left(\alpha_b\right)} \mathrm{d}t = \frac{t\Gamma\left(\alpha_b, \beta_b t\right)}{\Gamma\left(\alpha_b\right)} - \frac{\alpha_b \Gamma\left(\alpha_b+1, \beta_b t\right)}{\beta_b \Gamma\left(\alpha_b+1\right)}$$

We can now use $\mathbf{m}$ and $\mathbf{n}$, the sets of starts and ends of shifts respectively, and $T$, the total length of the rota pattern, to write an infinite sum that calculates $L(t)$, how long an individual would spend at work whilst infectious given that they became

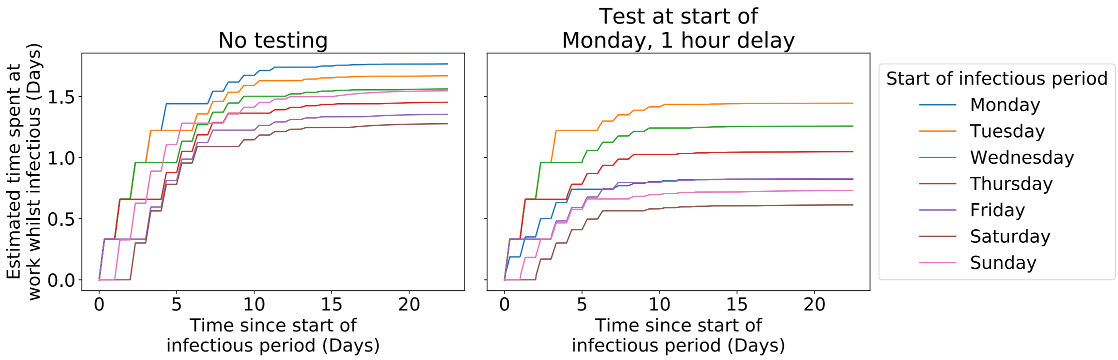


Figure 3.14. Cumulative estimate for the length of time spent at work whilst infectious as a function of the time since the beginning of an individual's infectious window. Each colour denotes a different day that the infectious period started on. In this case we use a 9-5 5 days a week work pattern, with the individual starting their infectious period at the start of their shift (or the equivalent time on rest days). The infectious period length follows a Gamma distribution with a mean of 6 days and a standard deviation of $\sqrt{12}$. The left-hand plot demonstrates how this estimate changes without testing while the right-hand plot includes a regular test occurring at the start of Monday shifts with an hour delay between test and results and a false negative rate of 0.5.

infectious at time $t$ (for convenience we will assume that $t \geq 0$):

$$\begin{aligned} L(t) = \sum_{j=0}^{\infty}\sum_{k=1}^{c} H\left(n_k + Tj - t\right) \Bigg[ & \frac{\left(n_k + Tj - t\right)\Gamma\left(\alpha_b, \beta_b\left(n_k + Tj - t\right)\right)}{\Gamma\left(\alpha_b\right)} \\ & + \frac{\alpha_b \Gamma\left(\alpha_b + 1, \beta_b \textbf{Max}\left[m_k + Tj - t, 0\right]\right)}{\beta_b \Gamma\left(\alpha_b + 1\right)} \\ & - \left(\frac{\textbf{Max}\left[m_k + Tj - t, 0\right]\Gamma\left(\alpha_b, \beta_b \textbf{Max}\left[m_k + Tj - t, 0\right]\right)}{\Gamma\left(\alpha_b\right)} \right. \\ & \left. + \frac{\alpha_b \Gamma\left(\alpha_b + 1, \beta_b\left(n_k + Tj - t\right)\right)}{\beta_b \Gamma\left(\alpha_b + 1\right)}\right) \Bigg] \end{aligned}$$

where $H(x)$ is the Heaviside step function[171] and $c$ is the total number of shifts in the rota pattern. Whilst this infinite sum does not have an analytical solution, we can truncate it to achieve any required degree of accuracy by using our knowledge of the Gamma distribution to calculate a suitable time after which the probability that the individual is still infectious past said point is negligible.

Including testing into this model is a trivial task. The survival function for the probability that an individual is still infectious needs to be reduced by a factor of the false negative rate of the test at each point where they could have had a positive test (accounting for delays between taking a test and receiving a result). This is because, for an individual to continue being infectious past this time, they must receive a false negative result, the probability of which is equal to the false negative rate. In turn, the integral of the survival function from that point onwards will also be reduced by a factor equal to the the false negative rate. As we are modelling a testing pattern that stays in line with the rota pattern, approximating our new $L(t)$ becomes a matter of calculating when these tests occur with respect to $t$ and reducing the subsequent integrals through multiplications by the false negative rate.

Figure 3.14 demonstrates an example of how regular testing could affect the length of time spent at work whilst infectious. In the left-hand plot we can see how an infected individual's estimated length of time spent at work whilst infectious would change for a given start of the infectious period. In this case we have chose a 9-5 rota and started their infectious period at 9 in the morning. Monday, the start of the shift pattern, would be the worst time to start their infectious period as this would on average result in the most time spent at work whilst infectious (as indicated by Monday's line ending the highest). In the right-hand plot we see what would happen if we tested at the start of each Monday shift, with an hour's delay between testing and results and a rate of false negative tests of 0.5. The end estimate for Monday has become much better, as half of individual's who start their infectious period on a Monday morning are now identified. Tuesday now becomes the worst time to start an infectious period, as it takes 6 days for any of them to be potentially identified as infectious. This was already the mean length of infectious period anyway, so removing half at this point has little effect.

We have now demonstrated how, for testing at a fixed point in a rota, we can estimate the mean length of time someone will spend at work given a specific start to their infectious period. We do not have to limit ourselves to one test in the rota as well, with the outcome being analytically similar, nor do different tests need to have the same false negative rate. Indeed, if we were tracking total force of infection rather than length of time at work, we could include a link between the false negative rate of a test and the individual's force of infection at time of testing. After all, there is an argument to be made that both are linked through viral or bacterial load. This is beyond the work of this thesis.

To demonstrate how effective a particular testing strategy would be we now need to calculate the probability of an individual starting their infectious period at a particular point in the rota $t$, multiply this by $L(t)$, the estimated length of time spent at work whilst infectious given their infectious period started at time $t$ and then integrate numerically across all possible values of $t$. It is with the distribution for the start of the infectious period that we can take advantage of Fourier analysis.

Recall that $R(t)$ was the repeating function representing an individual's probability of being infected at any point of a rota pattern, influenced by $u$ and $v$, the relative risk of being infected inside and outside of work. Rather than considering $\hat{R}(\omega)$, the Fourier transform of $R(t)$, we will start by considering $R(t)$ in terms of a Fourier series:

$$R(t) = a_0 + \sum_{j=1}^{\infty} a_j \cos\left[\frac{2\pi jt}{T}\right] + b_j \sin\left[\frac{2\pi jt}{T}\right]$$

where

$$\begin{aligned}
a_0 &= \frac{1}{T}\int_0^T R(t)\mathrm{d}t \\
a_j &= \frac{2}{T}\int_0^T R(t)\cos\left[\frac{2\pi j t}{T}\right]\mathrm{d}t \\
b_j &= \frac{2}{T}\int_0^T R(t)\sin\left[\frac{2\pi j t}{T}\right]\mathrm{d}t
\end{aligned}$$

Given that $R(t)$ is a probability density function for the time of infection in each complete rota, we know that between 0 and $T$ it must integrate to 1, making $a_0 = \frac{1}{T}$. As with the constants $C_{R,n}$, by remembering that the end of the last rest period in one rota is equivalent to the beginning to the first shift in the next pattern, we can write a simplified version of $a_n$ and $b_n$:

$$\begin{aligned}
a_j &= \frac{u-v}{A\pi j}\sum\left(\sin\left[\frac{2\pi j\mathbf{n}}{T}\right] - \sin\left[\frac{2\pi j\mathbf{m}}{T}\right]\right) \\
b_j &= \frac{v-u}{A\pi j}\sum\left(\cos\left[\frac{2\pi j\mathbf{n}}{T}\right] - \cos\left[\frac{2\pi j\mathbf{m}}{T}\right]\right)
\end{aligned}$$

Through trigonometric identities we can therefore show:

$$\begin{aligned}
R(t) = \frac{1}{T} + \frac{u-v}{A\pi}\sum_{j=1}^{\infty}\frac{1}{j}\Bigg[&\cos\left[\frac{2\pi j t}{T}\right]\left(\sum\sin\left[\frac{2\pi j\mathbf{n}}{T}\right] - \sin\left[\frac{2\pi j\mathbf{m}}{T}\right]\right) \\
&- \sin\left[\frac{2\pi j t}{T}\right]\left(\sum\cos\left[\frac{2\pi j\mathbf{n}}{T}\right] - \cos\left[\frac{2\pi j\mathbf{m}}{T}\right]\right)\Bigg]
\end{aligned}$$

Remembering that the Sine of the difference between two angles, $\sin[\alpha-\beta]$ can be written as $\sin[\alpha]\cos[\beta] - \cos[\alpha]\sin[\beta]$, this becomes:

$$R(t) = \frac{1}{T} + \frac{u-v}{A\pi}\sum_{j=1}^{\infty}\frac{1}{j}\sum\sin\left[\frac{2\pi j}{T}(\mathbf{n}-t)\right] - \sin\left[\frac{2\pi j}{T}(\mathbf{m}-t)\right]$$

and once again we make use of the difference between two Sine functions, $\sin[\alpha] - \sin[\beta] = 2\cos\left[\frac{\alpha+\beta}{2}\right]\sin\left[\frac{\alpha-\beta}{2}\right]$, so that we can make further simplifications:

$$R(t) = \frac{1}{T} + \sum_{j=1}^{\infty}\frac{2(u-v)}{A\pi j}\sum\left(\sin\left[\frac{\pi j(\mathbf{n}-\mathbf{m})}{T}\right]\cos\left[\frac{\pi j}{T}(2t-(\mathbf{n}+\mathbf{m})\right]\right)$$

This means we have a distribution for the time of infection with regards to the rota pattern in the form of a Fourier series. We want to know the probability that an individual starts their infectious period at any point in the rota pattern. If we say that the distribution for the disease's latent period can be represented by a Gamma distribution with input parameters $\alpha_a$ and $\beta_a$, then the probability that they start their infectious period at time $t$ is equivalent to the convolution for $R(t)$ with this Gamma distribution.

$$\begin{aligned}\mathbb{P}\left(t_i = t\right) =& \int_0^\infty \frac{\beta_a^{\alpha_a}}{\Gamma\left(\alpha_a\right)} x^{\alpha_a - 1} \exp\left[-\beta_a x\right] \left(\frac{1}{T}\right. \\ & \left. + \sum_{j=1}^{\infty} \frac{2(u-v)}{A\pi j} \sum \left(\sin\left[\frac{\pi j(\mathbf{n}-\mathbf{m})}{T}\right] \cos\left[\frac{\pi j}{T}(2(t-x) - (\mathbf{n}+\mathbf{m})\right]\right)\right) \mathrm{d}x \\ =& \frac{1}{T} + \sum_{j=1}^{\infty} \frac{2(u-v)\left(T\beta_a\right)^{\alpha_a}}{A\pi n\left(\left(T\beta_a\right)^2 + (2\pi j)^2\right)^{\frac{\alpha_a}{2}}} \\ & \times \sum \sin\left[\frac{\pi j(\mathbf{n}-\mathbf{m})}{T}\right] \cos\left[\frac{\pi j(\mathbf{n}+\mathbf{m}-2t)}{T} + \alpha_a \arctan\left[\frac{2\pi j}{T\beta_a}\right]\right]\end{aligned}$$

Once again, to simplify our notation, we can introduce a $\theta$ term, $\theta_{a,j}$ such that $\tan\left[\theta_{a,j}\right] = \frac{2\pi j}{T\beta_a}$:

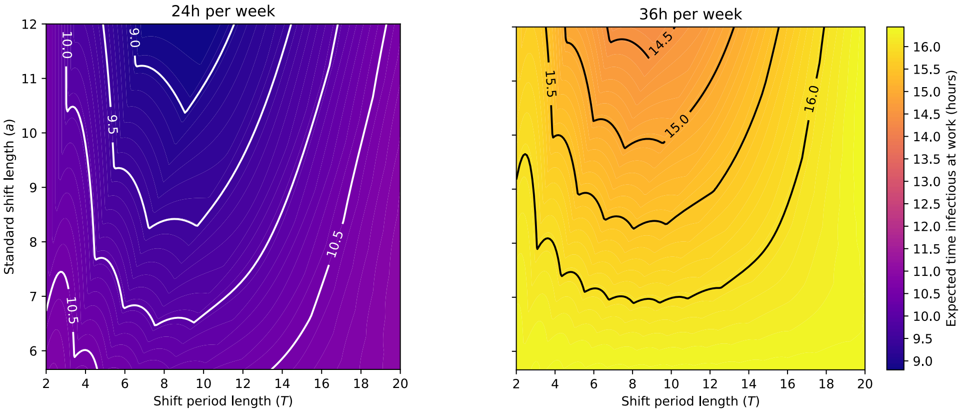
$$\begin{aligned}\tan\left[\theta_{a,j}\right] &= \frac{2\pi j}{T\beta_a} \\ \cos\left[\theta_{a,j}\right] &= \beta_a\left(\left(T\beta_a\right)^2 + (2\pi j)^2\right)^{-\frac{1}{2}} \\ \mathbb{P}\left(t_i = t\right) &= \frac{1}{T} + \sum_{j=1}^{\infty} \frac{2(u-v)\cos^{\alpha_a}\left[\theta_{a,j}\right]}{A\pi n} \\ & \quad \times \sum \sin\left[\frac{\pi j(\mathbf{n}-\mathbf{m})}{T}\right] \cos\left[\frac{\pi j(\mathbf{n}+\mathbf{m}-2t)}{T} + \alpha_a \theta_{a,j}\right]\end{aligned}$$

where $t_i$ is the time of the start of their infectious period.

An individual's expected length of time spent at work whilst infectious given a particular testing strategy is given by the multiplication of this probability by $L(t)$, integrated across all possible values of $t$ between 0 and $T$:

$$\mathbb{E}\left[\text{in-work infectious periods}\right] = \int_0^T \mathbb{P}(\text{Starting infectious period at time } t) \times L(t)\mathrm{d}t$$

As $L(t)$ is a periodic function, it can theoretically be represented as a Fourier series and calculating this integration numerically could be avoided. However, with the inclusion of testing we have been unable to find an analytical solution to $L(t)$ and so cannot find its Fourier series explicitly.



(a)

(b)

Figure 3.15. Demonstration of the expected time at work whilst infectious given different fixed rota patterns. In the observation equivalent study, we blocked out any rota pattern where an individual would not get at least 1 day off per rotation. We have decided to remove this constraint here.

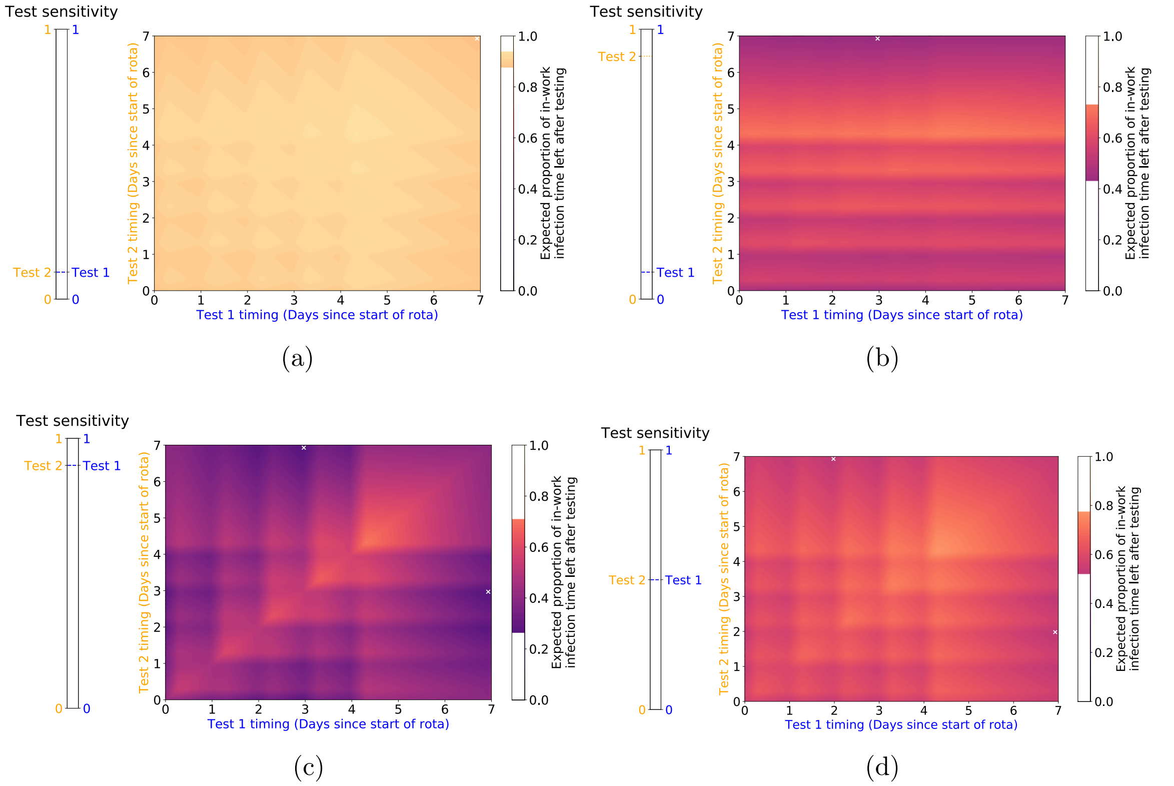
### 3.4.7 Results

Figure 3.15 shows the outcome of central estimates for the total at-work infectious times for an individual who works 24 hours per week and 36 hours per week depending on the structure of their rota. We can see that it has a lot in common with our initial observational estimates made in the previous chapter. If working 24 hours a week, these central estimates range between 8.8 and 10.95 hours, and if working 36 hours a week these range between 14.4 and 16.4 hours. The structure of the rota pattern can vary the expected at-work infectious time by as much as 2 hours.

As with the observational equivalent model, there appears to be an optimum length of rota pattern ($T$), although in our new analysis it appears to be somewhere between 8 and 10 days rather than just above 10 days as suggested by the analysis in the previous chapter. Increasing the length of the standard shift ($a$) consistently decreases the expected in-work infectious time. Finally, and somewhat unsurprisingly, increasing the proportion of time spent at work ($w$) increases the expected in work infectious period. However, this increase is not uniform and results in some shifting of the contour lines.

Figures 3.16 and 3.17 show some theoretical outputs for 2-test testing patterns across a 9-5 5 day a week rota. Both look at 9-5 5 days a week rota patterns with two regular tests, and a disease with the same parameter settings as in 3.15, but whilst Figure 3.16 shows two tests with the same delay of 1 hour between test and result, 3.17 shows the effect of having one high-sensitivity test with a whole day's delay between test and result and one low sensitivity test with only an hour's delay, a scenario designed to be analogous to a polymerase-chain-reaction (PCR) / Lateral Flow Test (LFT) combined testing strategy.

Consistently, the optimum strategy comes from the results of the higher sensitivity



(a)

(b)

(c)

(d)

Figure 3.16. Contour plots of expected in-work infectious time for a 9-5 5 day a week work pattern with tests that take 1 hour to return with the same disease dynamics as in the body of this chapter, as a function of the time (in days) after the start of the first shift when such tests are taken. Figure 3.16a shows the outcome of two low-sensitivity tests (sensitivity=0.1), Figure 3.16b shows the outcome from one low sensitivity test and one high sensitivity test (sensitivity = 0.1,0.9), Figure 3.16c shows the outcome from two high sensitivity tests (sensitivity = 0.9) and Figure 3.16d shows the outcome from two mid-range sensitivity tests (sensitivity = 0.6). White crosses indicate the optimum testing time for each combination of sensitivities.

test coming back in time for the start of the first shift of the week. This is still true when the results are delayed by a day (Figure 3.17). When there is disparity between the sensitivity of the two tests, the placement of the more sensitive test has a greater influence on the outcome of the strategy (Figure 3.16b). The optimum placement of the second test if heavily influenced by the sensitivity of the first test, as well as its own sensitivity. If the highest sensitivity is low enough, then the optimum placement for the second test is at the exact same time as the first one so that both results come back at the start of the week (Figure 3.16a). This assumes that the probability of a false negative in the second test is independent of a false negative from the first test taken at the same time, which may not be the case, depending on how each test is taken.

If the sensitivity of the first test is high enough, the second test is better off being placed elsewhere in the week, but still so that its results come in as the individual starts a shift. Which shift to choose depends on the sensitivity of both tests, as can be seen in Figures 3.16c and 3.16d. We can see that the timing of higher sensitivity tests has a greater influence on the outcome than the lower sensitivity tests and that careful consideration of each test's sensitivity needs to be applied when making this decision.

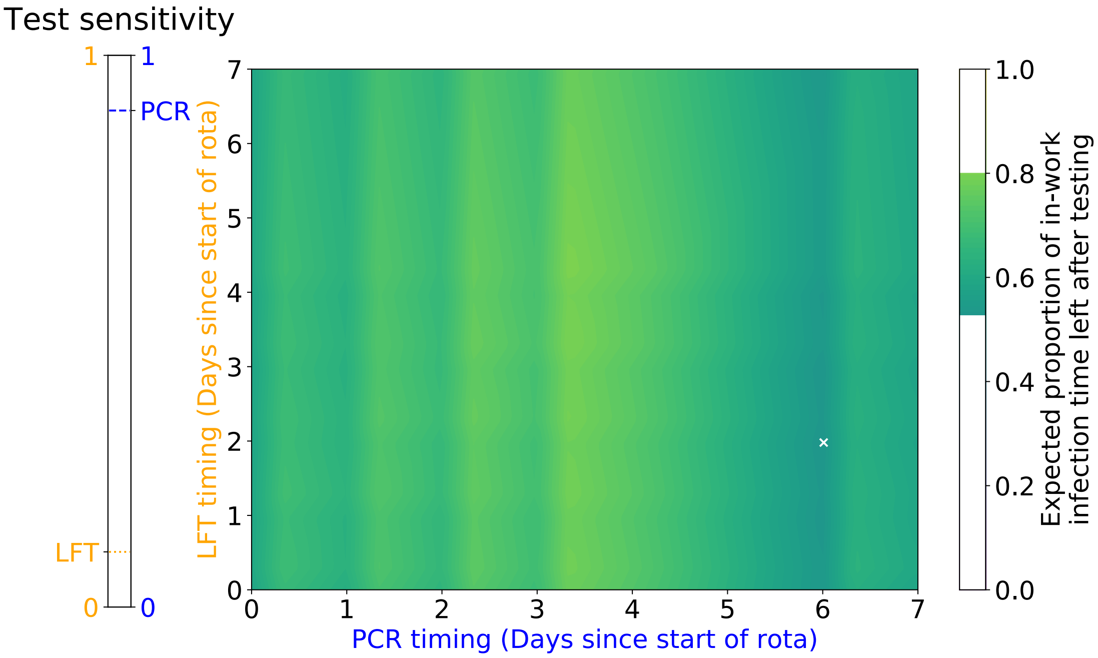


Figure 3.17. Contour plot showing the estimated in-work infectious time for a dual-test strategy where one test is sensitive but with a day's delay between test and result, and the other test has a lower sensitivity but only an hour's delay between test and result. These are analogous to PCR and LFT tests respectively, although extreme values of sensitivity have been chosen for demonstration purposes.

## 3.5 Discussion

From both analyses we can draw a few general rules regarding shift patterns:

1. A rota pattern with longer, fewer shifts results in a shorter length of time at work whilst infectious.
2. Whilst there does appear to be an optimum length of rota pattern when it comes to minimising the length of time at work whilst infectious, this optimum is less clear when looking at minimising the total start and end of shifts attended whilst infectious.
3. The optimum time to arrange regular testing is such that results from the test would come back at the start of the rota pattern.

Each of theses rules can be explained with careful consideration. If a person can work the majority of their hours in the short delay between exposure ($T_e$) and the beginning of the infectious period ($T_i$) then more of their infectious period will cross over with their rest period, reducing their average time at work whilst infectious. Rule 1 allows for this to occur, as the worker is able to do more hours of work in this delay period

If shift length and the ratio of days off to days on is maintained, as the rota period increases, the rest period increases. Initially, this increases the proportion of the infectious period spent during the rest period. However, the work period also increases.

In this model, the individual is only infected during a work shift. By increasing the number of days worked in a row, the average length of time from becoming exposed to starting their rest period increases, and so the probability of completing the entire incubation period before the rest period begins to increase. As a result, there is an optimum total rota length for any given ratio of days on to days off, as stated in Rule 2.

The second part of Rule 2, looking at the effect changing the rota pattern has on $\mathbb{E}[H]$, the expected number of handovers attended whilst infectious, can be explained by considering how changing the rota pattern will effect the number of handovers occurring per any time period. Increasing the length of individual shifts $a$ whilst maintaining the same length of rota pattern and proportion of time worked will result in fewer handovers over a given time period and therefore a lower $\mathbb{E}[H]$. However, changing the length of the rota pattern whilst maintaining the same length of individual shifts will result in approximately the same number of handovers per time period. The relationship between shift length and number of handovers attended whilst infectious is far stronger than the relationship between rota pattern length and number of handovers attended whilst infectious. Changing the length of shifts will not affect $T_h$ in the same way that changing the number of shifts does, resulting in Rule 2. We consider Rule 2 further in Section 3.5.1.

Rule 3 states that the optimum testing time results from testing so that the results come just at the start of an individual's first shift. In order to make the most effective reduction, we want to capture the most time an individual will be at work whilst infectious. It therefore seems reasonable that the time a detectable individual would have the most remaining time at work whilst infectious would be right at the start of the rota pattern. Indeed if the timing of infection $T_e$ and infectious period onset $T_i$ were Uniform with respect to the rota pattern then we would be certain that this is true. In that scenario, the remaining length of infectious period that a detectable individual left given that they are detected at time $t$ would be independent of time $t$. The amount of time they would have subsequently spent at work whilst infectious would have therefore been dictated solely by the proportion of time after their test that would have been at work. This is largest at the start of the rota pattern and therefore testing such that their results would have come back at the start of the pattern would result in the largest decrease in total time at work whilst infectious.

In this model, the timing of infection is not Uniform with respect to the rota pattern, as individuals could only be infected whilst at work. In turn, the timing of infectious period onset was not Uniform and the size of the remaining infectious period of a detectable individual would not be independent of the timing of the test. However, for our parameter sets this variability was not enough to change the optimum timing of test results. For a long enough pause between a run of shifts (i.e. a high enough value of $d$) we can imagine a scenario where this variability would be enough. If individuals can only be infected at work, for a high enough value of $d$, the probability that anyone who was infected in one rotation of the rota pattern would still be infectious come the start of the subsequent rota would be so small, and if they were still infectious their remaining infectious period would be so brief that it would be almost pointless to test so that the results come back at the start of the rota, and another optimum would need to be found.

Outside these rules, this analysis demonstrates more simply that in our model the rota pattern affects the length of time spent at work whilst infectious. To reduce workplace infections, interventions to the rota pattern can be considered.

Our model makes a number of assumptions. Most notably, the probability of being infected outside of work in our model is zero, whilst the probability of being infected at any time at work has a Uniform distribution. Both of these statements are likely not to be true. A more nuanced analysis may consider a lower but non-zero probability of being infected outside of work, or a varying probability of infection during work hours, for example an increase during known periods of increased human contact. An analysis of a model with a uniform probability of being infected throughout the week would in effect be independent to $T_e$ and would instead look at the relationship between $T_s - T_i$ and $T_e$.

This model also assumes to know the probability that an infected individual would develop symptoms and immediately remove themselves from the work place. This may be unrealistic. Multiple attempts have been made to calculate the exact rates of asymptomatic infection for SARS-Cov-2, with estimates ranging from 7.5 - 30.7%[134], [172], [173]. This remains an important data gap[174].

With regards to the handover period, whilst we count the total number of handovers, not all handovers are the same. They will vary in length and number of susceptible individuals attending from day to day, with a possible weekend effect. Additionally, if the individual developed symptoms during a shift and needed to go home, they would still need to handover details to a colleague. This is assumed not to happen in our model. It is difficult to know how someone might like to prioritise the number of handovers attended whilst infectious compared to the total length of time spent at work whilst infectious. Depending on the level of emphasis on cohorting and therefore the importance of handover time and total work time in the model, a score combining the to measures could be calculated.

Finally, we assume that length of time at work is analogous to number of work place infections. This may depend on the nature of the work involved. There may not be a linear relationship between time and number of infectious contacts made depending on the workplace contact pattern. If we look at a hospital ward for example, a ward round would result in an increased period of patient and clinician interaction and therefore a greater number of infectious contacts. Additionally, although we can see the mean length of time at work, this does not give us a complete distribution of $T_w$. A long tail may result in "super-spreader" events while a homogeneous spread would result in consistent transmission rates. This variability is not captured by the mean alone.

We have shown that our understanding of a disease process and parameters can be used to better optimise shift patterns to reduce transmission of a disease between staff whilst working the same total hours in a week. In this case, working longer shifts for fewer shifts in a row decreases the length of time a person would be at work and in handover infectious before developing symptoms. If the staff is not cohorted by shifts, handover may not be a concern, unless it has been demonstrated to be a time of increased disease transmission. However, with cohorting, handover would be a critical time to consider reducing the risk of transmission. One method of doing this could be looking at the shift patterns, as suggested by this investigation. We have shown that the effect on changing shift patterns on infectious handover time is not as uniform as it is on infectious work time. Therefore, caution should be practiced if trying to optimise shift patterns in this way.

Another important caution to consider is the ethical implications of changing rota patterns. We have shown an optimum solution is to lengthen work shifts. However, this may not be practical or appropriate given the individual key-worker. There is a big difference between working 8 hours 5 days a week and 10 hours 4 days a week even though on paper this results in the same hours worked. A socio-demographic study of key-workers in France revealed that they were more likely to be women, of a lower level of education and income and non-European immigrant workers. There may be multiple reasons why a key-worker's lifestyle may not be able to accommodate a change in their rota pattern and this should be accounted for before making a change to minimise SARS-CoV-2 impact in the workplace.

The first part of this chapter had to rely on numerical integration to calculate a central estimate for the total length of time at work whilst infectious and the total number of handovers attended during this time period. This can be a very slow technique and for this reason we have only investigated very simplistic rota patterns. It can also lead to numerical inaccuracy and if we look carefully at our figures we can begin to see the resolution in our results. In particular, rather than an expected smooth line in Figure 3.6, the multiple numerical intergrations have resulted in jagged lines. In comparison, the same analysis using the Fourier transform model showed smooth transitions as we changed the timing of the regular tests. Additionally, it was fast enough that we were able to perform the calculation for two tests in one rota pattern at once (Figure 3.16.

In doing so, we have shown an alternative method of calculating the expected time at work whilst infectious given a fixed rota pattern through Fourier transform. We also suggested a method of generating rota patterns that ensures a constant proportion of time $w$ is spent at work given any rota length $T$ and standard shift length $a$ (so long as $a$ is greater than $w$). It is gratifying to see that the conclusions from the previous numerical model and solution obtained through the use of Fourier transforms are largely in agreement. But we could have implemented the rota adjustment in our observational study, so the question becomes: “Other than confirming the conclusions of the observational model, what was the point of the solving the problem with Fourier transforms?”

In short, the major advantage of an analytical, Fourier-transform-based solution is accuracy. Unfortunately, due in part to the change in the calculation of $Y(j)$ depending if $\frac{j}{T}$ is an integer, an exact solution to the infinite sum for $\hat{W}(0)$ cannot be calculated with this structure of the rota pattern. However, we are guaranteed to get closer to the exact result as we sum to a larger value of $j$. In our results we calculated $\hat{W}(0)$ to $j = 100$, but we could have chosen a larger number for greater accuracy. With each increase of $j$ we alter our results less and less. An upper limit for $j$ could be chosen based on the accuracy required (i.e. stopping once increasing $j$ adjusts our results by some minimum).

Conversely, we calculated our observational model by numerically integrating and convolving across multiple Uniform distributions. We could increase our accuracy in this case by dividing time into smaller parts when integrating, but it would be difficult to know by what degree this would increase the accuracy, or, as an inverse of this statement, by what degree we are not representing the structure of our rota and infectious period. For a short enough shift length and a wide enough time-step it is not impossible that a shift could be completely missed in the numerical model. This problem would become more prominent with the “make-up” shifts (for a total rota length, total hours in a week worked and total hours in an unadjusted working day worked such that an integer total number of working days could not be found).

Additionally, increasing the accuracy of the time-steps in our observational model is more computationally taxing than increasing the maximum value of $j$. Increasing the maximum value of $j$ by one only requires adding this additional value to our current total, but increasing the number of time-steps by one will alter the position of all other time-steps, which in turn would all require re-calculating. In summary, with the Fourier solution we have a model whose accuracy we can increase with a linear increase in computational complexity, compared to the observational model, whose accuracy is difficult to be certain of without careful examination and requires complete recalculation when increased.

Another useful improvement obtained using Fourier transforms can be seen in the force of infection expression $\kappa(\tau)$. We wanted a function that expresses the probability that an individual was infectious at a particular time after they were infected. In the observational model we stated that both the latent and infectious periods fell on Gamma distributions with input parameters $\alpha_a, \beta_a$ and $\alpha_b, \beta_b$ respectively (ignoring the possibility of symptom onset resulting in an individual isolating themselves). The probability that an individual is infectious at time $\tau$ given that that are infected at time 0 is given by the following integration:

$$\mathbb{P}\left(\text{Infectious at time } \tau\right) = \int_0^\tau \frac{\beta_a^{\alpha_a}}{\Gamma\left(\alpha_a\right)} x^{\alpha_a - 1} \exp\left[-\beta_a x\right] \times \frac{\Gamma\left(\alpha_b, \beta_b(\tau - x)\right)}{\Gamma\left(\alpha_b\right)} \mathrm{d}x$$

That is, the probability that an individual is infectious at time $\tau$ is equivalent to the probability that they become infectious at time $x$ and then have an infectious period that is at least length $\tau - x$ for all possible values of $x$ between 0 and $\tau$. Unfortunately, this integral does not have an analytical solution and so we had to calculate it numerically, introducing more computational complexity and a greater possibility for numerical error (although, unlike with the rota pattern functions, it is reasonable to assume a relatively smooth curve for this probability function, meaning we could interpolate between calculated points without much concern for decreased accuracy).

We were able to calculate the exact function for the real part of the Fourier transform for this function, $\Re\left[\hat{\kappa}(\omega)\right]$ (the imaginary part was not relevant to our calculations, although could similarly be calculated). We started by considering it as a simple binary function which would only equal one during an individual's infectious period. Whilst this function poses challenges to express analytically, its Fourier transform is well described and through the linearity for Fourier transforms we could convert this function to express its expected value given Gamma distributions of the latent and infectious periods. This could be exactly combined with the Fourier transform for the time of infection given the rota pattern through multiplication rather than linear integration, once again reducing loss of accuracy.

But $\kappa(\tau)$ is just one of many functions we could have used in this model. When we initially defined the model, we were instead looking at $\lambda(\tau)$, which gave an individual's infectiousness at time $t = \tau$ given that they were infected at time $t = 0$. We could choose any function to represent an individual's changing force of infection and, as discussed earlier, the very constraints that would define it would mean that its Fourier transform must exist and therefore it can be incorporated into our model.

A final advantage of the Fourier model is how easy it is to adapt the rota. Any rota can be broken down into two vectors, **m** and **n**, denoting the starts and ends of shifts respectively and $T$, the length of the rota pattern. The resulting function $Y(j)$ is easy to adapt for any change in the rota pattern, be it an extra shift or a change in the length of the pattern. Adding an extra shift in the observational model involved including a new Uniform distribution to convolve, creating new room for numerical errors and increasing the computational complexity.

There are, however, two notable advantages to the observational when compared to the Fourier model. The observational model is easier to adapt to show the distribution of in-work infectious periods, rather than a central estimate. We are already numerically integrating across all possible outcomes to find the central estimate, so it does not require a large adaptation of our calculations to instead calculate an approximation of the cumulative density function and therefore a complete description of the in-work infectious period distribution. Whilst we can calculate the central estimate through Fourier transform with relative ease, gaining any other information about this distribution through similar analysis is analytically difficult.

The other major difficulty is including regular testing into the Fourier transforms approach. Through the observational model we were able to show that testing so that results would come back at the very start of the rota pattern would minimise the central estimate for the in-work infectious period. It would have been useful to prove this observation through some equivalent model based on Fourier transforms. However, this does not seem to be easy to achieve.

In the observational model, including a testing regime was as simple as introducing a fixed time point where there was a probability that anyone who was infectious would be removed from the system. As we were integrating numerically, it did not matter that this fundamentally changed the structure of the probability that an individual would be infectious at a particular time, as each case could be considered on its own.

The Fourier transforms approach relies on a fixed structure for the force of infection function $\lambda(\tau)$, or the probability of infectiousness function $\kappa(\tau)$. These function are independent of the time an individual is infected, meaning so too are their Fourier transforms. Their convolution becomes a simple matter of multiplication in the frequency-space. It is also because of this fixed structure that we can take advantage of the fact that the rota pattern influences both the time of infection and when an infectious individual would be at work, meaning we could substitute $Y(j)$ in our calculations.

Introducing regular testing removes this independence. An individual's force of infection and the probability that they are infectious would fundamentally depend on the timing of the regular swabs relative to their infection onset meaning the shape of these functions would be inextricably altered by timing of infection onset. Convolution of the infection onset distribution $R(t)$ and the force of infection distribution would no longer be simply equivalent to multiplication in the frequency domain.

This is a shame, because regular testing would seem to have the same cyclical properties as work rotas. Theoretically there exists a Fourier transforms approach that could incorporate regular testing, and even optimise for multiple tests across one rota pattern. A more complicated rota pattern may be more suited to a Fourier method of analysis than through observational modelling. However, this is beyond the scope of this investigation and leaves room for future research.

In summary, we have shown an alternative approach to investigating estimating an individual's time at work whilst infectious by taking advantage of the cyclical nature of rota pattern. We were able to corroborate observations made in the previous chapter, whilst improving the accuracy of the model used. To take this work further, one could look into including testing into the model.

### 3.5.1 Exploring an analytical reason for Rule 2

In our investigation of the effect of the rota pattern, one interesting conclusion we came to was that, for a given disease parameter set, there appeared to be an approximate optimum length of rota pattern to minimise the length of time an individual spent at work whilst infectious. Additionally, the minimum becomes more prominent for rota patterns with longer individual shifts, but remains fairly consistent. We called this observation Rule 2. For the parameter set we investigated, which was an approximate representation of the infectious career of an individual infected with SARS-CoV-2, we found that this optimum fell around approximately 10-11 days. However, our method of analysis provides us with little insight as to why this minimum occurs, why it becomes more prominent for longer individual shifts and why for our particular data set it was at around 10-11 days. In this section, we aim to drastically simplify our model in order to justify these observations.

We define a rota pattern of length $T$ where the total proportion of time spent at work during the entire rota pattern is $w$. For example, if an individual works a 24 hour week, $w$ will equal $\frac{1}{7}$ regardless of the value of $T$. The total time worked is $Tw$.

On a given working day, the proportion of the day an individual actually works is $a$. $a$ must be greater or equal to $w$ in order for it to be possible to have worked a total of $Tw$ over $T$ days. The individual starts work at the start of the rota pattern at time $t = 0$. They work all their shifts until they have worked a total of $Tw$ and then rest until $T$ and their work rota starts anew.

So far, this is very similar, if not effectively identical to the model discussed in the body of this chapter. The first change that we make to the model is to ignore shifts entirely. Rather than considering the rota as a series of days at work, with rests between shifts, in our abridged model, at the start of the rota our individual is constantly at work, accumulating work time at a rate of $a$ proportional to the time actually worked. Once they have accumulated a total of $Tw$ hours they leave work and do not return until $T$, when the rota pattern starts again. Figure 3.18 demonstrates how work hours are accumulated in this new simplified model, using a 9-5 rota pattern as an example.

If they accumulate work-time at a rate of $a$, it will take them until $t = \frac{Tw}{a}$ to accumulate a total work-time of $Tw$. This means in any given rota, they work until $\frac{Tw}{a}$, at which point they will be off until $T$. If they are infected in a given rota pattern that started at time $t = 0$, then without the structure of their rota pattern, the distribution for the time at which they were infected is a simple Uniform distribution $T_e \sim U\left(0, \frac{Tw}{a}\right)$.

We previously modelled both the latent period and infectious period according to

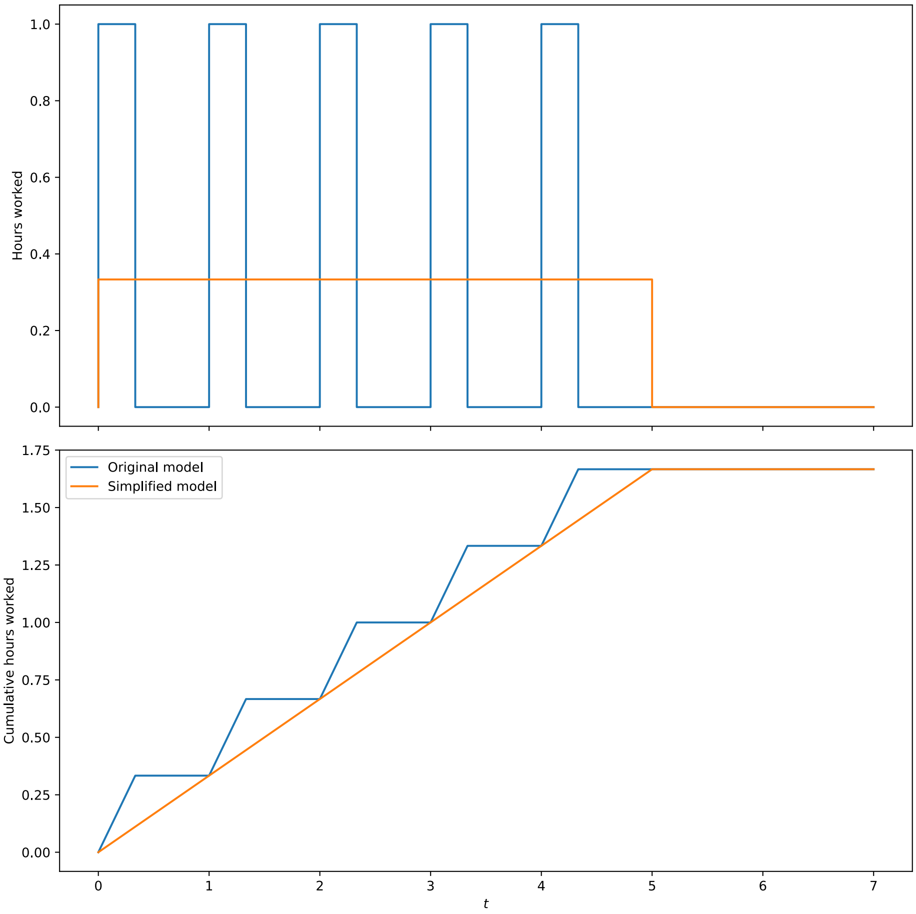


Figure 3.18. A graphical representation of simplified rota pattern model. In this case, we are looking at a 9-5 rota pattern starting at time $t = 0$. The top graph shows the rate at which work hours are being accumulated in the original model (blue) and the simplified model (orange) and the bottom graph shows the cumulative hours worked. By time $t = 5$, both models have worked the same length of time.

some probability distributions. However, we are now going to simplify things further. We shall say that both the latent period and infectious period are fixed at value of $x$ and $y$ respectively. This means for a given value of $T_e$, we can calculate the exact length of time an individual spends at work whilst infectious. Figure 3.19 demonstrates how this could appear if an individual was infected exactly halfway through their working shift $\left(T_e = \frac{Tw}{2a}\right)$. In this figure, the blue window shows the time the infected individual spends at work prior to their rest period. The orange window shows the length of time they spend at work whilst infectious after they return from their rest.

We are going to assume that $x$ and $y$ are large enough such that they definitely have not completed their infectious career prior to starting their rest period $\left(T_e + x + y > \frac{Tw}{a}\right)$. Conversely we assume that $x$ is not large enough that an individual could not have started their infectious period prior to the start of the second rota pattern $(T_e + x < T)$. Finally, we will also assume that $x$ and $y$ are not so large as to still be infectious going into a second rest period $\left(T_e + x + y < T + \frac{Tw}{a}\right)$. This means we only need to focus on the blue and orange windows in Figure 3.19.

The length of time accumulated at work after becoming infectious but prior to en-

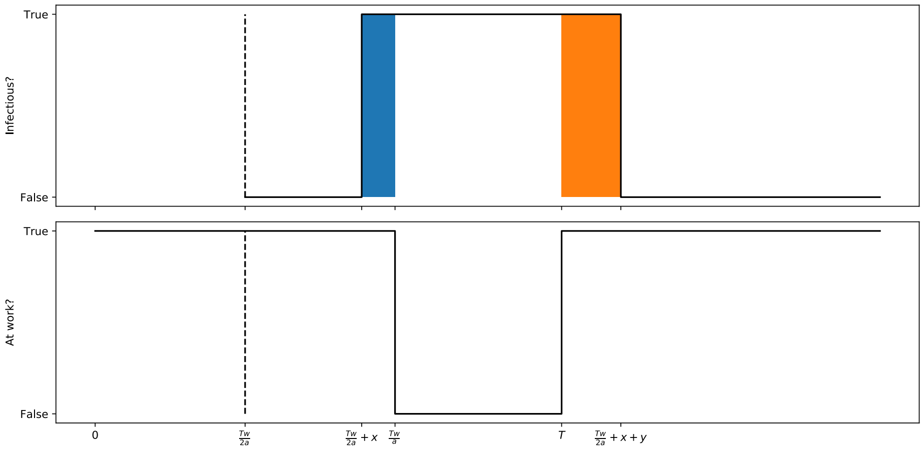


Figure 3.19. Simplified demonstration of possible in-work infectious windows. The individual is infected at time $\frac{Tw}{2a}$, represented by the vertical dashed line. The top graph shows the time at which the infected individual is infectious, while the bottom graph indicates when they are at work. The blue and orange windows show when these two time periods cross over.

tering the rest period (the blue window) is equal to the total length of the shift, $\frac{Tw}{a}$, minus the time it takes to become infectious, $T_e + x$, multiplied by $a$. Remember we multiply by $a$ because time in our simplified model time at work is accumulated at a rate of $a$. This is contingent on the individual becoming infectious prior to the rest period.

The length of time accumulated at work after coming back from the rest period but before the infectious window is over (the orange window) is equal to the time that the infectious window is complete, $T_e + x + y$, minus the start of the second rota cycle, $T$, again all multiplied by $a$. This is contingent on the individual still being infectious once the second rota pattern starts.

These two calculations allow us to write a conditional formula for $T_w$, the total time at work accumulated while infectious:

$$T_w = a \times \left( \left[ \begin{cases} \frac{Tw}{a} - (T_e + x)\,, & \frac{Tw}{a} > T_e + x \\ 0 & \text{otherwise} \end{cases} \right] + \left[ \begin{cases} T_e + x + y - T, & T_e + x + y > T \\ 0 & \text{otherwise} \end{cases} \right] \right)$$

As the probability distribution for $T_e$ falls on a Uniform distribution with a width of $\frac{Tw}{a}$, we can integrate through all possible values of $T_e$ to calculate a mean value of the time accumulated in the blue and orange windows. First though, we should clarify that if $x$ is greater than $\frac{Tw}{a}$ then no time can be accumulated in the blue window. The latent period would be too long so that even if the individual is infected right at the start of the rota pattern they would not be infectious before their rest period. Similarly, if $T$ is greater than $\frac{Tw}{a} + x + y$ they could never accumulate time in the orange window, Even if they were infected at the very last moment of their shift, they would already be no longer infectious by the time they return to work. The total expected time at work whilst infectious becomes the following:

$$\text{Blue area:} \quad \begin{cases} \frac{a^2}{Tw}\int_0^{\frac{Tw}{a}-x} \frac{Tw}{a} - (z+d)\mathrm{d}z, & \frac{Tw}{a} > x \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Orange area:} \quad \begin{cases} \frac{a^2}{Tw}\int_{T-y-x}^{\frac{Tw}{a}} z + x + y - T\mathrm{d}z, & T < \frac{Tw}{a} + x + y \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}\left[T_w\right] = \left[\begin{cases} \frac{(Tw-xa)^2}{2Tw}, & T > \frac{ax}{w} \\ 0 & \text{otherwise} \end{cases}\right] + \left[\begin{cases} \frac{(T(w-a)-a(x+y))^2}{2Tw}, & T < \frac{a(x+y)}{a-w} \\ 0 & \text{otherwise} \end{cases}\right] \tag{3.9}$$

Rule 2 stated that there appeared to be a length of rota pattern $T$ that resulted in a minimum length of time whilst infectious. We can find this minimum value in our simplified model first by differentiating $\mathbb{E}\left[T_w\right]$ with respect to $T$:

$$\frac{\mathrm{d}\mathbb{E}\left[T_w\right]}{\mathrm{d}T} = \left[\begin{cases} \frac{w}{2} - \frac{a^2x^2}{2T^2w}, & T > \frac{ax}{w} \\ 0 & \text{otherwise} \end{cases}\right] + \left[\begin{cases} \frac{(T(a-w))^2-(a(x+y))^2}{2T^2w}, & T < \frac{a(x+y)}{a-w} \\ 0 & \text{otherwise} \end{cases}\right] \tag{3.10}$$

The effect the rota length has on the expected length of time at work whilst infectious is dependent on two conditions: if $T > \frac{ax}{w}$ or if $T < \frac{a(x+y)}{a-w}$. It is possible for both or neither to be true, which means we now need to investigate four possible states:

1. Only the left hand condition is true: $T > \frac{ax}{w}$, $T \geq \frac{a(x+y)}{a-w}$
2. Only the right hand condition is true: $T \leq \frac{ax}{w}$, $T < \frac{a(x+y)}{a-w}$
3. Neither condition is true: $T \leq \frac{ax}{w}$, $T \geq \frac{a(x+y)}{a-w}$
4. Both conditions are true: $T > \frac{ax}{w}$, $T < \frac{a(x+y)}{a-w}$

If only the left hand condition is true then

$$\frac{\mathrm{d}\mathbb{E}\left[T_w\right]}{\mathrm{d}T} = \frac{w}{2} - \frac{a^2x^2}{2T^2w}$$

By rearranging the condition on $T$, we find

$$\frac{w}{2} > \frac{a^2x^2}{2T^2w}$$

The gradient of this slope must be positive. This means that there is a smaller value of $T$ that will result in a smaller estimated total time at work whilst infectious.

If only the right hand condition is true then

$$\frac{\mathrm{d}\mathbb{E}\left[T_w\right]}{\mathrm{d}T} = \frac{(T(a-w))^2 - (a(x+y))^2}{2T^2w}$$

Once again, we can rearrange the condition on $T$ to find out if this gradient is positive or negative (to be specific, when we are rearranging, we are assuming that all parameters are positive, in keeping with the model):

$$(T(a-w))^2 < (a(x+y))^2$$

The gradient is negative, meaning that for any value of $T$ that fulfills this criteria, there is a greater value that results in a smaller estimated value of $T_w$.

If neither condition is true, then the gradient is in fact 0. It may be initially unclear if this represent a minimum value of $\mathbb{E}[T_w]$ or a maximum. However, with careful consideration we can resolve this problem. If neither condition is true then:

$$\frac{a(x+y)}{a-w} \leq T \leq \frac{ax}{w}$$

If we decrease $T$ to the point where is it less than $\frac{a(x+y)}{a-w}$, this generates a situation where the right hand condition is true but the left hand condition is false. The gradient is negative, so any further decrease in the value of $T$ will result in an increase in the value of $\mathbb{E}[T_w]$. Similarly, if we were to increase the value of $T$ such that it becomes greater than $\frac{ax}{w}$ then the gradient becomes positive. Any further increase in the value of $T$ increases the estimated value of $T_w$. Therefore, if $\frac{a(x+y)}{a-w} < \frac{ax}{w}$, then the values for $T$ that results in a minimum value of $\mathbb{E}[T_w]$ are any values in this region.

Alternatively, we can observe what happens to the value of $\mathbb{E}[T_w]$ when neither equality is true from Equation 3.9. In this case, the estimated total length of time at work whilst infectious would be 0. Since this value must be non-negative, a value of 0 must represent a minimum.

Perhaps the most complicated to understand possibility is if both the left and right hand conditions are true. In this case we need to solve a polynomial equation to find when the gradient is equal to 0:

$$\begin{aligned}
\frac{\mathrm{d}\mathbb{E}[T_w]}{\mathrm{d}T} =& \frac{w}{2} - \frac{a^2x^2}{2T^2w} + \frac{(T(a-w))^2-(a(x+y))^2}{2T^2w} && = 0\\
& T^2w^2 - a^2x^2 + T^2(a-w)^2 - a^2(x+y)^2 && = 0\\
& T^2 - \frac{a^2(2x^2+2xy+y^2)}{a^2-2aw+2w^2} && = 0\\
& \left(T + a\sqrt{\frac{2x^2+2xy+y^2}{a^2-2aw+2w^2}}\right)\left(T - a\sqrt{\frac{2x^2+2xy+y^2}{a^2-2aw+2w^2}}\right) && = 0\\
T =& \pm\, a\sqrt{\frac{2x^2+2xy+y^2}{a^2-2aw+2w^2}}
\end{aligned}$$

We now have two possible minima for $T$. However, as we are only interested in positive values of $T$, we should scrutinise further to see if we can narrow our options down. Looking at the numerator in the fraction term, $2x^2 + 2xy + y^2$, this is certainly positive. The denominator, $a^2 - 2aw + 2w^2$, maybe more difficult to determine. However, if we realise that it is equal to $a^2 - 2aw + w^2 + w^2$ and that $a^2 - 2aw + w^2 = (a - w)^2$, then we can be certain that the denominator is also positive. In turn, we are only interested in the following value of $T$:

$$T = +a\sqrt{\frac{2x^2 + 2xy + y^2}{a^2 - 2aw + 2w^2}}$$

But does this minimum occur between our two thresholds? Proving

$$\frac{ax}{w} < a\sqrt{\frac{2x^2 + 2xy + y^2}{a^2 - 2aw + 2w^2}} < \frac{a(x + y)}{a - w}$$

seems like a fairly daunting task. Added to this, we have not actually shown that this is a minimum, only that it is an extremum. Instead, we will prove that there must be a minimum between the two thresholds, and as this is the only extremum that fulfills this criteria, it must be a minimum.

First, let us consider the gradients at the two thresholds, $T = \frac{ax}{w}$ and $T = \frac{a(x+y)}{a-w}$, when $\frac{ax}{w} < \frac{a(x+y)}{a-w}$. In the first case, the gradient with respect to $T$ is negative. The left hand condition of Equation 3.10 is false and so its contribution is equal to 0. The right hand condition is true and so its contribution is negative. What is useful to note is that even if we consider the left hand side to be true by relaxing the condition to $T \geq \frac{ax}{w}$, then inserting our value for $T$ into the calculation for the left hand side of this equation still results in 0. This part of the equation starts at 0 and smoothly transitions into being positive. This is important as it means this threshold does not represent a discrete change in the gradient but a smooth one. The gradient will remain negative immediately after this threshold.

A similar conclusion can be drawn when inserting $T = \frac{a(x+y)}{a-w}$ into Equation 3.10. This time, it is the right hand part of the equation that is equal to 0, regardless of if we take the threshold as $T < \frac{a(x+y)}{a-w}$ or $T \leq \frac{a(x+y)}{a-w}$ while the left hand side is true, resulting in a positive overall gradient. Again, as the right hand side is equal to 0 at this threshold, as we decrease the value of $T$ there must be a smooth, continuous increase in the value of the right hand side of the equation. The gradient immediately before the threshold must also be positive.

So, passing through this region, the gradient with respect to $T$ changes from negative to positive. In order for this to occur with a continuous curve, there must be at least one minimum in this region. Since $T = a\sqrt{\frac{2x^2+2xy+y^2}{a^2-2aw+2w^2}}$ is the only extremum that can fall in this region (the other option is negative), we know that it must fall in this region, and that it must represent a minimum.

We now have two possible solutions for a value of $T$ that results in a minimum value for the length of time spent at work whilst infectious. If

$$\frac{ax}{w} < \frac{a(x+y)}{a-w},$$

which rearranges to

$$\frac{a}{w} < 2 - \frac{x}{y},$$

then it is possible for both the left and right hand conditions to be true, meaning that the minimum value comes at $T = a\sqrt{\frac{2x^2+2xy+y^2}{a^2-2aw+2w^2}}$. Conversely, if $\frac{a}{w} \geq 2 - \frac{x}{y}$, then it is possible for neither left nor right hand condition to be true. When this occurs, $T_w = 0$ and the infectious individual spends no time at work whilst infectious. In this case the minimum value of $T$ is the region $\frac{a(x+y)}{a-w} \leq T \leq \frac{ax}{w}$.

These inequalities enable us to make two interesting observations by considering what the fractions $\frac{a}{w}$ and $\frac{x}{y}$ represent. The fraction $\frac{a}{w}$ is the proportion of time in a working day spent working over the proportion of time in a working week spent working. This must be greater or equal to 1, as in order to maintain a certain total proportion of time worked, $w$, the proportion of time worked on a working day, $a$, must be at least this value. In the case where $a = w$, the working individual would have to work every day of the rota pattern. In our simplified model there would be no minimum to be found, as the individual would always be at work.

The fraction $\frac{x}{y}$ is the length of an individual's latent period over their infectious period. It must be greater than 0, but can otherwise be any value, as $x$ and $y$ put no constraints on each other. As neither fraction can be negative, we can find two situations where the truth of the inequality is only dependent on one of the fractions.

Firstly, as $\frac{x}{y} > 0$ if $\frac{a}{w} > 2$, the inequality must be false. This means if the proportion of time spent working in a day is more than twice the total proportion of time spent working, then the optimum value of $T$ falls in a region between $\frac{a(x+y)}{a-w}$ and $\frac{ax}{w}$, regardless of the disease parameters.

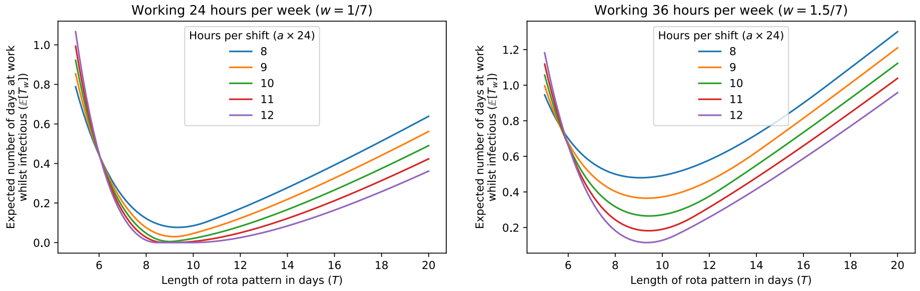
Secondly, as $\frac{a}{w} \geq 1$, if $\frac{x}{y}$ if greater than 1 then the inequality must also be false. This means if the latent period is longer than the infectious period then the optimum value for $T$ again falls in a region between $\frac{a(x+y)}{a-w}$ and $\frac{ax}{w}$. In both cases, for a large enough value of either $\frac{a}{w}$ or $\frac{x}{y}$, the inequality can never be true, regardless of the value of the other fraction.

As a final step, we can see how this simplified model compares to the output from the model in the body of the text. We investigated working a total of 24 and 36 hours a week ($w = \frac{1}{7}$ and $\frac{1.5}{7}$ respectively) with a daily shift length between 8 and 12 hours ($\frac{1}{3} \leq a \leq \frac{1}{2}$). We can take the mean values for the latent and infectious periods:

$$\begin{aligned} x =&2.84 \\ y = 6 \times 0.3 + 2 \times 0.7 =&3.2 \end{aligned}$$

(see Table 3.2).

Figure 3.20 shows the estimated length of time at work whilst infectious for our simplified model. As predicted in our analysis, this metric forms a minimum. If the length of the rota pattern $T$ is either too long or too short, $\mathbb{E}\left[T_w\right]$ increases. As $a$ increases, the minimum value of $\mathbb{E}\left[T_w\right]$ decreases to a lower limit of 0, at which point the minimum becomes a region.



(a)

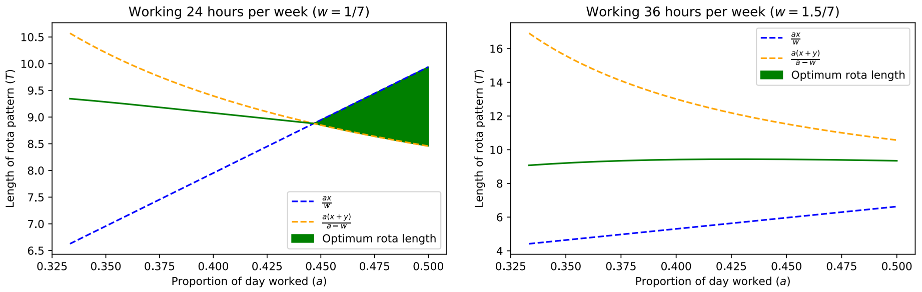
(b)

Figure 3.20. Estimated accumulated time at work whilst infectious in the simplified model. The left hand plot shows how this metric changes if an individual accumulates 24 hours of work a week ($w = \frac{1}{7}$), where the right hand plot shows the same metric for a 36 hours working week ($w = \frac{1.5}{7}$). In order to see the shape and minima nore clearly, we have decided to show this parameter through line plots rather than contour plots.

Figure 3.21 tracks the changing optimum value of $T$ for our simplified model across this parameter set. When this optimum is a single value (i.e. when $\frac{a}{w} < 2 - \frac{x}{y}$) this optimum remains fairly constant, approximately equal to 9 days. This is roughly in keeping with our observation in the body of the text of a minimum at between 10 and 11 days. However, when the optimum value of $T$ is instead a region ($\frac{a}{w} > 2 - \frac{x}{y}$), this region, whilst including our previously estimated optima, becomes relatively wide when compared to the variation in the minimum for lower values of $a$.

Through careful analysis of this simplified model we have been able to justify Rule 2 from the body of the text. We have found an analytical reason for an optimum length of rota pattern forming, that appears to be fairly constant as we vary the length of individual shifts. This reason is the trade-off between lengthening the rota pattern so that an infected individual is recovered by the time their next cycle starts and shortening the rota pattern so they are already in a rest period by the time they become infectious. Unintentionally, by looking at Figure 3.20 we have found further evidence for Rule 1 as well. As we increase the rate at which an individual accumulates work hours ($a$) we decrease $\mathbb{E}\left[T_w\right]$.

There are some notable differences between the results of our simplified model and



(a) (b)

Figure 3.21. Optimum rota lengths in a simplified rota model. The blue and the orange dotted lines show the left and right hand conditions from Equation 3.10 respectively. In turn the green line/region shows the value of $T$ that will minimise $\mathbb{E}\left[T_w\right]$. For any value of $a$ where $\frac{ax}{w} < \frac{a(x+y)}{a-x}$ (where the blue line is higher than the orange line), the optimum rota length takes a single value. However, past the threshold where this inequality is no longer true (where the blue and orange lines cross), the optimum value of $T$ is the region between these 2 lines. This does not occur on the right hand plot, because the value of $w$ is too high. If we extended the x-axis far enough we would expect to see these lines cross.

the numerical model in the body of the text. Of course, in our more complicated model, as in real life, it is very unlikely to find a rota pattern that will result in an individual spending exactly no time at work whilst infectious. By integrating through every possible latent, prodromal and infectious period we should always find some combination that results in an amount of infectious work-time. This may go some way to explain why we consistently underestimated the length of the optimum rota pattern in our simplified model. Another reason may well be our assumptions that neither the possibility of ending one's infectious period before the end of the first week's work, nor the possibility of not yet being infectious before the second week had begun, needed to be accounted for in our model (see Page 167). Both were kept out for analytical simplicity, but either one may affect the true location of the optimum value of $T$.

Overall, this simplified model has helped us understand a little more clearly the effect changing our parameter set has on our model. If we wanted to repeat this analysis for a different disease model, this simplified version may be a good place to start.

Summary:

We have taken two approaches to write a model that approximates the total expected length of time an infectious individual will spend at work given that they were infected at work. Through these models we have shown three key concepts to take away from these analyses. The first is that a rota pattern with longer, fewer shifts results in a shorter length of time at work whilst infectious. Secondly, there is optimum length of rota pattern when it comes to minimising the length of time at work whilst infectious. Finally, the optimum time to arrange regular testing is such that results from the test would come back at the start of the rota pattern.

Being able to express this model in terms of a Fourier transform opened up many opportunities for future analysis. In particular, it would be interesting to incorporate these findings into a full transmission model. It has also revealed the importance of rota patterns when considering work-place outbreaks, and suggests some mitigation that can be made in the future.

# Chapter 4

# Nosoco: A tool for up-to-date monitoring of nosocomial infections across a hospital

## 4.1 Introduction

Even prior to the SARS-CoV-2 global pandemic, nosocomial infections, meaning the transmission of an infectious disease in a healthcare setting[175], were a growing concern, with estimated total of 3.2-6.5% of all inpatients suffering from a hospital acquired infection at some point during their stay[176], [177], and this number potentially being higher in developing countries[178]. As they represent a collection of disparate infections, their aetiology and causative organisms vary depending on the infection. Possible nosocomial infections include, but are not limited to, lower respiratory tract infections resulting in pneumoniae, catheter related urinary tract infections, surgical site infections and central line infections. Through a point-prevalence study 2015, Magell et al. indicated that lower respiratory tract infections may be the leading nosocomial infection[176].

Once an infection is identified, defining it as nosocomial in origin is dependent mostly on history, in particular how long before a patient has been admitted to hospital (surgical site infections are an obvious exception to this). Individual guidance is available for treatment depending on the causative organism, although, in the case of bacterial infections, broad-spectrum antibiotics are often recommended owing to the high rate of multi-drug resistant nosocomial infections[179], [180]. This resistance, alongside an apparent raised mortality[181], [182], results in a considerable burden from nosocomial infections on hospitals.

According to Public Health England (PHE) guidance, in the early pandemic, a new diagnosis of SARS-CoV-2 infection in hospital was classified as a likely healthcare associated (nosocomial) infection should symptoms of the disease develop more than 8 days after admission and a definite healthcare associated infection if they start 14 or more days after admission. Additionally, an outbreak is defined as two or more such infections occurring in a single location (such as a ward)[183]. This robust method of identification leaves no room for doubt and allows for rapid identification of healthcare associated outbreaks. This advice continues to be the prevailing school of wisdom regarding SARS-CoV-2 nosocomial transmission identification, although earlier thresholds have at times been chosen for a higher sensitivity when detecting nosocomial transmissions.

Using the PHE definition of nosocomial Covid-19 infections, it has been estimated that currently nosocomial transmission represents approximately 17.6% of all transmission in England[184]. Zhou et al. conducted a meta-analysis of nosocomial outbreaks of similar coronavirus diseases (SARS and MERS)[185]. They estimated that in the early outbreaks of SARS-Cov-2, SARS and MERS, nosocomial transmissions contributed to as much as 44%, 36% and 56% of total transmissions respectively. They also noted the significant strain that hospital infections can cause through staff absences.

Outside of the ethical implications of iatrogenic infections, one reason the correct identification of nosocomial infections is important is their outcomes when compared to equivalent community acquired infections. On the one hand, it may be expected that patients who are infected in hospital will tend towards being more frail or having more co-morbidities (after all, they must at least already have one disease that has resulted in their hospitalisation in the first place) meaning they could be more vulnerable to severe sequelae of the infection in question. Conversely, it may be that a patient who is infected during their admission is more likely to be identified than if they had been infected with the same disease in the community due to regular screening, symptom monitoring and direct access to testing. If this were the case we could expect outcomes to appear to be worse for those infected outside of hospital. This external group would be more vulnerable to presentation bias, as individuals outside of hospitals who only experience mild-to-low may never appear the a dataset, where the equivalent individuals in hospital may be tested as a precaution to prevent further spread of disease.

Khan et al. performed a prospective cohort study comparing the 30-day mortality of those who were admitted due to SARS-CoV-2 infections, those who were admitted for unrelated reasons and coincidentally found to be SARS-CoV-2 positive and those who developed the infection at least seven days into their admission. This data was collected from three hospitals across one trust in Scotland, identifying 173 patients in total. They found no difference in the mortality of these three groups[186]. Conversely, the COPE-Nosocomial Study, an observational study performed in UK hospitals and one Italian hospital, identifying 1564 patients in total, found that nosocomial transmissions resulted in overall lower mortality rates when compared to external transmission, but no change in mortality across the initial seven days of infection[187]. Although the Khan et al. study did separate coincidental admissions with SARS-CoV-2 from admissions from SARS-CoV-2, neither study appeared to examine the biases inherent in monitoring infection rates in hospitals.

Whilst the exact nature of outcomes from nosocomial transmission of SARS-CoV-

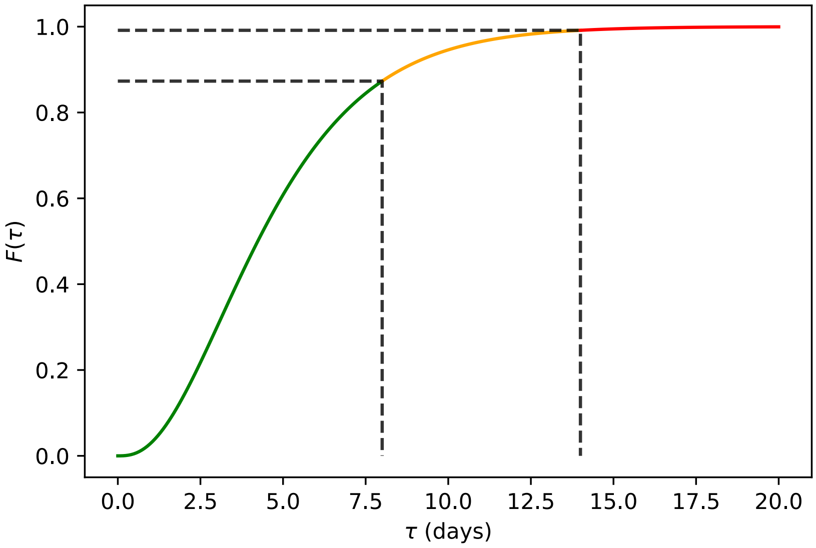


Figure 4.1. Cumulative density function ($F(\tau)$) for a SARS-Cov-2 incubation period represented by a Gamma distribution with a mean of 4.84 days and a standard deviation of 2.79 days. The grey dotted lines shows the cut-off point for nosocomial transmission as suggested by PHE of 8 and 14 days. Any person developing symptoms on their 8th day of admission has a 0.872 chance of have a incubation period shorter than their admission time, which, given an equal probability of infection inside and outside of hospital, is equivalent to saying they have a 0.872 probability of having been infected in hospital.

2 remains unclear, it is undeniable that it makes up an important proportion of all transmission. Identifying outbreak areas remains vital for understanding transmission methods and preventing further transmission. It may not be sensible, therefore, to rely on the PHE suggested 8-day or 14-day cut-off window alone.

The incubation period between infection and symptom onset can obfuscate the exact timing of infection. Overton et al. suggest that the way in which it varies can be mapped well to a Gamma distribution with a mean of 4.84 days and a standard deviation of 2.79 days[133]. With this distribution, there is approximately a 50% chance that someone developing symptoms was infected under 5 days ago. In the case of an individual developing symptoms five days into their hospital admission, this equates to a 50% chance of having been infected during their admission (assuming equal probabilities of infection inside and outside the hospital environment). Indeed, an individual developing symptoms on day 8, in keeping with a PHE diagnosis of likely nosocomial infection, would have a probability of 0.872 of having had an incubation period short enough that they were infected during their admission (see Figure 4.1).

Given this distribution, it is possible that an individual infected early in their admission may have a incubation period short enough to be assumed to have been infected outside the hospital. Similarly, with a long enough incubation period, an individual infected just before their admission date may be miss-classified as a nosocomial transmission.

Rather than classify an individual as one of the two binary options of having either a nosocomial or external transmission, we have developed a tool called Nosoco that is designed to approximate the probability that an individual was infected before or after their admission based on the timing of their symptom onset. This information on its own would not be informative on an individual basis. Nosoco combines each observed individual's probability of having been infected whilst in hospital across a population to indicate areas where and when it is highly probable that outbreaks have occurred. Although the basis of the mathematics that informs Nosoco is conceptually simple, to the best of our knowledge, no other equivalent tool exists for estimating the impact of nosocomial transmissions.

In Section 4.3 of this chapter, we aim to demonstrate the mathematics behind Nosoco using anonymised data from a multi-site NHS trust. We will show how knowing the timing of symptoms onset and pairing this with the timing of ward admissions can be used to estimate:

1. the total number of nosocomial transmissions occurring on any ward or any hospital
2. the probability that an outbreak may have occurred on a particular ward or hospital, according to PHE guidelines
3. the total rate of infection on a ward or in a hospital

In Section 4.4 we assess Nosoco in terms of its speed and how its results compare to more traditional methods of estimating the total number of nosocomial transmissions through random sampling across the pandemic. In Section 4.5, we will look at how data developed from Nosoco can be used to better understand nosocomial transmissions in general. In particular, we will focus on the information we can gain from estimating the total number of infections occurring in a hospital and comparing this to the total number that have occurred outside of hospital resulting in an admission. In Section 4.6 we will look at how the rate of infection has varied throughout the pandemic, how it varies between age groups, and how our method of calculating the rate of infection affects our conclusions.

## 4.2 Parameter and Function description

Table 4.1 is a list of relevant parameter and function definitions for this chapter. The reader may find it useful to refer back to this table as and when required.

| Parameter/Function | Description |
|---|---|
| $a_j$ | The maximum between time $x$ and the time the $j$th individual was admitted |
| $b_j$ | The minimum between time $y$ and the time the $j$th individual was discharged |
| $\delta_j$ | An indicator function showing if the $j$th individual had a first positive swab after $a_j$ |
| $F(t)$ | The cumulative density function for the incubation period of a disease |
| $g(t)$ | The probability density function for the length of an individual's ward admission |
| $G(t)$ | The cumulative density function for the length of an individual's ward admission |
| $\mathbb{L}(\theta\|x)$ | The likelihood function for parameter $\theta$ given some observation $x$ |
| $\Lambda$ | The rate of nosocomial infections |
| $m$ | The number of simulation iterations |
| $\nu$ | The rate of the Exponential distribution from which $\Lambda$ is drawn |
| $\mathbf{P}$ | The vector of all probabilities that individuals were infected in the observed ward during the observation period |
| $\pi(theta)$ | The prior distribution for some parameter $\theta$ |
| $\mathbf{q}$ | A matrix used to calculate $Q_m$ |
| $Q_m$ | The probability that at least $m$ individuals were infected during the observation period |
| $r$ | The total number of people infected during the observation period |
| $\bar{r}$ | Alpha parameter for the Gamma distribution that represents the central estimate for the posterior distribution for $\Lambda$ |
| $\mathbf{r}$ | The vector of different values of $r$ on simulation |
| $\tau$ | Time from when an individual was infected |
| $\boldsymbol{\tau}$ | The set of all observations $\mathbf{X}$ where $\delta_j = 1$ |
| $t_j$ | The timing of the $j$th individual's first positive swab if $\delta_j = 1$, otherwise a null value |
| $W$ | The total length of all admissions prior to first positive swab during the observation window |
| $\bar{W}$ | Beta parameter for the Gamma distribution that represents the central estimate for the posterior distribution for $\Lambda$ |
| $\mathbf{W}$ | The vector of different values for $W$ on simulation |
| $X$ | The set of all observations, made of vectors $\mathbf{a}$, $\mathbf{b}$, $\boldsymbol{\delta}$ and $\mathbf{t}$. |
| $x$ | The beginning of the observation window |
| $y$ | The end of the observation window |

Table 4.1. A complete description of parameters and functions used in this chapter

## 4.3 An explanation of the methods used by Nosoco

### 4.3.1 Structure

Nosoco is coded in Python as part of a Jupyter notebook. It is programmed to access data from an NHS trust database in the form of two SQL codes. These SQL codes can be changed depending on the trust database arrangement. We chose Python as the coding language owing to how easily the SQL commands in the Pandas package interacted with the particular databases we were extracting data from. We did briefly experiment with converting to C++ but found its SQL interactions less manageable.

We give Nosoco two input dates: a start date and an end date. It is between these two dates that we want to estimate the total number of nosocomial transmissions (as well as other transmission adjacent metrics). The two SQL requests extract hospital data between these two dates as well as over certain wider time frames.

The first SQL query, entitled `All movements.sql` identifies all ward admissions between the two input dates. It also looks at admissions for a time period prior to the requested window (as a default 2 months) to ensure nosocomial transmissions that are diagnosed in the requested window but actually occur prior to the requested window are observed as well. This is not strictly necessary for most standard analyses, but is useful when comparing Nosoco to other methods of identifying nosocomial transmissions. It records the ward name and site, the date at which the patient arrives and leaves, and the patient's name, date of birth, NHS Number and Hospital Number.

The second SQL query, entitled `Just all swabs.sql` collects all SARS-Cov-2 swabs performed between the same initial input dates. It also checks for swabs prior to the observation window, for either 90 days prior to the window or from 01/03/2020 to the start of the observation window, whichever is shortest. This is done to ensure that pre-existing SARS-CoV-2 infections are not identified as new infections during the observation window. It also requests all swabs at most 28 days after the observation period. This is done so that individuals who are infected during the observation window but only diagnosed afterwards are also observed. It records the same demographics of the patient as above, as well as the location of the swab, the timing of the swab (including the time at which the result is made accessible) and the result. It is advised that the input end date is more than 14 days prior to the current date for retrospective investigations as this will exclude individuals who have been infected but are yet to receive a positive test result.

The patients are identified first by their first names, last names and dates of birth. This is an optional function. They are then anonymised, each being assigned a random integer value. If the user wishes to know each patient's true identity (for the

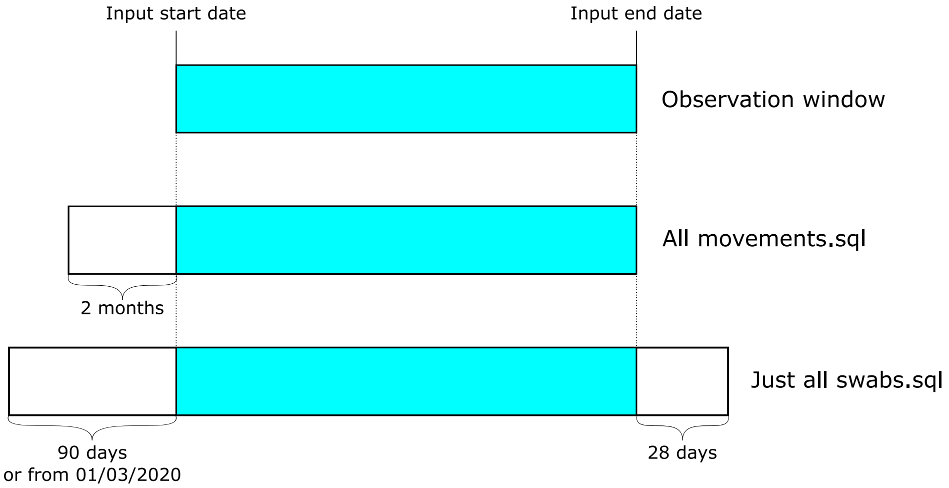


Figure 4.2. The time frames over which Nosoco performs SQL queries. `All movements.sql` checks two months prior to the requested observation window to see which patients were already admitted at the start of the observation window. `Just all swabs.sql` checks 90 days prior to the observation window up until 01/03/2020 to see if any positive cases in the observation window represent pre-existing diagnoses, and 28 days after the observation window to look for patients that were admitted during the observation window but diagnosed with SARS-CoV-2 after the input end date as they may represent nosocomial transmissions during the observation window. After its analysis, Nosoco only returns information regarding nosocomial transmissions it has estimated to have occurred during the requested observation window.

sake of contact tracing, for example) a dictionary that translates these integers back to their patient-identifiable details can be returned. Nosoco collates and orders all admissions and swabs of any patient identified through the SQL queries, removing any overlaps in admissions between wards. This generates two NumPy arrays, one containing the admission data and one containing the swab data. Details about each ward and a collated list of individuals who both appear in the admissions array and have a positive swab result in the swab array are also returned. Patients who have not received a positive diagnosis of SARS-CoV-2 are still recorded, as they represent an important denominator. For certain analyses, it may not be useful to just estimate the total number of nosocomial infections as a larger population size will inherently be more likely to see more nosocomial transmissions. For example, if we were comparing total nosocomial transmissions between age groups, we may see a higher total number of nosocomial transmissions among older populations. It would be unclear from this information alone if this would be due to a true increased risk or because older populations represent a larger proportion to the admitted population.

For each individual who was both admitted and had a positive swab, Nosoco can calculate the probability that they were infected over a particular time period by conflating the timing of their first positive swab with the timing of their symptom onset (see Section 4.4.3). By comparing a time period with how it overlaps with a patient's admission timing, it can also calculate the probability that an individual was infected over a particular time period in a particular location. The observed time period in which data was extracted (01/03/2020 to a chosen date) is divided into even chunks (default length of seven days). For each of these time periods for each ward Nosoco calculates three metrics:

1. Estimated total number of nosocomial infections
2. Probability that the total number of infections is greater or equal to 2 (this number can be changed depending on need)
3. Ongoing rate of infection (estimated through simulation assuming a constant rate of infection on any one ward over each chunk)

For each of these metrics, Nosoco generates heat-maps showing how these estimates change for each ward over time. A ward may be excluded from any heat-map if the probability of any nosocomial infection having occurred there is sufficiently low as to be irrelevant. Finally, it also optionally generates text files for the highest risk wards and time periods, including patients who were likely infected during these time periods, to aid manual investigation.

### 4.3.2 Calculating the metrics used by Nosoco

For any particular ward over a time period between times $x$ and $y$ we can find the set $\mathbf{X}$ of observations such that the $X_j$ represents the observation, between times $x$ and $y$, relative to the $j$th individual who was admitted on said ward. This includes individuals who were admitted prior to time $x$ as well as those who were discharged prior to time $y$, but excludes individuals who had a positive swab prior to time $x$ as we assume that an individual cannot be reinfected (specifically, we assume that they cannot be reinfected in under 90 days, so $\mathbf{X}$ excludes individuals who had a positive swab less than 90 days prior to $x$). The observation $X_j$ consists of four factors:

1. $a_j$ - The maximum between $x$ and the time individual $j$ was admitted to the ward
2. $\delta_j$ - An indicator function that is 1 if individual $j$ has their first positive swab after $a_j$ and 0 otherwise
3. $t_j$ - The time of the $j$th individual's first positive swab if $\delta_j = 1$, otherwise a null value
4. $b_j$ - The minimum between $y$, the time the $j$th individual is discharged from the ward and, if $\delta_j = 1$, $t_j$

We assume that an individual is only infected once, and that their first swab is taken at the time they develop symptoms, ignoring false negative swabs, delay between symptom onset and access to diagnostic tests and asymptomatic screening swabs. We also assume that infection detection is perfect and that the time between our observation period and the current time is sufficient enough that any individual that was infected during this time period would be detected by the time of our analysis (this is why we advise against using Nosoco to analyse a time later than 14 days prior to the current date).

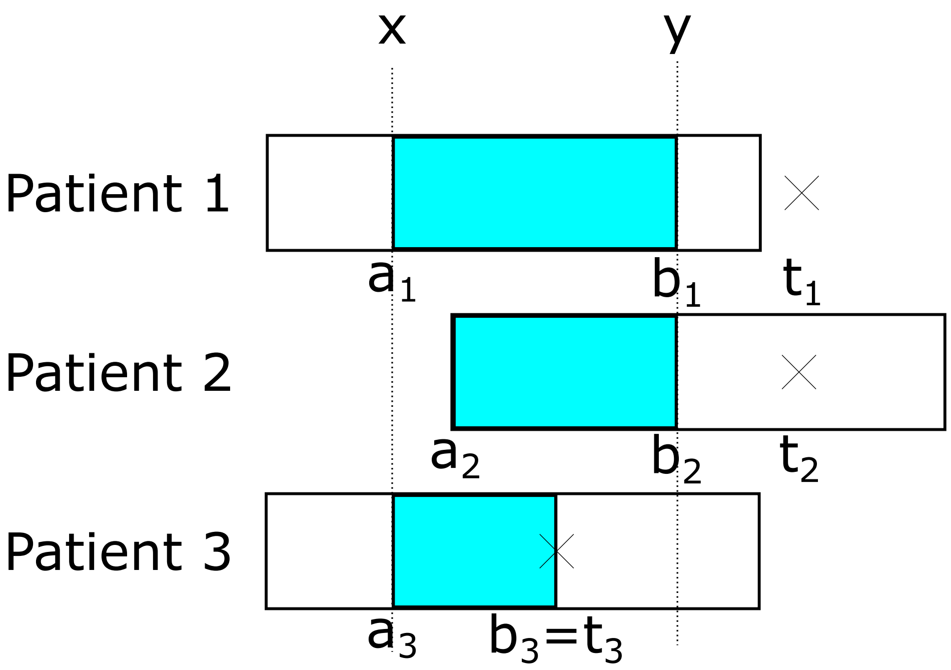


Figure 4.3. Three example observations made by Nosoco between times $x$ and $y$. Each patient's admission is marked by a black box and their first positive swab is marked by a cross (for each of these three patients, $\delta_j = 1$). Their observed period Nosoco considers is marked by a light blue box. Patient 1 was admitted before $x$ and discharged after $y$, so $a_1 = x$ and $b_1 = y$. Patient 2 was admitted after $x$ and discharged after $y$, so whilst $b_2 = y$, $a_2$ is equal to their admission time. Patient 3 was admitted before $x$ and discharged after $y$, but they had their first positive swab between $x$ and $y$. This means that $a_3 = x$ and $b_3 = t_3$, as he could have only been infected before testing positive.

This being the case, we can calculate the probability that an individual who is diagnosed at time $t$ was infected between times $a$ and $b$ as the probability that the same individual would have an incubation period of a length between $t - b$ and $t - a$. We consider the Gamma distributed incubation period from Overton et al. to calculate this probability[133]. Based on observation $X_j$, the probability that the $j$th individual was infected between times $x$ and $y$ can be written as:

$$\begin{aligned}\mathbb{P}(j \text{ infected between } x \text{ and } y) &= \mathbb{P}(j \text{ infected between } a_j \text{ and } b_j)\\ &= \delta_j \left(F(t_j - a_j) - F(t_j - b_j)\right)\end{aligned}$$

where $F(\tau)$ is the cumulative density function of the incubation period distribution. A different incubation period distribution could be chosen for different results. By calculating this probability for all observations in $\mathbf{X}$ we generate a new vector $\mathbf{P}$ of length $\sum_j \delta_j = n$ such that $P_j$ is the probability that the $j$th individual who had a positive swab after time $x$ was infected on the observed ward during the observation period.

The first metric, the estimated total number of nosocomial infections occurring in this time and place, $\mu$, is easily calculated as the sum of vector $\mathbf{P}$. This is equivalent to saying that the central estimate of the total number of infections is equal to the sum of the probability that each individual was infected during this time period. We can also calculate $\sigma^2$, the variance of the total number of infected individuals. In this case, variance means if we were to run simulations of these events based on $\mathbf{P}$, the observed variance of the total number of infected individuals should tend towards $\sigma^2$, as the observed mean tends towards $\mu$.

$$\begin{aligned}\mu &= \sum_{j=1}^{n} P_j \\ \sigma^2 &= \sum_{j=1}^{n} P_j(1 - P_j)\end{aligned}$$

The calculation of the second metric, $Q_m$, the probability that at least $m$ individuals were infected in the observed time and location ($m = 2$ is used over time periods of a week to indicate an outbreak on a ward), is slightly more involved. If our target minimum $m$ is equal to 1, then this is relatively easy. We recognise that the probability that at least one individual was infected is equivalent to one minus the probability that no individuals were infected, making the resulting calculation

$$Q_1 = 1 - \prod_{j=1}^{n}(1 - P_j)$$

However, no such shortcut exists for any value of $m$ greater than 1. Instead we calculate $Q_m$ for $m > 1$ through an iterative process starting with an array $\mathbf{q}$ of shape $(n + 1, m + 1)$. Row $q_1$ is an empty vector except for $q_{1,1}$ which equals 1. Each row corresponds to an inclusion in our calculations of another probability from the vector $\mathbf{P}$: i.e. row 2 contains the probabilities that individual 1 was or was not infected in the time period of interest, row 3 contains the probabilities that neither, either or both individuals 1 and 2 were infected, etc. Each column in $\mathbf{q}$ corresponds to a different total number of nosocomial infections, with the first column corresponding to no infections, the $m$th column corresponding to $m - 1$ infections, and the $(m + 1)$th column corresponding to $\geq m$ infections. We fill $\mathbf{q}$ iteratively, row by row. With each row, we consider a new probability from the vector $\mathbf{P}$. The value of $q_{j+1,k+1}$ is equal to the probability of there having been $k$ nosocomial infections given only the observations up to and including $P_j$. This means, usually, it is equal to the probability of there having been $k - 1$ positive cases prior to the $j$th observation and the $j$th observation being positive or there being $k$ positive cases prior to the $j$th observation and the $j$th observation being negative:

$$q_{j+1,k+1} = P_j \times q_{j,k} + (1 - P_j) \times q_{j,k+1}$$

The exceptions to this calculations are the first and last columns. It is not possible for there to be less than 0 nosocomial infections, meaning $q_{j+1,0}$ is equal to the probability that there were zero positive cases prior to the $j$th observation, and that the $j$th observation is also negative:

$$q_{j+1,0} = (1 - P_j) \times q_{j,0}$$

The final column represents the probability that at least $m$ nosocomial infections, which is equal to the probability $m - 1$ cases had been observed up until observation $j$ and that the $j$th observation is positive, or that at least $m$ positive cases had occurred prior to the $j$th observation, in which case we can disregard $P_j$:

$$\begin{aligned} q_{j+1,m+1} &= P_j \times q_{j,m} + (1 - P_j) \times q_{j,m+1} + P_j \times q_{j,m+1} \\ &= P_j \times q_{j,m} + q_{j,m+1} \end{aligned}$$

The final solution to this iterative process is $q_{n+1,m+1}$, which represents the probability of there having been at least $m$ nosocomial cases given all **P** observations. Table 4.2 shows this iterative process and Figure 4.4 graphically demonstrates how these probabilities can change with each iteration for 10 observation, calculating the probability of at least 5 nosocomial infections having occurred.

| | $q_{*,1}$ | $q_{*,2}$ | ... | $q_{*,m}$ | $q_{*,m+1}$ |
|---|---|---|---|---|---|
| $q_{1,*}$ | 1 | 0 | | 0 | 0 |
| $q_{2,*}$ | $1 - P_1$ | $P_1$ | | 0 | 0 |
| $q_{3,*}$ | $(1 - P_1)(1 - P_2)$ | $P_1(1 - P_2)$<br>$+(1 - P_1)P_2$ | | 0 | 0 |
| ... | | | ... | | |
| $q_{n,*}$ | $q_{n-1,1}(1 - P_{n-1})$ | $q_{n-1,2}(1 - P_{n-1})$<br>$+q_{n-1,1}P_{n-1}$ | | $q_{n-1,m}(1 - P_{n-1})$<br>$+qn - 1, m - 1P_{n-1}$ | $q_{n-1,m}P_{n-1}$<br>$+q_{n-1,m+1}$ |
| $q_{n+1,*}$ | $q_{n,1}(1 - P_n)$ | $q_{n,2}(1 - P_n)$<br>$+q_{n,1}P_n$ | | $q_{n,m}(1 - P_n)$<br>$+q_{n,m-1}P_n$ | $q_{n,m}P_n$<br>$+q_{n,m+1}$ $= Q_m$ |

Table 4.2. The iterative process required to calculate the probability that the total number of infected individuals is greater than or equal to $m$. This value, denoted by $Q_m$, is equal to the final value in the array, $q_{n+1,m+1}$. Each row demonstrates the next step in the process.

The final metric, and definitely the most involved calculation, is estimating the rate of infection on any one ward. This is a form of survival analysis. We want to estimate at what rate individuals would be infected given that they are on a ward indefinitely.

We assume that for a small enough time period, each ward has a constant force of infection $\Lambda$ during that time period. Patients on the ward who are yet to be infected

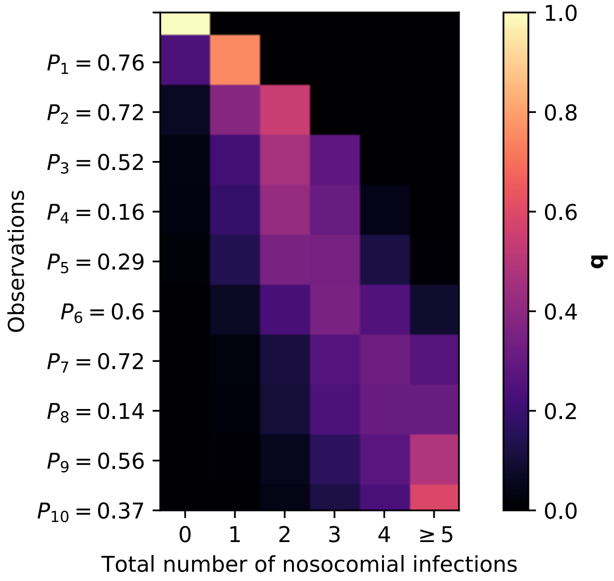


Figure 4.4. An example of the iterative process required to calculate $Q_m$. In this case, we observed 10 probabilities and wanted to know the probability of at least 5 successes. This value is demonstrated by the brightness of the bottom left square ($q_{11,6}$). In this case, $Q_5 = 0.593$. Each row shows how including a new probability changes our calculation.

are all equally exposed to this force of infection. Therefore, their time from admission to infection would follow the same exponential distribution with rate $\Lambda$. We initially want to take a Bayesian approach to calculate the most accurate posterior distribution which describes possible values of $\Lambda$ based on our observations of individuals on the ward. Later, to produce a meaningful heat map, we will have to reduce this posterior distribution down to a single value. In order to calculate a posterior distribution for $\Lambda$, we first need to describe a prior distribution, $\pi(\Lambda)$, and then describe a probability density function for our observations for any given value of $\Lambda$, $\mathbb{P}(\mathbf{X}|\Lambda)$. The posterior distribution for $\Lambda$ is proportional to the multiplication of these two terms:

$$\mathbb{P}(\Lambda|\mathbf{X}) \propto \pi(\Lambda) \times \mathbb{P}(\mathbf{X}|\Lambda)$$

Assume in what follows that we observed the exact time of infection of all individuals, and consider a generic observed individual, with time of infection $t_{inf}$. Then, for any admission between our observed times $x$ and $y$ prior to infection, we can calculate $\tau, \delta$, where $\tau$ is the length of time the observed individual spent admitted on the ward between times $a$ and $b$ prior to infection and $\delta$ is an indicator value such that $\delta = 1$ if the individual was infected during this time period, and $\delta = 0$ if the time period ended prior to infection (either the individual's admission ended or the observation period ended prior to discharge).

$$(\tau, \delta) = \begin{cases} (t_{inf} - a, 1) & \text{if infected during observation period} \\ (b - a, 0) & \text{otherwise} \end{cases}$$

Note that $b$, as defined at the start of this subsection, is the minimum between the end of the time interval of interest ($y$), the time of discharge, and the time when the individual has their first positive swab.

Because we have assumed we know $t_{inf}$, we can simplify our considerations by disregarding the time of first positive swab and redefining $b$ to be the minimum between $y$ and the time of discharge. The reason is that $t_{inf}$ is always earlier than the time of the first positive swab, so either the individual is infected during the time interval under consideration and we need not be concerned with what happens after $t_{inf}$, or because $t_{inf}$ falls after $b$ and so does the time of the first positive swab. Furthermore, if $t_{inf}$ occurs before $a$, the individual would not be susceptible during the interval of interest, so they would not be counted in these calculations.

The advantage, compared to before is that the interval $(a, b)$ is now independent on the infection or swab status of the individual, and we can imagine each individual being assigned a duration $b - a$ from a distribution that depends only on admission and discharge times, and the interval $(x, y)$ under consideration. Let us call this the censoring distribution (after the fact that the time $t_{inf}$ would be censored if it occurred after $b$), and let $g(t)$ be its PDF and $G(t)$ its survival function (i.e. $1 - G(t)$ is its CDF). We do not know the structure of this distribution, and although we could make some approximations based on admission times and hospital demographics, this is not necessary. In fact, given an individual observation $\tau, \delta$ and the censoring distribution function $g(x)$ and $G(x)$, we can write the likelihood function for $\Lambda$:

$$\begin{aligned}\mathbb{L}(\Lambda|\tau,\delta) &\propto \begin{cases}\Lambda \exp\left[-\Lambda\tau\right] G(\tau) & \delta = 1,\ \text{not censored}\\ \exp\left[-\Lambda\tau\right] g(\tau) & \delta = 0,\ \text{censored}\end{cases}\\ &\propto \Lambda^{\delta} \exp\left[-\Lambda\tau\right] \times \left(G(\tau)^{\delta} g(\tau)^{1-\delta}\right)\end{aligned}$$

If we consider this function across all possible values of $\Lambda$, we can see that the contribution of $g(\tau)$ and $G(\tau)$ remains constant regardless of the value of $\Lambda$ in this likelihood function. We can conclude that the likelihood that $\Lambda$ takes a specific value is independent of the shape or structure of our censoring distribution and as a result we can remove $g(\tau)$ and $G(\tau)$ from any further calculations.

Ignoring the censoring distribution, we could calculate a Jeffreys prior for the value of $\Lambda$, accounting for our censoring window between $a$ and $b$:

$$\pi\left(\Lambda\right) = \sqrt{-\mathbb{E}\left[\frac{\mathrm{d}^2 \ln\left[\mathbb{P}\left(\delta,\tau|\Lambda\right)\right]}{\mathrm{d}\Lambda^2}\right]}$$

$$
\begin{aligned}
&= \sqrt{-\begin{cases} \int_a^b \Lambda \exp\left[-\Lambda x\right] \frac{\mathrm{d}^2 \ln[\Lambda \exp[-\Lambda x]]}{\mathrm{d}\Lambda^2} \mathrm{d}x, & \delta = 1 \\ + \\ \exp\left[-\Lambda(b-a)\right] \frac{\mathrm{d}^2 \ln[\exp[-\Lambda(b-a)]]}{\mathrm{d}\Lambda^2}, & \delta = 0 \end{cases}} \\
&= \sqrt{-\begin{cases} \int_a^b -\Lambda^{-1} \exp\left[-\Lambda x\right] \mathrm{d}x \\ + \\ 0 \end{cases}} \\
&= \frac{\sqrt{\exp\left[-a\Lambda\right] - \exp\left[-b\Lambda\right]}}{\Lambda}
\end{aligned}
$$

where $a$ and $b$ are the beginning and end of an individual's observable window. This would be a perfect prior distribution if the values of $a$ and $b$ were fixed for all observations. However, as they are determined by when individuals are admitted and discharged, they can vary between individuals. We choose instead to ignore the censoring window. If we take $a = 0$ and $b = +\infty$, this gives us the prior distribution:

$$\pi(\Lambda) = \Lambda^{-1}$$

This would seem like an adequate prior distribution. However, it too has problems, which are more obvious once we use it to calculate our posterior distribution.

For a set of observations, $\boldsymbol{\tau}, \boldsymbol{\delta}$ of $n$ patients, we can write the probability of these observations for any given value of $\Lambda$:

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{\tau}, \boldsymbol{\delta}|\Lambda) &= \prod_{j=1}^{n} \Lambda^{\delta_j} \exp\left[-\Lambda \tau_j\right] \\
&= \Lambda^r \exp\left[-\Lambda W\right]
\end{aligned}
$$

where $r = \sum \boldsymbol{\delta}$ is the total number of infections observed during this time period and $W = \sum \boldsymbol{\tau}$ is the total time between admissions and positive tests across all observations. This notation has been well described[188]–[190].

We can now write a posterior function for $\Lambda$:

$$
\begin{aligned}
\mathbb{P}\left(\Lambda = \lambda | r, W\right) &\propto \pi(\Lambda) \times \mathbb{P}(r, W|\Lambda = \lambda) \\
&= \lambda^{r-1} \exp\left[-\lambda W\right]
\end{aligned}
$$

By finding a normalising constant (dividing by the integral of $\lambda$ for all possible values of $\Lambda$), we find that the resulting posterior distribution is usually proper, and is in fact a Gamma distribution with input parameters $\Gamma(r, W)$. This is possibly predictable as the Gamma distribution is the conjugate of many rate-based distributions.

However, we need to consider what happens if $r = 0$ or $r = 1$. In the case of $r = 0$, we write:

$$\mathbb{P}(\Lambda = \lambda | r, W) \propto \lambda^{-1} \exp\left[-\lambda W\right]$$

It is not possible to find normalising constant in this case, as the integral across all possible values of $\Lambda$ does not converge, meaning this is an improper posterior distribution.

The posterior distribution does converge when $r = 1$:

$$\begin{aligned}\mathbb{P}(\Lambda = \lambda | r, W) &= \exp\left[-\lambda W\right] \times \left[\int_0^\infty \exp\left[-kW\right] \mathrm{d}k\right]^{-1} \\ &= W \exp\left[-\lambda W\right]\end{aligned}$$

which is equivalent to $\Lambda$ following an exponential distribution of rate $W$. However, if we consider this distribution for a moment, we can see another problem. The mode value for an exponential distribution is 0. Despite having evidence that the rate must be greater than 0, as at least on infection has occurred, our posterior distribution still puts 0 as the mode rate of infection (the central estimate for the rate of infection would be $\frac{1}{W}$, which does at least stand up to some reasoning).

To resolve these issues, rather than a strictly Jeffreys prior of $\Lambda^{-1}$, we have opted to use the prior distribution $\Lambda^{-\frac{1}{2}}$. The resulting posterior distribution,

$$\mathbb{P}(\Lambda = \lambda | r, W) = \frac{W^{r+\frac{1}{2}}}{\Gamma\left(r + \frac{1}{2}\right)} \lambda^{r-\frac{1}{2}} \exp\left[-\lambda w\right]$$

is the Gamma distribution $\Lambda \sim \Gamma\left(r + \frac{1}{2}, W\right)$. This is a proper distribution when $r = 0$, with some density for values of $\Lambda$ greater than 0, allowing for some force of infection even if we have not observed an infection, and its mode value at $r = 1$ is $\frac{1}{2W}$.

We have shown how, by observing when individuals were (or in fact were not) infected, we can generate a posterior distribution for the rate of infection. Of course, we do not know the actual time of infection for any one individual, we only know their time of diagnosis. Using the distribution for the incubation period, we assign each infected individual a randomly sampled incubation period (intended as the time from infection to diagnosis) and therefore a time of infection. We can calculate a random value of $r$ and $W$ for any time period and location based on these infection times (we need to include individuals who do not ever receive a positive diagnosis in these calculations, whose contributions will be the same for any random iteration). We can therefore generate sets $\mathbf{r}$ and $\mathbf{W}$ of size $m$ by performing $m$ iterations of this randomisation.

We want to find $\tilde{r}$ and $\tilde{W}$, input values for a distribution that represents a central estimate for the distributions generated by $\mathbf{r}$ and $\mathbf{W}$. We find our central estimates for $\tilde{r}$ and $\tilde{W}$ by finding a Gamma distribution whose mean and variance have a minimum difference between each of the simulated posterior distributions. If the mean and variance of this new distribution is given by $\tilde{\mu}$ and $\tilde{\sigma^2}$, then:

$$
\begin{aligned}
Err_{\tilde{\mu}} &= \sum \left( \tilde{\mu} - \frac{\mathbf{r} + \boldsymbol{\frac{1}{2}}}{\mathbf{W}} \right)^2 \\
&= m\tilde{\mu}^2 - 2\tilde{\mu} \left( \sum \frac{\mathbf{r} + \frac{1}{2}}{\mathbf{W}} \right) + \sum \left( \frac{\mathbf{r} + \frac{1}{2}}{\mathbf{W}} \right)^2 \\
\frac{\mathrm{d}Err_{\tilde{\mu}}}{\mathrm{d}\tilde{\mu}} &= 2m\tilde{\mu} - 2 \sum \frac{\mathbf{r} + \frac{1}{2}}{\mathbf{W}} = 0 \\
\tilde{\mu} &= \frac{\sum \frac{\mathbf{r}+\frac{1}{2}}{\mathbf{W}}}{m} \\
Err_{\tilde{\sigma^2}} &= \sum \left( \tilde{\sigma^2} - \frac{\mathbf{r} + \frac{1}{2}}{\mathbf{W}^2} \right)^2 \\
&= m\tilde{\sigma^2}^2 - 2\tilde{\sigma^2} \left( \sum \frac{\mathbf{r} + \frac{1}{2}}{\mathbf{W}^2} \right) + \sum \left( \frac{\mathbf{r} + \frac{1}{2}}{\mathbf{W}^2} \right)^2 \\
\frac{\mathrm{d}Err_{\tilde{\sigma^2}}}{\mathrm{d}\tilde{\sigma^2}} &= 2m\tilde{\sigma^2} - 2 \sum \frac{\mathbf{r} + \frac{1}{2}}{\mathbf{W}} = 0 \\
\tilde{\sigma^2} &= \frac{\sum \frac{\mathbf{r}+\frac{1}{2}}{\mathbf{W}^2}}{m}
\end{aligned}
$$

In the event that $\mathbf{W}_j = 0$, which happens when either there are no patients on the ward or when all patients on the ward have been infected recently enough so as to no longer be susceptible, we assume that the rate of infection is 0, as there are 0 individuals on the ward to be infected. In the case that only a subset of the iterations result in $W = 0$, the means and variances of these iterations would also equal 0 when calculating a central estimate for these values across iterations.

From these central values, we can write a whole Gamma distribution, or we can return the mean of this distribution, $\tilde{\mu}$, depending on the analysis we want to perform. For simple heatmap outputs, Nosoco just returns $\tilde{\mu}$.

## 4.4 Evaluating Nosoco

### 4.4.1 Methods

We used anonymised patient data from a multi-site NHS trust to test Nosoco. We had two metrics that we wanted to investigate:

1. Speed
2. How its results compare to PHE definitions of "likely" and "definite" nosocomial infection.

#### Speed

We tested Nosoco's speed by running it multiple times with different observation window lengths (time between the input start date and the input end date, see Figure 4.2) Specifically, we looked at windows of length a week, two weeks, 28 days, 90 days and 365 days. For each window length, we randomly select 100 dates for the input start date such that the entire observation window would fall between 05/04/2021 and 01/9/2022. This start date approximately represents the date where admissions start to stablise post-pandemic. In Section 4.4.3 we show the same results if we started our analysis on 01/03/2020. The end date is the last date from which Nosoco was able to extract data from the hospital trust, owing to the trust switching to a new database system.

As an example, if 25/12/2021 was randomly selected for a 28 day analysis, then a Nosoco would receive the input dates 25/12/2021 and 22/01/2022. We record how long Nosoco takes from start to finish for each of these extractions, as well as the total number of patients that are recorded in this extraction.

We had to make one change to the code of Nosoco to test its speed. We did not plot any heatmap results. With each iteration, the code would skip the plotting line, but was otherwise identical. This was done in part as a data protection measure, but also because we did not want to cause a slow-down in our code due to limited file space as this would not have been relevant to our investigation. We do include an example heatmap in our results to give readers an idea as to the output from Nosoco. However, to maintain anonymity, the patients admissions and swab results were swapped so that each admission history was assigned a random swabbing history.

We ran our random sampling on a Friday during working hours to emulate data extraction during a working week.

We investigated the number of first positive swabs occurring on each day between 01/03/2020 and 01/09/2022. For each swab, we calculated the probability that they represented a nosocomial transmission, according to Nosoco, and recorded if they occurred more than 2, 8 or 14 days into the individual's admission. In this case we defined an admission as any string of recorded ward admissions such that the gap between the recorded ward admissions was no greater than 24 hours. From these calculations we were able to estimate the total number of nosocomial transmissions according to Nosoco and the total number according to these threshold values. We used the `numpy.polyfit` package with a degree of 1 to find a straight line relationship between these two measures.

We wanted to approximate what proportion of the Nosoco estimate would be ignored by a given threshold. Ignoring the negligible y-intercepts, the gradient of our lines gave the linear relationship between the Nosoco estimate and the total number of transmissions estimated by each threshold admission length:

$$y_n = kx$$

where $x$ is the Nosoco estimate, $k$ is the gradient and $y_n$ is the total number of first positive swabs that occurred after a threshold time into an individual's admission of $n$ days, from hereon known as the threshold$_n$ total. A gradient of $k = 1$ would indicate that, on average, Nosoco and the threshold$_n$ total estimate a similar number of nosocomial transmissions. The proportion missed by the threshold total is given be:

$$\begin{aligned}\frac{x - y_n}{x} &= \frac{x - kx}{x} \\ &= 1 - k\end{aligned}$$

It is possible for this value to be negative when the threshold$_n$ total is consistently greater than the Nosoco estimate.

Finally, we wanted to see how this relationship changes for different distributions of the incubation period. We repeated the calculations performed by Nosoco, but changed the Gamma distributed incubation period each time so that, while the $\alpha$ shape parameter remained the same, the $\beta$ location parameter was changed to result in a mean incubation period varying from 1 to 14 days. We calculated the gradient when comparing the threshold counts to the Nosoco estimates for each distribution. We also calculated a bootstrapped value for this gradient by sampling randomly from the pairs of $x$ and $y$ values generated from each incubation period calculation until we had the same number of pairs as in the original calculation and repeating the gradient calculations. By repeating this process 1000 times we were able to approximate 95% confidence intervals for the true value of the gradient for any given mean incubation period length.

In both analyses, the swabbing data we analysed was comprised only of PCR results and not lateral flow tests, with the PCR technique not recorded. We were unable to ascertain if the swabs were taken from symptomatic individual, as part of regular screening or as a result of an outbreak on the patient's ward. No clinical information was available for patients, nor background information that may indicate that they were more or less likely to have been infected prior to attending hospital. A swab was counted as a "first positive swab" and therefore a new infection only if they had not had another positive swab in the past 90 days.

### 4.4.2 Results

#### Speed

Figure 4.5 is an example heatmap output from Nosoco between 01/03/2020 and 09/01/2021, estimating the total number SARS-CoV-2 nosocomial infections occurring on the investigated wards over this time period. As stated earlier, to maintain anonymity patient and swab data were shuffled. Despite this fact, a first and second outbreak wave is still observable as there are such a comparatively high total number of positive swabs during these time periods.

Figure 4.6 is a scatter plot of the run-time of Nosoco for each of our test runs, plotted against the number of patients in the database for each run. Table 4.3 collates these results, grouping them together by the length of each investigation time period. As seen in Table 4.3 and Figure 4.6 there appears to be a positive relationship between the length of time being investigated, the number of patients in the observed database and the run-time of Nosoco. In general, a larger time period is associated with more patients, and more patients and a larger investigation period often results in a longer run-time. A slightly more subtle point, that can be observed in the scatter plot in particular, is that for the same analysis window, a greater number of patients does not appear to be directly associated with an increase in the run-time.

| Time period | Average number of patients (s.d.) | Average time taken (s.d.) |
|---|---|---|
| 7 days | 8447.64 (548.24) | 10.22 seconds (1.60) |
| 14 days | 13782.41 (743.74) | 11.58 seconds (1.31) |
| 28 days | 23357.63 (970.27) | 14.10 seconds (1.38) |
| 90 days | 59495.17 (920.15) | 25.19 seconds (1.56) |
| 365 days | 184158.70 (325.62) | 78.53 seconds (1.65) |

Table 4.3. Average run-times for Nosoco when investigating different time periods of different lengths.

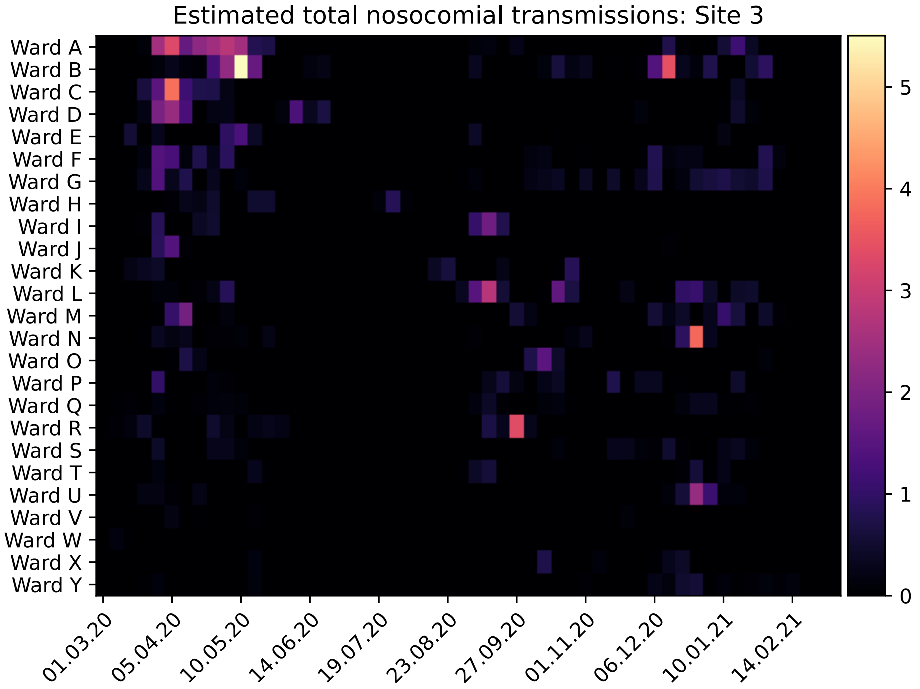


Figure 4.5. An example heatmap output for the estimated total number of nosocomial SARS-CoV-2 infections occurring on any one ward over a week period. This particular example looks at one anonymised site. The x-axis shows the passage of time and the y-axis shows each anonymised ward. In this investigation, positive swabs were randomly assigned to individuals to retain anonymity, meaning we are not actually observing real-world data. That being said, we can still observe the influence a high number of positive swabs can have on this estimate, as waves 1 and 2 of the SARS-CoV-2 pandemic can still be observed.

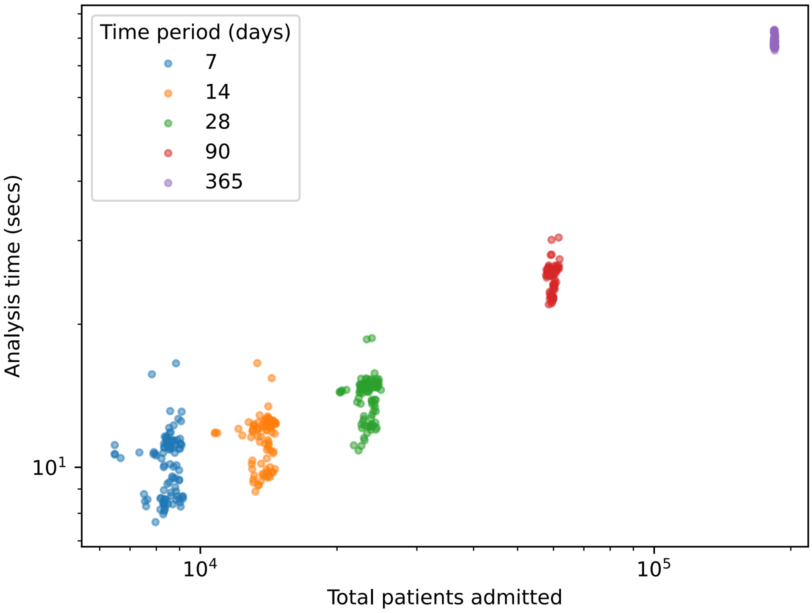


Figure 4.6. Scatter plots showing the relationship between the number of patients in a data-set and the run-time of Nosoco. Each point represents an individual run to Nosoco from a randomly sampled start point, with their colour representing the length of the time period being investigated. As the time period length could change the results by an order of magnitude, a log-log scale has been used.

Comparison with threshold method of nosocomial allocation

Figure 4.7 shows the relationship between the central estimate for the total number of nosocomial SARS-CoV-2 transmissions occurring over a time period as estimated by Nosoco and the number of nosocomial transmissions identified according to a cut-off threshold time from admission to diagnosis. Each point is calculated from all the first positive swabs occurring on a particular day, with the x-axis giving the Nosoco estimate and the y-axis giving the count of swabs that occurred past a certain threshold point in an individual′s admission ($\geq$ 2 days in blue, $\geq$ 8 days in orange and $\geq$ 14 days in orange). There is a positive relationship between the Nosoco estimate and each of the threshold counts. The total count when using a threshold of at least two days into an admission most closely matched the estimates from Nosoco, indicated by a slope for the fit closest to 1. Table 4.4 gives the mean daily counts for each metric, as well as the parameters for the lines of best fit.

In Figure 4.8 we can see by what proportion each threshold count is reduced when compared to Nosoco, and how this proportion would change if the true incubation period distribution was shorter or longer than the one used by Nosoco. In general, each threshold tends to result in a lower estimate of the total number of nosocomial transmissions when compared to Nosoco. If the mean incubation period were shorter (Nosoco uses a mean incubation period of 4.84 days), then this difference would be increased.

| | Daily count | | Relationship with Nosoco | |
|---|---|---|---|---|
| | Mean | St.d. | Gradient | Y-intercept |
| Nosoco | 2.45 | 3.36 | - | - |
| 2 day threshold | 2.53 | 3.30 | 0.94 | 0.23 |
| 8 day threshold | 1.29 | 2.11 | 0.58 | -0.12 |
| 14 day threshold | 0.80 | 1.51 | 0.38 | -0.12 |

Table 4.4. Daily nosocomial counts comparing Nosoco results to counts made based on admission length thresholds. The gradient and y-intercepts give the parameters for a line of best fit, obtained using the `numpy.polyfit` package. A gradient close to 1 indicates that on average the two measures observe the same number of nosocomial transmissions.

### 4.4.3 Discussion

From Figure 4.8, we can conclude that using a threshold of 8 days would estimate the total nosocomial transmissions as approximately half the estimate made by Nosoco, and using a threshold of 14 days would mean a reduction by nearly two thirds. If we went down to a threshold of 2 days, we would still see a lower estimate by approximately a tenth. Even if we accept that the distribution we use does not reflect the true distribution of time from infection to diagnosis in hospital, and that instead the average effective incubation period was 14 days long, we would still be missing a considerable proportion of transmissions by using either an 8 or 14 day threshold. If it is in fact shorter than the 4.84 days used by Nosoco, the threshold values would be considerably worse.

It would be tempting to simply change the threshold time to one that results in similar total estimates to Nosoco. However, doing so misses the main vulnerability of threshold methods: It makes a binary statement with complete certainty (ei-

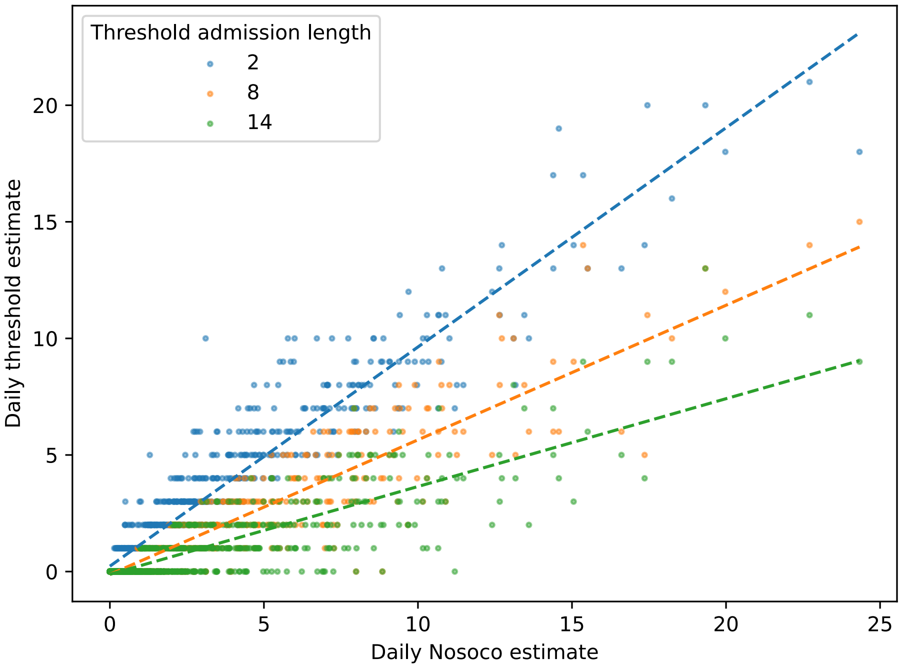


Figure 4.7. Scatter plot comparing the estimated total number of nosocomial transmissions from Nosoco to the total count based on an admission-to-diagnosis threshold time. The threshold values are discrete counts and as a result fall on integer lines. Each colour represents a different threshold. A line of best fit is shown for each threshold.

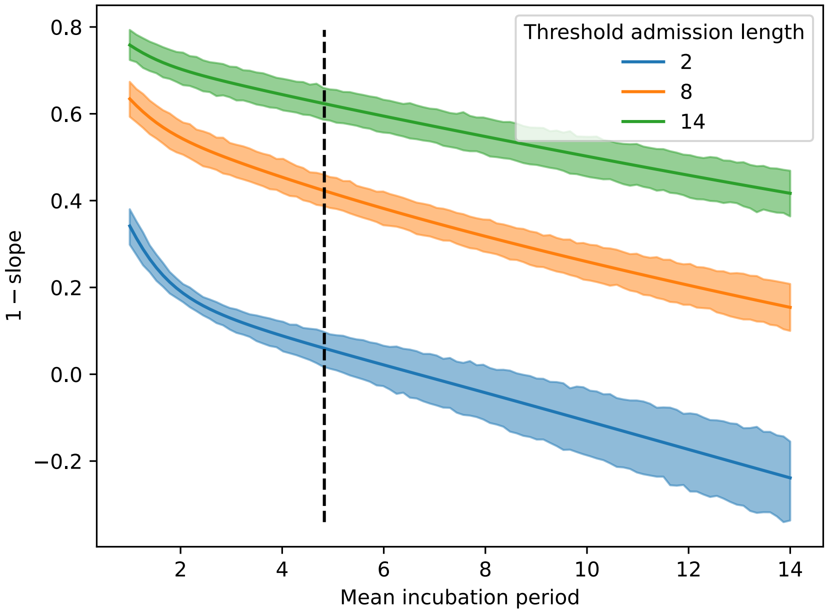


Figure 4.8. The estimated proportion of nosocomial transmissions that are not observed when using a threshold count method. This value is estimated as 1 minus the gradient of a line of best fit between the count estimated by Nosoco and the threshold count. As a result, it is possible for this value to be negative, indicating that the threshold method would estimate more nosocomial transmissions that Nosoco. Each colour represents a different threshold with the shaded areas showing the bootstrapped 95% confidence intervals. The x-axis shows how this estimate changes when the mean length of the incubation period distribution Nosoco uses is changed. The black dotted line indicated the distribution currently used by Nosoco.

ther the swab result in question does represent a nosocomial transmission or it does not) where there is in fact no certainty at all. By using a distribution to calculate the probability of a transmission having occurred at a particular time or a particular place, Nosoco allows us to incorporate that uncertainty into our assessment of any situation, with the added benefit that we are able to approximate not just that a transmission has occurred, but also where and when.

This is perhaps best demonstrated by the fact that the 2 day threshold total still only accounts for approximately 94% of the nosocomial transmissions estimated by Nosoco (see Table 4.4). This in itself is unexpected and remarkable. The distribution used by Nosoco would result in a slim chance of an individual being diagnosed within 2 days of infection (specifically a probability of 0.129). We might reasonably expect that having such a low threshold would mean that the number of nosocomial transmissions missed by having a short incubation period would be easily averaged out, if not surpassed, by the number of non-nosocomial transmissions that are misattributed as nosocomial transmissions owing to their long incubation periods. However, this assumes that individual's time of diagnosis is completely independent of the length of time they have been in hospital. If we are basing their time of diagnosis on time of first positive swab, then this is not true. A lot of hospitals, including the NHS trust involved in this analysis, use a screening program that makes it more likely for positive cases to be observed early in their admission (for the exact purpose of preventing nosocomial transmissions). A small proportion of these will represent nosocomial transmissions. They will be picked up by Nosoco but ignored by most threshold-based systems. If they represent a large enough proportion of all nosocomial transmissions, then a threshold system will always underestimate the number of nosocomial transmissions when compared to Nosoco.

For the hospital trust data we were investigating, Nosoco appears to be fast enough to investigate the prior month in under twenty seconds. From Figure 4.6 we can see that there initially appears to be a positive relationship between the run-time and the number of patients in the data-set. This would be unsurprising, as more patients will result in more calculations, as well as a longer SQL query. However, the investigation window shows us that this is a far more important indicator of run-time. For the same window lengths, varying the total number of patient admissions does not appear to affect the total run time (there is little increase in run time in data points of the same colour on the scatter plot in Figure 4.6). However, a larger window will usually result in a longer run time. This is likely because a larger window requires more partitions. For example, if each partition is a day long, then a 90 day window will have 90 partitions, compared to a 28 day window only having 28 partitions. Since Nosoco calculates the probability that each infected individual was infected in each of these partitions, a greater number of partitions may well result in a greater run-time, even if the number of patients remains the same.

There is room for improvement when it comes to the speed of Nosoco. For a large enough hospital trust investigated over a long enough time period, the run-time may become unmanageable. Past a certain point, local memory storage may mean that Nosoco stops being a viable option to investigate all nosocomial transmissions. This never occurred in our direct use of Nosoco, but has occurred when performing more complicated modelling of data generated by Nosoco. We suspect that, whilst easy to use, Python is not the optimum coding language for this task. As stated in the method, we did briefly investigate using C++. However, as Python was heuristically adequate, translating the tool into C++ will be a task for a future improvement of Nosoco.

"Heuristically" is an important word to consider. The PHE method of making a binary decision as to if an individual represented a nosocomial transmission based on their time of diagnosis was, at the time, heuristically adequate. It would not be possible or practical to perform the mathematics proposed by Nosoco by hand for every single patient, and so a simple way of indicating an increase in the number of nosocomial transmissions was required. However, as show in this analysis, there are more accurate methods of estimating the total number of nosocomial transmissions. Indeed, we can be certain that there are more accurate methods than Nosoco, even if they are not currently available. However, it is our hope the Nosoco provides medical practitioners with a more accurate, while equally easy, method of assessing the possible total number of nosocomial transmissions that have occurred.

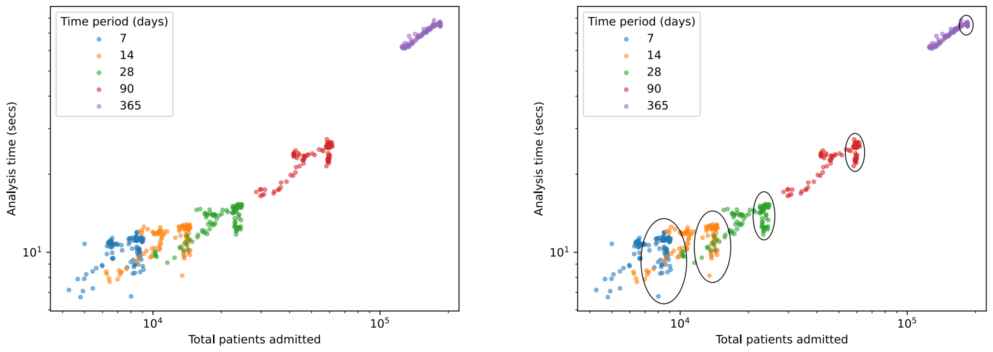
Analysis of speed starting at 1/3/2020

We chose to analyse Nosoco across dates between 05/04/2021 and 09/01/2022. We chose this end date as it marked the introduction of a new data management system in the NHS Trust we were using for the analysis. This meant that a different SQL request would be required after this date and matching data before and after this date could introduce errors. For this reason, we decided to only perform our analyses on hospital data prior to this time.

The SARS-CoV-2 global pandemic reached the United Kingdom at approximately 01/03/2020. We did not start our analysis until 05/04/2021, meaning there are approximately 400 days we have decided to exclude from our data-set. The reason we did this is because for about a year after the start of the pandemic, hospitals went through changes drastic enough as to effect the run-times of Nosoco.

Figure 4.9 shows the resulting plot of hospital population against run-time if we performed our analysis between 01/03/2020 and 01/09/2022. We can see some stark differences in the shape of this data if we compare it to the equivalent plot in the body of our text, Figure 4.6.

Firstly, for a given observation window, the range of total hospital admissions is a lot broader when we include the earlier parts of the pandemic. In Figure 4.6, the



(a)

(b)

Figure 4.9. A repeat of the run-time investigation of Nosoco from 01/03/2020 to 01/09/2022. Compare this plot to the equivalent plot in the body of the text, Figure 4.6. We can see the influence the changes in hospital during the early pandemic would have had on our conclusions regarding the run-time of Nosoco. Figure 4.9b circles the points on the scatter plots that represent data points from the same time period as Figure 4.6

range of total admissions was relatively narrow for each observation window length (each window length is represented with a different colour). There was little to no crossover between the observation window groups. In Figure 4.9, the ranges of total admissions for each observation window are broad enough that there are some crossovers. The week with the most total admissions (blue) saw more admissions than the fortnight with the least total admissions (orange) and, in turn, the fortnight with the most total admissions saw more admissions than the 28 days with the least total admissions from our data set (green).

This wider spread leads to a more dramatic observation: There appears to be a positive relationship between the number of admitted individuals and the length of time Nosoco takes to run. The points in each group appear to be forming a line. We previously noted that in each observation window group we could not see such a relationship. The points in each group in Figure 4.6 do not form a line and instead fall seemingly randomly in one region. How has changing our analysis date introduced a positive relationship between these two variables?

To understand this problem, we should first see if we can observe the run times from our initial analysis in this new plot. Indeed, by looking at the area of highest admission for each observation window, we can see regions where there appears to be no relationship between total admissions and run time. These regions are circled in Figure 4.9b. Our plots do not contradict each other. The positive relationship is only observable in the early period of the pandemic.

So why is there a positive relationship between Nosoco run time and total number of patients before 05/04/2020 that is no longer present after this time period. We suspect that this is relationship is an artifact of how the total number of admissions changed in the first year of the pandemic.

Figure 4.10 charts the total number of patients between 01/03/2020 and 01/09/2022.

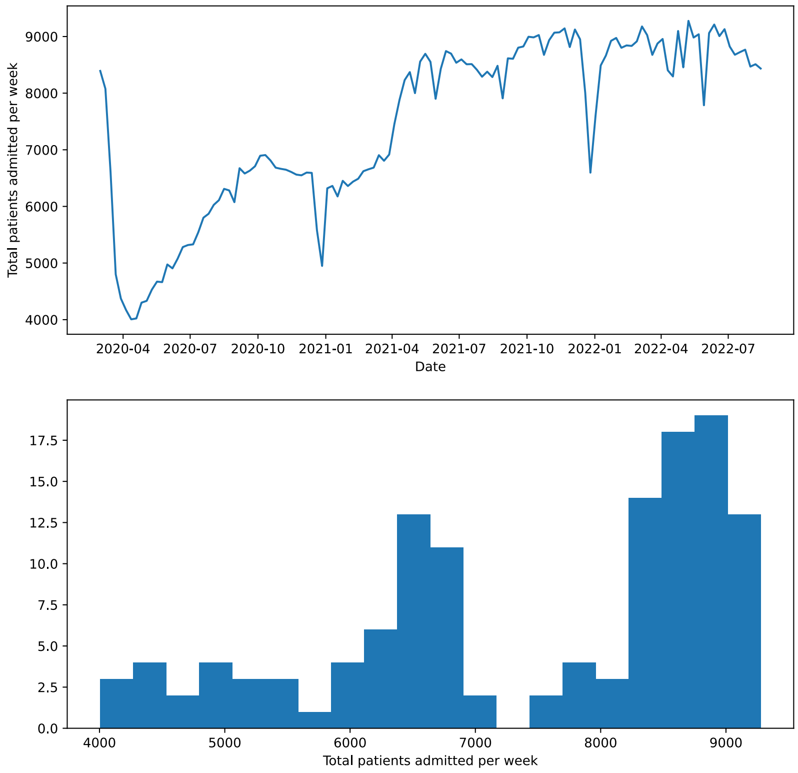


Figure 4.10. A count of the total daily number of patients admitted into the examined trust. The top plot shows how this changes over time and the bottom plot is a histogram of the totals. After approximately 05/04/2021, this count stabilises.

As we can see, at the start of the pandemic, there is a sharp drop in the total number of patients. This number steadily increases over the next 400 days and then stabilises to pre-pandemic levels. This may seem to be a surprising trend if we forget that this is a tally of the total number of different patients admitted each day, not the total number of patients admitted at any one time. This is not an indication that the trust was below capacity. Indeed, it may well be an indication that the trust was above capacity and not able to admit as many new patients.

It is the steady growth period that interests us. In the first year of the pandemic, a low total number of patients indicates that the sample was from earlier on in the year. As such, when there was a positive relationship between the total number of patients and the run time of Nosoco, there was also a positive relationship between the date of the Nosoco run and its run time. There are plenty of factors that may link the Nosoco run time to the date, for example the relative availability of SARS-CoV-2 testing throughout that first year. Additionally, when we choose a time period of analysis where there is no relationship between date and total number of patients (i.e. after 05/04/2021) we remove the association between total number of patients and total run time.

It is difficult to know what exactly causes this initial relationship, that then disappears after 05/04/2021. It may well be a combination of multiple factors that are difficult to account for. This is why, in order to avoid drawing incorrect conclusions regarding the relationship between the total number of patients and the Nosoco run time, we decided to perform our analysis between 05/04/2021 and 01/09/2022

## 4.5 The association between in-hospital and community rates of SARS-CoV-2 infections

The way in which an "outbreak" is defined is dependent on the user of Nosoco. For example, one may define an outbreak as more than one infection occurring in one location of the course of a week. Indeed, this is the definition initially proposed for a nosocomial outbreak of SARS-CoV-2 by PHE[183]. However, this tells us little about the risk to individuals on the ward where the outbreak is occurring, nor does it tell us anything about the size of the outbreak. Two nosocomial infections occurring in the same week on a ward of one hundred people is a very different picture to twenty out of twenty individuals being infected within a day of each other. Both will count as an outbreak (although we would hope that the latter would appear as a more likely outbreak according to our metric). For this reason, we will not be calculating the probability of an outbreak when analysing nosocomial infections in this chapter or the one following.

In Section 4.6 we will be looking specifically at the third metric of Nosoco, the rate of infection. This gives us the distribution for the time from an individual being admitted to becoming infected and it informs us of an individual's risk of infection whilst in hospital. We will use the rate of infection as a metric to see how this risk has changed over time and to see if age is an influencing factor. Instead, in this section we will be looking at the first of these metrics: the total number of nosocomial infections. At a cursory glance, it would seem that these two metrics are so intrinsically linked that there would be no need to examine them separately. There are, however, some important differences.

As mentioned above, the rate of infection shows the risk of infection to an individual. If we attempted to approximate this same risk using the total number of nosocomial infections alone, we would be at risk of introducing biases into our analyses. For example, if we wanted to investigate the risk of infection by age group and we observed that most nosocomial transmissions occur among the older hospital population, we may conclude that they are at greater risk of infection. However, this conclusion is going to be inherently influenced by the fact that the hospital population tends to be older owing to increased health needs and delayed discharges. Similarly, we may see a spike in risk of nosocomial infection among women of a child-bearing age, but this would likely only represent the fact that a proportion of all hospital admissions are for child-birth, and are intrinsically linked (with exceptions) to being a woman of child-bearing age. The aim of approximating the rate of nosocomial infection is to delineate these biases from true elevated risks of infection.

Total number of nosocomial infections does not, therefore, tell us of the risk to the individual. It does, however, inform us of the total burden nosocomial infections put on a ward or hospital. A subset of the hospital community may be at elevated risk of nosocomial infection (e.g. immunocompromised individuals). However, if they do not make up a large proportion of the hospital community, their elevated risk may not generate a large amount of burden on the hospital as a whole. In this way, total numbers of infections tells us more about the burden than the rate of infection can (although not everything: in the case of immunocompromised individuals, they may be at elevated risk of severe sequelae of infection, meaning total number of infections alone does not give us a complete picture of the resulting burden).

In this section, we are going to use Nosoco to investigate how the total number of nosocomial infections changed during the first two waves of the SARS-CoV-2 pandemic in a multi-site trust. In particular, we want to focus on how the internal nosocomial total compares to the total concurrent external transmissions.

Hospitals and hospital wards represent varied levels of closed environments. Theoretically, patients are not able to leave the ward and return freely. Any infection that they contract during their admission should be the result of a transmission occurring on the ward. Similarly, during the early stages of the pandemic, many hospitals banned patient visitors. These measures were all in place to try and limit the number of introductory events occurring in hospital, with PPE and patient isolation being used to limit spread of SARS-CoV-2 once it is introduced.

As no IPC (Infection Prevention and Control) plan can be perfect, we would expect to see some relationship between the external total number of infections and the total number of nosocomial infections. With more individuals infected outside of hospitals, there is a greater chance that one of these individuals is an asymptomatic health care worker, or is going to be admitted coincidentally as a patient, resulting in the start of a chain of nosocomial infections. The less variable these chains are, or the more regularly these introductory events occur, the closer these two values will be associated. Using data from a multi-site trust, we are going to show how this relationship changed during the first two waves. We are going to approximate both the external and internal total number of SARS-CoV-2 transmissions and use linear regression to show how this time period can be divided up according to this changing relationship.

### 4.5.1 Methods

As a reminder, we use Nosoco to generate an estimate for the total number of nosocomial infections based on each infected individual's probability of having been infected whilst in hospital. We observe the time at which they have their first positive swab taken and compare this to their admission periods prior to this time. In order for the initial transmission to have occurred whilst they are in hospital, they would need a time interval from transmission to first positive swab that is shorter than the time between admission and first positive swab. We say that the time from infection to first positive swab follows a known Gamma distribution. We use this distribution to calculate the probability that the individual has a time from infection to diagnosis in keeping with their infection occurring during the course of their hospital admission. By summing across all individuals' probability of having been infected in hospital on a particular day, we calculate a central estimate for the total number of nosocomial infections that occurred on that day.

We used Nosoco to extract and analyse swab and admissions data from 01/03/2020 to 31/12/2020 from a multi-site trust. Swabs taken after this time period were included in the analysis as they might refer to cases that were infected during the period of interest. Five sites were chosen from the trust, including two general hospitals, one general hospital with a zero-coronavirus policy (any possible infections were transferred to other sites), one women's health unit and a children's hospital. We calculated the total number of nosocomial infections for each day. We also approximated the daily external incidence from the same swab data by calculating the probability that each individual was infected on a particular day whilst not in hospital. In doing so, we likely underestimated the true total number of individuals infected outside of hospital, as it does not account for individuals who did not present to hospital for testing because they were not unwell enough to need to go to hospital, or they presented elsewhere (e.g. a different trust). We assumed that for a small enough time period, this bias will remain relatively constant. If this bias changes, it should change the gradient of the relationship between nosocomial and external rates of infection.

We aimed to fit the external incidence to the nosocomial incidence through linear regression, using the `numpy.polyfit` package. In order to accommodate for changing hospital policies, governmental policies, public health attitudes and environmental resistance, we divided our data-set into “before 01/04/2020", “01/04/2020 - 01/09/2020", “01/09/2020 - 01/11/2020" and “after 01/11/2020".

Care is needed when considering this linear regression, as explained in this and the next paragraph: although it is tempting to state that any correlation observed would prove a connection between external and internal incidence, in reality we are not looking at these values at all, but at their central estimates. An individual data point, defined as each infected individual's first positive swab, would be represented in this data-set across multiple days according to the incubation period distribution. It is possible that this could be vulnerable to misinterpretation, especially as one individual ends up contributing to the central estimates of both the external and nosocomial incidence.

Consider a scenario where we only have two individuals in our data set and they both have their first positive swab at the same time. The only different between the two is that whilst the first was admitted in hospital for many days before their first positive swab, the second had been admitted just before their first positive swab. The central estimate for the total number of individuals infected each day inside or outside of hospital is the sum of the probabilities of either individual being infected inside or outside hospital, respectively. These two sums would be identical for each day, as the probability that the first individual was infected on a particular day whilst in hospital would be identical to the probability that the second individual was infected outside of hospital on the same day owing to their identical first positive swab timing. The result would be a perfect correlation (i.e. dots, one for each day, all sitting along the main diagonal in a plot similar to Figure 4.11b), from which we would conclude that we have strong evidence that external and nosocomial total infections are closely correlated, despite only actually having two data points in our data set.

To compensate for this, we repeat our regression across 1000 iterations with each individual in each iteration assigned a specific random infection time according to their date of first positive swab and the distribution of the time from infection to diagnosis.

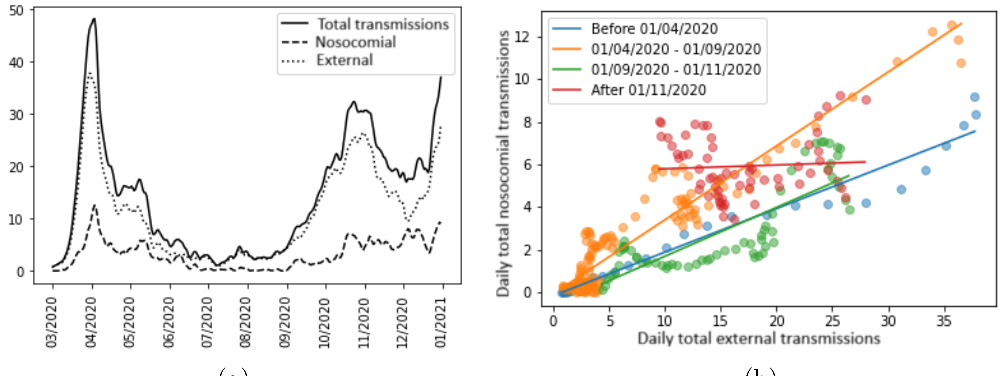
### 4.5.2 Results

Our investigations covered 106310 patients admitted during our observation window. 4197 individual had their first positive swabs during this time period, of which 1456 were admitted at one of our five sites on the trust under consideration at the time the swab was taken. 711 would be declared as “probable” nosocomial infections according to PHE guidance. We approximate the true number of nosocomial infections over this time period to be 959.89 (95% C.I. 937.32 - 981.49). Figure 4.11a shows how this incidence has evolved over time.

There appears to be a strong relationship between internal and external incidence at the start of the pandemic. However, we see no evidence of any connection after 01/11/2020. Table 4.5 details the results of these linear regressions in full. This observation is corroborated by Figure 4.12 and Table 4.6, which show the results of linear regressions across iterations of randomly assigned infection times.

| Time period | Constant | Coefficient | Standard error | T-score | P-value |
|---|---|---|---|---|---|
| Before 01/04/2020 | -0.165 | 0.205 | 0.008 | 25.803 | 0.000 |
| 01/04/2020 - 01/09/2020 | -0.077 | 0.347 | 0.010 | 36.021 | 0.000 |
| 01/09/2020 - 01/11/2020 | -0.611 | 0.230 | 0.021 | 10.820 | 0.000 |
| After 01/11/2020 | 5.620 | 0.018 | 0.037 | 0.480 | 0.633 |

Table 4.5. Linear regression results for the relationship between the total external and nosocomial transmissions



(a) (b)

Figure 4.11. Plots of the evolving nosocomial and external SARS-CoV-2 transmission totals as calculated from the timing of individual's first positive swabs. Figure 4.11a shows the ongoing changes in the number of transmissions, with the full line denoting total transmissions, the dotted line denoting external transmissions and the dashed line denoting nosocomial transmissions. Figure 4.11b shows the evolving relationship between external and nosocomial transmissions, with each point in the scatter plot representing the results from a particular day. These have been grouped into "Before 04/2020", "04/2020 - 09/2020", "09/2020 - 11/2020" and "After 11/2020", with a line of best fit to show the relationship in each sub-group.

| Time period | Constant (95% C.I.) | Coefficient (95% C.I.) | P-value (95% C.I.) |
|---|---|---|---|
| Before 01/04/2020 | 0.058 (-0.52 - 0.637) | 0.186 (0.128 - 0.245) | 0.000 (0.000 - 0.000) |
| 01/04/2020 - 01/09/2020 | 0.084 (-0.105 - 0.273) | 0.322 (0.257 - 0.386) | 0.000 (0.000 - 0.000) |
| 01/09/2020 - 01/11/2020 | -0.122 (-0.643 - 0.399) | 0.201 (0.123 - 0.279) | 0.000 (0.000 - 0.001) |
| After 01/11/2020 | 4.990 (2.670 - 7.310) | 0.054 (-0.029 - 0.137) | 0.420 (0.124 - 0.754) |

Table 4.6. Linear regression results for the relationship between the total external and nosocomial transmissions when infection times have been sampled across 1000 random iterations.

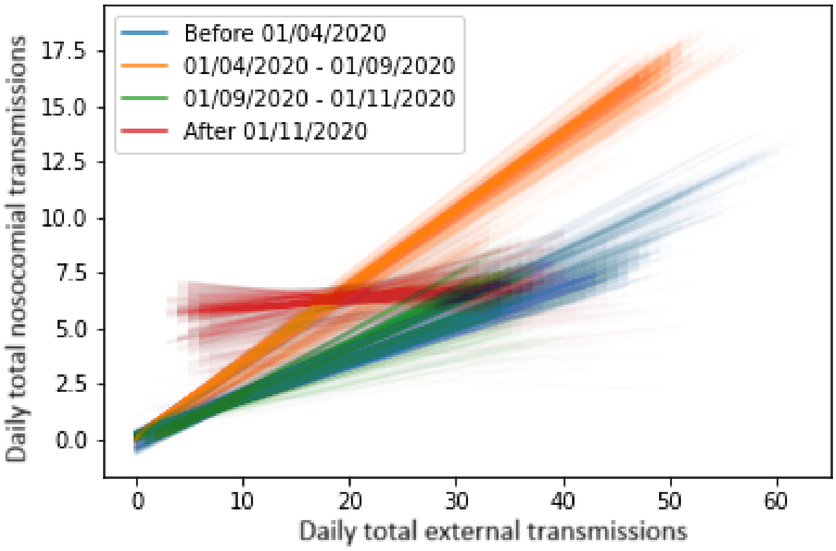


Figure 4.12. Multiple linear regressions mapping external and nosocomial transmission totals given randomly assigned transmission times. The same date groupings have been used as in Figure 4.11b, and the same observations remain true.

### 4.5.3 Discussion

Being able to discriminate between internal and external transmissions is a crucial part in understanding the spread of nosocomial infections in hospitals and ultimately limiting further outbreaks. Nosoco as a tool is designed to analyse and illustrate ongoing outbreaks. It would be tempting to just choose one metric to observe, such as total number of infections or rate of infection or even simply the total number of wards experiencing an outbreak. However, by defining each metric we gain the opportunity to draw interesting conclusions from each of them.

In this section we specifically wanted to illustrate the change in burden nosocomial infections put on the trust we were investigating over the first two waves of the pandemic. We used total number of nosocomial transmissions to demonstrate this burden.

Figure 4.11a illustrates how this burden appears to change over time. The highest peak of total nosocomial transmissions occurred during the first peak at the beginning of April 2020, in keeping with similar external transmission counts. The second external wave can be approximately divided into two sections, with a first initial peak in October 2020 followed by a trough leading into a peak in 2021. However, the total nosocomial transmissions does not follow this pattern, and instead steadily rises throughout this wave.

Figures 4.11b and 4.12 indicate a shift in relationship between external and internal rates of infection. Initially there appears to be a strong positive correlation between the two rates. Without careful examination, it is not clear in which direction this causation is occurring, and although it is reasonable to assume significant introduction of infection in hospitals from the community, evidence of community seeding from hospitals also exist (e.g. [191]).

This relationship continues all the way to November 2020, although the exact gradient appears to be dependent on if we are seeing an increase or decrease in external nosocomial transmissions, with a decrease in external transmission seeing a steeper gradient between the two (01/04/2020 - 01/09/2020). This may be indicative of a delay between the total number of external transmissions and the total number of nosocomial transmissions. If the number of nosocomial transmissions is dependent not on the current number of external transmissions but on the number from a few days prior, then a decreasing total number of external infections will result in an apparent increase in the gradient between the current number of external transmissions and the current number of nosocomial transmissions. This in turn implies a direction to this correlation, in that an increase in external transmissions results in an increase in nosocomial transmissions, and not the other way round.

However, our calculations from the start of November onward indicate no further relationship between external incidence and nosocomial transmission. There is no longer significant evidence of either external incidence placing a force of infection on admitted individuals or individuals infected in hospital seeding infections into the community.

There are many possible reasons for this change to occur that represent a true disconnection between external and nosocomial transmissions, including improved PPE use, better access to community testing or increased public awareness. Each of these could help reduce the extent to which external infections could effect nosocomial totals.

A compelling theory could be that our results show an improvement in SARS-CoV-2 diagnosis methods at the start of admissions. Infected individuals are correctly identified through swabbing earlier in their admission and as a result are not incorrectly identified as resulting from nosocomial transmission. Nosoco cannot identify individuals that have been diagnosed with a SARS-CoV-2 infection clinically. If there is delay between admission and a positive swab for these individuals, their apparent chance of representing a nosocomial infection will increase. Reducing this delay will have reduced the apparent correlation between external and internal rates of infection.

However, we would expect this reduction to be proportional and for there to still be an observable relationship. Instead we see a complete breakdown of any relationship. Looking at Figure 4.11a, we can see that this occurs during the trough midway through the second peak. This trough is likely a result of the second lockdown in the UK, and is reflected in national data[192] as well as the total incidence in the same plot. As it happens, the burden presented by nosocomial transmissions steadily increases. Without further analysis of nosocomial transmissions it is difficult to understand why.

Nosoco is by no means a perfect method of analysis. The above comments allude to one of its greatest weaknesses with the current design: Nosoco does not take the clinical picture into account. For multiple reasons it is possible for someone to be symptomatic long before they ever have a positive SARS-CoV-2 swab. Similarly, an individual may never develop symptoms but have a positive test as a result of regular screening or screening following exposure. Both options bring the use of the incubation distribution estimated by Overton et al.[133] as a proxy for the infection-to-diagnosis time interval into question. With a better understanding of the clinical picture, this may be prevented. One simple suggestion could be to record if a swab is taken for screening purposes or because the individual is symptomatic, and if they are symptomatic when these symptoms began. The distribution used could be altered based on this input. If this data is available, Nosoco is not currently arranged to access it.

Ultimately Nosoco only indicates trends and possible evidence of nosocomial transmission. It shows the most likely arrangement of nosocomial and external transmissions given diagnosis timings. These can be used for further statistical analysis of the hospital population as a whole, but for an individual knowing the probability that they were infected inside or outside the hospital does not equate to knowing if they do or do not represent a nosocomial transmission. There is a distinct reason why it is designed to output NHS Numbers for individuals who have been calculated to be high risk: for an individual, their probability needs clinical correlation.

Looking at the total number of transmissions in this way can be useful to get an idea for the magnitude of nosocomial transmissions, but it can leave a number of questions unanswered. In the next chapter, we will investigate what information the rate of nosocomial infection can provide us, compared to the total number of nosocomial infections. We will see if there is a true difference between the waves of SARS-CoV-2 outbreaks in terms of risk of nosocomial transmissions, and if age is a risk factor.

## 4.6 Estimating the changing rate of nosocomial transmission as a means of comparing outbreak waves

Understanding the rate of transmission can indicate areas in need of improvement in transmission prevention as well as help equate risks of elective hospital admissions and delayed discharges. In the previous section we were simply aiming to estimate the total number of nosocomial transmissions, as this can give insight as to the burden a hospital may fall under from nosocomial transmissions. We observed an interesting change in the relationship between nosocomial transmissions and external transmissions during the first two waves of the SARS-CoV-2 pandemic. In the first wave they were closely related, but in the second wave this relationship waned. Without more information, we were unable to delineate why this was happening.

Knowing the total number of nosocomial infections is useful for understanding the real-world burdens they can introduce. However, it does not necessarily inform us of the actual infectious processes occurring and it makes it very difficult to critically compare two scenarios without them being very similar. This method of calculation is predisposed to highlight areas of high transit. After all, with a constant rate of infection, wards or hospitals with more patients will generate more nosocomial infections. Additionally, depletion of susceptible individuals will appear as an improvement in management, rather than the reality: a natural part of any outbreak event. Instead we propose that time from admission to nosocomial transmission is distributed on an exponential distribution with rate $\Lambda$. A higher value of $\Lambda$ will indicate a higher individual risk of nosocomial transmission, regardless of the size of the ward or hospital. We demonstrate a method of estimating $\Lambda$ based on the total number of nosocomial transmissions and total pre-transmission time for all patients (including those that never get infected).

We propose two alternative approaches, one where the value of $\Lambda$ is fixed (this is the method used by Nosoco to assess the rate of nosocomial infections on wards over a short period of time) and one where it varies between individuals along an Exponential distribution.

We use each of these methods to approximate metrics for the rate of nosocomial transmission in an NHS trust between March 2020 and September 2022, as well as estimate the total number of infections for comparison. In particular, we want to contrast the rates of nosocomial transmission during five outbreak waves observed during this time period as well as the risk to individual age groups. We want to compare our conclusions to those we would have made if we looked at total number of nosocomial infections alone.

### 4.6.1 Model

We start with the assumption that we know the exact time every individual in our data-set is (or indeed is not) infected. We have already shown in Section 4.3.2 how we would calculate a posterior distribution for value of $\Lambda$, the rate of infection. As this will be useful to compare to our method for estimating the distribution a varying rate of infection follows, we will briefly reiterate our process here.

The observation of a single individual can be broken down into two values, $\tau$ and $\delta$. The indicator function $\delta$ tells us if the individual was infected during our observation, in which case $\delta = 1$ and otherwise $\delta = 0$. The continuous non-negative value $\tau$ tells us the length of time we observed the individual for before they were either infected or the observation period came to an end, which ever came first. When we assume each individual experiences the same constant rate of infection $\Lambda$, we are equivalently assuming that the distribution for their time from admission to infection is exponentially distributed with the rate $\Lambda$. As a result, we can write the likelihood function for $\Lambda$ given any observation $\tau$ and $\delta$:

$$\mathbb{L}(\Lambda|\tau, \delta) = \Lambda^{\delta} \mathsf{exp}\left[-\lambda\tau\right]$$

When we did observe an infection, $\delta = 1$ and this function is equal to the probability density function of the exponential distribution $\Lambda \mathsf{exp}\left[-\Lambda\tau\right]$. If our observation is censored before our individual is infected (for example, if they are discharged without being infected) then $\delta = 0$ and this function is equal to the survival function of the exponential distribution, $\mathsf{exp}\left[-\Lambda\tau\right]$.

If now we consider a set of $n$ paired observations, $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$, the likelihood function for $\Lambda$ given that each of these observations experienced the same constant rate of infection is the product of the likelihood function given each individual observation:

$$\mathbb{L}(\Lambda|\boldsymbol{\tau}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \Lambda^{\delta_i} \mathsf{exp}\left[-\Lambda\tau_i\right]$$

$$= \Lambda^{\sum_{i=1}^{n}\delta_i} \exp\left[-\Lambda \sum_{i=1}^{n} \tau_i\right]$$
$$= \Lambda^{r} \exp\left[-\Lambda W\right]$$

where $r = \sum_{i=1}^{n} \delta_i$ is the total number of observed infections and $W = \sum_{i=1}^{n} \tau_i$ is the total length of time observed across all observations. The posterior distribution for $\Lambda$ is proportional to some prior distribution for $\Lambda$ multiplied by the likelihood function for $\Lambda$ given our observations. We choose $\pi(\Lambda) = \Lambda^{-\frac{1}{2}}$ as our prior distributions for reasons discussed in Section 4.3.2, giving the posterior distribution:

$$\mathbb{P}(\Lambda = \lambda | \boldsymbol{\tau}, \boldsymbol{\delta}) = \lambda^{r-\frac{1}{2}} \exp\left[-\lambda W\right] \times c$$

where $c$ is a normalising constant. By integrating through all possible values of $\Lambda$ we find that $c = \frac{W^r}{\Gamma(r)}$ and that the posterior distribution for $\Lambda$ is the Gamma distribution $\Lambda \sim Gamma\left(r + \frac{1}{2}, W\right)$.

However, this posterior distribution assumes that all individuals in our observation set experience the same rate of infection. This may well not be a reasonable assumption. If there is an outbreak on a particular ward then observations coming from that ward would be expected to experience a higher rate than observations from other wards. Observations during a national wave of infections are likely to see a higher rate of infections than those in lull periods. More vulnerable individuals, such as immunocompromised individuals, could be more likely to be infected sooner and in effect see an elevated rate of infection. Whilst assuming a fixed, constant rate of infection may be adequate whilst assessing a ward's instantaneous infection rate, it is less useful and accurate when comparing broader data-sets.

For these reasons, we propose an alternative, one where the rate of infection itself falls on an unknown Exponential distribution. Rather than attempt to find a full posterior distribution for the parameters in these distributions, we find the maximum likelihood estimate (MLEs) for the rate parameter of the exponential distribution each individual's rate of nosocomial transmission is sampled from.

Part of our motivation for re-establishing our method for calculating a posterior distribution for $\Lambda$ is that it will provide a framework for us to calculate an MLE. We start by calculating the likelihood function for $\eta$, the rate of the exponential function from which $\Lambda$ is drawn, based on a single observation $\tau$ and $\delta$. In this case we integrate through every possible value of $\Lambda$:

$$\mathbb{L}(\eta | \tau, \delta) = \int_0^\infty \Lambda^{\delta} \exp\left[-\Lambda\tau\right] \times \eta \exp\left[-\Lambda\eta\right] \mathrm{d}\Lambda$$
$$= \frac{\eta}{(\tau + \eta)^{\delta+1}}$$

As with before, we now consider a set of $n$ paired observation, $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$. The likelihood function for $\eta$ given all of these observations is equal to the product of the likelihood function for $\eta$ given each of these observations:

$$
\begin{aligned}
\mathbb{L}(\eta|\boldsymbol{\tau},\boldsymbol{\delta}) &= \prod_{i=1}^{n} \mathbb{L}\left(\eta|\tau_i,\delta_i\right) \\
&= \prod_{i=1}^{n} \frac{\eta}{(\tau_i+\eta)^{\delta_i+1}} \\
&= \frac{\eta^n}{\prod_{i=1}^{n}(\tau_i+\eta)^{\delta_i+1}}
\end{aligned}
$$

Unfortunately, this product term does not reduce as elegantly as it did when we were calculating a posterior distribution for $\Lambda$. We are not trying to find a full posterior distribution this time. Instead, we are trying to find the value for $\eta$ that maximises this likelihood function. We can observe that a value that maximises a likelihood function will also maximise the logarithm of the likelihood function (the so-called log-likelihood function). The log-likelihood function for $\eta$ can be written as follows:

$$
\ln\left[\mathbb{L}(\eta|\boldsymbol{\tau},\boldsymbol{\delta})\right] = n\ln\left[\eta\right] - \sum_{i=1}^{n}(1+\delta_i)\ln\left[\tau_i+\eta\right]
$$

We will find this function's maximum at a point where its gradient with respect to $\eta$ is equal to 0:

$$
\begin{aligned}
\frac{\mathrm{d}\ln\left[\mathbb{L}(\eta)\right]}{\mathrm{d}\eta} = \frac{n}{\eta} - \sum_{i=1}^{n}\frac{1+\delta_i}{\eta+\tau_i} &= 0 \\
\sum_{i=1}^{n}\frac{1}{\eta} - \frac{1+\delta_i}{\eta+\tau_i} &= 0 \\
\sum_{i=1}^{n}\frac{\tau_i-\delta_i\eta}{\tau_i+\eta} &= 0 \\
\sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\eta} &= \eta\sum_{j=1}^{r}\frac{1}{\tilde{\tau}_j+\eta}
\end{aligned}
$$

where $\tilde{\boldsymbol{\tau}}$ is the set of all observations where $\delta = 1$ (i.e. where a transmission was observed). As $\eta$ increases, this gradient decreases, and for large enough values of $\eta$, the gradient will be negative so long as $r > 0$, meaning we are guaranteed to find a value of $\eta$ which is a solution to this equation. This solution will provide the MLE for $\eta$, the rate of the exponential distribution on which $\Lambda$ is in turn distributed.

### 4.6.2 Methods

Our data-set consists of all hospital admissions and positive SARS-CoV-2 swabs for patients in a trust from 01/03/2020 - 01/09/2022. From this time period, we focus on 5 sections, 11/3/2020-5/4/2020, 2/9/2020-2/1/2021, 11/12/2021-31/12/2021, 24/2/2022-31/3/2022 and 19/6/2022-9/7/2022, representing the start-to-peaks of five waves of nosocomial infections. Wave 1 represents the start of the SARS-CoV-2 pandemic, and Wave 3 approximately represents the start of the Omicron variant wave. These do not directly correlate to external waves during the pandemic, as shown in part in Section 4.5, but are instead named by observing the waves in out data. Initially we consider all patients susceptible. They are diagnosed (their incubation period ends) at the time of their first positive swab. An individual is not susceptible again until they have not had a positive swab for 90 days. After this point, they can be reinfected.

For each iteration, each occurrence of an infection is assigned a latent period based on random sampling from the Gamma distribution used in Nosoco. This means we have effectively assigned to each individual with a positive swab a time for their infection. For each iteration we can therefore define a set of observations $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$ from which we estimate the parameters $r$ and $W$ when assuming a fixed rate of infection, and additionally $\eta$ when assuming an Exponentially distributed rate of infection. We also approximate a range for the possible total number of infections based on the value of $r$ for each of our 1000 iterations.

We obtain central estimates for the mean and standard deviation of the rate of infection according to each of these distribution functions (in the case of the fixed rate of infection, we calculate the central estimate for the mean and standard deviation for the posterior distribution for $\Lambda$). From these averages we estimate a "central estimate distribution" for the rate of infection for each distribution type by generating a matching distribution with the corresponding mean and standard deviation.

We use this process to generate estimates for the rate of infection for each day across our analysis window. We repeat this analysis, but divide individuals up into the age groups <6 months, 6 months - 5 years, 5 year - 18 years, 18 years - 30 years, 30 years - 45 years, 45 years - 60 years, 60 years - 80 years and > 80 years. We define each individual's age by the date of birth listed on their admission.

Finally, we repeat this process, except rather than dividing our observation period into one-day time steps, we perform the analysis across the observation period, dividing the population only by their integer age in years from 0 to 80.

### 4.6.3 Results

Figure 4.13 shows the fluctuating estimates for the rate or total number of nosocomial infections over the first four waves of the pandemic. Wave 1, at the start of the pandemic, is the tallest, and sees the sharpest increase in infections. Wave 2 is different in nature to the other waves, taking a lot longer to reach the same height as the others. However, as it occurs over a much longer period of time, it represents a far higher total number of infections than the other waves. Waves 3, 4 and 5 could be argued to be part of the same wave. They never reach the same severity as the other waves, but their initial trajectories match closely to Wave 1.

Figure 4.14 shows the spaghetti plots for the same analysis performed on the first 50 random iterations for the assignment of infection times, with the first iteration in bold. This is how the analyses appear if we do not average across iterations.

Figure 4.15 shows an estimate for the daily number of transmissions when patients are separated up into age groups. In general, the 60-80 and >80 groups consistently have a higher total number of infections, followed by the 45-60 group. There is a small exception in Wave 3, where we see a small peak of 5-18 year olds being infected. Once we divide into these age groups, the central estimates for the daily number of nosocomial transmissions do not rise above 1 for any other than the two oldest age groups.

Figures 4.16 and 4.17 estimate the posterior distribution for a fixed rate of infection and the range of rates of infection experienced for the model with a varying rate of infection respectively. While the older age groups tend to have higher estimates for the rates of infection, there are few days if any when this difference is statistically significant and there are some days, specifically in Waves 3 and 4, when the highest rates of infection are experienced by younger age groups. Figures 4.18-4.20 show equivalent spaghetti plots.

Finally, we look at transmissions in yearly age groups for each age from 0 to 80. Figure 4.21 totals the number of transmissions for each of these ages across out analysis period, we 95% confidence intervals generated by multiple iterations of Nosoco. As we can see, most of the transmissions occur in the older ward populations. However, there is a slight up-tick in total nosocomial transmission among the youngest patients $< 2$ years.

This up-tick in total transmissions in the very young is not propagated through when estimating a shared or exponentially distributed rate of infection for each age (Figure 4.22). Instead, we generally see a higher rate for infection for all individuals above about 40 years old, although in the case of a varying rate of infection, this represents a window orders of magnitude wide that covers the central estimates for most ages.

### 4.6.4 Discussion

The rate and the total number of transmissions together can tell us a lot about how different waves and age groups compare with regards to the risk of nosocomial transmission. For example, looking at transmissions alone, it would appear that Wave 3 was initially nearly as severe as Wave 1 when it came to nosocomial transmissions (Figures 4.13 and 4.14). Indeed, in some ways it was, in that they posed a similar amount of burden on the trust as a whole (in the last chapter we discussed how the total number of transmissions is indicative of the total burden on a hospital setting). However, when comparing rates of infection, we see Wave 3 start to tail off far sooner than Wave 1. Similarly, we saw an up-tick in the total number of transmissions in the very youngest hospital populations that is not reflected in an overall raised rate of infection. When this occurs, it indicates that whilst we might in total be seeing more nosocomial transmissions from that age group or at that time, there is in truth limited evidence of a raised risk of infection to corresponding susceptible individuals.

In this section we have presented two options for estimating a rate of nosocomial infection: either by estimating a posterior distribution for the rate of infection that all individuals experience within the inspected subset or estimate an exponential rate from which each individual samples their own rate of infection. Looking at our results we see that they tend to approximate the same rate of infection, but estimating $\eta$ results in a broader confidence interval. The natural question to ask: which one should we use?

Considering real-world models of nosocomial transmissions, it seems reasonable to assume that we should not choose a model that assumes the same rate of transmission for each individual we are monitoring. We have shown that this rate varies dependent on age and the fact that the rate of nosocomial transmission may vary between wards is the very reason Nosoco was created. Similarly, we have seen that the rate varies over time with the different waves of nosocomial infections. If we are attempting to understand the rate of infection over a group varied enough in time, location, age or other unknown parameters, we can expect each individual's rate of nosocomial infection to be different, and so a model that reflects the variation would be more appropriate.

However, there are drawbacks to estimating $\eta$, the most notable of which is time. In Section 4.4 we focused on the run-time of Nosoco as one of its strengths. Calculating the posterior distribution for a shared constant rate of infection is simply a matter of calculating $r$ and $W$, the total number of observed infections and the total observation time across all patients respectively. We then form a Gamma distribution from these parameters as our posterior distribution and are able to perform any comparison or analysis we would like. With censored data there is no such closed form for calculating the value of $\eta$. Instead we must estimate its true value by maximising its likelihood function. This is a much slower process, and increases in length with each additional observation.

Additionally, we have no evidence to suggest that an Exponential distribution is the best description for the distribution from which each individual's rate of infection is sampled. Indeed by looking at Figure 4.22 we can start to see a problem with using an Exponential distribution in this way: its variance is inextricably linked to its mean. This means that the range of possible values is also linked to its central estimate. Looking at the ranges of possible rates of infection for individuals above the age of 40, they are all approximately the same, because so too are their central estimates. We did consider choosing a Gamma distribution for our rate of infection. However, in these calculations, the shape parameter $\alpha$ in the resulting Gamma distribution would often be so close to 0 that whilst a distribution with the correct mean rate was achieved, even the 95th percentile would be small enough as to be indistinguishable from 0.

Our final problem with using the Exponential distribution to investigate the range of rates of infection is one of communication. Nosoco is designed to be a tool for individuals who may not be well versed in statistics to use. We found we had difficulty communicating the central estimate for the total number of nosocomial transmissions and even more difficulty explaining the rate of infection. It is our worry that in introducing a distribution for the rate of infection, Nosoco would become too slow and too inaccessible as to be useful to the people it was designed for. For a short enough period of time (i.e. a day to a week), over an enclosed enough region (i.e. a ward) we feel it is appropriate to continue to assume that the rate of infection is constant.

Estimating the rate of nosocomial infection rather than the total number of nosocomial infections allows us a number of advantages when analysing hospital outbreaks. Most notably, the size of the hospital population (and susceptible population in particular) is constantly changing. A larger susceptible population will result in a larger total outbreak under the same exposure with the same preventative measures in use. It is important to know these total numbers for planning scenarios. Unfortunately, identical outbreak scenarios do not occur in real life, so in order to retrospectively compare interventions and scenarios we must use a method of standardisation. Estimating the exponential rate of nosocomial transmission per person can help with this estimation, as demonstrated with our example analysis.

Multiple assumptions had to be made due to lack of available data. We do not know the vaccination status of the individuals in our data set, nor can we guarantee that they have not been infected with SARS-CoV-2 without our knowledge. We instead must assume that they are all completely susceptible unless we have seen their positive swab. Similarly, we know that there are some issues in our data set with pairing swab results to hospital admissions. Unfortunately, not all hospitals use NHS Numbers to record both, so when not present, we used first name, last name and date of birth. This method of identification is liable to administrative error resulting in incorrect pairings of patients. Finally, and this will always be a problem, false negative tests and asymptomatic individuals will mean this data is always a slight underestimate of the true rates of nosocomial transmission. It is difficult to accurately mitigate for this effect, which likely will have changed over time with regular testing and improvement in symptom identification.

With nosocomial transmission making up an important proportion of all transmissions of SARS-CoV-2 (as well as other infectious diseases such as the influenzae) we need a way of meaningfully comparing real-world outbreaks in order to understand the efficacy of intervention. We have shown how our methods can be used to analyse ongoing nosocomial outbreaks. It is our hope that this method can be used to inform analysis of interventions in the future.

## 4.7 Conclusions

In light of the recent global pandemic and the growing evidence that nosocomial transmissions hold an important role in the spread of many communicable diseases, there is a clear need in healthcare for a tool like Nosoco that attempts to identify areas and times of high nosocomial transmission in the hospitals. Anecdotally, in the trust we worked with, the closest equivalent system was to manually enter each identified swab into a Windows Excel spread sheet and by eye attempt to identify areas of high transmission. This method is clearly not practical for large data-sets and is at risk of human error. We hope that Nosoco provides an easy-to-use equivalent system that eases this burden for large trusts.

Nosoco is a relatively simple tool to use. It runs off a Jupyter Notebook, which can be understood with a little instruction. The advantage of the Jupyter Notebook is that we can write this instruction in an accessible way, so that being able to understand the Python language is not a prerequisite for getting a meaningful output.

We feel that Nosoco is an adequately fast software, taking just over a minute to analyse a whole year′s worth of hospital data. For regular monitoring, where we would expect a user to look at most a month′s worth of data, Nosoco can run in under 15 seconds. We feel the Nosoco, or an analytical tool like it, is necessary.

An alternative approach when identifying nosocomial transmissions of SARS-CoV-2 in particular would be through genetic sequencing. The thought process would be that in an outbreak scenario, nosocomial transmissions would likely come from a similar or same lineage and therefore would be genetically similar enough to link together. One problem with this is that genetic sequencing and matching can take a long time, where Nosoco is able to run in a matter of seconds. The COG-UK-HOCI (COVID-19 Genomics UK Hospital Onset COVID-19 Infection) trial aimed to optimise genetic sampling of SARS-CoV-2. They hoped to see if there was a viable pipeline for using genetic data to rapidly assist in nosocomial transmission identification in the UK[193]. They were not able to show a significant change in nosocomial incidence through sequencing alone, but did find that sequencing would affect Infection Prevention and Control interventions if delivered in under 5 days. Closer investigation of the trial data suggests that a delay between positive PCR result and genetic samples reaching appropriate laboratories may in part be a cause of this lack of efficacy. Therefore relying on genetic sampling alone would not be appropriate intervention for hospital trusts with no close access to genomics laboratories[194]. Nosoco easily achieves the threshold window of 5 days and represents a cheap, readily available alternative.

A difficulty we found when designing Nosoco was pairing the hospital admission data with the swab data. In the investigated trust, these data existed on separate databases with varying quality in the accuracy of patient-identifying details. Also, in the same trust, but in different hospitals, different protocols were followed when it came to attaching patient-identifying details to swab results. For example, while one hospital would use a universal NHS Number, another may use a locally recognised District number, making patients transferred between hospitals vulnerable to mislabelling. This all leads to missing or mis-assigned data in Nosoco, which required careful planning and optimisation to minimise. We did not approach handling missing data from a statistical method (for example noting patients who had not had a single swab during their admission and assuming that their swab results must be mislabelled), but such an approach could be considered in the future.

Recently, the same trust has introduced a universal identifier system for all hospitals in the trust, meaning that admissions and swabs should be logged on the same database to the same individual. This should help improve the accuracy of Nosoco. However, data mis-assignment could still occur for patients who may be assigned the wrong or temporary NHS Numbers due to a mishearing of their name or date of birth when being entered into the hospital database. This singles out people for whom English is not a first language or who may have difficulty when speaking or writing clearly.

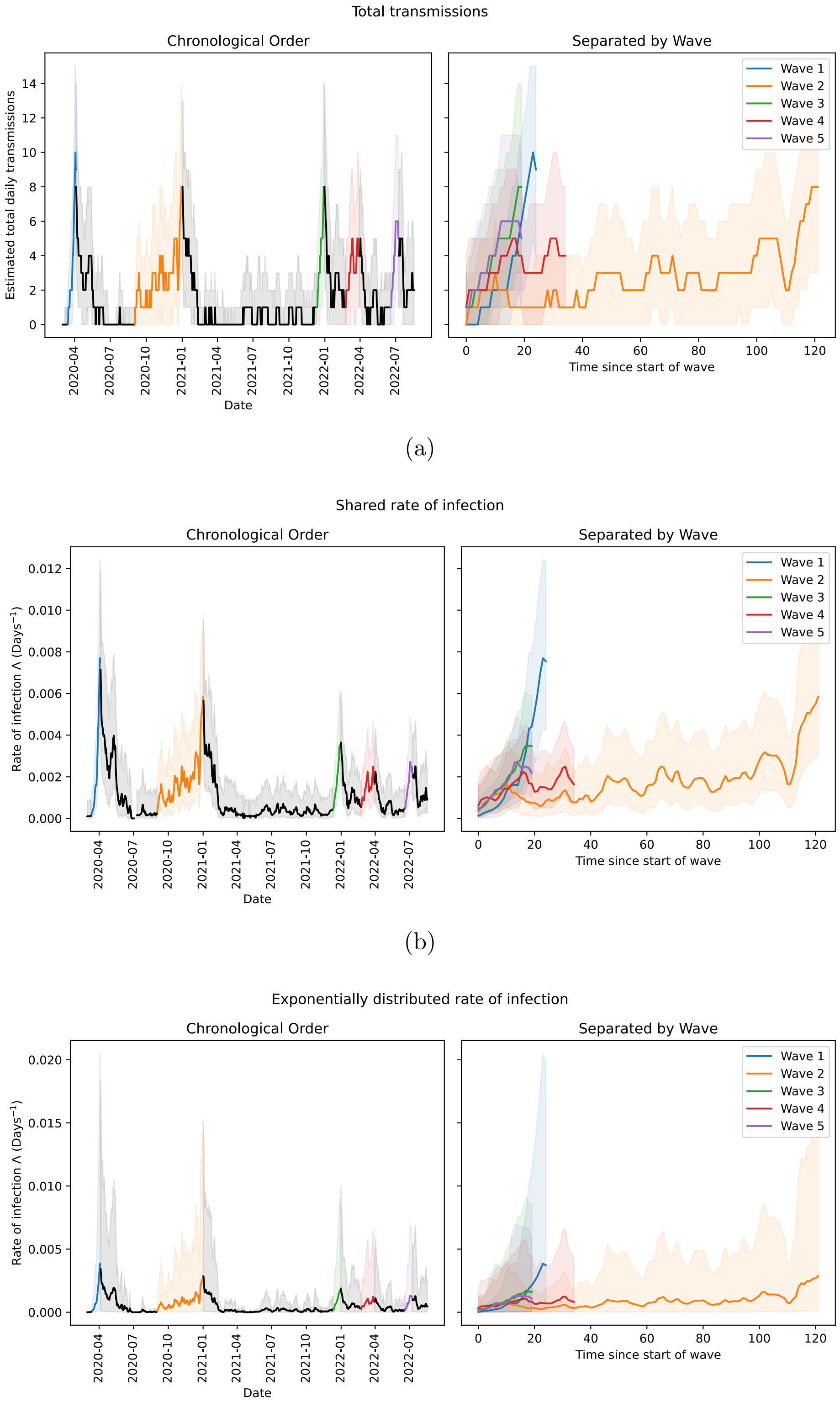
One of two major weakness of Nosoco is the choice of distribution that it uses. We chose to assume that the time from transmission to first positive swab is analogous to the incubation period of SARS-CoV-2 and in doing so justified the use of representing it with a Gamma distribution with a mean of 4.84 days and a standard deviation of 2.79 based on prior studies by Overton et al[133]. We know that this is not true. For example, in studying the early pandemic, Kraemar et al. showed that the average time from infection to diagnosis can decreased from 6.5 days (s.d. 4.2) to 4.8 days (s.d. 3.03) through more active surveillance[195]. Although this example is more looking at scenarios outside the hospital setting, it still shows how a change in monitoring policy can influence the time from infection to diagnosis. There are reasons why this time may be longer than the incubation period (e.g. delay between symptom onset and swabbing, delay in laboratory tests, availability in testing equipment) as well as reasons why it may be shorter (e.g. regular asymptomatic screening, screening in an outbreak scenario). Additionally, this difference may change over time with differing screening protocols and increased availability of tests. Nosoco calculates each probability based on this distribution, so if it is incorrect (and likely a more representative distribution does exist) then so too must be the calculations performed by Nosoco. However, finding the correct distribution is not easy, as in order to do so we would need a data-set where we knew for certain which individuals represent nosocomial transmissions and which represent community transmissions. In the future we hope to be able to access genetic data to perform this exact validation, but this data has not been available as part of our current study.

The second major weakness with Nosoco is true to any method of nosocomial transmission identification: it relies on a diagnosis being made in the first place. As discussed previously, SARS-CoV-2 can result in asymptomatic infections as well as false negative test results. In both cases, these infections would not appear in our data-set. Additionally, in the case of hospitals, there is a whole cohort that do not appear in Nosoco's data-set: hospital staff. Despite good PPE practice, there is no reason to assume that staff cannot become part of a nosocomial transmission tree. In fact, through modelling work performed by Evans et al. it has been shown that, in the UK at least, hospital staff act as vectors for between-patient transmission, and that this actually represents the most important method of nosocomial transmission for SARS-CoV-2[196]. It is not impossible to see how such data can be included in the Nosoco data-set by cross-referencing staff swab results with their shift assignments in the same way we cross-reference patient swab results and their ward admissions. However, this data was not available to us in this study. Also, there would be some difficulty when accounting for staff who can be assigned to multiple wards on one shift, such as porters or on-call doctors.

In both the choice of distribution and the missing data, genetic sampling has an advantage over Nosoco. We do not need to know a distribution for time from transmission to first positive swab to see how closely linked two genetic sequences match. Ellingford et al. developed a model for observing a chain of nosocomial transmission genetically by simply matching individuals in the same location who had a serial interval of 3-7 days[191]. By observing genetic similarities we can also account for missing steps in a chain of transmission such as staff or undiagnosed individuals. A secondary transmission (i.e. the effective transmission between one infected individual and another via a third vector individual) should still result in a close enough genetic similarity to be linked even if we do not have genetic data from the vector. However, if there is enough time between the two individual's first positive swabs, or if they were never spatially linked (i.e. they were never on the same ward at the same time) then they would never be linked by Nosoco as part of the same outbreak.

In an ideal environment, Nosoco could be used in conjunction with genetic testing. Its speed and coverage makes it useful as an early warning for outbreaks, with the slower, more expensive and more accurate genetic testing providing later confirmation of its results. Additionally, genetic testing could be used to continually validate and update probability distributions used by Nosoco. As it stands, Nosoco represents a cheap and fast method of outbreak detection that requires validation before becoming part of regular IPC practice.

Summary:
Nosoco is a Python based tool for estimating the total number of nosoc
that have occurred, based on hospital transmissions and the
timing of infected individual's first positive swab (in the case of
SARS-CoV-2 infection). Nosoco is a fast and effective alternative t
simply counting how many individuals are diagnosed after a certain thr
point in their admission, which is otherwise the standard method used.
relationship between total nosocomial transmissions and total
external transmissions appears to have changed over the first year o
the pandemic, which was likely a result of either improved preventat
measures or earlier diagnosis through improved access to testing. T
older hospital population appear to be more vulnerable to nosocomi
infection, although further work is required to investigate how this changes between individu
this chapter, we have shown that Nosoco is a tool that is useful to b
health care professionals and epidemiologists alike. It offers
real-world up-to-date insight into current outbreaks in hospitals,
as well as opportunities for more complex retrospective analysis
of past outbreaks.

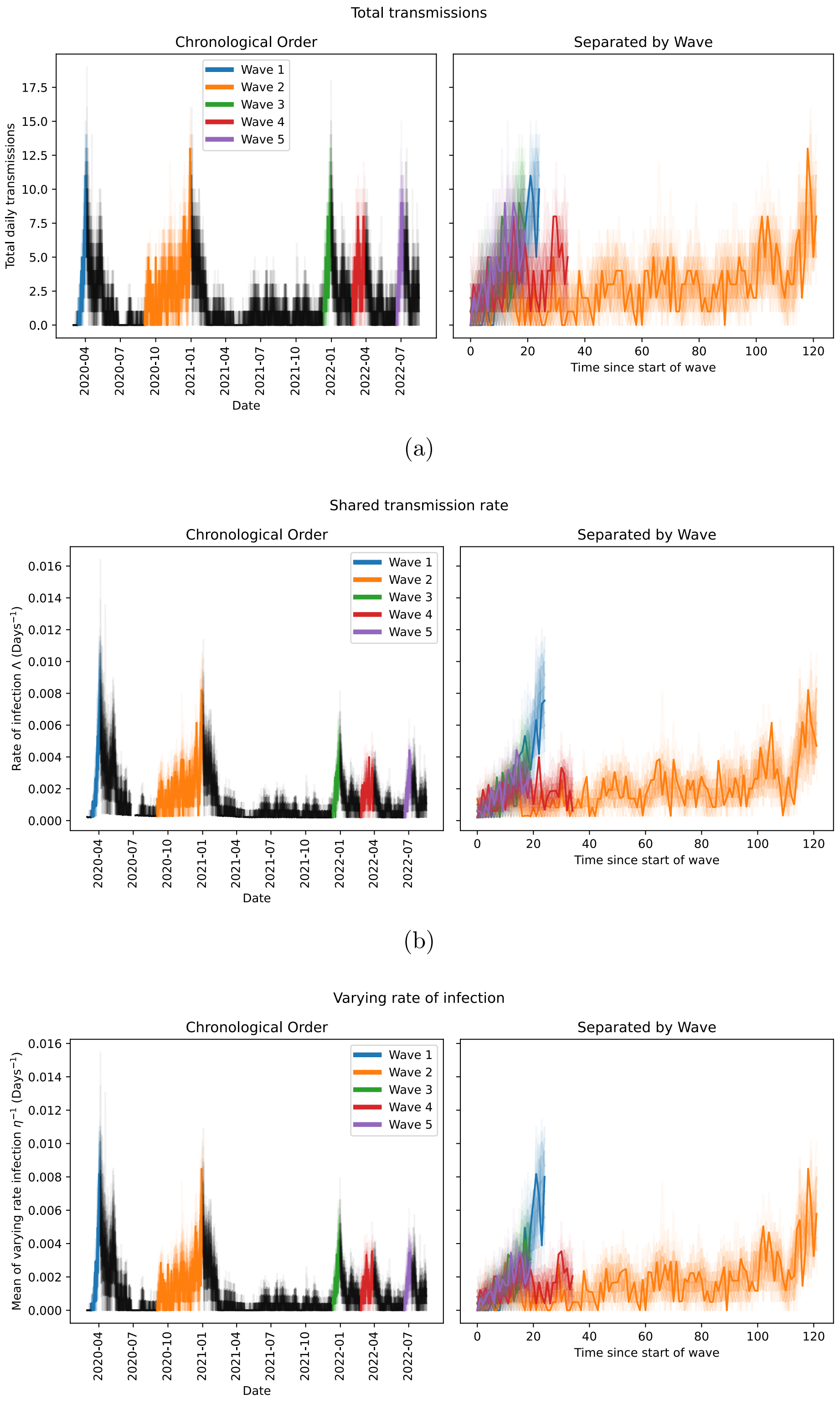


(a)

(b)

(c)

Figure 4.13. Plots of the varying total and rate of nosocomial infections during multiple waves of the SARS-CoV-2 pandemic. Figure 4.13a shows the total daily count, Figure 4.13b shows the median and 95% credible intervals for the rate of infection, assuming each admitted individual experiences the same shared rate of infection, and Figure 4.13c shows the estimated 95% range of rates experienced assuming the rate of nosocomial infection varies from person to person according to an Exponential distribution with rate $\eta$.



(a)

(b)

(c)

Figure 4.14. Spaghetti plots of analyses of the first fifty iterations of Nosoco, showing the total number of nosocomial infections and varying rate of nosocomial infections during multiple waves of the SARS-CoV-2 pandemic. Figure 4.14a shows the total daily count, Figure 4.14b shows the expected estimated rate of infection, $\frac{r+\frac{1}{2}}{W}$, assuming each admitted individual experiences the same rate, and Figure 4.14c shows the mean rate of nosocomial transmissions experienced by individuals, $\frac{1}{\eta}$, assuming they each sample their individual rate of infection from an Exponential distribution with a rate $\eta$. The results from the first iteration are shown in bold.

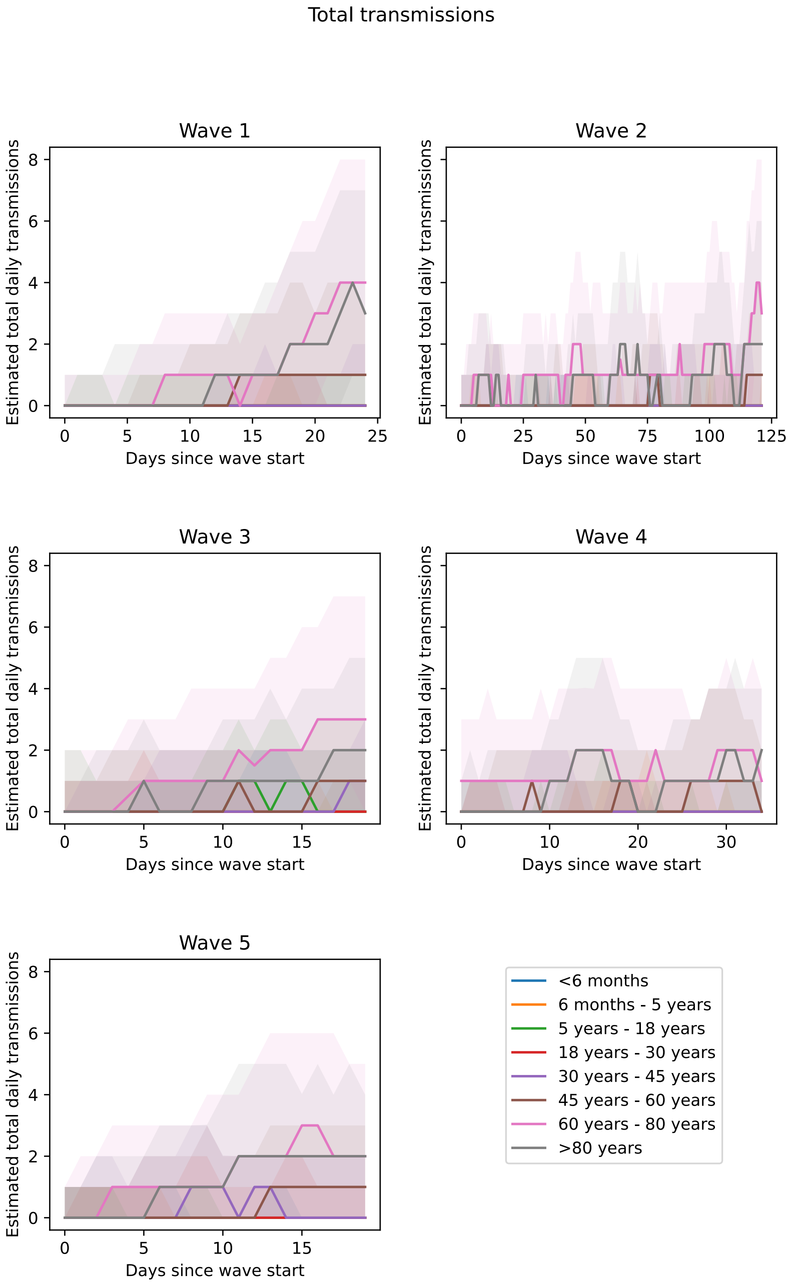


Figure 4.15. Total daily nosocomial transmissions, separated by wave and age group. 95% confidence intervals have been given by observing the variation between random iterations.

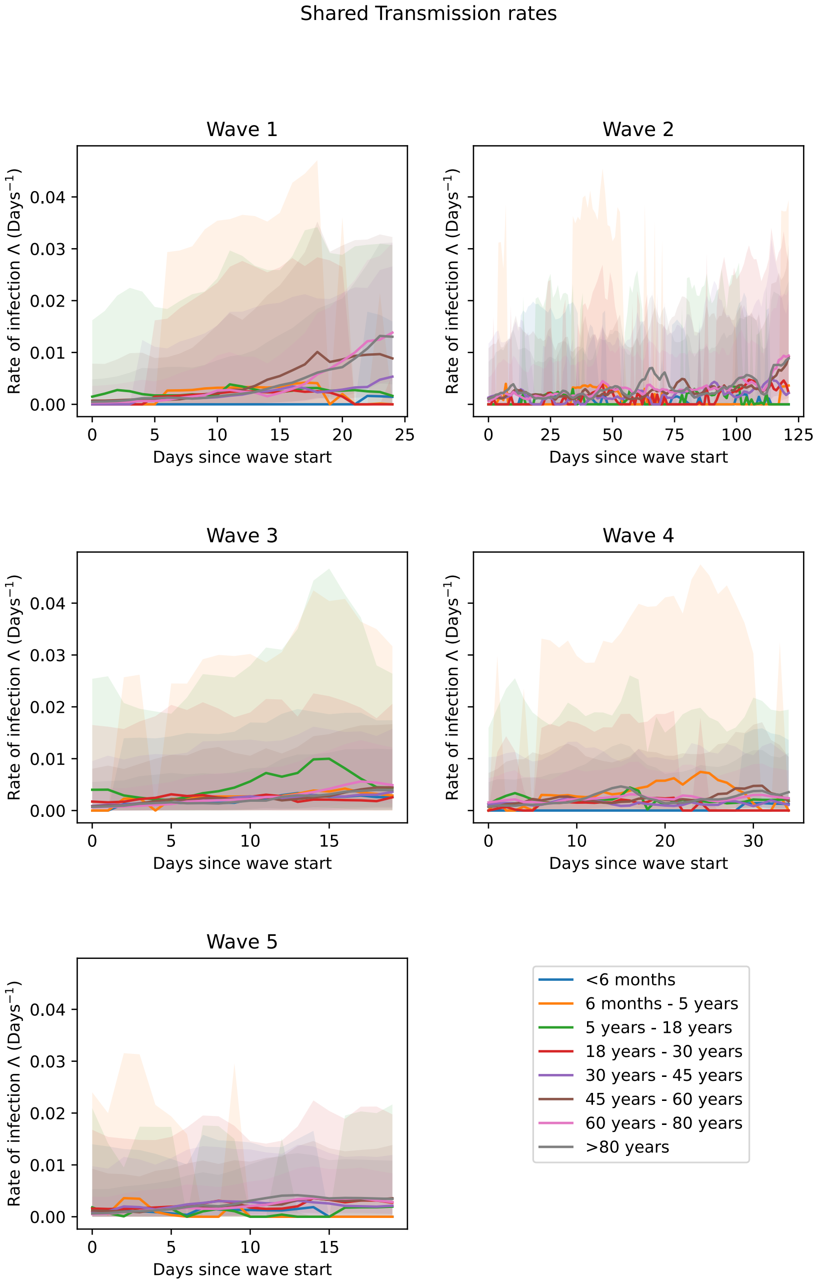


Figure 4.16. Posterior estimates for the shared rate of nosocomial infections experienced by individuals in the same age group. The lines show the median estimate for this shared rate, and windows show the 95% credible intervals.

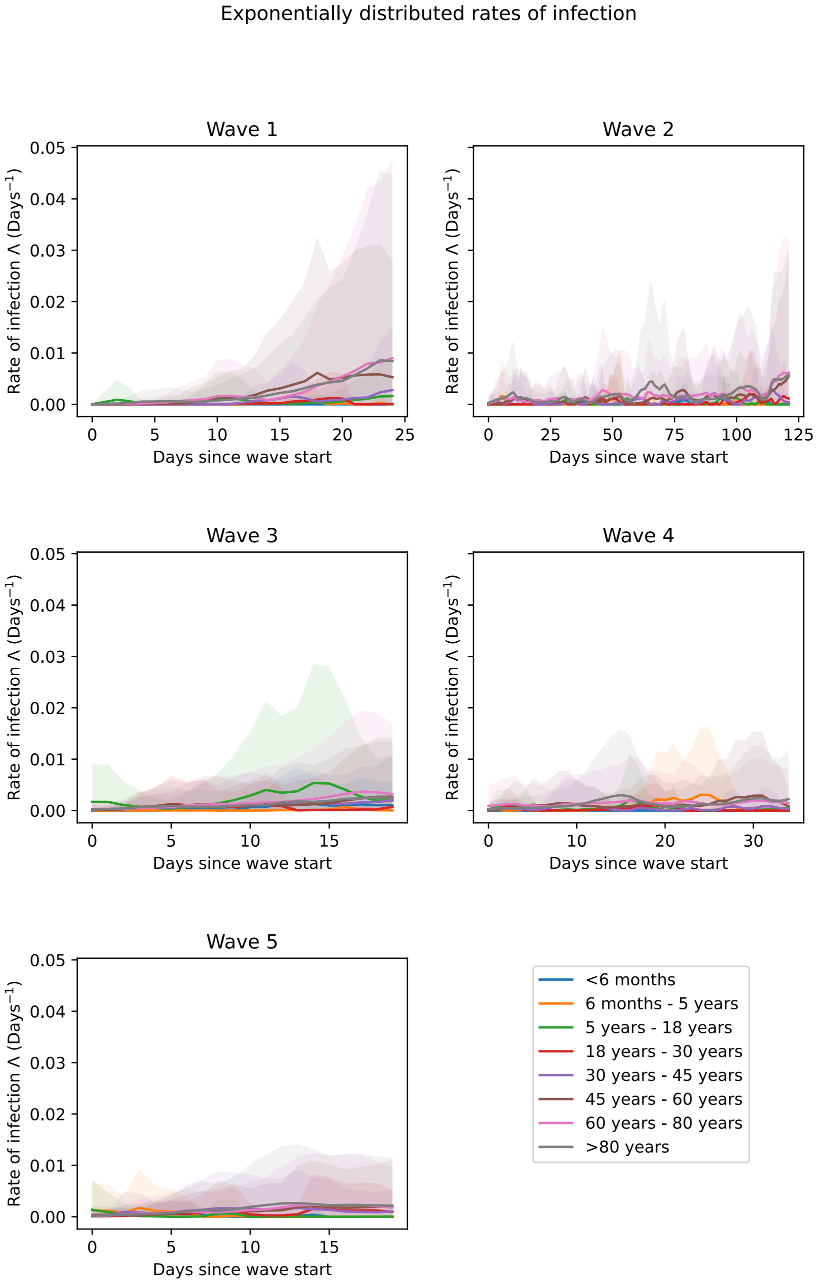


Figure 4.17. Exponentially varying rates of infection, separated by wave and age group. As stated previously, the 95% windows estimate the range of exponential rates of nosocomial infections individuals experience within each age group.

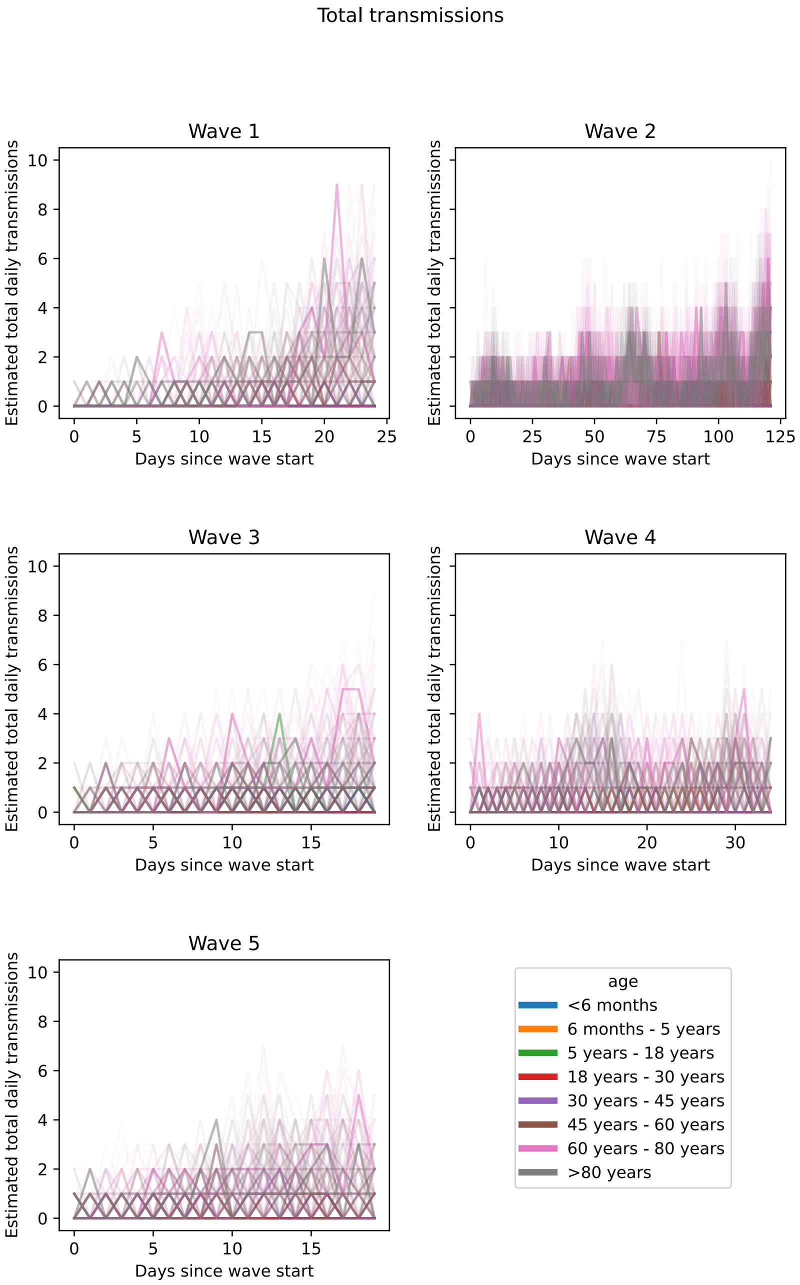


Figure 4.18. Spaghetti plots of the total number of nosocomial transmissions estimated by the first 50 iterations of Nosoco, divided by outbreak wave and age group. The results of the first iteration are shown in bold.

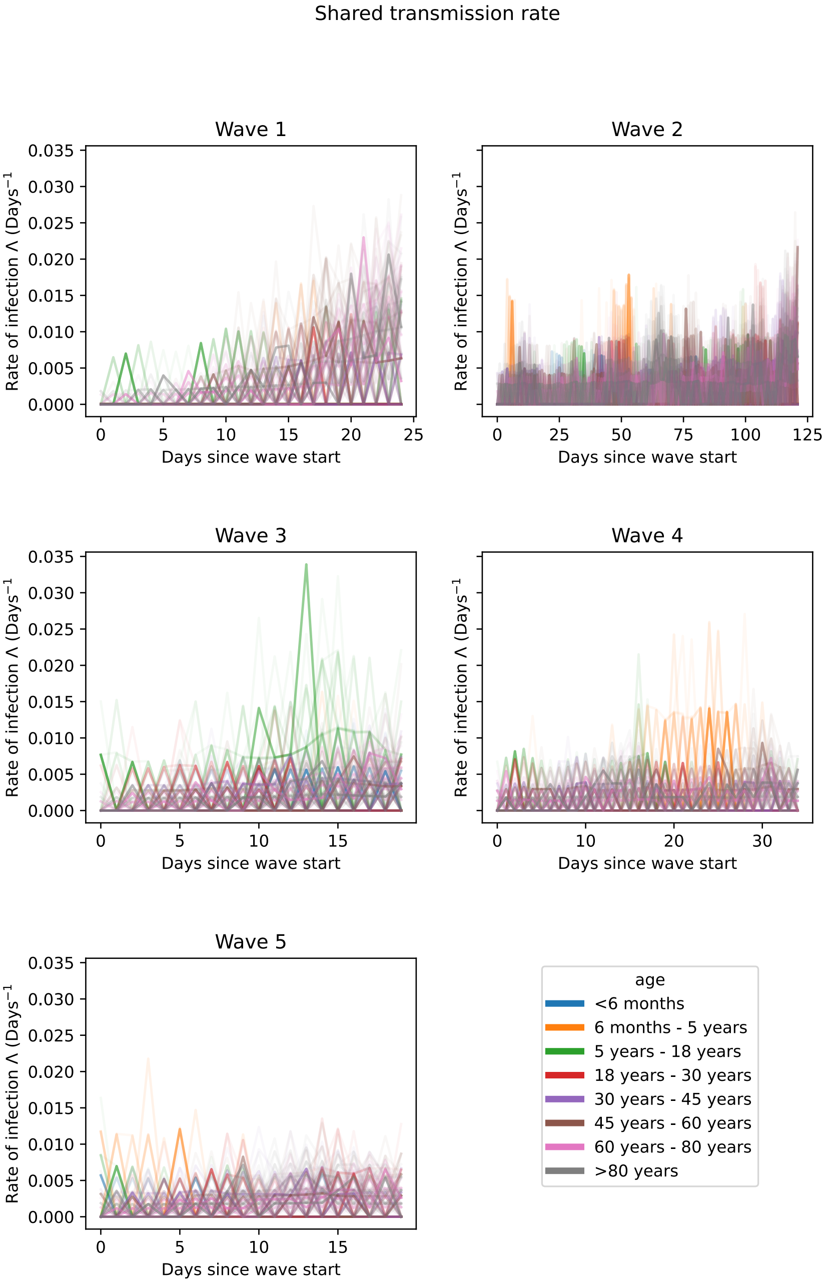


Figure 4.19. Spaghetti plots of the central estimates for the rate of infection shared within age groups. This shows results from the first 50 iterations of Nosoco, separated by wave. The results of the first iteration are shown in bold.

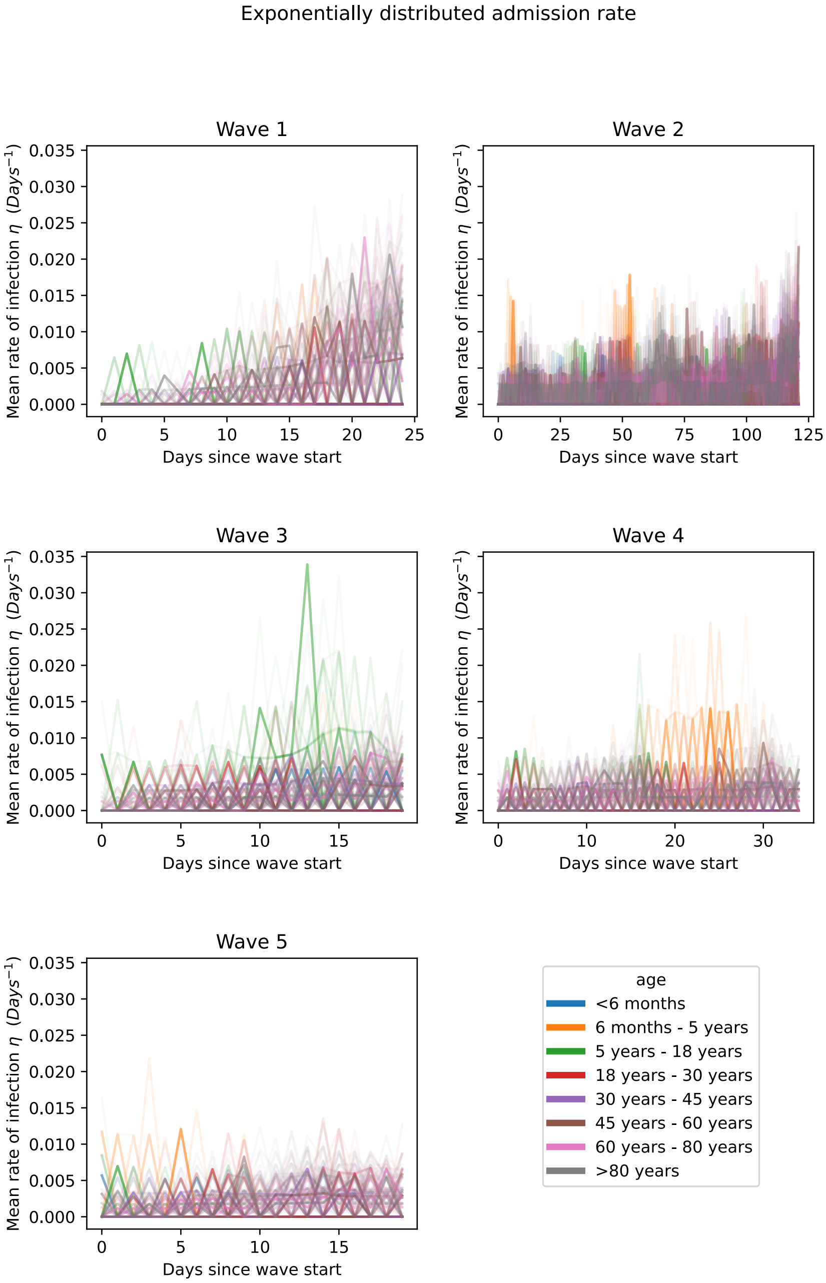


Figure 4.20. Spaghetti plots of the expected rate of infection experienced given it is sampled from a single exponential distribution within age groups. This shows results from the first 50 iterations of Nosoco, separated by wave. The results of the first iteration are shown in bold.

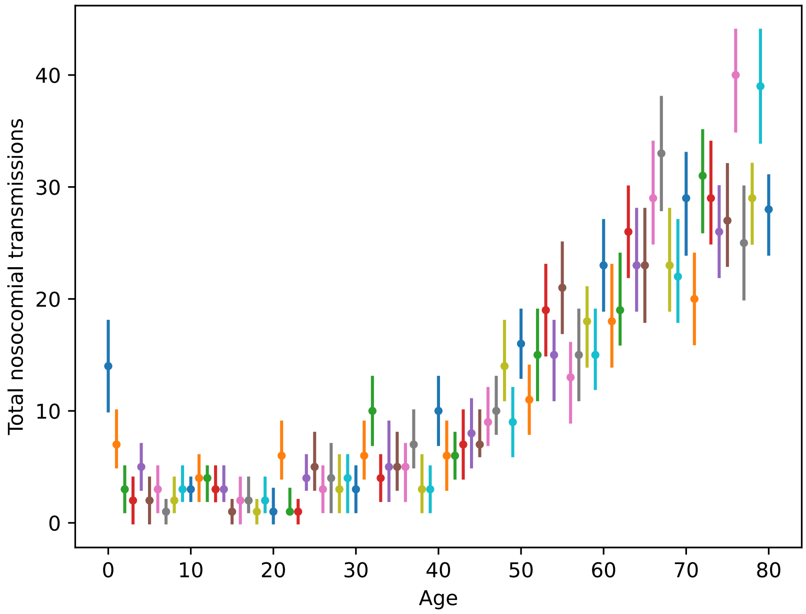


Figure 4.21. Estimated total number of nosocomial transmissions across our entire observation period, separated by age. 95% confidence intervals have been calculated through 1000 iterations of Nosoco.

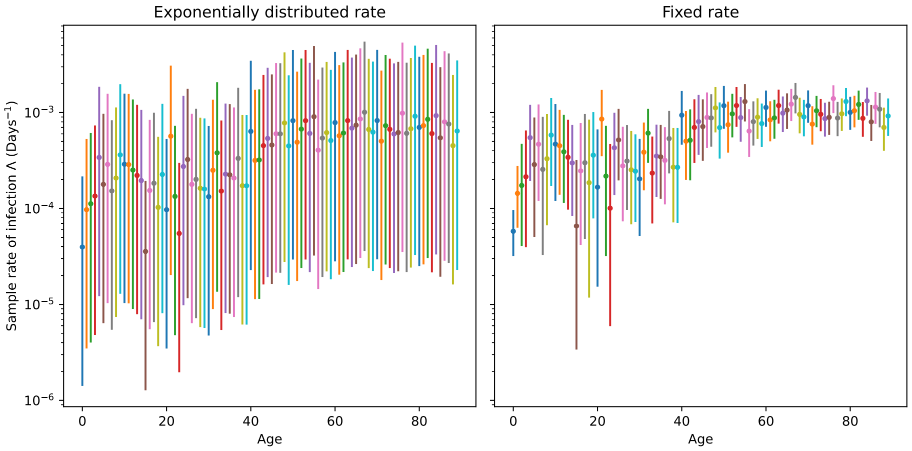


Figure 4.22. Comparison of the posterior distribution for a fixed shared rate of infection to the range of possible rates for an exponentially distributed distribution when separated in yearly age groups. The left-hand plot the estimated 95% range and the median for the rates of nosocomial infection an individual could have experienced during this time period given their age. The right-hand plot gives a posterior distribution for the one rate of nosocomial infection everyone of that age experienced during our observation window. Notice the logarithmic y-scale.

# Chapter 5

# Assessing incidence calculation in a closed environment for diseases with long incubation periods: A case study of Hepatitis C in prisons

## 5.1 Introduction

In terms of outbreak modelling, the prison environment falls somewhere between completely closed environments (such as ships in which no new individuals can enter once away from port) and completely open environments (such as cities and hospitals where new individuals are free to come and go as they please). Careful consideration must be made when modelling and understanding the spread of diseases in prisons.

A meta-analysis performed by Dolan et al. looked at published papers on the epidemiology of HIV, Hepatitis B, Hepatitis C and tuberculosis in prisons globally. They identified 299 publications covering 196 countries from 2005 to 2016. They demonstrated an elevated prevalence of each of these diseases within prisons almost universally when compared to the general public, with multiple individual outbreaks demonstrated of HIV, Hepatitis B and tuberculosis. They therefore showed the importance of these institutions when considering infectious disease modelling. Additionally they attempted to use this parameterisation to approximate the infection rate of HIV in prisons, and showed that reducing incarceration rates could contribute considerably to reducing the spread of this disease, a conclusion that could be applied to the other diseases as well[1].

Special effort has been made to look at the spread of infectious diseases in prisons among those who report injecting drugs. For example, Altice et al. observed a higher incidence of HIV and tuberculosis in Eastern European and Central Asia prisons among PWID[86]. They found that incarceration may responsible for up 75% of all TB incidence for people who inject drugs, and 28-55% of all new cases of HIV in the EECA over 15 years from 2016, a difference from the previous studies which needs to be explored further.

PWID are not the only high-risk group who are more likely to be incarcerated, and the over-representation of certain demographic groups inside of prison has been shown to have knock-on public health impacts outside of prisons. Adams et al. used TITAN, a network driven HIV spread modelling tool[49] to explore the effect of increased incarceration on the African American community in Philadelphia[87]. They found that an increased rate of incarceration in the male Africa American population had a knock-on effect of an increased HIV risk for the female African American community when they left prison, showing the importance of the external effects of closed environments.

This spill-over effect was investigated further by Mabud when looking at tuberculosis in prisons in Brazil[88]. They collected data looking at incidence of TB given time in prison, including incidence at time of release. They then used this data to parameterise a compartmental model looking at the effect of people moving in and out of hospital. Annual mass TB screening in prisons reduced this models in-prison TB incidence by 47.4% and out-of-prison incidence by 19.4%.

Additionally, the health outcomes for people with infectious diseases who are incarcerated have also been investigated. For example, Cohen et al. showed increased odds of mortality and poor HIV outcomes for women who are incarcerated[92].

Given the shown importance of incarceration in the spread of infectious diseases, Ndeffo-Mbah et al. performed a systematic review looking for papers published between 1970 and 2017 in English that model the spread of one or more diseases and include incarceration as part of their model[95]. In total they found 34 models published over this time period, some of which have already been mentioned in this paper, looking at HIV, TB, HCV and sexually transmitted diseases. The overview of these models showed the impact of incarceration. In communities of people who inject drugs, HIV prevalence greater than 5% resulted in incarcerations being linked to 12-55% of HIV incidence[1], [86]. They also noted that parameter uncertainty was only accounted for in 14 of the 34 models, of which only 8 performed a sensitivity analysis and model fitting and only 1 showed model validation. This gives a good indication as to what is required for future modelling of infectious disease models including incarceration incarceration.

Pitcher et al. performed a systematic review in 2019 looking at modelling efforts towards specifically the spread of Hepatitis C[2]. They collated papers looking at incarceration and co-infection with HIV. Models currently do not necessarily show a positive outlook for elimination in a decades time. However, incarceration takes a higher proportion of PWID and in turn Hepatitis C positive patients and therefore represents an opportunity for targeted treatment.

Martin et al. performed an extensive review of treatment of PWID with Hepatitis C over 2014 in 7 different sites[53]. They found a range of treatment levels (maximum 26/1000 PWID/year) and used this to support a model evaluating upcoming treatment changes. They found that the maximum treatment level (26/1000) is needed to effect any change over the next 10 years, but the more the better.

A more extensive and in-depth model was created later by the same group, in particular noting the increase in QALYs following multiple prison-based interventions[197]. This complicated model had over 1000 compartments based on age, injection history etc., giving it the risk of over fitting. Additionally, the parameters used were often estimates owing to multiple gaps in data. They found that increasing screening for Hepatitis C in prisons dramatically increased the efficacy of in-prison treatment plans. Although prisons represent a good opportunity to identify Hepatitis C positive individuals, a short incarceration time means they may be lost to follow-up and therefore not complete their treatment regimen. Martin et al. showed the effect of earlier in-prison diagnosis on preventing this from happening and therefore increasing overall treatment efficacy.

Another model looking specifically at increased screening rates of Hepatitis C in prisons was created by He et al[54]. Once again they found that increasing the screening rates in USA prisons, and screening indiscriminately with regards to age increased diagnosis and treatment in a cost-effective manner (more cost-effective than targeting those born between 1945-1965). They were limited in the parameters they knew compared to those they had to infer.

Alongside the study performed by Taylor et al. which is the focus of this paper[81], others have attempted to estimate the incidence of Hepatitis C in prisons around the world, including Australia[198] and Ireland[199]. The results from these studies appear to conflict greatly, with Crowley et al. finding limited evidence of any Hepatitis C transmissions and Cunningham et al. finding an incidence in keeping with demographic trends outside of prisons. The discrepancy could be caused by differences between countries. However, we should look more closely into how incidence is calculated. We will use the study by Taylor et al. to show how studies are at risk of underestimating incidence based on non-longitudinal studies if models of transmission are not considered.

In December 2012, Addiction published the study performed by Taylor et al. reporting the Low incidence of Hepatitis C virus among prisoners in Scotland[81]. In this study, 5187 prison occupants across 14 closed Scottish prisons were surveyed on various aspects of their prison life, including age, gender, drug use, sexual activity etc. They then were consented for blood tests for HCV RNA via PCR and HCV antibodies using a modified protocol for the Ortho HCV 3.0 SAVe enzyme-linked immunosorbent assay (ELISA) (product number940982; Ortho Diagnostics, Amersham, UK)[200]. Upon infection with HCV, it takes approximately 2 weeks for the viral RNA to be detectable via PCR. From this point it is unclear how long it takes for the body to establish antibodies, but it has been theorised that the time is somewhere between 51 and 75 days, so the study investigated both extremes.

With this gap in time between having detectable levels of HCV RNA and HCV antibodies, an HCV RNA positive individual without antibodies indicated a relatively recent infection. This information was used by Taylor et al. to estimate the incidence of Hepatitis C through the following formula:

$$I = \frac{\left(\frac{365}{T}\right) n}{(N - n) + \left(\frac{365}{T}\right) n} \tag{5.1}$$

where $I$ is the incidence among susceptible individuals, $T$ is the estimated duration of the RNA-positive antibody-negative window, $n$ is the total number of individuals found to be RNA-positive and antibody-negative when tested and $N$ is the total number of susceptible individuals. As at the time of testing they were only observing transmissions among people who were infected in a time period of length $T$, they scaled up this number, $n$ to the length of a year by multiplying it by $\frac{365}{T}$. This scaling up, whilst simple, does not take account of the exponential nature of rate-dependent occurrence such as endemic infection. We shall discuss an alternative method of estimating incidence in the rest of this chapter in the rest of this chapter.

Of the 5187 individuals tested, 2446 provided adequate blood samples that did not contain HCV antibodies, 3 of which were RNA positive, equating to an incidence between 0.006 and 0.009 infections per susceptible individual per year according to their calculations. 479 antibody-negative prisoners were PWID, of which two were RNA-positive, equating to an apparent incidence between 0.020 and 0.030. 91 antibody negative prisoners reported injecting drugs while in prison, of which 1 was RNA positive ($0.051 \leq I \leq 0.075$). The estimated incidence of HCV among Scottish people who inject drugs outside of prison is 0.120 per person per year, forcing the authors to conclude a reduced incidence inside prison. Given a high prevalence, this conclusion has had a number of implications around behaviour outside of prisons and where best to deploy interventions in order to reduce the spread of HCV.

Using this lowered incidence combined with the shown increased risk of injection related deaths following release from incarceration[83], [84], Stone et al. attempted to fit known PWID HCV incidence data to a deterministic model that mapped both a person's progress through prison systems as well as HCV contraction and treatment[85]. They supposed that if a person had an increased risk of injection related deaths immediately after leaving prison, they would also have an increased risk of contracting HCV. Their model suggested a 45% decrease of Hepatitis C incidence and chronic infection in Scotland should this risk be reduced compared to a 22% reduction in risk if injecting drugs was legalised. However, this relies on the assumption that there is an elevated risk of Hepatitis C infection immediately after incarceration.

| Parameter/Function | Description |
|---|---|
| $a$ | The number of antibody negative, RNA positive individuals in the data set (identical to $n$) |
| $b$ | The number of antibody negative, RNA negative individuals in the data set |
| $\delta$ | A comparison of the two methods of calculating incidence, defined as $\frac{I_{\text{Taylor}}}{I_\lambda}$ |
| $I$ | The annual incidence of Hepatitis C |
| $I_\lambda$ | The annual incidence of Hepatitis C as calculated through the method proposed in this chapter |
| $I_{\text{Taylor}}$ | The annual incidence of Hepatitis C as calculated through the method used by Taylor et al. |
| $\lambda$ | The rate of Hepatitis C infection |
| $n$ | The number of antibody negative, RNA positive individuals in the data set (identical to $a$) |
| $N$ | The number of antibody negative individuals in the data set |
| $P$ | The probability of still being incarcerated at the time of testing |
| $P_1$ | The probability of still being incarcerated at the time of testing given the results would be antibody negative, RNA positive |
| $P_2$ | The probability of still being incarcerated at the time of testing given the results would be antibody negative, RNA negative |
| $\pi(\theta)$ | The prior distribution for some parameter $\theta$ |
| $\rho$ | The probability that an individual is infected over a time period of length $T$ |
| $T$ | The length of time an infected individual is antibody negative, RNA positive |

Table 5.1. A complete description of parameters and functions used in this chapter

By considering transmission as an exponential process, we demonstrate how the method of scaling used by Taylor et al. potentially introduced errors into their calculation. We provide an alternative simple method of analysis that can be used with any non-longitudinal study to estimate incidence. In turn, we demonstrate that the Taylor et al. study shows no evidence of a reduced incidence of Hepatitis C in prisons.

## 5.2 Parameter and Function description

Table 5.1 is a list of relevant parameter and function definitions for this chapter. The reader may find it useful to refer back to this table as and when required.

## 5.3 Model

We start by considering what it would take to be identified as a positive case (RNA-positive, antibody-negative) in a cross-sectional Hepatitis C survey like the the Taylor et al. study. A susceptible individual is incarcerated. At some point during their stay they are infected with the Hepatitis C virus (either through shared needles, sexual intercourse or other means). Fourteen days pass and they are now RNA-positive, antibody-negative. This will be the case for somewhere between 51 and 75 days from the point of developing antibodies. It is somewhere in this window that they are tested as part of the study and then identified as a recent positive case.

Thought of in another way, an RNA-positive antibody-negative individual was infected some-when in a window between 14 days ago and either 65 or 89 days ago. The Taylor et al. study excluded anyone who had been incarcerated for less than 75 days in order to avoid mistaking an external infection for one that occurred inside the prison. Although there is still a slight overlap in the exclusion criteria, we are going to choose to ignore is and assume that everyone who is a positive case was infected whilst incarcerated.

This means that by observing a positive case, we are observing a 51 to 75 day window over which time they are infected. Conversely, any RNA-negative antibody-negative individual that we have observed have gone through this exact same time window and did not get infected. We can use the number of positive and negative cases we observe to estimate any individual's probability of being infected over this time period.

But if $\rho$ is the probability of being infected over either 51 or 75 days, how can this be extrapolated to estimate the annual incidence of Hepatitis C in prisons? This depends on your definition of incidence. Taylor et al. argue that it can be calculated by scaling up the number of positive cases. If $\rho = \frac{n}{N}$ calculates the probability of being infected over $T$ days, where $n$ is the number of positive cases and $N$ is the total number of antibody-negative individuals (the total number of prisoners that were susceptible at the start of the effective window of time that we are observing), then by scaling $n$ up by a factor of $\frac{365}{T}$ we get the annual incidence $I = \frac{(\frac{365}{T})n}{(N-n)+(\frac{365}{T})n}$ (see Equation 5.1).

If we assume that the incidence of Hepatitis C infections is relatively stable when averaged across multiple sites, a susceptible individual's time from start of incarceration to infection would fall on an exponential distribution. We use $\lambda$ to represent the rate of this exponential distribution, where $\frac{1}{\lambda}$ is the expected time from admission to infection in years. In fact, as exponential distributions are memory-less, the average time from any observed negative state to infection is $\frac{1}{\lambda}$. We can therefore calculate $\rho$ in terms of $\lambda$ and $T$:

$$\rho = 1 - \exp\left[-\lambda \frac{T}{365.25}\right]$$

We use 365.25 to average across leap years as well, although it is unlikely to have a dramatic effect on our calculations. If we can estimate $\rho$, we can also estimate $\lambda$.

The rate of infection is one possible measure of the incidence of Hepatitis C in prisons. However it is not entirely analogous to the proportion of susceptible individuals who will be infected after one year of exposure. For example, if $\lambda = 1$, resulting in an expected time from incarceration to infection of one year, the proportion of individuals who would be infected after 1 year would be $I = 1 - \exp[-1] = 0.632$. The rate of infection in terms of a rate of an exponential process has the advantage that it remains constant over any time period. However, it can be difficult to understand conceptually, and so converting it to be in terms of the proportion we would be expected to be infected over a year can be useful. In this case, $I = 1 - \exp[-\lambda]$. If we can calculate $\rho$ we can calculate $\lambda$ and if we can calculate $\lambda$ we can calculate $I$.

But how will we calculate $\rho$? Rather than taking Taylor et al.'s frequentest approach of assuming that we can observe the exact value of $\rho$, we approach from a Bayesian angle. This has the advantage that we can demonstrate a range of possible values for $\rho$ and account for any lack of statistical power. Our prior distribution for $\rho$ is a Jeffreys' prior $\pi(\rho) = \rho^{-\frac{1}{2}}(1-\rho)^{-\frac{1}{2}}$ which minimises the influence of our prior expectations of $\rho$ and remains invariant in our choice on parameterisation[145], [147] (as discussed in the Chapter 2). This means that the effect of our prior remains minimised when considering estimates for $\lambda$ and $I$.

If we observed $a$ RNA-positive antibody-negative individuals and $b$ RNA-negative antibody-positive individuals out of a particular subset, then given our prior distribution, the posterior distribution for $\rho$ falls on a Beta distribution with input values $a + \frac{1}{2}, b + \frac{1}{2}$. From this distribution we work out the 2.5, 50th and 97.5 centiles for $\rho$, where the 50th centile is the median values and the 2.5 and 97.5 centiles act as 96% confidence intervals. From these we calculate median and 95% credible intervals for $\lambda$ and $I$.

Additionally, we can observe how our estimate for the rate of infection may differ from the evaluation performed by Taylor et al.. We can insert $\lambda$ into Equation 5.1 to show their estimated incidence in terms of the rate of infection:

$$I_{\text{Taylor}} = \frac{1 - \exp[-\lambda T]}{1 - \exp[-\lambda T]\left(1 - \frac{T}{365.25}\right)}$$

In this case, $\lambda$ is given as a rate of infection per day to maintain similarity to the original equation, although this is low enough that in our results we give it terms of years. This calculation of the incidence of Hepatitis C is independent of $N$, the sample size. In other words, allowing for the reduction of stochastic random events that comes with increasing the sample size, this calculation is not directly influenced by a smaller or larger value of $N$. We can see how our two methods of calculating incidence compare by defining a parameter function $\delta$ as the estimated incidence divided by the modelled incidence ($I_\lambda$):

$$\delta = \frac{I_{\text{Taylor}}}{I_\lambda} = \frac{1 - \exp[-\lambda T]}{\left(1 - \exp[-\lambda T]\left(1 - \frac{T}{365.25}\right)\right)\left(1 - \exp[-\lambda 365.25]\right)}$$

In this case, if $\delta < 1$, the incidence estimated in the original paper will be less than the our estimate of the incidence and if $\delta > 1$, their estimated incidence will be greater than ours.

## 5.4 Results

Tables 5.2 and 5.3 show our estimates for the exponential rate of Hepatitis C infections and its annual incidence in prisons based on the study by Taylor et al.. Their study only found 3 RNA-positive antibody-negative individuals that fit their inclusion criteria. Of those three, only 2 reported ever injecting drugs, 1 reported injecting drugs in prison and none of them reported injecting drugs during their current incarceration. With a frequentest approach we would estimate the incidence to be 0 among people who inject drugs in prison. However, if we take a Bayesian approach to find a credible interval for the incidence, the study simply does not have a large enough sample size to conclude that the incidence of Hepatitis C among people who inject drugs in prisons is lower than those who inject drugs outside of prison (0.120). There is sufficient evidence to say that the in-prison incidence is reduced among people who report ever injecting drugs, but this is not true among people who report injecting drugs during this or any incarceration.

Indeed, if we were to assume that all the positive cases were due to injecting drugs, as was suggested in the original paper, and that the window of detection was 51 days rather than 75, there would be enough evidence to conclude with a greater than 95% certainty that the rate of infection is elevated while in prison. However, assuming that the positive cases falsely reported their drug-taking status needs to be coupled with the assumption that some negative cases also falsely reported their drug-taking status. With only three positive cases, we do not have enough data to make an accurate estimate of the probability that you would report drug taking given that you are drug taking. We therefore cannot accurately adjust the value of $b$ to match the changes our assumptions would make to the value of $a$.

Figure 5.1 shows the range for the estimated annual incidence of Hepatitis C among the different demographics investigated in prison. There is not enough evidence to show that individuals who report ever having injected in prison or individuals who report having injected drugs during their current incarceration have a decreased (or increased) incidence of Hepatitis C when compared to PWID outside of prison, as this external incidence is within the credible interval range for these demographics.

Figure 5.2 shows how the difference between our two methods of calculating incidence changes dependent on both the rate of infection, $\lambda$, and the length of the window of observation, $T$. This difference is given in terms of $\frac{I_{\text{Taylor}}}{I_\lambda}$, meaning that a value below 1 indicates the estimate method in the original paper is less than our method. We can see that the Taylor et al. method will tend to result in a lower estimate for the incidence of Hepatitis C infection. At an annual rate of approximately

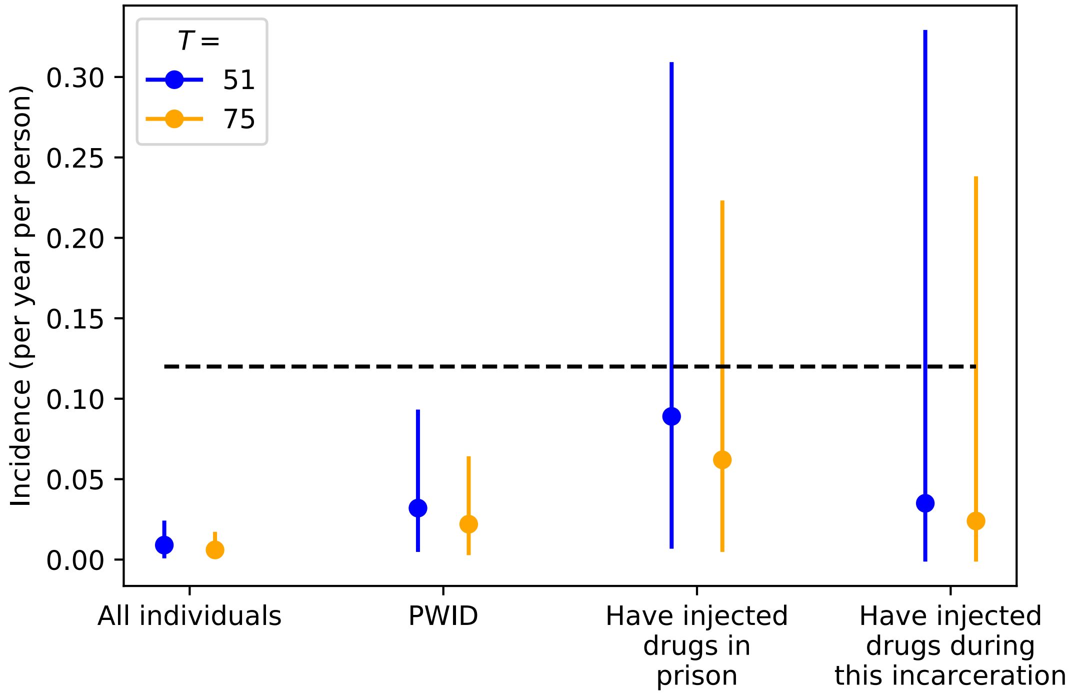


Figure 5.1. 95% Credible intervals for the annual incidence of Hepatitis in prisons according to the study performed by Taylor et al.. The spots show the median value for our posterior distributions. The black dotted line shows the estimated Hepatitis C incidence among people who inject drugs outside of prison at the time of the study. $T$ is the number of days an infected individual is expected to be RNA-positive antibody-negative, and therefore the window of time over which an infected individual can be identified as “recent”.

$\lambda = 2$, this difference is at its minimum (resulting at the greatest underestimation). Larger observation windows will tend to improve the accuracy of this estimation method.

## 5.5 Discussion

By and large, although we have taken a different approach, our central estimates for the incidence of Hepatitis C in prisons are in keeping with the estimates generated by Taylor et al.. The similarity is a product of how relatively low the true rate of infection is. Figure 5.2 shows how this difference changes with respect to $\lambda$, the rate of infection. If the rate of infection is 0, $\delta = 1$ (i.e. the estimated and true incidences are the same). This is expected, as if the rate is 0, the frequentest estimate for the incidence should also be 0 as no infections will be observed. For any rate of infection $\lambda$ greater than 0, the estimated incidence will be less than the true rate of infection, with a maximal difference at a value of $\lambda$ just below 2 years$^{-1}$. Regardless of the rate of infection, we would expect their estimate for the incidence to be within a fifth of our estimate. Dependent on the true value of $\lambda$, we would expect the $\delta$ value to be anywhere between 0.810 and 0.998. In general, this method of incidence estimation, whilst not mathematically optimal, does calculate the incidence to within an acceptable standard.

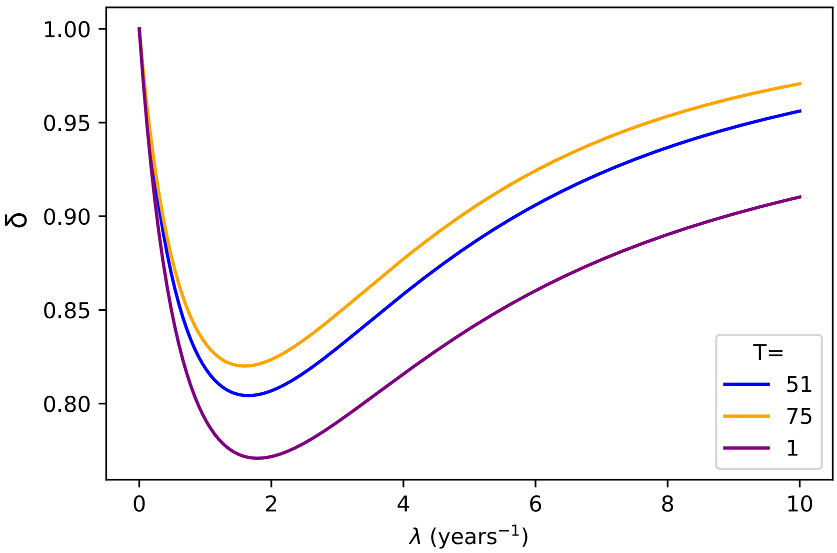


Figure 5.2. Difference in annual incidence estimation, as defined as the incidence calculated using the method in the original paper divided by the incidence calculated using a rate-based method, and its change with different observation windows ($T$) and rates of infection ($\lambda$).

It is the introduction of a Bayesian approach when interpreting the study results that has really improved our estimation of the annual incidence of Hepatitis C in prisons. In particular, it allows us to account for the low power result from the under-representation of certain demographics in the study (in particular the fact that only 45 individuals identified themselves as having injected drugs during their current incarceration). Without accounting for a low denominator, it is tempting to conclude that the rate of infection is zero. However, this would be wrong and implies that there is no risk of injecting drugs in prison. Instead, we conclude that there is no evidence of a decreased risk of Hepatitis C infection if one is injecting drugs in prison. There is enough evidence to conclude that there is a reduced risk among those who have ever injected drugs when compared to those who inject drugs outside of prison. We would expect some individuals who have ever injected drugs before to not be currently injecting drugs and therefore not currently be at risk of Hepatitis C inoculation through this method. This effect may be increased in prison due to lack of access to drugs. Of those who have ever injected drugs, only approximately a fifth ($\frac{90}{477}$) had ever injected drugs in prison. Self-reporting may be a confounding issue here, as an incarcerated individual may be more likely to report that they have ever injected drugs than that they have ever injected drugs in prison from fear to repercussions. If we ignore this factor, reduced injecting rates while in hospital would explain the reduction of risk among this group.

There may be a reason why even our calculation of the incidence is an underestimation. Figure 5.3 is a flow diagram demonstrating who appeared in the study data. As stated previously, testing for RNA-positive antibody-negative status is equivalent to testing if the individual was infected between 14 and $T - 14$ days ago. Of all of those who were incarcerated at the start of this time period, only those who were still incarcerated at the time of testing would be included in the study. We say that the probability of still being incarcerated is $P$, or $P_1$ for individuals who were infected during this time window and $P_2$ for those who were not.

On the surface, it is reasonable to assume that $P_1 = P_2 = P$ i.e. that there is no relationship between an individual's probability of being infected and probability of their incarceration period ending. There should be no direct influence between the two. However, if we consider injecting drugs as both a risk factor for Hepatitis C infection and an indication as to the type of crime that may have led to their incarceration, then this changes. It is not unreasonable to observe that PWID are more likely to have shorter sentences for "petty" crimes (to the author's best knowledge such studies have not been performed but are definitely needed). A shorter incarceration time will reduce $P_1$ when compared to $P_2$ resulting in an under-representation of infected individuals in the data-set. This is not the case when only observing individuals who actively inject drugs, but unfortunately the size of this observed group was too small to draw any meaningful conclusions from it.

As an additional wrinkle, it is unlikely that either group's incarceration time distributions will be Exponential, so any attempts to accurately account for this discrepancy would have to compare these distributions to the time delay between infection window and testing carefully. Indeed, the data from Taylor et al. show at very least an over-dispersed sentence length among the sampled population. The median sentence length was 0.79 years. However, 24% had a sentence length greater than 4 years while the remaining 76% had a sentence length less than one year. Of course there is not a 1:1 correlation between sentence length and incarceration time and by sampling current incarcerations we generate a bias towards longer incarcerations but this is still fairly indicative of a non-Exponential distribution. This information would be important when trying to estimate $P_1$ and $P_2$. In an Exponential distribution, the time until an individual leaves a system is independent of the time they have already been observed in system (i.e. if $T$ is Exponentially distributed, $\mathbb{P}(T > t+n | T > n)$ is a constant regardless of the value of $n$). In a non-Exponentially distributed system, such as length of prison incarceration, the rate of discharge is not independent of the length of time that has passed. Therefore in approximating $P_1$ and $P_2$ we would need to pay special attention to how much time has passed in our experiment.

The earliest occurrence we can find of the method used by Taylor et al. to estimate the incidence of Hepatitis C is a sero-surveillance study performed by Hope et al. in Bristol in 2006 (although the paper itself was published in 2010)[82]. In this paper they stated that the wide range of possible parameterisations for the observation window length (between 51-75 days) far out-stretched the uncertainty introduced by the confidence intervals in their data, so they did not need to report their confidence intervals. The study in question only included 299 participants, which is far out-stretched by the 2446 blood samples taken in the Taylor et al. study, so it would

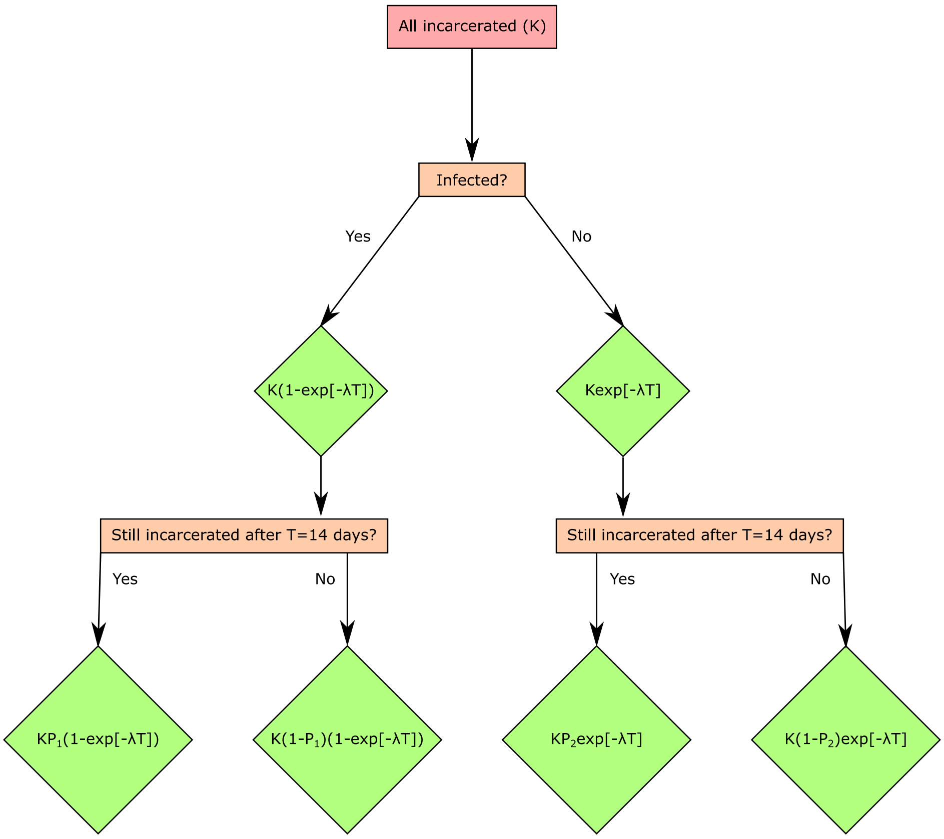


Figure 5.3. Flow diagram demonstrating who is included in the Taylor et al. study. Individuals are exposed over a time period of $T$. Anyone infected during this window who is tested at the testing time will be RNA-positive antibody negative. Given that they remain incarcerated during this exposure period, their probability of infection is $1 - \exp[-\lambda T]$. There is then a delay between the end of that window and the test time. Their over all probability of still being incarcerated at the testing time is $P_1$ for individuals infected in the window and $P_2$ for those that were susceptible but not infected in this window. We have assumed that $P_1 = P_2$ (i.e. that an individual's probability of still being incarcerated is independent of if they have been infected or not.

seem reasonable to conclude that Taylor et al. and other studies like it need not include their confidence intervals either. However, once we start breaking down demographics in this data set, such as only 45 individuals reporting actively injecting drugs during their current admission, we have shown this assertion no longer stands.

In total, after the Hope et al. study, this equation for estimating the incidence of Hepatitis C from the prevalence of RNA-positive antibody-negative individuals in a cross-sectional study has appeared in five other papers. McCauly et al. used it to estimate the incidence of Hepatitis C among individuals who inject Novel Psychoactive Substances (NPSs)[201]. Palmateer et al. showed a decrease in Hepatitis C incidence between 2008 to 2009 in Scotland, which they attributed in part to increase in Opioid Substitution Therapy (OST) and providing injecting equipment[202]. Both Taylor et al[81]. and Søholm et al[203]. used this equation to estimate Hepatitis C incidence in prisons, although in the case of Søholm et al., their study was looking at prisons in Denmark, and they actually found an elevated incidence among people who inject drugs whilst in prison. Finally, Antouri et al. used the equation to estimate the incidence of Hepatitis C from the HepCdetect, a cross-sectional study of individuals in Barcelona who inject drugs, but they adjusted the equation to account for false-negative results in their testing protocol[204]. Of these papers, the two looking at individuals in prisons are at particular risk of underestimating as discussed earlier.

For examples of cross-sectional Hepatitis C studies that consider Hepatitis C infection as occurring at an exponential rate, we can look at Leon et al.[205] and Sutton et al.[206]. Leon et al. used a compartmental model to model the flow of individuals from Susceptible to Infected, RNA-positive, antibody-negative to Infected antibody-positive in order to parameterise a rate of infection and therefore the incidence of infection among people in major cities of France who inject drugs. Their model breaks down the change of incidence in terms of the year (evaluating data from 2004-2011), as well as the age of the individual. Whilst this model is appropriate for a large data-set (the analysis covered multiple cross-sectional studies), without careful consideration of the stochasticity of small sample sizes, it may not be appropriate for the smaller data-sets seen in the Taylor et al. study. Sutton et al. estimate the total Force of Infection (FOI) an individual might experience by looking at multiple cross-sectional studies of Hepatitis-C prevalence among individuals from Glasgow who inject drugs. This is slightly different from other cross-sectional studies as rather than focus on indentifying recent infections, they parameterise this FOI by estimating the probability of infection given an individual's total exposure. The annual force of infection is equivalent to our estimation for $\lambda$, meaning theoretically Sutton et al. could have continued to estimate the incidence of Hepatitis C, a metric which is easier to conceptualise outside of the model. However, their model estimates the force of infection each year for each length of injecting career (i.e. that the force of infection experienced in 2010 by someone who has been injecting for 5 years will have been different to that experienced by someone who had only been injecting for two years), making the overall incidence difficult to extract.

The study performed by Taylor et al., along with others like it, are crucial when attempting to model, understand and ultimately prevent the spread of Hepatitis C inside and outside of the prison system. Prisons may well represent key opportunities in our continued goal of complete Hepatitis C eradication. While their estimation of incidence bares a striking similarity to our own based on their data, their methods could be adjusted to better represent the exponential process of infection. With a different disease with a higher rate of infection, the same calculation could be as much as 20% away from the true value. Additionally, by taking a Bayesian approach, we are able to demonstrate true credible intervals in our work and therefore prevent incorrect conclusions from low-powered observations. It is our hope that in correcting this calculation, we can encourage further research in this clearly important field of research, not just with regards to Hepatitis C, but other infectious diseases, as well as the accurate modeling of flow through incarceration centres.

Summary:

A study in 2012 used the rate at which individuals in Scottish prisons tested positive for Hepatitis C RNA but negative for Hepatitis C antibodies to calculate the incidence on HCV infections in prison. They concluded that their study showed a decreased incidence in prison. We demonstrate how to envelope exponential infection rates into incidence calculations for such scenarios with a Bayesian approach to provide an estimate for incidence with confidence intervals for this study and similar. We find no evidence of a decreased incidence of Hepatitis C among people who inject drugs in prison, which contradicts the study's conclusions. We also show that in the future it will be vital to understand incarceration times in order to improve the accuracy of this estimate. The work in this chapter reveals a key insight that contradicts previous conclusions regarding the incidence of Hepatitis C in prisons, as well as providing a reason behind why these conclusions were formed in the first place. It also provides a simple equation that can be used to estimate Hepatitis C incidence in the future.

| Demographic | $a$ | $b$ | $\lambda$ (years$^{-1}$) | | $\lambda$ if all acquired through injection | |
|---|---|---|---|---|---|---|
| | | | $T = 51$ days | $T = 75$ days | $T = 51$ days | $T = 75$ days |
| All incarcerated | 3 | 2443 | 0.009<br>(0.002-0.023) | 0.006<br>(0.002-0.016) | - | - |
| PWID | 2 | 477 | 0.033<br>(0.006-0.096) | 0.022<br>(0.004-0.065) | 0.047<br>(0.013-0.12) | 0.032<br>(0.009-0.081) |
| People who have ever injected drugs in prison | 1 | 90 | 0.093<br>(0.009-0.369) | 0.063<br>(0.006-0.251) | 0.248<br>(0.066-0.625) | 0.168<br>(0.045-0.425) |
| People who currently inject drugs in prison | 0 | 45 | 0.036<br>(0.0-0.398) | 0.024<br>(0.0-0.27) | 0.486<br>(0.129-1.227) | 0.331<br>(0.088-0.835) |

Table 5.2. Estimate for the exponential rate of Hepatitis C infection $\lambda$. If it is assumed that all infections in prisons occurred due to infecting drugs, then by definition all positive cases are individuals who are injecting drugs during their current admission

| Demographic | $a$ | $b$ | Incidence | | Incidence if all acquired through injection | |
|---|---|---|---|---|---|---|
| | | | $T = 51$ days | $T = 75$ days | $T = 51$ days | $T = 75$ days |
| All incarcerated | 3 | 2443 | 0.009<br>(0.002-0.023) | 0.006<br>(0.002-0.016) | 0.009<br>(0.002-0.023) | 0.006<br>(0.002-0.016) |
| PWID | 2 | 477 | 0.032<br>(0.006-0.092) | 0.022<br>(0.004-0.063) | 0.046<br>(0.013-0.113) | 0.032<br>(0.009-0.078) |
| People who have ever injected drugs in prison | 1 | 90 | 0.089<br>(0.008-0.308) | 0.062<br>(0.006-0.222) | 0.219<br>(0.064-0.465) | 0.155<br>(0.044-0.346) |
| People who currently inject drugs in prison | 0 | 45 | 0.035<br>(0.0-0.328) | 0.024<br>(0.0-0.237) | 0.385<br>(0.121-0.707) | 0.282<br>(0.084-0.566) |

Table 5.3. Estimate for the incidence of Hepatitis C infections in prisons $I$ as the proportion of susceptible incarcerated individuals who would be expected to be infected after 1 year of exposure

# Chapter 6

# Assessing the effect of cabin location on the spread of two infectious diseases on board a cruise-liner using novel methods

## 6.1 Introduction

When we have previously discussed close contact environments, and in particular closed environments, we highlighted how one particular aspect with ethical implications was the lack of choice. In both the case of hospitals and prisons, patients or inmates respectively have limited choice about where they are, either because they need to be in hospital to receive medical treatment or because they are forced to be in prison by law. This can be extended in some regards to care institutions where residents may need the elevated level of support provided.

An interesting exception to this lack of choice is cruise ships. In all but the most rare of circumstances, holiday makers on cruise ships have chosen to be there. They still, however, represent a closed, close-contact environment with the potential for outbreaks to occur. Additionally, because of the increased age range cruise trips appear to attract, which is increasing according to the Cruise Lines International Association 2021 report[207], there is a potential preponderance for its occupants to be more vulnerable to infectious diseases.

This closed environment has led to a lot of interest in modelling and understanding outbreaks aboard cruise ships. In particular, there have been multiple studies looking at the 2020 SARS-CoV-2 outbreak aboard the Diamond Princess cruise liner. Between January and February of that year, an outbreak occurred on the ship after an infected individual boarded in Hong Kong. Authorities in Japan enforced a quarantine on the ship that lasted the majority of February 2020 before the ship was allowed to dock in Yokohama Port.

Some attempts have been made to fully parameterise aspects of the outbreak, which at the time accounted for more than half the recorded SARS-CoV-2 cases outside of China[154]. For example, Nishiura estimated the ongoing cumulative incidence by comparing the symptom onset line list to a distribution for the incubation period of SARS-CoV-2[173]. In a Chapter 4 we explored a similar method for estimating the total number of in-hospital infections. Azimi et al. took advantage of the relatively small population size of the Diamond Princess to create multiple theoretical alternative outbreaks and in doing so showed the importance of aerosol based transmission of SARS-CoV-2 over fomite transmission[208], a growing concern when it came to outbreaks in closed environments[209].

SEIR models (Susceptible → Exposed → Infectious → Immune) models have been fitted to data from the Diamond Princess outbreak multiple times. With a relatively small population, it is important to include models that take into account the effect of a decreasing susceptible population. For example, Lai et al. used a Monte Carlo Markov Chain (MCMC) approach to parameterise their SEIR model of the Diamond Princess outbreak and found a remarkably high estimate for the initial reproduction number of 5.7[210]. In this case, the reproduction number represents the total number of infections an infectious individual would make throughout their entire infectious career with an entirely susceptible population, and is calculated by multiplying the rate of transmission by the expected length of their infectious period. Both Huang et al.[211] and Emery et al.[158] used an SEIR model when estimating the importance of asymptomatic transmission during the outbreak, and found that it accounted for a considerable proportion of all transmissions. Finally, Batista et al. used an SEIR model to show that while the quarantine was effective in reducing the number of infections, an increased immunity, such as through vaccination, would have been more effective at reducing the outbreak spread[212].

SEIR models are effective, in that they emulate the progress of an infectious disease. However, if not used with care, they fail to take into account the important effect of spatial factors. In a cruise in particular, we can get a good understanding of where people are likely to go and who they are likely to meet based on their cabin location. For example, Xu et al. used cabin allocation to specifically look for clustering of transmissions on the Diamond Princess after quarantine was in effect. They found that the majority of transmissions seemed to occur between individuals sharing the same cabin and that there was no clustering of outbreaks around certain cabins, suggesting that there was no evidence of spread through the ventilation system[213]. As an alternative approach, Liu et al. generated a full contact model of the Diamond Princess including spatial elements to show the importance of the quarantine[214].

A couple of studies analysed the genetic data coming from PCR results from the Diamond Princess outbreak. Sekizuka et al. investigated the genetic similarity of the swabs to prove that the outbreak could be traced to one exposure event[215]. Hoshino et al. took advantage of this relatively narrow genetic variance to estimate the mutation rate of SARS-CoV-2[216]. These are further examples of how relatively small outbreaks can be used to give insight on a disease as a whole.

The variety of approaches when analysing outbreaks on ships similar to the Diamond Princess has led to a number of reviews of such studies. Rosca et al. performed a systematic review of studies involving outbreaks of SARS-CoV-2 on cruise ships. In general they found the level of evidence to be poor and that there was a need for standardisation of such studies[217]. A broader review looking at interventions performed during SARS-CoV-2 outbreaks on cruise ships, transport ships and naval vessels was performed by Kordsmeyer et al.. They concluded that in general there was a need for ships to have made plans for outbreaks in advance of them occurring, such as having testing strategies and isolation plans in place before a voyage[218].

It is worth noting that outbreaks like the one seen onboard the Diamond Princess appear to be relatively rare. A review of all reported ship-based SARS-CoV-2 outbreaks between January and October of 2020 revealed 104 occurrences of a ship having at least one positive case during their voyage. This review, performed by Willebrand et al., showed a median attack rate for these outbreaks of just three individuals, with the Diamond Princess standing as a remarkable outlier, alongside another outbreak on the similarly named Ruby Princess[153].

One outbreak that does not appear frequently in literature for cruise-ship SARS-CoV-2 outbreaks is that of the focus of this chapter. In March 2020, a British holiday cruise-liner saw outbreaks of both SARS-CoV-2 and Norovirus at the same time. As SARS-CoV-2 is highly infectious, the ship was not allowed to dock at its nearest port, resulting in a completely closed outbreak of both disease. Of the 1085 total passengers on board, 19 individuals contracted Norovirus and 15 individuals contracted SARS-CoV-2. The cruise was allowed to dock on 20/3/2020, a week after the last case of either disease.

We have been provided with a line-list of diagnosis times and cabin numbers for the two concurrent outbreaks, along with a map of the ship. Due to the small size of the outbreak, it would be inappropriate to fit a deterministic SEIR model to the data-set. Additionally, as this outbreak occurred at a time before mass-screening was available, it is difficult to know to what extend asymptomatic transmissions were involved in this outbreak. Instead, we want to generate a simple model that investigates the spatial element of this outbreak.

By comparing the layout of the actual outbreak to the layout of random outbreaks, we aim to show an inverse link between distance between cabins and the probability of spread between cabins. We aim to show that the distance between infected cabins is shorter than would be expected if chosen at random and in doing so prove a link between cabin location and disease spread. Doing so will be useful as it will enable us to advise on spatial elements in future outbreaks on cruise-liners, as well as outbreaks in different but similar environments.

## 6.2 Methods

We start by investigating the layout of the cruise-liner. A copy of the cruise-liner's General Arrangement after dry dock in 2019 was used, provided through personal communication. Each of the 676 cabins was assigned a 3D Cartesian coordinate, with the z coordinate being the floor they were on. The x and y coordinates were based on the exact location of their cabin door hinge. We used the hinge because they are nearly universally present and demonstrate where individuals would enter a corridor and therefore interact with their neighbours. A few notable exceptions were found. Cabins on the 7th floor appear to have sliding doors. The door jambs were used as a proxy for a door hinge. There is a door between the Captain's and the Hotel Manager's cabins. This was ignored from the modelling as neither of them were affected by the outbreak. A number of cabin doors were missing on the plans. Their locations were approximated based on neighbouring cabins.

The cabins and dates of all passengers and crew members affected were provided through personal communication. We treat each cabin as an infectious unit. We classify them as symptomatic when the first individual in the cabin develops symptoms. A generation time of 5.2 days with a standard deviation of 1.72 days, using a Gamma distribution, was used for the Coronavirus outbreak[219], and a mean of 3.60 days and standard deviation of 1.70 for the Norovirus outbreak [220]. Using these distributions, two cabins were connected if their occupants contracted the same disease and if the generation times between the first people to contract the disease in each cabin would be between 0.025 and 0.975 for the cumulative density function of that disease's generation time.

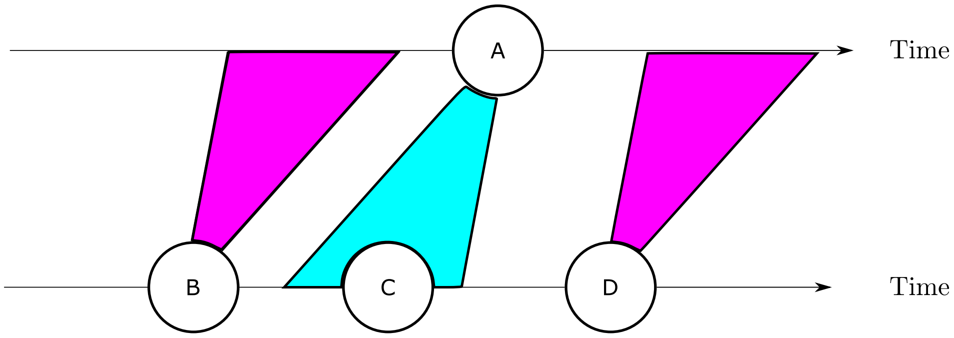


Figure 6.1. Four cabins are arranged by the time at which the first person in each cabin developed symptoms. The blue block indicates the window of time Cabin A's infector would have had to have developed symptoms, based on the time Cabin A developed symptoms. The purple blocks indicate the times Cabin A would have had to developed symptoms in in order for either Cabin B or D to be its infector. Cabin B developed symptoms too early, so it did not infect Cabin A, and we count it towards the list of failed transmissions. Cabin D developed symptoms too late. This does not count as a failed transmission as Cabin A must already have been infected before it was exposed to Cabin D. Cabin C is the only cabin that could have infected Cabin A.

If there were $n$ cabins involved in the outbreak, then $\mathbf{X}$ is an $n \times n$ sized array of the horizontal distances between each cabin. $X_{i,j}$ is a null value if their infection timings fail to align, such that the $i$th cabin could not have been infected by the $j$th cabin (this means that $X_{i,i}$ is also a null value as the $i$th cabin could not infect itself). For each infected cabin, we want to choose the closest possible infecting cabin with regards to horizontal distance. With regards to vertical distance, we have four decreasingly strict rules regarding if a cabin can infect another cabin dependent on the respective floors they are on. These are that one cabin can infect another cabin:

1. only if they are on the same floor.
2. only if they are on the same or adjacent floors.
3. only if they are both above or both below the fifth floor.
4. regardless of their floor number.

We separate cabins by the fifth floor in the third rule set because is this a communal area, as so we suspected that individuals from cabins above the fifth floor were unlikely to go below the fifth floor and vice versa.

For each rule-set, if the cabin $j$ could not infect cabin $i$ according to the rule, we give $X_{i,j}$ a null value. Each non-null value on the $i$th row therefore represents a possible infector of the $i$th cabin. We select the smallest of these value for each row to generate a transmission tree that minimises the horizontal distance between cabins (ignoring cases where the entire row is null values), assuming only one initial introductory event. We sum across the lengths of the edges of this transmission tree (defined as the horizontal distance between infector-infectee paired cabins), and take the natural logarithm of this value to calculate that transmission tree's score. We assume that infected cabins that had no infecting cabins according to our rule-set were infected by random homogeneous mixing in communal areas and are therefore not accounted for in our score.

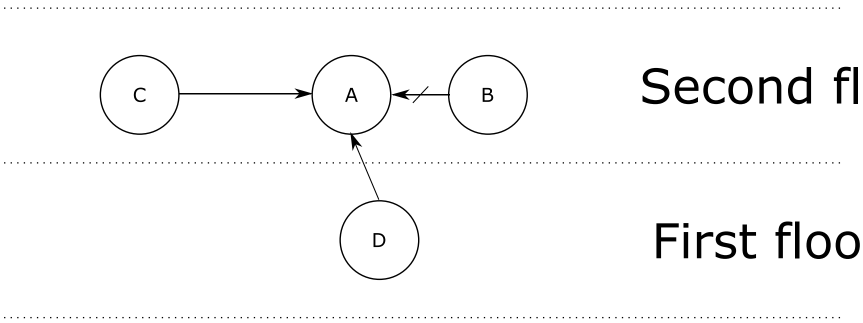


Figure 6.2. In this example, we want to choose the closest possible infecting cabin to Cabin A. Cabin B is on the same floor, and closer than Cabin C, but the timings for their symptom onsets do not line up (marked by the cross through their connecting arrow. Horizontally, Cabin D is closer to Cabin A than Cabin C is. However, it is on a different floor. For rule-sets that allow for infections between adjacent floors (Rule-sets 2-4), Cabin D will be chosen as Cabin A's infector. However, for Rule-set 1, where between-floor transmission is not allowed, Cabin C will be chosen.

These scores were then compared to 100000 alternative outbreaks. In these alternative simulated outbreaks, the same timings for the onset of symptoms in a cabin were used, but these were each allocated to a random cabin on the same floor as the initial symptomatic cabin: i.e. in each iteration, if occupants of a cabin on, say, the fourth floor developed symptoms at a particular time in the observed outbreak, we selected a random cabin on the same floor to develop symptoms at this time in the simulated outbreak. We repeated this analysis with staff and guest cabins separated, so that a true infection in a staff cabin would only be assigned to a random staff cabin on the same floor, and likewise for guest cabins. Staff and guest cabins were, by design, mostly separate, and so we wanted to avoid mis-attributing an outbreak among staff to the closeness of their cabins, rather than their shared work environments. We performed this simulation for each outbreak separately. In the observed outbreak, the was no recorded cabin that had been infected with both diseases. However, since each outbreak was relatively small compared to the total number of cabins, we felt it reasonable to assume this was coincidence rather than one disease being a protective factor for the other.

We compared the distribution of simulated scores to the observed score. If the distance between cabins in either outbreak was consistently shorter than expected, we would expect the score from the true outbreaks to be significantly smaller than from the randomly assigned outbreaks.

## 6.3 Results

The two outbreaks show a different relationship when it comes to the distance between affected cabins. Starting with Norovirus (shown in orange both in Figures 6.3 and 6.4), we essentially saw a total total distance between “infecting” cabins similar to the most likely value of what would be seen if the cabins had been selected from each floor at random. This was true both when our random selections mixed together guest and staff cabins (Figure 6.3) and when they kept them separate (Figure 6.4). In fact, for some rule-sets, the total distance between cabins is slightly bigger in the observed outbreak than expected (but not by any statistically significant margin).

Cabin distance appears to have had a greater effect on the spread of SARS-CoV-2. In particular, when we allow for transmission to occur between any cabin on the same side of the fifth floor, the distance between infecting cabins is smaller than expected under random chance. When we account for keeping staff and guest separate, this difference is close to being significant (p=0.052), and is significant when we do not limit which floor a cabin can infect (p=0.035). Tables 6.1 and 6.2 give these results in full.

## 6.4 Discussion

This investigation attempted to link the transmission of two infectious diseases in a closed environment to proximity, using cabin location as a proxy for location of infec-

|  |  | Expected score | Observed score | p-value |
|---|---|---|---|---|
| SARS-CoV-2 | Same floor | 5.48 | 5.24 | 0.203 |
|  | Within one floor | 5.72 | 5.69 | 0.453 |
|  | Same side of fifth floor | 5.94 | 5.65 | 0.105 |
|  | Across all floors | 5.78 | 5.43 | 0.09 |
| Norovirus | Same floor | 6.14 | 6.10 | 0.437 |
|  | Within one floor | 6.30 | 6.36 | 0.595 |
|  | Same side of fifth floor | 6.27 | 6.36 | 0.635 |
|  | Across all floors | 6.17 | 6.25 | 0.625 |

Table 6.1. The logarithm of the minimum total distance between transmitting cabins for each outbreak, with staff and guests treated indistinguishably in simulations. The expected score is the mean result across our simulations.

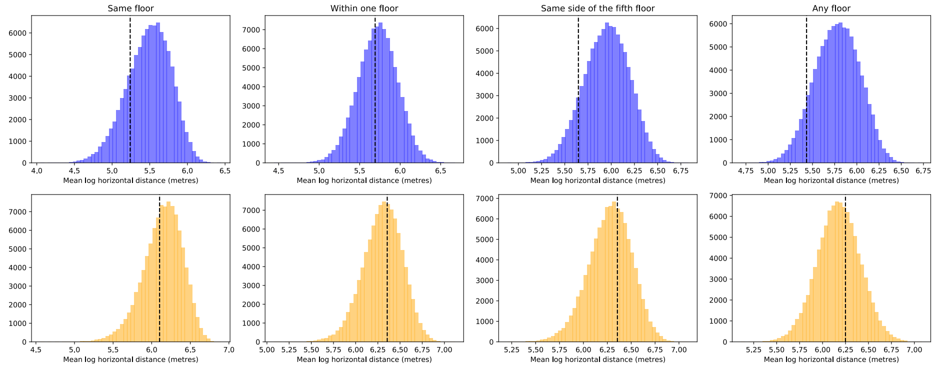


Figure 6.3. Histogram of random total distances between infected cabins. The blue plots show the SARS-CoV-2 results and the orange plots show the Norovirus results. The dotted lines show our observed results. In each of the SARS-CoV-2 cases the observed results are below the expected score if the cabins were randomly assigned, implying that there may be a relationship between distance and disease spread.

tors and infectees. Figures 6.3 and 6.4 show a possible inverse relationship between distance between cabins and the probability of transmission of SARS-CoV-2, but that of Norovirus.

This relationship was only truly observed when separating out staff and guests and allowing transmission between all floors. If we take this result at face value, it does not really point towards individuals infecting each other from the comfort of their own cabins. If it did, then we would expect to see a similar effect among individuals on the same or adjacent floors. Instead, one possible explanation for this observation is that we are seeing the effect of shared communal areas that are more likely to be used base on cabin location, such as stair wells or communal toilets. A deeper look into each cabin's route to the fifth deck may have indicated areas where people are likely to have crossed paths.

If there is a spatial element to the risk of spread, then it is very important that this is accounted for when attempting to parameterise a model of an outbreak like the one seen on the cruise-liner examined in this chapter. Given that there are far more cabins that are far away from each other than close together, if we enumerated each failed transmission (by which we mean each time an infectious cabin failed to infect another cabin) we would find that there are a lot more failed transmissions over a longer distance than over a shorter distance. We may initially see no difference be-

| | | Expected score | Observed score | p-value |
|---|---|---|---|---|
| SARS-CoV-2 | Same floor | 5.84 | 5.24 | 0.205 |
| | Within one floor | 5.87 | 5.69 | 0.286 |
| | Same side of fifth floor | 5.98 | 5.65 | 0.052 |
| | Across all floors | 5.81 | 5.43 | 0.035 |
| Norovirus | Same floor | 6.13 | 6.10 | 0.442 |
| | Within one floor | 6.30 | 6.36 | 0.598 |
| | Same side of fifth floor | 6.27 | 6.35 | 0.645 |
| | Across all floors | 6.17 | 6.25 | 0.633 |

Table 6.2. The logarithm of the minimum total distance between transmitting cabins for each outbreak, with staff and guests kept separate in our simulations. The expected score is the mean result across our simulations.

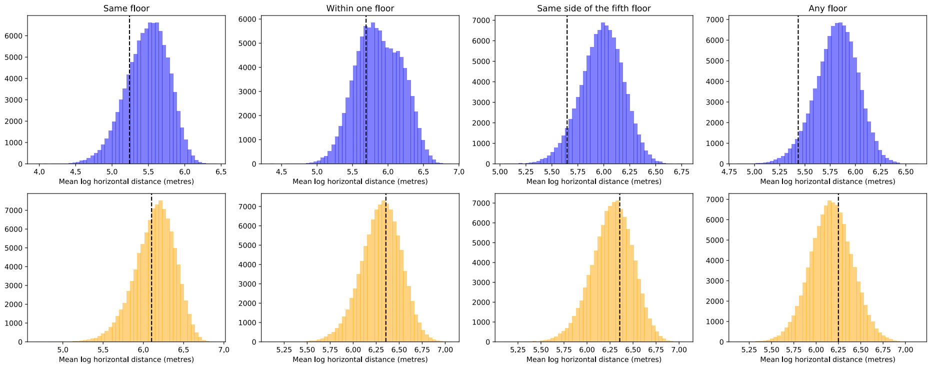


Figure 6.4. Histogram of random total distances between infected cabins when guests and staff have been treated separately. The blue plots show the SARS-CoV-2 results and the orange plots show the Norovirus results. The dotted lines show our observed results. In each of the SARS-CoV-2 cases the observed results are below the expected score if the cabins were randomly assigned, implying that there may be a relationship between distance and disease spread.

tween a model where we assume distance is irrelevant and one where distance between infector-infectee cabins is accounted for. However, as the number of susceptible cabins decreases, this similarity may change. For example, in a scenario where there are plenty of susceptible cabins left, but none are near infectious cabins, a model that accounts for distance may predict that the outbreak is coming to an end, while one that does not may allow it to continue. Conversely, if only a few susceptible cabins remain, but they are all close to infecting cabins, the model that accounts for distance may predict that the outbreak will continue, while the model that does not may say that there that the susceptible population it too small for the outbreak to continue. This could lead us to under- or overestimate the size of an outbreak. Multiple similar attempts have been made to infer spatial relationships of outbreaks in closed environments. Kirking et al. examined a Norovirus outbreak on a flight from Boston, Massachusetts, to Los Angeles, California. They used seat allocation as a proxy for location and through logistic regression found aisle seats and seat in close proximity to a particular group were positively linked to contracting Norovirus[221]. Han et al. directly modelled passenger movement and resulting droplet dispersion to simulate the spread of an in-flight SARS outbreak[222]. In order to understand in flight movement, Hertzberg et al. monitored passengers on 10 commercial flights and generated outbreak models based on their movements[223]. Over these shorter periods of time it may be easier to either directly model individual's movements or approximate them to a location such as their seat when modelling the spread of a disease in a close-contact environment. More unpredictable movement may be expected over the 23 days our outbreaks took place. We have shown it is still possible to extract positional aspects of these outbreaks through approximating individual's locations to their cabins.

There are elements to this investigation that require further analysis. Data was only available on the number of individuals in infected cabins. Therefore it was not possible to accurately investigate if this had an effect on a cabin's probability of being infected as it was not possible to compare infected cabin occupancy with non-infected cabin occupancy. Similarly, a cabin with multiple infected individuals is likely to be more infectious for longer which, again, was not accounted for in the model. Demographics of the individuals may be important for determining their risk of infection, in particular age. Distinguishing between passengers and crew members may elicit different interaction patterns and therefore different dependencies on cabin location (for example, it may be reasonable to assume that passengers on the top deck are more likely to interact with and therefore infect or be infected by crew staying on the bottom deck than passengers on the bottom deck). Accounting for this may increase the fidelity of the data.

One avenue of particular interest would be estimating the transmission rate between cabins, and seeing if this changes as the distance between cabins is increased, especially when they are on different floors. Seymour et al. approach a similar problem when trying to estimate the transmission rate of Foot and Mouth disease between farms in the UK[224] and avian flu in farms in the Netherlands[225]. In both cases, the farms act as infectious units and they attempted to find a non-parametric function that describes the transmission rate between two farms dependent on their distance. It is easy to see how this is analogous to considering cabins as infectious units and attempting to find a function that describes the between-cabin transmission rate dependent on their distance. Unfortunately, the data-points in the farm-based outbreaks are orders of magnitude larger than that of the outbreaks examined in this chapter. It is unlikely that we would have enough data to draw any meaningful conclusions.

A key assumption of this model is that there was only one introduction event for each disease. However, this may not be the case. If the Norovirus infection was related to food poisoning or the Coronavirus was introduced by passengers on shore, then the seeming first generation and resulting outbreak would be larger than expected when there was only one introduction. In the case of the Norovirus outbreak, the first four individuals to present with symptoms developed symptoms within 15 hours of each other, making it unlikely that they infected each other given the required generation time. This could explain why we do not see a spatial effect with this outbreak.

We also ignore the possibility of asymptomatic transmission. Ignoring the possibility of asymptomatic individuals or individuals who chose not to present with symptoms can increase the possibility of underestimating an outbreak size. We do not know the number of asymptomatic individuals during the actual outbreak. It can be argued that asymptomatic individuals can be accounted for in our model as a vector for disease transfer between cabins. This would require a different transmission probability, as it would represent the probability of two consecutive transmissions, as well as an increased generation time.

Additionally, there is an assumption that the rate of transmission between cabins did not change throughout the outbreak. We have not accounted for interventions such as quarantining and reduced mixing which may have changed in strictness as the outbreaks became more evident. There is circumstantial evidence for this being the case. There are seven days infection free days before the cruise-liner is evacuated, which was unlikely to occur with the rates of transmission seen at the beginning of the outbreak. This may indicate a reduced rate of transmission due to outbreak control, although disentangling it from the reduction in the susceptible population is not trivial.

Finally, this analysis has not investigated the effect of two concurrent outbreaks of different diseases. It may be that those already unwell with one disease have a reduced probability of contracting the other disease due to decreased mixing or an innate protective element of the disease, or an increased vulnerability due to increased frailty whilst unwell. Of note, no cabin reported having both diseases.

We have shown some weak effects of cabin location on the spread of SARS-CoV-2 in this cruise-liner outbreak. If the outbreak had been larger, it is possible that we would have been able to develop a full transmission model that could be generalised to other cruise-liners. As this study is retrospective, it cannot directly advise on or help mitigate the outbreak it is analysing. Further extensions of this investigation could look at how this model would interpret live data of an ongoing outbreak. With this information, it could be used to advise on swabbing and isolation protocols, indicating which cabins are more at risk before they have begun to develop symptoms, thus limiting the outbreak spread.

Summary:

In this chapter we investigate a dual outbreak of norovirus and SARS-CoV-2 on a cruise-liner. We propose a simple method of investigating spacial effects on the outbreak by treating each cabin as an individual infectious unit and comparing the horizontal distances between infected cabins. This would be an easy and effective way to show the importance (or lack of effect) of location on an outbreak, as well as inform where targeted screening could be focused. In this study, neither outbreak was large enough to show a significant effect of cabin location on who was infected.

# Chapter 7

# Summary

Through the course of this thesis, we have presented five different tools designed to help understand and mitigate against outbreaks in close-contact environments. With each tool we showed how it works, how it can be used and areas of possible research around these tools. These chapters also included full discussions regarding future work in each of these areas. As this thesis comes to an end, it is worth revisiting each of these tools in turn to summarise what we have learnt from them, as well as what there is still left to be learnt.

## 7.1 Estimating the size of a first generation of an outbreak

In Chapter 2 we investigated multiple variations of a scenario where a group of individuals were exposed to some infectious diseases. We wanted to know if we could infer the total number of individuals who had been infected based on how many were symptomatic before a certain time and how different aspects of the outbreak might affect our analysis. Indeed we could, as individuals becoming symptomatic earlier could be used to infer that there had been more samples of the underlying distribution that describes the disease's incubation period. We found that a greater proportion of symptomatic individuals led to a later certainty that we had seen all symptomatic individuals and that possibility of asymptomatic individuals stretch out our calculations in the time dimension. This knowledge could be incredibly useful in single exposure events or in early pandemics for advising when we can be certain that a group of individuals are not infectious.

As we allowed for prolonged exposure, we started having difficulties generalising our analysis to all diseases. The analysis involved convolving the incubation distribution with another distribution that was dependent on the model of prolonged exposure we were using. We broke down the model into a scenario where positive case occur with a probability $\rho$ and the probability of observing a positive case is described by the function $K(\rho)$. In our simplistic case of point exposure, $K(\rho)$ has a linear relationship with $\rho$. We could not find a non-linear function for $K(\rho)$ that generates an analytical solution for estimating the total number of positive cases.

Future work realising such a function would help generalise this method of estimation, and lead to estimates for the ongoing rate of infection, approximating the true total current prevalence and calculating the probability that an outbreak is over.

## 7.2 Adjusting rota patterns to reduce in-work infectious times

Chapter 3 showed a wholly unique line of enquiry investigating the role of work schedules in limiting the length time individuals are infectious whilst at work. To the best of our knowledge, this has not been looked at before and so there are plenty of variations and alternatives that are worth investigating further.

Our initial numerical analysis in revealed that, at least on paper, changing rota patterns can reduce the expected length of time an individual spends at work whilst infectious and the changing the timing of tests could be used to improve their efficacy. We then went further to represent rota patterns as Fourier series, enabling us to perform our analysis at higher speeds to a greater guaranteed precision.

There were three problems that remain open at the end of this thesis. The first is writing a complete distribution for the length of time whilst infectious. Currently, the system we proposed simply finds the central estimate for this value, but it should vary from person to person. Knowing how it varies will give us a better understanding how certain interventions would be. This can be done numerically, but would be at risk of introducing rounding errors, and would become progressively computationally cumbersome as the complexity of a rota pattern increased. Instead it would be useful to be able to calculate this distribution from our Fourier analysis.

The second problem also involves incorporating something numerically possible into the Fourier analysis: Testing patterns. In the body of this thesis we were unable to represent regular testing as part of a Fourier transformation, despite the fact that it is something that occurs regularly in sequence with the rota pattern with a predictable outcome. Instead we had to generate a work around which, whilst better than the plane numerical solution, still required some numerical integration. This was an incredibly frustrating problem, as it seems like a solution should be possible, although as of writing it is beyond this author's abilities.

The third problem (and a common problem with a lot of epidemiological models) is that this work is entirely theoretical. Whilst on paper we have shown the importance of rota pattern structure and test timing, it would be difficult to show in real life. We may only see differences on larger scales, and so studying the phenomenon in the real world may be impractical.

## 7.3 Using a probablistic method to estimate the total number of in-hospital infections

Nosoco, the focus of Chapter 4, is a tool intended to be used to estimate the total number of nosocomial transmissions that occurred during some observation window. As discussed in the body of this thesis, we saw a gap in analysis tools available in the UK, as the current standard was to declare an individual as representing a nosocomial transmission if they fulfilled some criteria regarding the timing of their diagnosis. Using the probabilistic approach of Nosoco allowed for some uncertainty and for evidence of nosocomial transmissions to accumulate between cases that would have been declared negative otherwise. It is a fast, cheap method of monitoring nosocomial transmissions without having to generate a full transmission model for inside and outside of hospitals.

There are improvements that can be made to Nosoco. The first, possibly most important improvement would be to validate the distribution Nosoco uses to describe the time from infection to diagnosis. Currently other researchers in our group are looking at genetic data of SARS-CoV-2 to assess if an individual's infection comes from within the hospital or occurred outside. We intend to compare this data to ward admission data to approximate the correct distribution for Nosoco to use. Ellingford et al. have already shown that RNA samples can be used to track SARS-CoV-2 outbreaks across wards and between hospital staff, so it seems reasonable that we should be able to approximate a time of infection.

Ideally, though, we would want to describe three separate distributions, dependent on why an individual was swabbed:

1. Swabbing because the patient had developed symptoms of SARS-CoV-2
2. Swabbing as a precaution because another individual on their ward had recently been diagnosed with SARS-CoV-2
3. Swabbing as part of a regularly scheduled swabbing program

Each of these reasons would likely detect a positive individual at a different stage of their infectious career and so come with a different probability that they were infected during their admission (i.e. that they represent a nosocomial transmission). If these details were available to use (which they are currently not - they can only be inferred) we get a more accurate picture of the distributions Nosoco needs.

Data handling is another large issue for Nosoco. We found that it was difficult to consistently pair swab and admission data between separate hospitals. Each hospital had their local protocol, which often did not match up. Fortunately, the trust has moved to a universal data system, which will hopefully improve the fidelity of Nosoco in the future. It would be useful to repeat our investigations using this new system.

Unfortunately, it does not apply to historic data, so would be unclear if difference came as the result of better data management or changes in outbreaks over time.

One area that we wanted to use Nosoco to investigate was between-ward transmissions. We never wanted to generate full transmission trees - this is not the purpose of Nosoco. However, we did want to see if there were any consistent temporal similarities between wards. We might expect wards to see some similarities. We showed an initial close relationship between external and internal rates of infections, so wards with elevated rates of infection would inevitably be in sync. However, what we were really interested in were wards where there was a consistent short delay between the two. This could have been indicative of infectious individuals on one ward resulting in infected individuals in the other. This would have been an incredibly useful piece of information to have to advise about transfers between wards. Unfortunately, we did not have the time to perform this analysis, which will have to be left to a later date.

## 7.4 Estimating the incidence of Hepatitis C in prisons

Chapter 5 raises some concerns regarding a common equation used to estimate the incidence of Hepatitis C from cross-sectional surveys:

$$I = \frac{\left(\frac{365}{T}\right) n}{(N - n) + \left(\frac{365}{T}\right)}$$

where $I$ is the incidence, $N$ is the total number of individuals in the survey who to not have Hepatitis C antibodies and $n$ is the subset of those individuals who do have Hepatitis C RNA when tested, which occurs in a narrow window of time of length $T$ days, close to the initial time of infection. We can see the logic with this analysis. A cross-sectional survey is only really assessing a $T$ length window, so we inflate the positive cases in that window to match how many we may expect in a year and calculate the incidence from that. However, as discussed in the chapter, this is not reflective of the infectious process. We instead show a simple calculation for approximating a Bayesian posterior for the rate of infection and from this calculate a central estimate and a credible interval for the incidence of Hepatitis C.

We particularly wanted to look at Hepatitis C in prisons. This calculation had twice been used to show that the incidence of Hepatitis C in prisons is less than one would expect among people who inject drugs. Once we used our formula, we showed no evidence of such discrepancy.

This was actually one of the original focuses of this thesis. We wanted to further investigate the incidence of Hepatitis C in prisons and see how early identification and treatment in prisons could help reduce the incidence inside and outside of prisons. One confounding factor is the unusual distribution for the length of stay in prisons. Some sentences can be exceedingly short, whilst others are orders of magnitude longer, and a sentence length may not be a direct indication as to how long an individual actually ends up staying in prison. For some people, their stay in prison may be far shorter than it takes to complete complete a course of treatment for Hepatitis C. Assuming that an effective treatment only occurs if they finish their course of treatment before leaving prison, understanding the distribution of prison stays (and particular if they are related to risky behaviours such as injecting drugs) would be vital to understanding how best to test and allocate treatment for the disease.

Unfortunately, of course, these plans were created in late 2019, before the SARS-CoV-2 pandemic. Whilst work continues on understanding and modelling Hepatitis C spread in prisons, there still has not been a full investigation into parameterising prison lengths of stay and inserting this into outbreak models.

## 7.5 Associating cabin distance with outbreak spread on a cruise-liner

Chapter 6 essentially showed a very loose way of detecting spatial clusters between fixed infected units. We approximated possible transmission trees for a dual outbreak on a cruise-liner and saw how that network compared to other networks generated randomly by random cabin allocation. We did find a slight spatial element, although it was more indicative of shared communal spaces rather than direct transmission between adjacent cabins.

We were reticent to perform any more in-depth analysis on the data-set simply due to its size. With outbreaks in the order of 20, compared to a possible 600, spaced out largely all over the ship, there simply was not enough data to generate a full transmission model without large enough margins of error such that our conclusions would be essentially meaningless. Outside of work written in this thesis, we did consider mapping physical routes around the cruise-liner to delineate between physical distance and route distance, but felt this may be better left for later research. Instead we have shown a fairly simplistic method of checking for clustering of outbreaks around cabins,

The tools we have presented in this thesis have varying and far-reaching applications. They are relatively easy to use and we hope we have presented them in a way which someone unfamiliar with the intricacies of infectious disease modelling can still use. There remains room for growth in this field and we can see multiple avenues down which these tools can be used to increase our understanding of and hopefully mitigate against further outbreaks.

# Bibliography

[1] K. Dolan, A. L. Wirtz, B. Moazen, et al., “Global burden of hiv, viral hepatitis, and tuberculosis in prisoners and detainees,” The Lancet, vol. 388, no. 10049, pp. 1089–1102, 2016. DOI: 10.1016/S0140-6736(16)30466-4. [Online]. Available: https://doi.org/10.1016/S0140-6736(16)30466-4 (cited on pp. 17, 33, 35, 230, 231).

[2] A. B. Pitcher, A. Borquez, B. Skaathun, and N. K. Martin, “Mathematical modeling of hepatitis c virus (hcv) prevention among people who inject drugs: A review of the literature and insights for elimination strategies,” Journal of Theoretical Biology, vol. 481, no. 21, pp. 194–201, 2018. DOI: 10.1016/j.jtbi.2018.11.013. [Online]. Available: https://doi.org/10.1016/j.jtbi.2018.11.013 (cited on pp. 17, 29, 231).

[3] A. J. Lotka, “Analytical note on certain rhythmic relations in organic systems,” Proceedings of the National Academy of Sciences, vol. 6, no. 7, pp. 410–415, Jul. 1920. DOI: 10.1073/pnas.6.7.410. [Online]. Available: https://doi.org/10.1073/pnas.6.7.410 (cited on p. 18).

[4] M. Keeling and P. Rohani, Modeling Infectious Diseases in Humans and Animals. Princeton University Press, 2011, ch. Introduction to Simple Epidemic Models. (Cited on p. 19).

[5] “A contribution to the mathematical theory of epidemics,” Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, vol. 115, no. 772, pp. 700–721, Aug. 1927. DOI: 10.1098/rspa.1927.0118. [Online]. Available: https://doi.org/10.1098/rspa.1927.0118 (cited on p. 20).

[6] PubMed, Pubmed search. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/?term=SIR+model%5C%5BTitle%5C%2FAbstract%5C%5D&filter=years.2020-2023&size=200 (visited on 03/16/2023) (cited on pp. 20, 40).

[7] D. Lu, S. Yang, J. Zhang, H. Wang, and D. Li, “Resilience of epidemics for sis model on networks,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 27, no. 8, p. 083 105, 2017. DOI: 10.1063/1.4997177. eprint: https:

//doi.org/10.1063/1.4997177. [Online]. Available: https://doi.org/10.1063/1.4997177 (cited on p. 20).

[8] F. Etbaigha, A. R. Willms, and Z. Poljak, "An seir model of influenza a virus infection and reinfection within a farrow-to-finish swine farm," PLoS One, vol. 13, no. 9, e0202493, 2017. DOI: 10.1371/journal.pone.0202493. [Online]. Available: https://doi.org/10.1371/journal.pone.0202493 (cited on p. 20).

[9] J. A. Lewnard, M. L. N. Mbah, J. A. Alfaro-Murillo, et al., "Dynamics and control of ebola virus transmission in montserrado, liberia: A mathematical modelling analysis," The Lancet Infectious Diseases, vol. 14, no. 12, pp. 1189–1195, Dec. 2014. DOI: 10.1016/s1473-3099(14)70995-8. [Online]. Available: https://doi.org/10.1016/s1473-3099(14)70995-8 (cited on p. 21).

[10] A. A. King, M. D. de Cellès, F. M. G. Magpantay, and P. Rohani, "Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola," Proceedings of the Royal Society B: Biological Sciences, vol. 282, no. 1806, p. 20 150 347, May 2015. DOI: 10.1098/rspb.2015.0347. [Online]. Available: https://doi.org/10.1098/rspb.2015.0347 (cited on p. 21).

[11] A. M. Al-Khani, M. A. Khalifa, A. Almazrou, and N. Saquib, "The SARS-CoV-2 pandemic course in saudi arabia: A dynamic epidemiological model," Infectious Disease Modelling, vol. 5, pp. 766–771, 2020. DOI: 10.1016/j.idm.2020.09.006. [Online]. Available: https://doi.org/10.1016/j.idm.2020.09.006 (cited on p. 22).

[12] T. Nguyen, M. Adnan, B. P. Nguyen, et al., "COVID-19 vaccine strategies for aotearoa new zealand: A mathematical modelling study," The Lancet Regional Health - Western Pacific, vol. 15, p. 100 256, Oct. 2021. DOI: 10.1016/j.lanwpc.2021.100256. [Online]. Available: https://doi.org/10.1016/j.lanwpc.2021.100256 (cited on p. 22).

[13] F. Song and M. O. Bachmann, "Vaccination against COVID-19 and society's return to normality in england: A modelling study of impacts of different types of naturally acquired and vaccine-induced immunity," BMJ Open, vol. 11, no. 11, e053507, Nov. 2021. DOI: 10.1136/bmjopen-2021-053507. [Online]. Available: https://doi.org/10.1136/bmjopen-2021-053507 (cited on p. 22).

[14] F. Ball and L. Shaw, "Estimating the within-household infection rate in emerging SIR epidemics among a community of households," Journal of Mathematical Biology, vol. 71, no. 6-7, pp. 1705–1735, Mar. 2015. DOI: 10.1007/s00285-015-0872-5. [Online]. Available: https://doi.org/10.1007/s00285-015-0872-5 (cited on p. 23).

[15] F. Ball, L. Pellis, and P. Trapman, "Reproduction numbers for epidemic models with households and other social structures II: Comparisons and implications for vaccination," Mathematical Biosciences, vol. 274, pp. 108–139, Apr. 2016. DOI: 10.1016/j.mbs.2016.01.006. [Online]. Available: https://doi.org/10.1016/j.mbs.2016.01.006 (cited on p. 23).

[16] M. J. Keeling and J. Ross, "Optimal prophylactic vaccination in segregated populations: When can we improve on the equalising strategy?" Epidemics, vol. 11, pp. 7–13, Jun. 2015. DOI: 10.1016/j.epidem.2015.01.002. [Online]. Available: https://doi.org/10.1016/j.epidem.2015.01.002 (cited on p. 23).

[17] Q. Zhang and D. Wang, "Assessing the role of voluntary self-isolation in the control of pandemic influenza using a household epidemic model," International Journal of Environmental Research and Public Health, vol. 12, no. 8, pp. 9750–9767, Aug. 2015. DOI: 10.3390/ijerph120809750. [Online]. Available: https://doi.org/10.3390/ijerph120809750 (cited on p. 23).

[18] L. Bioglio, M. Génois, C. L. Vestergaard, C. Poletto, A. Barrat, and V. Colizza, "Recalibrating disease parameters for increasing realism in modeling epidemics in closed settings," BMC Infectious Diseases, vol. 16, no. 1, Nov. 2016. DOI: 10.1186/s12879-016-2003-3. [Online]. Available: https://doi.org/10.1186/s12879-016-2003-3 (cited on p. 23).

[19] F. Ball and D. Sirl, "Evaluation of vaccination strategies for SIR epidemics on random networks incorporating household structure," Journal of Mathematical Biology, vol. 76, no. 1-2, pp. 483–530, Jun. 2017. DOI: 10.1007/s00285-017-1139-0. [Online]. Available: https://doi.org/10.1007/s00285-017-1139-0 (cited on p. 23).

[20] L. Otero, L. Shah, K. Verdonck, et al., "A prospective longitudinal study of tuberculosis among household contacts of smear-positive tuberculosis cases in lima, peru," BMC Infectious Diseases, vol. 16, no. 1, Jun. 2016. DOI: 10.1186/s12879-016-1616-x. [Online]. Available: https://doi.org/10.1186/s12879-016-1616-x (cited on p. 24).

[21] J. Zelner, M. Murray, M. Becerra, et al., "Protective effects of household-based TB interventions are robust to neighbourhood-level variation in exposure risk in lima, peru: A model-based analysis," International Journal of Epidemiology, vol. 47, no. 1, pp. 185–192, Sep. 2017. DOI: 10.1093/ije/dyx171. [Online]. Available: https://doi.org/10.1093/ije/dyx171 (cited on p. 24).

[22] O. Ali, A. Aseffa, A. B. Omer, et al., "Household transmission of neisseria meningitidis in the african meningitis belt: A longitudinal cohort study," The Lancet Global Health, vol. 4, no. 12, e989–e995, Dec. 2016. DOI: 10.1016/s2214-109x(16)30244-3. [Online]. Available: https://doi.org/10.1016/s2214-109x(16)30244-3 (cited on p. 24).

[23] C. E. Oldenburg, J. Bor, G. Harling, et al., "Impact of early antiretroviral therapy eligibility on HIV acquisition," AIDS, vol. 32, no. 5, pp. 635–643, Mar. 2018. DOI: 10.1097/qad.0000000000001737. [Online]. Available: https://doi.org/10.1097/qad.0000000000001737 (cited on pp. 24, 28).

[24] T. K. Tsang, L. L. Lau, S. Cauchemez, and B. J. Cowling, "Household transmission of influenza virus," Trends in Microbiology, vol. 24, no. 2, pp. 123–133, Feb. 2016. DOI: 10.1016/j.tim.2015.10.012. [Online]. Available: https://doi.org/10.1016/j.tim.2015.10.012 (cited on p. 24).

[25] J. N. Walker, J. V. Ross, and A. J. Black, "Inference of epidemiological parameters from household stratified data," PLOS ONE, vol. 12, no. 10, Y. Yang, Ed., e0185910, Oct. 2017. DOI: 10.1371/journal.pone.0185910. [Online]. Available: https://doi.org/10.1371/journal.pone.0185910 (cited on p. 24).

[26] N. Geard, K. Glass, J. M. McCaw, et al., "The effects of demographic change on disease transmission and vaccine impact in a household structured population," Epidemics, vol. 13, pp. 56–64, Dec. 2015. DOI: 10.1016/j.epidem.2015.08.002. [Online]. Available: https://doi.org/10.1016/j.epidem.2015.08.002 (cited on p. 24).

[27] Z. J. Madewell, Y. Yang, I. M. Longini, M. E. Halloran, and N. E. Dean, "Household transmission of SARS-CoV-2," JAMA Network Open, vol. 3, no. 12, e2031756, Dec. 2020. DOI: 10.1001/jamanetworkopen.2020.31756. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2020.31756 (cited on p. 24).

[28] F. Ball, T. Britton, T. House, et al., "Seven challenges for metapopulation models of epidemics, including households models," Epidemics, vol. 10, pp. 63–

67, Mar. 2015. DOI: 10.1016/j.epidem.2014.08.001. [Online]. Available: https://doi.org/10.1016/j.epidem.2014.08.001 (cited on p. 25).

[29] L. Ozella, F. Gesualdo, M. Tizzoni, et al., "Close encounters between infants and household members measured through wearable proximity sensors," PLOS ONE, vol. 13, no. 6, E. H. Y. Lau, Ed., e0198733, Jun. 2018. DOI: 10.1371/journal.pone.0198733. [Online]. Available: https://doi.org/10.1371/journal.pone.0198733 (cited on p. 25).

[30] J. Mossong, N. Hens, M. Jit, et al., "Social contacts and mixing patterns relevant to the spread of infectious diseases," PLoS Medicine, vol. 5, no. 3, S. Riley, Ed., e74, Mar. 2008. DOI: 10.1371/journal.pmed.0050074. [Online]. Available: https://doi.org/10.1371/journal.pmed.0050074 (cited on p. 25).

[31] P. Rohani, X. Zhong, and A. King, "Contact network structure explains the changing epidemiology of pertussis," Science, vol. 330, no. 6006, pp. 982–985, Nov. 2010. DOI: 10.1126/science.1194134. [Online]. Available: https://doi.org/10.1126/science.1194134 (cited on p. 25).

[32] F. Miura, T. Watanabe, K. Watanabe, K. Takemoto, and K. Fukushi, "Comparative assessment of primary and secondary infection risks in a norovirus outbreak using a household model simulation," Journal of Environmental Sciences, vol. 50, pp. 13–20, Dec. 2016. DOI: 10.1016/j.jes.2016.05.041. [Online]. Available: https://doi.org/10.1016/j.jes.2016.05.041 (cited on pp. 25, 31).

[33] K. Tokuda, Y. Yahata, and T. Sunagawa, "Prevention of secondary household transmission during shiga toxin-producing escherichia coli outbreaks," Epidemiology and Infection, vol. 144, no. 14, pp. 2931–2939, Jun. 2016. DOI: 10.1017/s0950268816001199. [Online]. Available: https://doi.org/10.1017/s0950268816001199 (cited on p. 25).

[34] J. Li and F. Brauer, "Continuous-time age-structured models in population dynamics and epidemiology," in Mathematical Epidemiology, Springer Berlin Heidelberg, 2008, pp. 205–227. DOI: 10.1007/978-3-540-78911-6_9. [Online]. Available: https://doi.org/10.1007/978-3-540-78911-6_9 (cited on p. 26).

[35] A. Bershteyn, D. J. Klein, and P. A. Eckhoff, "Age-dependent partnering and the HIV transmission chain: A microsimulation analysis," Journal of The Royal Society Interface, vol. 10, no. 88, p. 20 130 613, Nov. 2013. DOI: 10.

1098/rsif.2013.0613. [Online]. Available: https://doi.org/10.1098/rsif.2013.0613 (cited on pp. 26, 39).

[36] E. T. Richardson, S. E. Collins, T. Kung, et al., “Gender inequality and HIV transmission: A global analysis,” Journal of the International AIDS Society, vol. 17, no. 1, p. 19 035, Jan. 2014. DOI: 10.7448/ias.17.1.19035. [Online]. Available: https://doi.org/10.7448/ias.17.1.19035 (cited on p. 26).

[37] M. Meena, D. Rathee, P. Arora, et al., “Comparative study of clinico-bacterio-radiological profile and treatment outcome of smokers and nonsmokers suffering from pulmonary tuberculosis,” Lung India, vol. 33, no. 5, p. 507, 2016. DOI: 10.4103/0970-2113.188970. [Online]. Available: https://doi.org/10.4103/0970-2113.188970 (cited on p. 26).

[38] E. Hunter, B. M. Namee, and J. Kelleher, “An open-data-driven agent-based model to simulate infectious disease outbreaks,” PLOS ONE, vol. 13, no. 12, Y. E. Khudyakov, Ed., e0208775, Dec. 2018. DOI: 10.1371/journal.pone.0208775. [Online]. Available: https://doi.org/10.1371/journal.pone.0208775 (cited on pp. 26, 27).

[39] S. Venkatramana, B. Lewis, J. Chen, D. Higdon, A. Vullikanti, and M. Marathe, “Using data-driven agent-based models for forecasting emerging infectious diseases,” Epidemics, vol. 22, pp. 43–49, Mar. 2018. DOI: 10.1016/j.epidem.2017.02.010. [Online]. Available: https://doi.org/10.1016/j.epidem.2017.02.010 (cited on p. 26).

[40] K. Lum, S. Swarup, S. Eubank, and J. Hawdon, “The contagious nature of imprisonment: An agent-based model to explain racial disparities in incarceration rates,” Journal of The Royal Society Interface, vol. 11, no. 98, p. 20 140 409, Sep. 2014. DOI: 10.1098/rsif.2014.0409. [Online]. Available: https://doi.org/10.1098/rsif.2014.0409 (cited on p. 27).

[41] R. J. Rockett, A. Arnott, C. Lam, et al., “Revealing COVID-19 transmission in australia by SARS-CoV-2 genome sequencing and agent-based modeling,” Nature Medicine, vol. 26, no. 9, pp. 1398–1404, Jul. 2020. DOI: 10.1038/s41591-020-1000-7. [Online]. Available: https://doi.org/10.1038/s41591-020-1000-7 (cited on p. 27).

[42] R. T. Gilman, S. Mahroof-Shaffi, C. Harkensee, and A. T. Chamberlain, “Modelling interventions to control COVID-19 outbreaks in a refugee camp,” BMJ Global Health, vol. 5, no. 12, e003727, Dec. 2020. DOI: 10.1136/bmjgh-

2020-003727. [Online]. Available: https://doi.org/10.1136/bmjgh-2020-003727 (cited on p. 27).

[43] J. Szanyi, T. Wilson, S. Howe, et al., "Epidemiologic and economic modelling of optimal COVID-19 policy: Public health and social measures, masks and vaccines in victoria, australia," The Lancet Regional Health - Western Pacific, vol. 32, p. 100 675, Mar. 2023. DOI: 10.1016/j.lanwpc.2022.100675. [Online]. Available: https://doi.org/10.1016/j.lanwpc.2022.100675 (cited on p. 27).

[44] F. Galbusera, P. Côtè, and S. Negrini, "Expected impact of lockdown measures due to COVID-19 on disabling conditions: A modelling study of chronic low back pain," European Spine Journal, vol. 30, no. 10, pp. 2944–2954, Jul. 2021. DOI: 10.1007/s00586-021-06940-y. [Online]. Available: https://doi.org/10.1007/s00586-021-06940-y (cited on p. 27).

[45] C. C. Kerr, R. M. Stuart, D. Mistry, et al., "Covasim: An agent-based model of COVID-19 dynamics and interventions," PLOS Computational Biology, vol. 17, no. 7, M. Marz, Ed., e1009149, Jul. 2021. DOI: 10.1371/journal.pcbi.1009149. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1009149 (cited on p. 27).

[46] H. Chemaitelly and L. Abu-Raddad, "Characterizing HIV epidemiology in stable couples in cambodia, the dominican republic, haiti, and india," Epidemiology and Infection, vol. 144, no. 1, pp. 90–96, Apr. 2015. DOI: 10.1017/s0950268815000758. [Online]. Available: https://doi.org/10.1017/s0950268815000758 (cited on p. 28).

[47] Y.-J. Zhang, X.-X. Feng, Y.-G. Fan, et al., "HIV transmission and related risk factors among serodiscordant couples in liuzhou, china," Journal of Medical Virology, vol. 87, no. 4, pp. 553–556, Jan. 2015. DOI: 10.1002/jmv.24093. [Online]. Available: https://doi.org/10.1002/jmv.24093 (cited on p. 28).

[48] A. F. Fagbamigbe, S. B. Adebayo, and E. Idemudia, "Marital status and HIV prevalence among women in nigeria: Ingredients for evidence-based programming," International Journal of Infectious Diseases, vol. 48, pp. 57–63, Jul. 2016. DOI: 10.1016/j.ijid.2016.05.002. [Online]. Available: https://doi.org/10.1016/j.ijid.2016.05.002 (cited on p. 28).

[49] M. R. Group, Titan model, Dec. 2020. [Online]. Available: https://www.titanmodel.org/ (cited on pp. 28, 34, 37, 231).

[50] L. Martinez, M. E. C. Juliet N Sekandi, S. Zalwango, and C. C. Whalen, "Infectiousness of HIV-seropositive patients with tuberculosis in a high-burden african setting," American Journal of Respiratory and Critical Care Medicine, vol. 194, no. 9, pp. 1152–1163, Nov. 2016. DOI: 10.1164/rccm.201511-2146oc. [Online]. Available: https://doi.org/10.1164/rccm.201511-2146oc (cited on p. 28).

[51] K. Velen, J. J. Lewis, S. Charalambous, et al., "Household HIV testing uptake among contacts of TB patients in south africa," PLOS ONE, vol. 11, no. 5, E. M. Shankar, Ed., e0155688, May 2016. DOI: 10.1371/journal.pone.0155688. [Online]. Available: https://doi.org/10.1371/journal.pone.0155688 (cited on p. 28).

[52] World Health Organization, "Global health sector strategy on viral hepatitis 2016-2021. towards ending viral hepatitis," Technical documents, 2016, 53 p. (Cited on p. 28).

[53] N. Martin, G. Foster, J. Vilar, et al., "HCV treatment rates and sustained viral response among people who inject drugs in seven UK sites: Real world results and modelling of treatment impact," Journal of Viral Hepatitis, vol. 22, no. 4, pp. 399–408, Oct. 2014. DOI: 10.1111/jvh.12338. [Online]. Available: https://doi.org/10.1111/jvh.12338 (cited on pp. 29, 232).

[54] T. He, K. Li, M. S. Roberts, et al., "Prevention of hepatitis c by screening and treatment in u.s. prisons," Annals of Internal Medicine, vol. 164, no. 2, p. 84, Nov. 2015. DOI: 10.7326/m15-0617. [Online]. Available: https://doi.org/10.7326/m15-0617 (cited on pp. 29, 232).

[55] B. D. Smith, R. L. Morgan, G. A. Beckett, et al., "Recommendations for the identification of chronic hepatitis C virus infection among persons born during 1945-1965," en, MMWR Recomm. Rep., vol. 61, no. RR-4, pp. 1–32, Aug. 2012 (cited on p. 29).

[56] B. Smith, N. Patel, G. Beckett, A. Jewett, and W. Ward, "Hepatitis c virus antibody prevalence, correlates and predictors among persons born from 1945 through 1965," The Liver Meeting, vol. 2008, 1999 (cited on p. 29).

[57] B. D. Smith, R. L. Morgan, G. A. Beckett, Y. Falck-Ytter, D. Holtzman, and J. W. Ward, "Hepatitis c virus testing of persons born during 19451965: Recommendations from the centers for disease control and prevention," Annals of Internal Medicine, vol. 157, no. 11, p. 817, Dec. 2012. DOI: 10.7326/0003-

4819-157-9-201211060-00529. [Online]. Available: https://doi.org/10.7326/0003-4819-157-9-201211060-00529 (cited on p. 29).

[58] World Health Organisation, Dengue and severe dengue. world health organisation, Nov. 4, 2019. [Online]. Available: who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue (visited on 01/10/2020) (cited on p. 29).

[59] D. Murillo, A. Murillo, and S. Lee, “The role of vertical transmission in the control of dengue fever,” International Journal of Environmental Research and Public Health, vol. 16, no. 5, p. 803, Mar. 2019. DOI: 10.3390/ijerph16050803. [Online]. Available: https://doi.org/10.3390/ijerph16050803 (cited on p. 29).

[60] G. H. Stresman, J. Mwesigwa, J. Achan, et al., “Do hotspots fuel malaria transmission: A village-scale spatio-temporal analysis of a 2-year cohort study in the gambia,” BMC Medicine, vol. 16, no. 1, Sep. 2018. DOI: 10.1186/s12916-018-1141-4. [Online]. Available: https://doi.org/10.1186/s12916-018-1141-4 (cited on p. 29).

[61] J. Pinchoff, M. Mulenga, W. J. Moss, et al., “Predictive malaria risk and uncertainty mapping in nchelenge district, zambia: Evidence of widespread, persistent risk and implications for targeted interventions,” The American Journal of Tropical Medicine and Hygiene, vol. 93, no. 6, pp. 1260–1267, Dec. 2015. DOI: 10.4269/ajtmh.15-0283. [Online]. Available: https://doi.org/10.4269/ajtmh.15-0283 (cited on p. 29).

[62] A. N. Kabaghe, M. G. Chipeta, S. Gowelo, et al., “Fine-scale spatial and temporal variation of clinical malaria incidence and associated factors in children in rural malawi: A longitudinal study,” Parasites & Vectors, vol. 11, no. 1, Mar. 2018. DOI: 10.1186/s13071-018-2730-y. [Online]. Available: https://doi.org/10.1186/s13071-018-2730-y (cited on p. 30).

[63] N. C. P. Rodrigues, V. T. S. Lino, R. P. Daumas, et al., “Temporal and spatial evolution of dengue incidence in brazil, 2001-2012,” PLOS ONE, vol. 11, no. 11, Y.-H. Hsieh, Ed., e0165945, Nov. 2016. DOI: 10.1371/journal.pone.0165945. [Online]. Available: https://doi.org/10.1371/journal.pone.0165945 (cited on p. 30).

[64] A. Morales-Pérez, E. Nava-Aguilera, J. Legorreta-Soberanis, et al., “where we put little fish in the water there are no mosquitoes: a cross-sectional study on biological control of the aedes aegypti vector in 90 coastal-region communities of guerrero, mexico,” BMC Public Health, vol. 17, no. S1, May 2017. DOI: 10.

1186/s12889-017-4302-z. [Online]. Available: https://doi.org/10.1186/s12889-017-4302-z (cited on p. 30).

[65] T. K. Tsang, T.-M. Chen, I. M. Longini, M. E. Halloran, Y. Wu, and Y. Yang, "Transmissibility of norovirus in urban versus rural households in a large community outbreak in china," Epidemiology, vol. 29, no. 5, pp. 675–683, Sep. 2018. DOI: 10.1097/ede.0000000000000855. [Online]. Available: https://doi.org/10.1097/ede.0000000000000855 (cited on p. 30).

[66] J. P. Emont, A. I. Ko, A. Homasi-Paelate, N. Ituaso-Conway, and E. J. Nilles, "Epidemiological investigation of a diarrhea outbreak in the south pacific island nation of tuvalu during a severe la niñaassociated drought emergency in 2011," The American Journal of Tropical Medicine and Hygiene, pp. 16–0812, Jan. 2017. DOI: 10.4269/ajtmh.16-0812. [Online]. Available: https://doi.org/10.4269/ajtmh.16-0812 (cited on p. 30).

[67] M. N. Friedrich, T. R. Julian, A. Kappler, T. Nhiwatiwa, and H.-J. Mosler, "Handwashing, but how? microbial effectiveness of existing handwashing practices in high-density suburbs of harare, zimbabwe," American Journal of Infection Control, vol. 45, no. 3, pp. 228–233, Mar. 2017. DOI: 10.1016/j.ajic.2016.06.035. [Online]. Available: https://doi.org/10.1016/j.ajic.2016.06.035 (cited on p. 30).

[68] B. J. Paterson, M. D. Kirk, A. S. Cameron, C. D'Este, and D. N. Durrheim, "Historical data and modern methods reveal insights in measles epidemiology: A retrospective closed cohort study," BMJ Open, vol. 3, no. 1, e002033, 2013. DOI: 10.1136/bmjopen-2012-002033. [Online]. Available: https://doi.org/10.1136/bmjopen-2012-002033 (cited on p. 31).

[69] W. J. Moss and D. E. Griffin, "Global measles elimination," Nature Reviews Microbiology, vol. 4, no. 12, pp. 900–908, Nov. 2006. DOI: 10.1038/nrmicro1550. [Online]. Available: https://doi.org/10.1038/nrmicro1550 (cited on p. 31).

[70] A. J. Distasio and D. H. Trump, "The investigation of a tuberculosis outbreak in the closed environment of a u.s. navy ship, 1987," Military Medicine, vol. 155, no. 8, pp. 347–351, Aug. 1990. DOI: 10.1093/milmed/155.8.347. [Online]. Available: https://doi.org/10.1093/milmed/155.8.347 (cited on p. 31).

[71] D. M. Lister, D. Kotsanas, S. A. Ballard, et al., "Outbreak of vanB vancomycin-resistant enterococcus faecium colonization in a neonatal service,"

American Journal of Infection Control, vol. 43, no. 10, pp. 1061–1065, Oct. 2015. DOI: 10.1016/j.ajic.2015.05.047. [Online]. Available: https://doi.org/10.1016/j.ajic.2015.05.047 (cited on p. 32).

[72] J. Holmes and G. Simmons, "Gastrointestinal illness associated with a long-haul flight," Epidemiology and Infection, vol. 137, no. 3, pp. 441–447, Aug. 2008. DOI: 10.1017/s0950268808001027. [Online]. Available: https://doi.org/10.1017/s0950268808001027 (cited on p. 32).

[73] F. de Ritis, L. Mallucci, M. Coltorti, G. Giusti, and M. Caldera, "Anicteric virus hepatitis in a closed environment as shown by serum transaminase activity," en, Bull. World Health Organ., vol. 20, pp. 589–602, 1959 (cited on p. 32).

[74] A. Falchi, J. P. Amoros, C. Arena, et al., "Genetic structure of human a/h1n1 and a/h3n2 influenza virus on corsica island: Phylogenetic analysis and vaccine strain match, 20062010," PLoS ONE, vol. 6, no. 9, M. G. Semple, Ed., e24471, Sep. 2011. DOI: 10.1371/journal.pone.0024471. [Online]. Available: https://doi.org/10.1371/journal.pone.0024471 (cited on p. 32).

[75] Arakawa, Moriyama, Taira, et al., "Molecular analysis of hepatitis d virus infection in miyako island, a small japanese island," Journal of Viral Hepatitis, vol. 7, no. 5, pp. 375–381, Sep. 2000. DOI: 10.1046/j.1365-2893.2000.00244.x. [Online]. Available: https://doi.org/10.1046/j.1365-2893.2000.00244.x (cited on p. 32).

[76] Y. Sasaki, A. Kai, Y. Hayashi, et al., "Multiple viral infections and genomic divergence among noroviruses during an outbreak of acute gastroenteritis," Journal of Clinical Microbiology, vol. 44, no. 3, pp. 790–797, Mar. 2006. DOI: 10.1128/jcm.44.3.790-797.2006. [Online]. Available: https://doi.org/10.1128/jcm.44.3.790-797.2006 (cited on p. 32).

[77] T. Kageyama, M. Shinohara, K. Uchida, et al., "Coexistence of multiple genotypes, including newly identified genotypes, in outbreaks of gastroenteritis due to norovirus in japan," Journal of Clinical Microbiology, vol. 42, no. 7, pp. 2988–2995, Jul. 2004. DOI: 10.1128/jcm.42.7.2988-2995.2004. [Online]. Available: https://doi.org/10.1128/jcm.42.7.2988-2995.2004 (cited on p. 33).

[78] T. J. R. Finnie, I. M. Hall, and S. Leach, "Behaviour and control of influenza in institutions and small societies," Journal of the Royal Society of Medicine, vol. 105, no. 2, pp. 66–73, Feb. 2012. DOI: 10.1258/jrsm.2012.110249.

[Online]. Available: https://doi.org/10.1258/jrsm.2012.110249 (cited on p. 33).

[79] T. Finnie, V. Copley, I. Hall, and S. Leach, "An analysis of influenza outbreaks in institutions and enclosed societies," Epidemiology and Infection, vol. 142, no. 1, pp. 107–113, Apr. 2013. DOI: 10.1017/s0950268813000733. [Online]. Available: https://doi.org/10.1017/s0950268813000733 (cited on p. 33).

[80] C. Viboud, P.-Y. Boëlle, S. Cauchemez, et al., "Risk factors of influenza transmission in households," en, Br. J. Gen. Pract., vol. 54, no. 506, pp. 684–689, Sep. 2004 (cited on p. 33).

[81] A. Taylor, A. Munro, E. Allen, et al., "Low incidence of hepatitis c virus among prisoners in scotland," Addiction, vol. 108, no. 7, pp. 1296–1304, Mar. 2013. DOI: 10.1111/add.12107. [Online]. Available: https://doi.org/10.1111/add.12107 (cited on pp. 33, 232, 241).

[82] V. Hope, M. Hickman, S. Ngui, et al., "Measuring the incidence, prevalence and genetic relatedness of hepatitis c infections among a community recruited sample of injecting drug users, using dried blood spots," Journal of Viral Hepatitis, vol. 18, no. 4, pp. 262–270, Apr. 2010. DOI: 10.1111/j.1365-2893.2010.01297.x. [Online]. Available: https://doi.org/10.1111/j.1365-2893.2010.01297.x (cited on pp. 33, 240).

[83] E. L. Merrall, A. Kariminia, I. A. Binswanger, et al., "Meta-analysis of drug-related deaths soon after release from prison," Addiction, vol. 105, no. 9, pp. 1545–1554, Jun. 2010. DOI: 10.1111/j.1360-0443.2010.02990.x. [Online]. Available: https://doi.org/10.1111/j.1360-0443.2010.02990.x (cited on pp. 34, 233).

[84] F. Huber, A. Merceron, Y. Madec, et al., "High mortality among male HIV-infected patients after prison release: ART is not enough after incarceration with HIV," PLOS ONE, vol. 12, no. 4, G. Fischer, Ed., e0175740, Apr. 2017. DOI: 10.1371/journal.pone.0175740. [Online]. Available: https://doi.org/10.1371/journal.pone.0175740 (cited on pp. 34, 233).

[85] J. Stone, N. K. Martin, M. Hickman, et al., "Modelling the impact of incarceration and prison-based hepatitis c virus (HCV) treatment on HCV transmission among people who inject drugs in scotland," Addiction, vol. 112, no. 7, pp. 1302–1314, Mar. 2017. DOI: 10.1111/add.13783. [Online]. Available: https://doi.org/10.1111/add.13783 (cited on pp. 34, 233).

[86] F. L. Altice, L. Azbel, J. Stone, et al., “The perfect storm: Incarceration and the high-risk environment perpetuating transmission of HIV, hepatitis c virus, and tuberculosis in eastern europe and central asia,” The Lancet, vol. 388, no. 10050, pp. 1228–1248, Sep. 2016. DOI: 10.1016/s0140-6736(16)30856-x. [Online]. Available: https://doi.org/10.1016/s0140-6736(16)30856-x (cited on pp. 34, 35, 230, 231).

[87] J. W. Adams, M. N. Lurie, M. R. King, et al., “Potential drivers of HIV acquisition in african-american women related to mass incarceration: An agent-based modelling study,” BMC Public Health, vol. 18, no. 1, Dec. 2018. DOI: 10.1186/s12889-018-6304-x. [Online]. Available: https://doi.org/10.1186/s12889-018-6304-x (cited on pp. 34, 38, 231).

[88] T. S. Mabud, M. de Lourdes Delgado Alves, A. I. Ko, et al., “Evaluating strategies for control of tuberculosis in prisons and prevention of spillover into communities: An observational and modeling study from brazil,” PLOS Medicine, vol. 16, no. 1, J. Z. Metcalfe, Ed., e1002737, Jan. 2019. DOI: 10.1371/journal.pmed.1002737. [Online]. Available: https://doi.org/10.1371/journal.pmed.1002737 (cited on pp. 34, 231).

[89] H. A. A. Al-Darraji, A. Kamarulzaman, and F. L. Altice, “Isoniazid preventive therapy in correctional facilities: A systematic review [review article],” The International Journal of Tuberculosis and Lung Disease, vol. 16, no. 7, pp. 871–879, Jul. 2012. DOI: 10.5588/ijtld.11.0447. [Online]. Available: https://doi.org/10.5588/ijtld.11.0447 (cited on p. 34).

[90] M. Smieja, C. Marchetti, D. Cook, and F. M. Smaill, “Isoniazid for preventing tuberculosis in non-HIV infected persons,” Cochrane Database of Systematic Reviews, Jan. 1999. DOI: 10.1002/14651858.cd001363. [Online]. Available: https://doi.org/10.1002/14651858.cd001363 (cited on p. 34).

[91] C. Akolo, I. Adetifa, S. Shepperd, and J. Volmink, “Treatment of latent tuberculosis infection in HIV infected persons,” Cochrane Database of Systematic Reviews, Jan. 2010. DOI: 10.1002/14651858.cd000171.pub3. [Online]. Available: https://doi.org/10.1002/14651858.cd000171.pub3 (cited on p. 34).

[92] M. H. Cohen, K. M. Weber, N. Lancki, et al., “History of incarceration among women with HIV: Impact on prognosis and mortality,” Journal of Women’s Health, vol. 28, no. 8, pp. 1083–1093, Aug. 2019. DOI: 10.1089/jwh.2018.7454. [Online]. Available: https://doi.org/10.1089/jwh.2018.7454 (cited on pp. 35, 231).

[93] M. Erickson, K. Shannon, A. Sernick, et al., "Women, incarceration and HIV," AIDS, vol. 33, no. 1, pp. 101–111, Jan. 2019. DOI: 10.1097/qad.0000000000002036. [Online]. Available: https://doi.org/10.1097/qad.0000000000002036 (cited on p. 35).

[94] R. P. Westergaard, G. D. Kirk, D. R. Richesson, N. Galai, and S. H. Mehta, "Incarceration predicts virologic failure for HIV-infected injection drug users receiving antiretroviral therapy," Clinical Infectious Diseases, vol. 53, no. 7, pp. 725–731, Sep. 2011. DOI: 10.1093/cid/cir491. [Online]. Available: https://doi.org/10.1093/cid/cir491 (cited on p. 35).

[95] M. L. Ndeffo-Mbah, V. S. Vigliotti, L. A. Skrip, K. Dolan, and A. P. Galvani, "Dynamic models of infectious disease transmission in prisons and the general population," Epidemiologic Reviews, vol. 40, no. 1, pp. 40–57, 2018. DOI: 10.1093/epirev/mxx014. [Online]. Available: https://doi.org/10.1093/epirev/mxx014 (cited on pp. 35, 231).

[96] Y. Su and S. S. Yoon, "Epi info - present and future," en, AMIA Annu. Symp. Proc., p. 1023, 2003 (cited on p. 35).

[97] I. J. Schafer, E. Knudsen, L. A. McNamara, S. Agnihotri, P. E. Rollin, and A. Islam, "The epi info viral hemorrhagic fever (VHF) application: A resource for outbreak data management and contact tracing in the 20142016 west africa ebola epidemic," Journal of Infectious Diseases, vol. 214, no. suppl 3, S122–S136, Sep. 2016. DOI: 10.1093/infdis/jiw272. [Online]. Available: https://doi.org/10.1093/infdis/jiw272 (cited on p. 35).

[98] S. Spoto, M. Wiese, and M. Lyman, "Implementation of a facility based county surveillance system using epi info," Online Journal of Public Health Informatics, vol. 10, no. 1, May 2018. DOI: 10.5210/ojphi.v10i1.8919. [Online]. Available: https://doi.org/10.5210/ojphi.v10i1.8919 (cited on pp. 35, 36).

[99] E. Bumburidi, S. Ajeilat, A. Dadu, et al., "Progress toward tuberculosis control and determinants of treatment outcomes–kazakhstan, 2000-2002.," MMWR Suppl., vol. 55, no. 1, 2006 (cited on p. 35).

[100] B. Camp, J. K. Mandivarapu, N. Ramamurthy, et al., "A new cross-platform architecture for epi-info software suite," BMC Bioinformatics, vol. 19, no. S11, Oct. 2018. DOI: 10.1186/s12859-018-2334-8. [Online]. Available: https://doi.org/10.1186/s12859-018-2334-8 (cited on p. 36).

[101] J. H. Abramson and E. Peritz, Calculator programs for the health sciences. 1982 (cited on p. 37).

[102] J. H. Abramson, “Winpepi (pepi-for-windows): Computer programs for epidemiologists,” Epidemiologic Perspectives & Innovations, vol. 1, no. 1, p. 6, 2004. DOI: `10.1186/1742-5573-1-6`. [Online]. Available: `https://doi.org/10.1186/1742-5573-1-6` (cited on p. 37).

[103] Joseph H Abramson, “WINPEPI updated: Computer programs for epidemiologists, and their teaching potential,” Epidemiologic Perspectives & Innovations, vol. 8, no. 1, p. 1, 2011. DOI: `10.1186/1742-5573-8-1`. [Online]. Available: `https://doi.org/10.1186/1742-5573-8-1` (cited on p. 37).

[104] E. Ekpe and C. Eyo, “Determinants of mortality in chest trauma patients,” Niger J Surg., vol. 20, no. 1, pp. 30–34, 2014. DOI: `10.4103/1117-6806.127107`. [Online]. Available: `https://pubmed.ncbi.nlm.nih.gov/24665200/` (cited on p. 37).

[105] S. Levi, A. Zini, S. Fischman, and R. Czerninski, “Epidemiology of oral, salivary gland and pharyngeal cancer in children and adolescents between 1970 and 2011,” Oral Oncology, vol. 67, pp. 89–94, Apr. 2017. DOI: `10.1016/j.oraloncology.2017.02.010`. [Online]. Available: `https://doi.org/10.1016/j.oraloncology.2017.02.010` (cited on p. 37).

[106] A. Z. Win, C. Ceresa, K. Arnold, and T. Allison, “High prevalence of malnutrition among elderly veterans in home based primary care,” The journal of nutrition, health & aging, vol. 21, no. 6, pp. 610–613, Apr. 2017. DOI: `10.1007/s12603-017-0918-z`. [Online]. Available: `https://doi.org/10.1007/s12603-017-0918-z` (cited on p. 37).

[107] J. V. Madruga, D. Berger, M. McMurchie, et al., “Efficacy and safety of darunavir-ritonavir compared with that of lopinavir-ritonavir at 48 weeks in treatment-experienced, HIV-infected patients in TITAN: A randomised controlled phase III trial,” The Lancet, vol. 370, no. 9581, pp. 49–58, Jul. 2007. DOI: `10.1016/s0140-6736(07)61049-6`. [Online]. Available: `https://doi.org/10.1016/s0140-6736(07)61049-6` (cited on p. 37).

[108] J. W. Adams, M. N. Lurie, M. R. King, et al., “Decreasing HIV transmissions to african american women through interventions for men living with HIV post-incarceration: An agent-based modeling study,” PLOS ONE, vol. 14, no. 7, Y. E. Khudyakov, Ed., e0219361, Jul. 2019. DOI: `10.1371/journal.`

pone.0219361. [Online]. Available: https://doi.org/10.1371/journal.pone.0219361 (cited on p. 38).

[109] S. Raschka, The key differences between python 2.7.x and python 3.x with examples. 2014. [Online]. Available: https://sebastianraschka.com/Articles/2014_python_2_3_key_diff.html (visited on 12/12/2019) (cited on p. 38).

[110] A. Bershteyn, J. Gerardin, D. Bridenbecker, et al., "Implementation and applications of EMOD, an individual-based multi-disease modeling platform," Pathogens and Disease, vol. 76, no. 5, Jul. 2018. DOI: 10.1093/femspd/fty059. [Online]. Available: https://doi.org/10.1093/femspd/fty059 (cited on p. 38).

[111] K. A. McCarthy, E. A. Wenger, G. H. Huynh, and P. A. Eckhoff, "Calibration of an intrahost malaria model and parameter ensemble evaluation of a pre-erythrocytic vaccine," Malaria Journal, vol. 14, no. 1, Jan. 2015. DOI: 10.1186/1475-2875-14-6. [Online]. Available: https://doi.org/10.1186/1475-2875-14-6 (cited on p. 39).

[112] K. A. McCarthy, G. Chabot-Couture, M. Famulare, H. M. Lyons, and L. D. Mercer, "The risk of type 2 oral polio vaccine use in post-cessation outbreak response," BMC Medicine, vol. 15, no. 1, Oct. 2017. DOI: 10.1186/s12916-017-0937-y. [Online]. Available: https://doi.org/10.1186/s12916-017-0937-y (cited on p. 39).

[113] C. C. Kerr, "Is epidemiology ready for big software?" Pathogens and Disease, vol. 77, no. 1, Jan. 2019. DOI: 10.1093/femspd/ftz006. [Online]. Available: https://doi.org/10.1093/femspd/ftz006 (cited on p. 39).

[114] J. V. Douglas, S. Bianco, S. Edlund, et al., "STEM: An open source tool for disease modeling," Health Security, vol. 17, no. 4, pp. 291–306, Aug. 2019. DOI: 10.1089/hs.2019.0018. [Online]. Available: https://doi.org/10.1089/hs.2019.0018 (cited on p. 39).

[115] The United Nations, Member states. [Online]. Available: https://www.un.org/en/member-states/ (visited on 12/12/2019) (cited on p. 40).

[116] S. Edlund, M. Davis, J. V. Douglas, et al., "A global model of malaria climate sensitivity: Comparing malaria response to historic climate data based on simulation and officially reported malaria incidence," Malaria Journal, vol. 11, no. 1, p. 331, 2012. DOI: 10.1186/1475-2875-11-331. [Online]. Available: https://doi.org/10.1186/1475-2875-11-331 (cited on p. 40).

[117] A. Falenski, M. Filter, C. Thöns, et al., "A generic open-source software framework supporting scenario simulations in bioterrorist crises," Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science, vol. 11, no. S1, S134–S145, Sep. 2013. DOI: 10.1089/bsp.2012.0071. [Online]. Available: https://doi.org/10.1089/bsp.2012.0071 (cited on p. 40).

[118] E. Tye, T. Finnie, R. James, I. M. Hall, and S. Leach, "Pygom -a python package for simplifying modelling with systems of ordinary differential equations," en, 2018. DOI: 10.13140/RG.2.2.32293.86247. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.32293.86247 (cited on pp. 40, 41).

[119] F. Campbell, X. Didelot, R. Fitzjohn, N. Ferguson, A. Cori, and T. Jombart, "Outbreaker2: A modular platform for outbreak reconstruction," en, BMC Bioinformatics, vol. 19, no. Suppl 11, p. 363, Oct. 2018 (cited on p. 40).

[120] M. Abbas, A. Cori, S. Cordey, et al., "Reconstruction of transmission chains of SARS-CoV-2 amidst multiple outbreaks in a geriatric acute-care hospital: A combined retrospective epidemiological and genomic study," en, Elife, vol. 11, Jul. 2022 (cited on p. 41).

[121] K. E. Hjorleifsson, S. Rognvaldsson, H. Jonsson, et al., "Reconstruction of a large-scale outbreak of SARS-CoV-2 infection in iceland informs vaccination strategies," en, Clin. Microbiol. Infect., vol. 28, no. 6, pp. 852–858, Jun. 2022 (cited on p. 41).

[122] K. J. Siddle, L. A. Krasilnikova, G. K. Moreno, et al., "Transmission from vaccinated individuals in a large SARS-CoV-2 delta variant outbreak," en, Cell, vol. 185, no. 3, 485–492.e10, Feb. 2022 (cited on p. 41).

[123] R. N. Thompson, J. E. Stockwin, R. D. van Gaalen, et al., "Improved inference of time-varying reproduction numbers during infectious disease outbreaks," en, Epidemics, vol. 29, no. 100356, p. 100 356, Dec. 2019 (cited on p. 42).

[124] J. Wallinga and P. Teunis, "Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures," en, Am. J. Epidemiol., vol. 160, no. 6, pp. 509–516, Sep. 2004 (cited on p. 42).

[125] PubMed, Pubmed search. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/?term=epiestim&size=200 (visited on 03/13/2023) (cited on p. 42).

[126] A. Brizzi, M. O'Driscoll, and I. Dorigatti, "Refining reproduction number estimates to account for unobserved generations of infection in emerging epi-

demics," en, Clin. Infect. Dis., vol. 75, no. 1, e114–e121, Aug. 2022 (cited on p. 42).

[127] U. Wilensky, Netlogo, 1999. [Online]. Available: http://ccl.northwestern.edu/netlogo/ (visited on 12/20/2019) (cited on p. 43).

[128] W. M. Getz and E. R. Dougherty, "Discrete stochastic analogs of erlang epidemic models," Journal of Biological Dynamics, vol. 12, no. 1, pp. 16–38, Nov. 2017. DOI: 10.1080/17513758.2017.1401677. [Online]. Available: https://doi.org/10.1080/17513758.2017.1401677 (cited on p. 43).

[129] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al., "The FAIR guiding principles for scientific data management and stewardship," en, Sci. Data, vol. 3, no. 1, p. 160 018, Mar. 2016 (cited on p. 44).

[130] M. Barker, N. P. Chue Hong, D. S. Katz, et al., "Introducing the FAIR principles for research software," en, Sci. Data, vol. 9, no. 1, p. 622, Oct. 2022 (cited on p. 44).

[131] J. Declan. "Safety match." (2020), [Online]. Available: https://juandelcan.com/matchsticks (visited on 03/03/2023) (cited on p. 46).

[132] B. Hollingsworth, K. W. Okamoto, and A. L. Lloyd, "After the honeymoon, the divorce: Unexpected outcomes of disease control measures against endemic infections," en, PLoS Comput. Biol., vol. 16, no. 10, e1008292, Oct. 2020 (cited on p. 47).

[133] C. E. Overton, H. B. Stage, S. Ahmad, et al., "Using statistics and mathematical modelling to understand infectious disease outbreaks: COVID-19 as an example," Infectious Disease Modelling, Jul. 2020. DOI: 10.1016/j.idm.2020.06.008. [Online]. Available: https://doi.org/10.1016/j.idm.2020.06.008 (cited on pp. 50, 58, 122, 178, 184, 208, 218).

[134] K. Mizumoto, K. Kagaya, A. Zarebski, and G. Chowell, "Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020," Eurosurveillance, vol. 25, no. 10, Mar. 2020. DOI: 10.2807/1560-7917.es.2020.25.10.2000180. [Online]. Available: https://doi.org/10.2807/1560-7917.es.2020.25.10.2000180 (cited on pp. 50, 161).

[135] S. A. Lauer, K. H. Grantz, Q. Bi, et al., "The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application," Annals of Internal Medicine, vol. 172, no. 9, pp. 577–582,

May 2020. DOI: 10.7326/m20-0504. [Online]. Available: https://doi.org/10.7326/m20-0504 (cited on p. 50).

[136] J. Qin, C. You, Q. Lin, T. Hu, S. Yu, and X.-H. Zhou, “Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study,” Science Advances, vol. 6, no. 33, Aug. 2020. DOI: 10.1126/sciadv.abc1202. [Online]. Available: https://doi.org/10.1126/sciadv.abc1202 (cited on p. 50).

[137] J. A. Backer, D. Klinkenberg, and J. Wallinga, “Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from wuhan, china, 2028 january 2020,” Eurosurveillance, vol. 25, no. 5, Feb. 2020. DOI: 10.2807/1560-7917.es.2020.25.5.2000062. [Online]. Available: https://doi.org/10.2807/1560-7917.es.2020.25.5.2000062 (cited on p. 50).

[138] J. R. Egan and I. M. Hall, “A review of back-calculation techniques and their potential to inform mitigation strategies with application to non-transmissible acute infectious diseases,” en, J. R. Soc. Interface, vol. 12, no. 106, p. 20 150 096, May 2015 (cited on pp. 50, 52).

[139] H. L. Kirking, J. Cortes, S. Burrer, et al., “Likely transmission of norovirus on an airplane, october 2008,” Clinical Infectious Diseases, vol. 50, no. 9, pp. 1216–1221, May 2010. DOI: 10.1086/651597. [Online]. Available: https://doi.org/10.1086/651597 (cited on p. 50).

[140] X. Ren, Y. Li, X. Yang, et al., “Evidence for pre-symptomatic transmission of coronavirus disease 2019 (COVID-19) in china,” Influenza and Other Respiratory Viruses, vol. 15, no. 1, pp. 19–26, Aug. 2020. DOI: 10.1111/irv.12787. [Online]. Available: https://doi.org/10.1111/irv.12787 (cited on p. 50).

[141] S. Kim, Y. B. Seo, and E. Jung, “Prediction of COVID-19 transmission dynamics using a mathematical model considering behavior changes,” Epidemiology and Health, e2020026, Apr. 2020. DOI: 10.4178/epih.e2020026. [Online]. Available: https://doi.org/10.4178/epih.e2020026 (cited on p. 50).

[142] A. Akinbi, M. Forshaw, and V. Blinkhorn, “Contact tracing apps for the COVID-19 pandemic: A systematic literature review of challenges and future directions for neo-liberal societies,” Health Information Science and Systems, vol. 9, no. 1, Apr. 2021. DOI: 10.1007/s13755-021-00147-7. [Online]. Available: https://doi.org/10.1007/s13755-021-00147-7 (cited on p. 51).

[143] J. A. M. López, B. A. Garca, P. Bentkowski, et al., “Anatomy of digital contact tracing: Role of age, transmission setting, adoption, and case detection,” Science Advances, vol. 7, no. 15, Apr. 2021. DOI: 10.1126/sciadv.abd8750. [Online]. Available: https://doi.org/10.1126/sciadv.abd8750 (cited on p. 51).

[144] A. Pratt, E. Bennett, J. Gillard, S. Leach, and I. Hall, “Dose-response modeling: Extrapolating from experimental data to real-world populations,” en, Risk Anal., vol. 41, no. 1, pp. 67–78, Jan. 2021 (cited on p. 51).

[145] “An invariant form for the prior probability in estimation problems,” Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, vol. 186, no. 1007, pp. 453–461, Sep. 1946. DOI: 10.1098/rspa.1946.0056. [Online]. Available: https://doi.org/10.1098/rspa.1946.0056 (cited on pp. 54, 236).

[146] E. L. Lehmann and G. Casella, Theory of Point Estimation, Second. New York, NY, USA: Springer-Verlag, 1998, p. 116 (cited on p. 54).

[147] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, Bayesian Data Analysis. Chapman and Hall/CRC, Jun. 1995. DOI: 10.1201/9780429258411. [Online]. Available: https://doi.org/10.1201/9780429258411 (cited on pp. 54, 105, 236).

[148] M. Abramowitz, I. A. Stegun, and R. H. Romer, “Handbook of mathematical functions with formulas, graphs, and mathematical tables,” American Journal of Physics, vol. 56, no. 10, pp. 958–958, Oct. 1988. DOI: 10.1119/1.15378. [Online]. Available: https://doi.org/10.1119/1.15378 (cited on p. 58).

[149] L. V. Kantorovich, “Mathematical methods of organizing and planning production,” Management Science, vol. 6, no. 4, pp. 366–422, Jul. 1960. DOI: 10.1287/mnsc.6.4.366. [Online]. Available: https://doi.org/10.1287/mnsc.6.4.366 (cited on p. 61).

[150] C. E. Shannon, “A mathematical theory of communication,” Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, Jul. 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x. [Online]. Available: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x (cited on p. 61).

[151] C. E. Shannon, “A mathematical theory of communication,” Bell System Technical Journal, vol. 27, no. 4, pp. 623–656, Oct. 1948. DOI: 10.1002/j.1538-7305.1948.tb00917.x. [Online]. Available: https://doi.org/10.1002/j.1538-7305.1948.tb00917.x (cited on p. 61).

[152] T. G. Martin, B. A. Wintle, J. R. Rhodes, et al., "Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations," Ecology Letters, vol. 8, no. 11, pp. 1235–1246, Oct. 2005. DOI: 10.1111/j.1461-0248.2005.00826.x. [Online]. Available: https://doi.org/10.1111/j.1461-0248.2005.00826.x (cited on p. 74).

[153] K. S. Willebrand, L. Pischel, A. A. Malik, S. M. Jenness, and S. B. Omer, "A review of COVID-19 transmission dynamics and clinical outcomes on cruise ships worldwide, january to october 2020," Eurosurveillance, vol. 27, no. 1, Jan. 2022. DOI: 10.2807/1560-7917.es.2022.27.1.2002113. [Online]. Available: https://doi.org/10.2807/1560-7917.es.2022.27.1.2002113 (cited on pp. 74, 248).

[154] M. Belam, B. Quinn, and A. Rourke, "Cruise ship accounts for more than half of virus cases outside china  as it happened," The Guardian, Feb. 20, 2020. [Online]. Available: https://www.theguardian.com/world/live/2020/feb/20/coronavirus-live-updates-diamond-princess-cruise-ship-japan-deaths-latest-news-china-infections (visited on 02/03/2023) (cited on pp. 74, 246).

[155] D. P. Oran and E. J. Topol, "Prevalence of asymptomatic SARS-CoV-2 infection," Annals of Internal Medicine, vol. 173, no. 5, pp. 362–367, Sep. 2020. DOI: 10.7326/m20-3012. [Online]. Available: https://doi.org/10.7326/m20-3012 (cited on p. 79).

[156] S. Lee, T. Kim, E. Lee, et al., "Clinical course and molecular viral shedding among asymptomatic and symptomatic patients with SARS-CoV-2 infection in a community treatment center in the republic of korea," JAMA Internal Medicine, vol. 180, no. 11, p. 1447, Nov. 2020. DOI: 10.1001/jamainternmed.2020.3862. [Online]. Available: https://doi.org/10.1001/jamainternmed.2020.3862 (cited on p. 79).

[157] M. A. Almadhi, A. Abdulrahman, S. A. Sharaf, et al., "The high prevalence of asymptomatic SARS-CoV-2 infection reveals the silent spread of COVID-19," International Journal of Infectious Diseases, vol. 105, pp. 656–661, Apr. 2021. DOI: 10.1016/j.ijid.2021.02.100. [Online]. Available: https://doi.org/10.1016/j.ijid.2021.02.100 (cited on p. 79).

[158] J. C. Emery, T. W. Russell, Y. Liu, et al., "The contribution of asymptomatic SARS-CoV-2 infections to transmission on the diamond princess cruise ship,"

eLife, vol. 9, Aug. 2020. DOI: 10.7554/elife.58699. [Online]. Available: https://doi.org/10.7554/elife.58699 (cited on pp. 79, 247).

[159] N. G. Davies, P. Klepac, Y. Liu, et al., "Age-dependent effects in the transmission and control of COVID-19 epidemics," Nature Medicine, vol. 26, no. 8, pp. 1205–1211, Jun. 2020. DOI: 10.1038/s41591-020-0962-9. [Online]. Available: https://doi.org/10.1038/s41591-020-0962-9 (cited on pp. 80, 121, 122).

[160] Q. Bi, Y. Wu, S. Mei, et al., "Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in shenzhen, china: A retrospective cohort study," The Lancet Infectious Diseases, vol. 20, no. 8, pp. 911–919, Aug. 2020. DOI: 10.1016/s1473-3099(20)30287-5. [Online]. Available: https://doi.org/10.1016/s1473-3099(20)30287-5 (cited on p. 80).

[161] F. Riccardo, M. Ajelli, X. D. Andrianou, et al., "Epidemiological characteristics of COVID-19 cases and estimates of the reproductive numbers 1 month into the epidemic, italy, 28 january to 31 march 2020," Eurosurveillance, vol. 25, no. 49, Dec. 2020. DOI: 10.2807/1560-7917.es.2020.25.49.2000790. [Online]. Available: https://doi.org/10.2807/1560-7917.es.2020.25.49.2000790 (cited on p. 80).

[162] B. Doherty and D. Phillips, "Coronavirus: Cruise passengers stranded as countries turn them away," The Guardian, Mar. 2020. [Online]. Available: theguardian.com/world/2020/mar/16/cruise-ships-scramble-to-find-safe-harbour-amid-covid-19-crisis-as-countries-turn-them-away (cited on p. 117).

[163] B. Davies, M. Araghi, M. Moshe, et al., "Acceptability, usability, and performance of lateral flow immunoassay tests for severe acute respiratory syndrome coronavirus 2 antibodies: REACT-2 study of self-testing in nonhealthcare key workers," Open Forum Infectious Diseases, vol. 8, no. 11, Oct. 2021. DOI: 10.1093/ofid/ofab496. [Online]. Available: https://doi.org/10.1093/ofid/ofab496 (cited on p. 118).

[164] C.-C. Topriceanu, A. Wong, J. C. Moon, et al., "Impact of lockdown on key workers: Findings from the COVID-19 survey in four UK national longitudinal studies," Journal of Epidemiology and Community Health, vol. 75, no. 10, pp. 955–962, Apr. 2021. DOI: 10.1136/jech-2020-215889. [Online]. Available: https://doi.org/10.1136/jech-2020-215889 (cited on p. 118).

[165] A. J. Page, A. E. Mather, T. Le-Viet, et al., "Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management," Microbial Genomics, vol. 7, no. 6, Jun. 2021. DOI: 10.1099/mgen.0.000589. [Online]. Available: https://doi.org/10.1099/mgen.0.000589 (cited on p. 118).

[166] P. C. Moroni-Zentgraf, C. Keller, M. Mahmoudi, et al., "Pilot study of an occupational healthcare program to assess the SARS-CoV-2 infection and immune status of employees in a large pharmaceutical company," Current Medical Research and Opinion, vol. 37, no. 6, pp. 939–947, Apr. 2021. DOI: 10.1080/03007995.2021.1914943. [Online]. Available: https://doi.org/10.1080/03007995.2021.1914943 (cited on p. 118).

[167] M. K. Vasconcelos, C. Marazia, M. Koniordou, H. Fangerau, I. Drexler, and A. A.-A. Awuah, "A conceptual approach to the rationale for SARS-CoV-2 vaccine allocation prioritisation," Pathogens and Global Health, vol. 115, no. 5, pp. 273–276, Jun. 2021. DOI: 10.1080/20477724.2021.1932136. [Online]. Available: https://doi.org/10.1080/20477724.2021.1932136 (cited on p. 118).

[168] Z. Hu, C. Song, C. Xu, et al., "Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in nanjing, china," Science China Life Sciences, vol. 63, no. 5, pp. 706–711, Mar. 2020. DOI: 10.1007/s11427-020-1661-4. [Online]. Available: https://doi.org/10.1007/s11427-020-1661-4 (cited on pp. 119, 122).

[169] N. G. Davies, A. J. Kucharski, R. M. Eggo, et al., "Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: A modelling study," The Lancet Public Health, vol. 5, no. 7, e375–e385, Jul. 2020. DOI: 10.1016/s2468-2667(20)30133-x. [Online]. Available: https://doi.org/10.1016/s2468-2667(20)30133-x (cited on pp. 121, 122).

[170] A. R. Tuite, D. N. Fisman, and A. L. Greer, "Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of ontario, canada," Canadian Medical Association Journal, vol. 192, no. 19, E497–E505, Apr. 2020. DOI: 10.1503/cmaj.200476. [Online]. Available: https://doi.org/10.1503/cmaj.200476 (cited on pp. 121, 122).

[171] M. Abramowitz, I. A. Stegun, and R. H. Romer, "Handbook of mathematical functions with formulas, graphs, and mathematical tables," American Journal

of Physics, vol. 56, no. 10, p. 1020, Oct. 1988. DOI: 10.1119/1.15378. [Online]. Available: https://doi.org/10.1119/1.15378 (cited on p. 153).

[172] J. A. Al-Tawfiq, "Asymptomatic coronavirus infection: MERS-CoV and SARS-CoV-2 (COVID-19)," Travel Medicine and Infectious Disease, vol. 35, p. 101 608, May 2020. DOI: 10.1016/j.tmaid.2020.101608. [Online]. Available: https://doi.org/10.1016/j.tmaid.2020.101608 (cited on p. 161).

[173] H. Nishiura, "Backcalculating the incidence of infection with COVID-19 on the Diamond Princess," Journal of Clinical Medicine, vol. 9, no. 3, p. 657, Feb. 2020. DOI: 10.3390/jcm9030657. [Online]. Available: https://doi.org/10.3390/jcm9030657 (cited on pp. 161, 247).

[174] J. Stock, "Data gaps and the policy response to the novel coronavirus," Tech. Rep., Mar. 2020. DOI: 10.3386/w26902. [Online]. Available: https://doi.org/10.3386/w26902 (cited on p. 161).

[175] Z. F. Sikora A, Nosocomial Infections. Treasure Island (FL): StatPearls Publishing, 2022 (cited on p. 176).

[176] S. S. Magill, E. O'Leary, S. J. Janelle, et al., "Changes in prevalence of health care–associated infections in u.s. hospitals," en, N. Engl. J. Med., vol. 379, no. 18, pp. 1732–1744, Nov. 2018 (cited on p. 176).

[177] C. Suetens, K. Latour, T. Kärki, et al., "Prevalence of healthcare-associated infections, estimated incidence and composite antimicrobial resistance index in acute care hospitals and long-term care facilities: Results from two european point prevalence surveys, 2016 to 2017," en, Euro Surveill., vol. 23, no. 46, Nov. 2018 (cited on p. 176).

[178] B. Allegranzi, S. B. Nejad, C. Combescure, et al., "Burden of endemic health-care-associated infection in developing countries: Systematic review and meta-analysis," en, Lancet, vol. 377, no. 9761, pp. 228–241, Jan. 2011 (cited on p. 176).

[179] J. A. Jernigan, K. M. Hatfield, H. Wolford, et al., "Multidrug-resistant bacterial infections in u.s. hospitalized patients, 2012-2017," en, N. Engl. J. Med., vol. 382, no. 14, pp. 1309–1319, Apr. 2020 (cited on p. 176).

[180] D. M. Sievert, P. Ricks, J. R. Edwards, et al., "Antimicrobial-resistant pathogens associated with healthcare-associated infections summary of data reported to the national healthcare safety network at the centers for disease control and prevention, 2009–2010," en, Infect. Control Hosp. Epidemiol., vol. 34, no. 1, pp. 1–14, Jan. 2013 (cited on p. 176).

[181] J.-L. Vincent, "The prevalence of nosocomial infection in intensive care units in europe," JAMA, vol. 274, no. 8, p. 639, Aug. 1995 (cited on p. 176).

[182] L. Soufir, J. F. Timsit, C. Mahe, J. Carlet, B. Regnier, and S. Chevret, "Attributable morbidity and mortality of catheter-related septicemia in critically ill patients: A matched, risk-adjusted, cohort study," en, Infect. Control Hosp. Epidemiol., vol. 20, no. 6, pp. 396–401, Jun. 1999 (cited on p. 176).

[183] P. I. Ruth May Steve Powis and P. Phillip, Healthcare associated covid-19 infections  further action, 2020. [Online]. Available: england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2020/06/Healthcare-associated-COVID-19-infections-further-action-24-June-2020.pdf (cited on pp. 176, 202).

[184] C. Heneghan, D. Howdon, J. Oke, and J. T, "The ongoing problem of UK hospital acquired infections," The Centre for Evidence-Based Medicine, 2021. [Online]. Available: https://www.cebm.net/covid-19/the-ongoing-problem-of-hospital-acquired-infections-across-the-uk/ (cited on p. 177).

[185] Q. Zhou, Y. Gao, X. Wang, et al., "Nosocomial infections among patients with COVID-19, SARS and MERS: A rapid review and meta-analysis," Annals of Translational Medicine, vol. 8, no. 10, pp. 629–629, May 2020. DOI: 10.21037/atm-20-3324. [Online]. Available: https://doi.org/10.21037/atm-20-3324 (cited on p. 177).

[186] K. Khan, H. Reed-Embleton, J. Lewis, J. Saldanha, and S. Mahmud, "Does nosocomial COVID-19 result in increased 30-day mortality? a multi-centre observational study to identify risk factors for worse outcomes in patients with COVID-19," Journal of Hospital Infection, vol. 107, pp. 91–94, Jan. 2021. DOI: 10.1016/j.jhin.2020.09.017. [Online]. Available: https://doi.org/10.1016/j.jhin.2020.09.017 (cited on p. 177).

[187] B. Carter, J. Collins, F. Barlow-Pay, et al., "Nosocomial COVID-19 infection: Examining the risk of mortality. the COPE-nosocomial study (COVID in older PEople)," Journal of Hospital Infection, vol. 106, no. 2, pp. 376–384, Oct. 2020. DOI: 10.1016/j.jhin.2020.07.013. [Online]. Available: https://doi.org/10.1016/j.jhin.2020.07.013 (cited on p. 177).

[188] P. Diaconis and D. Ylvisaker, "Conjugate priors for exponential families," The Annals of Statistics, vol. 7, no. 2, Mar. 1979. DOI: 10.1214/aos/1176344611.

[Online]. Available: https://doi.org/10.1214/aos/1176344611 (cited on p. 189).

[189] L. D. Brown, “Fundamentals of statistical exponential families with applications in statistical decision theory,” Lecture Notes-Monograph Series, vol. 9, pp. i–279, 1986, ISSN: 07492170. [Online]. Available: http://www.jstor.org/stable/4355554 (cited on p. 189).

[190] A. Wiranto, A. Kurniawan, D. A. Fitria, Suliyanto, and N. Chamidah, “Estimation of type i censored exponential distribution parameters using objective bayesian and bootstrap methods (case study of chronic kidney failure patients),” Journal of Physics: Conference Series, vol. 1397, no. 1, p. 012 060, Dec. 2019. DOI: 10.1088/1742-6596/1397/1/012060. [Online]. Available: https://doi.org/10.1088/1742-6596/1397/1/012060 (cited on p. 189).

[191] J. M. Ellingford, R. George, J. H. McDermott, et al., “Genomic and healthcare dynamics of nosocomial SARS-CoV-2 transmission,” eLife, vol. 10, Mar. 2021. DOI: 10.7554/elife.65453. [Online]. Available: https://doi.org/10.7554/elife.65453 (cited on pp. 207, 219).

[192] UKHSA, Gov.uk coronavirus (covid-19) in the uk. [Online]. Available: https://coronavirus.data.gov.uk/details/cases?areaType=nation&areaName=England (visited on 02/27/2023) (cited on p. 208).

[193] J. Blackstone, O. Stirrup, F. Mapp, et al., “Protocol for the COG-UK hospital-onset COVID-19 infection (HOCI) multicentre interventional clinical study: Evaluating the efficacy of rapid genome sequencing of SARS-CoV-2 in limiting the spread of COVID-19 in UK NHS hospitals,” BMJ Open, vol. 12, no. 4, e052514, Apr. 2022. DOI: 10.1136/bmjopen-2021-052514. [Online]. Available: https://doi.org/10.1136/bmjopen-2021-052514 (cited on p. 218).

[194] O. Stirrup, J. Blackstone, F. Mapp, et al., “Effectiveness of rapid SARS-CoV-2 genome sequencing in supporting infection control for hospital-onset COVID-19 infection: Multicentre, prospective study,” eLife, vol. 11, Sep. 2022. DOI: 10.7554/elife.78427. [Online]. Available: https://doi.org/10.7554/elife.78427 (cited on p. 218).

[195] M. U. G. Kraemer, C.-H. Yang, B. Gutierrez, et al., “The effect of human mobility and control measures on the COVID-19 epidemic in china,” Science, vol. 368, no. 6490, pp. 493–497, May 2020. DOI: 10.1126/science.abb4218.

[Online]. Available: https://doi.org/10.1126/science.abb4218 (cited on p. 218).

[196] S. Evans, J. Stimson, D. Pople, et al., “Quantifying the contribution of pathways of nosocomial acquisition of COVID-19 in english hospitals,” International Journal of Epidemiology, vol. 51, no. 2, pp. 393–403, Dec. 2021. DOI: 10.1093/ije/dyab241. [Online]. Available: https://doi.org/10.1093/ije/dyab241 (cited on p. 219).

[197] N. K. Martin, P. Vickerman, I. F. Brew, et al., “Is increased hepatitis c virus case-finding combined with current or 8-week to 12-week direct-acting antiviral therapy cost-effective in UK prisons? a prevention benefit analysis,” Hepatology, vol. 63, no. 6, pp. 1796–1808, Mar. 2016. DOI: 10.1002/hep.28497. [Online]. Available: https://doi.org/10.1002/hep.28497 (cited on p. 232).

[198] E. B. Cunningham, B. Hajarizadeh, N. A. Bretana, et al., “Ongoing incident hepatitis c virus infection among people with a history of injecting drug use in an australian prison setting, 2005-2014: The HITS-p study,” Journal of Viral Hepatitis, vol. 24, no. 9, pp. 733–741, Apr. 2017. DOI: 10.1111/jvh.12701. [Online]. Available: https://doi.org/10.1111/jvh.12701 (cited on p. 232).

[199] D. Crowley, G. Avramovic, W. Cullen, et al., “New hepatitis c virus infection, re-infection and associated risk behaviour in male irish prisoners: A cohort study, 2019,” Archives of Public Health, vol. 79, no. 1, Jun. 2021. DOI: 10.1186/s13690-021-00623-2. [Online]. Available: https://doi.org/10.1186/s13690-021-00623-2 (cited on p. 232).

[200] A. Judd, J. Parry, M. Hickman, et al., “Evaluation of a modified commercial assay in detecting antibody to hepatitis c virus in oral fluids and dried blood spots,” Journal of Medical Virology, vol. 71, no. 1, pp. 49–55, Jul. 2003. DOI: 10.1002/jmv.10463. [Online]. Available: https://doi.org/10.1002/jmv.10463 (cited on p. 232).

[201] A. McAuley, A. Yeung, A. Taylor, S. J. Hutchinson, D. J. Goldberg, and A. Munro, “Emergence of novel psychoactive substance injecting associated with rapid rise in the population prevalence of hepatitis c virus,” International Journal of Drug Policy, vol. 66, pp. 30–37, Apr. 2019. DOI: 10.1016/j.drugpo.2019.01.008. [Online]. Available: https://doi.org/10.1016/j.drugpo.2019.01.008 (cited on p. 241).

[202] N. E. Palmateer, A. Taylor, D. J. Goldberg, et al., “Rapid decline in HCV incidence among people who inject drugs associated with national scale-up

in coverage of a combination of harm reduction interventions," PLoS ONE, vol. 9, no. 8, W. Ho, Ed., e104515, Aug. 2014. DOI: 10.1371/journal.pone.0104515. [Online]. Available: https://doi.org/10.1371/journal.pone.0104515 (cited on p. 241).

[203] J. Søholm, D. K. Holm, B. Mössner, et al., "Incidence, prevalence and risk factors for hepatitis c in danish prisons," PLOS ONE, vol. 14, no. 7, A. J. Santella, Ed., e0220297, Jul. 2019. DOI: 10.1371/journal.pone.0220297. [Online]. Available: https://doi.org/10.1371/journal.pone.0220297 (cited on p. 241).

[204] A. Antuori, V. Montoya, D. Piñeyro, et al., "Characterization of acute HCV infection and transmission networks in people who currently inject drugs in catalonia: Usefulness of dried blood spots," Hepatology, vol. 74, no. 2, pp. 591–606, Aug. 2021. DOI: 10.1002/hep.31757. [Online]. Available: https://doi.org/10.1002/hep.31757 (cited on p. 242).

[205] L. Leon, S. Kasereka, F. Barin, et al., "Age- and time-dependent prevalence and incidence of hepatitis c virus infection in drug users in france, 20042011: Model-based estimation from two national cross-sectional serosurveys," Epidemiology and Infection, vol. 145, no. 5, pp. 895–907, Dec. 2016. DOI: 10.1017/s0950268816002934. [Online]. Available: https://doi.org/10.1017/s0950268816002934 (cited on p. 242).

[206] A. J. Sutton, S. A. McDonald, N. Palmateer, A. Taylor, and S. J. Hutchinson, "Estimating the variability in the risk of infection for hepatitis c in the glasgow injecting drug user population," Epidemiology and Infection, vol. 140, no. 12, pp. 2190–2198, Mar. 2012. DOI: 10.1017/s0950268812000489. [Online]. Available: https://doi.org/10.1017/s0950268812000489 (cited on p. 242).

[207] C. L. I. A. (CLIA), "2021 global marketing report," 2021. [Online]. Available: https://cruising.org/-/media/clia-media/research/2022/2021-1r-clia-001-overview-global.ashx (cited on p. 246).

[208] P. Azimi, Z. Keshavarz, J. G. C. Laurent, B. Stephens, and J. G. Allen, "Mechanistic transmission modeling of COVID-19 on the diamond princess cruise ship demonstrates the importance of aerosol transmission," Proceedings of the National Academy of Sciences, vol. 118, no. 8, Feb. 2021. DOI: 10.1073/pnas.2015482118. [Online]. Available: https://doi.org/10.1073/pnas.2015482118 (cited on p. 247).

[209] G. Correia, L. Rodrigues, M. G. da Silva, and T. Goncalves, “Airborne route and bad use of ventilation systems as non-negligible factors in SARS-CoV-2 transmission,” Medical Hypotheses, vol. 141, p. 109 781, Aug. 2020. DOI: 10.1016/j.mehy.2020.109781. [Online]. Available: https://doi.org/10.1016/j.mehy.2020.109781 (cited on p. 247).

[210] C.-C. Lai, C.-Y. Hsu, H.-H. Jen, A. M.-F. Yen, C.-C. Chan, and H.-H. Chen, “The bayesian susceptible-exposed-infected-recovered model for the outbreak of COVID-19 on the diamond princess cruise ship,” Stochastic Environmental Research and Risk Assessment, Jan. 2021. DOI: 10.1007/s00477-020-01968-w. [Online]. Available: https://doi.org/10.1007/s00477-020-01968-w (cited on p. 247).

[211] L.-S. Huang, L. Li, L. Dunn, and M. He, “Taking account of asymptomatic infections: A modeling study of the COVID-19 outbreak on the diamond princess cruise ship,” PLOS ONE, vol. 16, no. 3, Q. Zeng, Ed., e0248273, Mar. 2021. DOI: 10.1371/journal.pone.0248273. [Online]. Available: https://doi.org/10.1371/journal.pone.0248273 (cited on p. 247).

[212] B. Batista, D. Dickenson, K. Gurski, M. Kebe, and N. Rankin, “Minimizing disease spread on a quarantined cruise ship: A model of COVID-19 with asymptomatic infections,” Mathematical Biosciences, vol. 329, p. 108 442, Nov. 2020. DOI: 10.1016/j.mbs.2020.108442. [Online]. Available: https://doi.org/10.1016/j.mbs.2020.108442 (cited on p. 247).

[213] P. Xu, W. Jia, H. Qian, et al., “Lack of cross-transmission of SARS-CoV-2 between passenger’s cabins on the diamond princess cruise ship,” Building and Environment, vol. 198, p. 107 839, Jul. 2021. DOI: 10.1016/j.buildenv.2021.107839. [Online]. Available: https://doi.org/10.1016/j.buildenv.2021.107839 (cited on p. 247).

[214] F. Liu, X. Li, and G. Zhu, “Using the contact network model and metropolis-hastings sampling to reconstruct the COVID-19 spread on the diamond princess,” Science Bulletin, vol. 65, no. 15, pp. 1297–1305, Aug. 2020. DOI: 10.1016/j.scib.2020.04.043. [Online]. Available: https://doi.org/10.1016/j.scib.2020.04.043 (cited on p. 247).

[215] T. Sekizuka, K. Itokawa, T. Kageyama, et al., “Haplotype networks of SARS-CoV-2 infections in the diamond princess cruise ship outbreak,” Proceedings of the National Academy of Sciences, vol. 117, no. 33, pp. 20 198–20 201, Jul.

2020. DOI: 10.1073/pnas.2006824117. [Online]. Available: https://doi.org/10.1073/pnas.2006824117 (cited on p. 247).

[216] K. Hoshino, T. Maeshiro, N. Nishida, et al., “Transmission dynamics of SARS-CoV-2 on the diamond princess uncovered using viral genome sequence analysis,” Gene, vol. 779, p. 145 496, May 2021. DOI: 10.1016/j.gene.2021.145496. [Online]. Available: https://doi.org/10.1016/j.gene.2021.145496 (cited on p. 247).

[217] E. C. Rosca, C. Heneghan, E. A. Spencer, et al., “Transmission of SARS-CoV-2 associated with cruise ship travel: A systematic review,” Tropical Medicine and Infectious Disease, vol. 7, no. 10, p. 290, Oct. 2022. DOI: 10.3390/tropicalmed7100290. [Online]. Available: https://doi.org/10.3390/tropicalmed7100290 (cited on p. 248).

[218] A.-C. Kordsmeyer, N. Mojtahedzadeh, J. Heidrich, et al., “Systematic review on outbreaks of SARS-CoV-2 on cruise, navy and cargo ships,” International Journal of Environmental Research and Public Health, vol. 18, no. 10, p. 5195, May 2021. DOI: 10.3390/ijerph18105195. [Online]. Available: https://doi.org/10.3390/ijerph18105195 (cited on p. 248).

[219] T. Ganyani, C. Kremer, D. Chen, et al., “Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, march 2020,” Eurosurveillance, vol. 25, no. 17, Apr. 2020. DOI: 10.2807/1560-7917.es.2020.25.17.2000257. [Online]. Available: https://doi.org/10.2807/1560-7917.es.2020.25.17.2000257 (cited on p. 249).

[220] J. C. Heijne, P. Teunis, G. Morroy, et al., “Enhanced hygiene measures and norovirus transmission during an outbreak,” Emerg Infect Dis., vol. 15(1), pp. 24–30, Jan. 2009 (cited on p. 249).

[221] H. L. Kirking, J. Cortes, S. Burrer, et al., “Likely transmission of norovirus on an airplane,” Clinical Infectious Diseases, vol. 50, no. 9, pp. 1216–1221, May 2010, ISSN: 1058-4838. DOI: 10.1086/651597. eprint: https://academic.oup.com/cid/article-pdf/50/9/1216/894913/50-9-1216.pdf. [Online]. Available: https://doi.org/10.1086/651597 (cited on p. 253).

[222] Z. Han, G. N. S. To, S. C. Fu, C. Y.-H. Cha, W. Weng, and Q. Huang, “Effect of human movement on airborne disease transmission in an airplane cabin: Study using numerical modeling and quantitative risk analysis,” BMC Infect Dis, vol. 14, no. 434, Jul. 2014. DOI: 10.1186/1471-2334-14-434. eprint: https://bmcinfectdis.biomedcentral.com/track/pdf/10.1186/1471-

2334-14-434. [Online]. Available: https://doi.org/10.1186/1471-2334-14-434 (cited on p. 253).

[223] V. S. Hertzberg, H. Weiss, L. Elon, W. Si, and S. L. Norris, "Behaviors, movements, and transmission of droplet-mediated respiratory diseases during transcontinental airline flights," Proceedings of the National Academy of Sciences, vol. 115, no. 14, H. Baker, M. Brouillette, S. Campillo, et al., Eds., pp. 3623–3627, 2018, ISSN: 0027-8424. DOI: 10.1073/pnas.1711611115. eprint: https://www.pnas.org/content/115/14/3623.full.pdf. [Online]. Available: https://www.pnas.org/content/115/14/3623 (cited on p. 253).

[224] R. G. Seymour, T. Kypraios, and P. D. O'Neill, "Bayesian nonparametric inference for heterogeneously mixing infectious disease models," Proceedings of the National Academy of Sciences, vol. 119, no. 10, Mar. 2022. DOI: 10.1073/pnas.2118425119. [Online]. Available: https://doi.org/10.1073/pnas.2118425119 (cited on p. 254).

[225] R. G. Seymour, T. Kypraios, P. D. O'Neill, and T. J. Hagenaars, "A bayesian nonparametric analysis of the 2003 outbreak of highly pathogenic avian influenza in the netherlands," Journal of the Royal Statistical Society Series C: Applied Statistics, vol. 70, no. 5, pp. 1323–1343, Nov. 2021. DOI: 10.1111/rssc.12515. [Online]. Available: https://doi.org/10.1111/rssc.12515 (cited on p. 254).