WARWICK
THE UNIVERSITY OF WARWICK

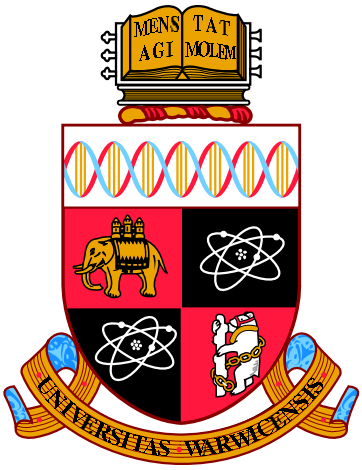
A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:
http://wrap.warwick.ac.uk/101938/

Copyright and reuse:
This thesis is made available online and is protected by original copyright.
Please scroll down to view the document itself.
Please refer to the repository record for this item for information to help you to cite it.
Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications



Mathematical and statistical challenges for the surveillance of gastroenteritis

by

Elizabeth Buckingham-Jeffery

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy in Interdisciplinary Mathematics and Complexity Science

Centre for Complexity Science

February 2018

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

My first, and biggest, thank you must go to my supervisor Thomas House for all his advice, training, and encouragement throughout my time as a student and beyond. His seemingly eternal optimism towards our science perfectly counteracts my pessimism; without this I would have certainly given up on this PhD many years ago.

I am grateful to Professor Valerie Isham of University College London for her support with chapter 2 and to Dr Sebastian Funk and Dr Clare Wenham of the London School of Hygiene and Tropical Medicine for providing data for section 4.3.

I would like to thank all members of the Real-time Syndromic Surveillance Team for their extensive support during this PhD and for making me feel incredibly welcome during my time there. Much of this work would not have been possible without my secondment there. In addition, I thank the data providers who work with the Real-time Syndromic Surveillance Team to supply the data for daily health surveillance activities and thus for the work presented in chapter 3 and section 4.3:

- the clinicians in each emergency department and other staff within each Trust for their help and continued involvement in the Emergency Department Syn-

dromic Surveillance System, L2S2 Ltd for undertaking the daily extraction and transfer of anonymised attendance data from all participating emergency departments, and EMIS Health for facilitating data extraction at the relevant Emergency Department Syndromic Surveillance System sites.

- NHS 111 and NHS Digital for their assistance and support in providing the anonymised call data that underpin the Remote Health Advice Syndromic Surveillance System.

- the University of Nottingham, ClinRisk and the contribution of EMIS and EMIS practices to version 1 of the QSurveillance database for the GP In-hours Syndromic Surveillance System. Additionally, TPP, ResearchOne and the SystmOne GP practices contributing to this surveillance system.

- Advanced Health and Care and the GP out-of-hours and unscheduled care service providers who have kindly agreed to participate in the GP Out-of-hours Syndromic Surveillance System.

I would like to thank Professor Christl Donnelly of Imperial College London and Dr Simon Spencer of the University of Warwick for examining my PhD. My viva was, despite my nerves, a positive experience and their corrections have improved this thesis.

I am thankful to the EPSRC for their complete funding of this PhD, via the Complexity Science Doctoral Training Centre; to the European Society for Mathematical and Theoretical Biology and the Royal Statistical Society for further financial support to attend conferences; to the University of Warwick for a bursary to enable some of my regular travel to Public Health England; and to my parents for additional financial support throughout.

Finally, I must thank my friends, family, and treasured colleagues past and present for their support and encouragement throughout the last four years.

LIST OF ABBREVIATIONS

AIC Akaike information criterion

ARIMA autoregressive integrated moving average

CDC Centers for Disease Control and Prevention

CI confidence interval

df degrees of freedom

E exposed

ED emergency department

EM Euler-Maruyama

GP general practitioner

GPIH GP in hours

GPOOH GP out-of-hours

I infectious

IID1 first study of infectious intestinal disease in the community

IID2 second study of infectious intestinal disease in the community

ILI influenza-like illness

KL Kullback-Leibler

MAE mean absolute error

ODE ordinary differential equation

OU Ornstein-Uhlenbeck

PHE Public Health England

R removed

RAMMIE rising activity, multi-level mixed effects, indicator emphasis

ReSST Real-time Syndromic Surveillance Team

S susceptible

SDE stochastic differential equation

SSE error sum of squares

SSS syndromic surveillance system

Tukey's HSD Tukey's honest significant difference

DECLARATIONS

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy in Interdisciplinary Mathematics and Complexity Science. It has been composed by myself and has not been submitted in any previous application for any degree. In particular, the opinions expressed herein do not necessarily reflect the views of the Real-time Syndromic Surveillance Team or any part of Public Health England.

Parts of this thesis have been published:

1. Chapter 2 has been accepted for publication as:

   E. Buckingham-Jeffery, V. Isham, T. House. (2018) *Gaussian process approximations for fast inference from infectious disease data.* Mathematical Biosciences (in press).

2. Section 3.4.2 from chapter 3 has been published as:

   E. Buckingham-Jeffery, R. Morbey, T. House, A. J. Elliot, S. Harcourt, G. E. Smith. (May 2017) *Correcting for day of the week and public holiday effects: improving a national daily syndromic surveillance service for detecting public health threats.* BMC Public Health, 17:477, DOI: 10.1186/s12889-017-4372-y

ABSTRACT

Gastroenteritis, causing vomiting and diarrhoea, is very common all over the world. Viral causes, such as norovirus and rotavirus, are the most frequent, although some bacteria, parasites and fungi can also lead to gastroenteritis. Many countries operate surveillance systems of diseases, including gastroenteritis or specific gastroenteritis causing pathogens. Typically, statistical methods are used to analyse surveillance data and alert public health authorities of unexpectedly high levels of illness. These methods use historical data to predict the expected value of current data.

In this thesis, we address some of the challenges that remain when analysing gastroenteritis surveillance data, with a particular focus on syndromic surveillance data. We work with both mechanistic and statistical modelling approaches in an attempt to bridge the gap between the statistical methods that are used in practice for syndromic surveillance and mechanistic models that are used to model infectious diseases.

In particular, we address three challenges. In chapter 2 we present a flexible framework for deriving approximations of stochastic mechanistic models of epidemics for fast inference. In chapter 3 we investigate day of the week and public holiday effects in syndromic indicators of gastroenteritis from syndromic surveillance systems operated by Public Health England in order to improve existing surveillance methodologies. In chapter 4 we identify and analyse additional online datasets for gastroenteritis, and in particular norovirus, surveillance.

CHAPTER 1

INTRODUCTION

1.1 Gastroenteritis

A case of gastroenteritis can be defined as *"an individual with three or more loose stools, or any vomiting in 24 hours"* (Majowicz et al. 2008 [1]) but excluding individuals with existing medical conditions known to cause these same symptoms (such as Crohn's disease) or symptoms caused by drugs, alcohol, or pregnancy. Many cases of gastroenteritis are mild and self-limiting. However, serious complications can arise, particularly in the elderly and young children, the most frequent and dangerous of which is dehydration [2].

Most mortalities due to gastroenteritis occur in low-income countries with reports of more than 25% of deaths due to gastroenteritis occurring in Africa and south-east Asia [3]. Gastroenteritis is also a leading cause of morbidity in developed countries, resulting in increased healthcare costs and productivity loss due to time off work and school [4, 5]. A study by Lopman et al. (2004, [6]) in 2002-2003 estimated that healthcare-associated outbreaks of gastroenteritis cost the English NHS £115 million in that year. A further study by Danial et al. (2011, [7]) estimated that gastroenteritis cost just one region of NHS Scotland £1.2 million between 2007 and 2009.

There are many causes of gastroenteritis, including viruses, bacteria, parasites, and fungi. However, viral causes, particularly rotavirus and norovirus, are the most

common [2, 8].

Rotavirus mainly causes gastroenteritis in young children; the most common age of infection is between six months and two years [8, 9]. A systematic review of rotavirus associated mortality (Tate et al. 2012, [10]) concluded that in 2008 rotavirus was responsible for around 40% of diarrhoea deaths and 5% of all deaths in children aged under five globally. Over the last ten years a rotavirus vaccine, recommended for use in babies from six weeks old, has been introduced in many countries across the world [11]. There is now strong evidence that the vaccine has reduced hospital admissions and deaths in many countries including Malawi (Bar-Zeev et al. 2015, [9]), Mexico (Richardson et al. 2010, [12]), the US (Tate et al. 2011, [13]), Australia (Buttery et al. 2011, [14]), and England (Bawa et al. 2015, [15]).

Norovirus is reported to be associated with nearly a fifth of all gastroenteritis cases globally [3]. Across the world it has a high health and economic burden [5]. In England, Harris et al. (2014 [16]) report that norovirus in a hospital setting is associated with, on average, 8,900 days of ward closure each year at a loss of over 15,500 bed-days.

Norovirus is endemic in the UK, and cases can occur throughout the year [17]. However, there is a clear seasonality in the number of norovirus cases with regular, annual winter outbreaks. Frequently, norovirus is associated with outbreaks in closed populations, such as in hospitals and care facilities [18], on cruise ships [19], and due to contaminated food and water [20]. However, we are interested here in the endemic, population-level burden.

1.2 Disease surveillance

The *International Health Regulations* (2005, third edition [21]) define disease surveillance as "the systematic ongoing collection, collation and analysis of data for public health purposes and the timely dissemination of public health information for assessment and public health response as necessary". Effective disease surveillance is used to determine vaccination formulation, design vaccination strategies, and to evaluate vaccination effectiveness; identify outbreaks and inform control strategies; and guide clinical practices and the best allocation of resources [22].

Information about diseases was reportedly collected and analysed as long ago as

Ancient Egypt [23]. However, Choi (2012, [23]) reports that the first public health action as a result of collecting data was in Europe in response to plague; travellers arriving from plague-infected areas were quarantined. Since then, disease surveillance has developed into a complex and global network of surveillance systems using a wide range of data sources [24]. Data sources for surveillance now include the more traditional, such as mortality data and laboratory reports, and those that are more modern, such as statistics on the use of healthcare services and over-the-counter drug sale data [24].

Surveillance data are often analysed using statistical methods to identify unusual activity in both space and time. Typically, within public health authorities, this analysis involves comparing the current data with an expected value obtained from a model fitted to previous data [24]. Frequently used techniques include regression approaches and time series methods, incorporating both space and time characteristics, and multivariate detection methods that make use of more than one dataset [25].

A commonly used, simple, statistical method for analysing surveillance time series data and identifying abnormal levels was developed by Stroup et al. (1989, [26]), and this has been used by the Centers for Disease Control and Prevention for the surveillance of notifiable diseases [27]. In this simple method, data are aggregated into four week blocks (which Stroup et al. call a month for simplicity) and each month is compared to the mean of fifteen baseline values. These are the data from the same month and each of the surrounding months from the past five years. A 95% prediction interval on the mean of the baseline values is computed, assuming normality, and the data are considered unusual if it is outside this interval.

Farrington et al. (1996, [28]) developed a statistical, automated, surveillance algorithm that is used to analyse data from public health laboratories in England and Wales. This method uses a regression to derive both an estimated value for the current week's data and a prediction interval. The threshold is defined as the upper value of this prediction interval. Data over this threshold are flagged as being unusual [28]. The regression is based on 35 baseline values from the same week, and the surrounding three weeks, from the past five years in order to account for seasonality. This has recently been extended to also account for reporting delays in the data [29].

More recently, also within Public Health England, the 'rising activity, multi-level

mixed effects, indicator emphasis' (RAMMIE) method (Morbey et al. 2015, [30]) was developed to detect unusual activity in syndromic surveillance systems. This method was designed to work robustly with multiple syndromic surveillance systems that have a wide range of data volumes, and it is additionally able to prioritise alarms so that there are a manageable number each day. The following overview of the RAMMIE method is based on the methods described by Morbey et al. in [30].

The RAMMIE model is made up of three components: a multi-level mixed effects model giving a prediction of the days activity for each indicator using historical data, baseline and spike thresholds which if exceeded generate alarms, and priority rules so that the user is not overwhelmed with alarms. The model is multi-level as each indicator is modelled at the national level as well as at a regional and local level. A mixed effects model contains both fixed and random effects. In this case, the fixed effects include day of the week and month and random effects include a contribution from the region to models at the local level.

RAMMIE uses a combination of methods to generate and prioritise alarms - an indication that the current activity level is higher than expected and should be investigated further. Historical data and the regression model generate a prediction of the activity level of the current day. From this an upper threshold is obtained. Two types of alarms can be generated. Historical alarms are generated if the current day's activity level is higher than the threshold. In addition, the current activity may be above or below the model generated prediction for an extended period of time (for example, following the introduction of a new vaccine). During this period, an historical alarm would sound continuously or never sound. Therefore, RAMMIE also generates spike alarms to identify recent increases in activity compared to the past week regardless of the comparison with the threshold.

Surveillance algorithms have also been developed to identify anomalous cases in space, or space-time. For example, Besag and Newell (1991, [31]) used significance tests to identify small clusters of cases of a rare disease over a large geographical area. A spatial scan statistic was developed by Kulldorff (1997, [32]) and later extended to a spatial-temporal scan statistic (Kulldorff 2001, [33]). Finally, Bayesian methods for spatial, and spatial-temporal, disease surveillance have been developed such as those used by Spencer et al. (2011, [34]) to identify outbreaks of campylobacteriosis in New Zealand.

Robust computational methods for data transfers, data processing, and analysis

are a key-stone for successful, regular, reliable disease surveillance. The R-package *Surveillance* provides a 'test-bench' for surveillance algorithms and includes the algorithms by Stroup and Farrington described above, among others [27, 35].

1.3 Gastroenteritis surveillance in the UK

Public Health England operate multiple systems that contribute to gastroenteritis surveillance in the UK. This includes analysing data from stool samples submitted for laboratory tests and maintaining four syndromic surveillance systems that include gastroenteritis, diarrhoea or vomiting syndromes (more details of which will be given in chapter 3 of this thesis).

However, there are two aspects to norovirus, and more generally gastroenteritis, surveillance: the surveillance of geographically focussed, relatively small outbreaks and the surveillance of population level endemic illness. Hall et al. (2013, [17]) identified many published reports of small laboratory-confirmed norovirus outbreaks, but fewer publications focussed on assessing the burden of endemic, or community level, norovirus. They comment that one problem contributing to this may be the difficulty in obtaining data from community cases due to "low health-care seeking rates of patients with gastroenteritis" [17]. This is, however, to be expected given that many cases of gastroenteritis are mild and self-limiting. We comment on this difficulty in section 4.3 of chapter 4 of this thesis.

One approach to community surveillance is to carry out community based surveys. There have been two prospective monitoring studies to estimate community gastroenteritis burden in England (*IID1* study, Wheeler et al. 1999, [36] and *IID2* study, O'Brien et al. 2010, [37]). These, whilst providing a valuable snapshot into the number of gastroenteritis cases, do not however provide continual community surveillance.

1.4 Infectious disease modelling

Methods for analysing infectious disease data are more extensive and broad than the statistical surveillance methods used regularly by public health authorities and described in section 1.2. In this section we will briefly describe some examples of

infectious disease modelling. Generally, these analysis methods fall into two groups: statistical (phenomenological) models and mechanistic models. However, the field of infectious disease modelling is very broad and it is beyond the scope of this thesis to provide a full review.

The work of Daniel Bernoulli, in the late 18th century, is often described as one of the earliest examples of a mathematical model of disease [38]. The analysis was developed in order to quantify the benefits of smallpox inoculation, and the work estimated that this intervention increased average life expectancy by three years [39] (although from just 26 years and 7 months to 29 years and 9 months [38]!). However, this model treats incidence of infection as a constant and so does not deal with transmission dynamics.

A transmission model developed by Kermack and McKendrick in 1927 is often described as the most influential contribution to mathematical modelling of infectious diseases [38]. Kermack and McKendrick developed the ordinary differential equations (ODEs) that form the SIR compartmental model of disease transmission. The SIR model has since been extensively developed and extended (see *Mathematical Epidemiology* by Brauer et al. for example [40]) perhaps because these systems of ODEs are now relatively easy to work with and solve computationally with modern computers.

These models are deterministic mechanistic models of the process of disease transmission. However, as stated by Bailey in his paper *'A simple stochastic epidemic'* (1950, [41]): *"a considerable degree of chance enters into the conditions under which fresh infections take place, and it is clear that for a more precise analysis we ought to take these statistical fluctuations into account. In short, we require 'stochastic' models to supplement existing deterministic ones"*. Nevertheless, stochastic models are relatively more difficult to work with than deterministic ones, or at least become quickly too complex for basic analysis techniques. This paper by Bailey is considered one of the earliest contributions to stochastic models of disease [38] and introduces a relatively simple stochastic model which would now be referred to as an SI model (individuals can become infected but there are no further dynamics). Stochastic disease models have also been extensively developed and extended, in particular to include different scales of contact structure and as the corresponding theoretical analysis approaches for stochastic processes have developed (see, for example, [42]).

In addition, developing simultaneously in time with mechanistic modelling, statis-

tical models have been used to analyse many aspects of epidemics. For example, early work considered fitting distributions to small epidemiological datasets, such as fitting the beta-binomial distribution to cases of "*the common cold*" by Griffiths (1973, [43]). More recently, a simple generalised model which had previously been used in demography was fitted to the early growth of a variety of epidemics by Viboud et al (2016, [44]). This flexible statistical model was able to fit successfully to epidemic data demonstrating a range of growth scales, from very slow to near-exponential. Statistical models can be particularly useful when, for example, disease dynamics are not well understood and so a mechanistic model would contain many uncertainties.

One key success of infectious disease modelling, in all forms, is its use to inform policy decisions. This often relates to recommendations of control measures to mitigate outbreaks. These control measures include a broad range of possible interventions whose effectiveness may not be immediately evident without modelling (such as closing a school during a pandemic [45]). Assessing control measures drives the continual development of models of infectious diseases as we try to mitigate their impact on populations across the globe.

Early successful examples of this are for livestock diseases, perhaps due to the larger choice of control measures. For example, one early analysis by Anderson et al. (1996, [46]) gave recommendations on culling protocols to control the BSE epidemic in cattle in Great Britain. A further example is the work of Howard and Donnelly (2000, [47]) who assessed the impact of quick culling of animals on a farm in response to identifying the presence of foot and mouth disease in a herd. This went on to inform further policy-motivated modelling in response to the outbreak in the UK in 2001, such as the effects of vaccinating by Keeling et al. (2003, [48]) and culling by Ferguson et al. (2001, [49]).

1.5 This thesis

This thesis will contain three studies each tackling an outstanding mathematical or statistical challenge in the field of gastroenteritis surveillance. This is in an attempt to start bridging the gap between the statistical techniques used practically for gastroenteritis surveillance and mechanistic modelling approaches.

The analysis methods used daily for syndromic surveillance by public health au-

thorities do not typically include any disease mechanisms. We attempted to fit an SEIRS ODE model to syndromic surveillance data of gastroenteritis (analysis not reported here). However, we found that the parameters of this model were poorly identifiable from these data and that this was not a simple problem. Therefore, we start this thesis by describing our development of approximating methods for stochastic mechanistic models that could be applied to the large datasets used in syndromic surveillance (chapter 2). Next, we realise that mechanistic models can 'idealise' data; when fitting these models it is important to be aware of any regular signals in data that are not directly due to disease levels. Therefore, we investigate and highlight reporting artefacts in daily syndromic surveillance data from a variety of healthcare services (chapter 3). Finally, we suspect that models which combine data sources may be more successful at modelling these processes than models which rely on a single data source. Therefore, we investigate additional online data sources of gastroenteritis (chapter 4).

Most of the work in chapters 3 and 4 was undertaken during a secondment at the Real-time Syndromic Surveillance Team of Public Health England during this PhD. This placement gave the author the opportunity to interact with syndromic surveillance data on a daily basis and experience the challenges of working with these data first-hand. Additionally, this secondment gave limited access to the data used in this work, which are covered by governance and contractual agreements that limit their use for Public Health England surveillance activities only. The data are, therefore, not available for sharing. Additionally, the opinions expressed in this thesis are the author's own do not necessarily reflect the views of the Real-time Syndromic Surveillance Team or any part of Public Health England.

Further introductory details of each chapter will now follow.

1.5.1 Gaussian process approximations for fast inference from epidemic data

The first, and most mathematically technical, work chapter concerns the development of approximation methods for stochastic models of infectious diseases for fast inference with epidemiological data. Many of the methods routinely used for regular surveillance by public health authorities are statistical methods, such as regression models and time series analyses. However, mechanistic models of infectious disease are well established. Surveillance systems obtaining regular, frequent updates need

these data to be analysed quickly. Yet many of the approaches designed to work with non-linear stochastic models of infectious diseases can be computationally intensive and slow. We investigate Gaussian process approximations so that we can exploit useful simple properties of the multivariate normal distribution for fast inference.

1.5.2 Day of the week and public holiday effects in syndromic surveillance data

The second work chapter contains a statistical investigation into the artefacts left by weekends and public holidays in daily surveillance data from a variety of healthcare services. We demonstrate improvements to current surveillance methods that can be made based upon this investigation. Surveillance data of gastroenteritis capture many different trends including, but certainly not limited to, the levels of gastroenteritis circulating in the population. For example, behaviours surrounding seeking healthcare, healthcare availability, and the ability to self-treat all influence the number of gastroenteritis cases recorded by each healthcare system. In this chapter, we investigate two social constructs–weekends and public holidays–that may impact on surveillance systems that analyse daily data. There has not previously been any formal investigation into day of the week and public holiday effects in the syndromic surveillance data used by Public Health England.

1.5.3 Online surveillance of gastroenteritis

The third, and final, work chapter investigates potential additional data sources for gastroenteritis surveillance in the UK. Surveillance systems can be validated by making use of as much data as possible. We investigate data from search engines, from webpage use, and from an online survey designed for influenza surveillance. As gastroenteritis is often self-limiting only some small proportion of cases report to healthcare services. Therefore, not all cases end up in traditional surveillance datasets. This chapter makes use of existing statistical methodologies that have previously been used to verify and compare novel surveillance datasets to investigate additional datasets that have not previously been used for these purposes.

CHAPTER 2

GAUSSIAN PROCESS APPROXIMATIONS FOR FAST INFERENCE FROM EPIDEMIC DATA

In this chapter we describe and investigate Gaussian process approximations of stochastic epidemic models. This chapter has been accepted for publication as:

E. Buckingham-Jeffery, V. Isham, T. House. (2018) *Gaussian process approximations for fast inference from infectious disease data.* Mathematical Biosciences (in press).

This chapter will be structured as follows. First, an introduction with the necessary background material, followed by a numerical comparison of different approximation methods. We then perform inference on synthetic and real data, and finally we analytically compare the approximation methods.

2.1 Introduction and background

Analysing infectious disease data in real time allows us to learn about diseases, to estimate key parameters to understand disease dynamics, and to evaluate interventions. Often we have imperfect, incomplete observations that we would like to analyse quickly so our results can be useful to public health authorities. We can either make use of generic statistical data analysis methods or, as we do here, consider a problem-specific approach with a transmission-dynamic epidemic model.

The epidemic models that are typically used are non-linear. Many methods exist to fit these models to data depending on the model and data involved [50]. Here we are interested in supporting public health surveillance teams. Therefore, we consider the case of daily or weekly prevalence or cumulative incidence data where there is no tractable closed-form likelihood.

Recently, many different methods to deal with intractable likelihoods have been developed [51–56]. However, typically these approaches are computationally intensive and can be difficult to implement which limits their practical use. On the other hand, fitting deterministic ordinary differential equation (ODE) models for epidemics to data, for example by least-squares, is often an ill-posed inverse problem; it is widely accepted that stochastic effects need to be included in epidemic models for inference and prediction to be reliable [57–59].

Therefore, in this chapter we investigate Gaussian process approximations of stochastic epidemic models to speed-up real time analysis of disease data when other methods are too complex. We aim to demonstrate how these approximations can be used for fast inference with outbreak data.

Note that we are not the first to consider Gaussian process approximations for epidemic inference. For example, Ross et al. [60] considered parameter estimation for the SIS model, Fearnhead et al. [61] considered Gaussian approximations based on the linear-noise approximation, and Ball and House [62] considered inference for the SIR epidemic on a network using a Gaussian process approximation.

2.1.1 Aims and objectives

The purpose of this chapter is to show how approximations of stochastic epidemic models can be used with disease data for fast inference to support public health authorities. Our aims are therefore to:

- describe a general structure for Gaussian process approximations of stochastic compartmental models, specifically applied to the SIR model as an example,
- numerically compare different Gaussian process approximations,
- apply the approximations to outbreak data in order to make fast inference of epidemic parameters and unobserved time series,

- derive bounds on the errors of the approximations.

2.1.2 The SIR model

The stochastic SIR model is a simple individual-based stochastic model of infectious disease [63]. Individuals are assumed to be identical and belong to one of three classes: susceptible (S), infectious (I), or removed (R). The number of individuals in each class changes with time as new infections occur (movement of an individual from the susceptible to the infectious class) and as individuals recover or are quarantined or die (movement from the infectious to the removed class). In this case, we will always consider a closed population with no births, deaths, immigration, or other changes to the total population size which remains constant. This compartmental structure has been widely used for infectious disease modelling and has also been extensively expanded, for example by Keeling et al. (2003, [64]) to incorporate vaccination strategies, by Conlan and Grenfell (2007, [65]) to include latent infection and seasonality, and by Riley et al. (2003, [66]) to include coupled metapopulations in different locations.

Pure jump Markov chain

Let $N_S(t)$, $N_I(t)$ and $N_R(t)$ denote the random integer number of people who are susceptible, infectious and removed respectively at time t. The vector $\mathbf{N}(t) = (N_S(t), N_I(t), N_R(t))$ is a continuous-time Markov chain with the following events and rates:

$$\begin{aligned} &(N_S, N_I, N_R) \to (N_S - 1, N_I + 1, N_R) \text{ at rate } \beta \frac{N_S N_I}{N} \ , \\ &(N_S, N_I, N_R) \to (N_S, N_I - 1, N_R + 1) \text{ at rate } \gamma N_I \ . \end{aligned} \tag{2.1}$$

These correspond to an infection event and a removal event respectively, where $N = N_S + N_I + N_R$ is the constant population size and the constants β and γ are the model parameters.

Diffusion approximation

Using the convergence results of Kurtz ([67, 68]), the Markov chain defined by (2.1) is well approximated by the solution $\mathbf{X}(t)$ of the stochastic differential equation (SDE)

$$\mathrm{d}\mathbf{X} = \mathbf{F}(\mathbf{X})\,\mathrm{d}t + \sqrt{\boldsymbol{V}(\mathbf{X})}\,\mathrm{d}\mathbf{W}, \tag{2.2}$$

where

$$\mathbf{X}(t) = \begin{pmatrix} S(t) \\ I(t) \end{pmatrix} \;, \quad \mathbf{F}(\mathbf{X}) = \begin{pmatrix} -(\beta/N)SI \\ (\beta/N)SI - \gamma I \end{pmatrix} \;, \quad \boldsymbol{V}(\mathbf{X}) = \begin{pmatrix} (\beta/N)SI & -(\beta/N)SI \\ -(\beta/N)SI & (\beta/N)SI + \gamma I \end{pmatrix} \;,$$

and $\mathbf{W} = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$ where W_1 and W_2 are two independent Wiener processes. Note that due to the constant population size we can ignore the removed individuals.

The distribution of $\mathbf{X}(\Delta t)|\mathbf{X}(0)$ given by equation (2.2) will not, in general, be Gaussian.

Deterministic approximation

The deterministic approximation of the SIR SDE (equation (2.2)) is given by

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\frac{\beta}{N}si \;, \qquad \frac{\mathrm{d}i}{\mathrm{d}t} = \frac{\beta}{N}si - \gamma i \;, \tag{2.3}$$

where $s(t)$ and $i(t)$ are the numbers of susceptible and infectious individuals respectively at time t that satisfy this deterministic model.

2.1.3 SDE approximations

Most SDEs cannot be solved analytically [69]. Therefore, approximate SDEs which can be solved more easily must be used to obtain approximate solutions of the mean and variances-covariances of the stochastic system.

Linear SDE approximation

According to Archambeau et al. (2007, [70]), the following linear SDE

$$\mathrm{d}\mathbf{x} = (\boldsymbol{A}(t)\mathbf{x} + \mathbf{b}(t))\,\mathrm{d}t + \sqrt{\boldsymbol{U}(t)}\,\mathrm{d}\mathbf{W} \ , \tag{2.4}$$

will have a Gaussian process solution, $\mathrm{GP}(\mathbf{m}(t), \boldsymbol{C}(t))$, with mean and variance-covariance matrix satisfying

$$\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} = \boldsymbol{A}\mathbf{m} + \mathbf{b} \ , \qquad \frac{\mathrm{d}\boldsymbol{C}}{\mathrm{d}t} = \boldsymbol{A}\boldsymbol{C} + \boldsymbol{C}\boldsymbol{A}^{\top} + \boldsymbol{U} \ . \tag{2.5}$$

Therefore, to obtain a Gaussian process approximation to the stochastic SIR model we must choose the time-varying matrices $\boldsymbol{A}(t)$, $\mathbf{b}(t)$, and $\boldsymbol{U}(t)$ in the linear SDE (equation (2.4)) so that it approximates the SIR SDE (equation (2.2)) and hence the full stochastic system.

For the SIR model, based on equation (2.2), there is one obvious choice for the matrix $\boldsymbol{U}(t)$:

$$\boldsymbol{U}(t) = \begin{pmatrix} (\beta/N)s(t)i(t) & -(\beta/N)s(t)i(t) \\ -(\beta/N)s(t)i(t) & (\beta/N)s(t)i(t) + \gamma i(t) \end{pmatrix} \ ,$$

where $s(t)$ and $i(t)$ are as defined in the deterministic approximation (equation (2.3)).

However, there are many choices for the matrix $\boldsymbol{A}(t)$ and the vector $\mathbf{b}(t)$ that will lead to equation (2.4) having a good approximation of the mean behaviour of equation (2.2) but which differ in their approximation of the variance. We define some in the sections to follow and compare their behaviour in the rest of this chapter.

To summarise, let $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ denote the number of susceptible and infectious people respectively in the Gaussian process approximation at time t. These follow a Gaussian process $\mathrm{GP}(\mathbf{m}(t), \boldsymbol{C}(t))$ with mean and variance-covariance matrix

$$\mathbf{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}[\mathcal{X}(t)] \\ \mathbb{E}[\mathcal{Y}(t)] \end{pmatrix} \ , \qquad \boldsymbol{C} = \begin{pmatrix} C_{11} & C_{21} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} \mathrm{var}(\mathcal{X}(t)) & \mathrm{cov}(\mathcal{X}(t), \mathcal{Y}(t)) \\ \mathrm{cov}(\mathcal{X}(t), \mathcal{Y}(t)) & \mathrm{var}(\mathcal{Y}(t)) \end{pmatrix} \ ,$$

whose behaviour is given by the ODEs in equation (2.3).

We will now give some choices of $\boldsymbol{A}$ and $\mathbf{b}$ that correspond to both approximations already discussed in the literature and new approximations not previously named.

Linear-noise / Ornstein-Uhlenbeck approximation

To derive the Ornstein-Uhlenbeck (OU) [71] or linear-noise [72] approximation we start with the SIR SDE (equation (2.2)). We assume both S and I separate into the deterministic part and fluctuations around this, and then we linearise around the deterministic solution. That is, we write

$$S(t) = s(t) + \tilde{S}(t) \ , \qquad I(t) = i(t) + \tilde{I}(t) \ ,$$

where the quantities $\tilde{S}, \tilde{I}$ are assumed to be small. We ignore quadratic terms, $O(\tilde{S}^2, \tilde{S}\tilde{I}, \tilde{I}^2)$, and keeping only the linear terms get

$$\mathrm{d}\begin{pmatrix} S(t) \\ I(t) \end{pmatrix} \approx \begin{pmatrix} -(\beta/N)(s(t)I(t) + S(t)i(t) - s(t)i(t)) \\ (\beta/N)(s(t)I(t) + S(t)i(t) - s(t)i(t)) - \gamma I(t) \end{pmatrix} \mathrm{d}t + \sqrt{\begin{pmatrix} (\beta/N)s(t)i(t) & -(\beta/N)s(t)i(t) \\ -(\beta/N)s(t)i(t) & (\beta/N)s(t)i(t) + \gamma i(t) \end{pmatrix}} \ \mathrm{d}\mathbf{W} \ .$$

This is a special case of the linear SDE Gaussian process approximation described above with

$$\boldsymbol{A}(t) = \begin{pmatrix} -(\beta/N)i(t) & -(\beta/N)s(t) \\ (\beta/N)i(t) & (\beta/N)s(t) - \gamma \end{pmatrix} \ , \qquad \mathbf{b}(t) = \begin{pmatrix} (\beta/N)s(t)i(t) \\ -(\beta/N)s(t)i(t) \end{pmatrix} \ .$$

Other special cases

As well as considering the above two existing approximations in the literature we also introduce other special cases that we do not believe are yet named.

We name the two special cases that we consider in an obvious way:

'A noise': $\boldsymbol{A}(t) = \begin{pmatrix} -(\beta/N)i(t) & 0 \\ 0 & (\beta/N)s(t) - \gamma \end{pmatrix}$, $\mathbf{b}(t) = \mathbf{0}$.

'b noise': $\boldsymbol{A}(t) = \mathbf{0}$, $\mathbf{b}(t) = \begin{pmatrix} -(\beta/N)s(t)i(t) \\ (\beta/N)s(t)i(t) - \gamma i(t) \end{pmatrix}$.

For each of these, and for the approximations from the literature, we obtain a set of five ODEs (from equation (2.5)), which we can solve numerically to give a Gaussian

process approximation of the mean, variances, and covariances of the stochastic SIR model.

For example, for the A noise approximation we get:

$$\begin{aligned}
\frac{\mathrm{d}m_1}{\mathrm{d}t} &= -\frac{\beta}{N} i(t) m_1 \ , \qquad \frac{\mathrm{d}m_2}{\mathrm{d}t} = \left(\frac{\beta}{N} s(t) - \gamma \right) m_2 \ , \\
\frac{\mathrm{d}C_{11}}{\mathrm{d}t} &= \frac{\beta}{N} s(t) i(t) - 2 \frac{\beta}{N} i(t) C_{11} \ , \\
\frac{\mathrm{d}C_{12}}{\mathrm{d}t} &= -\frac{\beta}{N} s(t) i(t) + \left(\frac{\beta}{N} (s(t) - i(t) - \gamma \right) C_{12} \ , \\
\frac{\mathrm{d}C_{22}}{\mathrm{d}t} &= \frac{\beta}{N} s(t) i(t) + \gamma i(t) + 2 \left(\frac{\beta}{N} s(t) - \gamma \right) C_{22} \ ,
\end{aligned}$$

and for b noise:

$$\begin{aligned}
\frac{\mathrm{d}m_1}{\mathrm{d}t} &= -\frac{\beta}{N} s(t) i(t) \ , & \frac{\mathrm{d}m_2}{\mathrm{d}t} &= \frac{\beta}{N} s(t) i(t) - \gamma i(t) \ , \\
\frac{\mathrm{d}C_{11}}{\mathrm{d}t} &= \frac{\beta}{N} s(t) i(t) \ , & \frac{\mathrm{d}C_{12}}{\mathrm{d}t} &= -\frac{\beta}{N} s(t) i(t) \ , \\
\frac{\mathrm{d}C_{22}}{\mathrm{d}t} &= \frac{\beta}{N} s(t) i(t) + \gamma i(t) \ .
\end{aligned}$$

Linear stochastic process approximation

For comparison with approximations from the linear SDE, we consider two other approximations from the literature. In the linear stochastic process (LSP) approximation it is assumed that the susceptible population evolves deterministically but that the infectious population is normally distributed. This was introduced in the context of the SIR model by Isham (1991, [73]). This approach removes the non-linearity from the SIR model (which is only introduced through the product $S(t)I(t)$). The susceptible, rather than the infectious, population is assumed to evolve deterministically as the initial size of infectious population is typically small whereas the initial size of the susceptible population is typically close to the total population size.

The LSP approximation gives the following set of three ODEs for the evolution of the deterministic susceptible population, $s(t)$, and the mean, $\mu_Y = \mathbb{E}[Y(t)]$, and

variance, $\sigma_{YY} = \text{var}(Y(t))$, of the infectious population:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= -\frac{\beta}{N} s\mu_Y \ , \\ \frac{\mathrm{d}\mu_Y}{\mathrm{d}t} &= \frac{\beta}{N} s\mu_Y - \gamma\mu_Y \ , \\ \frac{\mathrm{d}\sigma_{YY}}{\mathrm{d}t} &= \frac{\beta}{N} \left(2s\sigma_{YY} + s\mu_Y\right) - \gamma(2\sigma_{YY} - \mu_Y) \ . \end{aligned} \tag{2.6}$$

Multivariate normal moment closure

For the final comparative approximation we consider the multivariate normal (MVN) moment closure method, developed by Isham (1991, [73]), which can also be used to obtain an approximation of the stochastic SIR model. In this method, we assume that the joint distribution of the susceptible and infectious populations can be approximated by a bivariate normal distribution. We can then derive a set of five ODEs for the mean, variances, and covariances of the susceptible and infectious populations.

To derive the MVN moment closure approximation we begin from the Markov chain of the SIR model (equation (2.1)) and write down the master equation for the probabilities $p_t(N_S, N_I)$ that there are N_S and N_I susceptible and infectious people at time t:

$$\begin{aligned} \frac{\mathrm{d}p_t(N_S, N_I)}{\mathrm{d}t} =& \frac{\beta}{N}(N_S + 1)(N_I - 1)p_t(N_S + 1, N_I - 1) + \gamma(N_I + 1)p_t(N_S, N_I + 1) \\ & - \frac{\beta}{N} N_S N_I p_t(N_S, N_I) - \gamma N_I p_t(N_S, N_I) \ . \end{aligned}$$

Let $X(t)$ and $Y(t)$ denote for the number of susceptible and infectious people respectively in the MVN moment closure approximation of the SIR model at time t. We need to compute each of $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X]$, $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[Y]$, $\frac{\mathrm{d}}{\mathrm{d}t}\text{var}(X)$, $\frac{\mathrm{d}}{\mathrm{d}t}\text{var}(Y)$, and $\frac{\mathrm{d}}{\mathrm{d}t}\text{cov}(X, Y)$. We then apply the MVN moment closure assumption in order to close this set of ODEs.

The MVN moment closure assumption gives us that $X(t)$ and $Y(t)$ follow a Gaussian process $\text{GP}(\boldsymbol{\mu}(t), \boldsymbol{\sigma}(t))$ with mean and variance-covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X(t)] \\ \mathbb{E}[Y(t)] \end{pmatrix} \ , \quad \boldsymbol{\sigma} = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{YY} \end{pmatrix} = \begin{pmatrix} \text{var}(X(t)) & \text{cov}(X(t), Y(t)) \\ \text{cov}(X(t), Y(t)) & \text{var}(Y(t)) \end{pmatrix} \ ,$$

and that all higher cumulates are zero [74].

We will show in detail the method for deriving $\frac{\mathrm{d}}{\mathrm{d}t}\mu_X$ and $\frac{\mathrm{d}}{\mathrm{d}t}\sigma_{XX}$. The rest are very similar. First consider $\frac{\mathrm{d}}{\mathrm{d}t}\mu_X$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mu_X =& \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X] = \frac{\mathrm{d}}{\mathrm{d}t}\sum_{\substack{X=0,\\Y=0}}^{N,N} X p_t(X,Y) = \sum_{\substack{X=0,\\Y=0}}^{N,N} X\frac{\mathrm{d}}{\mathrm{d}t}p_t(X,Y)\\ &= \sum_{\substack{X=0,\\Y=0}}^{N,N} X\frac{\beta}{N}(X+1)(Y-1)p_t(X,Y) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X\frac{\beta}{N}XYp_t(X,Y)\\ &+ \sum_{\substack{X=0,\\Y=0}}^{N,N} X\gamma(Y+1)p_t(X,Y+1) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X\gamma Y p_t(X,Y)\ .\end{aligned}$$

Considering just the terms from the infection event:

$$\begin{aligned}&\sum_{\substack{X=0,\\Y=0}}^{N,N} X\frac{\beta}{N}(X+1)(Y-1)p_t(X+1,Y-1) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X\frac{\beta}{N}XYp_t(X,Y)\\ =& \sum_{\substack{X=1,\\Y=-1}}^{N+1,N-1} (X-1)\frac{\beta}{N}XYp_t(X,Y) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X\frac{\beta}{N}XYp_t(X,Y)\\ =& \sum_{\substack{X=0,\\Y=0}}^{N,N} (X-1)\frac{\beta}{N}XYp_t(X,Y) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X\frac{\beta}{N}XYp_t(X,Y)\\ =& -\frac{\beta}{N}\sum_{\substack{X=0,\\Y=0}}^{N,N} XYp_t(X,Y)\\ =& -\frac{\beta}{N}\mathbb{E}[XY]\ ,\end{aligned}$$

and from the removal event:

$$\begin{aligned}
&\sum_{\substack{X=0,\\Y=0}}^{N,N} X\gamma(Y+1)p_t(X,Y+1) - \sum_{\substack{X=0,\\Y=0}}^{N,N} \gamma XY p_t(X,Y) \\
=&\sum_{\substack{X=0,\\Y=1}}^{N,N+1} \gamma XY p_t(X,Y) - \sum_{\substack{X=0,\\Y=0}}^{N,N} \gamma XY p_t(X,Y) \\
=&\sum_{\substack{X=0,\\Y=0}}^{N,N} \gamma XY p_t(X,Y) - \sum_{\substack{X=0,\\Y=0}}^{N,N} \gamma XY p_t(X,Y) \\
=&0 \ .
\end{aligned}$$

Putting this together gives $\frac{\mathrm{d}}{\mathrm{d}t}\mu_X = -\frac{\beta}{N}\mathbb{E}[XY] = -\frac{\beta}{N}(\mathbb{E}[X]\mathbb{E}[Y] + \mathrm{cov}(X,Y)) = -\frac{\beta}{N}(\mu_X\mu_Y + \sigma_{XY})$.

For $\frac{\mathrm{d}}{\mathrm{d}t}\sigma_{XX}$ we need to compute $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X^2]$:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X^2] = \frac{\mathrm{d}}{\mathrm{d}t}\sum_{\substack{X=0,\\Y=0}}^{N,N} X^2 p_t(X,Y) = \sum_{\substack{X=0,\\Y=0}}^{N,N} X^2 \frac{\mathrm{d}}{\mathrm{d}t} p_t(X,Y) \\
=&\sum_{\substack{X=0,\\Y=0}}^{N,N} X^2 \frac{\beta}{N}(X+1)(Y-1)p_t(X+1,Y-1) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X^2 \frac{\beta}{N} XY p_t(X,Y) \\
+&\sum_{\substack{X=0,\\Y=0}}^{N,N} X^2\gamma(Y+1)p_t(X,Y+1) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X^2\gamma Y p_t(X,Y) \ .
\end{aligned}$$

Again, we begin by considering just the terms from the infection event separately:

$$\begin{aligned}
&\sum_{\substack{X=0,\\Y=0}}^{N,N} X^2\frac{\beta}{N}(X+1)(Y-1)p_t(X+1,Y-1) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X^2\frac{\beta}{N}XYp_t(X,Y)\\
=&\sum_{\substack{X=1,\\Y=-1}}^{N+1,N-1} (X-1)^2\frac{\beta}{N}XYp_t(X,Y) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X^2\frac{\beta}{N}XYp_t(X,Y)\\
=&\sum_{\substack{X=0,\\Y=0}}^{N+1,N-1} (X^2-2X+1)\frac{\beta}{N}XYp_t(X,Y) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X^2\frac{\beta}{N}XYp_t(X,Y)\\
=&-2\frac{\beta}{N}\sum_{\substack{X=0,\\Y=0}}^{N+1,N-1} XXYp_t(X,Y) + \frac{\beta}{N}\sum_{\substack{X=0,\\Y=0}}^{N+1,N-1} XYp_t(X,Y)\\
=&-2\frac{\beta}{N}\mathbb{E}[X^2Y] + \frac{\beta}{N}\mathbb{E}[XY]\ ,
\end{aligned}$$

and secondly the removal event terms:

$$\begin{aligned}
&\sum_{\substack{X=0,\\Y=0}}^{N,N} X^2\gamma(Y+1)p_t(X,Y+1) - \sum_{\substack{X=0,\\Y=0}}^{N,N} X^2\gamma Yp_t(X,Y)\\
=&\sum_{\substack{X=0,\\Y=1}}^{N,N+1} \gamma X^2Yp_t(X,Y) - \sum_{\substack{X=0,\\Y=0}}^{N,N} \gamma X^2Yp_t(X,Y)\\
=&\sum_{\substack{X=0,\\Y=0}}^{N,N} \gamma X^2Yp_t(X,Y) - \sum_{\substack{X=0,\\Y=0}}^{N,N} \gamma X^2Yp_t(X,Y)\\
=&0\ .
\end{aligned}$$

This gives $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X^2] = -2\frac{\beta}{N}\mathbb{E}[X^2Y] + \frac{\beta}{N}\mathbb{E}[XY]$.

As the third order cumulant is equal to the third central moment, the MVN moment closure assumption gives us that

$$\begin{aligned}
&\mathbb{E}[(X-\mathbb{E}[X])^2(Y-\mathbb{E}[Y])] = 0\\
\Rightarrow&\mathbb{E}[X^2Y] = \mathbb{E}[X^2]\mathbb{E}[Y] + 2\mathbb{E}[XY]\mathbb{E}[X] - 2\mathbb{E}[X]^2\mathbb{E}[Y].
\end{aligned}$$

Putting this together with the chain rule gives

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\sigma_{XX} &= \frac{\mathrm{d}}{\mathrm{d}t}(\mathbb{E}[X^2] - \mathbb{E}[X]^2) \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X^2] - 2\mathbb{E}[X]\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X] \\
&= -2\frac{\beta}{N}\mathbb{E}[X^2Y] + \frac{\beta}{N}\mathbb{E}[XY] - 2\mathbb{E}[X](-\frac{\beta}{N}(\mathbb{E}[X]\mathbb{E}[Y] + \mathrm{cov}(X,Y))) \\
&= \frac{\beta}{N}\left(\mathbb{E}[X]\mathbb{E}[Y] + \mathrm{cov}(X,Y) - 2\mathbb{E}[X]\mathrm{cov}(X,Y) - 2\mathbb{E}[Y]\mathrm{var}(S)\right) \\
&= \frac{\beta}{N}(\mu_X\mu_Y + \sigma_{XY} - 2\mu_X\sigma_{XY} - 2\mu_Y\sigma_{XX}).
\end{aligned}$$

Overall, we get that $X(t)$ and $Y(t)$ follow a Gaussian process $\mathrm{GP}(\boldsymbol{\mu}(t), \boldsymbol{\sigma}(t))$ with mean and variance-covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X(t)] \\ \mathbb{E}[Y(t)] \end{pmatrix}, \qquad \boldsymbol{\sigma} = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{YY} \end{pmatrix} = \begin{pmatrix} \mathrm{var}(X(t)) & \mathrm{cov}(X(t),Y(t)) \\ \mathrm{cov}(X(t),Y(t)) & \mathrm{var}(Y(t)) \end{pmatrix},$$

that obey equations

$$\begin{aligned}
\frac{\mathrm{d}\mu_X}{\mathrm{d}t} &= -\frac{\beta}{N}(\mu_X\mu_Y + \sigma_{XY}), \\
\frac{\mathrm{d}\mu_Y}{\mathrm{d}t} &= \frac{\beta}{N}(\mu_X\mu_Y + \sigma_{XY}) - \gamma\mu_Y, \\
\frac{\mathrm{d}\sigma_{XX}}{\mathrm{d}t} &= \frac{\beta}{N}(\mu_X\mu_Y + \sigma_{XY} - 2\mu_X\sigma_{XY} - 2\mu_Y\sigma_{XX}), \\
\frac{\mathrm{d}\sigma_{XY}}{\mathrm{d}t} &= \frac{\beta}{N}(\mu_X(\sigma_{XY} - \sigma_{YY}) + \mu_Y(\sigma_{XX} - \sigma_{XY}) - \mu_X\mu_Y - \sigma_{XY}) - \gamma\sigma_{XY}, \\
\frac{\mathrm{d}\sigma_{YY}}{\mathrm{d}t} &= \frac{\beta}{N}(2\mu_X\sigma_{YY} + 2\mu_Y\sigma_{XY} + \mu_X\mu_Y + \sigma_{XY}) - \gamma(2\sigma_{YY} - \mu_Y),
\end{aligned}$$

which we can solve numerically.

Note that there are other moment closures, such as log-normal [75] and beta-binomial [76], which we do not currently consider.

2.2 Numerical comparisons of approximation methods

In order to assess the performance of each of the Gaussian process approximations defined above we numerically compare each approximation to simulations of the stochastic process for a range of epidemiological model parameter values and pop-

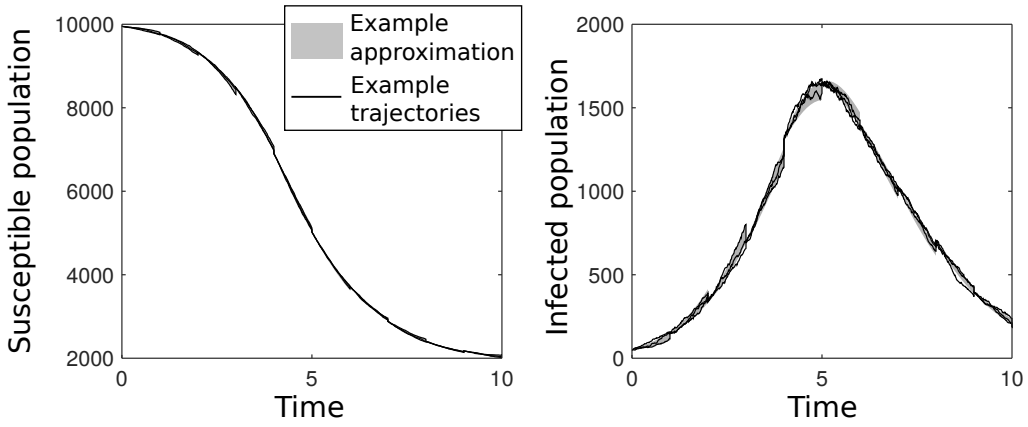


Figure 2.1: A typical example of stochastic trajectories and one Gaussian process approximation. This example was generated with parameters $\beta = 2$, $\gamma = 1$, and $N = 1 \times 10^4$ and the MVN moment closure approximation. The shaded approximation region corresponds to the mean plus/minus one standard deviation.

ulation sizes N, with 50 initial infectious individuals.

Due to our interest in regularly spaced data, we simulate trajectories of both the susceptible and infectious populations of the stochastic SIR model at regular time intervals between known data points using the tau-leap algorithm [77]. We do this for each set of model parameters of interest. Figure 2.1 shows an example of these trajectories for a specific set of parameter values. Simulating 10^4 trajectories gives a distribution at each time point to which we compare each approximating Gaussian distribution.

For each approximation we computed the mean and variance of the size of the susceptible and infectious populations forward from the current data point until the next. The mean of the approximation was then reset to the data point, and the variances to zero. Figure 2.1 shows an example of this for one specific set of parameter values and one specific approximation.

2.2.1 Kullback-Leibler divergence

We compared the approximations numerically to the stochastic simulations using the Kullback-Leibler (KL) divergence, a measure of difference between two probability distributions [78]. The KL divergence can only be used to compare two distributions with the same support. Therefore, we discretised the Gaussian distributions of the

approximations in order to compare with the discrete numerical distribution given by the stochastic simulations. It is appropriate to discretise as we wish to work on a discrete time-scale for this approximate inference.

For discrete probability distributions P and Q the KL divergence is defined as

$$D_{\mathrm{KL}}(P||Q) = \sum_i P(i)(\ln P(i) - \ln Q(i)) \ .$$

A better approximation will result in a smaller KL divergence [78]. The KL divergence was computed each time data were obtained and before the simulations and approximations were reset to the data value.

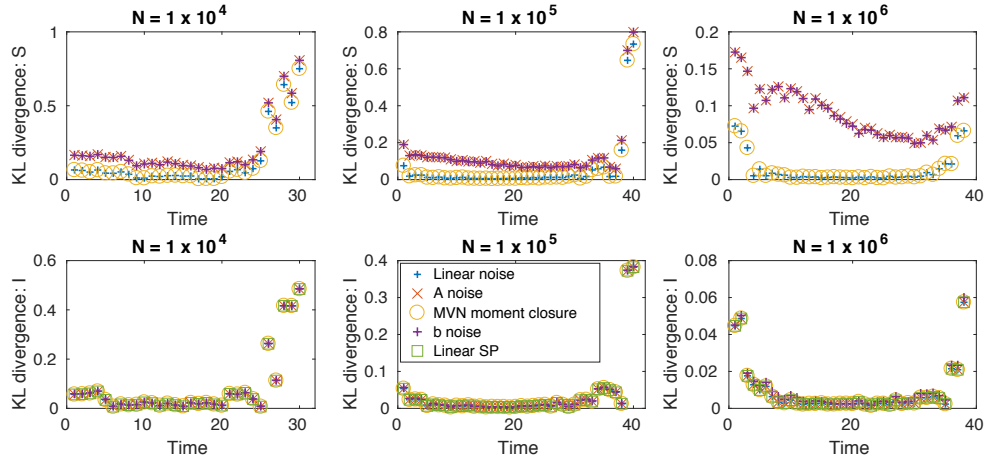
Note that we could not compute the KL divergence for a comparison of the LSP with the stochastic simulations for the size of the susceptible population as in the LSP the susceptible population evolves deterministically. Additionally, on occasion the KL divergence could not be computed for other approximations because one distribution took a value very close to zero. However, as displayed in figures 2.2 and 2.3, this does not occur very often.

2.2.2 Results

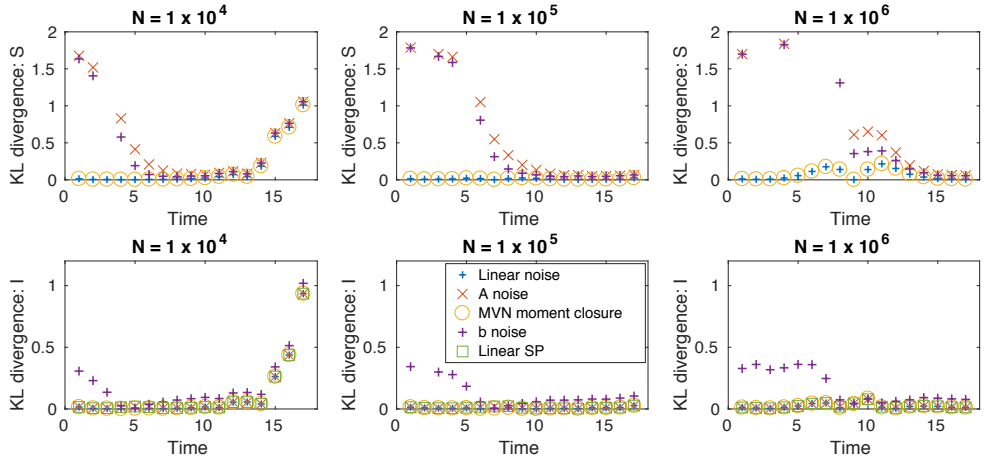
Figure 2.2 demonstrates these comparisons for three examples of epidemiological model rate parameters that we have chosen to cover a range of R_0 values. The MVN moment closure and LN approximations consistently have the smallest KL divergence in both the size of the susceptible population and the size of the infectious population (figure 2.2). Additionally, and in particular for larger population sizes, the A noise and LSP approximations also approximate well the size of the infectious population. The b noise approximation does not approximate the size of the infectious class as well, in particular at the start and end of the epidemic.

For approximating the size of the susceptible population, the A noise and b noise approximations perform adequately but not as well as the other approximations particularly at the start and end of the epidemic.

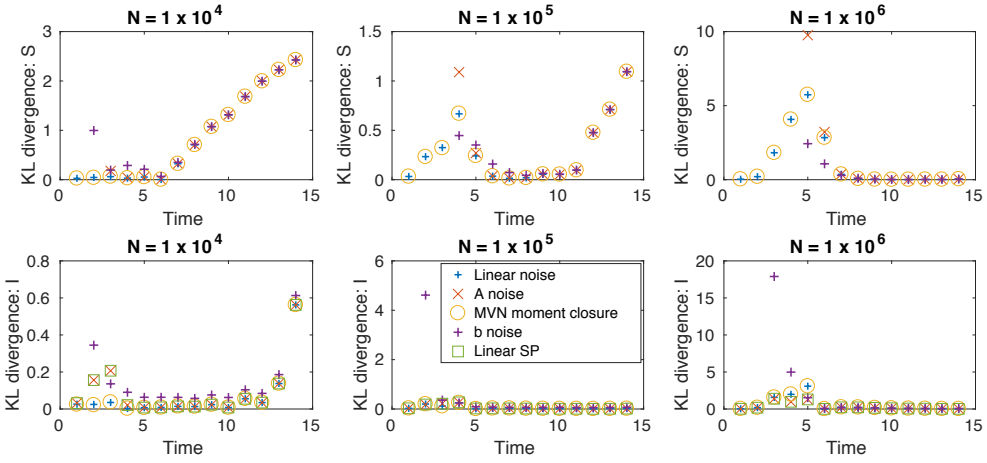
We performed the same analysis over a longer time step where we did not obtain any new data throughout the epidemic (figure 2.3). We saw similar results; the MVN moment closure and OU approximations are best, with the A noise and LSP approximations also good approximations for the infectious population. However,



(a) $\beta = 0.6$, $\gamma = 0.5$ $(R_0 = 1.2)$



(b) $\beta = 2$, $\gamma = 1$ $(R_0 = 2)$



(c) $\beta = 3$, $\gamma = 0.5$ $(R_0 = 6)$

Figure 2.2: Numerical comparisons of the approximation schemes with stochastic simulations of the SIR model using the KL divergence when new data are obtained each day. Within each subplot (a-c) different rate constant parameter values were used to generate stochastic simulations for comparison to each Gaussian approximation. The size of the susceptible population is compared on the top line and the size of the infectious population on the bottom line, for three population sizes (increasing from left to right).

the A noise approximation became a much less good approximation of the susceptible population over this longer time step.

The ODEs that define the approximations were solved numerically. The b noise approximation has the simplest set of ODEs and so is fastest to solve (figure 2.4). The A noise and LSP approximations are slower and, finally, the MVN moment closure and OU approximations take longest (figure 2.4).

2.2.3 Conclusions

In conclusion, these numerical comparisons show that the A noise Gaussian process approximation can perform comparably to the MVN moment closure and OU approaches, in particular for large population sizes and for the infectious population size, while being computationally faster. We expect this computational advantage to become much more pronounced for more complex compartmental models.

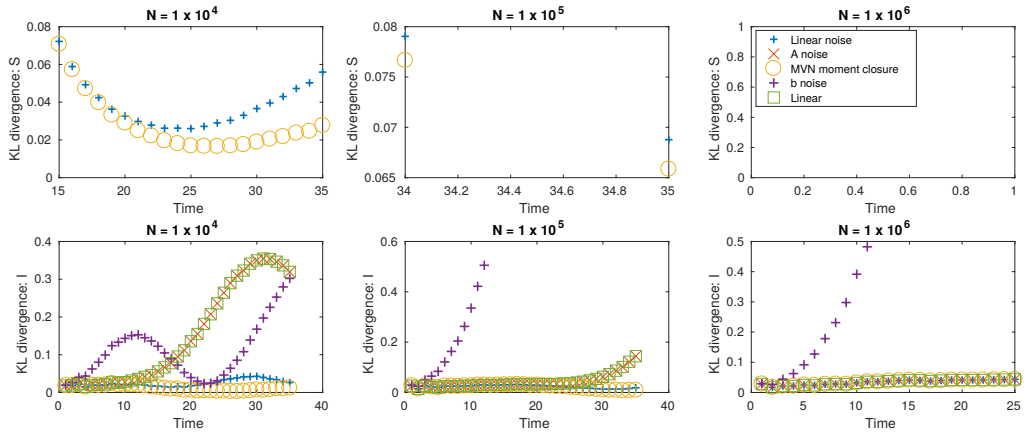
In addition, we note the ease with which the A noise approximation can be derived (in essence, just written down) in comparison to the MVN moment closure and OU approaches and that, again, this will become more pronounced for more complex compartmental models (this is demonstrated further in section 2.3.2 where we apply these approximations to the SEIR model).

This near-comparable performance, along with the mathematical and computational advantages, means that we consider the A noise approximation with the MVN moment closure and OU approximations for inference with data in the next section.

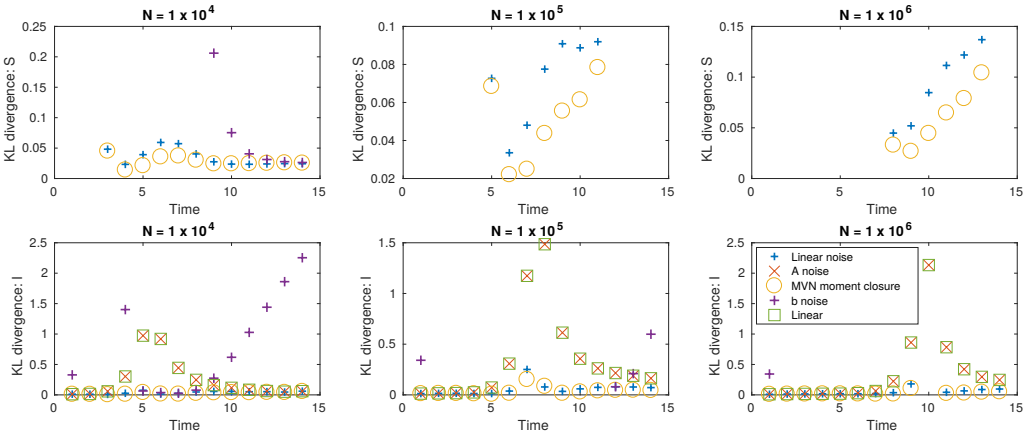
2.3 Inference

We will now demonstrate how these approximation methods can be used for fast inference of partially observed epidemics. We will just use the approximation methods that gave the best results in the previous section; namely, the MVN moment closure, OU, and A noise approximations.

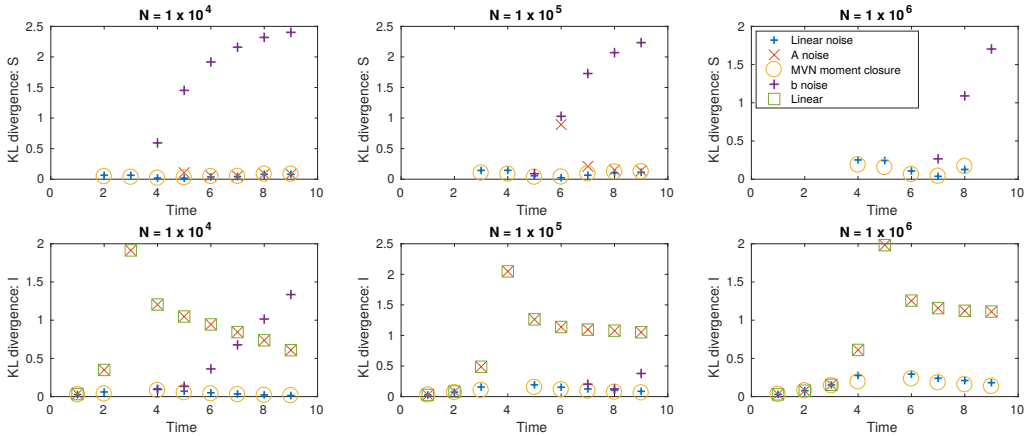
This section will proceed as follows. We initially apply the Gaussian process approximations to synthetic data from the SIR model to demonstrate that the size of the susceptible population can be recovered from weekly prevalence measurements. Secondly, we consider real data from a norovirus outbreak with the SEIR



(a) $\beta = 0.6$, $\gamma = 0.5$ ($R_0 = 1.2$)



(b) $\beta = 2$, $\gamma = 1$ ($R_0 = 2$)



(c) $\beta = 3$, $\gamma = 0.5$ ($R_0 = 6$)

Figure 2.3: Numerical comparisons of the approximation schemes with stochastic simulations of the SIR model using the KL divergence where no new data are obtained throughout the epidemic. Within each subplot (a-c) the size of the susceptible population is compared on the top line and the size of the infectious population on the bottom line, for three population sizes (increasing from left to right).

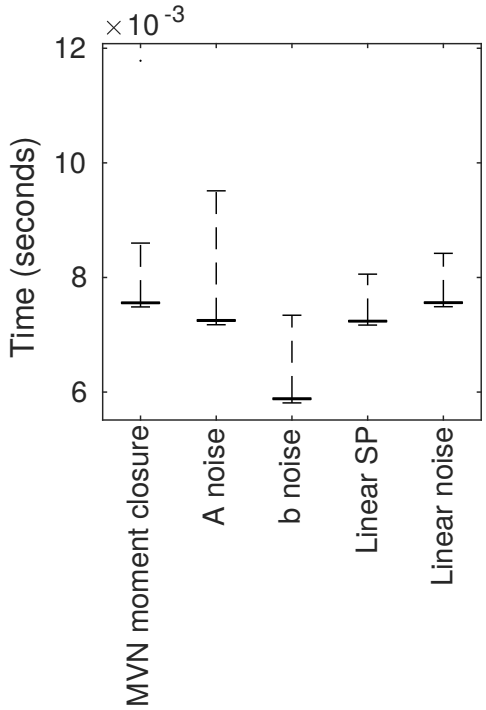


Figure 2.4: Running times for the sets of differential equations for each of the approximation methods with $R_0 = 2$ and $N = 1 \times 10^6$. The ODEs were solved many times for each approximation. The box denotes the median, lower quartile, and upper quartile of the running times. Whiskers extend to the maximum and minimum. (Note that the boxes display as simply thick lines because the interquartile ranges are all very small in this case.)

(susceptible-exposed-infectious-removed) model to demonstrate that it is straightforward to use these approximations with real data and more complex models.

2.3.1 Simulated prevalence from the SIR model

Consider a disease that is well approximated by the SIR model with constant transmission rate β and constant removal rate γ. Suppose we have data of the form of a set of times $\{t_i\}_{i=0}^n$ together with associated measurements of the number of infecteds $\{y_i\}_{i=0}^n$.

We have Gaussian process approximations of the SIR model such that given susceptible and infectious populations of size x_0 and y_0 respectively at the start of a time interval of length Δt, at the end of that time interval the mean and variance-covariance matrix are $\boldsymbol{\mu}(\Delta t; x_0, y_0, \beta, \gamma)$ and $\boldsymbol{\Sigma}(\Delta t; x_0, y_0, \beta, \gamma)$ respectively.

If we also had measurements of the susceptible population, $\{x_i\}_{i=0}^n$, then we could write the likelihood function for the parameters of the approximating model given the data as

$$L(\beta, \gamma; \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} \mathcal{N}\left((x_i, y_i); \boldsymbol{\mu}(t_i - t_{i-1}; x_{i-1}, y_{i-1}, \beta, \gamma), \boldsymbol{\Sigma}(t_i - t_{i-1}; x_{i-1}, y_{i-1}, \beta, \gamma)\right) \ .$$

In practice, however, the data on the susceptible population are not readily available. Instead, we can impute this information using the marginal, and marginal conditional, distributions of the MVN distribution. These can be explicitly computed as follows [79].

For random vector (x, y) with MVN distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, y has marginal probability density function

$$f(a) = \mathcal{N}(a; \mu_2, \Sigma_{22}) \ , \tag{2.7}$$

and, conditional on an observation $y = a$, the random variable x has marginal conditional probability density function

$$f(x; y = a) = \mathcal{N}(x; \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_i - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \ . \tag{2.8}$$

We can use these rules to build up a likelihood from the product of terms such as equation (2.7). Also, at each observation point we can use equation (2.8) to

update the mean vector and covariance matrix for the Gaussian process approximation.

Synthetic data were obtained from one run of the stochastic SIR model using parameter values $\beta = 2$, $\gamma = 1$, and $N = 1 \times 10^4$ with one initial infected.

Using each of the MVN moment closure, OU, and A noise approximations, we were reliably able to recover the epidemiological model parameters (figure 2.5). Using maximum likelihood optimisation, the three approximation methods all gave the same parameter estimates (estimates $\hat{\beta} = 2.04$ and $\hat{\gamma} = 1.01$). Additionally, they gave similar results for the inference of the susceptible population from regular data on the number of infecteds.

2.3.2 Cumulative incidence of a real norovirus outbreak with the SEIR model

In section 2.3.1 we considered the case the data available to us are the number of infecteds at regular time points. An alternative, and common, situation is when only illness onset times and not removal times are available. For an SIR model this corresponds to having measurements of the cumulative incidence, which is equal to $N - S(t)$.

In this section we will be using an SEIR (susceptible-exposed-infectious-removed) modelling framework. Therefore, data of this type are not necessarily measurements of $N - S(t)$. However, we assume in this case that they are. We make the somewhat conservative assumption that newly diagnosed individuals are no longer susceptible but could potentially be in any of the E, I, or R states so that our data are effectively values of $N - S(t)$. A different approach could easily be taken within this inferential framework; the assumption we have made is simply intended to demonstrate that our inference methods are working effectively.

We consider real data from an outbreak of norovirus on a cruise ship visiting the British Isles as reported by Vivancos et al. (2010, [80]). This report gives us data on the number of new reported norovirus cases per day during this outbreak in a small, closed population of 1714 individuals. A single norovirus outbreak in a closed population is commonly assumed to follow the SEIR framework [81–84]. After infection individuals enter a latent state, E, that they leave at rate ω on becoming infectious. The stochastic differential equation for the SEIR framework, equivalent

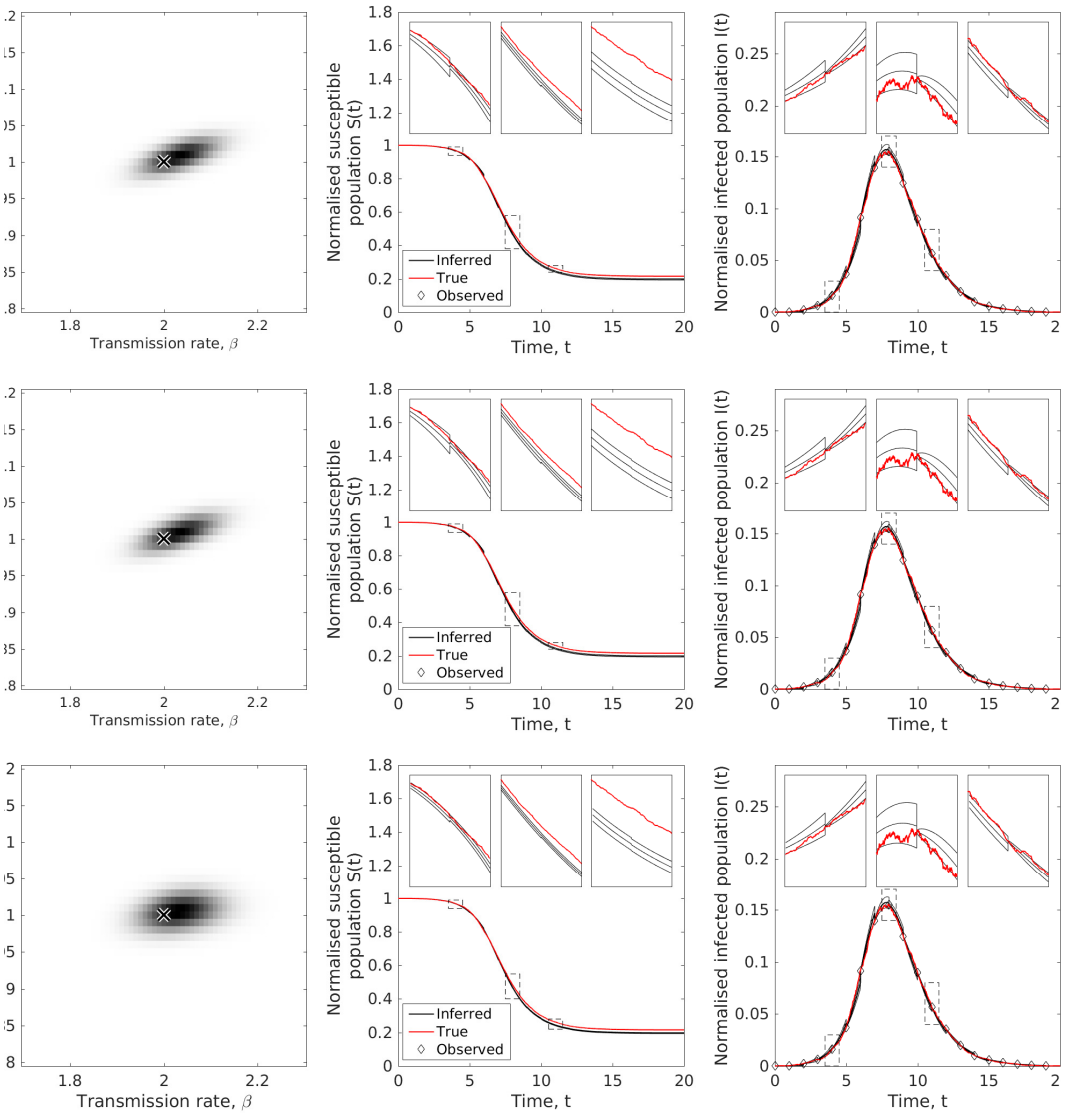


Figure 2.5: Inference of the susceptible population using the MVN moment closure (top), OU (middle), and A noise (bottom) Gaussian process approximations. Left: Likelihood (density plot) is concentrated around the true value (cross). Centre: Data on the number of infecteds allows for good reconstruction of the unobserved susceptibility over time. Shown are the synthetic data ('True', red), the mean of the approximation scheme using the inferred parameter values ('Inferred', black), and the mean plus/minus one standard deviation (black) in both the main figures and insets. Dashed rectangles on the main figures show the locations of the insets from left to right. Right: Data ('Observed', diamonds) allow for good reconstruction of the number of infecteds over time. The red lines, black lines, and dashed rectangles are as before.

to equation (2.2) for the SIR framework, is given by $\mathrm{d}\mathbf{X} = \mathbf{F}(\mathbf{X})\,\mathrm{d}t + \sqrt{\boldsymbol{V}(\mathbf{X})}\,\mathrm{d}\mathbf{W}$ with

$$\begin{aligned} \mathbf{X}(t) &= \begin{pmatrix} S(t) \\ E(t) \\ I(t) \end{pmatrix}, \quad \mathbf{F}(\mathbf{X}) = \begin{pmatrix} -\beta SI/N \\ \beta SI/N - \omega E \\ \omega E - \gamma I \end{pmatrix}, \\ \boldsymbol{V}(\mathbf{X}) &= \begin{pmatrix} \beta SI/N & -\beta SI/N & 0 \\ -\beta SI/N & \beta SI/N + \omega E & -\omega E \\ 0 & -\omega E & \omega E + \gamma I \end{pmatrix}. \end{aligned} \tag{2.9}$$

The deterministic approximation of the stochastic SEIR model is given by the ODEs

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\frac{\beta}{N} si\,, \qquad \frac{\mathrm{d}e}{\mathrm{d}t} = \frac{\beta}{N} si - \omega e \qquad \frac{\mathrm{d}i}{\mathrm{d}t} = \omega e - \gamma i\,, \tag{2.10}$$

where, as before, $s(t)$, $e(t)$, and $i(t)$ are the numbers of susceptible, exposed, and infected individuals respectively at time t given by the deterministic model.

As before, we can impute the unobserved time series $E(t)$ and $I(t)$ using the conditional rules of the MVN distribution and use the marginal rules to perform maximum likelihood estimation on the parameter values β, γ, and $S(0)$. Note that we fit $S(0)$ instead of taking it to be $N - 1$ because we do not know the infection history and contact structure of the population. For example, some of the population may have been previously recently exposed to norovirus and therefore not currently susceptible. Some groups of passengers may not mix due to cabin location, excursion choice and control measures in place [80]. Finally, evidence indicates that there may be some level of genetic immunity to norovirus which may protect some passengers [85].

Additionally, some care must be taken because ω is poorly identifiable from this cumulative incidence data, and our attempts to fit it alongside the other three parameters produced unrealistically large estimates motivating us to fix this parameter from other data. We found that the literature gives the latent, or incubation, period of norovirus to be between 0.5 and 2 days. For example, an SEIR model fitted to an outbreak in a long-term care facility estimated the latent period of norovirus as 1.3 days [82]. A systematic review of the incubation period of norovirus genogroups I and II gives it as 1.2 days (95% confidence interval 1.1–1.2) [86]. The CDC report that the incubation period of norovirus is between 0.5 and 2 days [87]. Finally, a large dataset of norovirus outbreaks showed the incubation period to have a mean and median of 1.4 (95% confidence interval 1.3–1.4) days. Since ω is the reciprocal of the latent period, we therefore chose $\omega = 2$ days^{-1} as the largest value consistent with existing knowledge about norovirus.

A noise with SEIR

To use the A noise Gaussian process approximation with the SEIR model we simply need again to choose the time-varying matrices, $\boldsymbol{A}$, $\mathbf{b}$, and $\boldsymbol{U}$ in equation (2.4) so that it approximates the SEIR SDE (equation (2.9)).

For the A noise approximation we can see that one choice will simply be:

$$\mathbf{A}(t) = \begin{pmatrix} 0 & 0 & -\beta s(t)/N \\ 0 & -\omega & \beta s(t)/N \\ 0 & \omega & -\gamma \end{pmatrix} , \quad \mathbf{b}(t) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} ,$$
$$\boldsymbol{U}(t) = \begin{pmatrix} \beta s(t)i(t)/N & -\beta s(t)i(t)/N & 0 \\ -\beta s(t)i(t)/N & \beta s(t)i(t)/N + \omega e(t) & -\omega e(t) \\ 0 & -\omega e(t) & \omega e(t) + \gamma i(t) \end{pmatrix} .$$

This gives the following ODEs for the approximation of the means, variances, and covariances:

$$\begin{aligned}
&\frac{\mathrm{d}m_1}{\mathrm{d}t} = -\frac{\beta}{N} s m_3 , \quad \frac{\mathrm{d}m_2}{\mathrm{d}t} = \frac{\beta}{N} s m_3 - \omega m_2 , \quad \frac{\mathrm{d}m_3}{\mathrm{d}t} = \omega m_2 - \gamma m_3 , \\
&\frac{\mathrm{d}C_{11}}{\mathrm{d}t} = -2\frac{\beta}{N} s C_{13} + \frac{\beta}{N} s i , \\
&\frac{\mathrm{d}C_{12}}{\mathrm{d}t} = \frac{\beta}{N} s (C_{13} - C_{23} - i) - \omega C_{12} , \\
&\frac{\mathrm{d}C_{13}}{\mathrm{d}t} = -\frac{\beta}{N} s C_{33} + \omega C_{12} - \gamma C_{13} , \\
&\frac{\mathrm{d}C_{22}}{\mathrm{d}t} = \frac{\beta}{N} s (2C_{23} + i) + \omega (e - 2C_{22}) , \\
&\frac{\mathrm{d}C_{23}}{\mathrm{d}t} = \omega (C_{22} - C_{23} - e) - \gamma C_{23} + \frac{\beta}{N} s C_{33} , \\
&\frac{\mathrm{d}C_{33}}{\mathrm{d}t} = \omega (2C_{23} + e) + \gamma (i - 2C_{33}) .
\end{aligned}$$

Note that writing down $\boldsymbol{A}$, $\mathbf{b}$, and $\boldsymbol{U}$ in this case is very straightforward, and we suspect that this will continue to be the case as we consider the A noise approximation with more complex compartmental models. As the matrices can simply be written down, the chances for derivation errors are minimised. Additionally, we do not need to use any additional computer algebra packages that may be necessary to support the derivation of some of the other approximations.

OU with SEIR

The OU approximation for the SEIR model can be derived in the same way as for the SIR model (see section 2.1.3). This gives the following set of ODEs for the means, variances, and covariances of the Gaussian process:

$$
\begin{aligned}
&\frac{\mathrm{d}m_1}{\mathrm{d}t} = -\frac{\beta}{N} m_1 m_3 \ , \quad \frac{\mathrm{d}m_2}{\mathrm{d}t} = \frac{\beta}{N} m_1 m_3 - \omega m_2 \ , \quad \frac{\mathrm{d}m_3}{\mathrm{d}t} = \omega m_2 - \gamma m_3 \ , \\
&\frac{\mathrm{d}C_{11}}{\mathrm{d}t} = \frac{\beta}{N}(m_1 m_3 - 2m_3 C_{11} - 2m_1 C_{13}) \ , \\
&\frac{\mathrm{d}C_{12}}{\mathrm{d}t} = \frac{\beta}{N}(m_3(C_{11} - C_{12}) - m_1 C_{23}) - \omega C_{12} \ , \\
&\frac{\mathrm{d}C_{13}}{\mathrm{d}t} = -\frac{\beta}{N}(m_3 C_{13} + m_1 C_{33}) + \omega C_{12} - \gamma C_{13} \ , \\
&\frac{\mathrm{d}C_{22}}{\mathrm{d}t} = \frac{\beta}{N}(m_1 m_3 + 2m_1 C_{23} + 2m_3 C_{12}) + \omega(m_2 - 2C_{22}) \ , \\
&\frac{\mathrm{d}C_{23}}{\mathrm{d}t} = \frac{\beta}{N}(m_3 C_{13} + m_1 C_{33}) - \omega(m_2 - C_{22} + C_{23}) - \gamma C_{23} \ , \\
&\frac{\mathrm{d}C_{33}}{\mathrm{d}t} = \omega(m_2 + 2C_{23}) + \gamma(m_3 - 2C_{33}) \ .
\end{aligned}
$$

Note that this derivation required more work than for the A noise approximation where $\boldsymbol{A}$, $\mathbf{b}$, and $\boldsymbol{U}$ were just able to be written down.

MVN moment closure with SEIR

The MVN moment closure approximation of the SEIR model is derived in exactly the same way as for the SIR model (see section 2.1.3). Note that this was quite an involved procedure and certainly not as simple as writing down the A noise approximation.

We get that $X(t)$, $Y(t)$, and $Z(t)$ follow a Gaussian process $\mathrm{GP}(\boldsymbol{\mu}(t), \boldsymbol{\sigma}(t))$ with

mean and variance-covariance matrix that obey equations

$$\begin{aligned}
\frac{\mathrm{d}\mu_X}{\mathrm{d}t} &= -\frac{\beta}{N}(\mu_X\mu_Z + \sigma_{XZ}) \ , \quad \frac{\mathrm{d}\mu_Y}{\mathrm{d}t} = \frac{\beta}{N}(\mu_X\mu_Z + \sigma_{XZ}) - \omega\mu_Y \ , \\
\frac{\mathrm{d}\mu_Z}{\mathrm{d}t} &= \omega\mu_Y - \gamma\mu_Z \ , \\
\frac{\mathrm{d}\sigma_{XX}}{\mathrm{d}t} &= \frac{\beta}{N}(\mu_X\mu_Z + \sigma_{XZ} - 2\mu_X\sigma_{XZ} - 2\mu_Z\sigma_{XX}) \ , \\
\frac{\mathrm{d}\sigma_{XY}}{\mathrm{d}t} &= \frac{\beta}{N}(-\sigma_{XZ} - \mu_X\mu_Z + \mu_X\sigma_{XZ} + \sigma_{XX}\mu_Z - \mu_X\sigma_{YZ} - \sigma_{XY}\mu_Z) - 2\omega\sigma_{XY} \ , \\
\frac{\mathrm{d}\sigma_{XZ}}{\mathrm{d}t} &= -\frac{\beta}{N}(\mu_X\sigma_{ZZ} + \mu_Z\sigma_{XZ}) - \gamma\sigma_{XZ} + \omega\sigma_{XY} \ , \\
\frac{\mathrm{d}\sigma_{YY}}{\mathrm{d}t} &= \frac{\beta}{N}(\mu_X\mu_Z + \sigma_{XZ} + 2\mu_X\sigma_{YZ} + 2\sigma_{XY}\mu_Z) + \omega(\mu_Y - 2\sigma_{YY}) \ , \\
\frac{\mathrm{d}\sigma_{YZ}}{\mathrm{d}t} &= \frac{\beta}{N}(\mu_X\sigma_{ZZ} + \mu_Z\sigma_{XZ}) - (\omega+\gamma)\sigma_{YZ} - \omega(\mu_Y - \sigma_{YY}) \ , \\
\frac{\mathrm{d}\sigma_{ZZ}}{\mathrm{d}t} &= \omega(\mu_Y + 2\sigma_{YZ}) + \gamma(\mu_Z - 2\sigma_{ZZ}) \ .
\end{aligned}$$

Results

Working at two significant figures or zero decimal places as appropriate, parameter estimates and 95% confidence intervals from fitting the data with each of these approximations are given in table 2.1.

Table 2.1: Epidemic model parameter estimates and 95% confidence intervals from maximum likelihood. Note that the confidence intervals are truncated at zero for rate parameters.

Approximation	β	γ (days^{-1})	$S(0)$
MVN moment closure	21 [8.4, 33]	1.7 [0, 3.9]	258 [159, 357]
A noise	23 [0.81, 44]	1.5 [0, 4.7]	241 [125, 357]
OU	18 [8.6, 27]	1.1 [0, 3.1]	237 [137, 336]

The average infectious periods (estimated from $1/\gamma$ as 0.59, 0.91, and 0.67 days for the MVN moment closure, OU, and A noise approximations respectively) are shorter than the natural history of norovirus would indicate, which is likely to be due to control measures in place upon the ship [80] limiting the time period during which cases are able to infect others. Additionally, $S(0)$ is estimated as much smaller than N, which could be due to pre-existing immunity, control measures in place on board the ship, and non-homogeneous mixing (through excursion choice and cabin

location), as discussed previously [80].

The standard error estimates for the confidence intervals were taken from the leading diagonal of an approximate covariance matrix of the parameter estimates. The approximate covariance matrices were computed as the negative inverse of an approximation to the Hessian of the log-likelihood at the maximum likelihood estimates obtained using finite differences (from the MATLAB function `mlecov()`). The correlation matrices between the parameters for each approach are, respectively,

$$\boldsymbol{R}_{\mathrm{MVN}} = \begin{pmatrix} 1.00 & 0.77 & 0.37 \\ 0.77 & 1.00 & 0.87 \\ 0.37 & 0.87 & 1.00 \end{pmatrix} ,$$

$$\boldsymbol{R}_{\mathrm{LN}} = \begin{pmatrix} 1.00 & 0.79 & 0.57 \\ 0.79 & 1.00 & 0.94 \\ 0.57 & 0.94 & 1.00 \end{pmatrix} ,$$

$$\boldsymbol{R}_{\mathrm{A}} = \begin{pmatrix} 1.00 & 0.91 & 0.71 \\ 0.91 & 1.00 & 0.93 \\ 0.71 & 0.93 & 1.00 \end{pmatrix} .$$

The overall picture, from these confidence intervals and correlation matrices, is as would be expected when fitting a complex non-linear stochastic model to limited data: highly correlated parameters with relatively large marginal confidence intervals.

Results for learning the time series of $S(t)$, $E(t)$ and $I(t)$ are shown in Figure 2.6, which shows general agreement on mean behaviour, but differences in the uncertainty. Note that, particularly for the LN approximation, near the start of the outbreak two standard deviations below the mean gives a value of less than zero for the number of exposed and infectious individuals. This implies that perhaps the Gaussian approximation is not as suitable when the number of individuals is very close to zero.

In the results presented so far, the full dataset has been used to estimate the parameters of the epidemic model, before the time series were estimated as the epidemic progressed. We show here, in figure 2.7, the results of also estimating β, γ, and S_0 as the epidemic progressed, beginning from day nine. Note that in this instance, these methods did not work with fewer than nine days of data. These datapoint-by-datapoint estimates remain consistent over the epidemic.

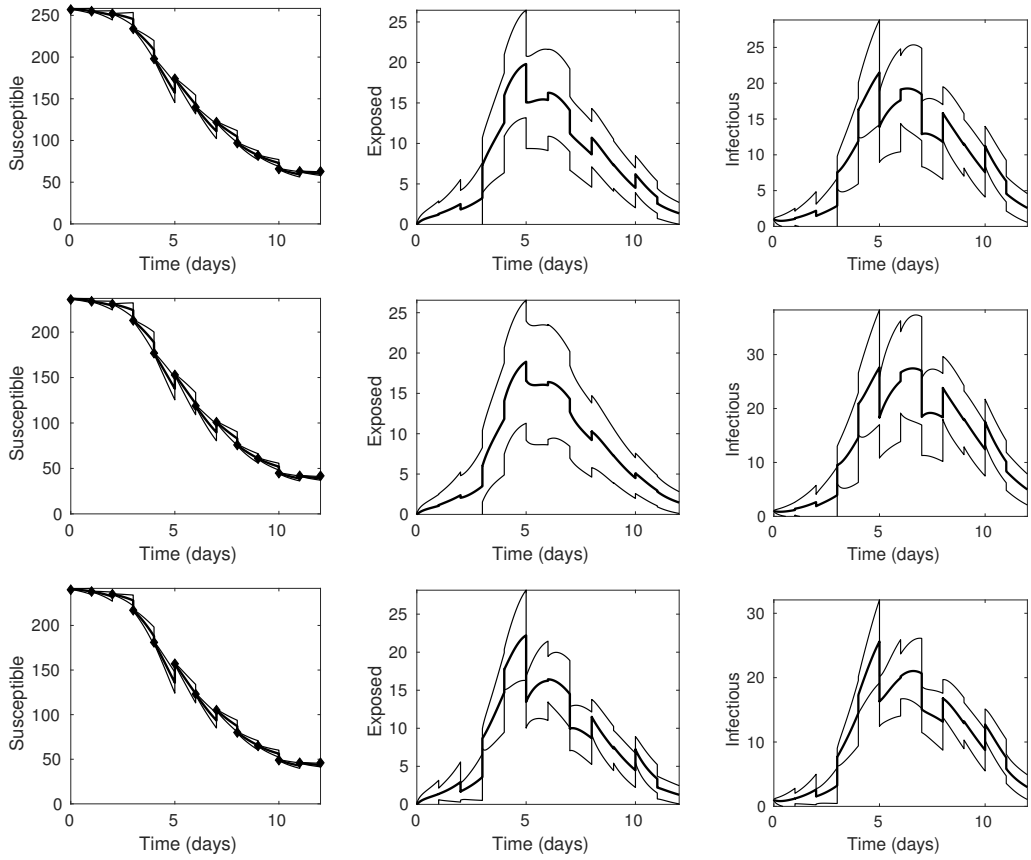


Figure 2.6: Inference of the susceptible (left), exposed (centre), and infectious (right) population using the MVN moment closure approximation (top), the linear noise approximation (middle), and the A noise approximation (bottom) from data of the number of new cases of norovirus per day on a cruise ship. The black diamonds (left) are our known values which we obtain from the data reported by [80], as described in the text. The dark lines are the mean and light lines are the mean plus/minus one standard deviation.

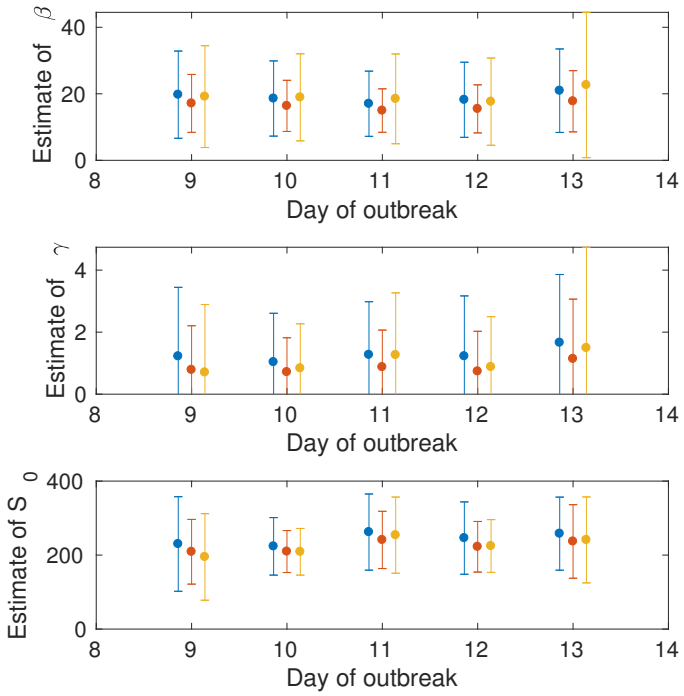


Figure 2.7: Estimates of the model parameters β, γ, and S_0 as the epidemic progresses for the multivariate normal moment closure approximation (blue), the linear noise approximation (red), and the A noise approximation (yellow). Points are maximum likelihood estimates and bars indicate 95% confidence intervals (truncated at zero for rate parameters).

We conclude that even in this common case where there is a small dataset of symptom onset times, our Gaussian process approach can be applied and gives epidemiologically reasonable answers in little computational time.

2.4 Analytical comparisons

Each of the Gaussian models previously described is chosen on the basis of different *a priori* assumptions. However, we would like to find a way to compare analytically the errors introduced by each approximation model. For this, we will use the framework depicted in figure 2.8.

Since we are interested in regularly-spaced, frequent observations of data, the relevant control parameter is the timestep Δt. Our starting point is the stochastic differential equation of the stochastic SIR model in the regime where the susceptible population is approximately constant (for example, at N when this is close enough to its starting value). This is chosen to simplify calculations, although note that Cauchemez et al. [88] suggest that this approximation of constant susceptible population can be made throughout the epidemic if the time period, Δt, over which it is made is relatively small. With this assumption, the SDE of interest is

$$\mathrm{d}I = rI\,\mathrm{d}t + \sqrt{\rho I}\,\mathrm{d}W\ , \tag{2.11}$$

where $r = \beta - \gamma$ and $\rho = \beta + \gamma$. We will use a stochastic Taylor method to expand this equation in Δt.

We also have Gaussian process approximations, whose mean vectors and variance-covariance matrices are given by ODEs (equation (2.5)). We will Taylor expand these and compare to the expansion of the SDE in order to describe the accuracy of the approximations (figure 2.8).

2.4.1 Stochastic Taylor expansion

Stochastic Taylor expansion is the stochastic analogue of the classical Taylor expansion which is used to obtain numerical solutions of ODEs. Stochastic Taylor schemes approximate SDEs locally in time. There are many such schemes, for example the Euler-Maruyama (EM) scheme is the most simple of these methods and

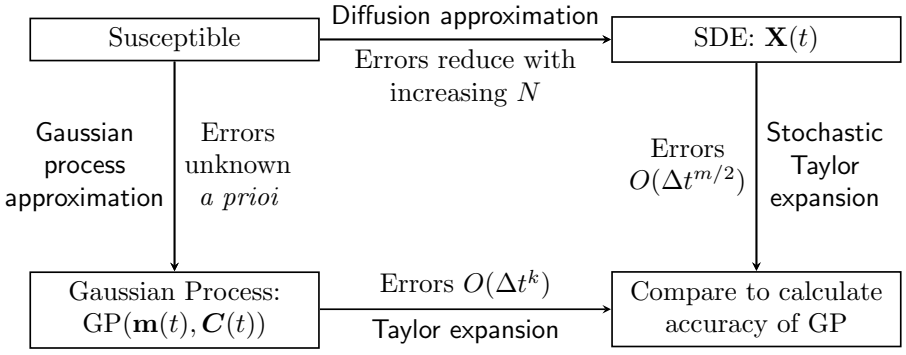


Figure 2.8: The overall scheme we will use to assess the accuracy of a given stochastic approximation. Errors are controlled by the inverse of N, the population size, and by the time-step Δt. We use k and m to stand for the integer order of errors in the time-step to be determined.

is widely used [89]. For an SDE $\mathrm{d}X_t = a(X_t)\,\mathrm{d}t + b(X_t)\,\mathrm{d}W_t$ the EM scheme gives the following approximate solution, $Y_{\Delta t}$, after a small time step Δt starting from Y_0:

$$Y_{\Delta t} = Y_0 + a(Y_0)\Delta t + b(Y_0)\Delta W \ ,$$

where $\Delta W \sim \mathcal{N}(0, \Delta t)$. This simple scheme is easy to apply however, Kloeden and Platen [89] state that "*in general, however, [the EM scheme] is not particularly satisfactory and the use of higher order schemes is recommended.*"

We use the weak order-3 scheme given by Kloeden and Platen [89]. This scheme has a rather complex general form, however for the specific SDE we have, (2.11), subject to initial condition $I(0) = I_0 \gg 1$ we obtain the following result:

$$I(\Delta t) = I_0\left(1 + r\Delta t + \frac{1}{2}r^2\Delta t^2\right) + \left(\rho I_0 \Delta t\left(1 + \frac{3}{2}r\Delta t + \frac{7}{6}r^2\Delta t^2\right)\right)^{1/2} U + O(\Delta t^3, {I_0}^0) \ , \tag{2.12}$$

where $U \sim \mathcal{N}(0, 1)$ is a standard normal random variable. This has the following mean and variance:

$$\begin{aligned} \mathrm{mean}(I(\Delta t)) &= I_0\left(1 + r\Delta t + \frac{1}{2}r^2\Delta t^2\right) \ , \\ \mathrm{var}(I(\Delta t)) &= \rho I_0 \Delta t\left(1 + \frac{3}{2}r\Delta t + \frac{7}{6}r^2\Delta t^2\right) \ . \end{aligned}$$

2.4.2 Taylor expand ODEs giving the approximations

Next, we Taylor expand the ODEs giving the mean and variance-covariances of the Gaussian process approximations so that we may compare to the result of the expansion in the previous section.

For the Gaussian process approximations that arise from the linear SDE approach (section 2.1.3), we consider the ODEs for the mean and variance-covariances (equation (2.5)) again in the limit where the size of the susceptible population is approximately constant. This gives the following ODEs for the mean, m_2, and variance, C_{22}, of the size of the infectious population

$$\frac{\mathrm{d}m_2}{\mathrm{d}t} = A_{21}(t)N + A_{22}(t)m_2(t) + b_2(t) \ , \qquad \frac{\mathrm{d}C_{22}}{\mathrm{d}t} = 2A_{22}(t)C_{22}(t) + \rho i(t) \ .$$

Taylor expanding these, subject to initial conditions $m_2(0) = i(0) = I_0$, $C_{22}(0) = 0$, we get, for each considered model, that the mean is

$$m_2(\Delta t) = I_0 \left(1 + r\Delta t + \frac{1}{2}r^2\Delta t^2 + \cdots\right) \ .$$

For the variance, for b noise we have

$$C_{22}(\Delta t) = \rho I_0 \left(1 + \frac{1}{2}r\Delta t + \frac{1}{6}r^2\Delta t^2 + \cdots\right)\Delta t \ ,$$

and for the A noise and OU approximations we have

$$C_{22}(\Delta t) = \rho I_0 \left(1 + \frac{3}{2}r\Delta t + \frac{7}{6}r^2\Delta t^2 + \cdots\right)\Delta t \ .$$

For the MVN moment closure approximation, the mean, $\mu_Y(t)$, and variance, σ_{YY}, of the epidemic at constant susceptible population are given by the following ODEs:

$$\frac{\mathrm{d}\mu_Y}{\mathrm{d}t} = r\mu_Y, \qquad \frac{\mathrm{d}\sigma_{YY}}{\mathrm{d}t} = \rho\mu_Y + 2r\sigma_{YY}.$$

We also Taylor expand these subject to $\mu_Y(0) = I_0$, $\sigma_{YY}(0) = 0$ to obtain

$$\mu_Y(\Delta t) = I_0 \left(1 + r\Delta t + \frac{1}{2}r^2\Delta t^2 + \cdots\right) \ ,$$
$$\sigma_{YY}(\Delta t) = \rho I_0 \left(1 + \frac{3}{2}\Delta t r + \frac{7}{6}r^2\Delta t^2 + \cdots\right)\Delta t \ .$$

2.4.3 Bounding the errors of the Gaussian process

Putting the results of the previous two sections together, we see that the MVN moment closure, the linear noise, and the A noise approximations are all consistent with the SDE equation (2.11) expanded as in equation (2.12). This justifies our continued work with these approximations through section 2.3. However, the b noise approximation is less accurate. This, again, justifies our decision to not take it further into section 2.3.

To see why errors at this order represents a significant improvement on other possible approaches, consider the EM approximation to the SDE (equation (2.11)),

$$J_{\Delta t} = (1 + r\Delta t)I_0 + \sqrt{\rho I_0}U \ ,$$

where $U \sim \mathcal{N}(0, \Delta t)$ is, again, a standard Gaussian random variable. Comparing this to equation (2.12), we see that errors to this appear at $O(\Delta t)$. This tells us that the ODE-based Gaussian process approximations we have described and analysed are more accurate than, the frequently used, Euler-Maruyama steps by several orders of magnitude in Δt.

2.5 Discussion and conclusions

In this chapter we have investigated Gaussian process approximations of stochastic models of epidemics with an aim to provide results that will allow these approximation techniques to become more routinely used in disease surveillance and epidemiology.

Throughout this chapter we have applied these approximations to the stochastic SIR model and additionally to the SEIR model when we look at real data. However, one strength of this approximation framework, and the A and b noise approximations discussed, is how straightforward it is to write down the approximation for a more complex compartmental model. This approach may even be easy to use with models from outside of epidemiology.

Our analytical approach for quantifying the accuracy of the Gaussian process approximations (section 2.4) is only in the specific regime where S is approximately constant. As we have previously stated, Cauchemez et al. [88] suggests this approx-

imation is appropriate throughout the epidemic if the time period Δt is relatively small. However, this may be a weakness of our approach.

2.5.1 Further work

We compare approximations from the linear stochastic differential equation with the MVN moment closure approximation. We note, in section 2.1.3, that other moment closure approximations, such as the log-normal, also exist although we do not implement them in the current study. We could do this in the future to quantify how these compare to the approximations already discussed.

We update our current estimates of the population in each epidemic class when new data are obtained. We take the mean to the observed data point and the variance to zero. In the future, we would like to work from the assumption that these data may not be perfect observations. We would not update the variance to zero, but consider the observation to be a sample from a normal distribution with small variance. This would also change how the unobserved classes were updated using the MVN marginal conditional rules.

For future, more long term, plans we would like to incorporate these approximations into online real-time, robust systems to help improve disease surveillance (this is discussed further in section 5.2).

2.5.2 Conclusions

In this chapter we have presented a flexible framework for deriving Gaussian process approximations of non-linear stochastic models of epidemics using the SIR model as our initial example. We have numerically and analytically compared a variety of approximations reported in the literature and additional examples we do not believe have previously been named. We have shown how these approximations can be used to perform quick maximum likelihood inference for the underlying parameters of the epidemic model given population measures of incidence or prevalence at given time points. Finally, we also show how the unobserved processes can be inferred at the same time as the underlying parameters. This work goes some way in demonstrating how, with appropriate approximations, stochastic epidemic models can be used for fast inference and so these kinds of models could be used more routinely with

regularly updated surveillance data.

CHAPTER 3

DAY OF THE WEEK AND PUBLIC HOLIDAY EFFECTS IN SYNDROMIC SURVEILLANCE DATA

3.1 Introduction

Anecdotal evidence of day of the week and public holiday effects in daily data from a range of syndromic surveillance systems operated by Public Health England (PHE) was informally reported by analysts in the Real-time Syndromic Surveillance Team (ReSST). The purpose of this chapter is to formally describe these effects in order to understand one of the systematic causes of bias in healthcare data. This will, therefore, improve procedures for the current analysis and presentation of these data which give a lot of information on gastroenteritis burden in England.

This work was undertaken during a secondment at the ReSST of PHE during this PhD. The data used in this work are covered by governance and contractual agreements that limit their use for PHE surveillance activities only. The data are therefore not available for sharing. It should also be emphasised that the opinions expressed herein do not necessarily reflect the views of the ReSST or any part of PHE.

The work presented in this chapter has been used to improve surveillance methodologies at PHE. In particular, it has been used to improve the visualisation of data (section 3.4.2). Additionally, section 3.4.2 has been published as:

- **E. Buckingham-Jeffery, R. Morbey, T. House, A. J. Elliot, S. Harcourt, G. E. Smith.** (May 2017) *Correcting for day of the week and public holiday effects: Improving a national daily syndromic surveillance service for detecting public health threats.* BMC Public Health, 17:477.

The chapter will be structured as follows. First, a review of the background knowledge needed to approach this problem, a precise statement of the aims, and a description of the data. Second, an investigation into day of the week effects in the data followed by a similar investigation into public holiday effects. Finally, descriptions of two ways in which the knowledge obtained in the previous two sections can be used to improve current syndromic surveillance systems at PHE.

3.1.1 Background to syndromic surveillance at Public Health England

Traditionally, disease surveillance was based on monitoring a set of pre-determined diseases with laboratory confirmation. However, these systems can be slow and unable to detect novel, unexpected diseases [90].

The European wide *Triple S* project defines syndromic surveillance as the "*real-time (or near real-time) collection, analysis, interpretation, and dissemination of health-related data to enable the early identification of the impact (or absence of impact) of potential human or veterinary public health threats that require effective public health action*" [90]. The notable difference between this definition of syndromic surveillance and the definition of disease surveillance is the requirement of timeliness. In order to do this, syndromic surveillance systems report data from signs and symptoms to infer the presence a disease before clinical or laboratory confirmation [90, 91].

PHE has responsibility for disease surveillance within England. In particular, the ReSST coordinates several national syndromic surveillance systems (SSSs). There are four key systems and data will be used from each of these in this chapter [92]. Each system monitors the daily number of contacts or attendances with a healthcare service for a wide range of syndromes and are fully described in the given references.

The remote health advice SSS monitors the number of phone calls to the NHS 111 non-emergency telehealth number (we will refer to this system as the 111 SSS) [93]. The GP in hours (GPIH) SSS monitors the number of visits to GPs during regular

surgery hours (which are, typically, normal working hours from Monday to Friday, excluding any public holidays) [94]. It covers approximately 55% of England's population. The GP out-of-hours (GPOOH) SSS monitors the number of unscheduled contacts (visits and calls) to GPs during evenings, overnight, on weekends, and on public holidays [95]. It covers approximately 80% of England's population. Finally, the emergency department (ED) SSS monitors visits to a sentinel network of emergency departments [96].

Contacts or attendances associated with a particular syndrome is called a (syndromic) indicator. Data from an indicator are a daily time series of the number of contacts with or attendances at the particular healthcare service with the particular syndrome. There are indicators in each SSS for either gastroenteritis or vomiting and diarrhoea. As introduced in section 1.2, data are analysed each day using the rising activity, multi-level mixed effects, indicator emphasis (RAMMIE) method developed by Morbey et al. (2015, [30]) to detect unusual activity that could require further investigation.

3.1.2 Background to day of the week and public holiday effects

The seven-day week is often out of agreement with other calendar features. In particular, the start of a new year always coincides with the start of a new month. But days of the year and days of the week are independent of one another; the 1st of January does not always fall on the same day of the week and a month is not made of a whole number of weeks [97].

However, despite the complications arising from the lack of synchrony between the week, the month and the year, the seven-day week has become engrained in our society. Many activities in our lives are given routine and structure by the week. In particular, the traditional Monday to Friday working week followed by a weekend has established a "*labour and rest rhythm*" [98]. It is, therefore, not surprising that a regular seven day periodicity is seen in time series of the number of occurrences of many events and activities. We call a statistically significant difference in the number of occurrences of an event on a particular day of the week, or set of days, a *day of the week effect*.

This often, although not always, manifests as a difference between the days of the working week (Monday to Friday) and the weekend days (Saturday and Sunday).

However, according to the most recent *European Working Conditions Survey* (2016, [99]), 22% of UK workers work at least one night a month and 59% work at least once at the weekend per month.

We define a *public holiday effect* as a statistically significant difference in the number of occurrences of an event on, or near to, a public holiday compared with other similar days. This often manifests as public holiday days having similar properties to weekend days, as many workplaces close at these times (however there is no legal requirement in the UK for workers to be given these days off).

Other time periodicities are also common in time series data, such as monthly or seasonal effects. However, this investigation specifically studies just day of the week and public holiday effects in syndromic surveillance data. Timeliness is key in syndromic surveillance, and thus it is important to understand patterns on these short time-scales. Additionally, the fast availability and analysis of daily data updates is a relatively new practice, only available due to recent improvements in computing systems.

Statistical analysis of day of the week effects first became commonplace in the econometrics literature during the 1980s. French (1980, [100]) was one of the first to statistically compare mean stock market returns on each day of the week and reported that the average returns on Mondays were negative and lower than the average returns on the other days of the working week in an American stock market index. Day of the week effects were then found in similar markets worldwide, but not necessary with lower returns on Mondays [101, 102]. The investigation was extended to emerging markets, with significant day of the week effects later found in, amongst others, Turkish, Singaporean, Malaysian, Taiwanese, Thai, Hong Kongese, and Bangladeshi markets [103–106]. However, more recent studies are reporting that even though a day of the week effect was clear in markets in the 1980s, these have been reducing over time and perhaps even reversing [107].

Similar analyses have also discovered public holiday effects in stock market returns. In particular, significantly higher returns on the day preceding a public holiday have been found in markets across the world (note that markets are closed on public holidays) [108–111]. However, similarly to the day of the week effect, these effects are reported to be diminishing over time [112]. Further studies consider the public holidays separately and discover, for example, that the Christmas and New Year holidays have the largest effect [112].

Similar statistical methods have since been used to identify day of the week and public holiday effects in a broad range of applications beyond the financial sector. More papers are submitted to the *Journal of the Serbian Chemical Society* on Wednesday than any other day, but a higher proportion of submissions made on a Tuesday are accepted [113]. Maximum levels of particulate pollution in California occur at the end of the working week and minimum levels on Sunday [114]. Significantly more crimes are reported in underground stations in Stockholm on holidays compared to days of the working week [115]. The length of time taken to fix bugs in the Ubuntu Linux distribution differs by the day of the week that the bug was reported [116]. And in Ontario, Canada, young male drivers had marginally significantly more accidents on Fridays and Saturdays compared to other days of the week [117].

Day of the week and public holiday effects have also been identified in health related applications, for example in data relating to food and exercise. Weekend days are associated with higher levels of physical activity [118], and lower levels of dangerous physical inactivity [119]. In a study of children in Atlanta, U.S., fruit and vegetables were most frequently consumed on days of the working week compared to weekend days [120]. An increased intake of energy, protein, and many micro-nutrients on weekend days, particularly Sundays, was reported for an elderly population in Norwich, UK [121]. Higher rates of alcohol were drunk on Thursdays compared to other days of the working week by U.S. college students [122], and presentations to emergency departments with acute alcohol intoxication increased substantially in Australia on the day before a public holiday [123].

Day of the week and public holiday effects have also been identified in mental health data. Day of the week effects were reported in both self-reported measures of mood and sentiment analysis generated measures of mood from *Facebook* statuses [124, 125]. There are higher levels of binging and purging behaviours in people with bulimia on Sundays [126]. There are, reportedly, an increased number of suicides on Mondays [127–129]. It is also widely acknowledged that there are fewer suicides or suicide attempts than expected before major public holidays, but more than expected afterwards [130, 129, 131–133] leading to the theory that some suicides are 'postponed' until after these major events.

Additionally, and perhaps now more controversially, day of the week and public holiday effects have been found in the health outcomes of patients admitted to hospital, being born in hospital, or receiving an operation [134–137], with weekend days typically having a higher proportion of worse outcomes. This has led to heated

political discussions, with some politicians attributing this effect to fewer doctors being available on weekend days [138, 139]. However, the cause of this weekend effect is unclear and unknown. Many other possibilities have been proposed such as inconsistent data and the severity of illness of patients admitted on a weekend day [140–142].

We are interested, however, in day of the week and public holiday effects in the number of patients using healthcare services. We have a particular focus on those presenting with gastrointestinal symptoms, however we will also look at total service use and other syndromes for comparison. Previous studies have identified some statistically significant day of the week and public holiday effects in healthcare service presentation, both in general and for specific conditions. This includes in emergency services across the U.S. [143–145] and in a sexual health clinic in Australia [146]. The results of studies of particular conditions show increased attendances at emergency departments for severe asthma on Sundays and Mondays [147] and, perhaps the most widely reported, increased cardiac difficulties on Mondays [148–150]. There has not previously been a comprehensive analysis of day of the week and public holiday effects in syndromic surveillance data from healthcare services in England.

There may be multiple, complex, mechanisms driving these day of the week and public holiday effects. Some of the statistical analyses mentioned in this section hypothesise the cause of the effects they report. However, this rapidly becomes difficult to prove and requires specialist knowledge of the application area. We comment, briefly, throughout this chapter on possible causes but do not wish to delve too much into this.

Day of the week and public holiday effects in presentation data to healthcare services have, we hypothesise, two components. Firstly, the effects due to actual changes in levels of illness. For example, increased stress on Mondays has been linked to increased cardiac problems on this day of the week [151]. And secondly, the effects due to the timing of reporting of illness *rather than* the timing of illness itself.

It is not possible to disentangle these two effects with the data used in this study. When discussing day of the week effects in healthcare data, Zerubavel (1989, [98]) states that *"only a non-medical, sociological explanation can account for these findings"*. However, an intrinsic day of the week effect in infectious disease levels should not be discounted. The difference in contact patterns (for example as reported by

Edmunds, 1997 [152] and in the *POLYMOD* study, 2008 [153]) and lifestyle between days of the working week and weekend days would be expected to impact on the spread of disease and on an individual's inherent susceptibility.

Many people with illness do not seek help from healthcare services. There are reported differences in the uptake of healthcare based on gender [154], ethnicity, and socio-economic status [155]. However, many factors can influence when, or if ever, an individual reports their symptoms including the availability of services (for example opening times, waiting times, systems for obtaining appointments) and the impact of the illness on an individual's everyday life [156, 157]. These will contribute to any observed day of the week and public holiday effects.

3.1.3 Aims and objectives

The aim of this chapter is to improve the current analysis and interpretation of syndromic surveillance data of gastroenteritis by describing the regular biases in reporting symptoms to healthcare services in England caused by day of the week and public holidays. This will be achieved by:

- Formally identifying and describing the day of the week and public holiday effects in syndromic surveillance data of gastroenteritis from different syndromic surveillance systems.

- Formally identifying and describing the day of the week and public holiday effects in syndromic surveillance data of other key conditions and of total healthcare use for comparison.

- Describing how the RAMMIE method can be improved based on this knowledge of day of the week and public holiday effects.

- Developing improved methods for the visualisation of syndromic data based on this knowledge of day of the week and public holiday effects. These visualisations are used for the analysis of trends and for presentation to the public.

3.1.4 Data

Data have been made available for this analysis from each syndromic surveillance system managed by PHE (table 3.1). There is no gastroenteritis indicator in the 111 SSS, but the diarrhoea and vomiting indicators were chosen as a close comparison.

The difficulty breathing/wheeze/asthma, or severe asthma, indicator was chosen as it can also be compared across multiple syndromic surveillance systems. In comparison to gastroenteritis, which is often self-limiting, asthma is a health complaint for which it is more necessary to regularly visit a healthcare professional [158]. This makes for an interesting comparison condition.

Two further indicators (cardiac from the ED SSS and herpes zoster from the GPIH SSS) were chosen based on anecdotal evidence of unusual or interesting day of the week and public holiday effects.

Finally, the total number of contacts or attendances with the 111, GPOOH, and ED SSSs was also analysed for day of the week and public holiday effects. This will help us understand how reporting gastroenteritis symptoms differs from general reporting of poor health. Note that this data is not available in the GPIH SSS.

3.2 Day of the week effects

In this section we will present the methods used to investigate the magnitude of day of the week effects in syndromic data and the results obtained from this analysis. Where statistical tests have been used, they were separately applied in the same way to each indicator (indicators as described in section 3.1.4).

3.2.1 Exploring the data

As the time series data from the syndromic surveillance systems span multiple years and seasons there are annual, and other longer term, trends present. We are not interested in these other effects and they make it inappropriate to identify day of the week effects by simply comparing the average contacts or attendances on each day of the week across the full time series. We, therefore, use an adjusted time series for this analysis which is constructed by subtracting from each day's contacts

Table 3.1: A summary of the syndromic indicators used for this analysis from the syndromic surveillance systems operated by Public Health England.

Syndromic surveillance system	Dates data available	Indicator
GP out of hours	09/01/2012 - 11/01/2015	Total contacts Difficulty breathing/wheeze/asthma Gastroenteritis
111	30/09/2013 – 11/01/2015	Total contacts Difficulty breathing Diarrhoea Vomiting
GP in hours	02/04/2012 – 11/01/2015	Severe asthma Gastroenteritis Herpes zoster
Emergency department	17/09/2012 – 11/01/2015	Total attendances Difficulty breathing/wheeze/asthma Gastroenteritis Cardiac

or attendances the mean of the week's activities. Mathematically, for time series $X = \{X_t\}$ the adjusted time series $\tilde{X} = \{\tilde{X}_t\}$ is computed as

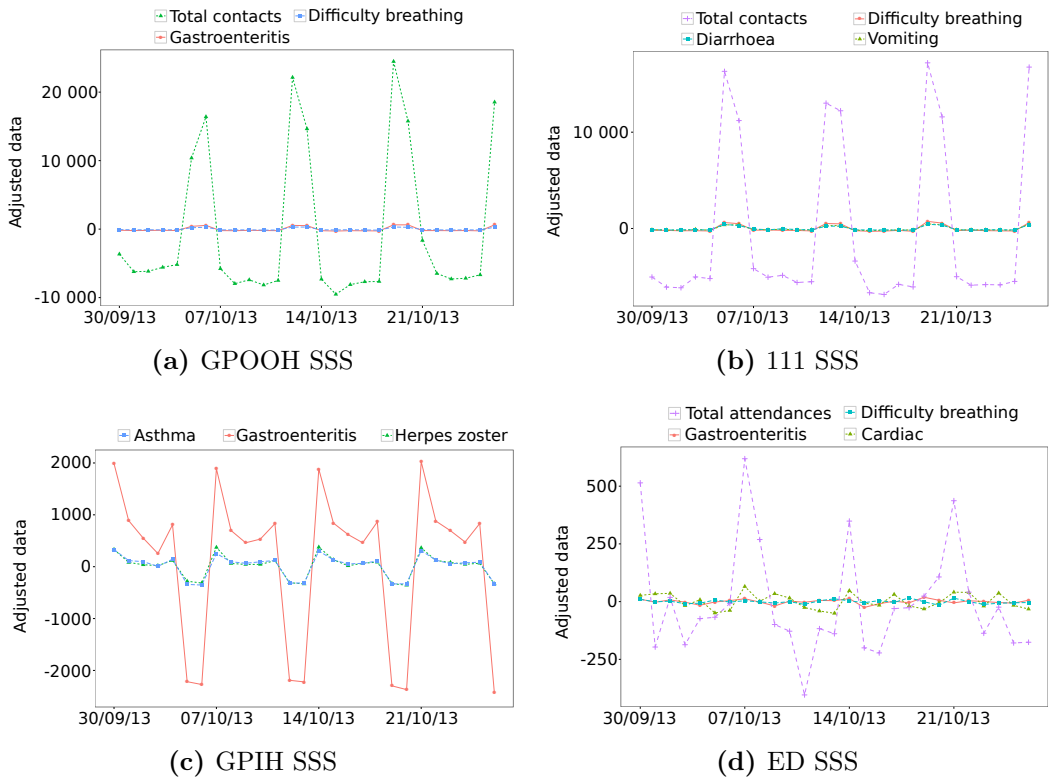
$$\tilde{X}_t = X_t - \frac{1}{7} \sum_{i \in W_t} X_i \ ,$$

where W_t is the week (Monday to Sunday) containing day t.

Additionally, we suspect (and will show in section 3.3) that weeks containing public holidays have different numbers of attendances than typical weeks. Therefore, we exclude these weeks from all analysis of day of the week effects.

A sample of four weeks of the adjusted time series data from each indicator studied in the four syndromic surveillance systems is presented as an example of the data being analysed (figure 3.1). A summary of each adjusted time series is shown in a box plot (figures 3.2 to 3.5).

The samples of data implies that there is a clear day of the week effect in the data from the GPOOH, 111, and GPIH syndromic surveillance systems (figure 3.1), and this is confirmed by the box plots (figures 3.2 to 3.4). In both the GPOOH and



(a) GPOOH SSS **(b)** 111 SSS

(c) GPIH SSS **(d)** ED SSS

Figure 3.1: A four week sample of the adjusted data from each indicator.

111 syndromic surveillance systems, there are more contacts on weekend days than on average for the week. In the GPIH SSS there are more contacts on days of the working week than on weekend days. The daily distributions look roughly symmetric (as shown by the roughly symmetric boxes and whiskers) and the interquartile ranges are small. This indicates that the day of the week effects are quite consistent over the full time period of the data. However, there are outliers in each dataset. These patterns are consistent across all the indicators analysed from the GPOOH, 111, and GPIH systems.

Considering the opening hours of GP practices, and the aim that out of hours services should supplement them during closures, these day of the week effects are exactly what we expect. It is not immediately obvious that there are further day of the week effects in the GPOOH and 111 syndromic surveillance systems beyond the weekend and working week divide. However, there appears to be further day of the week effects within the days of the working week in data from the GPIH SSS (figure 3.4). Further analysis will investigate these, potentially more subtle, day of the week effects and confirm and quantify the obvious day of the week effects.

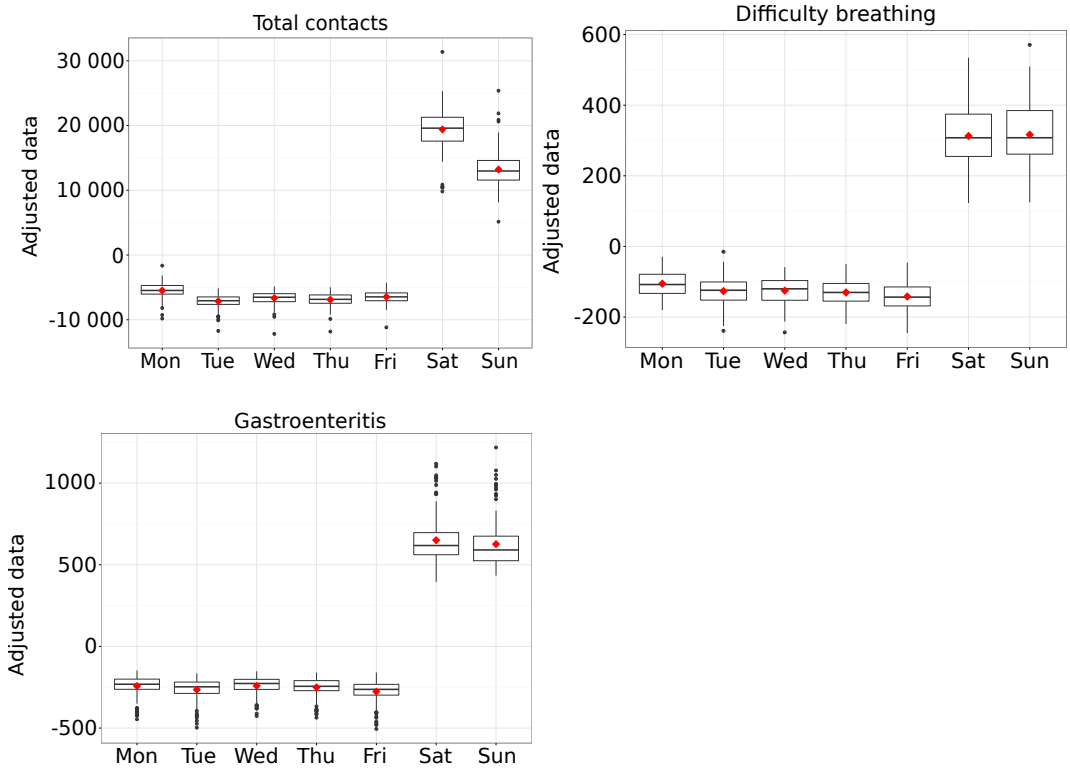


Figure 3.2: Box plots of the adjusted data from the GPOOH SSS. The lower whisker is at the smallest data point within 1.5 times the interquartile range of the first quartile, and the upper whisker is at the largest data point within 1.5 times the interquartile range of the third quartile. Data points outside of this range are individually plotted. The lower line of the box is at the first quartile, the middle line is at the median, and the upper line of the box is at the third quartile. The interquartile range is the difference between the third and first quartiles. The mean is additionally marked with a red diamond.

It is clear that there are more attendances at emergency departments on Mondays compared to the rest of the week (figure 3.1 and figure 3.5 top left). However, it is not clear whether this effect is also seen in the data from the indicators studied (figure 3.5). Further analysis will fully investigate day of the week effects in the ED SSS as they are not immediately obvious from these summary plots.

Based on this initial data visualisation, the rest of this section will proceed as follows. First, a thorough investigation into day of the week effects in the ED SSS, followed by an investigation of whether further day of the week effects exist in the GPOOH, 111, and GPIH data beyond the weekend effect already identified.

Note that the statistical methodologies that we have decided to use are just one

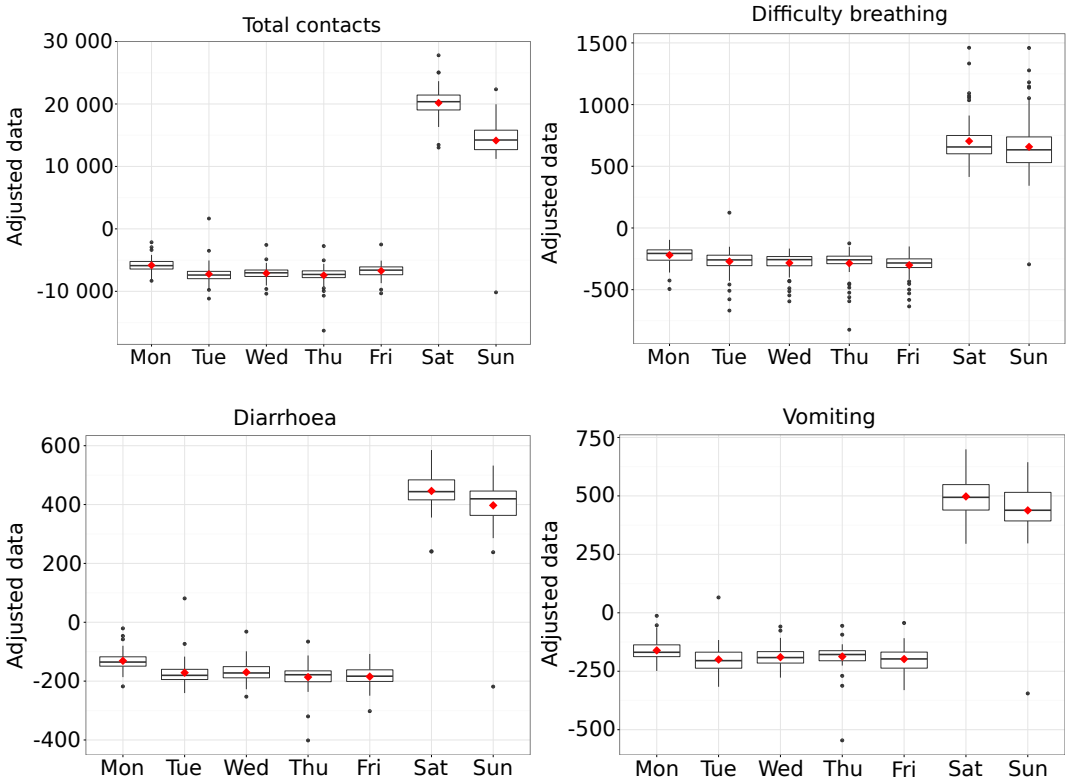


Figure 3.3: Box plots of the adjusted data from the 111 SSS. The lower whisker is at the smallest data point within 1.5 times the interquartile range of the first quartile, and the upper whisker is at the largest data point within 1.5 times the interquartile range of the third quartile. Data points outside of this range are individually plotted. The lower line of the box is at the first quartile, the middle line is at the median, and the upper line of the box is at the third quartile. The interquartile range is the difference between the third and first quartiles. The mean is additionally marked with a red diamond.

possible way to achieve the aims stated in section 3.1.3. This will be discussed further in section 3.2.4.

3.2.2 The emergency department syndromic surveillance system

Based on the visualisations in section 3.2.1, it is not immediately clear that there are day of the week effects in presentations to emergency departments. In order to identify what, if any, day of the week effects are statistically significant in the data from the ED SSS we will perform a simple regression analysis using dummy variables, followed by an F-test for regression, and Tukey's honest significant difference

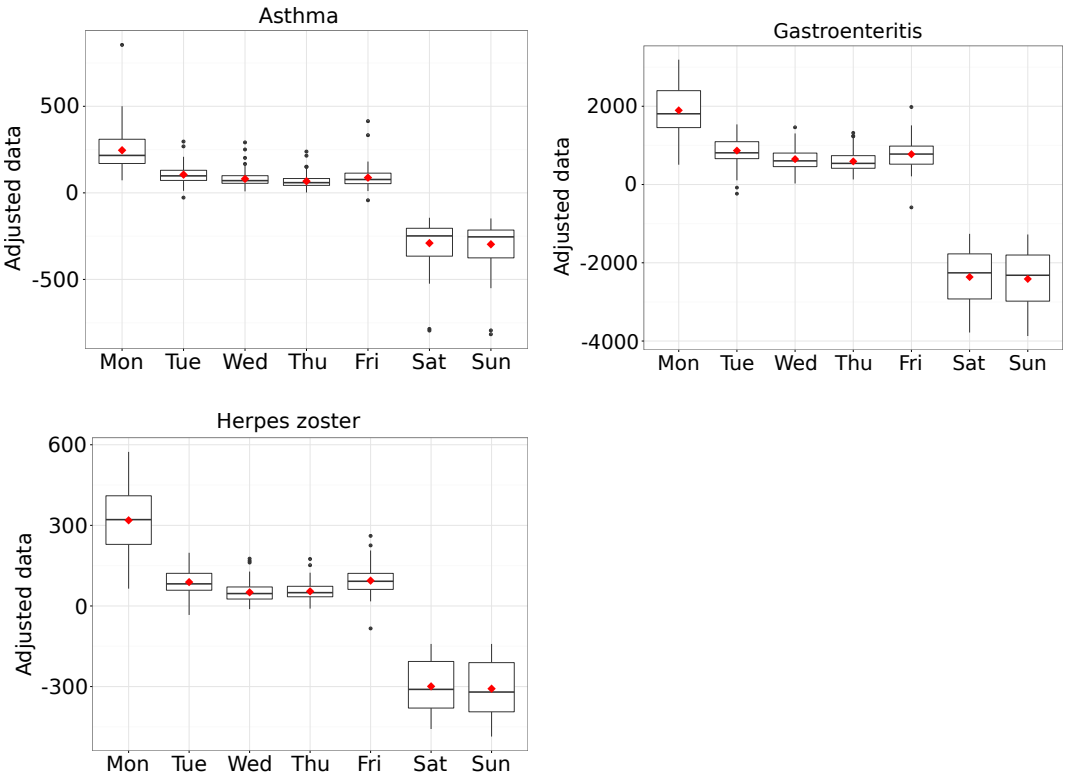


Figure 3.4: Box plots of the adjusted data from the GPIH SSS. The lower whisker is at the smallest data point within 1.5 times the interquartile range of the first quartile, and the upper whisker is at the largest data point within 1.5 times the interquartile range of the third quartile. Data points outside of this range are individually plotted. The lower line of the box is at the first quartile, the middle line is at the median, and the upper line of the box is at the third quartile. The interquartile range is the difference between the third and first quartiles. The mean is additionally marked with a red diamond.

(Tukey's HSD) test to compare pairs of days. We will also present the effect size of the difference between days because statistical significance does not always mean practical significance, in particular as we have large sample sizes.

This type of statistical analysis has been widely used in the econometrics literature to investigate day of the week effects in stock exchanges [100, 103–105, 107, 159], and a range of similar regression analyses and hypothesis tests have been used to comment on day of the week effects in emergency department attendances [143], other daily health data [118, 124, 126, 147], deaths and car accidents [117, 127–129], food and drink intake [120–122], code bug reports [116], and pollution levels [114].

We will use the adjusted time series for this analysis, as described in section 3.2.1,

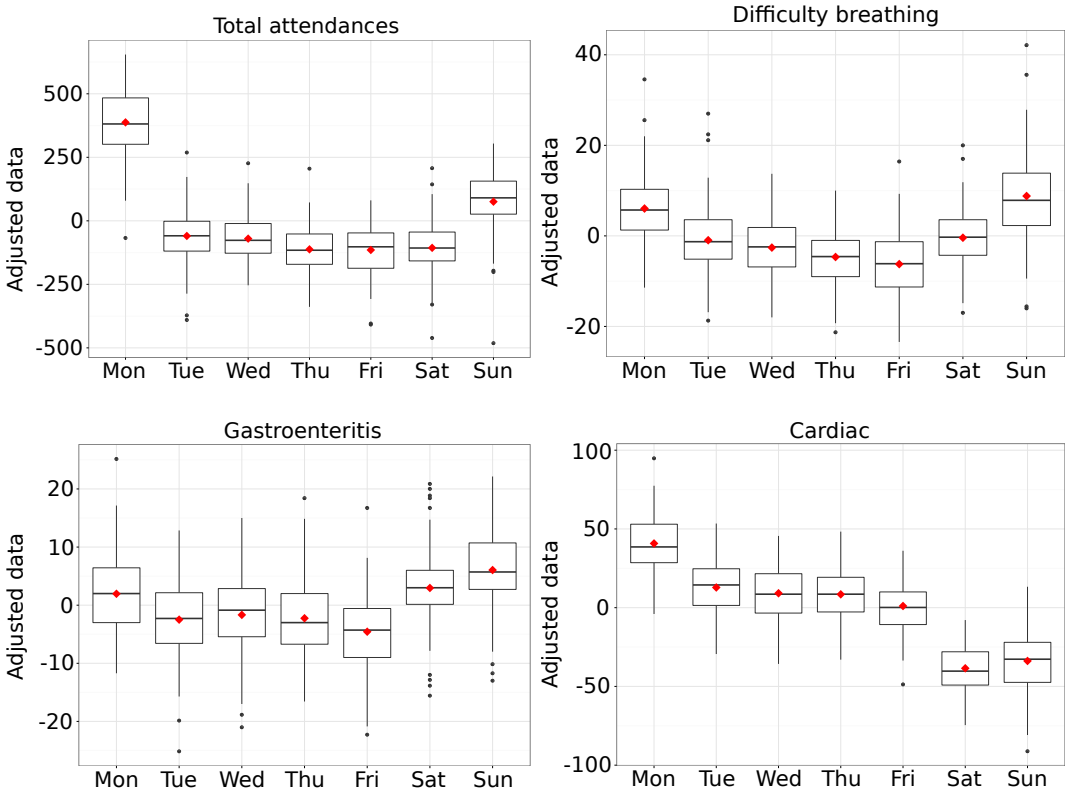


Figure 3.5: Box plots of the adjusted data from the ED SSS. The lower whisker is at the smallest data point within 1.5 times the interquartile range of the first quartile, and the upper whisker is at the largest data point within 1.5 times the interquartile range of the third quartile. Data points outside of this range are individually plotted. The lower line of the box is at the first quartile, the middle line is at the median, and the upper line of the box is at the third quartile. The interquartile range is the difference between the third and first quartiles. The mean is additionally marked with a red diamond.

where the mean number of attendances each week has been subtracted so that seasonal effects are removed.

Methods: Dummy variable regression

Rutherford, 2013 [160] has been used as the main reference for the statistical methodology that will be outlined in this section.

We fit the following simple linear dummy variable regression model to each adjusted

dataset $\tilde{X}$:

$$\tilde{X}_t = \alpha + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6 + \beta_7 D_7 + \epsilon_t \ , \tag{3.1}$$

where α is the intercept which gives the expected proportion of the week's contacts or attendances that occur on Monday (the base group), and ϵ_t is the error. D_2, ..., D_7 are dummy variables for the remaining six days of the week; that is if day t is a Tuesday then $D_2 = 1$ and $D_j = 0$ for $j = 3, ..., 7$. Finally, β_2, ..., β_7 are the regression coefficients which give the expected difference between α and the proportion of contacts or attendances on each of the other days of the week. Note, this type of simple linear regression where all the explanatory variables are dummy variables is equivalent to a one-way ANOVA test.

Fitting this model requires the calculation of the mean of the set of adjusted data points for each day of the week. That is

$$\alpha = \frac{1}{n_w} \sum_{i=1,8,15,...} \tilde{X}_i \ , \qquad \alpha + \beta_2 = \frac{1}{n_w} \sum_{i=2,9,16,...} \tilde{X}_i \ ,$$

and similarly for $\beta_3, ..., \beta_7$, where n_w is the number of weeks in the dataset (and so the number of data points for each day of the week).

An F-test of the overall significance of the regression model tests the null hypothesis that the fit of the model stated by equation (3.1) (which we call the full model) and the intercept-only model (which we call the reduced model) are equal against the alternative hypothesis that the fit of the reduced model is significantly reduced compared to the fit of the full model. Framing this as an ANOVA problem, this is equivalent to the hypothesis test with null hypothesis that there is no day of the week effect, that is $\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$, against the alternate hypothesis that there is a day of the week effect, that is at least one $\beta_i \neq 0$.

If the full model provides a better description of the data than the reduced model then it should have a smaller error component. Therefore, the F-test statistic compares the error sum of squares (SSE) of the two models

$$F = \frac{(\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}})/(df_{\text{reduced}} - df_{\text{full}})}{\text{SSE}_{\text{full}}/df_{\text{full}}} \ , \tag{3.2}$$

where df is the degrees of freedom. The reduced model has $n-1$ degrees of freedom, where n is the total number of data points. The full model has $n - 7$ degrees of

freedom. Under the null hypothesis, this test statistic has the F-distribution on $((df_{\text{reduced}} - df_{\text{full}}), df_{\text{full}}) = (6, n-7)$ degrees of freedom. Based on this, we compute a p-value so that we may comment on the significance of a day of the week effect.

In order for this analysis to be valid, the following three assumptions should be upheld:

1. **Errors are independent.** We will check the independence of residuals in time graphically.

2. **Errors exhibit common variance across all groups of the independent variable.** We will check this graphically by looking at box plots of the residuals for each day of the week. The test is robust to violations of this assumption when the sample sizes are equal, as in our case. A rule of thumb states that if the ratio of the largest variance to the smallest variance does not exceed 3 or 4 this assumption is probably satisfied [161].

3. **Errors are normally distributed.** We will check this graphically using a Q-Q plot comparing a theoretical normal distribution with the distribution of the residuals of the regression. However, this analysis is robust with respect to violations of this assumption. In particular, in our situation when the sample sizes for each group are equal and large [160].

Results: Dummy variable regression

We start by reporting the results of the assumption checks. Plots of the residuals by time do not show any obvious correlations (figure 3.6). Q-Q plots show approximate normality for the four datasets (figure 3.7). Box plots show some small variability in the spread of the residuals across the days of the week (figure 3.8). However, the ratio of residual variances between the day of the week with the largest residual variance and the day of the week with the smallest residual variance are all smaller than 2.6 (table 3.2). Therefore, we proceed with the analysis.

The F-statistic given by the regression for each dataset gives a very small p-value (table 3.3). This implies that there is a significant day of the week effect present in each dataset from the ED SSS. Further investigation in the following sections will reveal the size of these effects, and therefore how *practically significant* they are, and between exactly which days.

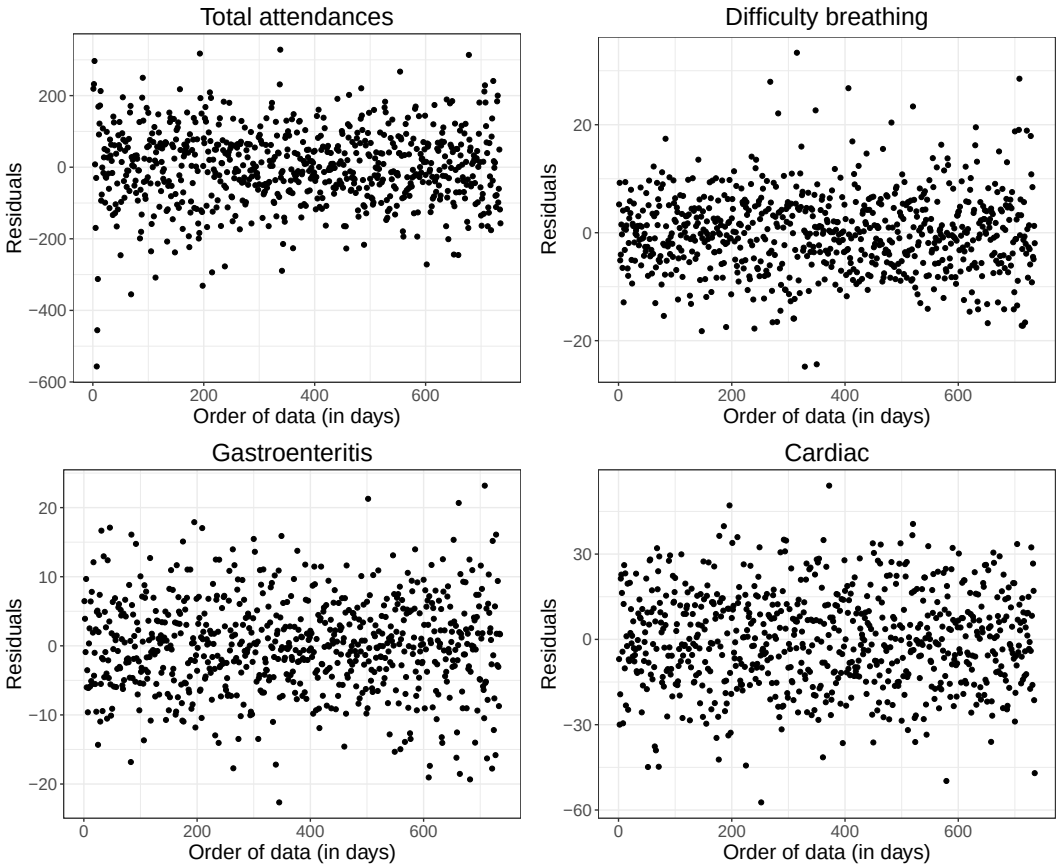


Figure 3.6: Autocorrelation plots of the residuals against time for total attendances and indicators from the ED SSS.

Methods: Post-hoc tests (Tukey's HSD) [162, 163]

The previous analysis indicates that, for each indicator studied from the ED SSS, the number of attendances on at least one day of the week differs from the other days. However, it cannot tell us which days differ from each other. To do that, we compare all days in a pairwise manner using Tukey's HSD test.

A basic hypothesis test to compare two means is a t-test. We wish to compare seven means pairwise, which could be achieved with 21 individual t-tests. However, this type of multiple hypothesis testing is not advised [162]. Suppose we are working with a significance level of 0.05. Therefore, when completing one hypothesis test the probability of getting a false positive (rejecting the null hypothesis when it is in fact true) is 5%. However, when completing 21 hypothesis tests the probability of getting one false positive is $1 - (1 - 0.05)^{21}$, so around 66%. Tukey's HSD test, on

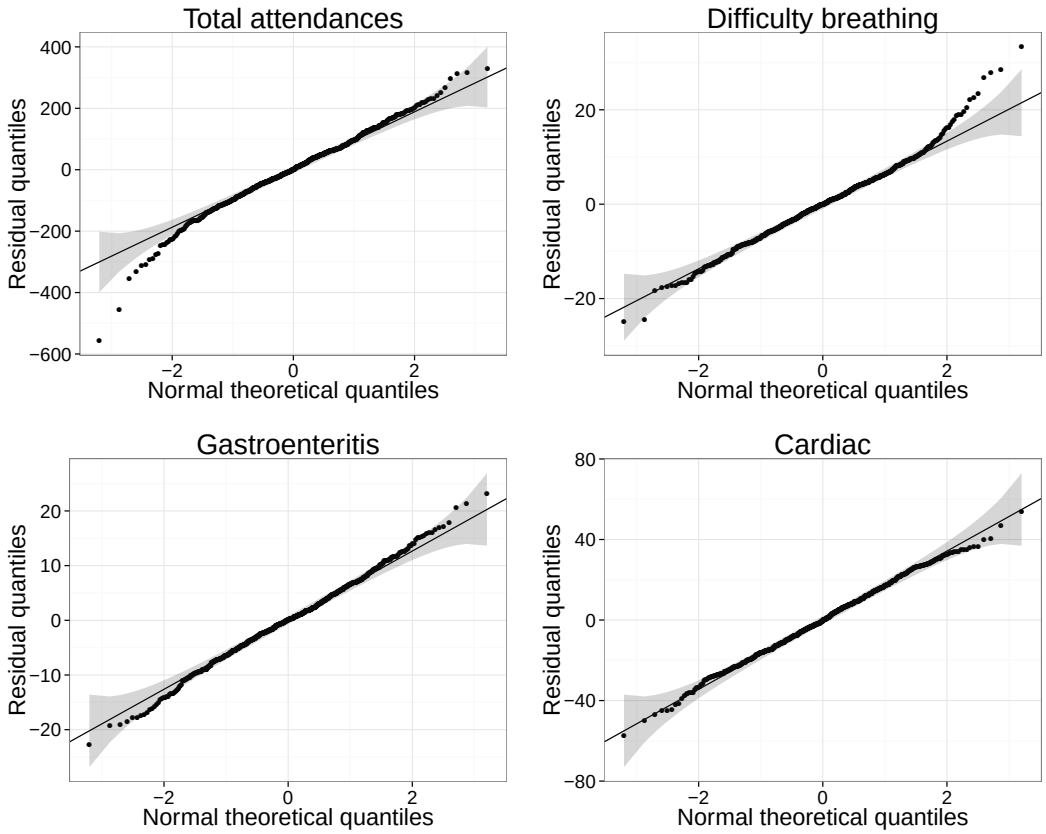


Figure 3.7: Normal Q-Q plots comparing the distribution of the residuals with the normal distribution. Reference line (black line) and 99% confidence region (grey) for total attendances and indicators from the ED SSS.

the other hand, makes an adjustment to account for these multiple comparisons.

The assumptions of Tukey's HSD test are the same as those required for the previous F-test (normally distributed with equal variances) and so there is no need to check these again.

To test the difference between days of the week d and d' using Tukey's HSD test we will construct a confidence interval on the difference of the means of the adjusted data on these days, $m_d - m_{d'}$, using the overall significance level of 95%. If this confidence interval does not span 0 this implies that there is a difference between the number of attendances at emergency departments on days d and d'. The confidence interval is constructed as

$$m_d - m_{d'} \pm q_{0.05;n-7,7}\sqrt{\frac{MSE}{n_w}}\ , \tag{3.3}$$

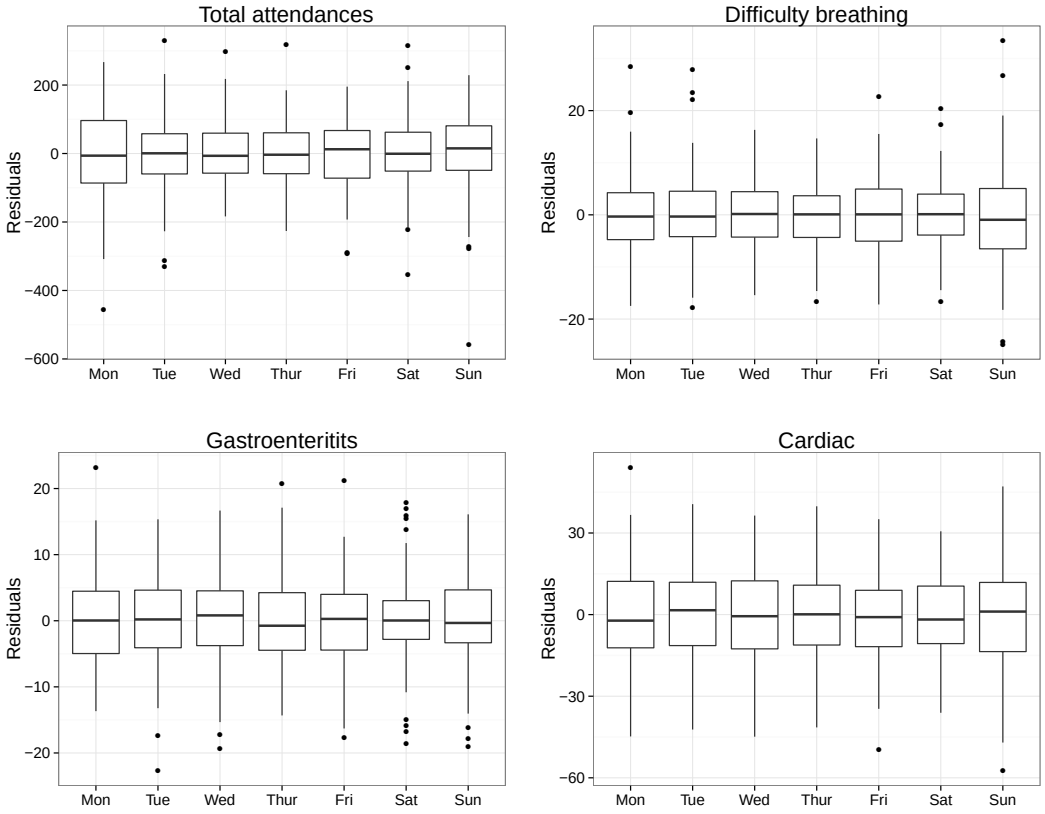


Figure 3.8: Box plots of the model residuals on each day of the week demonstrating similar variability for total attendances and indicators from the ED SSS.

where $q_{0.05;n-7,7}$ is the critical value of the *Studentised range* distribution with parameters $n-7$ and 7, and MSE is the mean squared error which is obtained by dividing the error sum of squares by the degrees of freedom [162, 163]. Notice that the width of the confidence interval is the same for each pair of days being compared.

Methods: Effect size (Cliff's delta)

Null hypothesis testing has received some criticism, in particular as a tiny, trivial difference can be found to be statistically significant if the sample size is large enough [160, 164, 165]. Therefore, we supplement our reporting of the results from statistical significance tests with further discussions of effect size which is a measure of practical significance. An effect size comments on the *size* of the difference between groups whereas a p value simply comments on the *existence* of a difference. With a large enough sample size, only a very small difference is needed for a statis-

Table 3.2: The ratio of the largest residual variance to the smallest residual variance across days of the week for total attendances and indicators from the ED SSS.

Indicator	Ratio
Total attendances	2.310
Difficulty breathing/wheeze/asthma	2.581
Gastroenteritis	1.459
Cardiac	1.597

Table 3.3: The results of the F-test for total attendances and each indicator from the ED SSS.

Indicator	F-statistic (df 6, 728)	p-value
Total attendances	315.60	< 0.0001
Difficulty breathing/wheeze/asthma	59.44	< 0.0001
Gastroenteritis	32.23	< 0.0001
Cardiac	285.70	< 0.0001

tically significant p value. Very small differences, even those that are statistically significant, are not often useful, meaningful, or important [165, 166].

For comparing days d and d' with mean number of attendances m_d and $m_{d'}$ respectively we report the difference $m_d - m_{d'}$, which is just a number of attendances so easy to interpret. We also attempt to compare our results with any day of the week effects already reported in the literature and to compare effect sizes between indicators and between syndromic surveillance systems by reporting Cliff's delta. This is a standardised, non-parametric effect size, introduced by statistician Norman Cliff (1996, [167]), that computes the degree to which samples overlap. In our context, Cliff's delta is computed by counting the number of times attendances on day d are larger than attendances on day d', and vice versa [168]. The calculation is as follows

$$\delta = \frac{\#(X_{id} > X_{jd'}) - \#(X_{id} < X_{jd'})}{n_w^2} \ .$$

Recall that n_w is the number of weeks in the dataset, and as such the number of data points for each day of the week.

We also, loosely, bear in mind guidance given to interpret the effect size from Cliff's delta as a small, medium, and large effect (table 3.4) [168].

There are parametric effect size statistics, such as Cohen's d, that would have been

Table 3.4: Cliff's delta small, medium, and large effect sizes, as reported by Romano et al. in [168].

Effect size	Cliff's delta
Small	$0.330 > \|\delta\| \geq 0.147$
Medium	$0.474 > \|\delta\| \geq 0.330$
Large	$\|\delta\| \geq 0.474$

suitable to report for this analysis (Cohen, 1992 [169]). However, in later analyses (section 3.2.3) we see that data from the other surveillance systems do not satisfy the assumptions of normality and homogeneous variances across the days of the week. Therefore, in order to be able to compare the sizes of the day of the week effects between syndromic surveillance systems we consistently report this non-parametric effect size throughout the chapter.

Results: Post-hoc tests (Tukey's HSD) & effect size (Cliff's delta)

Total attendances: The most notable significant day of the week effect identified in the total number of attendances at emergency departments was more attendances on Mondays compared to every other day of the week (figure 3.9). This had a very large effect size. Cliff's delta was always computed as 1, meaning every data point from a Monday was larger than every data point from the other days. On average, 312 more people per day attended emergency services on Mondays compared to Sundays, and on average at least 447 more people attended on Monday compared to the rest of the week. To put this into context, during a typical day in 2014 there would be around 7200 attendances recorded per day by the ED SSS. 312 attendances is 4.3% of a days attendances and 447 is 6.2%.

There were also more attendances than expected on Sundays compared to the remaining days of the week (not Monday). This also had a large effect size, with Cliff's delta computed as at least 0.644 and a difference of at least 135 attendances. Tuesdays, Wednesdays, Thursdays, Fridays, and Saturday were virtually indistinguishable (figure 3.9).

Difficulty breathing: The largest day of the week effect was similar to the day of the week effect in total attendances. There were more attendances at emergency departments for difficulty breathing on both Sundays and Mondays compared to the remaining days of the week (figure 3.10). This had a large effect size with Cliff's

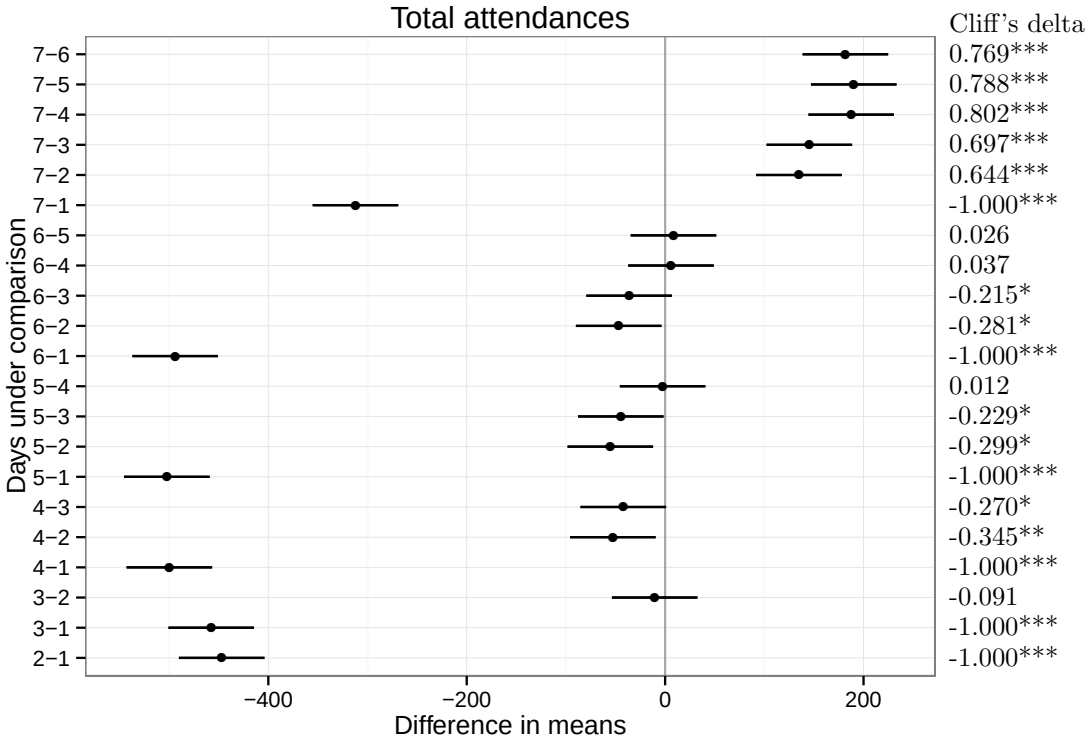


Figure 3.9: Pairwise comparisons of total attendances from the ED SSS. Each row gives the results from comparing a pair of days, where days are numbered 1 through 7 for Monday through Sunday. The difference between the means of the two days of the week is given by the black dot, with a 95% overall confidence interval from Tukey's HSD, along with Cliff's delta, a standardised effect size (where * is a small effect size, ** is a medium effect size, and *** is a large effect size).

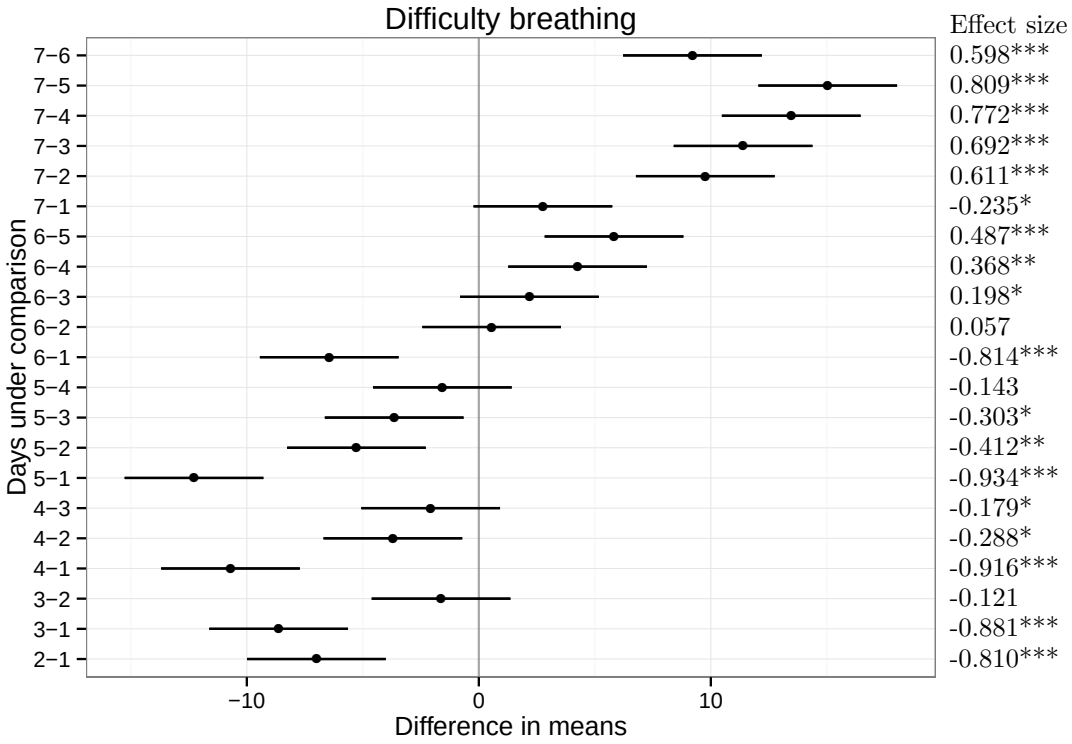


Figure 3.10: Pairwise comparisons of the difficulty breathing indicator from the ED SSS. Each row gives the results from comparing a pair of days, where days are numbered 1 through 7 for Monday through Sunday. The difference between the means of the two days of the week is given by the black dot, with a 95% overall confidence interval from Tukey's HSD, along with Cliff's delta, a standardised effect size (where * is a small effect size, ** is a medium effect size, and *** is a large effect size).

delta computed as at least 0.598 and a difference of at least 6 attendances. During a typical day in 2014 there would be around 70 attendances recorded to the difficulty breathing indicator and 6 is 8.5% of this.

There were also significant differences between the remaining days of the week. These can be split into two groups where the number of attendances within the group are not significantly different but the number of attendances between the groups are (almost always) significantly different. The first group is Thursday with Friday, and the second consists of Saturday, Tuesday and Wednesday. However, these effect sizes were a lot smaller.

Gastroenteritis: Sunday was identified as having significantly more emergency de-

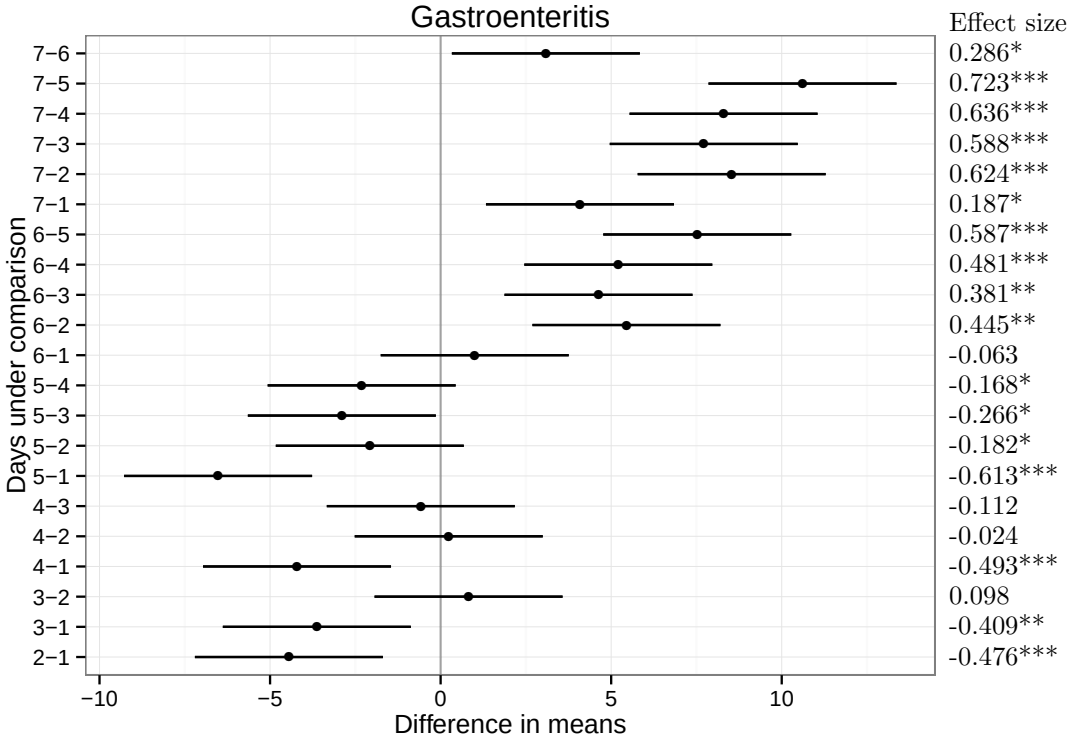


Figure 3.11: Pairwise comparisons of the gastroenteritis indicator from the ED SSS. Each row gives the results from comparing a pair of days, where days are numbered 1 through 7 for Monday through Sunday. The difference between the means of the two days of the week is given by the black dot, with a 95% overall confidence interval from Tukey's HSD, along with Cliff's delta, a standardised effect size (where * is a small effect size, ** is a medium effect size, and *** is a large effect size).

partment attendances for gastroenteritis than all other days of the week (figure 3.11). However, the differences between Sunday and Monday, and Sunday and Saturday were only small. The other effect sizes for Sunday comparisons were large, with a difference of at least 8 emergency department attendances. There are around 85 attendances per day coded to the gastroenteritis indicator, of which 8 attendances is 9.4%. Cliff's delta was at least 0.588.

Monday and Saturday had more attendances than the remaining days (not Sunday). These differences had medium to large effect sizes. Tuesdays, Wednesdays, Thursdays, and Fridays were virtually indistinguishable (figure 3.11).

Cardiac: Monday had significantly more emergency department attendances for

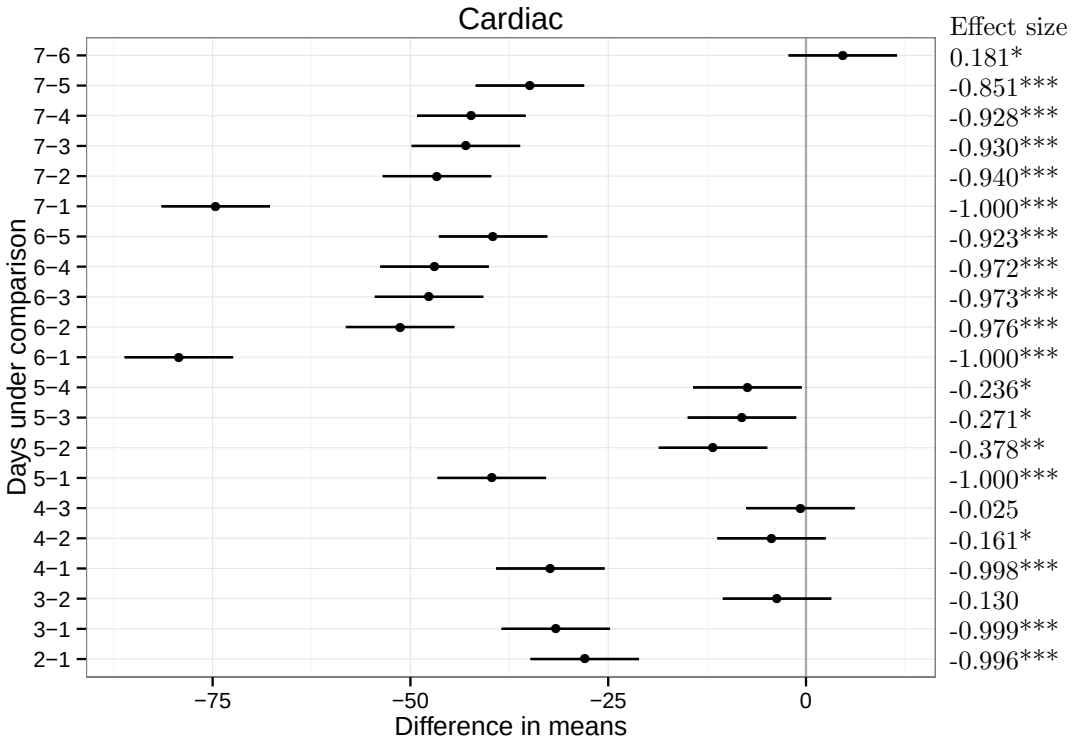


Figure 3.12: Pairwise comparisons from the cardiac indicator from the ED SSS. Each row gives the results from comparing a pair of days, where days are numbered 1 through 7 for Monday through Sunday. The difference between the means of the two days of the week is given by the black dot, with a 95% overall confidence interval from Tukey's HSD, along with Cliff's delta, a standardised effect size (where * is a small effect size, ** is a medium effect size, and *** is a large effect size).

cardiac than all other days of the week (figure 3.12). This had a very large effect size, corresponding to an additional 28 attendances at emergency departments and with Cliff's delta of at least 0.996. There were around 280 attendances in the cardiac indicator each day, of which 28 is 10%.

Interestingly, the other notable significant day of the week effect was fewer cardiac attendances on both Saturday and Sunday than the other days of the week (again with large effect sizes) and fewer attendances on Friday than the remaining days (Tuesday, Wednesday, Thursday) although this was a much smaller difference (figure 3.12). This is different to day of the week effects identified in the other indicators and in total attendances.

Overall: The largest day of the week effects were in the total number of atten-

dances. The day of the week effects in the cardiac indicator were also very large, and interestingly did not follow the same pattern as the total attendances.

The ED SSS day of the week effects: Comparisons with existing studies

There have been some previous reports of day of the week effects in emergency department attendances in North America. In particular, there are significant day of the week effects in total attendances and in a selection of conditions including gastrointestinal and respiratory indicators in the Indiana Public Health Emergency Surveillance System [143] and in attendances for asthma to emergency departments in Ontario [147]. In both these studies the authors report, similarly to us, that the largest proportion of the week's attendances were on Mondays and Sundays.

There has been quite a wealth of previous research on day of the week effects in cardiac problems. For example, a meta-analysis of excess cardiac mortality found that many studies report an increase of cardiac death on Monday, and that some report lower levels on weekends [150]. We also found higher number of attendances for the cardiac indicators on Monday and lower numbers of attendances on the weekends.

We would like to compare the effect sizes from our results with previously reported results in the literature. That is why we report Cohen's d, a standardised effect size. Unfortunately, we were not able to find reports in the literature that also gave effect sizes or even clearly gave standard deviations so that effect sizes could be subsequently calculated.

3.2.3 The GP out of hours, 111, and GP in-hours syndromic surveillance systems

Due to observing clear evidence of a weekend effect in the adjusted data from the GPOOH, 111, and GPIH syndromic surveillance systems (section 3.2.1) it is not necessary to do an omnibus test (F-test or similar) for overall day of the week effects in these systems. We do, however, look for further, less obvious day of the week effects.

Methods: Less obvious day of the week effects

We wish to identify whether there are any further, less obvious, day of the week effects beyond the weekend effect. In order to do this we will compare, pairwise, the days of the week using Cliff's delta as a standardised effect size. We will also report the difference between the mean number of attendances on each pair of days, with an error bar given by the pooled standard deviation of the two days, calculated for days d and d' as

$$s_{\text{pooled}} = \sqrt{\frac{s_d^2 + s_{d'}^2}{2}} \ .$$

We are unable to use Tukey's HSD test for pairwise comparisons for these systems as we observed many violations of the assumptions of normality and equal variances. For example, it is clear from figures 3.2 and 3.3 that the spread of the data from midweek days is smaller than the variance from weekend days in the GPOOH and 111 syndromic surveillance systems. In fact, if we consider the ratio of the largest variance to the smallest variance we get values as large as 13. Cliff's delta is a non-parametric measure of effect size that does not make assumptions of normality or homogeneous variances. In addition, we have decided not to transform the data to attempt to make it satisfy the assumptions of Tukey's HSD in order to keep the analysis relatively simple and easy to generalise across further syndromic indicators and future additional surveillance systems. This is discussed further in section 3.2.4.

Results: Less obvious day of the week effects

There are many figures of results from this analysis. To save space we have included one from each syndromic surveillance system here and the rest are contained in appendix A at the end of this thesis.

GPOOH SSS: The large weekend effect is clearly present in the total contacts and all indicators from the GPOOH SSS (figures 3.13, A.1 and A.2). Each weekend day has many more contacts than each day of the working week. These all have effect size 1 indicating that every Saturday and Sunday in each of these data sets had more contacts than every week day.

There was potentially an additional, smaller day of the week effect in the GPOOH SSS. Monday had more contacts than the other days of the working week. This had

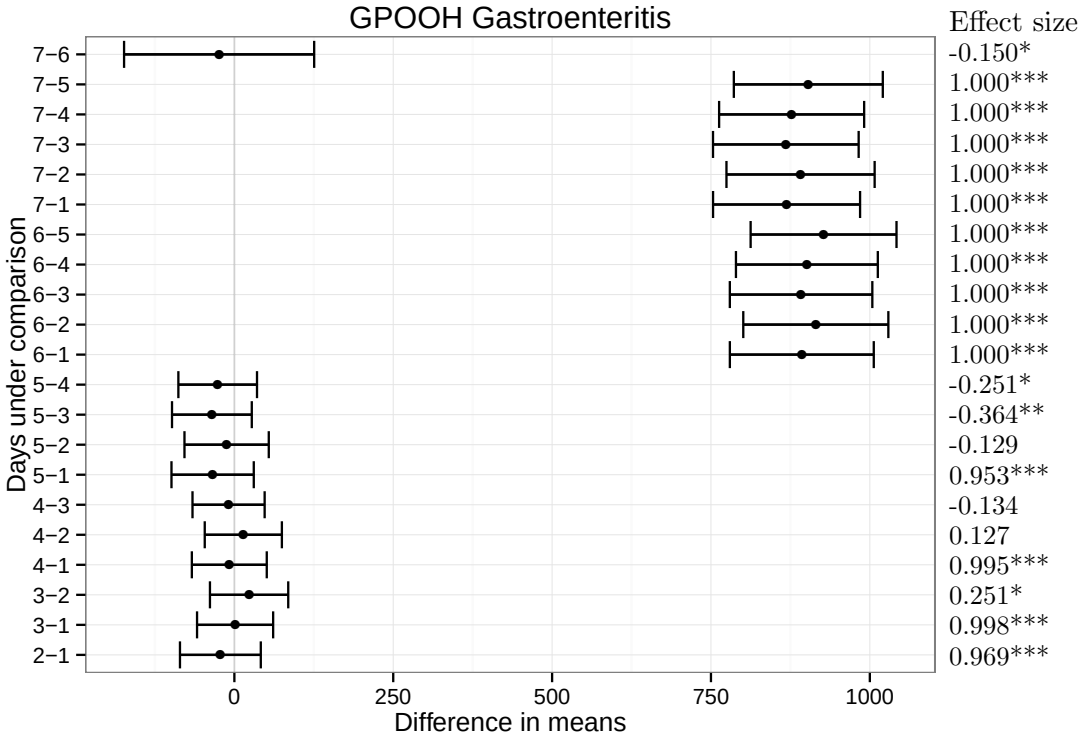


Figure 3.13: Each row gives the results from comparing a pair of days from the gastroenteritis indicator of the GPOOH SSS, where days are numbered 1 through 7 for Monday through Sunday. The difference between the means of the two days of the week is given by the black dot, with an error bar given by +/- one pooled standard deviation, along with Cliff's delta (where * is a small, ** a medium, and *** a large effect size).

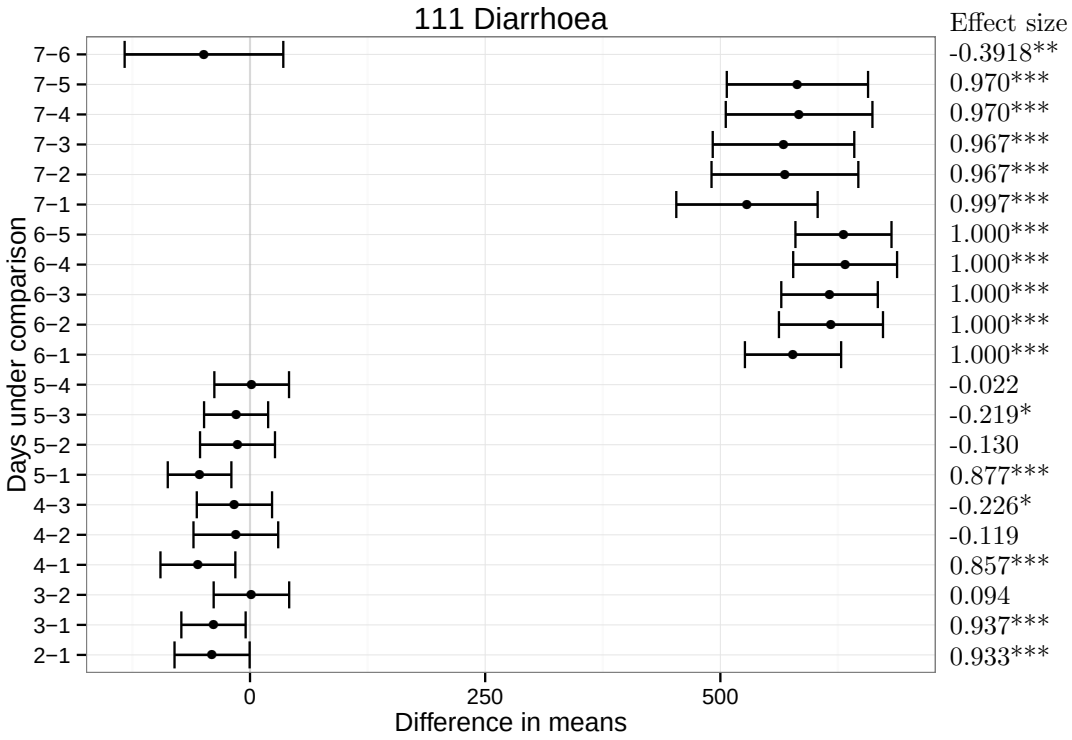


Figure 3.14: Each row gives the results from comparing a pair of days from the diarrhoea indicator of the 111 SSS, where days are numbered 1 through 7 for Monday through Sunday. The difference between the means of the two days of the week is given by the black dot, with an error bar given by +/- one pooled standard deviation, along with Cliff's delta (where * is a small, ** a medium, and *** a large effect size).

a large effect size, however the error bars sometimes overlapped zero particularly for the difficulty breathing and gastroenteritis indicators.

Finally, there were more total contacts on Saturday than there were on Sunday, but this effect was not present in the indicators (figure A.1)

111 SSS: The results from the 111 SSS were very similar to the results from the GPOOH SSS. The large weekend effect is clearly present in the total contacts and all indicators (figures 3.14 and A.3 to A.5). Each weekend day has many more contacts than each day of the working week and these all have large effect sizes. The effect sizes when comparing Saturday to the days of the working week were always 1. However, the effect sizes when comparing Sunday were slightly smaller at consistently less than 1.

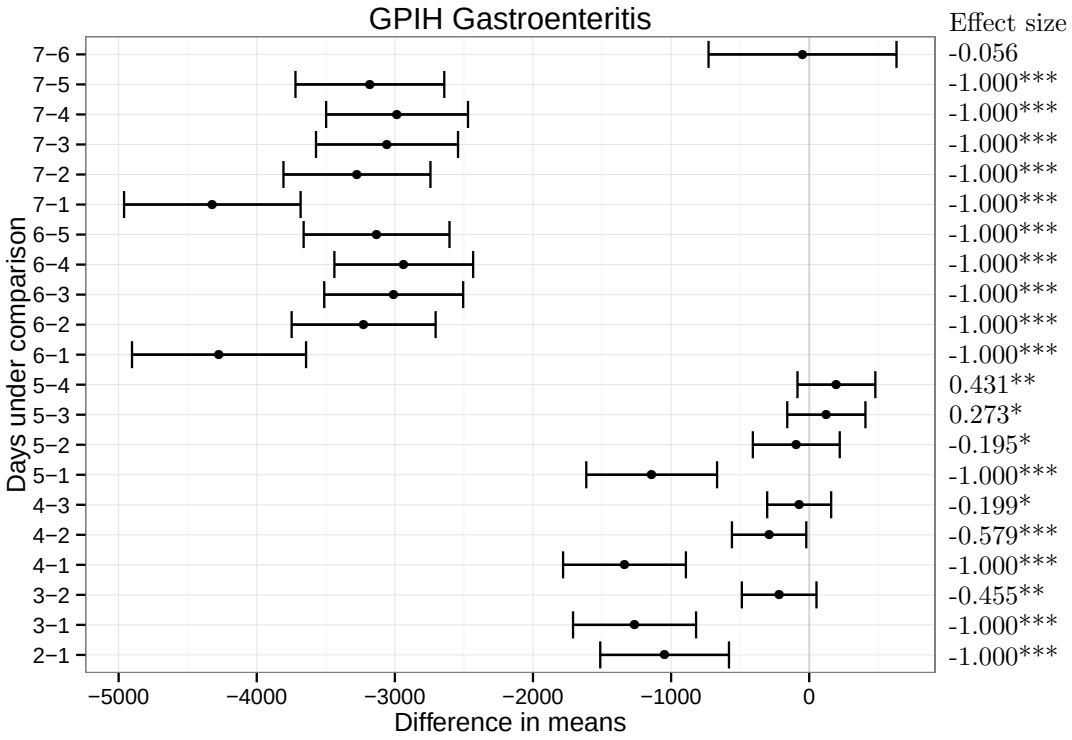


Figure 3.15: Each row gives the results from comparing a pair of days from the gastroenteritis indicator of the GPIH SSS, where days are numbered 1 through 7 for Monday through Sunday. The difference between the means of the two days of the week is given by the black dot, with an error bar given by +/- one pooled standard deviation, along with Cliff's delta (where * is a small, ** a medium, and *** a large effect size).

The same, smaller day of the week effects from the GPOOH SSS were observed in the 111 SSS. Monday had more contacts than the other days of the working week. This had a large effect size, however the error bars sometimes overlapped zero particularly for the difficulty breathing and gastroenteritis indicators. There were more total contacts on Saturday than there were on Sunday, but this effect was not present in the indicators (figure A.3)

GPIH SSS: The large weekend effect is clearly present in the GPIH SSS; each weekend day has far fewer attendances than each day of the working week for all indicators (figures 3.15, A.6 and A.7). This always had a very large effect size of 1.

There was a further clear day of the week effect in all indicators; there were more attendances on Monday than on any of the other day of the working week. This

also, almost always, had an effect size of 1.

Finally, there were more attendances on Tuesday than on Wednesday and Thursday. However this effect was not as large with effect sizes varying between 0.393 and 0.579.

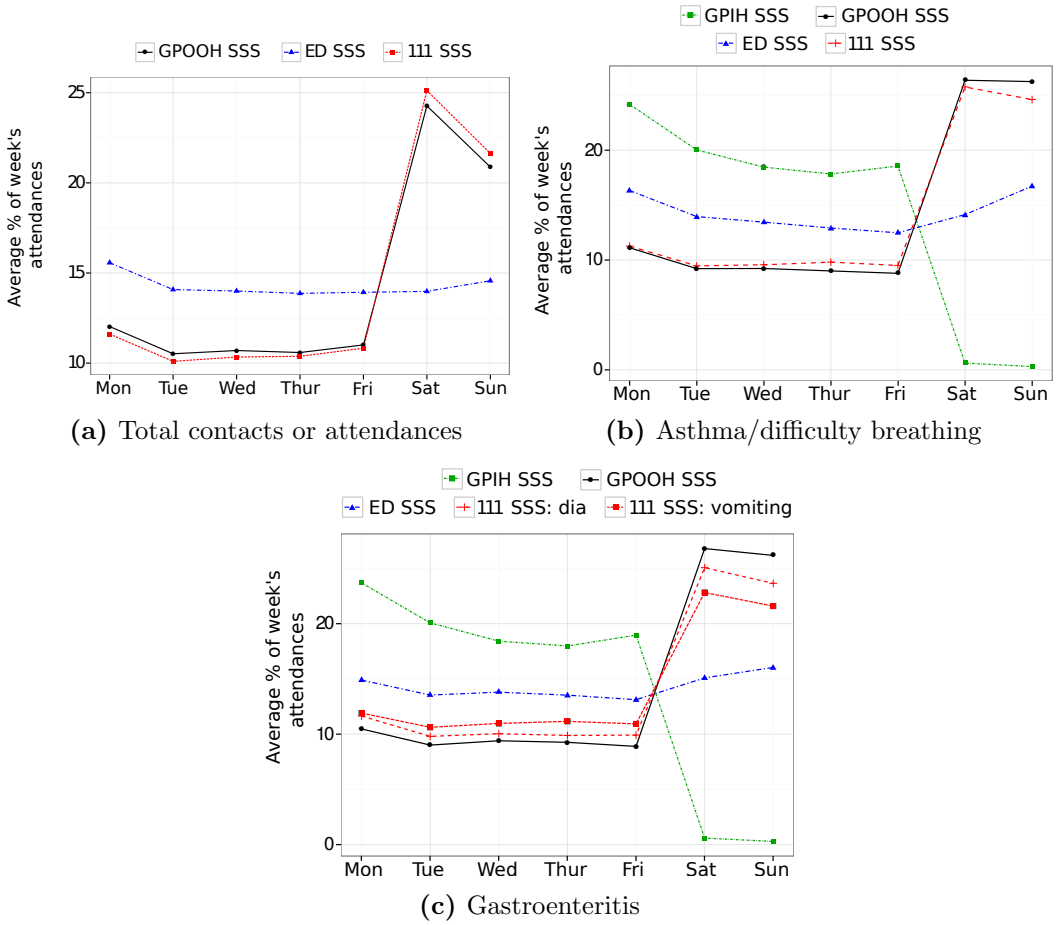
3.2.4 Day of the week effects: Discussion and conclusions

In summary, we found evidence of day of the week effects in each of the syndromic surveillance systems operated by PHE. Most of these were obvious effects which were to be expected due to the availability and purpose of the different healthcare services. However, this is the first formal description of these.

In the GPOOH, 111, and GPIH syndromic surveillance systems the day of the week effects were consistent across the selection of indicators studied. However, they were not consistent across the ED SSS. As expected, the day of the week effects in the GPOOH and 111 systems complemented the effects in the GPIH system (figure 3.16). Finally, the effects in the ED SSS are much smaller than in the other systems (figure 3.16). This is also, perhaps, to be expected as this healthcare service is designed to be available at all times.

The types of statistical tests applied here are used quite frequently to identify day of the week effects. However, there are a multitude of other methods that could have been used. In particular, formal non-parametric hypothesis tests or a transformation of the data could have been used when non-normality and heterogeneous variances were observed and time-series models, such as an autoregressive integrated moving average (ARIMA) model, could have been fitted to identify seven day periodicities. However, the approach we have taken appears to have been successful, is relatively simple, and is easy to apply to any syndromic indicator with daily data thus making it easy for public health professionals to independently carry out the same analysis on other datasets. Finally, access to these data was only available while the author was on a secondment at the ReSST, and therefore further statistical testing now is not easily possible.

During this investigation we assumed that day of the week effects remain constant throughout time. The only concession to this assumption is that we have removed weeks containing public holidays from the analysis as we suspect the impact of public holidays may interfere with day of the week effects. An investigation into whether the effects described here change with time is left to future work. In particular, it



(a) Total contacts or attendances

(b) Asthma/difficulty breathing

(c) Gastroenteritis

Figure 3.16: The adjusted data for total contacts or attendances, the difficulty breathing indicator, and the gastroenteritis indicator so that the different syndromic surveillance systems can be easily compared.

may be interesting to see if seasonal disease outbreaks impact on the size of day of the week effects.

3.3 Public holiday effects

Common sense suggests that the way people behave and report illness will be atypical on public holidays; some public services will be closed on this day and many people have the day off work. In this section we will investigate the magnitude of this change on public holidays. We will also investigate whether the public holiday effect changes the number of contacts or attendances on days other than the public holiday itself. We will do this by looking at the contacts or attendaces on each day surrounding a public holiday and comparing it with typical examples of those days.

3.3.1 Public holiday effects: Methods

We need to confirm and quantify the difference in the average number of contacts or attendances with each healthcare service on public holidays compared to the average number on a typical day. We also want to see whether the effect is the same for each public holiday. And finally, we want to see if the public holiday effect is solely restricted to the public holiday itself. However, we need to use methods that also take into account the day of the week effects that we now know about.

We used the non-public holiday weeks neighbouring a public holiday week to give an estimate of the expected number of attendances during each day of the public holiday week if it had not contained any public holidays. Based on this we could identify any unusual effects during the public holiday weeks.

In detail, we first split the data into *typical* and *public holiday* weeks. A public holiday week was defined as any including a public holiday or immediately preceding a week beginning with a public holiday. Note that this choice was made so that all final, and first, working days before, and after, a public holiday were in public holiday weeks. In England, all public holidays on a Friday are immediately followed by a week containing more public holidays. All other weeks were defined as typical weeks. We obtained an expected total number of contacts or attendances for each public holiday week by taking a linear interpolation between the total contacts or attendances in the previous and the next typical weeks. This gives an estimate of the number of contacts or attendances that would have occurred during the public holiday week had it been a typical one. Mathematically, consider public holiday week i, surrounded by typical weeks i^- and i^+. Note that i^- and i^+ may not be adjacent to week i as there can be consecutive public holiday weeks. Week i consists

of data points $X_{i_1}, ..., X_{i_7}$ with total $T_i = \sum_{j=1}^{7} X_{i_j}$. The expected total number of contacts or attendances had week i not contained a public holiday, $\overline{T}_i$, was computed as $\overline{T}_i = \frac{T_{i^+} - T_{i^-}}{i^+ - i^-}(i - i^-) + T_{i^-}$.

Next, we split the estimate $\overline{T}_i$ of the expected number of contacts or attendances in public holiday week i had it not contained a public holiday into estimates for each day within week i. We also compute tolerance intervals on these estimates. If the data on any day of the public holiday week fall outside the tolerance interval we consider this day to be unusual. In this way we can measure any unusual effects on public holidays themselves and the days surrounding them. We can also separately look at the different types of holidays (the Christmas period, the Easter weekend, and single public holiday Mondays).

In order to construct the tolerance interval, we transformed the raw data from every typical week into a daily percentage of the week's total contacts or attendances. Mathematically, for time series $X = \{X_t\}$ this gives the adjusted time series $\tilde{X} = \{\tilde{X}_t\}$ computed as

$$\tilde{X}_t = 100 \frac{X_t}{\sum_{i \in W_t} X_i} ,$$

where W_t is the week (Monday to Sunday) containing day t.

For each day of the week, the collection of these adjusted data points forms a sample from the true distribution of the percentage of the week's contacts or attendances on that day.

Based on these samples we construct $(95\%, 99\%)$ tolerance intervals for the percentage of the week's contacts or attendances on each day of the week. This interval tells us, with 99% confidence, that data for at least 95% of weeks will fall within this range [170]. Based on the estimates $\overline{T}_i$ we transform the tolerance intervals from percentages to numbers for each day of each public holiday week.

We use a non-parametric tolerance interval in order to make minimal assumptions about the samples of adjusted data for each day of the week. The literature warns that slight non-normality, and particularly skewness, can cause tolerance intervals based on the normal distribution to "*give very erroneous results*" [171]. Non-parametric tolerance intervals are usually wider than parametric tolerance intervals [170]. Therefore, our identification of days with an atypical number of contacts or attendances will be quite conservative.

We use the R function `nptol.int` from the `tolerance` package with the Wilks method to compute the tolerance intervals [171, 172].

In order to easily compare across the different systems and indicators, we compute an exceedance score for each day of the public holiday weeks. This is the difference between the data and the expected value as a proportion of the size of the prediction interval. In detail, the exceedance score is the ratio of two differences: the difference between the data and the expected number of contacts or attendances and the difference between the maximum of the tolerance interval and the expected number of contacts or attendances [28]. A value larger than 1 indicates that there were more contacts or attendances on a public holiday than on a typical day and a value smaller than -1 indicates there were fewer contacts or attendances on a public holiday than on a typical day.

3.3.2 Public holiday effects: Results

Public holiday days: For the GPOOH, 111, and GPIH syndromic surveillance systems the number of contacts or attendances on the public holiday days themselves was always outside the tolerance interval (figures 3.17 to 3.19). For the GPOOH and 111 SSS the data always exceeded the tolerance interval and for the GPIH SSS were always below. This is what we expect based on typical GP opening hours on public holidays and the expectation that out of hours services receive patients that would otherwise attend the GP. The amount by which the data was outside the tolerance interval differed by system and by indicator, but also by holiday type within systems. The exceedance in the GPOOH system was larger than the exceedance in the 111 and GPIH systems. In the GPOOH system the exceedances were larger for the gastroenteritis indicator than for the total contacts and difficulty breathing indicator. The total contacts and indicators were similar within the 111 system and within the GPIH system. For the GPOOH and 111 syndromic surveillance systems the exceedance was always smallest on Christmas Day compared to all other public holidays. The size (absolute value) of the exceedance in the GPIH system was similar for each type of public holiday, but slightly larger for those in the Christmas period.

The only public holiday on which there was a public holiday effect in ED attendances was Christmas Day; the number of total ED attendances on Christmas Day was consistently below the tolerance interval (figure 3.20). However, this was not the

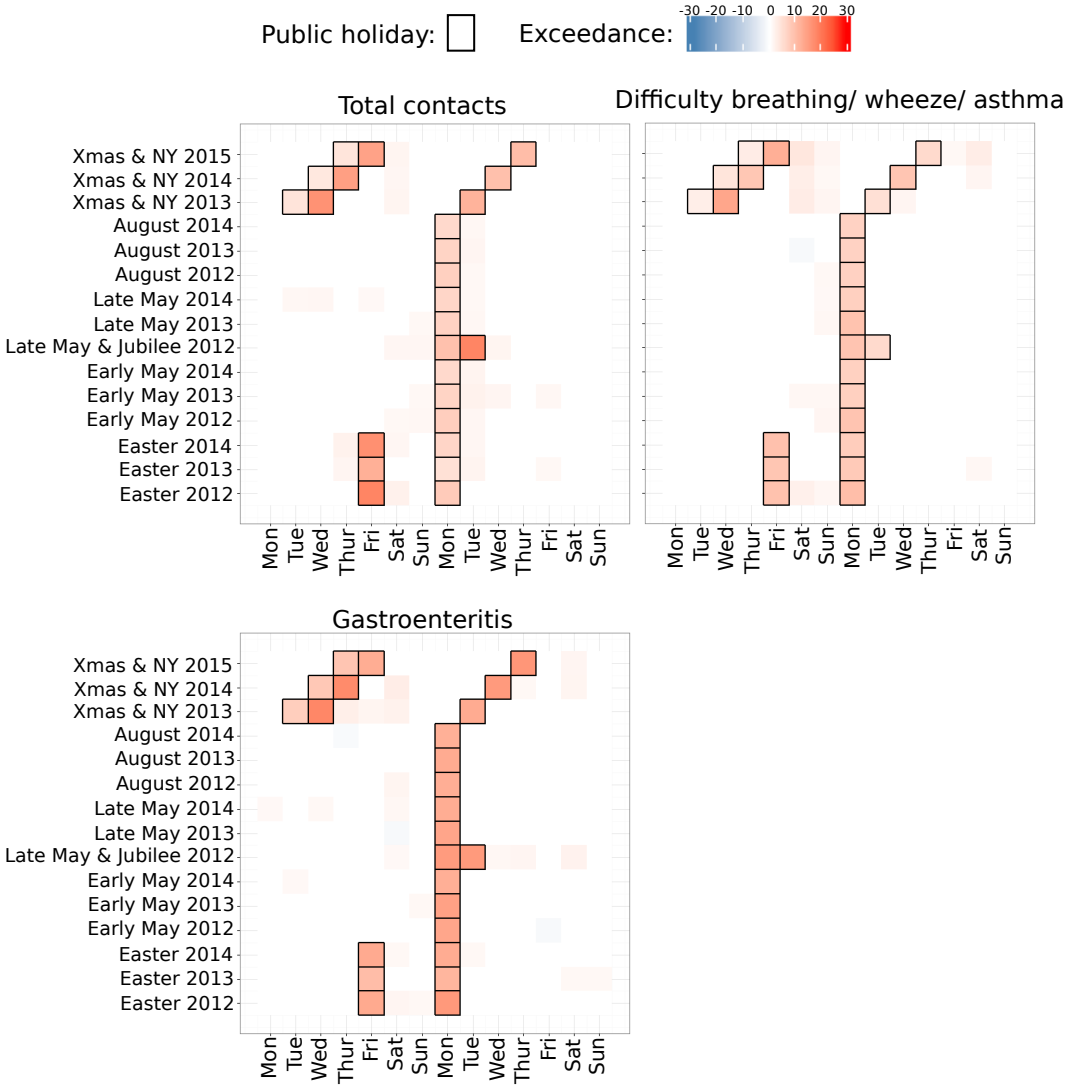


Figure 3.17: The exceedance score for each day of each public holiday week for the GPOOH SSS. A red square shows that the number of contacts was above the tolerance interval and a blue square shows that the number of contacts was below the tolerance interval. Public holiday days are marked by black boxes.

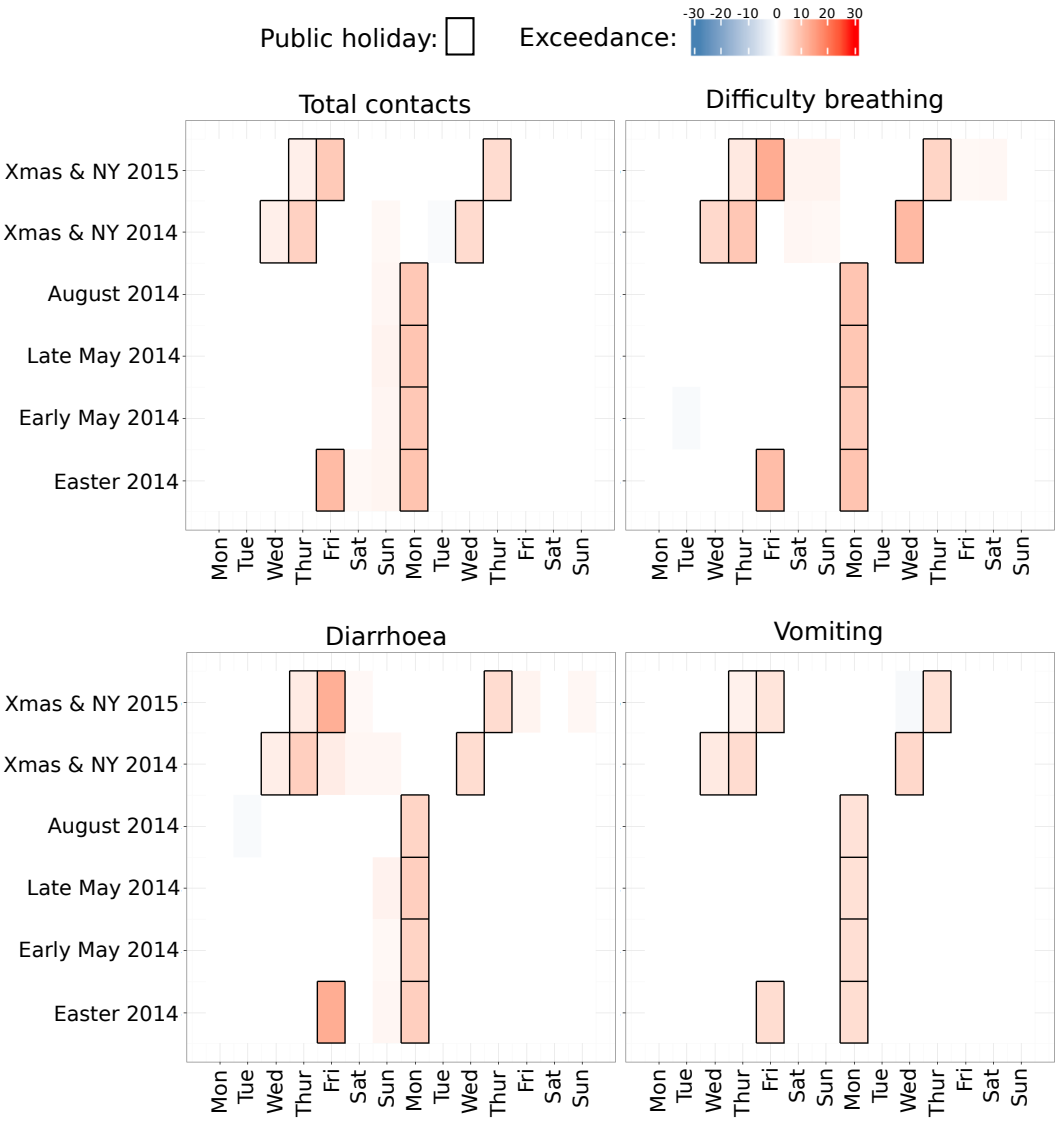


Figure 3.18: The exceedance score for each day of each public holiday week for the 111 SSS. A red square shows that the number of contacts was above the tolerance interval and a blue square shows that the number of contacts was below the tolerance interval. Public holiday days are marked by black boxes.

case for the asthma indicator (where the attendances on Christmas Day was within the tolerance interval) or the gastroenteritis indicator (where the attendances on Christmas Day was above the tolerance interval on two of the three years). For the cardiac indicator, the number of attendances on Christmas Day and most Monday public holidays was below the tolerance interval.

Extended public holiday effects: Within the GPOOH SSS there were further exceedances of the tolerance interval on the days surrounding some holidays (figure 3.17). The total number of contacts on the Saturday between Christmas and New Year consistently exceeded the tolerance interval, and so did the Saturday in most 4 day weekends. However for the difficulty breathing indicator, this was only the case for the weekend between Christmas and New Year. This effect was not seen consistently for any public holiday weekend in the gastroenteritis indicator. Additionally, the total number of contacts on the first working day after a Monday holiday consistently exceeded the tolerance interval. However this was not consistently seen for the difficulty breathing and gastroenteritis indicators.

Within the 111 SSS, the Sunday immediately before a Monday public holiday consistently exceeded the tolerance interval for the total number of contacts and the number of contacts for diarrhoea, and was very close to the upper end of the tolerance interval for the difficulty breathing indicator (figure 3.18). However, there was no consistent increase at this time in the vomiting indicator.

Within the GPIH SSS there was a clear effect on the first working day after a Monday holiday (figure 3.19). For all Monday holidays (including Easter) for the asthma and herpes zoster indicators the number of attendances on Tuesday was above the tolerance interval. For the gastroenteritis indicator this was the case for most Monday holidays. The number of attendances on the Tuesday was often in line with the tolerance interval for the Monday (the first day of a typical working week). However, after most Easters it exceeded the Monday tolerance interval as well. The results for the Christmas period were slightly different. There was not always a large number of attendances on the day after the public holidays, particularly when the Christmas Day and Boxing Day holidays were towards the end of the week.

Within the emergency department SSS there were some impacts of the holidays on the days surrounding them (figure 3.20). The total number of attendances on both Christmas Eve and New Years Eve was below the tolerance interval. In contrast, the number of attendances on New Years Day was always above, or very close to,

the maximum of the tolerance interval. However, this effect was not seen for any of the indicators studied where we did not identify any consistent extended public holiday effects.

3.3.3 Public holiday effects: Discussion and conclusions

Strengths and weaknesses

The methods used give us the flexibility to compare all public holidays separately and, as such, identify different effects due to different types of public holidays. Additionally, and importantly, this method takes into account the day of the week effects previously described.

However, we assume that the number of contacts or attendances during the neighbouring weeks of a public holiday week is a good indication of the number of contacts or attendances during the public holiday week itself; we assume that there are smooth changes over the time period of a few weeks.

We use a conservative non-parametric tolerance interval so there may be additional smaller effects that we have not been able to identify at this stage. However, this non-parametric tolerance interval makes few assumptions of the data.

Finally, we did not undertake a computation of the sample size requirements in order to achieve the level of precision we desired for the tolerance interval, which a rigorous statistical analysis could. However, with sample sizes of at least 67 weeks, and over 150 weeks for some SSSs, we are still reasonably confident in our conclusions. Although this calculation would have been nice to do, we do not feel it is crucial for delivering the message about the public holiday effects we described, in particular as we focussed on those with large exceedance scores.

Comparisons with existing studies

There are some existing studies that comment on public holiday effects in emergency department visits. These studies report, however, conflicting evidence about the use of emergency departments on these days. Some report more attendances on public holidays [173, 174] and some report fewer attendances on public holidays [175]. We have not been able to find comparable studies on GP, out of hours, and teleheath

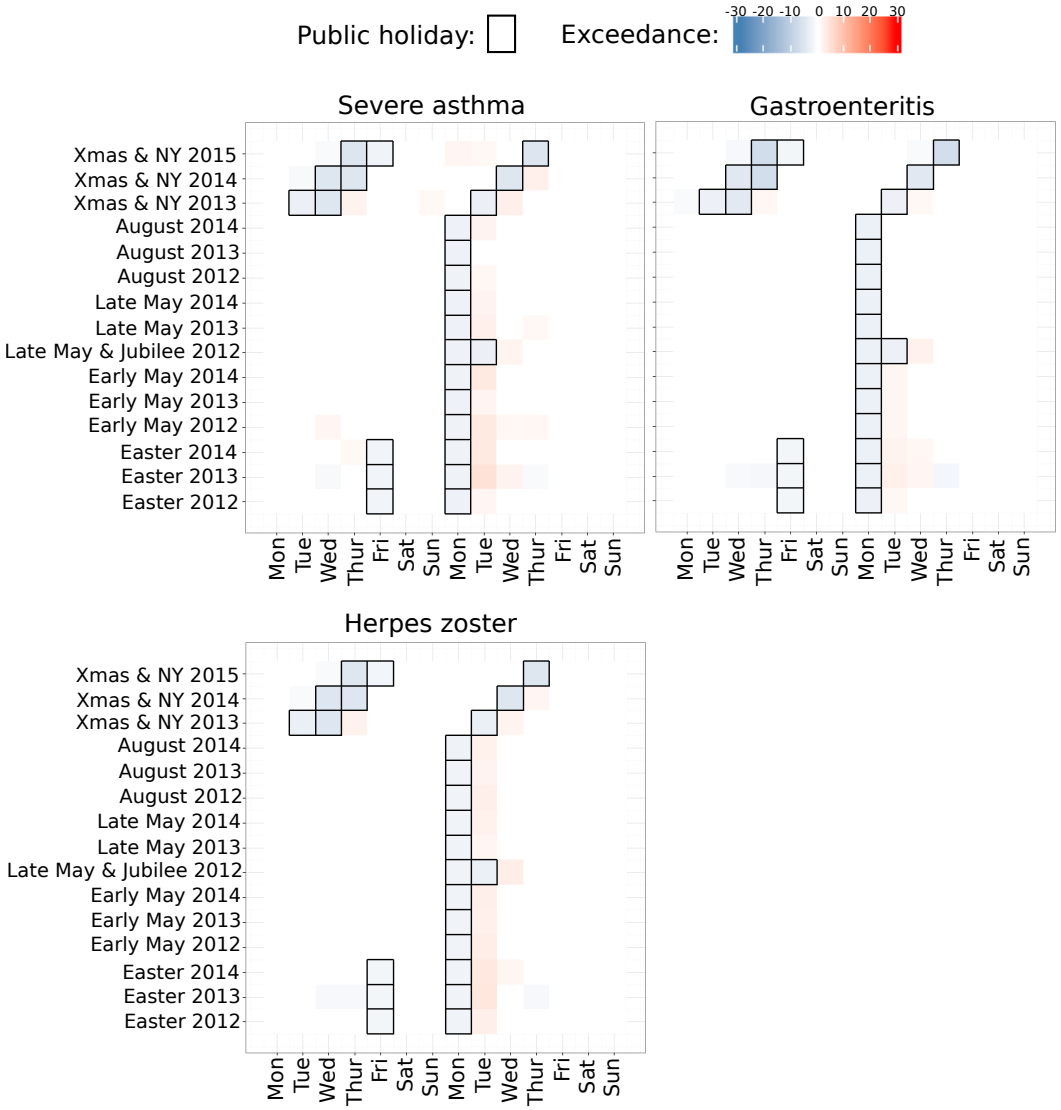


Figure 3.19: The exceedance score for each day of each public holiday week for the GPIH SSS. A red square shows that the number of attendances was above the tolerance interval and a blue square shows that the number of attendances was below the tolerance interval. Public holiday days are marked by black boxes.

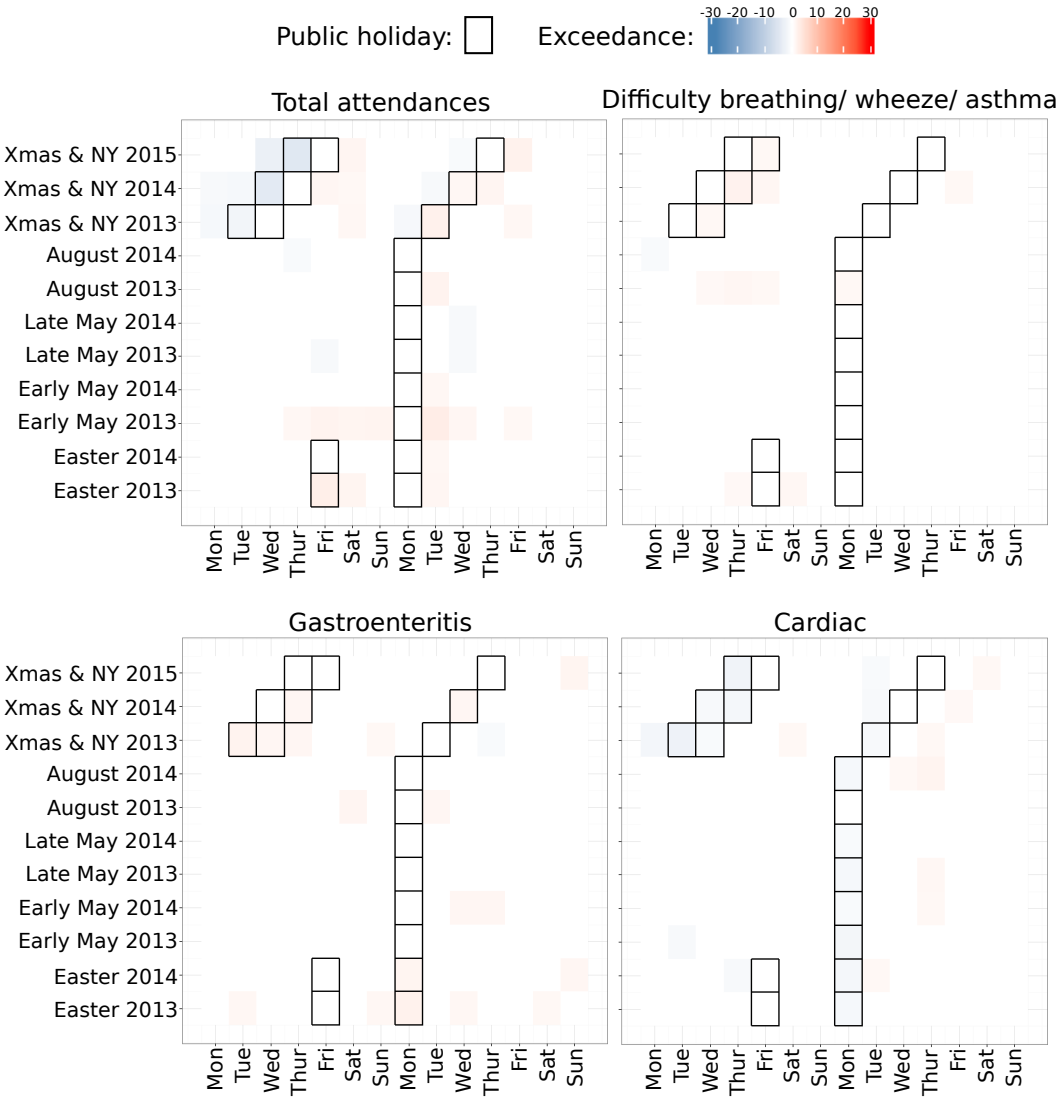


Figure 3.20: The exceedance score for each day of each public holiday week for the ED SSS. A red square shows that the number of attendances was above the tolerance interval and a blue square shows that the number of attendances was below the tolerance interval. Public holiday days are marked by black boxes.

use during public holiday periods.

A study of emergency department usage in Hong Kong in 2000 [176] reported that the second most common reason for patients to visit emergency departments with conditions that could be treated by GPs was feeling sick on a public holiday. We suspect that the closure of GP services on a public holiday drives some of these public holiday effects that we have observed.

Conclusions

In addition to the obvious public holiday effects, suspected to be driven by availability of GP services, we found additional public holiday effects in the syndromic datasets. However, these were smaller and less consistent. In particular, we identified a public holiday effect in the GPIH SSS on the first working day after a public holiday. In the GPOOH and 111 SSS, in addition to increased contacts at the weekends due to the day of the week effects previously described, there were even more contacts than expected on weekends adjacent to public holidays. We observed that not all public holidays were the same; often the public holiday effects at Christmas were different. Finally, we did not find large differences between reports of gastroenteritis to healthcare services over public holiday periods and general reports of poor health.

3.4 Putting knowledge into action

In this section we suggest potential changes that can be made to the current working practices of the ReSST at PHE as a result of this investigation into the day of the week and public holiday effects. The two areas we suggest changes in are the statistical method for detecting unusual activity levels (number of contacts or attendances) and the smoothing method used in presentations of the daily syndromic data from the GPIH SSS. The improved smoothing method we describe is now in use by the ReSST.

3.4.1 Improving statistical regression methods to detect unusual activity

The purpose of syndromic surveillance is to identify abnormally elevated disease levels as early as possible so that action can be taken to minimise the problem [177]. Statistical methodologies are used in order to identify whether the current activity levels are unusual. A range of methods have been developed, some of which were described in section 1.2 and reviewed by Unkel et al. (2012, [178]).

As previously described (section 1.2), the RAMMIE method was developed by the ReSST to detect unusual activity in all their syndromic surveillance systems [30].

The day of the week and whether the current day is a public holiday are included in the RAMMIE model as fixed effects. Additionally, whether it is the day after a public holiday is also included when RAMMIE is used with the GPIH SSS. The RAMMIE method was developed before the statistical analysis of day of the week and public holiday effects described in this chapter took place. Therefore, these variables were included based on anecdotal evidence and retained based on significant p-values from the regression.

We suggest here some potential improvements that could be made to the RAMMIE method based on the results obtained from our day of the week and public holiday effect investigation.

Suggested improvement: Christmas day

All public holidays are treated in the same way by the RAMMIE method; Morbey et al. (2015, [30]) state that a fixed effect (dummy variable) is included for "whether or not the day was a bank holiday." Historical data from all previous public holidays are used to give the predicted activity level for a future public holiday.

However, in section 3.3.2 we demonstrated that Christmas Day shows a smaller increase in activity than the other public holidays for the GPOOH and 111 SSS. As all previous public holidays will contribute to the expected activity on the next public holiday, the inclusion of Christmas Day in this calculation will give a lower estimate than if it were not included. This could be leading to false alarms on public holidays which will waste investigator's time. Additionally, as all the other public holidays are used to give the estimated syndromic activity baseline for Christmas

Day, this will be higher than it should be. This could be causing unusual increases on Christmas Day to be missed.

Consistent public holiday effects were only identified in total emergency department attendances on Christmas Day, but not on any other public holiday (section 3.3.2). The current RAMMIE model treats all public holidays the same and will therefore over-estimate the baseline on Christmas Day, potentially leading to missed alarms.

Therefore, we argue that it may be better to not treat all public holidays in the same way. To avoid these false or missed alarms an additional independent variable could be included in the RAMMIE model for Christmas Day. However, there would be minimal data to fit this variable to.

Suggested improvement: After a public holiday in the GPIH SSS

When used with the GPIH SSS, the RAMMIE method includes a fixed effect (dummy variable) for the day after a public holiday. We demonstrated, in section 3.3.2, that there was a higher than usual number of attendances on the first working day after a public holiday Monday but no consistent effect after public holidays on other days of the week (Christmas and New Year). Therefore, the inclusion of the 'day after a public holiday' variable could be leading to missed alarms on the days after the Christmas and New Year holidays, and false alarms on the day after a Monday public holiday. To avoid these false or missed alarms, the 'day after a public holiday' variable could be adjusted to only apply to the first working day after a Monday public holiday. The first working day after Christmas and New Year public holidays would be treated as a standard day.

RAMMIE improvements: Discussion

The RAMMIE method was developed with day of the week and public holiday effects already in mind. Our analysis confirms that this was necessary. However, we suggest some refinements relating to more subtle public holiday effects that could potentially improve the RAMMIE method.

However, if the days of the year are subset into smaller groups then there will be fewer historical data points to contribute to predictions. Therefore, these suggested changes may only be beneficial when there is a large amount of historical data.

These suggestions could be tested by comparing the original RAMMIE model with an adapted model using model selection techniques. The RAMMIE method was validated by reporting specificity, sensitivity, positive predictive value, and timeliness against known historic incidents [30]. These same measures could be computed for an adapted model using the same historic data and compared. This is beyond the scope of this thesis but, as far as we are aware, is something that the ReSST plans to investigate as a result of this work.

3.4.2 Improving trend identification and data visualisations

Line graphs of time series data offer a simple and effective way to review data and undertake exploratory analysis [179, 180]. They are used, in addition to automated statistical alarms, by the ReSST to investigate, interpret, and display the current trends in syndromic data and for comparisons of the current data with previous years to identify changes from the norm. Regular, large fluctuations at small time-scales can, however, make it difficult to identify longer time-period trends in time series graphs. These difficulties can be overcome by adding to the graph a smooth trend curve which takes into account these known day-to-day fluctuations [181]. Therefore, in order to mitigate for any difficulties caused by the previously described day of the week and public holiday effects, the ReSST add smooth trend curves to graphs of syndromic data.

These graphs of syndromic data are used in two main ways. Firstly, although much of the syndromic surveillance systems are automated, statistical alarms are created that require manual, in-depth investigation [30]. Effective data visualisations should be used in order for the manual investigation stage not to become the bottle-neck of the real-time data analysis process [182]. Secondly, graphs of the syndromic indicators are presented to the public and wider audiences in weekly bulletins [183]. Therefore, it is important that the current trend in illness levels can be clearly interpreted from the graph without additional data or expert knowledge.

Based on the knowledge we have obtained on day of the week and public holiday effects in syndromic surveillance systems, we have developed the *extended working day moving average* smoothing method to display trends in syndromic indicators from the GPIH SSS. This smoothing method takes the expected day of the week and public holiday effects into account simultaneously and displays no trend due to these predictable variations. It has been applied to time series graphs to enhance

visual analysis of daily GP attendance data for syndromic surveillance.

We developed this method for the GPIH SSS as we observed that the previous smoothing method in use at PHE was unable to successfully account for the observed public holiday effects. This is the only healthcare system monitored by PHE that has a five-day working week. Therefore, the smoothing method used was unique to this system. The smoothing method we have developed is also unique to this system. However, its development demonstrates that both day of the week and public holiday effects must be considered simultaneously when smoothing data with calendar effects. This motivates careful further work by the ReSST to ensure that the smooth trend curves used with other surveillance systems are also adequate. If the systematic changes in the number of contacts with or attendances at healthcare systems due to day of the week and public holidays are not accounted for, they could mask real increases in disease levels, create false alarms, and delay decision making over public holiday periods as more data are required to understand the current trend. It is important to try and distinguish the expected changes in attendances due to day of the week or public holiday effects from unexpected changes due to potential public health threats.

This section will continue as follows. We will first discuss the existing smoothing methods to account for day of the week and public holiday effects in healthcare data. We will then describe the extended working day moving average. We will describe the data that we will use to demonstrate the method. Then we will present an evaluation of the extended working day moving average, with comparison to the more simple smoothing methods. Finally, the strengths and limitations of the extended working day moving average and the impact on public health practice will be discussed.

Existing smoothing methods for day of the week and public holiday effects

Smoothing to remove day of the week effects and visualise trends has been noted as being important for analysis of healthcare data [184–188], although few smoothing methodologies have specifically been developed to enhance visual interpretations in this context. However, there are examples of both model-based and data-driven smoothing methods that remove day of the week and/or public holiday effects as part of more complex detection algorithms [189]. Many published methodologies smooth day of the week effects but do not also consider public holiday effects [143, 186, 189].

However, we will demonstrate that both day of the week and public holiday effects should be considered simultaneously to enable continued, effective surveillance of GP attendance data during and around public holidays.

A seven-day moving average is the simplest data-driven smoothing approach to remove a day of the week effect. No adjustment is made for public holiday effects in this method. A moving average is a series of averages of subsets of the time series of syndromic data. The first element of a seven-day moving average is the average of the first seven data points. The second element is the average of the second to eighth data point. This is continued so that each set of seven consecutive data points is averaged. Seven days was chosen in this context as day of the week effects have seven-day periodicity. We will use the seven-day moving average as a simple comparison to our newly developed smoothing method.

The working day moving average method was previously developed by PHE to take both day of the week and public holiday effects into account when visualising data from the GPIH syndromic surveillance system.

The working day moving average is constructed as follows. Very few routine in-hours GP attendances occur on public holidays. Therefore, public holidays are grouped with weekend days, and a moving average is computed that takes into account the number of working days. Let n denote the number of working days within the current block of seven days being considered to give an element of the moving average. In the GPIH SSS this is typically five. However, in blocks containing public holidays it will be fewer. Instead of simply computing the average of the number of attendances on the seven days, the sum of the number of attendances on working days was multiplied by $\frac{1}{5}$ and the sum of the number of attendances on non-working days was multiplied by $\frac{2}{7-n}$. The sum of these totals was then divided by five, the typical number of working days in the GPIH SSS.

For a block of seven days with no public holidays, this calculation just gives $\frac{1}{5}$ times the sum of the number of attendances on the seven days in question: a basic moving average. For blocks of seven days containing public holidays this calculation weights the working days slightly more than the simple sum and the non-working days slightly less. This is to account for the expected reduction in total attendances in the week due to the public holiday.

The extended working day moving average

Data from healthcare services reflect the time at which patients sought healthcare advice. This does not necessarily correspond with date of symptom onset. In particular, patients with milder illnesses may not present unless the symptoms become more severe or complications develop [190, 191]. Therefore, the number of healthcare attendances is not a simple measure of illness in the population but rather a combination of illness levels, severity of the illness, availability of healthcare services, and ability or willingness to seek healthcare [189]. Based on this, we developed the extended working day moving average. This is a data-driven smoothing method that uses scaling factors to take both day of the week and public holiday effects into account.

In the extended working day moving average, each different day of the week and each day affected by a public holiday is assigned a scaling factor. This simultaneously takes into account changes in the number of healthcare attendances on days surrounding public holidays, changes in the number of attendances on the public holiday itself, and the day of the week effect. Data from one complete year, excluding any weeks containing public holidays, were used to give the scaling factors of the extended working day moving average for a syndromic indicator from the GPIH SSS. Therefore, the scaling factors will be different for each syndromic indicator.

Data from the previous year is used to compute the scaling factors for an indicator. The proportion of each weeks activity (Monday - Sunday) on each day was calculated. These were averaged over all weeks not containing public holidays to give an average proportion of the weekly activity on each day of the week. These average proportions were multiplied by five, the number of working days in a typical week in the GPIH SSS, to give the initial scaling factors. Additional scaling factors were developed based on the public holiday effects. Each public holiday was assigned the same scaling factor as a typical Sunday, and the first working day after a public holiday was given the same scaling factor as a typical Monday. This was based on the observed public holiday effects (described previously in section 3.3). These scaling factors reflect the typical number of attendances on each day of the week; a value larger than one reflects a day with typically a higher than average number of attendances.

To construct the extended working day moving average, the sum of each seven-day block was divided by the sum of the corresponding scaling factors. Note that

the extended working day moving average for a seven-day block without a public holiday is simply the sum of attendances divided by five, giving a basic moving average during these periods.

Improving trend identification and data visualisations: Data

The extended working day moving average has been developed for smoothing data from the GPIH SSS. However, the dynamics of the diseases that generate the syndromic data are complex, and the recorded number of attendances are affected by system coverage fluctuations, data collection changes, and other unknown influences on top of the day of the week and public holidays effects [30]. This can make it difficult to clearly compare and evaluate the different smoothing methods. Therefore, the smoothing method was first applied to synthetic data with the same public holiday and day of the week effects as the GPIH SSS but without longer-term trends and noise.

We constructed synthetic data for a period of four weeks. Based on historic data, we considered a total of 2900 attendances per week and split this into 696 attendances on Monday (24% of the weeks attendances), 522 (18%) on each of Tuesday to Friday, and 58 (2%) on weekend days. This gave a day of the week effect similar to what has been previously described (section 3.2). In order to incorporate a public holiday effect, the third Monday of the synthetic data was denoted as a public holiday. This day was given the same number of attendances as a Sunday (58 attendances, or 2.4% of the public holiday week's attendances). The Tuesday immediately after was given the same number of attendances as the typical Mondays (696 attendances, or 28.6%). The number of attendances on all other days in this week was left unchanged (522, or 21.4%, on the remaining days of the working week and 52, or 2.4%, on the weekend days). There were fewer attendances overall in the week containing the public holiday. This synthetic data can be seen in figure 3.21.

We also considered actual data from the GPIH SSS for 52 weeks, from 13th January 2014 to 11th January 2015. The indicators severe asthma and gastroenteritis, which we have previously worked with, were chosen as examples.

Table 3.5: Scaling factors for indicators from the GPIH SSS for the extended working day moving average. These scaling factors for Monday – Sunday were based on 52 weeks of data (13th January 2014 – 11th January 2015) using the method outlined in the main text. The scaling factors for public holidays and their surrounding days were based on observations made of the GPIH SSS over multiple years and described in section 3.3

	Scaling factors: severe asthma	Scaling factors: gastroenteritis
Monday	1.30	1.25
Tuesday	0.95	0.95
Wednesday	0.91	0.91
Thursday	0.87	0.90
Friday	0.93	0.95
Saturday	0.03	0.02
Sunday	0.01	0.01
Public holiday	0.01	0.01
First working day after public holiday	1.30	1.25

An evaluation of the extended working day moving average

The extended working day moving average was applied to synthetic data and the severe asthma and gastroenteritis syndromic indicators from the GPIH SSS. The seven-day and working day moving averages were also applied for comparison.

Using the percentages 2%, 18%, and 24%, described in the Data section just above, the scaling factors for the extended working day moving average applied to the synthetic data were calculated as 0.1 for weekend days and public holidays, 1.2 for typical Mondays and the first working day after a public holiday, and 0.9 for all other days of the working week. The scaling factors calculated from the severe asthma and gastroenteritis indicator data are given in table 3.5.

The extended working day moving average showed a no-trend line when applied to the synthetic data, as the combination of day of the week and public holiday effects were taken into account (figure 3.21). The extended working day moving average also continued to display the trends in the syndromic data throughout public holiday periods (figure 3.22).

In the absence of public holidays, the seven-day moving average applied to the synthetic data smoothed the regular day of the week effect to highlight the current

trend. However, there is a dip in the smoothing trend curve for seven days around the public holiday (figure 3.21). These synthetic data followed the expected behaviour of no-trend syndromic data around a public holiday. With real data, this dip in the smoothing curve could mask an actual increase in disease levels over this time period. However, this change is entirely expected due to the change in healthcare service provision on public holidays. Additionally, the seven-day moving average was lower than the average number of attendances on a working day. It is more useful that the smooth trend curve gives an indication of the number of healthcare contacts or attendances on a typical working day. These same results were also seen when the seven-day moving average was applied to surveillance data for the severe asthma and gastroenteritis indicators (figure 3.22).

The working day moving average applied to synthetic data gave a better smooth curve than the seven-day moving average (figure 3.21). However, a drop three days before and a peak four days after public holidays were still present in the smoothing curve when applied to both synthetic and real data (figures 3.21 and 3.22). These were due to the combination of the day of the week and public holiday effects. The drop was caused by that seven-day sum not including a typical Monday, and the peak was caused by that seven-day sum including both a typical Monday and the elevated Tuesday directly after the public holiday.

In the absence of big day of the week effects, the working day moving average would smooth a simple public holiday effect. However, the interaction between day of the week and public holiday effects, and extended holiday effects such as a change in the number of attendances on the first working day after a public holiday, are not accounted for.

Smoothing trend curves are used to help investigators visually identify current unusual activity during daily surveillance of syndromic disease data. It is easy to retrospectively look at the smoothing curve given by the working day moving average and identify the spikes as clearly spurious due to their short duration. However, in order to emphasise how misleading the seven-day and working day moving averages can be, we applied all the smoothing methods to the dataset that would be available a week after a Monday public holiday. This graph would be used to assess the current trend in the number of severe asthma attendances (figure 3.23). The trend one week after a public holiday would be noted as increasing if either the seven-day or working day moving averages were used. This could lead to unnecessary alarm. The extended working day moving average did not show an increasing trend and,

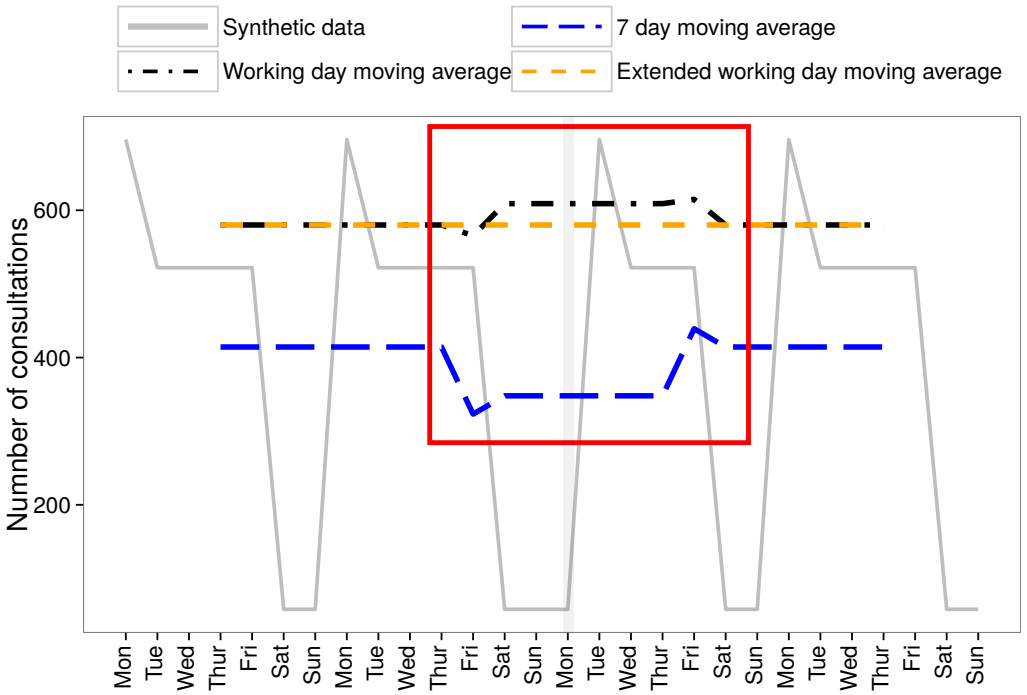


Figure 3.21: The extended working day moving average applied to synthetic data, with the seven-day and working day moving averages for comparison. Synthetic data were generated for 28 days, containing day of the week and public holiday effects representative of those observed in the GPIH SSS, but without noise and longer term trends. The synthetic data included a public holiday Monday. This is indicated by the grey vertical line and easily identifiable by the negligible number of attendances on this day. The extended working day moving average was applied to these data with the seven-day and working day moving average shown for comparison. The red box highlights the pre- and post- public holiday period of interest.

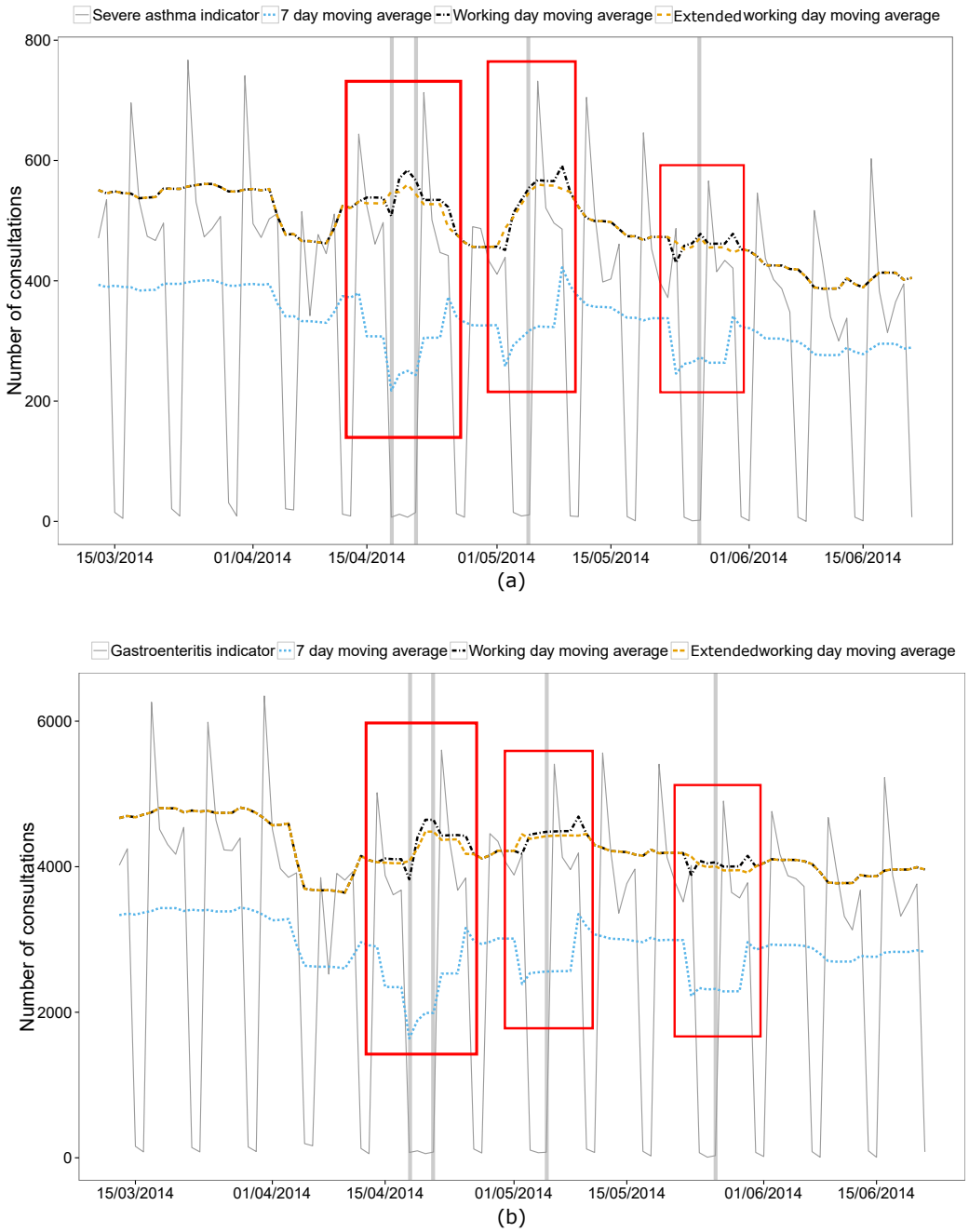


Figure 3.22: The extended working day moving average applied to indicators from the GPIH SSS, with the seven-day and working day moving averages for comparison. The number of (a) severe asthma and (b) gastroenteritis attendances from the GPIH SSS with the extended working day moving average. The seven-day and working day moving averages are also included for comparison. The grey vertical lines indicate public holidays. The red boxes highlight the pre- and post- Monday public holiday dips and peaks in the seven-day and working day moving average and their removal in the extended working day moving average.

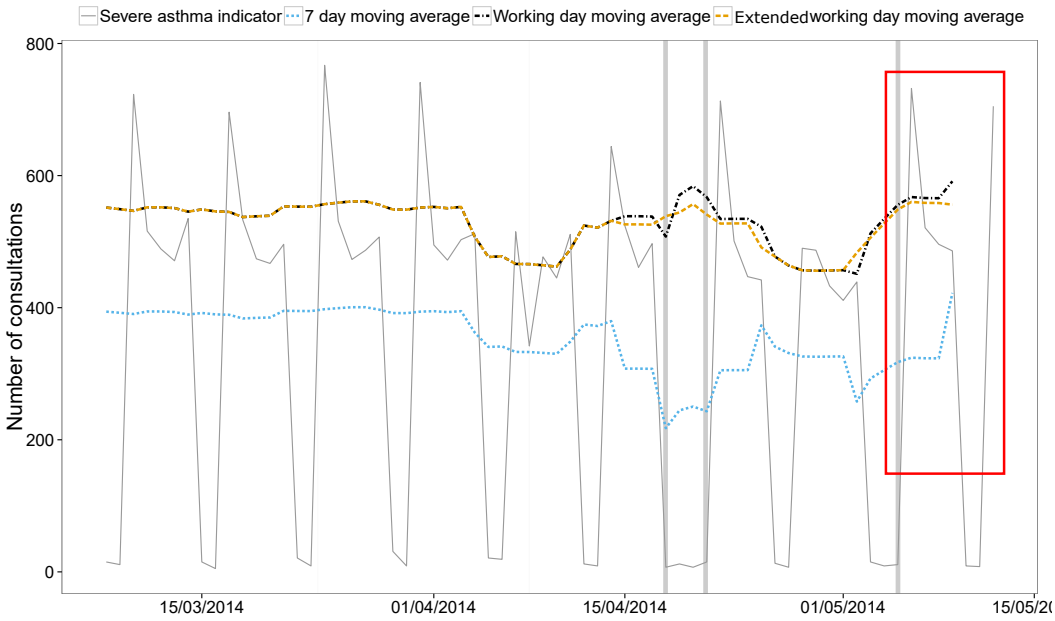


Figure 3.23: A comparison of the current trend given by each of the smoothing methods for the severe asthma indicator from the GPIH SSS. This graph displays the data that is available one week after a Monday public holiday (public holidays indicated by grey vertical lines). A smoothing method would be used to display the current trend (the area of interest inside the red box). Both the seven-day and working day moving averages show a currently increasing trend. The extended working day moving average and, importantly, the data do not.

more importantly, neither did the data. The extended working day moving average would make it easier for investigators to identify unusual activity during this period.

Improving trend identification and data visualisations: Discussion

It is widely acknowledged, and we demonstrated in sections 3.2 and 3.3, that day of the week and public holiday effects exist in healthcare data used for syndromic surveillance and that this can disguise anomalies in the data when visually inspecting it [143, 30, 184–189]. In this section, we have used the knowledge obtained previously about day of the week and public holiday effects in the GPIH SSS to develop a smoothing method where both day of the week and public holiday effects are taken into account simultaneously. We demonstrated how the extended working day moving average can be used to aid interpretation of the trends in real-time syndromic surveillance data from GP services, thereby improving the public

health action resulting from the analysis. The extended working day moving average method retains the ability to display unusual changes in the trends of syndromic indicators from the GPIH SSS during public holiday periods, and it removes the potentially misleading spikes observed in the working day moving average. This reduces the potential for delays in the detection of public health threats during this time.

In this section, data-driven smoothing methods were considered. Syndromic surveillance uses large, varied data sets, and it is desirable for syndromic surveillance reporting systems to be as automated as possible. A straightforward data-driven smoothing approach ensures sufficient flexibility so that smoothing methods can be applied to a wide range of indicators in an automated way [187]. This study shows that both day of the week and public holiday effects should be considered simultaneously to create adequately smooth daily healthcare data. We have addressed this problem in the context of in-hours GP attendance data used for daily syndromic surveillance in England, and we have focused on methods to improve time series graphs used for daily risk assessments by investigators.

The extended working day moving average described here was developed for use with just one particular syndromic surveillance system. Further work in this area could be to investigate whether the extended working day moving average could be applied to other surveillance systems. In particular, whether it is valid for those which monitor attendances at seven-day healthcare services. If the day of the week and public holiday effects in other surveillance systems are not as large as those observed in the GPIH SSS a more simple method could be sufficient.

The main limitation of the extended working day moving average is that historical data are needed to compute the scaling factors. In particular, sufficient data are needed to learn how the number of attendances changes around each public holiday. On the other hand, the working day moving average and seven-day moving average do not require historical data and, therefore, can be used immediately with new syndromic surveillance systems.

The extended working day moving average is now in use in the GPIH SSS at PHE. It has led to enhanced visualisations of these data during the analysis phase and in weekly public health bulletins [183]. Based on this work, it is recommended that analysis and visualisation methods for syndromic data carefully take both day of the week and public holiday effects into account.

3.5 Overall discussion and conclusions

Syndromic surveillance data are a near real-time source of healthcare information. There are, of course, many potential biases that affect these data. In this chapter we have described one of the main additional problems of daily data compared to weekly data.

These results show that corrections should be made for the day of the week, public holidays, and days surrounding public holidays when analysing, visualising, and modelling daily syndromic data of gastroenteritis and other health conditions. The analysis highlights the importance of being aware of the potential trends and patterns in healthcare data due to changes in behaviour rather than changes in actual disease levels. These results are of practical value for anyone analysing and interpreting healthcare data, on gastroenteritis and other conditions, at a daily time granularity.

CHAPTER 4

ONLINE SURVEILLANCE OF GASTROENTERITIS

The burden of gastroenteritis on the general population is not well understood. The purpose of this chapter is to explore modern, additional sources of online data to complement more traditional surveillance data of gastroenteritis in an attempt to provide better estimates of the number of cases.

Section 4.3 was undertaken during a secondment at the ReSST of PHE during this PhD. The data from the ReSST used in this work are covered by governance and contractual agreements that limit their use for PHE surveillance activities only. These data are therefore not available for sharing. It should also be emphasised that the opinions expressed herein do not necessarily reflect the views of the ReSST or any part of PHE.

4.1 Aims and objectives

The aim of this chapter is to investigate whether novel datasets found online can complement existing surveillance of gastroenteritis in the UK. This will be tackled in two ways and, as such, this chapter is split into two main sections. In particular, we will:

1. Investigate whether laboratory surveillance of norovirus can be supplemented with online data.

1a. Describe the necessary background material.

1b. Extract data on gastroenteritis from Google search engine queries and Wikipedia page views.

1c. Compare the novel datasets with existing laboratory surveillance data.

1d. Assess whether the novel datasets can be used for enhanced forecasting of the laboratory data.

2. Investigate reports of gastroenteritis from Flusurvey: an online cohort survey designed for influenza-like illness (ILI) surveillance.

2a. Describe the necessary background material.

2b. Describe data on gastroenteritis symptoms from Flusurvey.

2c. Compare Flusurvey data on gastroenteritis to existing syndromic surveillance systems of gastroenteritis, and discuss whether this information adds value to existing surveillance.

4.2 Surveillance of norovirus using search engine queries and page view data

In this section we will investigate the potential of search engine query data from Google and web page view data from Wikipedia to supplement laboratory surveillance of norovirus.

We started by extracting the necessary data from Google and Wikipedia. We compared these data to the weekly number of positive norovirus laboratory reports in England and Wales. Finally, we compared a selection of models for forecasting or nowcasting the norovirus laboratory data to see if the addition of these new data sources leads to better predictive ability.

We start with a review of existing studies and of the statistical methods we will use.

4.2.1 Search engine queries and page view data: Background

Suspected cases of norovirus in England and Wales can be confirmed with laboratory testing of stool samples in PHE laboratories [192]. However, laboratory based

surveillance systems do not report cases immediately. The number of days between a sample being collected and the laboratory report being generated is not trivial; this reporting delay was given by Noufaily et al. (2016, [193]) as around 11 days for norovirus. There will be additional delays between a patient first experiencing symptoms and their report to a healthcare service. The annual winter norovirus peak is not entirely regular: it frequently shifts by weeks or months from year to year [192]. Therefore, it is important to have timely surveillance of norovirus to identify the onset of this annual, seasonal outbreak.

We hypothesise that indications of an outbreak may be seen in other datasets before being identified in the laboratory data. We may be able to use these indications to give an early warning of the onset of the winter norovirus season. This would enable healthcare services and the public to prepare and take preventative action.

Finally, laboratory surveillance systems are expensive to run, upgrade, and expand. Cheaper, more readily available data could be used to supplement traditional methods to enhance disease surveillance.

One recent expanding area of research is using user-generated information from the internet for disease surveillance, see for example the *Perspectives* article in the *New England Journal of Medicine* from Brownstein et al. (2009, [194]) "Digital Disease Detection- Harnessing the Web for Public Health Surveillance". These novel web-based sources of data are one approach to developing surveillance systems that give an early warning of outbreaks using cheap, timely, readily available data. The data used in this type of surveillance fall, roughly, into three categories: search engine query data, web page view data and social media posts. We will give a brief outline of the way each has previously been used for gastroenteritis surveillance. Note, however, that few research articles in this area have explored surveillance of gastroenteritis and gastroenteritis causing pathogens - studies of influenza are more common [195]. Afterwards, we will give a brief outline of the statistical methods that will be used for our analysis.

Literature: Search engine queries

It has been reported that both in the U.S. and in Europe more than 50% of the population use the internet to find health information [196, 197]. We often start our search for information online from a search engine. A study from 2003 esti-

mated that around 5% of search engine queries are health related, corresponding to about 7 million health related searches on Google per day [198]. Based on this, researchers have hypothesised that the daily volume of search engine queries for health-related information can be used as a proxy for the number of people who are ill and, therefore, for disease surveillance.

An early, and well known, example of using search engine query data for disease surveillance was *Google Flu Trends* [199]. Google Flu Trends sought to rapidly predict ILI prevalence using search terms that gave a close fit to historical ILI data from the U.S. Centers for Disease Control and Prevention (CDC). However, after a successful start, Google Flu Trends predicted double the amount of ILI in 2013 than was observed by the CDC [200] and, in 2015, stopped publishing ILI predictions. Nevertheless, Google stil provide the service *Google Trends* which makes search volume data on any search term available for a subsection of Google searches. This has subsequently been used in many other areas of healthcare research [201].

Google Trends (`https://trends.google.co.uk/trends/`) gives search volume data for search terms entered into Google, relative to the total search volume in that region, on an almost real-time basis. However, the raw data on the number of Google searches are not publicly available. Data for the number of searches are extracted from an unbiased sample of all Google searches. Each data point is divided by the total number of searches in order to account for changes in search engine use over time and between locations. The data are then scaled to a range of 0 to 100 for any time period requested and this is the only data given to users. No misspellings, spelling variations, synonyms, plural or singular versions of the requested search term are included in the results. Repeated searches from the same person in a short period of time are excluded. The results given for any search term are searches relating to the specific term and any broadly matched search terms. These cannot be separated. For example, the Google Trends Help pages state that if the term 'banana sandwich' is entered into Google Trends, the results given "*include searches for banana sandwich as well as 'banana for lunch' and 'peanut butter sandwich'* " (`https://support.google.com/trends/answer/4359550?hl=en&ref_topic=4365530`). These scaled, sampled data given by Google Trends is typically referred to as search interest data. See *Google Trends Help* for full details of the service (available at `https://support.google.com/trends#topic=6248052`).

We have found three studies using Google queries to investigate gastroenteritis incidence. A previous version of Google Trends, *Google Insights for Search*, was used

to find search terms that correlated highly with acute diarrhoea surveillance data from the French Sentinel Network [202]. In this study, Pelat et al. (2009, [202]) concluded that just *"one well chosen query was sufficient to provide time series of searches highly correlated with incidence"*. Google Insights for Search was also used to investigate whether selected search terms correlated with norovirus outbreak and hospitalisation data from the U.S. and could be used as an early indication of elevated disease activity [203]. In this study, Desai et al. found that search terms that best correlated with nationwide norovirus outbreak data were those such as 'stomach flu' and 'stomach bug'. Finally, and more recently in 2017, Google searches for dysphagia, vomiting, and diarrhoea in the U.S. were seen to correlate with a large dataset of inpatient visits for the same symptoms between 2008 and 2010 [204]. However, there were no notable correlations with a dataset of outpatient visits for the symptoms.

Search queries submitted specifically to health websites rather than to Google may serve as a better measure of how many people are ill. Search terms submitted to a health related website in Sweden were seen to correlate well with laboratory reports for norovirus in the country between 2006 and 2013 [205, 206]. The search data detected the start of the winter norovirus season two to three weeks earlier than using the laboratory reports. They found that using the specific search term 'winter vomiting disease' gave better results than the general search term 'vomiting'. However, web query data from the same Swedish website were not found to identify local norovirus outbreaks [207].

We have not been able to identify any existing studies of gastroenteritis surveillance using search engine query data in the UK.

Literature: Web page view data

If searches on the internet for health related information start at a search engine you would imagine they would lead to a web page giving advice. Therefore, the number of times specific web pages are viewed has also been considered as a proxy for the number of people that are ill. There are fewer studies in this area, we suspect because these data are not often publicly available.

We believe that the first study in this area aimed to determine whether the number of times pages containing information about influenza on the small health website

Healthlink were viewed correlated with influenza data from the CDC (Johnson et al. 2004, [208]). Moderately strong correlations between the web page view data and the 'gold standard' influenza surveillance data were found.

All further studies, that we have been able to find, using web page view data for disease surveillance consider *Wikipedia* page view statistics, which are freely available to download. Wikipedia (`https://en.wikipedia.org`) is an open-access online encyclopaedia written by its users [209]. It contains many articles on health-related topics (among many other). A Wikipedia page was found to be among the first ten results in more than 70% of searches for health-related keywords on a selection of search engines tested in a study by Laurent and Vickers (2009, [209]), and they conclude that it is a "*prominent source of online health information*".

Three studies applied statistical methods, including regression, Pearson's correlations, and Bayesian change point detection, to Wikipedia page view data to investigate ILI, cholera, dengue, Ebola, HIV, plague, and tuberculosis surveillance using these data [210–212]. Additionally, Wikipedia page view data have been incorporated into a mechanistic model of disease spread (SEIR-type model) to produce reasonably good forecasts of the influenza season in the U.S. [213].

A simple study of the norovirus Wikipedia page view data compared the number of views in January 2008 with the number of views in June 2008 using a t-test to identify a seasonal change in page use that broadly corresponds with the seasonality of norovirus [209].

A key difficulty of using these Wikipedia page view data is that no location information is provided. Generous et al. (2014, [211]) use article language as a proxy for location, although clearly this is not possible for English language articles. McIver et al. (2014, [210]) make estimates of ILI in the U.S. only, noting that 41% of Wikipedia English language article views come from the U.S.

As with search engine queries, it is impossible to discern whether the web page views are from people suffering from an illness or just information seeking in response to, for example, increased news coverage of a health event. To this end, in addition to being used as a proxy for the number of ill people, Wikipedia page view data have been used as a proxy for concern and 'public anxiety' in a population during the 2009 H1N1 influenza outbreak [214].

We have not been able to identify any studies of gastroenteritis surveillance using

web page view data beyond the simple winter-summer comparison made by Laurent et al. (2009, [209]), and we have not been able to find any uses of Wikipedia page view data for disease surveillance in the UK.

Literature: Social media posts

Social media platforms are places for people to post about their experiences, views, and opinions. The text posted on these sites has been considered as a dataset for disease surveillance. In particular, *Twitter* is the social media site used most often for this analysis [215]. Perhaps this is due to the fact it is simple to collect millions of tweets from the Twitter Application Programming Interface [216]. Initial studies in this area, again, focussed on ILI surveillance [215].

We have found some studies using Twitter data for gastroenteritis surveillance. A basic investigation into norovirus related keyword use on Twitter reported, perhaps unsurprisingly, that the keyword 'diarrhoea' was not reported as frequently as other gastroenteritis related keywords such as 'fever,' 'norovirus,' and 'sick' [217].

Within the UK, there is an ongoing (and as yet, we believe, unpublished) study by the Food Standards Agency investigating the ability of Tweets containing norovirus related keywords to predict norovirus outbreaks (a description of the project is available at `http://blogs.nhs.uk/choices-blog/2016/02/12/guest-blog-using-twitter-to-predict-norovirus-outbreaks/`). Data collected from Twitter are compared to laboratory reports for norovirus. Preliminary results show that the Twitter data can give good predictions of the laboratory data.

The analysis of social media data is slightly different from the analysis of search engine queries and web page view data as it requires text analysis. It is likely that not all relevant tweets are classified correctly due to common misspellings, slang, and abbreviations. Additionally, only a very small percentage of tweets are geotagged. This makes it difficult to restrict the data to a country or area of interest. Finally, Krieck et al. (2011, [218]) report that the majority (51%) of the tweets they collected containing symptoms or disease names contained news reports and information as opposed to descriptions of personal symptoms. Due to these reasons, and the existence of the Food Standards Agency project described above, we will not analyse social media posts in this work.

Method: Cross-correlation

The cross-correlation of two time series is a measure of their similarity at different time lags. It is used to compare two time series when it is suspected that there is a delay in similar trends between them. To find the cross correlation at lag k, Pearson's correlation coefficient of the two time-series is computed with the first time series shifted forward k time steps. Mathematically, the cross-correlation at lag k of time series $x = \{x_t\}$ and $y = \{y_t\}$, both of length n, is

$$\rho_{x,y}(k) = \frac{\sum_{t=1}^{n-k}(x_{t+k} - \overline{x}^{(k)})(y_t - \overline{y}^{(-k)})}{\sqrt{\sum_{t=1}^{n-k}(x_{t+k} - \overline{x}^{(k)})^2}\sqrt{\sum_{t=1}^{n-k}(y_t - \overline{y}^{(-k)})^2}} \ ,$$

where

$$\overline{x}^{(k)} = \frac{\sum_{t=1}^{n-k} x_{t+k}}{n-k} \qquad \text{and} \qquad \overline{y}^{(-k)} = \frac{\sum_{t=1}^{n-k} y_t}{n-k} \ .$$

If the cross-correlation of x and y has a large value (close to 1) at lag k, where k is a positive number, we say that x *lags* y. If the cross-correlation of x and y has a large value at lag k, where k is a negative number, we say that x *leads* y. We use the `ccf` function in the statistical computing programming language R to compute cross-correlations [219].

Method: Autocorrelation

The autocorrelation of a time series is simply the cross-correlation with itself. This can be used to identify periodicities in the data; a large value of the autocorrelation at lag k indicates a periodicity of length k in the time series.

Method: Serfling method

The Serfling method is a harmonic regression model developed by Robert E. Serfling in 1963 [220] to estimate the number of excess deaths due to influenza. It is used with seasonal data to extract the activity due to a seasonal outbreak from usual baseline activity levels and to identify the onset of an epidemic. Although there are more technical approaches for this purpose, the Serfling method has been widely used to establish a baseline measure of influenza activity by Public Health groups around the world [221].

Initially, a harmonic regression model is fitted to the out-of-season data to obtain a baseline measure of activity. This harmonic model assumes that the seasonal pattern of disease activity remains stationary over the years. Only the out-of-season data is used to establish this baseline to prevent epidemic activity from raising it. A simple alternative is to construct a flat baseline, for example using the mean of all out-of-season data, however we prefer the harmonic model that takes into account a gentle seasonality and highlights that the more extreme seasonality is due to winter outbreaks.

We fitted a regression model of the form

$$y_i = a_0 + a_1 t + a_2 \cos\left(\frac{2\pi}{52} t_i\right) + a_3 \sin\left(\frac{2\pi}{52} t_i\right) + \epsilon_i \ ,$$

using the `HarmonicRegression` package in R [222] to only the out-of-season data. We defined the out-of-season period of norovirus as June to October inclusive.

The standard deviation of the residuals, σ_{res}, gives an estimate of the variation in the regression model fit [223]. Assuming the residuals follow a normal distribution around zero, an approximate 95% prediction interval on a predicted value x is

$$[x - 1.96\sigma_{\text{res}}, x + 1.96\sigma_{\text{res}}] \ .$$

Excess activity due to a seasonal outbreak is defined as activity above the upper bound of this prediction interval. We will call this upper bound the *Serfling threshold*.

Method: ARIMA [224]

An ARIMA model is a time series modelling method made up of three components: an autoregressive component, a differenced component, and a moving average component.

The differenced series, $y' = \{y'_t\}$, of a time series $y = \{y_t\}$ is formed from the change in consecutive observations of y:

$$y'_t = y_t - y_{t-1} \ .$$

Differencing is used to make a time series stationary. We may need to difference

more than once to obtain a stationary time series. The number of times differencing is applied is called the *degree of differencing*.

An autoregressive model is a linear regression on previous values of the dependent variable. The number of previous values included is called the *order of the model*.

A moving average regression model is a linear regression on previous forecast errors. The number of previous errors included is called the *order of the moving average*. Combining these gives an ARIMA model.

Let p be the order of the autoregressive part, d the degree of differencing, and q the order of the moving average. For the differenced time series y' an ARIMA(p, d, q) model is defined as:

$$y'_t = c + \alpha_1 y'_{t-1} + ... + \alpha_p y'_{t-p} + \beta_1 e_{t-1} + ... + \beta_q e_{t-q} + e_t \ ,$$

where c, α_i, and β_i are the regression coefficients and e_i the errors.

The model orders (values of p, d, q) can be chosen by selecting the ARIMA model that gives the smallest Akaike information criterion (AIC). The AIC is an estimator of the relative quality of a model for a given set of data. We use the `auto.arima` function from the `forecast` package in R to fit ARIMA models [224]. This automatically selects the ARIMA model with smallest AIC, removing the need for manual model selection.

An ARIMA model can be extended to also include exogenous dependent variables. Given independent time series $y = \{y_t\}$ and dependent time series $x = \{x_t\}$, a regression model with ARIMA errors is written as

$$y_t = \gamma x_t + n_t \ , \qquad n_t = \alpha_1 n_{t-1} + ... + \alpha_p n_{t-p} + \beta_1 e_{t-1} + ... + \beta_q e_{t-q} + e_t \ .$$

We can also ensure that this regression model with ARIMA errors includes any seasonal knowledge that we may have about the dataset by regressing on k Fourier terms, where k is chosen in advance. A regression model with Fourier terms and ARIMA noise is written as

$$y_t = \sum_{i=1}^{k} \gamma_i \sin\left(\frac{2\pi i t}{\omega}\right) + \sum_{i=1}^{k} \lambda_i \cos\left(\frac{2\pi i t}{\omega}\right) + n_t \ ,$$
$$n_t = \alpha_1 n_{t-1} + ... + \alpha_p n_{t-p} + \beta_1 e_{t-1} + ... + \beta_q e_{t-q} + e_t \ ,$$

where ω is the period length.

An ARIMA model can also be extended to a seasonal ARIMA model. This involves adding terms to a standard ARIMA model which are shifted by the seasonal period (the number of data-points until the seasonal pattern repeats again). In detail, the seasonal part of the model contains, again, an autoregressive component, a differencing, component, and a moving average component but these all operate on a lag which is a multiple of the number of periods in a season. Therefore, the seasonal ARIMA model has orders p, d, and q as before and additionally P, D, and Q for order of the seasonal autoregressive component, the seasonal differencing, and the seasonal moving average component respectively, and m which gives the number of periods per season.

Method: Time series cross-validation

Cross-validation is a technique to assess the performance of predictive models. The available data are split into subsets. The model is run on a particular subset (*training set*) and the results are used to predict the rest of the values (*test set*). The predictions can be compared to the actual data. This is repeated for many different training and test sets.

Time series cross-validation applies this principle to time series data. In particular, we will be considering forecasts only one week (which is just one data point) ahead. In this context, our time-series cross validation will involve initially a training set of the first m points of the time series and test set of point $m + 1$. Then a training set of points 2 to $m + 1$ and testing on point $m + 2$, and so on.

4.2.2 Ground truth

Novel surveillance data sources are typically validated by comparison to data from existing surveillance systems, referred to as the 'ground truth' or 'gold standard'. Success is typically measured by the success of the novel data source to replicate the patterns and trends of this ground truth.

For this study, we consider the ground truth to be the weekly number of confirmed laboratory reports of norovirus from PHE. These laboratory reports are available freely online in pdf format (at `https://www.gov.uk/government/publications/`

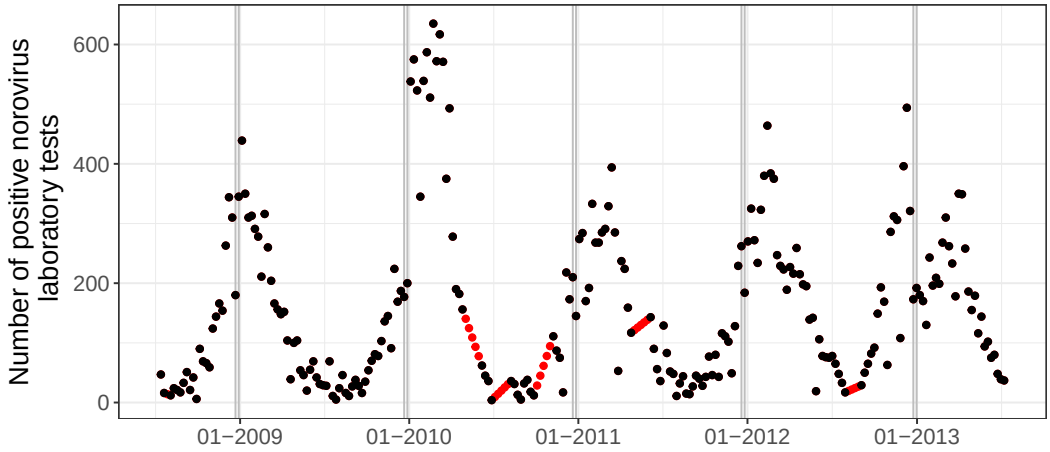


Figure 4.1: The number of positive norovirus laboratory reports each week (black), missing data estimated by linear interpolation between the surrounding two non-missing data points (red), and the weeks of Christmas and New Year (grey lines).

`common-gastrointestinal-infections-in-england-and-wales-laboratory-reports-in-2017`). These reports contain data from English and Welsh laboratories on norovirus identified in stool samples collected from outbreak situations and from patients reporting to doctors.

There are some missing data points (figure 4.1). We estimate these with a linear interpolation between the previous and next non-missing data points. This will be the time series used for all further analysis. Notice that during the week of Christmas and New Year there appears to be fewer reports than in the neighbouring weeks. This is not surprising as we have already discussed how reporting to healthcare services changes over periods containing public holidays (chapter 3). The data are clearly seasonal. The autocorrelation plot shows strong annual periodicity (figure 4.2).

These laboratory data have previously been used as the ground truth for investigations into alternative data sources for gastroenteritis surveillance, including the previously described project by the Food Standards Agency with Twitter data and by Loveridge et al. (2010, [225]), as part of the ReSST at PHE, to evaluate data from a national healthcare telephone service.

Finally, a study by Lopman et al. (2009, [192]) looked at forecasting these data using a Poisson regression model taking into account weather conditions and the emergence of new norovirus variants. For their study, the laboratory data were available

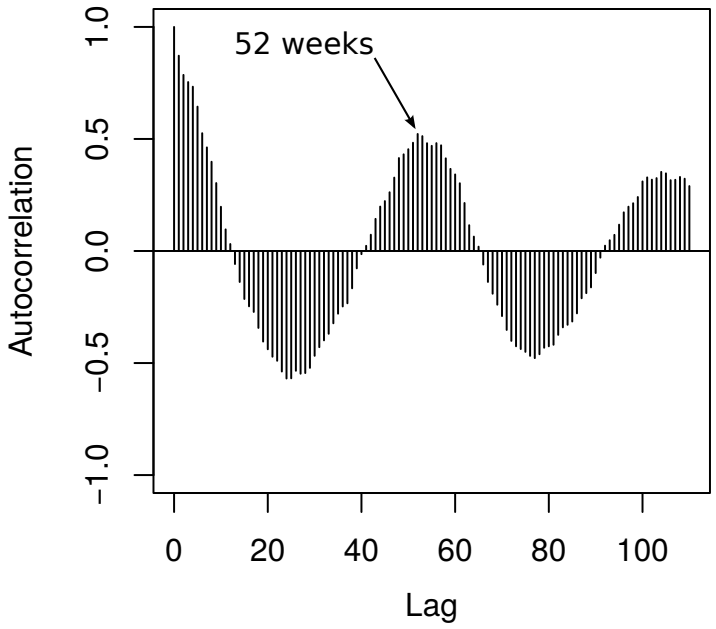


Figure 4.2: The autocorrelation plot (lag in weeks) of the norovirus laboratory reports showing a clear annual periodicity.

on a daily time granularity (as opposed to weekly in our case). In a Poisson regression model it is assumed that the dependent variable has a Poisson distribution, and when this approach is used to model count data a log link function is applied. Lopman et al. additionally adapted the Poisson regression modelling approach to be appropriate for time series data by adding autoregressive and background seasonality terms.

4.2.3 Search engine queries: Google Trends

The data

Google Trends (introduced in section 4.2.1) gives results at limited time aggregations and over a limited time period. Data are available for long time periods, from 2004 to the present, but only aggregated by month. For data aggregated into weekly search interest, we need to restrict our data requests to periods of at most a year. Therefore, we considered each norovirus season separately. We extracted weekly data from Google Trends for week 29 of one year to week 28 of the following. Note

that because Google Trends scales each dataset to be between 0 and 100, this gives different data compared to if we were able to extract all seasons at once. This dataset is suitable for most analyses, however not for a comparison through time of the severity of each season.

Google Trends aggregates data into weekly counts from Sunday to Saturday, inclusive. The norovirus laboratory data are aggregated into weeks from Monday to Sunday, inclusive. However, as the laboratory data typically come from stool samples submitted via doctors, we assume that the majority of these data is generated during the working week, Monday to Friday. Therefore, the slightly mismatched weeks should not have a large impact on this analysis.

Google Trends data are computed from an unbiased sample of Google searches, and therefore the data that are available changes slightly from week to week as the same sample is not always used [226]. We overcame this by downloading data for each search term on four different occasions. For each search term, we averaged the four datasets to give one Google Trends timeseries.

We considered the following four search terms restricted to searches in the UK: 'norovirus', 'winter vomiting bug', 'gastroenteritis', 'diarrhoea + vomiting'. Note that the '+' sign gives results for searches containing either of the words diarrhoea or vomiting.

Previous studies using Google Trends data for surveillance of norovirus and other causes of gastroenteritis have used a variety of search terms [202–204]. We inspected all of the terms used in previous analyses, and included in our analysis those that appeared, at least vaguely seasonal, and had time series that did not consist of mostly zero over the time period of interest. This gave us the terms 'norovirus', 'gastroenteritis', 'diarrhoea + vomiting'. Additionally, we included 'winter vomiting bug' as it is a common British synonym for norovirus and this is the first study to use Google Trends data for surveillance of norovirus in the UK.

Similarity of all data

The laboratory and search interest data were standardised to account for longer term trends. Each year of data (from week 29 one year to week 28 the next year) was treated separately. In the standardised datasets, the data point for each week was expressed as a proportion of the total number of laboratory notifications or

search interest in the entire year. To additionally smooth the data we used a five week moving average. Note that this smoothing was just used for this one test of similarity; it was not used throughout the rest of this section. This follows the methodology used by Edelstein et al. (2014, [206]) to use search engine data to identify the norovirus season in Sweden.

Graphically, the smoothed, standardised laboratory data and search interest data appeared to be similar as both have annual winter peaks (figure 4.3). The search terms 'gastroenteritis' and 'diarrhoea + vomiting' appeared most dissimilar to the lab data with less obvious seasonality. The peaks for 'norovirus' and 'winter vomiting bug' search interest seemed sharper than the seasonal peaks in the laboratory data. Finally, a double peak can be seen in the laboratory data during the 2007-2008 and 2012-2013 seasons, but does not appear in the search term data.

We computed the cross-correlations of the standardised, smoothed laboratory data with each Google search interest dataset (figure 4.4). The 'norovirus', 'winter vomiting bug', and 'diarrhoea + vomiting' search interest data each have high correlations with the laboratory data at either no lag or a lead of 2 weeks. The 'gastroenteritis' data were not well correlated with the lab data.

Timing of the seasonal outbreak

The laboratory data have a strong annual periodicity, with winter outbreaks each year (section 4.2.2). In order to compare the timings of the winter outbreaks in the datasets we will compute two measures: the week of season onset and the week of peak activity.

The week of peak activity was simply the week in the year with most norovirus cases or search interest. The week of season onset was computed as the first week exceeding the Serfling threshold (section 4.2.1). This method has been previously used to detect the onset time of the norovirus season in Swedish datasets [206].

It only makes sense to consider season timing for those datasets which show annual periodicity. We use autocorrelation plots to determine that there is not strong annual periodicity in the 'gastroenteritis' search volume data (figure 4.5). We will therefore not include this dataset in the analysis of season timing.

Upon visual inspection, it appears that the harmonic regression of the Serfling

method has fitted well to the out-of-season data (June to October) from the norovirus lab reports and the 'norovirus' and 'winter vomiting bug' Google search interest data (figure 4.6). However, it does not appear to be as well fitted to the Google searches for 'diarrhoea + vomiting'.

The week of season onset calculated from the 'diarrhoea + vomiting' Google search interest data always came after the week of season onset calculated from the laboratory data (range: +1 to +15 weeks) (figure 4.6). The week of season onset calculated from the 'norovirus' and 'winter vomiting bug' Google search terms data was more varied: the season onset from both search terms preceded the season onset from lab data during the 2010-2011 and 2011-2012 seasons (range: -4 to -1 weeks), but came after the lab data season onset week otherwise (range: +1 to +5 weeks). The week of peak activity from the 'norovirus' and 'winter vomiting bug' search term data preceded the peak week in the laboratory data in all except the 2012-2013 season (range: -14 to +1 weeks). The peak week in the 'diarrhoea + vomiting' Google search term data was more variable.

In conclusion, from this analysis there does not seem to be much consistency in whether the peak and onset weeks from the Google search interest data precede or follow the peak and onset weeks seen in the laboratory data, and there is variability in the amount they each lag or lead from year to year.

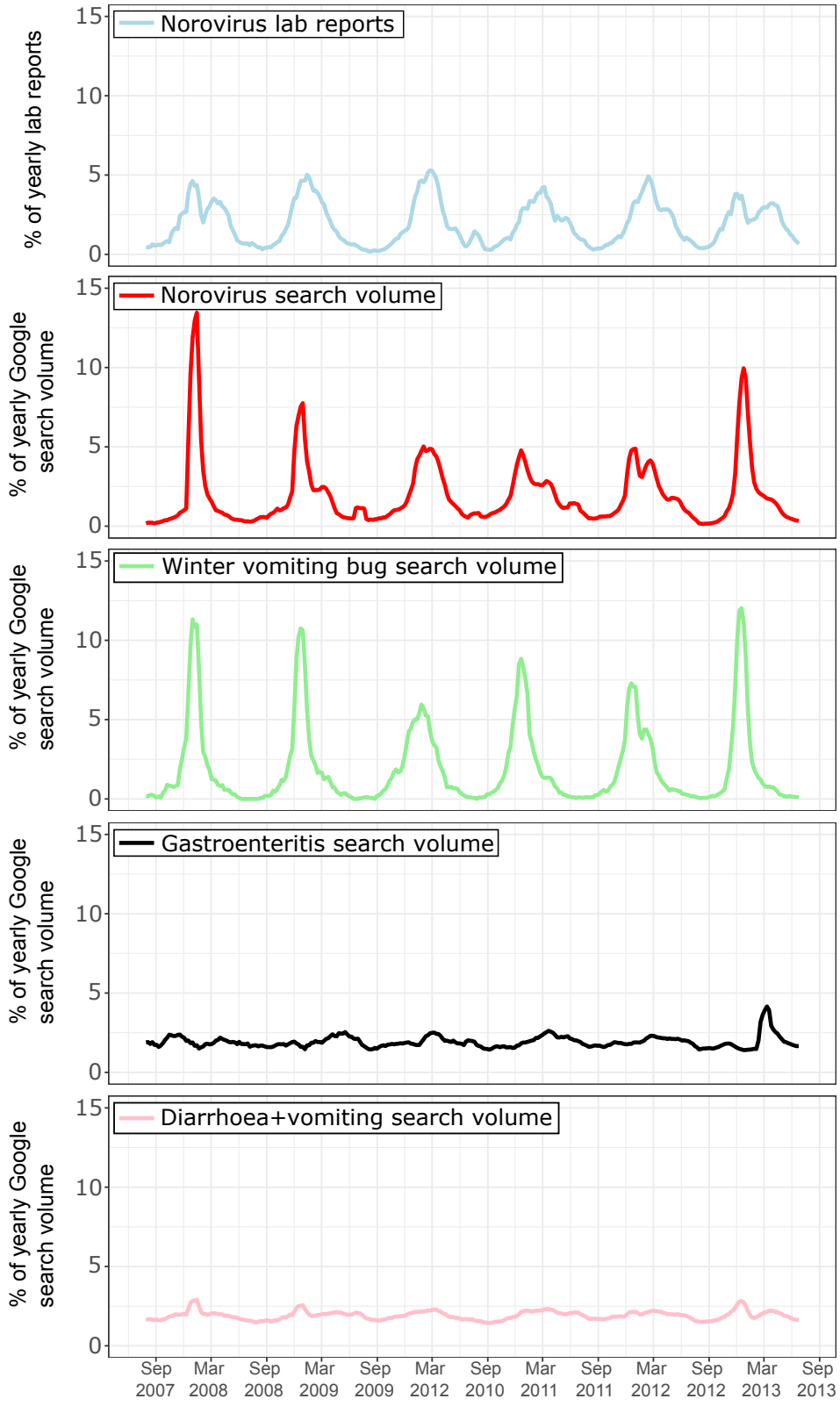


Figure 4.3: Smoothed, standardised laboratory and Google search interest data

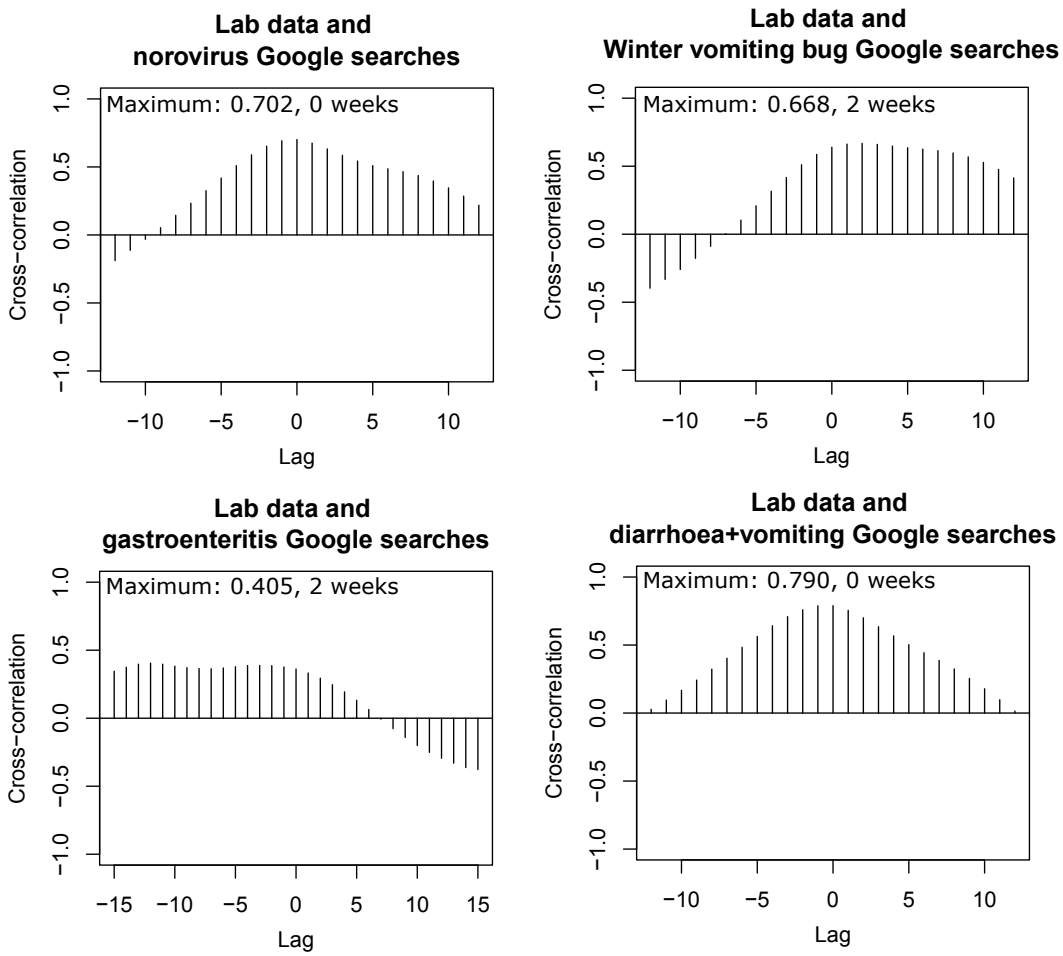


Figure 4.4: Cross-correlations of the norovirus laboratory data with the Google search volume datasets annotated with the maximum correlation and lag at which this is seen. Note that here a lag of -1 weeks corresponds to laboratory data at week 0 being compared with search engine data at week 1 (for example, trends are seen first in laboratory data and secondly in search engine data a week later). A lag of 1 week means laboratory data at week 1 being compared with search engine data at week 0 (for example, trends are seen first in search engine data and secondly in laboratory data a week later).

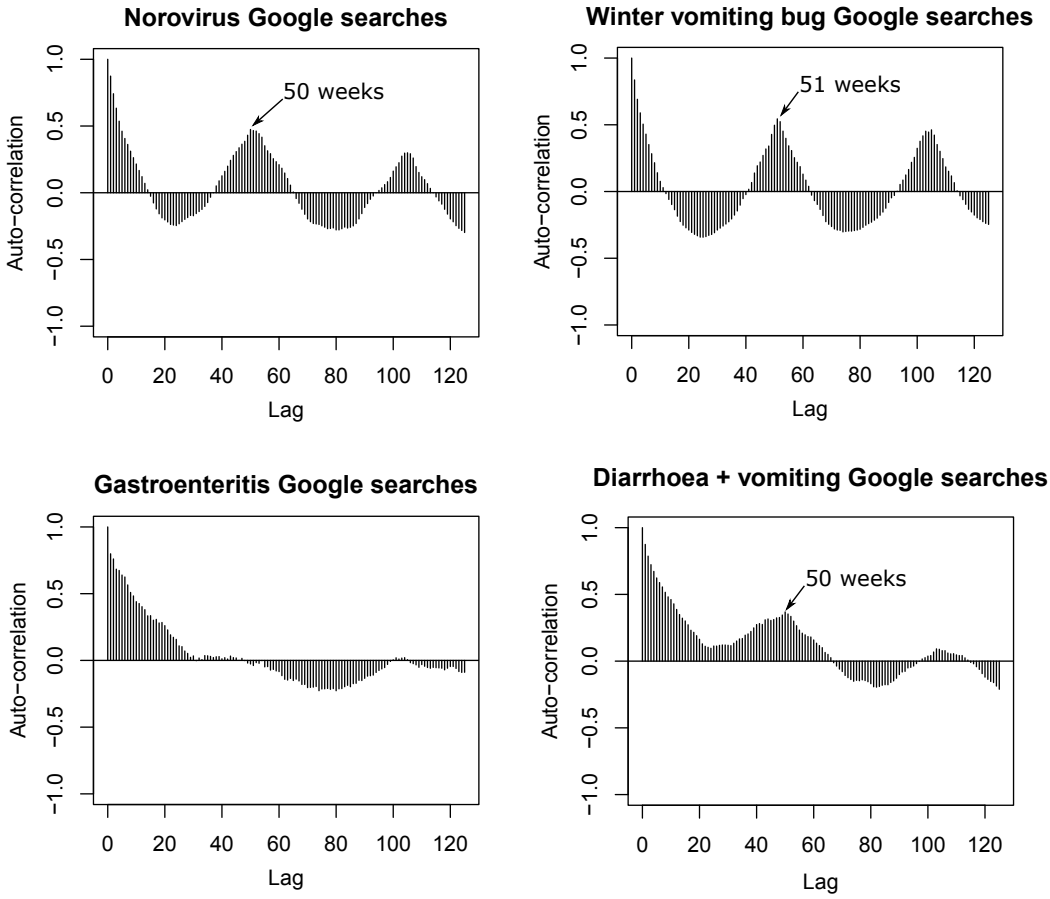


Figure 4.5: Autocorrelation plots of the Google search interest datasets (lag in weeks). Annual seasonality is seen in the 'norovirus', 'winter vomiting bug', and 'diarrhoea + vomiting' search term data. No clear seasonality is seen in the 'gastroenteritis' search term data.

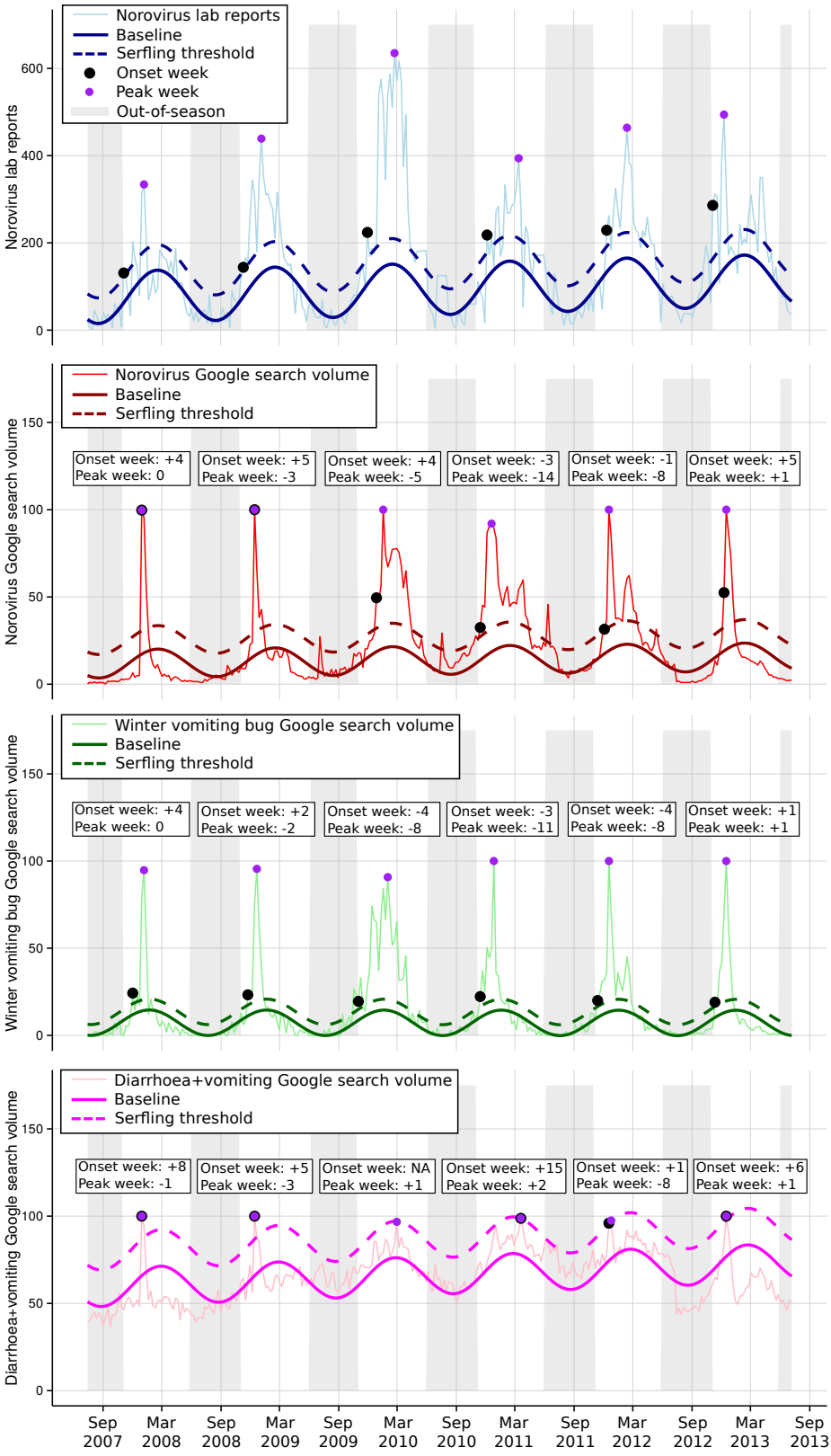


Figure 4.6: Analysis of season onset and peak times from norovirus laboratory data and Google search interest datasets. Baselines were fitted using a Serfling regression model and seasonal outbreak periods defined as activity over the Serfling threshold. Numbers describe the difference in the onset, or peak, week from Google data with the onset, or peak, week in the laboratory data (positive number indicates that the week was later in the Google data and vice versa for negative).

4.2.4 Web page view data: Wikipedia

The data

Wikipedia makes freely available data on the number of times each page was loaded. Until January 2016 this was via the website `stats.grok.se` but is now via the `toollabs` project `pageviews` (`https://tools.wmflabs.org/pageviews`). Note that this analysis was completed before the `pageviews` tool was available.

We downloaded daily page view data from 15th July 2007 to 7th July 2013. Data are available from 10th December 2007 [210]. However, norovirus typically has a peak of activity in the winter therefore we only collected data for which the full winter season was available. No data are available from 13th July 2007 to 31st July 2007 [210]. However, as this fell outside of the typical norovirus season we recorded these dates as zeros, and we do not feel this impacted too greatly on the analysis. The data were aggregated to weekly levels by summing the page views over seven days (Monday - Sunday).

Note that from these page view statistics we cannot infer anything about why the page was loaded, how long the user remained on the page, or whether they read any of the content. However, these counts of page views have been previously used as a proxy for human views of the page by Generous et al. (2014, [211]).

In order to account for the changing use of Wikipedia over time, we normalised the time series of views for each Wikipedia page by dividing by the number of times the main Wikipedia page (`https://en.wikipedia.org/wiki/Main_Page`) was viewed in the same time period.

We analysed the page view statistics of the following Wikipedia pages: norovirus, Norwalk virus, gastroenteritis, diarrhea, and vomiting. We chose these using the same approach detailed by Generous et al. (2014, [211]): articles linked from the page for the disease itself were listed and those on relevant symptoms, synonyms, and epidemiology were chosen along with the page on the disease itself. However, we additionally added vomiting, the second major symptom of norovirus, even though this was not identified by the structured method. Note that norovirus was originally named Norwalk virus, after an outbreak in Norwalk, Ohio in the U.S., and that the name Norwalk virus is sometimes used synonymously with the name norovirus.

A Wikipedia *redirect* page has no content itself but instead just points to another

article (the *target page*). These account for synonyms and misspellings [211]. However, when a redirect page is used a page view is not registered for the target page, but instead for the redirect page. The sum of the redirect page views and the target page views should give the total number of target page views. However, as stated by Generous et al. (2014, [211]), *"reliably mapping redirects to targets is a non-trivial problem because this mapping changes over time"*. Therefore, we do not consider redirects in this analysis.

Note that the introduction of the new `toollabs Pageviews` project includes the *Redirect Views* tool which gives page view statistics of a page and all its redirects. However, this was not available at the time of the analysis and only gives data from July 2015 so cannot even be retrospectively used to re-do this analysis.

Similarity of all data

The page view data were standardised to account for longer term trends. As before, each year of data (from week 29 of one year to week 28 of the next) was treated separately. In the standardised time series, the number of page views for each week was expressed as a proportion of the total number of page views in the entire year. This was additionally smoothed using a five week moving average. Note that this smoothing was only applied for this one test of similarity; it is not used throughout the rest of this section. This, again, follows the methodology used by Edelstein et al. [206].

Visually, the laboratory data and page view data from the norovirus and Norwalk virus pages were most similar due to the shared obvious annual peaks (figure 4.7). The data for the gastroenteritis page showed some small seasonality. The data from the vomiting and diarrhoea pages appear very similar to each other, but without regular annual spikes in activity do not appear similar to the norovirus lab reports.

Indeed, these observations are corroborated by the cross-correlations (figure 4.8). The highest correlation is between the lab data and the views of the norovirus page, with the Wikipedia data leading the lab data by one week. There are additionally reasonably high correlations between the norovirus laboratory data and both the Norwalk virus page views and the gastroenteritis page views at leads of 2 and 3 weeks respectively.

Timing of the season

The laboratory data demonstrate considerable seasonality (section 4.2.2). To continue our analysis, we will restrict ourselves to just those Wikipedia pages that also demonstrate annual seasonality. The autocorrelation of the Wikipedia page view data shows that the views of the norovirus and Norwalk virus pages are seasonal and that the views of the gastroenteritis page are perhaps weakly seasonal (figure 4.9). However, there is no clear seasonality in the views of the vomiting and diarrhoea pages. We, therefore, continue this analysis with just the norovirus and Norwalk virus page view datasets.

Again, we discuss the timing of the season using the onset week, calculated as the first week to exceed the Serfling threshold, and the peak week, defined as the week with most activity. As part of the Serfling method, a harmonic regression model was fitted to the out of season data. There appears, through visual inspection, to be a good fit to both the out-of-season (June to October) norovirus and Norwalk virus Wikipedia page view data (figure 4.10).

For all but the 2012-2013 season, the norovirus Wikipedia page view peak week coincided with or preceded the norovirus lab data peak week. However, the range of the difference was reasonably large (range: -11 to 0 weeks). The onset week from the norovirus page views sometimes preceded and sometimes followed the lab report onset week. This was the same for the Norwalk virus peak week. Finally, the Norwalk virus page view onset week always followed the norovirus lab reports onset week by at least 7 weeks (range: $+7$ to $+10$ weeks).

In conclusion, there is no evidence of an association between whether outbreaks seen in the page view data precede or follow the outbreaks seen in the lab report data based on the current methodology used to analyse these datasets.

Severity of the season

The severity of the season was assessed by calculating the percentage of reports or page views on the peak week of each season and by also calculating the percentage of the season's reports or page views considered as excess (above the Serfling threshold).

The first measure gives an idea of the 'sharpness' of the epidemic peak and the second gives a measure of the total size of the seasonal outbreak. These are similar

Table 4.1: Two measurements of the severity of each season. The peak column gives the percentage of the reports or searches on the peak week of the season. The outbreak column gives the percentage of the reports or searches above the Serfling threshold.

	Lab Reports (%)		Norovirus page views (%)		Norwalk virus page views (%)	
	Peak	Outbreak	Peak	Outbreak	Peak	Outbreak
2008-09	6.272	18.616	5.728	5.578	3.848	0.394
2009-10	5.756	37.485	4.932	10.522	3.917	0.017
2010-11	5.284	11.153	4.057	10.753	4.978	0
2011-12	5.927	13.016	2.748	37.986	4.874	0.278
2012-13	5.798	16.445	2.581	42.613	1.249	64.490

to measures used by Olson et al. (2013, [227]) to assess epidemic intensity for seasonal influenza.

Generally, the peak week consisted of between 3% and 6% of the season's reports or page views (table 4.1). In particular, the peak week percentage of norovirus laboratory reports was very consistently between 5.3% and 6.3% of the season's total. The Wikipedia page view peak week percentages were not as consistent. The season with maximum severity, as measured by the peak week, coincided for the norovirus lab reports and norovirus Wikipedia page views. The peak week percentage of Norwalk virus Wikipedia page views was particularly low during the 2012-13 season. This was due to a large number of weeks each having a large number of views: a broad rather than 'peaky' outbreak.

The total outbreak severity in the Norwalk virus page views was very low except for in the 2012-13 season. This season also had the most severe outbreak in norovirus page views. Conversely, this was not a severe season for the outbreak in norovirus lab reports. The most severe outbreak for the norovirus lab reports was during the 2009-10 season. This was not a remarkable year in either the norovirus or Norwalk virus Wikipedia page views.

Overall, the most severe norovirus laboratory report seasons did not correspond with the most severe Wikipedia page view seasons and vice-versa using this analysis method.

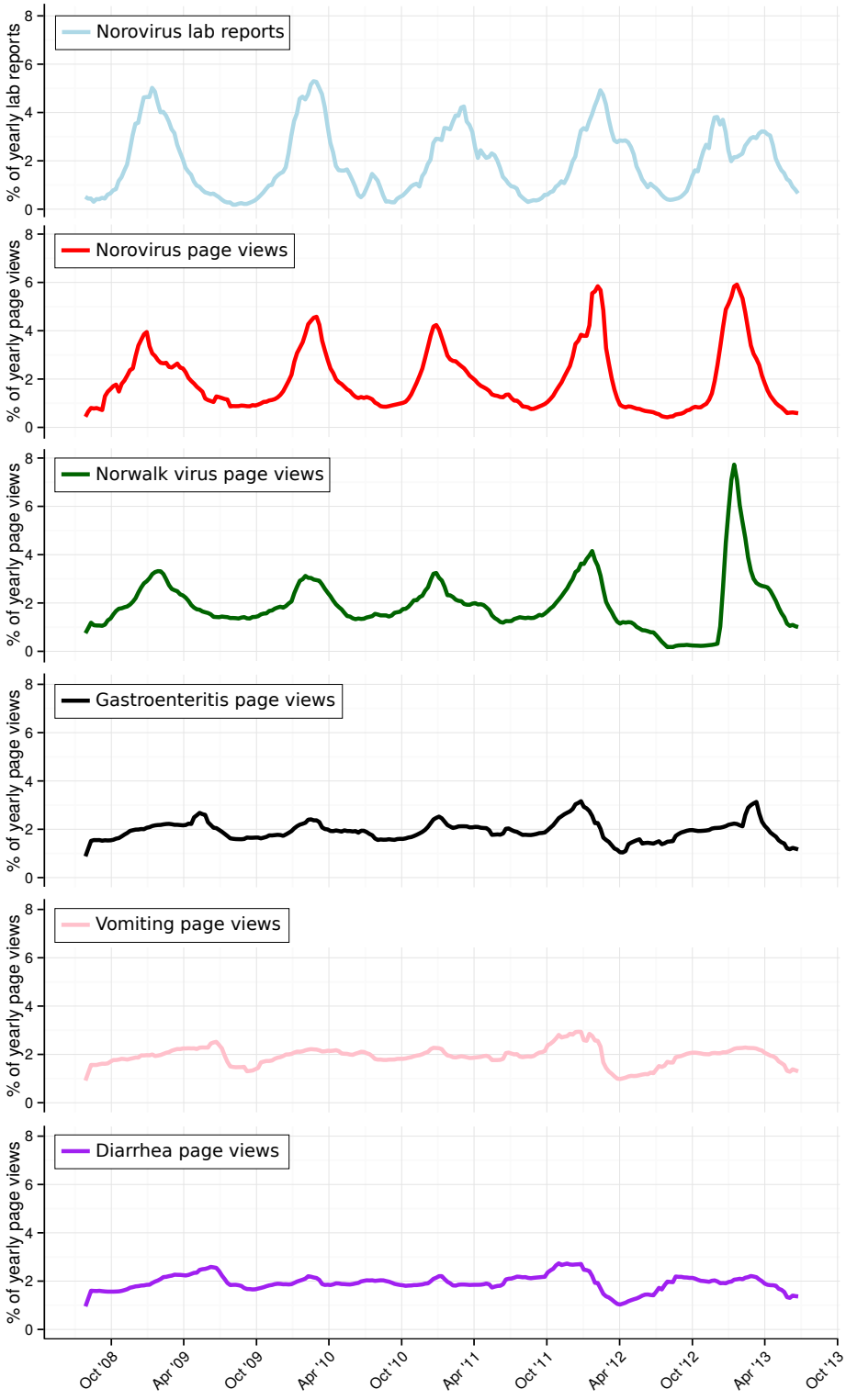


Figure 4.7: Smoothed, standardised norovirus laboratory reports and Wikipedia page view statistics.

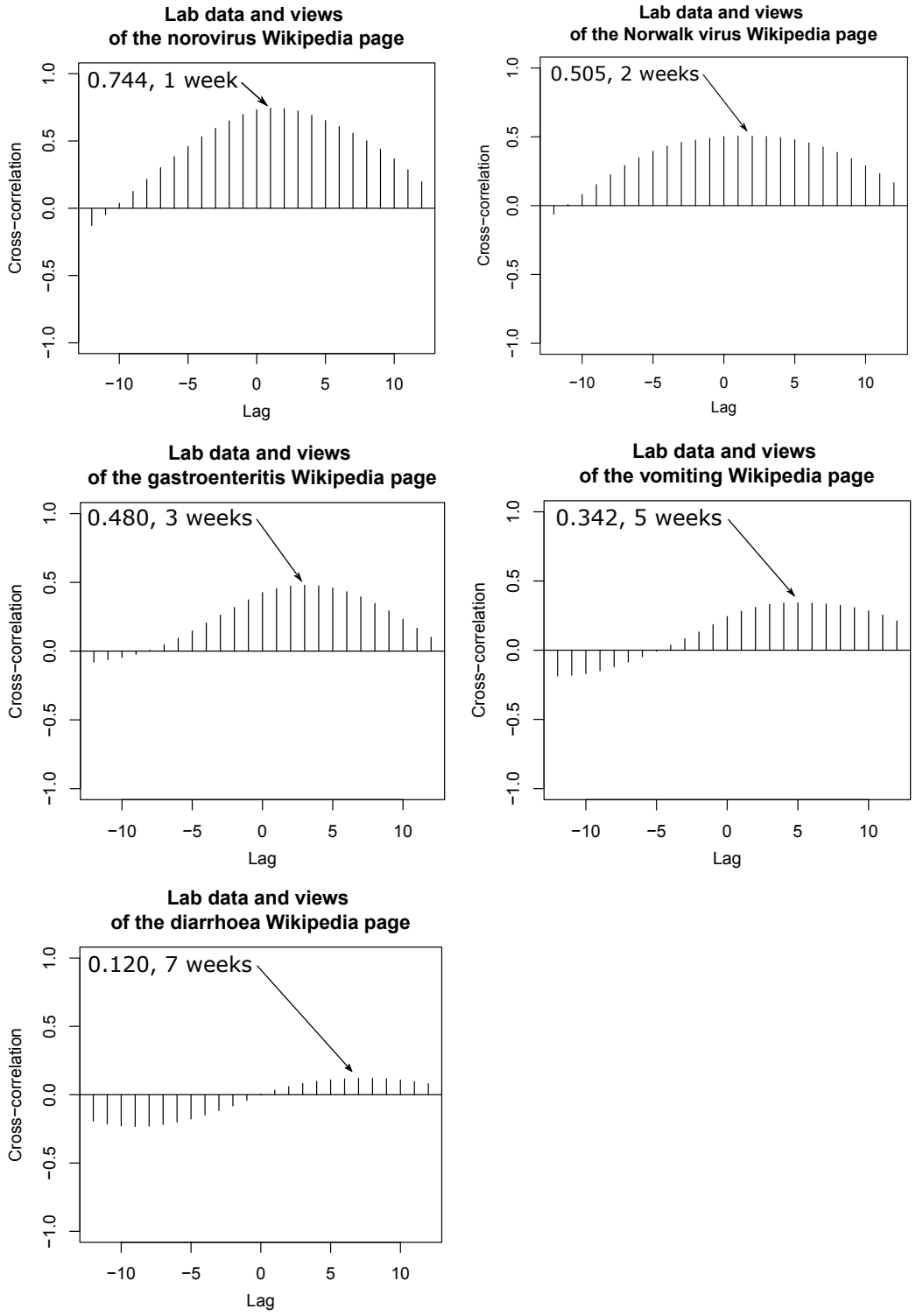


Figure 4.8: Cross-correlations between the norovirus lab data and the Wikipedia page views annotated with the maximum correlation. Note that here a lag of -1 weeks corresponds to laboratory data at week 0 being compared with page view data at week 1 (for example, trends are seen first in laboratory data and secondly in the page view data a week later). A lag of 1 week means laboratory data at week 1 being compared with page view data at week 0 (for example, trends are seen first in page view data and secondly in laboratory data a week later).

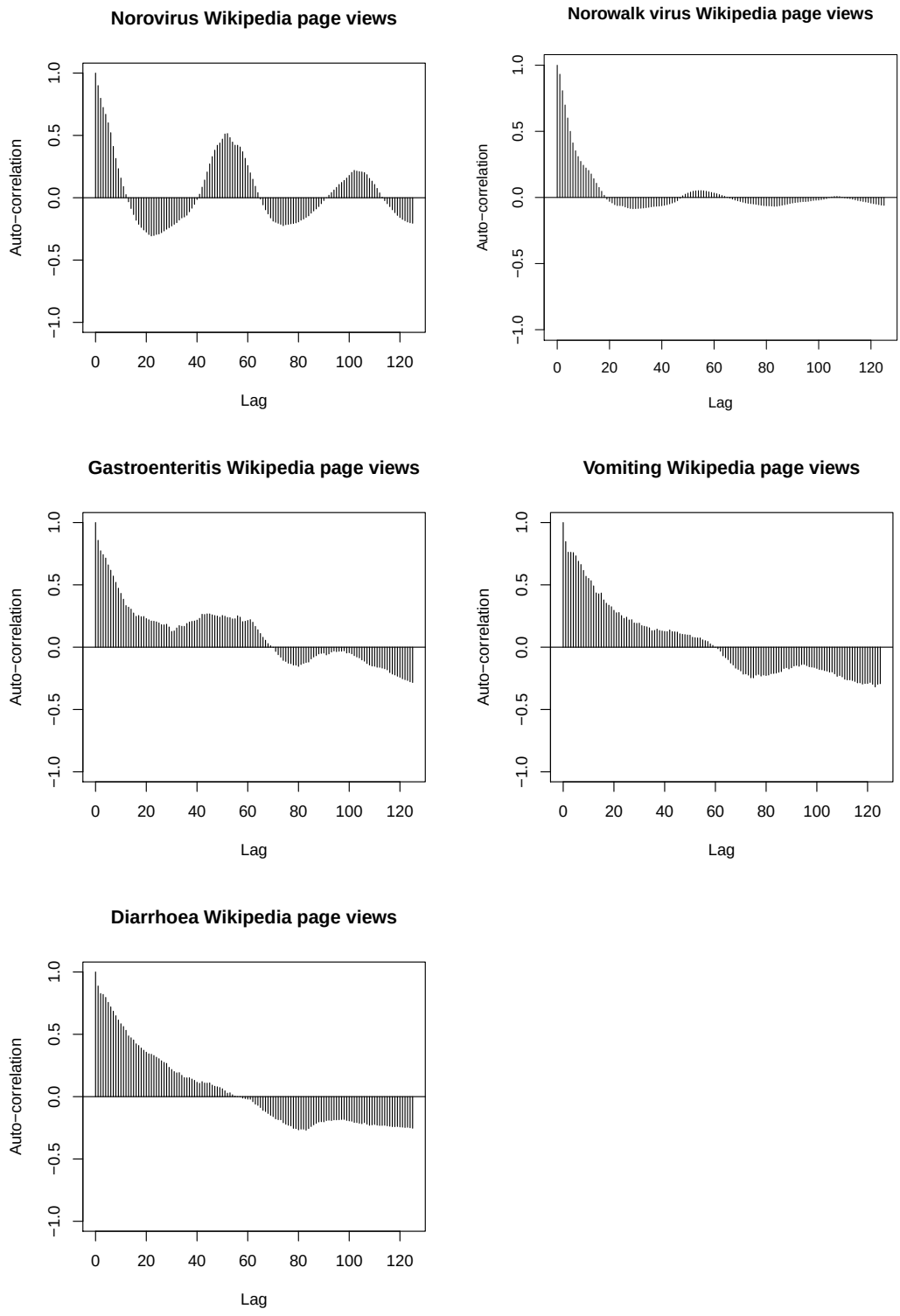


Figure 4.9: Autocorrelation plots of the Wikipedia page view data in order to identify periodicities (lag in weeks).

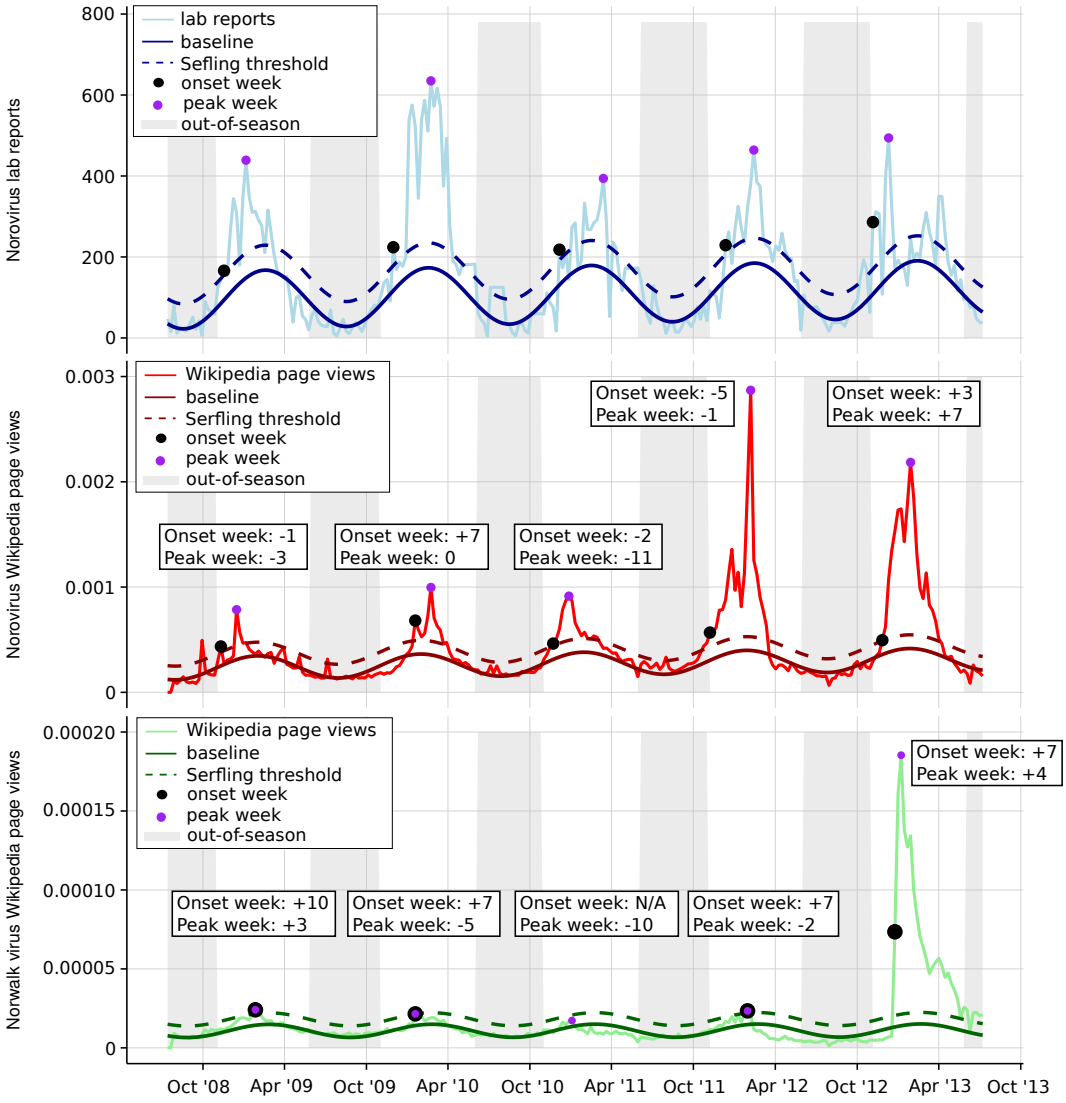


Figure 4.10: Analysis of season onset and peak times from norovirus laboratory data and Wikipedia page view datasets. Baselines were fitted using a Serfling regression model and seasonal outbreak periods defined as activity over the Serfling threshold. The numbers indicate the difference in onset, or peak, week identified in the Wikipedia data from the onset, or peak, week identified by the laboratory data. A positive number indicates that the onset, or peak, week was later in the Wikipedia data. A negative number indicates that the onset, or peak, week was earlier in the Wikipedia dataset.

4.2.5 Forecasting and nowcasting

In this section we will describe attempts to incorporate these new data sources into statistical forecasts or nowcasts of norovirus laboratory reports. For comparison, we also implement forecasting methods that do not make use of the additional data sources. We would like to see if the additional data improve our forecasting/nowcasting ability.

Predictive models that do not use the online datasets will forecast the next week of laboratory data. However, predictive models that use the online datasets will nowcast the next week of laboratory data as the current week of online data will be used for prediction. This is because there is typically a delay in when the laboratory data is available, due to the time taken to collate samples and test them, whereas the online data is available is near-to real time.

Naïve model

Simple forecasting methods are computationally cheap and can be surprisingly effective [228]. Therefore, we implement the most simple forecasting model for seasonal data as a comparison for the more complex models. The seasonal naïve model is a simple forecasting model for seasonal data (described by Hyndman and Athanasopoulos 2013, [228]). Each forecast value is equal to the previous observed value for the same season. The norovirus laboratory data are seasonal with a 52 week period. Therefore, the naïve forecast for week i is the observed value from week $i-52$. This model does not make use of the additional online data.

Reduced ARIMA model

The second method we consider that does not use the additional online data is an ARIMA model, as introduced in section 4.2.1. ARIMA models are widely used in time series forecasting [228]. We use the `auto.arima` function from the `forecast` package in R to fit this model [224].

Seasonal reduced ARIMA model

The next method we consider that does not use the additional online data is a seasonal ARIMA model (section 4.2.1) with period 52 as we have weekly data and a clear annual seasonality. We again use the `auto.arima` function from the `forecast` package in R to fit this model, however we force $D = 1$ and $m = 52$ to ensure seasonality [224].

Fourier terms with ARIMA errors

The final method that we consider that does not use the additional online data but does make use of the fact we know there is annual seasonality is a regression with Fourier terms and ARIMA errors (section 4.2.1). We do this because the seasonal period (52 weeks) is quite long. We pick the number of Fourier terms by minimising the AIC and, again, the orders of the ARIMA model are chosen with the `auto.arima` function from the `forecast` package in R.

Online only model

The most simple way to use the Wikipedia page view and Google search interest datasets to nowcast the norovirus laboratory data is to use a linear multiple regression model, as per Generous et al. (2014, [211]). We fit the linear regression model on historical laboratory data and use the current week's Wikipedia and Google data to give an estimate of the current number of lab reports. We additionally use a fixed effect (dummy variable) to distinguish the two weeks of holiday over Christmas and New Year from the rest of the year, based on the observation in section 4.2.2 that these weeks have fewer reports. We use the `lm` function from R to fit this model [219].

Full ARIMA model

A more complex way to incorporate the search interest and page view datasets into a predictive model is to include them as exogenous variables in an ARIMA model (as defined in section 4.2.1). We also include the same fixed effect for the Christmas and New Year holiday period as described in the online only model section. We make use

of the analysis of sections 4.2.3 and 4.2.4 to incorporate the exogenous variables in an educated way; we only include those variables with a maximum cross-correlation of at least 0.6, and we include them at the lead which gave this maximum. Again, we use the `auto.arima` function from the `forecast` package in R to fit this model.

Results

The reduced ARIMA, seasonal reduced ARIMA, Fourier model with ARIMA errors, online only, and full ARIMA models were fitted to rolling subsets of data so that time series cross-validation could be used to assess their performances. Initially these subsets were of two years. For each two year window and for each model a prediction was obtained (figures 4.12 and 4.13). The orders of the fitted ARIMA models are all quite small (figure 4.11); each model is relatively simple. However, note that they are not always consistent; a slightly different model is fitted for each rolling subset of training data. It appears that many of the predictions were within the 80% prediction interval. To formally compare the predictions to the actual data, the absolute errors between the predictions and the data points were calculated. The models were compared by computing the mean of these absolute errors (table 4.2). A better model has a smaller mean absolute error (MAE).

The naïve and online only models gave the largest MAEs. The reduced and full ARIMA models had similar, and the smallest, MAEs. The seasonal reduced ARIMA model and the model with Fourier terms had intermediate MAEs. Therefore, we conclude that including differencing and previous values gives a better prediction of the number of norovirus lab reports, but that also explicitly including annual seasonality and information from the online data sources does not change the predictive ability of this kind of model. The mean number of norovirus reports per week during the time period we made forecasts was 148, and the maximum was 494. Therefore, an average error of around 40 norovirus cases is reasonably small.

We also considered other lengths of training period. These gave similar results. However for a training period of just one year, the full ARIMA model did not perform as well as the reduced ARIMA model (MAE full ARIMA = 56.4, MAE reduced ARIMA = 48.4 for one year training period).

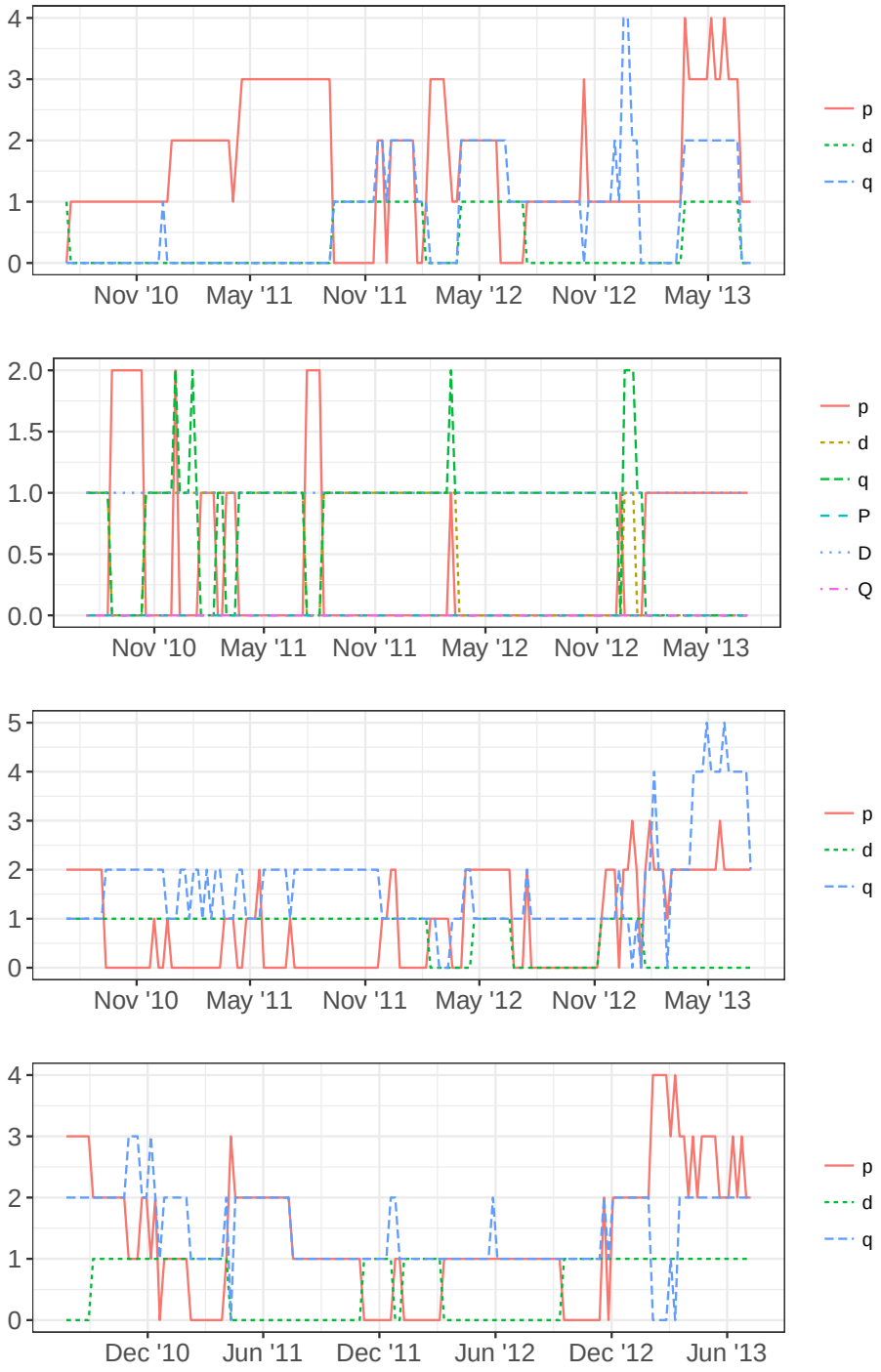


Figure 4.11: The orders of the ARIMA models compared for forecasting or nowcasting the number of norovirus lab reports. From top to bottom: reduced ARIMA model, seasonal reduced ARIMA model, Fourier model with ARIMA errors, full ARIMA model.

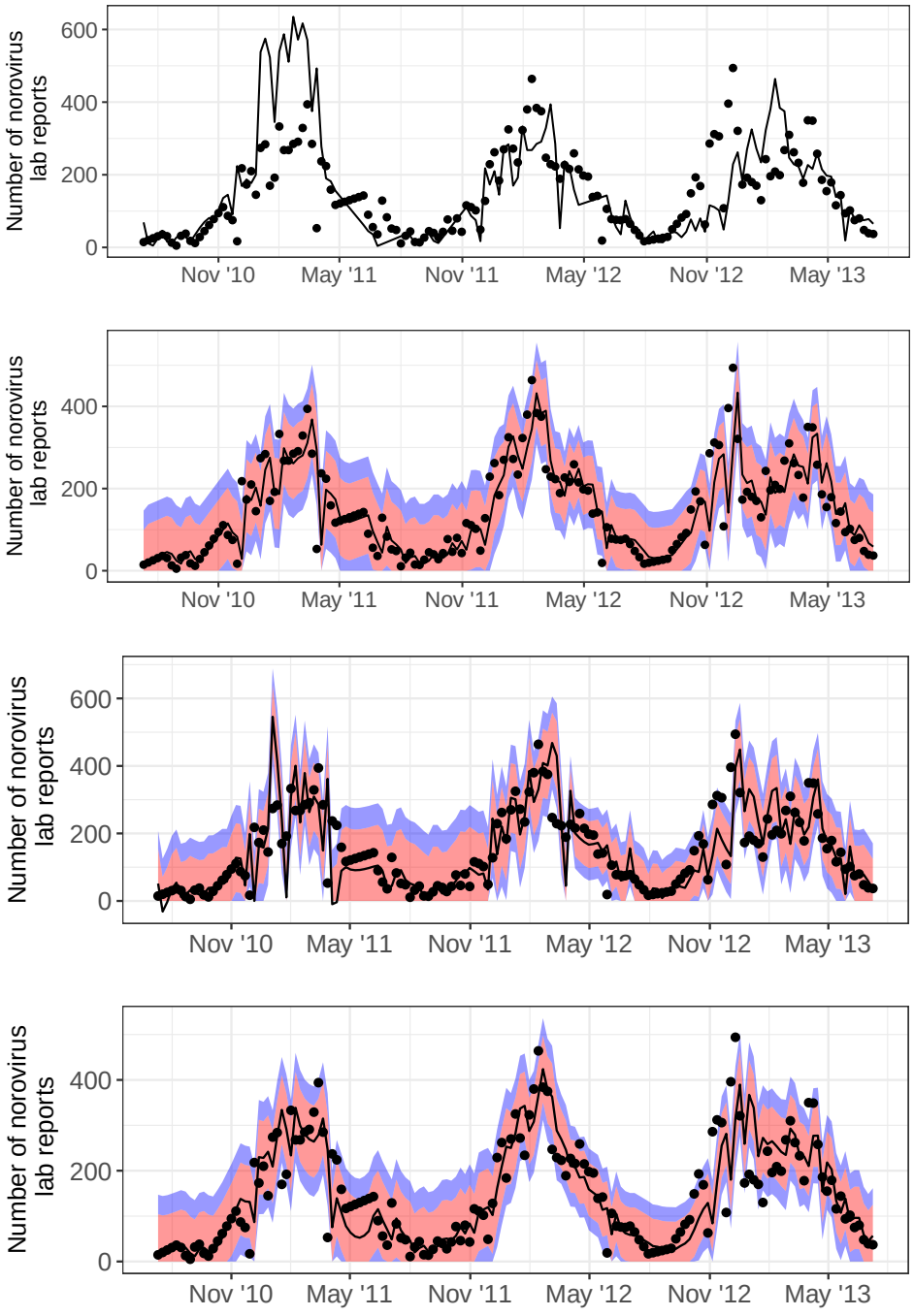


Figure 4.12: The number of norovirus lab reports (black dots) with the predictions from the forecasting/nowcasting models (black lines) based on fitting to the previous two years of data with 80% (red) and 95% (blue) prediction intervals. From top to bottom: naïve model, reduced ARIMA model, seasonal reduced ARIMA model, Fourier model with ARIMA errors. Figure continued on next page.

Table 4.2: Mean absolute error of the models for a training period of two years

Model	Mean absolute error (MAE)
Naïve	69.5
Reduced ARIMA	42.8
Seasonal reduced ARIMA	54.0
Fourier terms with ARIMA errors (reduced)	47.1
Online only	63.5
Full ARIMA	44.0

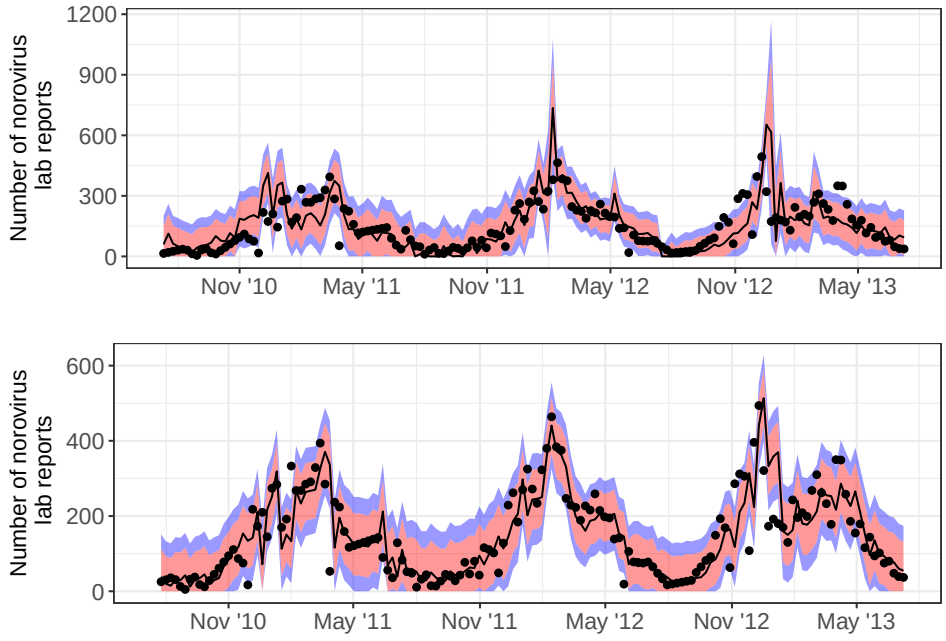


Figure 4.13: Figure continued from previous page. The number of norovirus lab reports (black dots) with the predictions from the forecasting/nowcasting models (black lines) based on fitting to the previous two years of data with 80% (red) and 95% (blue) prediction intervals. From top to bottom: online only model, full ARIMA model.

4.2.6 Search engine queries and page view data: Discussion and conclusions

We have found that data from Google search interest and Wikipedia page views have some similarity to the time series of norovirus laboratory reports. However, no one particular time series from these online datasets is well matched to all aspects of the laboratory data. The addition of these new data to forecasting models do not improve the predictive ability and simply result in more complex models. In conclusion, based on this investigation data from Google search interest and Wikipedia page views will not add value to norovirus laboratory surveillance.

The idea to do this analysis, and the majority of the work described in this section, was completed in spring and summer 2014. This field has progressed rapidly since then. In 2014 this work was more novel than it may, perhaps, be considered now. Indeed, the first key papers using Wikipedia page view data for healthcare surveillance were published in April and November 2014 [210, 211].

Many analyses of internet based data for surveillance of an illness compare a novel data source to existing traditional data (the ground truth, described in section 4.2.2) with the aim of identifying data and models that will give earlier and easier detection of changes in patterns. However, this assumes that the ground truth accurately reports the levels of illness in the population of interest. In some limited cases, for example severe notifiable diseases with specific syndromes or for surveillance of a small, well monitored population, this may be true. However, we do not feel the norovirus laboratory dataset is a good ground truth dataset for the total burden of norovirus.

In particular, positive norovirus laboratory reports only reflect a small proportion of community cases [229, 230]. We expect this to consist of the most severe cases, cases associated to outbreaks undergoing investigation, and cases in at risk populations such as the elderly and children. These cases may not have the same incidence patterns over time as the full burden of norovirus. Additionally, there are anecdotal reports of sampling and reporting behaviours for norovirus changing over time. For example, once doctors are aware that the norovirus season has begun they may be less likely to request laboratory confirmation of symptoms. This will create an artificially more peaky dataset.

Further work: search engine queries and page view data

Immediate further work on this study could be to consider a wider range of search terms and Wikipedia pages; we only considered here a limited selection. However, this should be done in a principled way so that only terms connected to norovirus are included. The process of simply considering all search terms and looking for a good match is flawed and has faced criticism when used in the past for influenza surveillance by Google Flu, for example from Lazer et al. (2014, [200]). In particular, this is a challenge when the strongest signal in the data is a winter peak, like we have in the norovirus laboratory data, as there are many winter seasonal activities.

The Serfling method, with a harmonic regression model, was used to define a baseline activity level for norovirus and a threshold over which we consider to be outbreak activity. This was fitted to the out-of-season data only, and upon visual inspection fitted reasonably well in most instances. This was not the most simple option for fitting a baseline, for example a constant baseline could have been defined as the mean of all out-of-season data. However, we could do further work to improve this model. For example, we could more rigorously assess the fit of the baseline to the out-of-season data in order to be able to formally describe the performance of the baseline. We could additionally add in more harmonic terms to the regression model to take into account more seasonality.

We were not able to compare Google search interest data over multiple years as the data are only available weekly for one year at a time and each download is scaled to be between 0 and 100. However, if we did want some measure of which seasons were measured as severe from this data, we could analyse monthly data which are available for the full time period of interest.

There are further forecasting methods that we have not considered, such as exponential smoothing and neural network models. Additionally, for a more formal comparison of forecasting methods we could have used a statistical test, such as the Diebold-Mariano test, to compare forecast errors. This would have also allowed us to formally test many lengths of forecast window. Finally, we only considered nowcasts or forecasts one week ahead. We could investigate the ability of the different models to forecast further into the future, and in particular we suspect that the models with explicit annual seasonality may perform well in this context.

Finally, we believe that using these analysis methods with page view data from

the NHS website will give interesting results as this is a trusted source of healthcare information in the UK. However, this requires collaborative efforts that were beyond the scope of this PhD.

Data weaknesses

We believe that the main weaknesses of this study were due to restricted access to these datasets.

PHE makes the weekly number of positive laboratory test results for norovirus publicly available online. However, there are some periods of missing data in the public reports. We used the most simple method to overcome this, by estimating missing data points with a linear interpolation between the nearest non-missing data points. However, as the missing data points are clustered in time, this leads to periods of estimated unrealistic linear activity. Fortunately however, most of the missing data points are not in the main norovirus outbreak season.

Google does not make raw data available. There were mismatches in both timing and location between the Google data we could access and the laboratory data. As previously discussed, the Google week begins on Sunday and the laboratory week begins on Monday. Additionally, the Google data were restricted to searches within the UK whereas the laboratory data covers reports from England and Wales. However, as the other parts of the UK have a similar climate and social mixing to England and Wales we suspect that the patterns of disease in these areas will be similar to those of England and Wales.

There was, unfortunately, no location information associated with the Wikipedia data, as previously described. This is a current major limitation of using these data for disease surveillance. McIver et al. (2014, [210]) reported that 41% of Wikipedia English language article views come from the U.S. This certainly will have had some impact on the results of this analysis.

A full discussion of the values of open data and data sharing within disease surveillance is beyond the scope of this PhD.

Conclusions: search engine queries and page view data

We have demonstrated that data from Google Trends and Wikipedia page views do not currently add value to forecasts of the norovirus laboratory time series. However, without knowing the motivation of the people creating these data, and without an actual measure of norovirus incidence for comparison, we cannot say whether these data make any other measurement of norovirus burden.

4.3 Surveillance of gastroenteritis using an online participatory influenza surveillance system

In the remainder of this chapter we will investigate reports of gastroenteritis from the community that may have not been reported to healthcare services, using the online participatory influenza surveillance system *Flusurvey*. This will extend our picture of gastroenteritis burden beyond that which is measured by existing PHE surveillance systems.

Many aspects of disease surveillance in the UK are considered world leading [231]. For the surveillance of gastroenteritis, there is syndromic surveillance of consultations with general practitioners (GPs) and national laboratory surveillance systems that record the number of stool samples that test positive for norovirus (reportedly the leading cause of gastrointestinal disease in the UK [232]).

However, the incidence of gastroenteritis is underestimated by these surveillance systems since not everyone presents to healthcare services when they have gastrointestinal symptoms [192, 230]. It is acknowledged that those who do seek healthcare advice for gastroenteritis, and get reported to national surveillance, is a biased sample of the population [230]. For example, patients with more severe illness, recent foreign travel, and lower socio-economic status are over-represented in gastroenteritis cases reported to GPs (IID1 study, 2003, [229]). Additionally, it is generally acknowledged that healthcare facilities are used more by young children, and that men can be more reluctant to seek professional healthcare than women [233–235].

Therefore, in use in order to record cases of gastroenteritis in patients who do not actively seek healthcare services it is necessary to extend the current surveillance systems. This is particularly important for gastroenteritis which is self-limiting and as NHS advice is to avoid going to GP services [232].

In this section we will explore reports of gastroenteritis made by the Flusurvey cohort. These are reports of gastroenteritis from the community and will not necessarily have been reported to any other national surveillance system. We will compare these reports with positive laboratory specimens for norovirus and with gastroenteritis syndrome reports from GP based syndromic surveillance systems. We start with a review of existing studies and of the statistical methods we will use.

4.3.1 Flusurvey: Background and methods

Background: Community cohort studies

Community cohort studies can give an estimate of the incidence of disease in the community. In a community cohort study a sample of the population is recruited for monitoring over time [236]. A key advantage of community cohort studies is that cases of illness are recorded even if individuals do not present to healthcare services. Cohort studies have been used in all areas of healthcare. For example, to measure the prevalence of urinary incontinence [237], the mortality rate of individuals with schizophrenia [238], and future healthcare attendances of patients with pneumonia [239].

However, recruiting, regularly contacting, and maintaining the cohort can be a costly and time-consuming endeavour [240, 241]. Therefore, large-scale, prospective community cohort studies of gastroenteritis are uncommon, particularly those measuring community incidence through laboratory confirmation. The first and second studies of infectious intestinal disease in the community (IID1 and IID2) were, however, two such studies in the UK [36, 37]. IID1 was completed during the 1990s and consisted of, among several other related studies, a population based community cohort study with weekly postal reports of either no symptoms or stool samples for laboratory testing for six months. IID2 was completed during the first decade of the 21st century. It again consisted of several related studies including a retrospective community cohort telephone survey and a prospective weekly community cohort study with monitoring carried out via email or post for one year and stool samples taken for laboratory testing upon the onset of symptoms. Using the data collected from the studies, estimates were made of the burden of infectious intestinal disease in England and the factor by which the number of cases identified by surveillance systems should be multiplied to estimate the actual number of infections in the community. Similar studies have also taken place outside of the UK. For example, *Sensor* was a population based cohort study on gastroenteritis incidence in the Netherlands [242].

Studies such as those described above provide one-off estimates of disease prevalence as opposed to continual, real time (or near-to real time) disease surveillance. In the last decade, community cohort surveys using a regularly completed online questionnaire have become more common. Members of the general public volunteer to join and regularly report symptom information, typically weekly. The symptom

reports are used to infer the presence of a particular illness [243].

These systems still have the benefit of a traditional community cohort study, in so much that data will be collected from cases that do not necessarily report to healthcare services, but due to having no laboratory confirmation and being fully online they are less costly and more timely [243]. Additionally, validation checks can be incorporated into the questionnaire so that important fields are not left empty, data are collected electronically so do not need to be entered manually after paper collection, and participants have been reported to typically reply rapidly to email requests [244, 245]. There are, of course, downsides to these systems. The people that chose to take part may not be representative of the general population (in particular, children and the elderly are less likely to use the internet) and disease presence can never be confirmed as the only data collected are self-reported symptoms [243].

While there is no long-running web-based cohort study primarily measuring the community incidence of self-reported gastroenteritis in the UK, *Flusurvey* (`https://flusurvey.org.uk/`) is an internet-based participatory surveillance system that has been running since 2009 to monitor ILI in the UK [246, 247]. Participants are recruited into Flusurvey for each influenza season (November to April). Any member of the UK public can take part. Participants are initially required to complete a background questionnaire and then each week an email is sent asking them to indicate which, if any, of a given list of symptoms they had experienced in the last week. Participants are asked to submit an empty symptom survey if they had no symptoms. The symptoms diarrhoea and vomiting are included within these weekly surveys, but as of now these data have not been collated or analysed. Any Flusurvey participants reporting symptoms are additionally asked if they sought healthcare advice [233]. These data can be used to give an estimate of how to scale traditional healthcare based surveillance systems to give an estimate of community burden. Further full details of the Flusurvey system are reported elsewhere [246].

Flusurvey is part of a consortium of similar online participatory ILI surveillance systems, called *Influenzanet*, covering approximately 10 countries in Europe [248]. Similar systems also exist in Australia [249], Mexico (`http://reporta.c3.org.mx/`), and in the U.S. [250]. The first system of this type designed for surveillance of an illness other than ILI was *Dengue Na Web*, based in Brazil, to monitor dengue activity [243]. *SaludBoricua*, in Puerto Rico, monitors dengue, ILI, leptospirosis, and chikungunya symptoms simultaneously, demonstrating the flexibility of these systems to monitor multiple illnesses (`https://saludboricua.org`).

However, these systems do not seem to be widely used for gastroenteritis surveillance. We have only found one such example (published years after the analysis reported in this chapter had taken place); the Influenzanet system in Sweden was used to estimate the community incidence of gastrointestinal illness, as well as ILI and acute respiratory illness (Pini et al. 2017, [251]). Data from this online system were compared to data from search terms to a medical website, reasons for calling a medical advice hotline, and laboratory notifications for norovirus using Spearman correlation coefficients.

Method: non-parametric bootstrap confidence intervals

Non-parametric bootstraps can be used to compute confidence intervals on parameters, such as means, estimated from data [252, 253].

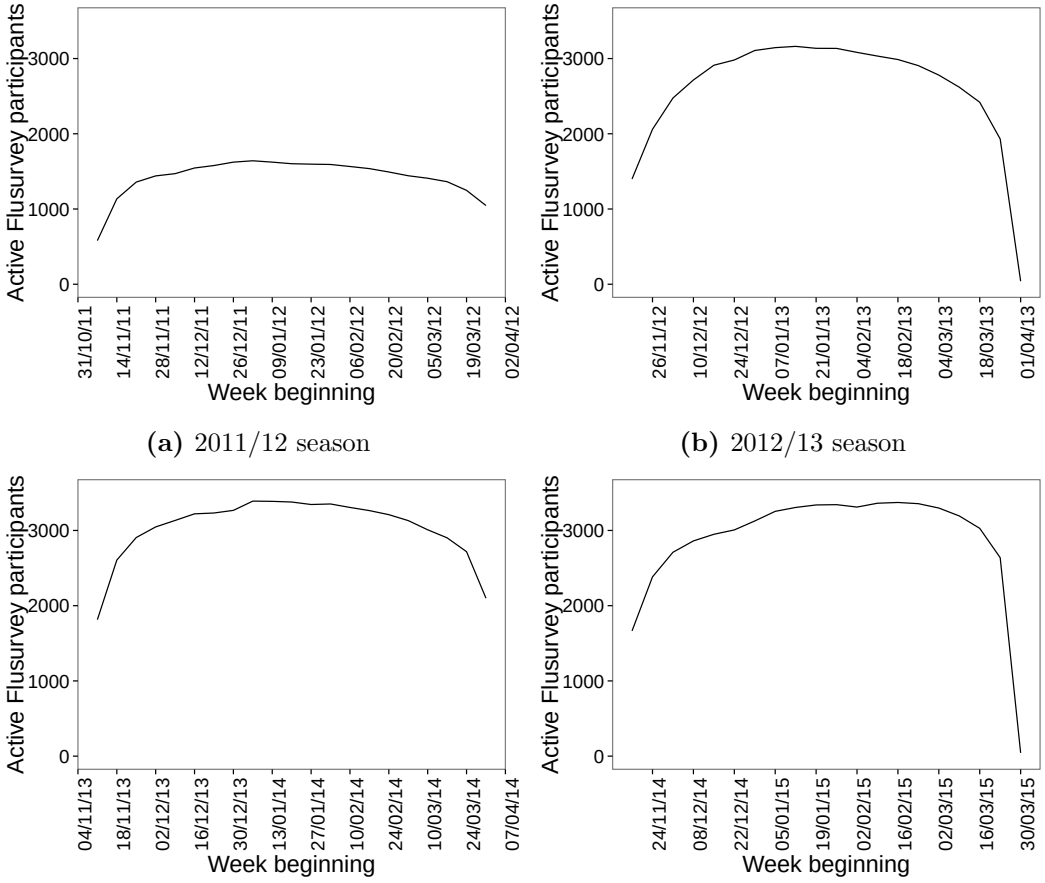
Given a collection of data points, $x_1, ..., x_n$, a bootstrap sample, $x_1^*, ..., x_n^*$, is simply obtained by randomly sampling n times with replacement from the original data points. To calculate, for example, a 95% confidence interval on the mean of $x_1, ..., x_n$ we need an estimate for how much the distribution of the sample mean $\overline{x}$ varies around the population mean μ, that is $\delta = \overline{x} - \mu$. In a bootstrap confidence interval, δ is approximated by $\delta^* = \overline{x}^* - \overline{x}$, where $\overline{x}^*$ is the sample mean computed from a bootstrap sample. We compute many bootstrap samples and δ^* for each. The 2.5^{th} and 97.5^{th} percentile of the collection of δ^* give the confidence interval.

4.3.2 Gastroenteritis from Flusurvey

Gastroenteritis reports

An extraction of the cleaned Flusurvey dataset was provided for this analysis. The data were cleaned in the same way as in previous analyses of Flusurvey reports (see, for example Tilston et al. 2010, [246]). The first symptom report of each participant was excluded and only participants who then had submitted at least two further symptom reports were included in the analysis. This is to reduce the effect of participants who sign up just in response to their current symptoms.

A participant was considered active on any week between their first and last symptom reports in a season. Therefore, on any given week, the number of active partic-



(a) 2011/12 season **(b)** 2012/13 season

(c) 2013/14 season **(d)** 2014/15 season

Figure 4.14: The number of active Flusurvey participants each week for each season.

ipants was not necessarily equal to the number of symptom reports submitted. We used the number of active participants as the denominator for computing incidence rates. We do this even though participants were asked to submit empty symptom reports if they had no symptoms as it is probable that participants were more motivated to complete reports on those weeks when they experienced symptoms. This assumption has been used in previous analyses of Flusurvey data [246].

Data were provided for the 2011/12 season (week beginning 07/11/2011 to week beginning 26/03/2012), the 2012/13 season (week beginning 19/11/2012 to week beginning 01/04/2013), the 2013/14 season (week beginning 11/11/2013 to week beginning 31/03/2014), and the 2014/2015 season (week beginning 17/11/2014 to week beginning 30/03/2015).

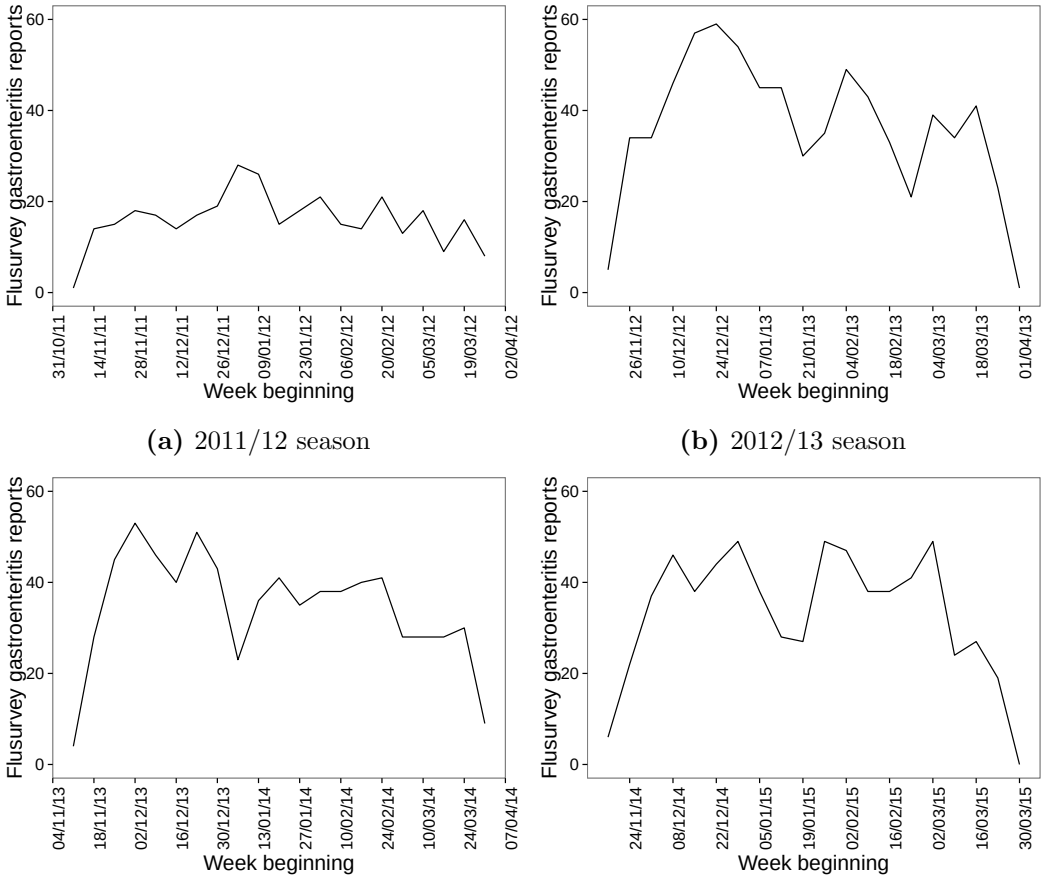
During the 2011/12 season the mean number of active Flusurvey participants per week was 1423 (95% confidence interval (CI) [1307, 1516]). The was much lower than the subsequent three seasons, which had on average 2601 (95% CI [2237, 2886]), 3034 (95% CI [2843, 3193]), and 2877 (95% CI [2495, 3158]) active participants per week respectively (figure 4.14). The number of active participants was reasonably stable each season, except for a steep increase, and drop off, in the first, and last, few weeks respectively.

The suggested standard symptom-based case definition of gastroenteritis given by Majowicz et al. (2008, [1]) is "*an individual with three of more loose stools, or any vomiting, in any 24 hours*" and excluding those with pre-existing medical conditions, for example irritable bowel syndrome, and causes due to drugs, alcohol, or pregnancy. We are unable to assign causes to any symptoms recorded by Flusurvey participants. Therefore we could only use an adjustment of this definition of gastroenteritis. We classified all Flusurvey participants who recorded at least one of the symptoms diarrhoea and vomiting in a symptom survey as reporting gastroenteritis.

The mean weekly number of symptom surveys classified as gastroenteritis was 16 (95% CI [14, 18]) during 2011/12, 36 (95% CI [30, 43]) during 2012/13, 35 (95% CI [29, 39]) during 2013/14, and 33 (95% CI [27, 39]) during 2014/15 (figure 4.15). For comparison, the mean weekly number of Flusurvey symptom surveys classified as ILI was 35 (95% CI [30, 41]) during 2011/12, 71 (95% CI [58, 83]) during 2012/13, and 68 (95% CI [58, 77]) during 2013/14, and 76 (95% CI [63, 88]) during 2014/15.

In order to take into account changes in the size of the reporting cohort we computed the gastroenteritis incidence rate from Flusurvey using the number of active participants each week (figure 4.16). The gastroenteritis rate quickly increased to a peak rate during December, and then remained at either around this rate, or slightly lower, for the rest of the season. The average gastroenteritis incidence rate from Flusurvey was 0.011 (95% CI [0.009, 0.012]) in 2011/12, 0.014 (95% CI [0.012, 0.016]) in 2012/13, 0.011 (95% CI [0.010, 0.013]) in 2013/14, and 0.011 (95% CI [0.009, 0.013]) in 2014/15. In 2011/12 the maximum rate occurred at week 9 of the Flusurvey season. In 2012/13 it was week 20 and in both 2013/14 and 2014/15 it was week 4.

Upon registration, Flusurvey participants completed a background questionnaire giving their age. Based on this, we calculated the gastroenteritis incidence rate by age group (under 19 years, 19-45 years, 46-65 years, and over 65 years) (figure 4.17



(a) 2011/12 season

(b) 2012/13 season

(c) 2013/14 season

(d) 2014/15 season

Figure 4.15: The number of Flusurvey symptom reports classified as gastroenteritis each week and for each season.

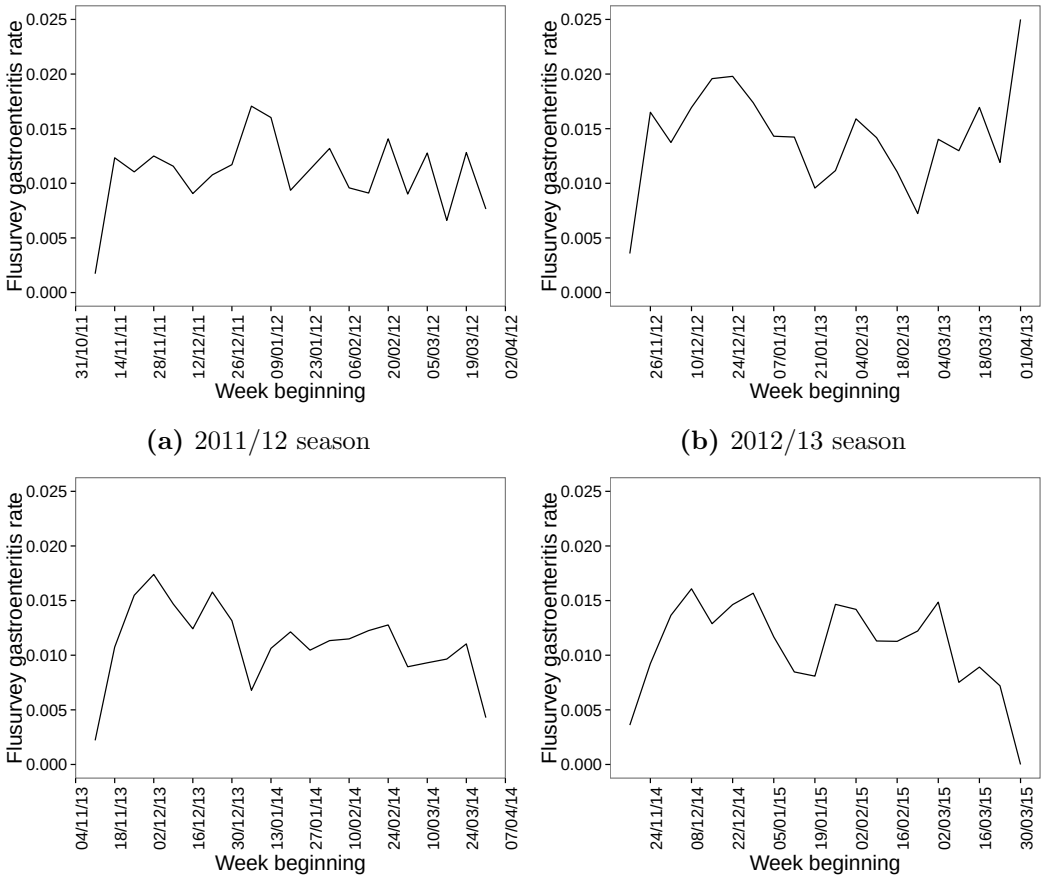
with averages by season given in table 4.3). Those aged under 19 had the highest gastroenteritis incidence rate recorded by Flusurvey (figure 4.17). In particular during November and December 2012 the rate in this age group was nearly double the Flusurvey gastroenteritis rate in any age group at any other time. The over 65 years age group had a lower incidence rate than the other age groups.

Proportion seeking healthcare advice

Flusurvey participants reporting any symptoms were asked subsequent questions on whether they contacted a medical professional. This included contact with GP services, out of hours services, or hospital services either by telephone, on the internet, or in person. In order to calculate the proportion of gastroenteritis cases that sought

Table 4.3: The average gastroenteritis incidence rate for Flusurvey participants by age group and for each season.

Season	Age group (years)	Mean rate (95% CI)	Maximum rate	Week of maximum rate
2011/12	Under 19	0.010 [0.004, 0.018]	0.067	8
	19 - 45	0.013 [0.011, 0.015]	0.024	16
	46 - 65	0.009 [0.007, 0.011]	0.018	5
	Over 65	0.007 [0.005, 0.010]	0.019	20
2012/13	Under 19	0.026 [0.020, 0.033]	0.063	2
	19 - 45	0.016 [0.014, 0.019]	0.025	20
	46 - 65	0.012 [0.009, 0.014]	0.022	5
	Over 65	0.009 [0.006, 0.011]	0.023	6
2013/14	Under 19	0.017 [0.013, 0.021]	0.033	4
	19 - 45	0.014 [0.011, 0.016]	0.022	4
	46 - 65	0.009 [0.007, 0.011]	0.018	16
	Over 65	0.006 [0.004, 0.008]	0.019	13
2014/15	Under 19	0.017 [0.013, 0.022]	0.041	11
	19 - 45	0.014 [0.012, 0.016]	0.021	4
	46 - 65	0.010 [0.008, 0.011]	0.016	6
	Over 65	0.006 [0.004, 0.009]	0.017	5



(a) 2011/12 season

(b) 2012/13 season

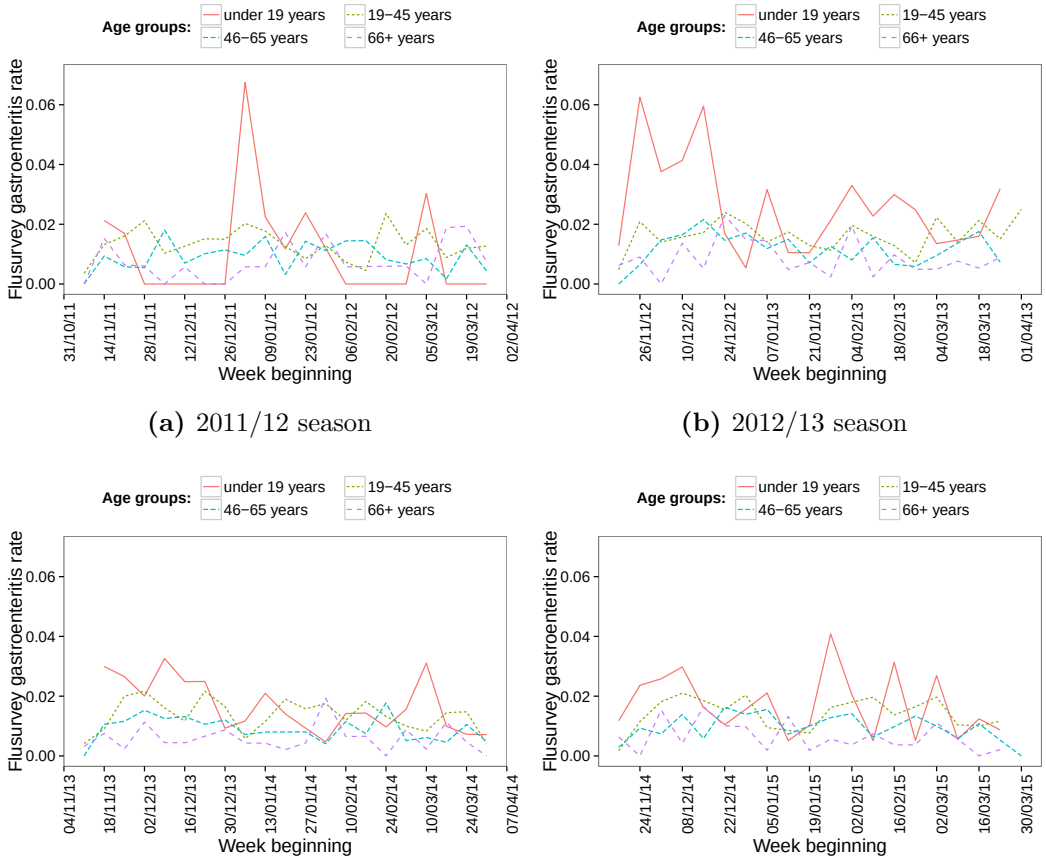
(c) 2013/14 season

(d) 2014/15 season

Figure 4.16: The Flusurvey gastroenteritis rate from dividing the number of gastroenteritis symptom reports by the number of active participants each week and for each season.

healthcare advice we aggregated the responses to these questions from participants who reported gastroenteritis symptoms. The mean percentage of gastroenteritis cases seeking healthcare advice, and 95% bootstrapped confidence intervals, were computed for each of the Flusurvey seasons. We were unable to stratify by age due to small numbers of reports in some age groups.

13.9% (95% CI [12.2%, 15.7%]) of Flusurvey participants with gastroenteritis sought medical attention (figure 4.18). A higher percentage of participants sought healthcare advice in the 2013/14 and 2014/15 Flusurvey seasons compared to the 2011/12 and 2012/13 seasons (10.9%, 95% CI [8.0%, 13.9%] in 2011/12, 11.7%, 95% CI [8.9%, 14.4%] in 2012/13, 17.3%, 95% CI [13.5%, 21.9%] in 2013/14, and 15.8%, 95% CI [13.0%, 18.9%] in 2014/15).



(a) 2011/12 season

(b) 2012/13 season

(c) 2013/14 season

(d) 2014/15 season

Figure 4.17: The Flusurvey gastroenteritis rate each week broken down by age group and for each season.

4.3.3 Comparisons with other surveillance systems

We compared the gastroenteritis incidence rate measured by the Flusurvey system with the incidence of gastroenteritis measured by three surveillance systems maintained by PHE: the GP in-hours (GPIH) syndromic surveillance system (SSS) (described in section 3.1.1), the GP out-of-hours (GPOOH) SSS (described in section 3.1.1), and laboratory confirmed reports of norovirus (described in section 4.2.2). We could only undertake this analysis for the 2012/13 and 2013/14 Flusurvey seasons due to limitations in the available data from the comparison systems.

We extracted daily data on the number of consultations coded as gastroenteritis from both the GPIH and GPOOH SSSs, and summed these syndromic indicators

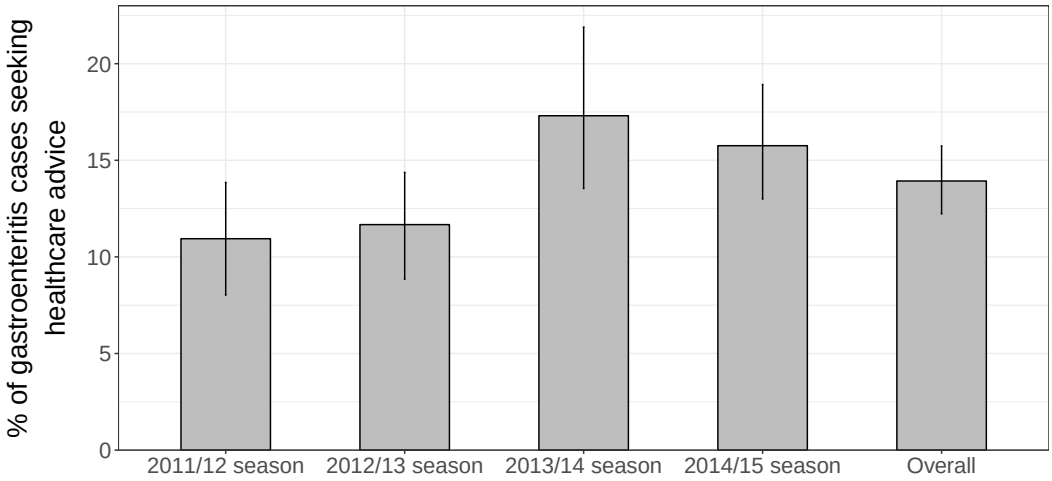


Figure 4.18: The proportion of Flusurvey participants with syndromic gastroenteritis that sought healthcare advice during the 2012/13 and 2013/14 seasons. The bar gives the mean and the vertical lines indicates 95% confidence intervals on the mean, obtained by bootstrapping.

into weekly measures for comparison with the Flusurvey gastroenteritis incidence. From the GPOOH SSS we also extracted the total number of consultations per day and constructed a weekly rate of gastroenteritis, to account for the changing use of the system throughout time [95]. The gastroenteritis rate was additionally calculated for separate age groups (<1 years, 1-4 years, 5-14 years, 15-24 years, 25-44 years, 45-64 years, 65-74 years, and 75+ years).

It is not possible to obtain a total number of consultations per day from the GPIH system. Instead the total number of patients registered with GP services taking part in the system was extracted and used to compute a gastroenteritis rate per 100,000 registered patients. This accounts for changes in the number of GP practices reporting to the system each day, but cannot account for day-to-day changes in the use of the system due factors other than illness levels. This is standard practice within the GPIH SSS [94]. The rate per 100,000 registered patients was additionally calculated by age group (<1 years, 1-4 years, 5-14 years, 15-44 years, 45-64 years, 65-74 years, 75+ years).

Norovirus is recognised as the leading cause of gastrointestinal illness in the UK. This motivates comparison of the Flusurvey gastroenteritis rate with the norovirus laboratory data. The number of norovirus laboratory reports was also extracted by age group (<5 years, 5-64 years, 65+ years).

We compared the datasets visually and also more formally by computing the cross-correlation of the different measures of gastroenteritis up to plus and minus 6 weeks (cross-correlation defined in section 4.2.1). A parametric bootstrap was used to construct 95% confidence intervals on the correlation at each lag. Each value in the two time series being compared was assumed to be a sample from a Poisson distribution, with mean at the observed value. 1000 bootstrap samples were generated from these distributions, and the cross-correlation between each pair of bootstrap samples was used to give an estimate of the 95% confidence interval on the correlations. This computationally intensive approach was chosen to make minimal assumptions about the underlying latent process from which we assume the data were sampled. A Poisson distribution was chosen as we have count data.

The general trend of the weekly gastroenteritis rate per 100,000 registered patients from the GPIH SSS showed a peak over winter 2012/13, a drop during summer 2013 and then an increase through the rest of the data (figure 4.19 A). Over the two years, there was a slight increase in the gastroenteritis rate given by this system, as opposed to the Flusurvey rate which was higher during 2012/13 than 2013/14. Those aged up to 5 years had the highest, and most seasonal, GPIH gastroenteritis rate per registered population (figure 4.19 B). Most of the consultations for gastroenteritis with the GPIH services were in these age groups, whereas the bulk of the adult population (those aged 15 to 64) had only a very small rate per registered population.

There were reasonably large positive correlations between the Flusurvey gastroenteritis rate and the GPIH rate during the 2012/13 season at a lag of 1 and 2 weeks (figure 4.22 A). There were no significant correlations to note between these two rates in the 2013/14 season (figure 4.22 B). This indicates some similarity between the two rates but no consistent patterns.

The weekly rate given by the gastroenteritis indicator from the GPOOH SSS peaked twice during the 2012/13 winter season and was more flat throughout 2013/14 (figure 4.20 A). Overall, it was more obviously seasonal than the Flusurvey rate. The three youngest age groups (under 1 year old, 1-4 years old, and 5-14 years old) had the highest rates, similarly to the gastroenteritis rate from Flusurvey, however there was a bigger difference between age groups in the GPOOH rate (figure 4.20 B). Due to restricted data access we were, unfortunately, not able to compute the cross-correlation of the Flusurvey gastroenteritis rate with the GPOOH data.

The number of laboratory confirmed cases of norovirus was clearly seasonal with

a peak of cases each winter (figure 4.21 A). There were significantly more cases in 2012/13 than 2013/14. The peak laboratory confirmed norovirus incidence in 2012/13 was approximately twice as high as the peak incidence in 2013/14. The 2012/13 season incidence was highest during December, which coincided with the high Flusurvey gastroenteritis rate in those aged under 19 years old. Most of the cases of norovirus were in those aged 65 years and over, in contrast to the Flusurvey gastroenteritis rate which was lowest in this age group (figure 4.21 B). There was a very weak positive correlation between the Flusurvey gastroenteritis rate and the laboratory confirmed norovirus incidence in 2012/13 at a positive lag and in 2013/14 at a lag of 2 weeks, indicating a very small amount of similarity between the trends in gastroenteritis incidence measured by the two systems (figure 4.22 C and D).

4.3.4 Flusurvey data: Discussion and conclusions

The data and analysis presented here show that the online community cohort study Flusurvey receives sufficient reports of gastroenteritis symptoms to be considered for use for syndromic gastroenteritis surveillance.

There were some broad similarities between the Flusurvey gastroenteritis rate and the rates from other surveillance systems. In particular, the GPOOH and norovirus laboratory surveillance had higher incidence early in the 2012/13 winter season than in the 2013/14 season, which corresponded with high Flusurvey gastroenteritis rates for those aged under 19 years old. At this time a new norovirus genotype was circulating in the UK [254, 255].

However, there are some stark differences in the age distributions and the extent of the seasonality between the different surveillance systems of gastroenteritis. The rate given by the Flusurvey symptom reports was less peaked than the rates seen in other surveillance systems. The Flusurvey gastroenteritis rate in children was higher than the rate for other age groups, although the differences between age groups was less pronounced compared to the other surveillance systems. The clear majority of laboratory confirmed norovirus cases were in people over 65 years, and this was very seasonal. This may be due to laboratory confirmed cases of norovirus primarily arising from samples collected in hospitals and nursing homes where the elderly are over-represented [256]. We know that individuals aged between 35 and 64 years are over-represented in the Flusurvey cohort compared to the UK population [247]. However, the Flusurvey cohort gives the opportunity for surveillance

of this section of the population, which is potentially under-represented in other surveillance systems.

The correlations between the Flusurvey gastroenteritis incidence rate and the other measures of gastroenteritis were relatively low, indicating again that there are differences between the different measures. However, only we used a Poisson parametric bootstrap for the cross-correlation computations. For future work, we will investigate whether the data were overdispersed. If so, we could use, for example, a negative-binomial distribution for the parametric bootstrap intervals.

We estimate from the Flusurvey data that around 14% of those with gastroenteritis symptoms seek healthcare advice. This additional knowledge of healthcare seeking behaviour can help determine the number of community cases from existing GP surveillance systems. For comparison, the IID2 study (Tam et al. 2012, [230]) found that for every 1 case of infectious intestinal disease presenting to GPs there were 15.5 in the community, corresponding to a reporting percentage of 6.5%. However, the definition of infectious intestinal disease used in the IID2 study explicitly excluded non-infectious causes of vomiting and diarrhoea, which we were unable to do. Additionally, we included face-to-face, internet, and telephone contact with medical professionals, including GP services, hospital services, and out of hours services, so would expect to find a higher rate than in the IID2 study.

For an additional comparison, around 35% of Flusurvey participants reporting ILI symptoms report seeking healthcare advice [257], but this varies over time in particular rising to 43% at the start of the 2009 H1N1 influenza epidemic [233]. Additional reports from the 2009 epidemic in the U.S. give the percentage of persons seeking medical care for influenza as between 42% and 52% [258]. These percentages are larger than we found for gastroenteritis. We theorise that this may be related to the nature of the symptoms associated with each condition.

Strengths and limitations

A strength of internet-based surveillance of gastroenteritis over existing surveillance systems is that we are able to access people who do not actively seek healthcare advice. This helps build a wider picture of community based infections as these symptoms are not reported by any other means.

Additionally, it is easy and relatively cheap to expand the Flusurvey system to

record more symptoms and to include more participants. However, there is an acknowledged bias in the age and gender distribution of Flusurvey participants. Women are over-represented in the cohort, and people aged under 25 and over 65 are under-represented [246].

A strength of the Flusurvey system for surveillance compared with surveillance using GP consultations is that reports to the Flusurvey system are less affected by changes in healthcare seeking behaviour. For example, there is a drop in the GPIH gastroenteritis rate during the week containing Christmas, due to a change in availability of healthcare and healthcare seeking behaviour (see chapter 3 for details). As the Flusurvey reports are submitted online they are less liable to being affected by this behavioural and availability change.

A limitation of using Flusurvey for gastroenteritis surveillance is that data collection only occurs from November to April. Although there is acknowledgement of winter seasonality in norovirus infections, the IID1 study found no seasonality in cases of infectious intestinal disease in the community [259, 260]. However extending the Flusurvey system to the full year could lead to participant fatigue and dropping participation levels; the number of active participants is already seen to drop towards the end of each season.

We were unable to assign any cause to the symptoms reported in the Flusurvey symptom surveys. This limits the definition of syndromic gastroenteritis we are able to use as we are unable to exclude non-infectious causes of diarrhoea and vomiting. Due to this, we potentially record a higher incidence level than studies with more stringent definitions of gastroenteritis. As Flusurvey is designed for surveillance of ILI, the background survey asks about pre-existing conditions that can cause respiratory symptoms. If this were extended to ask about conditions relating to gastroenteritis symptoms this would improve future analysis.

We were not able to stratify the proportion of gastroenteritis cases seeking healthcare services by age, due to some sample sizes becoming too small. If the Flusurvey cohort were larger, we would be able to see if this proportion differs by age and over time.

Conclusions: Flusurvey data

The analysis presented here suggests that the pre-existing internet-based surveillance system for ILI, Flusurvey, also captures data on gastroenteritis incidence and

gives an estimate of the usage of healthcare services by those with gastroenteritis symptoms. The gastroenteritis incidence rate from Flusurvey was less seasonal, and the rates more similar between different age groups, than the trends seen in existing surveillance systems of gastroenteritis and gastroenteritis-causing pathogens. These differences show that further surveillance is required if the burden of gastroenteritis in the community is to be fully understood.

Internet-based surveillance is a timely and relatively cheap way to monitor disease incidence, and collects data from people who do not necessarily actively contact healthcare services. As with most disease surveillance systems, there are reporting biases due to the nature of the reporting cohort. However, Flusurvey offers an additional tool that could be used to complement existing gastroenteritis surveillance systems, each with its own reporting biases influencing the trends reported. This analysis shows that there is the potential to extend current internet based influenza surveillance systems, which exist in many countries, to include gastroenteritis surveillance without many additional resources.

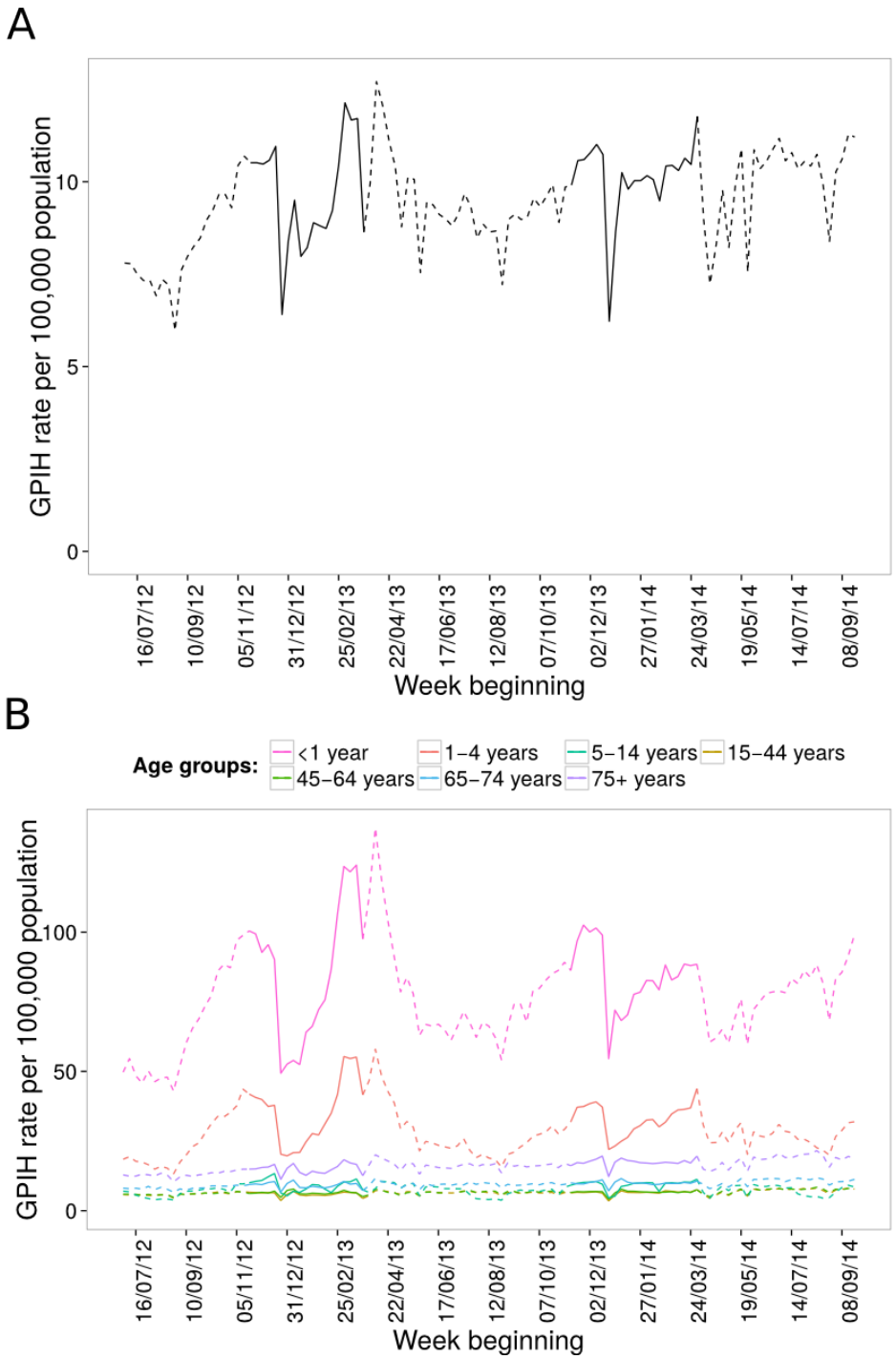


Figure 4.19: (A) The weekly gastroenteritis rate per 100,000 registered patients given by the gastroenteritis indicator from the GPIH SSS; (B) stratified by age. The time period for which we also have Flusurvey data for comparison is indicated by solid lines.

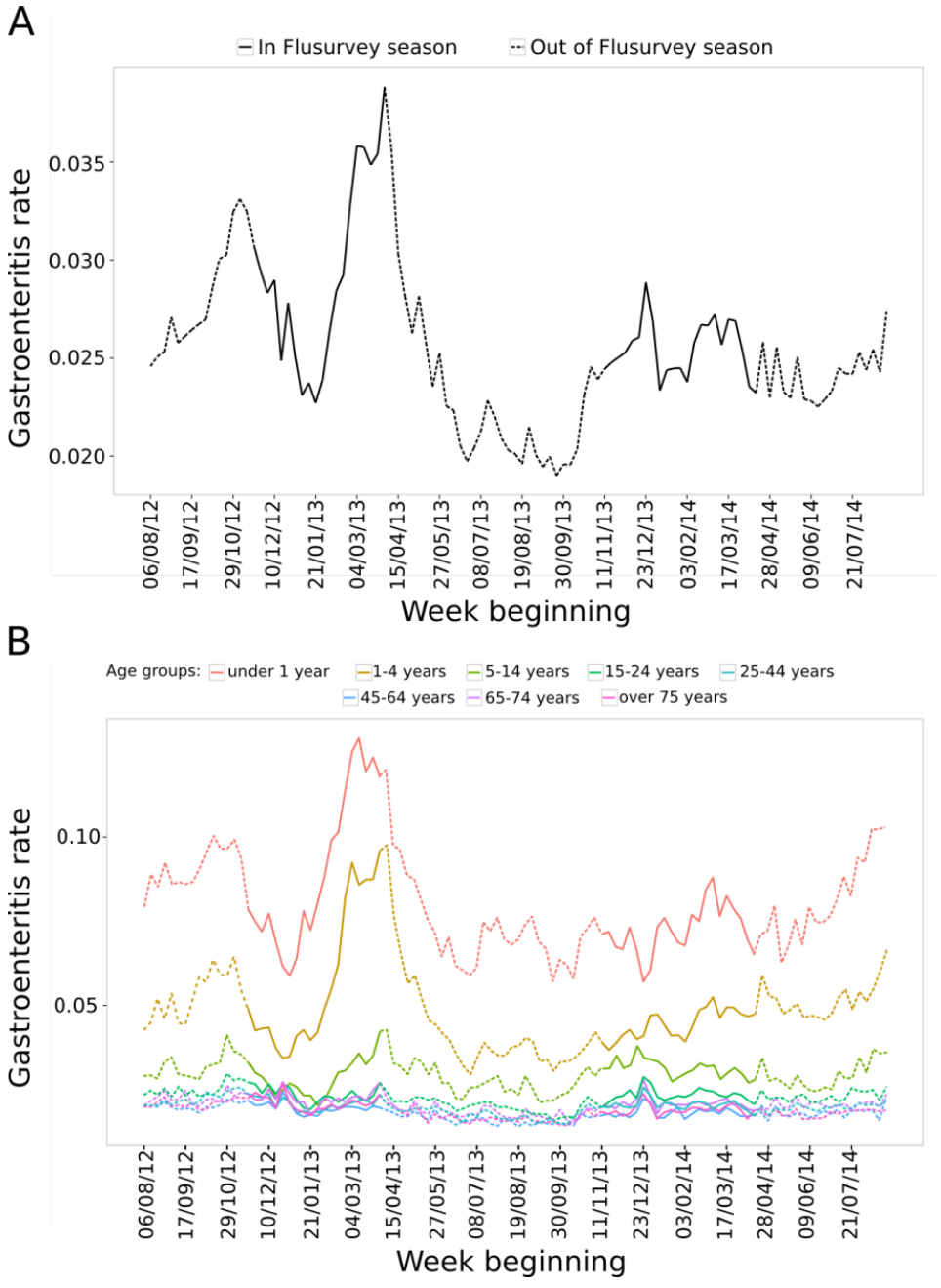


Figure 4.20: (A) The weekly gastroenteritis rate given by the gastroenteritis indicator from the GPOOH SSS; (B) stratified by age. The period of time for which we also have Flusurvey data for comparison is indicated by solid lines.

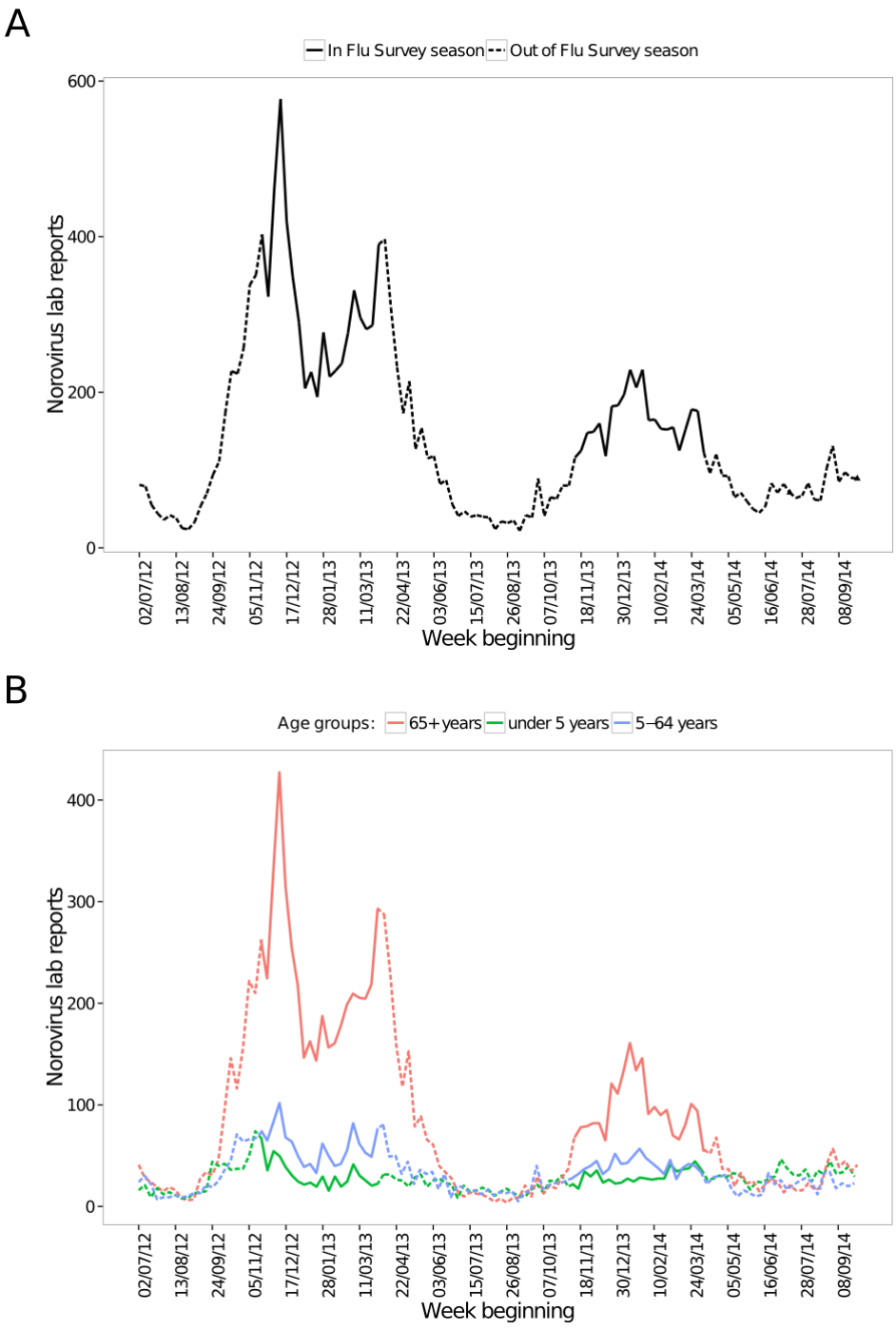


Figure 4.21: (A) The weekly number of laboratory reports for specimens positive with norovirus; (B) and stratified by age. The period of time for which we also have Flusurvey data for comparison is indicated by solid lines.

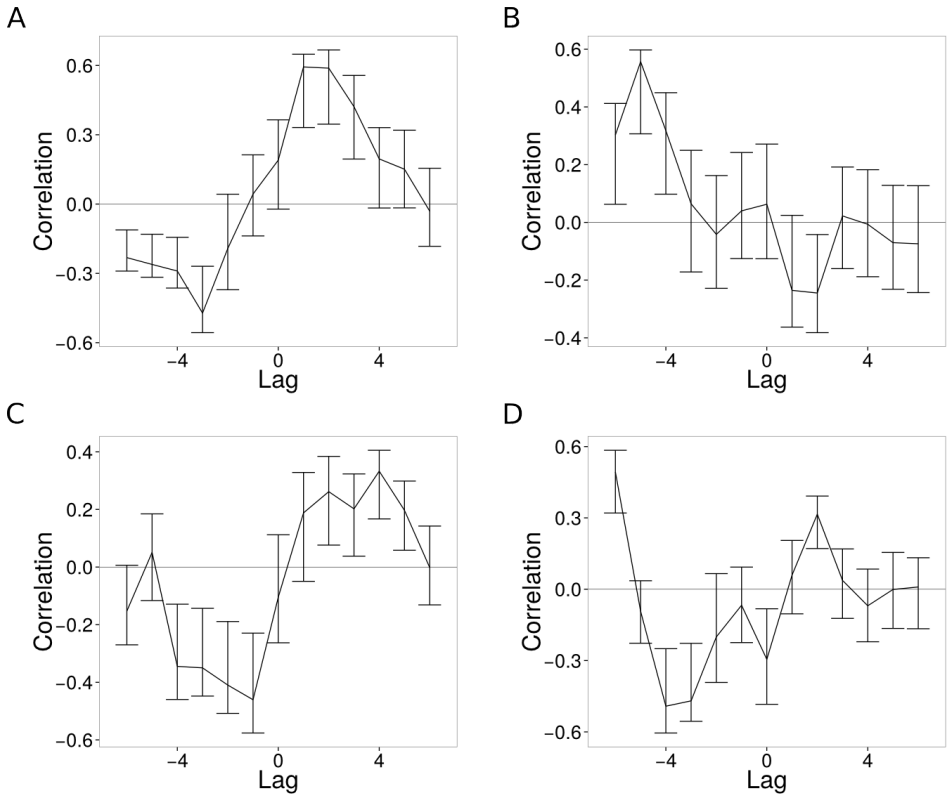


Figure 4.22: Cross correlations (line) of measures of gastroenteritis with 95% confidence intervals obtained via parametric bootstrapping (bars). In all plots lag 1 means that the Flusurvey gastroenteritis rate is being compared with the other measure of gastroenteritis from the previous week. (A) Flusurvey gastroenteritis rate with the GPIH gastroenteritis rate per 100,000 registered patients, 2012/13. (B) Flusurvey gastroenteritis rate with the GPIH gastroenteritis rate per 100,000 registered patients, 2013/14. (C) Flusurvey gastroenteritis rate with norovirus lab reports, 2012/13. (D) Flusurvey gastroenteritis rate with norovirus lab reports, 2013/14.

CHAPTER 5

CONCLUSIONS

5.1 Summary

Many aspects of disease surveillance in the UK are world-leading, and across the world disease surveillance has developed over the last 50 years into a vital public health resource. However, there are still outstanding mathematical and statistical questions in this field, in particular as computational resources develop and new datasets become available. In this thesis, we have investigated three data-heavy challenges that can each be used to develop the surveillance of gastroenteritis.

In chapter 2 we presented a flexible framework for deriving Gaussian process approximations of stochastic models of epidemics and compared a variety of approximations. We preformed fast inference on both synthetic and real epidemic data, with the real data coming from an outbreak of norovirus on a cruise ship in British waters. We derived good estimates for the parameter values of the epidemic models and inferred the unobserved processes. We, therefore, demonstrated that these approximation methods could be used for routine fast inference of epidemiological surveillance data.

In chapter 3 we found strong evidence of day of the week and public holiday effects in syndromic indicators of gastroenteritis from syndromic surveillance systems operated by PHE. Most of these effects were to be expected given the availability and purpose of the different healthcare services. However, this is the first formal descrip-

tion of these. We did not find large differences in the day of the week and public holiday effects in reports of gastroenteritis compared to the total number of reports of poor health. Next, we used the knowledge that we had obtained about day of the week and public holiday effects to suggest improvements to current syndromic surveillance methods. We suggested refinements to the regression method used by PHE to analyse syndromic data in order to take into account some of the more subtle public holiday effects. We also developed a smoothing method where both day of the week and public holiday effects are taken into account simultaneously. This can aid the interpretation of trends in daily data from GP services. This smoothing method is now in use by PHE.

Our analysis demonstrated that corrections must be made for the day of the week, public holidays, and days surrounding public holidays when analysing, visualising, and modelling daily syndromic data of gastroenteritis. We have highlighted the importance of being aware of potential trends in healthcare data due to changes in behaviour rather than changes in actual disease levels.

In chapter 4 we found that data from Google searches and Wikipedia use relating to gastroenteritis have some similarities with the number of positive laboratory reports for norovirus, but that these online data cannot be used to add value to simple forecasting/nowcasting models of laboratory reports. We then went on to show that the ILI surveillance system Flusurvey receives sufficient reports of gastroenteritis symptoms to be considered for use for syndromic gastroenteritis surveillance. This would not require the addition of many extra resources to this system. The gastroenteritis incidence rate from Flusurvey was less seasonal, and the rates more similar between different age groups, than the trends seen in existing surveillance systems of gastroenteritis and gastroenteritis-causing pathogens. These differences demonstrated that further surveillance is required if the true burden of gastroenteritis in the community is to be fully understood.

5.2 Further work

Conclusions have been obtained from each of these three pieces of work but, of course, with additional time and resources there are many extensions that could be investigated. Additionally, there are many other features of gastroenteritis surveillance that we could have tackled during the past four years. Some smaller, more

explicit suggestions for future work have been described in the discussion and conclusion section of each chapter. Therefore, here we mainly give broad directions for bigger future projects.

The motivation for beginning this work arose from the fact that mechanistic models of infectious diseases are not routinely used for syndromic surveillance of illnesses, specifically of gastroenteritis. The literature, and application, of these mechanistic models is extensive and varied. A mechanistic model of disease allows us to infer unobserved processes, the key one in this case being the size of the susceptible population, and this feeds into predictions of incidence. This can lead to, for example, a prediction of fewer norovirus cases in the winter after a number of years of many cases (due to susceptible depletion). A statistical model that assumes future activity will reflect past activity would not give this type of prediction. However, syndromic surveillance requires analysis methods that are fast and robust and, therefore, it is not simple to incorporate mechanistic models.

We began our investigation by attempting to fit an SEIRS ODE model to syndromic data of gastroenteritis (analysis not shown here). However, we found that the parameters were poorly identifiable. This motivated our work on stochastic models and approximations in order to perform the fast inference that would be required for surveillance. Therefore, the next major step would be to take the approximations investigated in chapter 2, develop them, and incorporate them into a real-time surveillance system. This would require these approximations to be more robust so they can be applied, in a reasonably automated way, to the variety of different syndromes and situations that surveillance systems work with.

The investigation into day of the week and public holiday effects in chapter 3 was quite thorough. However, there are other regular effects that may also influence these data for which similar statistical analyses could be undertaken. For example, reports to emergency departments for alcohol intoxication were elevated in Australia after major sporting events [123]. Sporting events, social events, and other mass gatherings may impact healthcare usage in the UK as well. In particular, these effects may be noticeable in surveillance data from a smaller geographical region as opposed to aggregated data from the whole country. acho2013

In chapter 4 we briefly mentioned that analysing page view data of specific healthcare websites may provide more insight into community burden of illnesses. The NHS website contains a wealth of information and is a trusted source of advice for people

in the UK. The use of the pages relating to gastroenteritis on this website will give interesting information about the current levels of information seeking on this condition in the UK.

The second part of chapter 4 investigated gastroenteritis reports made to Flusurvey. The natural extension as a result of this work is to slightly adapt the branding, the choice of language, and the background survey of this system so that it can carry on collecting information more generally about symptoms and can be used for gastroenteritis surveillance each winter. Conclusions have been made from the Flusurvey system about risk factors, vaccination effectiveness, and quality of life impact of ILI [247, 261, 262]. There is the potential to do the same for gastroenteritis. This work will become increasingly important as norovirus vaccines become more viable.

5.3 Concluding remark

In this thesis, we have worked with both mechanistic and statistical techniques to address some of the challenges that remain for both of these approaches when analysing syndromic gastroenteritis surveillance data. Statistical surveillance models do not consider known biological mechanisms, such as the depletion of the susceptible population, and instead assume that the relationships demonstrated in past data will persist. On the other hand, mechanistic models often require idealised datasets, whereas we have shown artefacts, such as public holidays, leave strong signals that should be taken into account. Finally, we identified other sources of data on gastroenteritis burden. We have also offered some further directions for bridging this gap between syndromic surveillance and mechanistic disease modelling and hope to see further progress made on this.

APPENDIX A

ADDITIONAL FIGURES FOR LESS OBVIOUS DAY OF THE WEEK EFFECTS

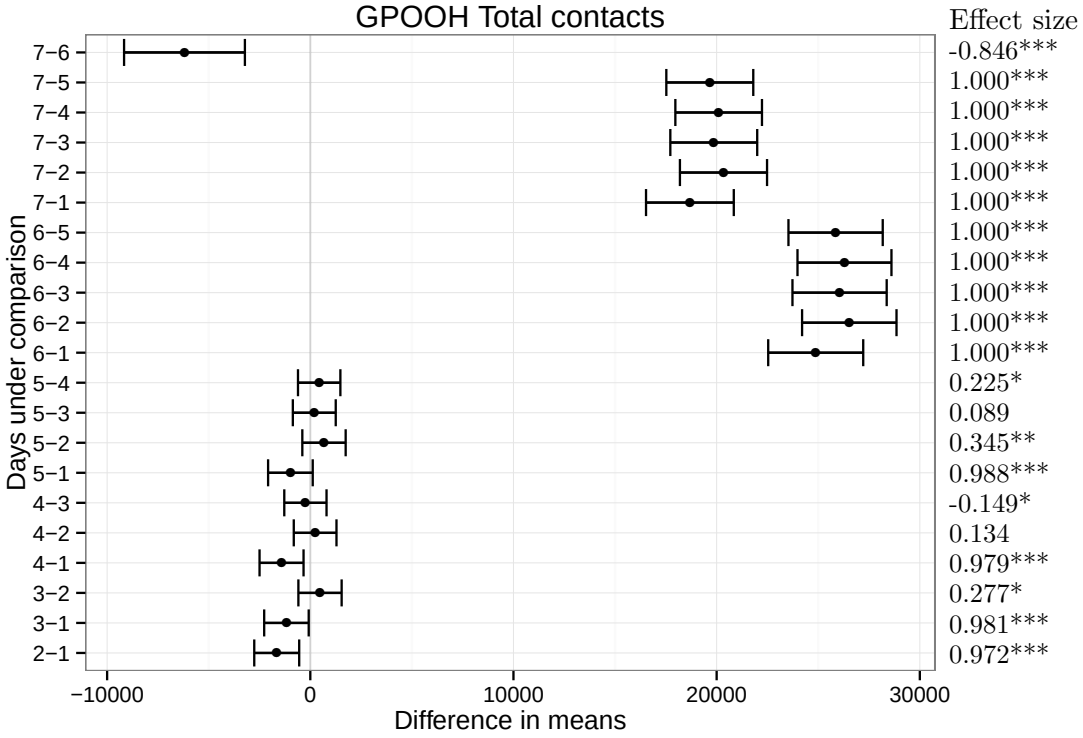


Figure A.1: GPOOH total contacts. Each row gives the results from comparing a pair of days (numbered 1 - 7 for Monday - Sunday). The difference between the means of the two days is given by the black dot, the error bar is +/- one pooled standard deviation, with Cliff's delta (where * is a small, ** a medium, and *** a large effect size).

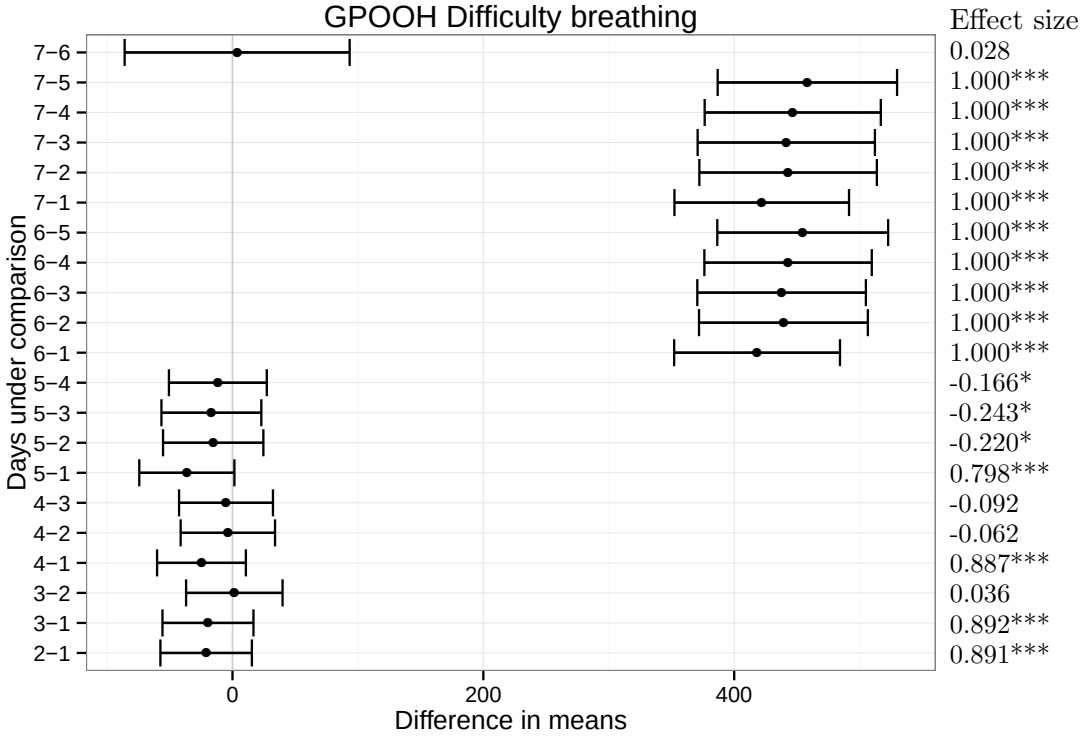


Figure A.2: GPOOH difficulty breathing. Each row gives the results from comparing a pair of days (numbered 1 - 7 for Monday - Sunday). The difference between the means of the two days is given by the black dot, the error bar is +/- one pooled standard deviation, with Cliff's delta (where * is a small, ** a medium, and *** a large effect size).

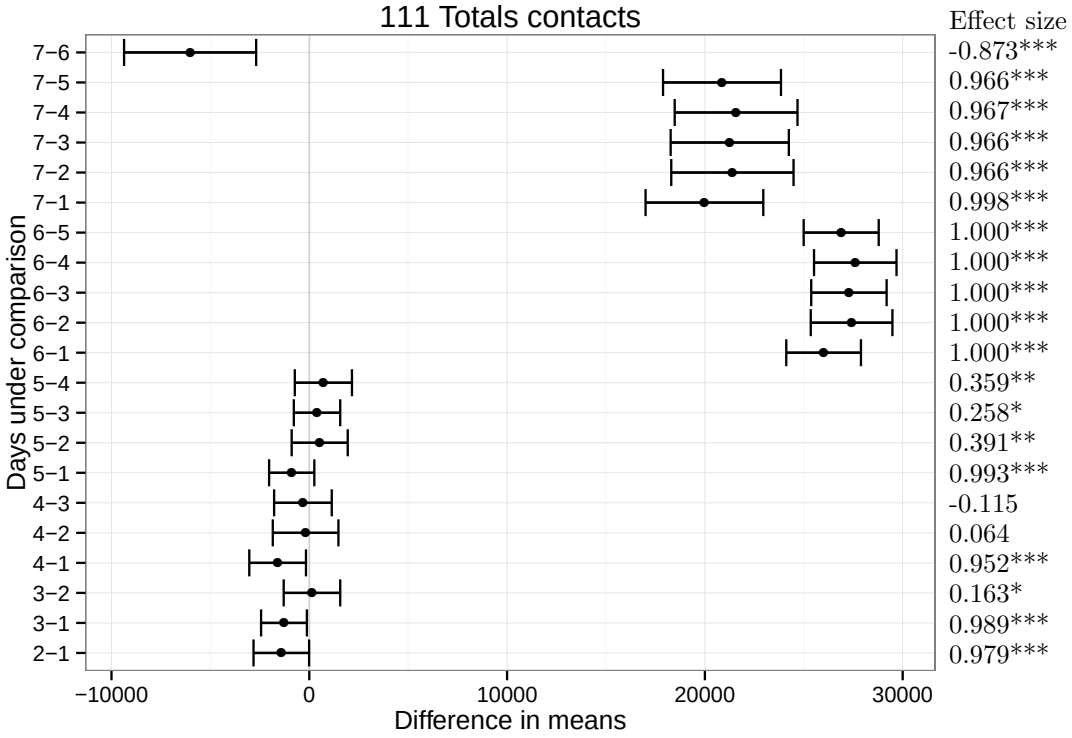


Figure A.3: 111 total contacts. Each row gives the results from comparing a pair of days (numbered 1 - 7 for Monday - Sunday). The difference between the means of the two days is given by the black dot, the error bar is +/- one pooled standard deviation, with Cliff's delta (where * is a small, ** a medium, and *** a large effect size).

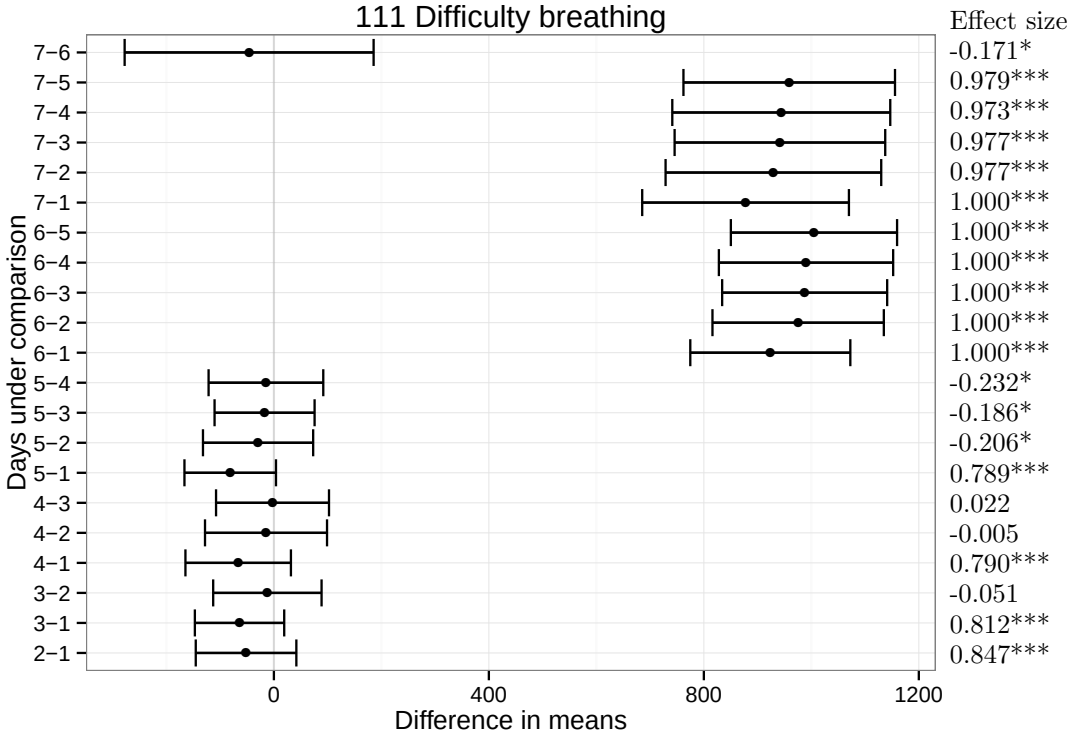


Figure A.4: 111 difficulty breathing. Each row gives the results from comparing a pair of days (numbered 1 - 7 for Monday - Sunday). The difference between the means of the two days is given by the black dot, the error bar is +/- one pooled standard deviation, with Cliff's delta (where * is a small, ** a medium, and *** a large effect size).

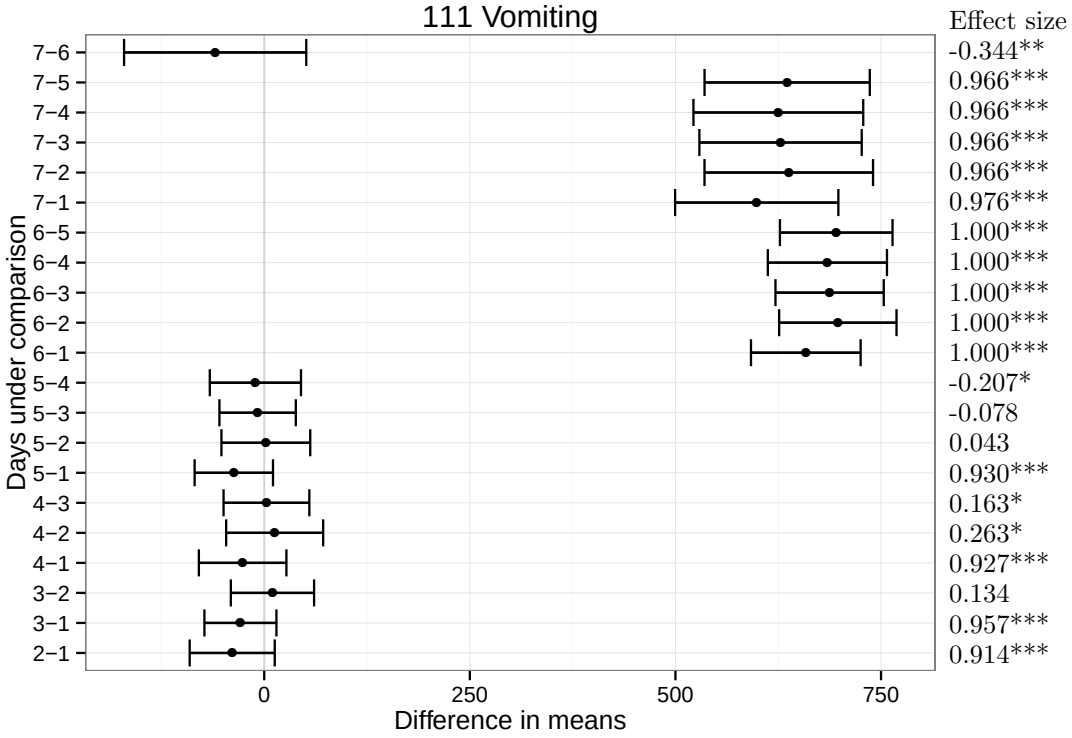


Figure A.5: 111 vomiting. Each row gives the results from comparing a pair of days (numbered 1 - 7 for Monday - Sunday). The difference between the means of the two days is given by the black dot, the error bar is +/- one pooled standard deviation, with Cliff's delta (where * is a small, ** a medium, and *** a large effect size).

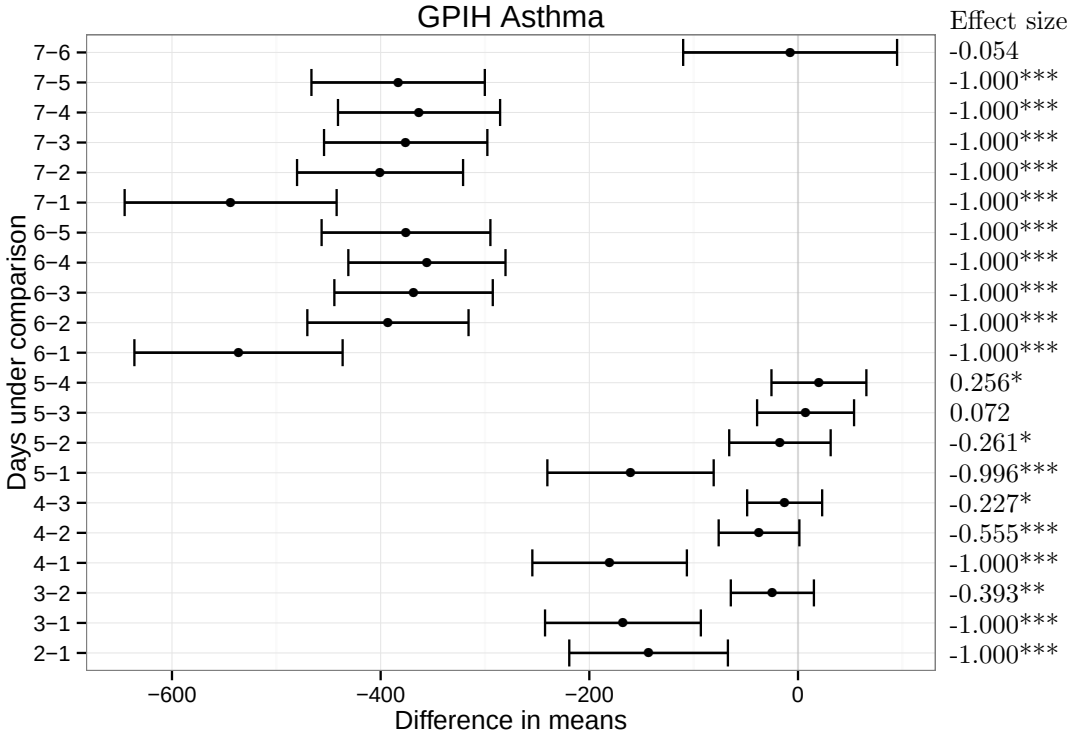


Figure A.6: GPIH asthma. Each row gives the results from comparing a pair of days (numbered 1 - 7 for Monday - Sunday). The difference between the means of the two days is given by the black dot, the error bar is +/- one pooled standard deviation, with Cliff's delta (where * is a small, ** a medium, and *** a large effect size).

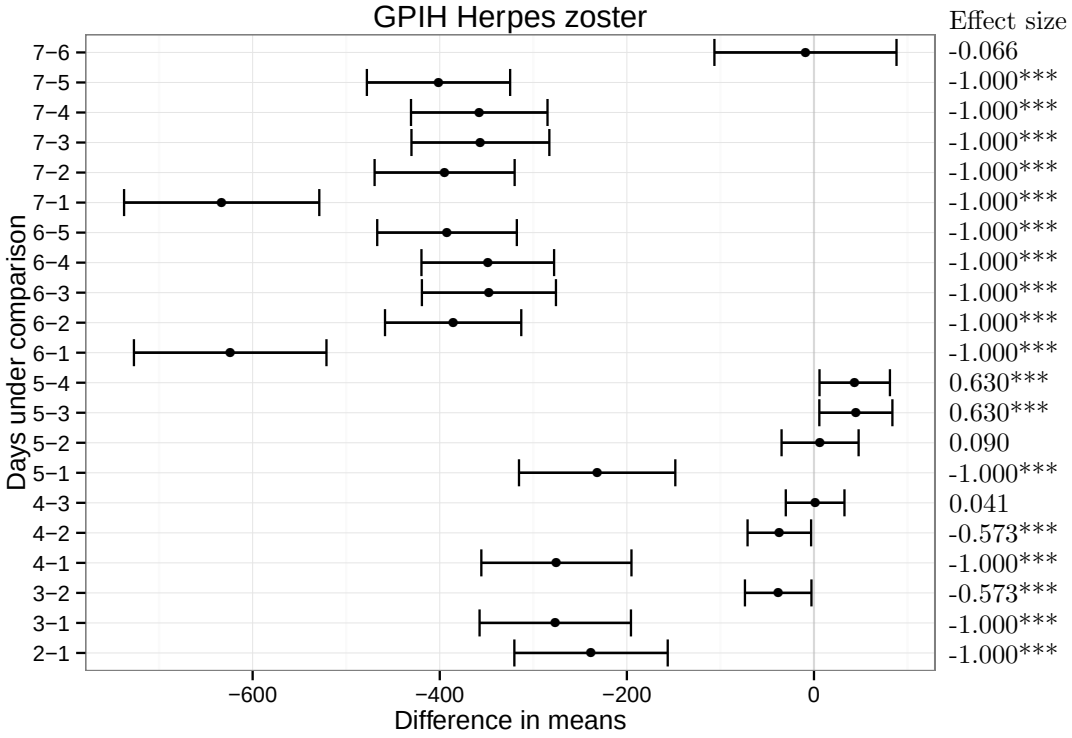


Figure A.7: GPIH herpes zoster. Each row gives the results from comparing a pair of days (numbered 1 - 7 for Monday - Sunday). The difference between the means of the two days is given by the black dot, the error bar is +/- one pooled standard deviation, with Cliff's delta (where * is a small, ** a medium, and *** a large effect size).

BIBLIOGRAPHY

[1] S. E. Majowicz, G. Hall, E. Scallan, G. K. Adak, C. Gauci, T. F. Jones, S. O'Brien, O. Henao, and P. N. Sockett. A common, symptom-based case definition for gastroenteritis. *Epidemiology and Infection*, 136(7):886–94, 2008. [Cited on pages 1 and 143.]

[2] Amandeep Singh and Michelle Fleurat. Pediatric Emergency Medicine Practice Acute Gastroenteritis - An Update. *Pediatric Emergency Medicine Practice*, 7(7):1–26, 2010. [Cited on pages 1 and 2.]

[3] Sharia M. Ahmed, Aron J. Hall, Anne E. Robinson, Linda Verhoef, Prasanna Premkumar, Umesh D. Parashar, Marion Koopmans, and Benjamin A. Lopman. Global prevalence of norovirus in cases of gastroenteritis: A systematic review and meta-analysis. *The Lancet Infectious Diseases*, 14(8):725–730, 2014. [Cited on pages 1 and 2.]

[4] Nicole Pfeil, Ulrike Uhlig, Karel Kostev, Rita Carius, Helmut Schröder, Wieland Kiess, and Holm H. Uhlig. Antiemetic medications in children with presumed infectious gastroenteritis-Pharmacoepidemiology in Europe and Northern America. *Journal of Pediatrics*, 153(5), 2008. [Cited on page 1.]

[5] Sarah M. Bartsch, Benjamin A. Lopman, Sachiko Ozawa, Aron J. Hall, and Bruce Y. Lee. Global economic burden of norovirus gastroenteritis. *PLoS ONE*, 11(4):1–16, 2016. [Cited on pages 1 and 2.]

[6] Ben A. Lopman, Mark H. Reacher, Ian B. Vipond, Dawn Hill, Christine Perry, Tracey Halladay, David W. Brown, W. John Edmunds, and Joyshri Sarangi.

Epidemiology and cost of nosocomial gastroenteritis, Avon, England, 2002-2003. *Emerging Infectious Diseases*, 10(10):1827–1834, 2004. [Cited on page 1.]

[7] J. Danial, J. A. Cepeda, F. Cameron, K. Cloy, D. Wishart, and K. E. Templeton. Epidemiology and costs associated with norovirus outbreaks in NHS Lothian, Scotland 2007-2009. *Journal of Hospital Infection*, 79(4):354–358, 2011. [Cited on page 1.]

[8] Elizabeth Jane Elliott. Acute gastroenteritis in children. *BMJ Clinical Review*, 334:35–40, 2007. [Cited on page 2.]

[9] Naor Bar-Zeev, Lester Kapanda, Jacqueline E. Tate, Khuzwayo C. Jere, Miren Iturriza-Gomara, Osamu Nakagomi, Charles Mwansambo, Anthony Costello, Umesh D. Parashar, Robert S. Heyderman, Neil French, Nigel A. Cunliffe, James Beard, Amelia C. Crampin, Carina King, Sonia Lewycka, Hazzie Mvula, Tambosi Phiri, Jennifer R. Verani, and Cynthia G. Whitney. Effectiveness of a monovalent rotavirus vaccine in infants in Malawi after programmatic roll-out: An observational and case-control study. *The Lancet Infectious Diseases*, 15(4):422–428, 2015. [Cited on page 2.]

[10] Jacqueline E. Tate, Anthony H. Burton, Cynthia Boschi-Pinto, A. Duncan Steele, Jazmin Duque, and Umesh D. Parashar. 2008 estimate of worldwide rotavirus-associated mortality in children younger than 5 years before the introduction of universal rotavirus vaccination programmes: A systematic review and meta-analysis. *The Lancet Infectious Diseases*, 12(2):136–141, 2012. [Cited on page 2.]

[11] World Health Organization. Rotavirus vaccines WHO position paper - January 2013. *Weekly Epidemiological Record*, 88(5):49–64, 2013. [Cited on page 2.]

[12] Vesta Richardson, Joselito Hernandez-Pichardo, Manjari Quintanar-Solares, Marcelino Esparza-Aguilar, Brian Johnson, Cesar Misael Gomez-Altamirano, Umesh Parashar, and Manish Patel. Effect of rotavirus vaccination on death from childhood diarrhea in Mexico. *The New England Journal of Medicine*, 362(4):299–305, 2010. [Cited on page 2.]

[13] Jacqueline E. Tate, Margaret M. Cortese, Daniel C. Payne, Aaron T. Curns, Catherine Yen, Douglas H. Esposito, Jennifer E. Cortes, Benjamin A. Lopman, Manish M. Patel, Jon R. Gentsch, and Umesh D. Parashar. Uptake, impact, and effectiveness of rotavirus vaccination in the United States: Review of the

first 3 years of postlicensure data. *The Pediatric Infectious Disease Journal*, 30(1):S56–S60, 2011. [Cited on page 2.]

[14] Jim P. Buttery, Stephen B. Lambert, Keith Grimwood, Michael D. Nissen, Emma J. Field, Kristine K. Macartney, Jonathan D. Akikusa, Julian J. Kelly, and Carl D. Kirkwood. Reduction in rotavirus-associated acute gastroenteritis following introduction of rotavirus vaccine into Australia's National Childhood vaccine schedule. *The Pediatric Infectious Disease Journal*, 30:S25–S29, 2011. [Cited on page 2.]

[15] Zharain Bawa, Alex J. Elliot, Roger A. Morbey, Shamez Ladhani, Nigel A. Cunliffe, Sarah J. O'Brien, Martyn Regan, Gillian E. Smith, and Robert A. Weinstein. Assessing the likely impact of a rotavirus vaccination program in England: The contribution of syndromic surveillance. *Clinical Infectious Diseases*, 61(1):77–85, 2015. [Cited on page 2.]

[16] J. P. Harris, N. L. Adams, B. A. Lopman, D. J. Allen, and G. K. Adak. The development of web-based surveillance provides new insights into the burden of norovirus outbreaks in hospitals in England. *Epidemiology and Infection*, 142(08):1590–1598, 2014. [Cited on page 2.]

[17] Aron J. Hall, Ben A. Lopman, Daniel C. Payne, Manish M. Patel, Paul A. Gastañaduy, Jan Vinjé, and Umesh D. Parashar. Norovirus disease in the United States. *Emerging Infectious Diseases*, 19(8):1198–1205, 2013. [Cited on pages 2 and 5.]

[18] Miren Iturriza-Gómara and Benjamin Lopman. Norovirus in healthcare settings. *Current Opinion in Infectious Diseases*, 27(5):437–43, 2014. [Cited on page 2.]

[19] Elmira T. Isakbaeva, Marc Alain Widdowson, R. Suzanne Beard, Sandra N. Bulens, James Mullins, Stephan S. Monroe, Joseph Bresee, Patricia Sassano, Elaine H. Cramer, and Roger I. Glass. Norovirus transmission on cruise ship. *Emerging Infectious Diseases*, 11(1):154–157, 2005. [Cited on page 2.]

[20] Maria Lysén, Margareta Thorhagen, Maria Brytting, Marika Hjertqvist, Yvonne Andersson, and Kjell Olof Hedlund. Genetic diversity among foodborne and waterborne norovirus strains causing outbreaks in Sweden. *Journal of Clinical Microbiology*, 47(8):2411–2418, 2009. [Cited on page 2.]

[21] World Health Organisation. International Health Regulations - Third Edition. *WHO*, 2005:84, 2016. [Cited on page 2.]

[22] N. M. M'ikanatha, R. Lynfield, C. A. Van Beneden, and H. de Valk. Infectious Disease Surveillance: A Cornerstone for Prevention and Control. In Nkuchia M. M'ikanatha, Ruth Lynfield, Kathleen G. Julian, Chris A. Van Beneden, and Henriette de Valk, editors, *Infectious Disease Surveillance*. John Wiley & Sons Ltd, Oxford, UK, second edition, 2007. [Cited on page 2.]

[23] Bernard C. K. Choi. The past, present, and future of public health surveillance. *Scientifica*, 2012:1–26, 2012. [Cited on page 3.]

[24] S. Declich and A. O. Carter. Public health surveillance: Historical origins, methods and evaluation. *Bulletin of the World Health Organization*, 72(2):285–304, 1994. [Cited on page 3.]

[25] Steffen Unkel, C. Paddy Farrington, Paul H. Garthwaite, Chris Robertson, and Nick Andrews. Statistical methods for the prospective detection of infectious disease outbreaks : a review. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 175(1):49–82, 2012. [Cited on page 3.]

[26] Donna F. Stroup, G. David Williamson, Joy L. Herndon, and John M. Karon. Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in Medicine*, 8(3):323–329, 1989. [Cited on page 3.]

[27] Michael Hhle, Andrea Riebler, and Michaela Paul. Getting started with outbreak detection. `https://cran.r-project.org/web/packages/surveillance/vignettes/surveillance.pdf`, 2007. [Cited on pages 3 and 5.]

[28] C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(3):547–563, 1996. [Cited on pages 3 and 78.]

[29] Angela Noufaily, Yonas Ghebremichael-Weldeselassie, Doyo Gragn Enki, and Paul Garthwaite. Modelling reporting delays for outbreak detection in infectious disease data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 178(1):205–222, 2015. [Cited on page 3.]

[30] Roger A. Morbey, Alex J. Elliot, Andre Charlett, Neville Q. Verlander, Nick Andrews, and Gillian E. Smith. The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE) method for syndromic surveillance in England. *Bioinformatics*, 31(22):3660–3665, 2015. [Cited on pages 4, 46, 86, 88, 92, and 97.]

[31] Julian Besag and James Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(1):143–155, 1991. [Cited on page 4.]

[32] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496, 1997. [Cited on page 4.]

[33] Martin Kulldorff. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 164:61–72, 2001. [Cited on page 4.]

[34] Simon E. F. Spencer, Jonathan Marshall, Ruth Pirie, Donald Campbell, and Nigel P. French. The detection of spatially localised outbreaks in campylobacteriosis notification data. *Spatial and Spatio-temporal Epidemiology*, 2(3):173–183, 2011. [Cited on page 4.]

[35] Salmon Maëlle, Schumacher Dirk, and Höhle Michael. Monitoring count time series in R: Aberration detection in public health surveillance. *Journal Of Statistical Software*, 70(10), 2016. [Cited on page 5.]

[36] J. G. Wheeler, D. Sethi, J. M. Cowden, P. G. Wall, L. C. Rodrigues, D. S. Tompkins, M. J. Hudson, and P. J. Roderick. Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance. *BMJ*, 318(7190):1046–50, 1999. [Cited on pages 5 and 139.]

[37] Sarah J O'Brien, Greta Rait, Paul R Hunter, James J Gray, Frederick J Bolton, David S Tompkins, Jim McLauchlin, Louise H Letley, Goutam K Adak, John M Cowden, Meirion R Evans, Keith R Neal, Gillian E Smith, Brian Smyth, Clarence C Tam, and Laura C Rodrigues. Methods for determining disease burden and calibrating national surveillance data in the United Kingdom: the second study of infectious intestinal disease in the community (IID2 study). *BMC Medical Research Methodology*, 10(39):1–13, 2010. [Cited on pages 5 and 139.]

[38] Ivo M. Foppa. *A Historical Introduction to Mathematical Modeling of Infectious Diseases*. Academic Press, Boston, 2017. [Cited on page 6.]

[39] J. A. P. Heesterbeek and M. G. Roberts. How mathematical epidemiology became a field of biology: a commentary on Anderson and May (1981) 'The

population dynamics of microparasites and their invertebrate hosts'. *Philosophical Transactions of the Royal Society of London. Series B (Biological sciences)*, 370(1666):20140307, 2015. [Cited on page 6.]

[40] Fred Brauer, Pauline van den Driessche, and Jianhong Wu, editors. *Mathematical Epidemiology*. Springer-Verlag Berlin Heidelberg, 2008. [Cited on page 6.]

[41] Norman T. J. Bailey. A simple stochastic epidemic. *Biometrika*, 37(3/4):193–202, 1950. [Cited on page 6.]

[42] M. J. Keeling and J. V. Ross. On methods for studying stochastic disease dynamics. *Journal of The Royal Society Interface*, 5(19):171–181, 2008. [Cited on page 6.]

[43] D. A. Griffiths. Maximum likelihood estimation for the Beta-Binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, 29(4):637–648, 1973. [Cited on page 7.]

[44] Cécile Viboud, Lone Simonsen, and Gerardo Chowell. A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. *Epidemics*, 15:27–37, 2016. [Cited on page 7.]

[45] Simon Cauchemez, Alain-Jacques Valleron, Pierre-Yves Boëlle, Antoine Flahault, and Neil M. Ferguson. Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, 452(7188):750–754, 2008. [Cited on page 7.]

[46] R. M. Anderson, C. A. Donnelly, N. M. Ferguson, M. E. J. Woolhouse, C. J. Watt, H. J. Udy, S. MaWhinney, S. P. Dunstan, T. R. E. Southwood, J. W. Wilesmith, J. B. M. Ryan, L. J. Hoinville, J. E. Hillerton, A. R. Austin, and G. A. H. Wells. Transmission dynamics and epidemiology of BSE in British cattle. *Nature*, 382(6594):779–788, 1996. [Cited on page 7.]

[47] S. C. Howard and C. A. Donnelly. The importance of immediate destruction in epidemics of foot and mouth disease. *Research in Veterinary Science*, 69(2):189–96, 2000. [Cited on page 7.]

[48] M. J. Keeling, M. E. J. Woolhouse, R. M. May, G. Davies, and B. T. Grenfell. Modelling vaccination strategies against foot-and-mouth disease. *Nature*, 421(6919):136–142, 2003. [Cited on page 7.]

[49] Neil M. Ferguson, Christl A. Donnelly, and Roy M. Anderson. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature*, 413(6855):542–548, 2001. [Cited on page 7.]

[50] Philip D. O'Neill. Introduction and snapshot review: Relating infectious disease transmission models to data. *Statistics in Medicine*, 29(20):2069–2077, 2010. [Cited on page 11.]

[51] E. L. Ionides, C. Bretó, and A. A. King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 103(49):18438–18443, 2006. [Cited on page 11.]

[52] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P H Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6:187–202, 2009. [Not cited.]

[53] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010. [Not cited.]

[54] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos. SMC 2 : an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):397–426, 2013. [Not cited.]

[55] Trevelyan J. McKinley, Joshua V. Ross, Rob Deardon, and Alex R. Cook. Simulation-based Bayesian inference for epidemic models. *Computational Statistics and Data Analysis*, 71:434–447, 2014. [Not cited.]

[56] Peter Neal and Chien Lin Terry Huang. Forward simulation Markov Chain Monte Carlo with applications to stochastic epidemic models. *Scandinavian Journal of Statistics*, 42(2):378–396, 2015. [Cited on page 11.]

[57] Jacob Leander, Torbjörn Lundh, and Mats Jirstrand. Stochastic differential equations as a tool to regularize the parameter estimation problem for continuous time dynamical systems given discrete time measurements. *Mathematical Biosciences*, 251:54–62, 2014. [Cited on page 11.]

[58] Thomas House. For principled model fitting in mathematical biology. *Journal of Mathematical Biology*, 70(5):1007–1013, 2015. [Not cited.]

[59] Aaron A. King, Matthieu Domenech de Celles, Felicia M. G. Magpantay, and Pejman Rohani. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society B: Biological Sciences*, 282(20150347), 2015. [Cited on page 11.]

[60] J. V. Ross, T. Taimre, and P. K. Pollett. On parameter estimation in population models. *Theoretical Population Biology*, 70(4):498–510, 2006. [Cited on page 11.]

[61] Paul Fearnhead, Vasilieos Giagos, and Chris Sherlock. Inference for reaction networks using the linear noise approximation. *Biometrics*, 70(2):457–466, 2014. [Cited on page 11.]

[62] Frank Ball and Thomas House. Heterogeneous network epidemics: real-time growth, variance and extinction of infection. *Journal of Mathematical Biology*, 75(3):1–43, 2017. [Cited on page 11.]

[63] Hakan Andersson and Tom Britton. *Stochastic epidemic models and their statistical analysis*. Springer-Verlag New York, New York, NY, 2000. [Cited on page 12.]

[64] M. J. Keeling, M. E. J. Woolhouse, R. M. May, G. Davies, and B. T. Grenfell. Modelling vaccination strategies against foot-and-mouth disease. *Nature*, 421(6919):136–142, 2003. [Cited on page 12.]

[65] Andrew J K Conlan and Bryan T Grenfell. Seasonality and the persistence and invasion of measles. *Proceedings of the Royal Society B: Biological Sciences*, 274(1614):1133–1141, 2007. [Cited on page 12.]

[66] Steven Riley, Christophe Fraser, Christl A. Donnelly, Azra C. Ghani, Laith J. Abu-Raddad, Anthony J. Hedley, Gabriel M. Leung, Lai-Ming Ho, Tai-Hing Lam, Thuan Q. Thach, Patsy Chau, King-Pan Chan, Su-Vui Lo, Pak-Yin Leung, Thomas Tsang, William Ho, Koon-Hung Lee, Edith M. C. Lau, Neil M. Ferguson, and Roy M. Anderson. Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science*, 300(5627):1961–1966, 2003. [Cited on page 12.]

[67] Thomas G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov Processes. *Journal of Applied Probability*, 7(1):49–58, 1970. [Cited on page 13.]

[68] Thomas G. Kurtz. Limit theorems for sequences of jump Markov Processes approximating ordinary differential processes. *Journal of Applied Probability*, 8(2):344–356, 1971. [Cited on page 13.]

[69] Eckhard Platen and Nicola Bruti-Liberati. *Numerical solution of stochastic differential equations with jumps in finance*. Springer-Verlag Berlin Heidelberg, 2010. [Cited on page 13.]

[70] Cedric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor. Gaussian process approximations of stochastic differential equations. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 1:1–16, 2007. [Cited on page 14.]

[71] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the Brownian motion. *Physical Review*, 36(5):823–841, 1930. [Cited on page 15.]

[72] Andrew J. Black and Alan J. McKane. Stochastic amplification in an epidemic model with seasonal forcing. *Journal of Theoretical Biology*, 267(1):85–94, 2010. [Cited on page 15.]

[73] Valerie Isham. Assessing the variability of stochastic epidemics. *Mathematical Biosciences*, 107(2):209–224, 1991. [Cited on pages 16 and 17.]

[74] Björn Holmquist. Moments and cumulants of the multivariate normal distribution. *Stochastic Analysis and Applications*, 6(3):273–278, 1988. [Cited on page 18.]

[75] M. J. Keeling. Multiplicative moments and measures of persistence in ecology. *Journal of Theoretical Biology*, 205(2):269–281, 2000. [Cited on page 21.]

[76] Isthrinayagy Krishnarajah, Alex Cook, Glenn Marion, and Gavin Gibson. Novel moment closure approximations in stochastic epidemics. *Bulletin of Mathematical Biology*, 67(4):855–873, 2005. [Cited on page 21.]

[77] D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716–1733, 2001. [Cited on page 22.]

[78] David J C MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 7.2 edition, 2005. [Cited on pages 22 and 23.]

[79] Morris L. Eaton. The Normal distribution on a vector space. In *Multivariate Statistics: A Vector Space Approach*. Lecture Notes–Monograph Series: Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007. [Cited on page 28.]

[80] Roberto Vivancos, Alex Keenan, Will Sopwith, Ken Smith, Catherine Quigley, Ken Mutton, Evdokia Dardamissis, Gordon Nichols, John Harris, Christopher Gallimore, Linda Verhoef, Qutub Syed, and John Reid. Norovirus outbreak in a cruise ship sailing around the British Isles: Investigation and multi-agency management of an international outbreak. *Journal of Infection*, 60(6):478–485, 2010. [Cited on pages 29, 31, 34, 35, and 36.]

[81] Kirsten Simmons, Manoj Gambhir, Juan Leon, and Ben Lopman. Duration of immunity to norovirus gastroenteritis. *Emerging Infectious Diseases*, 19(8):1260–1267, 2013. [Cited on page 29.]

[82] J. Vanderpas, J. Louis, M. Reynders, G. Mascart, and O. Vandenberg. Mathematical model for the control of nosocomial norovirus. *Journal of Hospital Infection*, 71(3):214–222, 2009. [Cited on page 31.]

[83] M. O. Milbrath, I. H. Spicknall, J. L. Zelner, C. L. Moe, and J. N. S. Eisenberg. Heterogeneity in norovirus shedding duration affects community risk. *Epidemiology and Infection*, 141(8):1572–1584, 2013. [Not cited.]

[84] Eamon B. O'Dea, Kim M. Pepin, Ben A. Lopman, and Claus O. Wilke. Fitting outbreak models to data from many small norovirus outbreaks. *Epidemics*, 6:18–29, 2014. [Cited on page 29.]

[85] Jacques Le Pendu, Nathalie Ruvoën-Clouet, Elin Kindberg, and Lennart Svensson. Mendelian resistance to human norovirus infections. *Seminars in Immunology*, 18(6):375–386, 2006. [Cited on page 31.]

[86] Rachel M. Lee, Justin Lessler, Rose A. Lee, Kara E. Rudolph, Nicholas G. Reich, Trish M. Perl, and Derek A. T. Cummings. Incubation periods of viral gastroenteritis: A systematic review. *BMC Infectious Diseases*, 13(1):446, 2013. [Cited on page 31.]

[87] Centers for Disease Control Division of Viral Diseases, National Center for Immunization and Respiratory Diseases and Prevention. Updated norovirus outbreak management and disease prevention guidelines. *Centers for Disease Control and Prevention Morbidity and Mortality Weekly Report*, 60(3):1–18, 2011. [Cited on page 31.]

[88] Simon Cauchemez and Neil M. Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *Journal of the Royal Society Interface*, 5(25):885–97, 2008. [Cited on pages 38 and 41.]

[89] Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*. Springer-Verlag New York, New York, NY, 3rd edition, 1992. [Cited on page 39.]

[90] Marta Sala Soler, Anne Fouillet, Anne Catherine Viso, Loic Josseran, Gillian E. Smith, Alex J. Elliot, Jim McMenamin, Alexandra Ziemann, and Thomas Krafft. Assessment of syndromic surveillance in Europe. *The Lancet*, 378(9806):1833–1834, 2011. [Cited on page 45.]

[91] Kelly J. Henning. Overview of Syndromic Surveillance: What is Syndromic Surveillance? *Centers for Disease Control and Prevention: Morbidity and Mortality Weekly Report*, 53:7–11, 2004. [Cited on page 45.]

[92] Public Health England. Syndromic surveillance: systems and analyses. `https://www.gov.uk/government/collections/syndromic-surveillance-systems-and-analyses`, 2014. [Cited on page 45.]

[93] S. E. Harcourt, R. A. Morbey, P. Loveridge, L. Carrilho, D. Baynham, E. Povey, P. Fox, J. Rutter, P. Moores, J. Tiffen, S. Bellerby, P. McIntosh, S. Large, J. McMenamin, A. Reynolds, S. Ibbotson, G. E. Smith, and A. J. Elliot. Developing and validating a new national remote health advice syndromic surveillance system in England. *Journal of Public Health*, 39(1):184–192, 2017. [Cited on page 45.]

[94] S. E. Harcourt, G. E. Smith, A. J. Elliot, R. Pebody, A. Charlett, S. Ibbotson, M. Regan, and J. Hippisley-Cox. Use of a large general practice syndromic surveillance system to monitor the progress of the influenza A(H1N1) pandemic 2009 in the UK. *Epidemiology and Infection*, 140(1):100–5, 2012. [Cited on pages 46 and 148.]

[95] S. E. Harcourt, J. Fletcher, P. Loveridge, A. Bains, R. Morbey, A. Yeates, B. McCloskey, B. Smyth, S. Ibbotson, G. E. Smith, and A. J. Elliot. Developing a new syndromic surveillance system for the London 2012 Olympic and Paralympic Games. *Epidemiology and Infection*, 140(12):2152–2156, 2012. [Cited on pages 46 and 148.]

[96] Alex J. Elliot, Helen E. Hughes, Thomas C. Hughes, Thomas E. Locker, Tony Shannon, John Heyworth, Andy Wapling, Mike Catchpole, Sue Ibbotson, Brian McCloskey, and Gillian E. Smith. Establishing an emergency department syndromic surveillance system to support the London 2012 Olympic and Paralympic Games. *Emergency Medicine Journal*, 29(12):954–60, 2012. [Cited on page 46.]

[97] Leofranc Holford-Strevens. *History of Time: A Very Short Introduction*. Oxford University Press, USA, 2005. [Cited on page 46.]

[98] Eviatar Zerubavel. *The seven day circle: The history and meaning of the week*. The University of Chicago Press, 1989. [Cited on pages 46 and 49.]

[99] Eurofound. Sixth European Working Conditions Survey - Overview report. Technical Report November, Luxembourg, 2016. [Cited on page 47.]

[100] Kenneth R. French. Stock returns and the weekend effect. *Journal of Financial Economics*, 8(1):55–69, 1980. [Cited on pages 47 and 56.]

[101] Jeffrey Jaffe and Randolph Westerfield. The week-end effect in common stock returns: The international evidence. *The Journal of Finance*, 40(2):433–454, 1985. [Cited on page 47.]

[102] L. Condoyanni, J. O'Hanlon, and C. W. R. Ward. Day of the week effects on stock returns: International evidence. *Journal of Business Finance & Accounting*, 14(2):159–174, 1987. [Cited on page 47.]

[103] Ercan Balaban. Day of the week effects: new evidence from an emerging stock market. *Applied Economics Letters*, 2(5):139–143, 1995. [Cited on pages 47 and 56.]

[104] Chris Brooks and Gita Persand. Seasonality in Southeast Asian stock markets: Some new evidence on day-of-the-week effects. *Applied Economic Letters*, 8(3):155–158, 2001. [Not cited.]

[105] K. A. Wong, T. K. Hui, and C. Y. Chan. Day-of-the-week effects: Evidence from developing stock markets. *Applied Financial Economics*, 2(1):49–56, 1992. [Cited on page 56.]

[106] Rakibul Islam and Nadira Sultana. Day of the week effect on stock return and volatility: Evidence from Chittagong stock exchange. *European Journal of Business and Management*, 7(3):165–173, 2015. [Cited on page 47.]

[107] G. Kohers, N. Kohers, V. Pandey, and T. Kohers. The disappearing day-of-the-week effect in the world's largest equity markets. *Applied Economics Letters*, 11(3):167–171, 2004. [Cited on pages 47 and 56.]

[108] Charles Bram Cadsby and Mitchell Ratner. Turn-of-month and pre-holiday effects on stock returns: Some international evidence. *Journal of Banking and Finance*, 16(3):497–509, 1992. [Cited on page 47.]

[109] Robert A. Ariel. High stock returns before holidays: Existence and evidence on possible causes. *The Journal of Finance*, 45(5):1611–1626, 1990. [Not cited.]

[110] Tian Yuan and Rakesh Gupta. Chinese Lunar New Year effect in Asian stock markets, 1999-2012. *Quarterly Review of Economics and Finance*, 54(4):529–537, 2014. [Not cited.]

[111] Vicente Meneu and Angel Pardo. Pre-holiday effect, large trades and small investor behaviour. *Journal of Empirical Finance*, 11(2):231–246, 2004. [Cited on page 47.]

[112] Olga Dodd and Alex Gakhovich. The holiday effect in Central and Eastern European financial markets. *Investment Management and Financial Innovations*, 8(4):29–35, 2011. [Cited on page 47.]

[113] Marcel Ausloos, Olgica Nedic, and Aleksandar Dekanski. Day of the week effect in paper submission/acceptance/rejection to/in/by peer review journals. *Physica A*, 456:197–203, 2016. [Cited on page 48.]

[114] Nehzat Motallebi, Hien Tran, Bart E Croes, and Lawrence C Larsen. Day-of-week patterns of particulate matter and its chemical components at selected sites in California. *Journal of the Air & Waste Management Association*, 53(7):876–88, 2003. [Cited on pages 48 and 56.]

[115] Vania Ceccato and Adriaan Cornelis Uittenbogaard. Space-Time Dynamics of Crime in Transport Nodes. *Annals of the Association of American Geographers*, 104(1):131–150, 2014. [Cited on page 48.]

[116] Prasanth Anbalagan and Mladen Vouk. "Days of the week" effect in predicting the time taken to fix defects. *Proceedings of the 2nd International Workshop on Defects in Large Software Systems: Held in conjunction with the ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2009)*, pages 29–30, 2009. [Cited on pages 48 and 56.]

[117] Sean T. Doherty, Jean C. Andrey, and Carolyn MacGregor. The situational risks of young drivers: The influence of passengers, time of day and day of week on accident rates. *Accident Analysis and Prevention*, 30(1):45–52, 1998. [Cited on pages 48 and 56.]

[118] Carmen Peiró-Velert, Jose Devís-Devís, Vicente J. Beltrán-Carrillo, and Kenneth R. Fox. Variability of Spanish adolescents' physical activity patterns by seasonality, day of the week and demographic factors. *European Journal of Sport Science*, 8(3):163–171, 2008. [Cited on pages 48 and 56.]

[119] Charles E. Matthews, Barbara E. Ainsworth, Raymond W. Thompson, and David R. Bassett Jr. Sources of variance in daily physical activity levels as measured by an accelerometer. *Medicine & Science in Sports & Exercise*, 34(8):1376–1381, 2002. [Cited on page 48.]

[120] T. Baranowski, M. Smith, M. D. Hearn, L. S. Lin, J. Baranowski, C. Doyle, K. Resnicow, and D. T. Wang. Patterns in children's fruit and vegetable consumption by meal and day of the week. *Journal of the American College of Nutrition*, 16(3):216–223, 1997. [Cited on pages 48 and 56.]

[121] S. Maisey, J. Loughridge, S. Southon, and R. Fulcher. Variation in food group and nutrient intake with day of the week in an elderly population. *The British Journal of Nutrition*, 73(3):359–373, 1995. [Cited on page 48.]

[122] Phillip K. Wood, Kenneth J. Sher, and Patricia C. Rutledge. College student alcohol consumption, day of the week, and class schedule. *Alcoholism: Clinical and Experimental Research*, 31(7):1195–1207, 2007. [Cited on pages 48 and 56.]

[123] Belinda Lloyd, Sharon Matthews, Michael Livingston, Harindra Jayasekara, and Karen Smith. Alcohol intoxication in the context of major public holidays, sporting and social events: A time-series analysis in Melbourne, Australia, 2000-2009. *Addiction*, 108(4):701–709, 2013. [Cited on pages 48 and 160.]

[124] Alpaslan Akay and Peter Martinsson. Sundays are blue: aren't they? The day-of-the-week effect on subjective well-being and socio-economic status. *The Institute for the Study of Labor*, Discussion, 2009. [Cited on pages 48 and 56.]

[125] Azizah Abu Bakar, Antonios Siganos, and Evangelos Vagenas-Nanos. Does mood explain the Monday effect? *Journal of Forecasting*, 33(6):409–418, 2014. [Cited on page 48.]

[126] Joshua M. Smyth, Stephen A. Wonderlich, Martin J. Sliwinski, Ross D. Crosby, Scott G. Engel, James E. Mitchell, and Rachel M. Calogero. Ecological momentary assessment of affect, stress, and binge-purge behaviors: Day of week and time of day effects in the natural environment. *International Journal of Eating Disorders*, 42(5):429–436, 2009. [Cited on pages 48 and 56.]

[127] K. A. Bollen. Temporal variations in mortality: a comparison of U.S. suicides and motor vehicle fatalities, 1972-1976. *Demography*, 20(1):45–59, 1983. [Cited on pages 48 and 56.]

[128] Helen Johnson, Anita Brock, Clare Griffiths, and Cleo Rooney. Mortality from suicide and drug-related poisoning by day of the week in England and Wales, 1993-2002. *Health Statistics Quarterly*, 26 & 27:13–16, 2005. [Not cited.]

[129] Nishi Motoi, Miyake Hirotsugu, Oksmoto Hiroyuki, Goto Youhei, and Sakai Toshirou. Relationship between suicide and holidays. *Journal of Epidemiology*, 10(5):317–320, 1994. [Cited on pages 48 and 56.]

[130] David P. Phillips and Judith Liu. The frequency of suicides around major public holidays: Some surprising findings. *Suicide and Life-Threatening Behavior*, 10(1):41–50, 1980. [Cited on page 48.]

[131] G. Jessen, B. F. Jensen, E. Arensman, U. Bille-Brahe, P. Crepet, D. De Leo, K. Hawton, C. Haring, H. Hjelmeland, K. Michel, A. Ostamo, E. Salander-Renberg, A. Schmidtke, B. Temesvary, and D. Wasserman. Attempted suicide and major public holidays in Europe: Findings from the WHO/EURO Multicentre Study on Parasuicide. *Acta Psychiatrica Scandinavica*, 99(6):412–8, 1999. [Cited on page 48.]

[132] G. Jessen and B. F. Jensen. Postponed suicide death? Suicides around birthdays and major public holidays. *Suicide & Life-Threatening Behavior*, 29(3):272–283, 1999. [Not cited.]

[133] Ramune Kalediene and Jadvyga Petrauskiene. Inequalities in daily variations of deaths from suicide in Lithuania: identification of possible risk factors. *Suicide & Life-Threatening Behavior*, 34(2):138–146, 2004. [Cited on page 48.]

[134] Stacy Smith, Ananda Allan, Nicola Greenlaw, Sian Finlay, and Chris Isles. Emergency medical admissions, deaths at weekends and the public holiday effect. Cohort study. *Emergency Medicine Journal: EMJ*, 31(1):30–4, 2014. [Cited on page 48.]

[135] Paul K. Edwards, Kristie B. Hadden, Jacob O. Connelly, and C. Lowry Barnes. Effect of total joint arthroplasty surgical day of the week on length of stay and readmissions: A clinical pathway approach. *The Journal of Arthroplasty*, 31(12):2726–2729, 2016. [Not cited.]

[136] C. D. Mathers. Births and perinatal deaths in Australia: variations by day of week. *Journal of Epidemiology and Community Health*, 37:57–62, 1983. [Not cited.]

[137] Chaim M. Bell and Donald A. Redelmeier. Mortality among patients admitted to hospitals on weekends as compared with weekdays. *New England Journal of Medicine*, 345(9):663–668, 2001. [Cited on page 48.]

[138] Fiona Godlee. A seven day NHS. *BMJ*, 352(i1248):1–2, 2016. [Cited on page 49.]

[139] Mohammed A Mohammed, Muhammad Faisal, Donald Richardson, Robin Howes, Kevin Beaston, Kevin Speed, and John Wright. Adjusting for illness severity shows there is no difference in patient mortality at weekends or weekdays for emergency medical admissions. *JAMA: the Journal of the American Medical Association*, 316(24):2593–2594, 2016. [Cited on page 49.]

[140] Jacqui Wise. Study casts more doubt on weekend effect. *BMJ*, 355(i5675):1, 2016. [Cited on page 49.]

[141] Martin McKee, Ben Bray, Rhona Buckingham, and Chris Boulton. The weekend effect: now you see it, now you don't. *BMJ (Clinical research ed.)*, 353(i2750):1–2, 2016. [Not cited.]

[142] Laura Anselmi, Rachel Meacock, Søren Rud Kristensen, Tim Doran, and Matt Sutton. Arrival by ambulance explains variation in mortality by time of admission: retrospective study of admissions to hospital following emergency department attendance in England. *BMJ Quality & Safety*, 0:1–9, 2016. [Cited on page 49.]

[143] Ross Maciejewski, Stephen Rudolph, Shaun J. Grannis, and David S. Ebert. The day-of-the-week effect: A study across the Indiana public health emergency surveillance system. *International Society for Disease Surveillance Annual Conference Advances in Disease Surveillance*, 5(44), 2008. [Cited on pages 49, 56, 69, 89, and 97.]

[144] H. Batal, J. Tench, S. McMillan, J. Adams, and P. S. Mehler. Predicting patient visits to an urgent care clinic using calendar variables. *Academic Emergency Medicine*, 8(1):48–53, 2001. [Not cited.]

[145] Donald R. Holleman and Renee L. Bowling. Predicting daily visits to a walk-in clinic and emergency department using calendar and weather data. *Journal of General Internal Medicine*, 11(4):237–239, 1996. [Cited on page 49.]

[146] Nimal Gamagedara, Jane S. Hocking, Mathew Law, Glenda Fehler, Marcus Y. Chen, Catriona S. Bradshaw, and Christopher K. Fairley. What are seasonal and meteorological factors associated with the number of attendees at a sexual health service? An observational study between 2002-2012. *Sexually Transmitted Infections*, 90(8):635–640, 2014. [Cited on page 49.]

[147] Akerke Baibergenova, Lehana Thabane, Noori Akhtar-Danesh, Mitchell Levine, Amiram Gafni, Rahim Moineddin, and Indra Pulcins. Effect of gender, age, and severity of asthma attack on patterns of emergency department visits due to asthma by month and day of the week. *European Journal of Epidemiology*, 20(11):947–956, 2005. [Cited on pages 49, 56, and 69.]

[148] C. Evans, J. Chalmers, S. Capewell, A. Redpath, A. Finlayson, J. Boyd, J. Pell, J. McMurray, K. Macintyre, and L. Graham. "I don't like Mondays" - day of the week of coronary heart disease deaths in Scotland: study of routinely collected data. *BMJ (Clinical research ed.)*, 320(7229):218–9, 2000. [Cited on page 49.]

[149] S. N. Willich, H. Löwel, M. Lewis, A. Hörmann, H. R. Arntz, and U. Keil. Weekly variation of acute myocardial infarction. Increased Monday risk in the working population. *Circulation*, 90(1):87–93, 1994. [Not cited.]

[150] D. R. Witte, D. E. Grobbee, M. L. Bots, and A. W. Hoes. A meta-analysis of excess cardiac mortality an monday. *European Journal of Epidemiology*, 20(5):401–406, 2005. [Cited on pages 49 and 69.]

[151] Roberto Manfredini, Rodolfo Citro, Mario Previtali, Olga Vriz, Quirino Ciampi, Marco Pascotto, Ercole Tagliamonte, Gennaro Provenza, Fabio Manfredini, and Eduardo Bossone. Monday preference in onset of takotsubo cardiomyopathy. *American Journal of Emergency Medicine*, 28(6):715–719, 2010. [Cited on page 49.]

[152] W. J. Edmunds, C. J. O'Callaghan, and D. J. Nokes. Who mixes with whom? A method to determine the contact patterns of adults that may

lead to the spread of airborne infections. *Proceedings of the Royal Society B*, 264(1384):949–57, 1997. [Cited on page 50.]

[153] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, Janneke Heijne, Malgorzata Sadkowska-Todys, Magdalena Rosinska, and W John Edmunds. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, 5(3):e74, 2008. [Cited on page 50.]

[154] Paul M. Galdas, Francine Cheater, and Paul Marshall. Men and health help-seeking behaviour: Literature review. *Journal of Advanced Nursing*, 49(6):616–623, 2005. [Cited on page 50.]

[155] Joy Adamson, Yoav Ben-Shlomo, Nish Chaturvedi, and Jenny Donovan. Ethnicity, socio-economic position and gender - Do they affect reported healthcare seeking behaviour? *Social Science and Medicine*, 57(5):895–904, 2003. [Cited on page 50.]

[156] Irving Kenneth Zola. Pathways to the doctor - From person to patient. *Social Science and Medicine*, 7(9):677–689, 1973. [Cited on page 50.]

[157] Geraldine M. Leydon, Sheila Turner, Helen Smith, and Paul Little. The journey from self-care to GP care: A qualitative interview study of women presenting with symptoms of urinary tract infection. *British Journal of General Practice*, 59(564):490–495, 2009. [Cited on page 50.]

[158] P. G. Gibson, H. Powell, J. Coughlan, A. J. Wilson, M. Abramson, P. Haywood, A. Bauman, M. J. Hensley, and E. H. Walters. Self-management education and regular practitioner review for adults with asthma. *Cochrane Database of Systematic Reviews*, 3:CD001117, 2002. [Cited on page 51.]

[159] Michael R Gibbons and Patrick Hess. Day of the Week Effects and Asset Returns. *The Journal of Business*, 54(4):579–596, 1981. [Cited on page 56.]

[160] Andrew Rutherford. *ANOVA and ANCOVA: A GLM Approach: Second Edition*. Wiley, 2013. [Cited on pages 57, 59, and 62.]

[161] Angela M. Dean and Daniel Voss. *Design and analysis of experiments*. Springer-Verlag New York, New York, NY, 1999. [Cited on page 59.]

[162] Juliet Popper Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995. [Cited on pages 60 and 62.]

[163] John A. Rafter, Martha L. Abell, and James P. Braselton. Multiple comparison methods for means. *SIAM review*, 44(2):259–278, 2002. [Cited on pages 60 and 62.]

[164] Jessica Middlemis Maher, Jonathan C. Markey, and Diane Ebert-May. The other half of the story: Effect size analysis in quantitative research. *CBE Life Sciences Education*, 12(3):345–351, 2013. [Cited on page 62.]

[165] Gail M Sullivan and Richard Feinn. Using effect size - or why the p value Is not enough. *Journal of Graduate Medical Education*, 4(3):279–82, 2012. [Cited on pages 62 and 63.]

[166] James H. Ware, Frederick Mosteller, Fernando Delgado, Christl Donnelly, and Joseph A. Ingelfinger. p-Values. In John C. Bailar and David C. Hoaglin, editors, *Medical Uses of Statistics*, chapter 10, pages 181–200. CRC Press, 2nd edition, 1992. [Cited on page 63.]

[167] Norman Cliff. Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, 31(3):331–350, 1996. [Cited on page 63.]

[168] Jeanine Romano, Jeffrey D. Kromrey, Jesse Coraggio, Jeff Skowronek, and Linda Devine. Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices? *Annual meeting of the Southern Association for Institutional Research*, 2006. [Cited on pages 63 and 64.]

[169] Jacob Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992. [Cited on page 64.]

[170] Gerald J Hahn and William Q Meeker. Overview of different types of statistical intervals. In *Statistical Intervals: A Guide for Practitioners*, chapter 2, pages 27–40. John Wiley & Sons, Inc, 1991. [Cited on page 77.]

[171] S. S. Wilks. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941. [Cited on pages 77 and 78.]

[172] Derek S. Young. tolerance: An R package for estimating tolerance intervals. *Journal of Statistical Software*, 36(5):1–39, 2010. [Cited on page 78.]

[173] Chia-Chun Tai, Chien-Chang Lee, Chung-Liang Shih, and Shyr-Chyr Chen. Effects of ambient temperature on volume, specialty composition and triage

levels of emergency department visits. *Emergency Medicine Journal : EMJ*, 24(9):641–4, 2007. [Cited on page 82.]

[174] Yan Sun, Bee Hoon Heng, Yian Tay Seow, and Eillyne Seow. Forecasting daily attendances at an emergency department to aid resource planning. *BMC Emergency Medicine*, 9:1, 2009. [Cited on page 82.]

[175] Kiran A. Faryar. The effects of weekday, season, federal holidays, and severe weather conditions on emergency department volume in Montgomery County, Ohio. *Wright State University, Dayton, Ohio*, 2013. [Cited on page 82.]

[176] A. Lee, F. L. Lau, C. B. Hazlett, C. W. Kam, P. Wong, T. W. Wong, and S. Chow. Factors associated with non-urgent utilization of Accident and Emergency services: a case-control study in Hong Kong. *Social Science & Medicine*, 51(7):1075–85, 2000. [Cited on page 85.]

[177] Kenneth D. Mandl, J. Marc Overhage, Michael M. Wagner, William B. Lober, Paola Sebastiani, Farzad Mostashari, Julie A. Pavlin, Per H. Gesteland, Tracee Treadwell, Eileen Koski, Lori Hutwagner, David L. Buckeridge, Raymond D. Aller, and Shaun Grannis. Implementing syndromic surveillance: A practical guide informed by the early experience. *Journal of the American Medical Informatics Association*, 11(2):141–150, 2004. [Cited on page 86.]

[178] Steffen Unkel, C. Paddy Farrington, Paul H. Garthwaite, Chris Robertson, and Nick Andrews. Statistical methods for the prospective detection of infectious disease outbreaks: A review. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 175(1):49–82, 2012. [Cited on page 86.]

[179] Logan Hauenstein, Richard Wojcik, Wayne Loschen, Raj Ashar, Carol Sniegoski, and Nathaniel Tabernero. Putting it together: The biosurveillance information system. *Disease Surveillance: A Public Health Informatics Approach*, pages 193–261, 2006. [Cited on page 88.]

[180] Wolfgang Müller and Heidrun Schumann. Visualization for modeling and simulation: visualization methods for time-dependent data - an overview. *Proceedings of the 2003 Winter Simulation Conference*, pages 737–745, 2003. [Cited on page 88.]

[181] Bircan Erbas and Rob J. Hyndman. Data visualisation for time series in environmental epidemiology. *Journal of Epidemiology and Biostatistics*, 6(6):433–443, 2001. [Cited on page 88.]

[182] Kieran M Moore, Graham Edge, and Andrew R Kurc. Visualization techniques and graphical user interfaces in syndromic surveillance systems. Summary from the Disease Surveillance Workshop, Sept. 11-12, 2007; Bangkok, Thailand. *BMC Proceedings*, 2(Suppl 3):S6, 2008. [Cited on page 88.]

[183] Public Health England. GP in hours: weekly bulletins. `https://www.gov.uk/government/publications/gp-in-hours-bulletin`, 2014. [Cited on pages 88 and 98.]

[184] K. Bollaerts, J. Antoine, E. Robesyn, L. Van Proeyen, J. Vomberg, E. Feys, E. De Decker, and B. Catry. Timeliness of syndromic influenza surveillance through work and school absenteeism. *Archives of Public Health*, 68:115–120, 2010. [Cited on pages 89 and 97.]

[185] H.S. Burkom, S.P. Murphy, and G. Shmueli. Automated time series forecasting for biosurveillance. *Statistics in Medicine*, 26(22):4817–4834, 2007. [Not cited.]

[186] J. J. Van Wijk and E. R. Van Selow. Cluster and calendar based visualization of time series data. *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99)*, pages 1–6, 1999. [Cited on page 89.]

[187] G. Shmueli and H. Burkom. Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics*, 52(1):39–51, 2010. [Cited on page 98.]

[188] Laura Forsberg, Caroline Jeffery, Al Ozonoff, and Marcello Pagano. *A spatiotemporal analysis of syndromic data for biosurveillance*, pages 173–191. Springer New York, New York, NY, 2006. [Cited on page 89.]

[189] Weng-Keen Wong and Andrew W. Moore. Classical time-series methods for biosurveillance. In Michael M. Wagner, Andrew W. Moore, and Ron M. Aryel, editors, *Handbook of biosurveillance*, chapter 14, pages 217 – 234. Academic Press, Burlington, 2006. [Cited on pages 89, 91, and 97.]

[190] Pete Riley, Angelia A Cost, and Steven Riley. Intra-weekly variations of influenza-like illness in military populations. *Military Medicine*, 181(4):364–368, 2016. [Cited on page 91.]

[191] D. M. Fleming and A. J. Elliot. Lessons from 40 years' surveillance of influenza in England and Wales. *Epidemiology and Infection*, 136(7):866–75, 2008. [Cited on page 91.]

[192] Ben Lopman, Ben Armstrong, Christina Atchison, and Jim J Gray. Host, weather and virological factors drive norovirus epidemiology: time-series analysis of laboratory surveillance data in England and Wales. *PloS ONE*, 4(8):e6671, 2009. [Cited on pages 101, 102, 111, and 138.]

[193] Angela Noufaily, Paddy Farrington, Paul Garthwaite, Doyo Gragn Enki, Nick Andrews, and Andre Charlett. Detection of infectious disease outbreaks from laboratory data with reporting delays. *Journal of the American Statistical Association*, 111(514):488–499, 2016. [Cited on page 102.]

[194] John S. Brownstein, Clark C. Freifeld, and Lawrence C. Madoff. Digital disease detection - Harnessing the web for public health surveillance. *The New England Journal of Medicine*, 360(21):2153–2157, 2009. [Cited on page 102.]

[195] Theresa Marie Bernardo, Andrijana Rajic, Ian Young, Katie Robiadek, Mai T. Pham, and Julie A. Funk. Scoping review on search queries and social media for disease surveillance: A chronology of innovation. *Journal of Medical Internet Research*, 15(7):1–13, 2013. [Cited on page 102.]

[196] Per Egil Kummervold, Catherine E. Chronaki, Berthold Lausen, Hans Ulrich Prokosch, Janne Rasmussen, Silvina Santana, Andrzej Staniszewski, and Silje Camilla Wangberg. eHealth trends in Europe 2005-2007: A population-based survey. *Journal of Medical Internet Research*, 10(4):1–10, 2008. [Cited on page 102.]

[197] Susannah Fox, Lee Rainie, John Horrigan, Amanda Lenhart, Tom Spooner, Maura Burke, Oliver Lewis, and Cornelia Carter. The online health care revolution: How the Web helps Americans take better care of themselves. *Pew Internet & American Life Project*, pages 1–23, 2000. [Cited on page 102.]

[198] G. Eysenbach and Ch. Kohler. What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the internet. *AMIA Symposium Proceedings*, pages 225–9, 2003. [Cited on page 103.]

[199] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–4, 2009. [Cited on page 103.]

[200] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. Big data. The parable of Google Flu: traps in big data analysis. *Science*, 343:1203–1205, 2014. [Cited on pages 103 and 135.]

[201] Sudhakar V. Nuti, Brian Wayda, Isuru Ranasinghe, Sisi Wang, Rachel P. Dreyer, Serene I. Chen, and Karthik Murugiah. The use of google trends in health care research: A systematic review. *PLoS ONE*, 9(10), 2014. [Cited on page 103.]

[202] Camille Pelat, Clément Turbelin, Avner Bar-Hen, Antoine Flahault, and Alain-Jacques Valleron. More diseases tracked by using Google Trends. *Emerging Infectious Diseases*, 15(8):1327–1328, 2009. [Cited on pages 104 and 113.]

[203] Rishi Desai, Aron J. Hall, Benjamin A. Lopman, Yair Shimshoni, Marcus Rennick, Niv Efron, Yossi Matias, Manish M. Patel, and Umesh D. Parashar. Norovirus disease surveillance using google internet query share data. *Clinical Infectious Diseases*, 55(8):75–78, 2012. [Cited on page 104.]

[204] Benjamin G. Hassid, Lukejohn W. Day, Mohannad A. Awad, Justin L. Sewell, E. Charles Osterberg, and Benjamin N. Breyer. Using search engine query data to explore the epidemiology of common gastrointestinal symptoms. *Digestive Diseases and Sciences*, 62(3):588–592, 2017. [Cited on pages 104 and 113.]

[205] Anette Hulth, Yvonne Andersson, Kjell-Olof Hedlund, and Mikael Andersson. Eye-opening approach to norovirus surveillance. *Emerging Infectious Diseases*, 16(8):131, 2010. [Cited on page 104.]

[206] Michael Edelstein, Anders Wallensten, Inga Zetterqvist, and Anette Hulth. Detecting the norovirus season in Sweden using search engine data–meeting the needs of hospital infection control teams. *PloS ONE*, 9(6):e100309, 2014. [Cited on pages 104, 114, and 121.]

[207] T Andersson, P Bjelkmar, A Hulth, J Lindh, S Stenmark, and M Widerström. Syndromic surveillance for local outbreak detection and awareness: evaluating outbreak signals of acute gastroenteritis in telephone triage, web-based queries and over-the-counter pharmacy sales. *Epidemiology and Infection*, 142:303–13, 2014. [Cited on page 104.]

[208] H. A. Johnson, M. M. Wagner, W. R. Hogan, W. Chapman, R. T. Olszewski, J. Dowling, and G. Barnas. Analysis of web access logs for surveillance of influenza. *Proceedings of the 11th World Congress on Medical Informatics*, 107(2):1202–1206, 2004. [Cited on page 105.]

[209] Michaël R. Laurent and Tim J. Vickers. Seeking health information online: does Wikipedia matter? *Journal of the American Medical Informatics Association*, 16(4):471–9, 2009. [Cited on pages 105 and 106.]

[210] David J. McIver and John S. Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLoS Computational Biology*, 10(4):e1003581, 2014. [Cited on pages 105, 120, 134, and 136.]

[211] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. Global disease monitoring and forecasting with Wikipedia. *PLoS Computational Biology*, 10(11), 2014. [Cited on pages 105, 120, 121, 129, and 134.]

[212] J .Danielle Sharpe, Richard S. Hopkins, Robert L. Cook, and Catherine W. Striley. Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: A comparative analysis. *JMIR Public Health and Surveillance*, 2(2):e161, 2016. [Cited on page 105.]

[213] Kyle S. Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M. Hyman, Alina Deshpande, and Sara Y. Del Valle. Forecasting the 2013-2014 influenza season using Wikipedia. *PLoS Computational Biology*, 11(5):1–29, 2015. [Cited on page 105.]

[214] Yla Tausczik, Kate Faasse, James W. Pennebaker, and Keith J. Petrie. Public Anxiety and Information Seeking Following the H1N1 Outbreak: Blogs, Newspaper Articles, and Wikipedia Visits. *Health Communication*, 27(2):179–185, 2012. [Cited on page 105.]

[215] Lauren E. Charles-Smith, Tera L. Reynolds, Mark A. Cameron, Mike Conway, Eric H Y Lau, Jennifer M. Olsen, Julie A. Pavlin, Mika Shigematsu, Laura C. Streichert, Katie J. Suda, and Courtney D. Corley. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PLoS ONE*, 10(10):1–20, 2015. [Cited on page 106.]

[216] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Processing*, 150(12):1–6, 2009. [Cited on page 106.]

[217] Ahmed H. Youssefagha, Wasantha P. Jayawardene, and David K. Lohrmann. Role of social media in early warning of norovirus outbreaks: A longitudinal Twitter-based infoveillance. In *WORLDCOMP'13 9th International Conference on Data Mining*, 2013. [Cited on page 106.]

[218] M. Krieck, J. Dreesman, L. Otrusina, and K Denecke. A new age of public health: Identifying disease outbreaks by analyzing Tweets. *Proceedings of*

Health WebScience Workshop, ACM Web Science Conference, 2011. [Cited on page 106.]

[219] R Core Team. R: A Language and Environment for Statistical Computing. `http://www.r-project.org/`, 2017. [Cited on pages 107 and 129.]

[220] Robert E. Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports*, 78(6):494–506, 1963. [Cited on page 107.]

[221] Xiaoli Wang, Shuangsheng Wu, C. Raina MacIntyre, Hongbin Zhang, Weixian Shi, Xiaomin Peng, Wei Duan, Peng Yang, Yi Zhang, and Quanyi Wang. Using an adjusted serfling regression model to improve the early warning at the arrival of peak timing of influenza in beijing. *PLoS ONE*, 10(3):1–14, 2015. [Cited on page 107.]

[222] Sarah Lück, Kevin Thurley, Paul F. Thaben, and Pål O. Westermark. Rhythmic degradation explains and unifies circadian transcriptome and proteome data. *Cell Reports*, 9(2):741–751, 2014. [Cited on page 108.]

[223] Camille Pelat, Pierre-Yves Boëlle, Benjamin J. Cowling, Fabrice Carrat, Antoine Flahault, Séverine Ansart, and Alain-Jacques Valleron. Online detection and quantification of epidemics. *BMC Medical Informatics and Decision Making*, 7(29):1–9, 2007. [Cited on page 108.]

[224] Rob J. Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for R. *Journal Of Statistical Software*, 27(3):C3–C3, 2008. [Cited on pages 108, 109, 128, and 129.]

[225] P. Loveridge, D. Cooper, A. J. Elliot, J. Harris, J. Gray, S. Large, M. Regan, G. E. Smith, and B. Lopman. Vomiting calls to NHS Direct provide an early warning of norovirus outbreaks in hospitals. *The Journal of Hospital Infection*, 74(4):385–93, 2010. [Cited on page 111.]

[226] Hyunyoung Choi and Hal Varian. Predicting the present with Google Trends. *Economic Record*, 88(suppl.1):2–9, 2012. [Cited on page 113.]

[227] Donald R. Olson, Kevin J. Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLoS Computational Biology*, 9(10), 2013. [Cited on page 123.]

[228] R.J. Hyndman and G. Athanasopoulos. Forecasting: principles and practice. `http://otexts.org/fpp/`, 2013. [Cited on page 128.]

[229] Clarence C. Tam, Laura C. Rodrigues, and Sarah J. O'Brien. The study of infectious intestinal disease in England: what risk factors for presentation to general practice tell us about potential for selection bias in case-control studies of reported cases of diarrhoea. *International Journal of Epidemiology*, 32(1):99–105, 2003. [Cited on pages 134 and 138.]

[230] Clarence C. Tam, Laura C. Rodrigues, Laura Viviani, Julie P. Dodds, Meirion R. Evans, Paul R. Hunter, Jim J. Gray, Louise H. Letley, Greta Rait, David S. Tompkins, and Sarah J. O'Brien. Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut*, 61(1):69–77, 2012. [Cited on pages 134, 138, and 151.]

[231] Public Health England Transition Team. Public Health Surveillance: Towards a Public Health Surveillance Strategy for England. *Department of Health*, 2012. [Cited on page 138.]

[232] John P. Harris. Norovirus surveillance: An epidemiological perspective. *Journal of Infectious Diseases*, 213(Suppl 1):S8–S11, 2016. [Cited on page 138.]

[233] Ellen Brooks-Pollock, Natasha Tilston, W John Edmunds, and Ken T D Eames. Using an online survey of healthcare-seeking behaviour to estimate the magnitude and severity of the 2009 H1N1v influenza epidemic in England. *BMC Infectious Diseases*, 11(68), 2011. [Cited on pages 138, 140, and 151.]

[234] Mariam Bibi, Richard W. Attwell, Richard J. Fairhurst, and Susan C. Powell. Variation in the usage of NHS Direct by age, gender and deprivation level. *Journal of Environmental Health Research*, 4(2):63–68, 2005. [Not cited.]

[235] I. Banks. No man's land: men, illness, and the NHS. *BMJ (Clinical research ed.)*, 323(7320):1058–60, 2001. [Cited on page 138.]

[236] Moyses Szklo. Population-based cohort studies. *Epidemiologic Reviews*, 20(1):81–90, 1998. [Cited on page 139.]

[237] Rosebud Roberts, Steven Jacobsen, Thomas Rhodes, W. Terence Reilly, Cynthia Girman, Nicholas J. Talley, and Michael M. Lieber. Urinary incontinence in a community based cohort: prevalence and healthcare-seeking. *Journal of the American Geriatrics Society*, 46:467–472, 1998. [Cited on page 139.]

[238] Steve Brown, Miranda Kim, Clemence Mitchell, and Hazel Inskip. Twenty-five year mortality of a community cohort with schizophrenia. *British Journal of Psychiatry*, 196(2):116–121, 2010. [Cited on page 139.]

[239] Jordi Adamuz, Diego Viasus, Paula CampreciÓs-RodrÍguez, Olga Cañavate-Jurado, Emilio Jiménez-MartÍnez, Pilar Isla, Carolina GarcÍa-Vidal, and Jordi CarratalÇ. A prospective cohort study of healthcare visits and rehospitalizations after discharge of patients with community-acquired pneumonia. *Respirology*, 16(7):1119–1126, 2011. [Cited on page 139.]

[240] Jae W. Song and Kevin C. Chung. Observational studies: Cohort and case-control studies. *Plastic and Reconstructive Surgery*, 126(6):2234–2242, 2011. [Cited on page 139.]

[241] P. Sedgwick. Prospective cohort studies: advantages and disadvantages. *BMJ*, 347, 2013. [Cited on page 139.]

[242] M. A. S. de Wit, M. P. G. Koopmans, L. M. Kortbeek, W. J. B. Wannet, J. Vinjé, F. van Leusden, A. I. M. Bartelds, and Y. T. H. P. van Duynhoven. Sensor, a population-based cohort study on gastroenteritis in the Netherlands: Incidence and etiology. *American Journal of Epidemiology*, 154(7):666, 2001. [Cited on page 139.]

[243] Oktawia P. Wojcik, John S. Brownstein, Rumi Chunara, and Michael A. Johansson. Public health for the people: participatory infectious disease surveillance in the digital age. *Emerging Themes in Epidemiology*, 11:7, 2014. [Cited on page 140.]

[244] Marleen M. H. J. Van Gelder, Reini W. Bretveld, and Nel Roeleveld. Web-based questionnaires: The future in epidemiology? *American Journal of Epidemiology*, 172(11):1292–1298, 2010. [Cited on page 140.]

[245] Alexandra Ekman, Paul W. Dickman, Åsa Klint, Elisabete Weiderpass, and Jan Eric Litton. Feasibility of using web-based questionnaires in large population-based epidemiological studies. *European Journal of Epidemiology*, 21(2):103–111, 2006. [Cited on page 140.]

[246] Natasha L. Tilston, Ken T. D. Eames, Daniela Paolotti, Toby Ealden, and W. John Edmunds. Internet-based surveillance of Influenza-like-illness in the UK during the 2009 H1N1 influenza pandemic. *BMC Public Health*, 10(1):650, 2010. [Cited on pages 140, 141, 142, and 152.]

[247] Alma J. Adler, Ken T. D. Eames, Sebastian Funk, and W. John Edmunds. Incidence and risk factors for influenza-like-illness in the UK: online surveillance using Flusurvey. *BMC Infectious Diseases*, 14:232, 2014. [Cited on pages 140, 150, and 161.]

[248] D. Paolotti, A. Carnahan, V. Colizza, K. Eames, J. Edmunds, G. Gomes, C. Koppeschaar, M. Rehn, R. Smallenburg, C. Turbelin, S. Van Noort, and A. Vespignani. Web-based participatory surveillance of infectious diseases: The Influenzanet participatory surveillance experience. *Clinical Microbiology and Infection*, 20(1):17–21, 2014. [Cited on page 140.]

[249] Craig B. Dalton, Sandra J. Carlson, Michelle T. Butler, John Feisa, Elissa Elvidge, and David N. Durrheim. Flutracking weekly online community survey of influenza-like illness annual report, 2010. *Communicable Diseases Intelligence Quarterly Report*, 35(4):288–293, 2011. [Cited on page 140.]

[250] Mark S. Smolinski, Adam W. Crawley, Kristin Baltrusaitis, Rumi Chunara, Jennifer M. Olsen, Oktawia Wójcik, Mauricio Santillana, Andre Nguyen, and John S. Brownstein. Flu near you: Crowdsourced symptom reporting spanning 2 influenza seasons. *American Journal of Public Health*, 105(10):2124–2130, 2015. [Cited on page 140.]

[251] A. Pini, H. Merk, A. Carnahan, I. Galanis, E. Van Straten, K. Danis, M. Edelstein, and A. Wallensten. High added value of a population-based participatory surveillance system for community acute gastrointestinal, respiratory and influenza-like illnesses in Sweden, 2013-2014 using the web. *Epidemiology and Infection*, 145(6):1193–1202, 2017. [Cited on page 141.]

[252] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall., New York, NY, 1993. [Cited on page 141.]

[253] Rabi Bhattacharya, Lizhen Lin, and Victor Patrangenaru. *A Course in Mathematical Statistics and Large Sample Theory*. Springer-Verlag New York, New York, NY, 2016. [Cited on page 141.]

[254] S. Bennett, A. MacLean, R. S. Miller, C. Aitken, and R. N. Gunson. Increased norovirus activity in Scotland in 2012 is associated with the emergence of a new norovirus GII.4 variant. *Eurosurveillance*, 18(2):2011–2012, 2013. [Cited on page 150.]

[255] J. van Beek, K. Ambert-Balay, N. Botteldoorn, J. S. Eden, J. Fonager, J. Hewitt, N. Iritani, A. Kroneman, H. Vennema, J. Vinje, P. A. White, and

M. Koopmans. Indications for worldwide increased norovirus activity associated with emergence of a new variant of genotype II.4, late 2012. *Euro surveillance: Rapid Communications*, 18(1):8–9, 2013. [Cited on page 150.]

[256] Ben A. Lopman, Mark Reacher, Chris Gallimore, Goutam K. Adak, Jim J. Gray, and David W. G. Brown. A summertime peak of "winter vomiting disease": Surveillance of noroviruses in England and Wales, 1995 to 2002. *BMC Public Health*, 4:1–4, 2003. [Cited on page 150.]

[257] Sander P. van Noort, Cláudia T. Codeço, Carl E. Koppeschaar, Marc van Ranst, Daniela Paolotti, and M. Gabriela M. Gomes. Ten-year performance of Influenzanet: ILI time series, risks, vaccine effects, and care-seeking behaviour. *Epidemics*, 13:28–36, 2015. [Cited on page 151.]

[258] Carrie Reed, Frederick J. Angulo, David L. Swerdlow, Marc Lipsitch, Martin I. Meltzer, Daniel Jernigan, and Lyn Finelli. Estimates of the prevalence of pandemic (H1N1) 2009, United States, april-july 2009. *Emerging Infectious Diseases*, 15(12):2004–2007, 2009. [Cited on page 151.]

[259] Anthony W. Mounts, Tamie Ando, Marion Koopmans, Joseph S. Bresee, Jacqueline Noel, and Roger I. Glass. Cold Weather seasonality of gastroenteritis associated with Norwalk-like viruses. *The Journal of Infectious Diseases*, 181(s2):S284–S287, 2000. [Cited on page 152.]

[260] The Food Standards Agency. Report of the study of infectious intestinal disease in England. *Communicable Disease Report*, 10(51), 2000. [Cited on page 152.]

[261] K. T. D. Eames, E. Brooks-Pollock, D. Paolotti, M. Perosa, C. Gioannini, and W. J. Edmunds. Rapid assessment of influenza vaccine effectiveness: analysis of an internet-based cohort. *Epidemiology and Infection*, 140(07):1309–1315, 2012. [Cited on page 161.]

[262] Anton Camacho, Ken Eames, Alma Adler, Sebastian Funk, and John Edmunds. Estimation of the quality of life effect of seasonal influenza infection in the UK with the internet-based Flusurvey cohort: an observational cohort study. *The Lancet*, 382:S8, 2013. [Cited on page 161.]