Real-time modelling of a Pandemic Influenza Outbreak

Paul J. Birrell[1], **Richard G. Pebody**[2], **André Charlett**[2], **Xu-Sheng Zhang**[2], **and Daniela De Angelis***[1,2]

[1]Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK

[2]Centre for Infectious Disease Surveillance and Control, Public Health England, 61 Colindale Avenue, London, NW9 5EQ, UK

***Corresponding author**: Daniela De Angelis, Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK. E-mail: daniela.deangelis@mrc-bsu.cam.ac.uk

Competing interests: None declared

Keywords: real-time modelling, pandemic influenza, Bayesian statistics, sequential Monte Carlo, syndromic surveillance, GP consultations, serological data, spatial modelling

Type of study: Qualitative and mixed methods

Abstract

Background

Real-time modelling is an essential component of the public health response to an outbreak of pandemic influenza in the United Kingdom. A model for epidemic reconstruction based on realistic epidemic surveillance data has been developed but needs enhancing to provide spatially disaggregated epidemic estimates whilst ensuring real-time implementation is feasible.

Objectives

To advance state-of-the-art real-time pandemic modelling by:

- developing an existing epidemic model to capture spatial variation in transmission;
- devising efficient computational algorithms for the provision of timely statistical analysis;
- incorporating the above in freely available software;

Methods

Markov chain Monte Carlo (MCMC) is used to derive Bayesian statistical inference using 2009 pandemic data from two candidate modelling approaches: a parallel-region (PR) approach, splitting the pandemic into non-interacting epidemics occurring in spatially disjoint regions; and a meta-region (MR) approach, treating the country as a single meta-population with long-range contact rates informed by census data on commuting. Model discrimination is performed through posterior mean deviance statistics alongside more practical considerations.

In a real-time context, the use of Sequential Monte Carlo (SMC) algorithms to carry out real time analyses is investigated as an alternative to MCMC, using simulated data designed to sternly test both algorithms. SMC-derived analyses are compared against "gold-standard" MCMC-derived inferences in terms of estimation quality and computational burden.

Results

The PR approach provides a better, timelier, fit to the epidemic data. Estimates of pandemic quantities of interest are consistent across approaches, and, in the PR approach, across regions (e.g. R_0 is consistently estimated to be 1.76-1.80, dropping by 43%-50% during an over-summer school holiday).

An SMC approach is developed that required some tailoring to tackle a sudden "shock" in the data resulting from a pandemic intervention. This semi-automated SMC algorithm outperforms MCMC, both in terms of precision of estimates and their timely provision.

Software implementing all findings has been developed and installed within PHE with key staff trained in its use.

Limitations

The PR model lacks the predictive power to forecast the spread of infection in the early stages of a pandemic, whereas the MR model may be limited by its dependence on commuting data to describe transmission routes.

As demand for resources increases in a severe pandemic, data from GPs and on hospitalisations may become unreliable or biased.

The SMC algorithm developed is semi-automated, some statistical literacy is required to achieve optimal performance.

Conclusions

Following the study objectives, timely, spatially disaggregate, real-time pandemic inference is feasible and an implementing system has been developed that assumes data as per pandemic preparedness plans.

Future Work

Modelling studies investigating: the impact of pandemic interventions (e.g. vaccination and school closure); the utility of alternative data sources (e.g. internet searches) to augment traditional surveillance; the correct handling of test sensitivity and specificity in serological data, propagating this uncertainty into the real-time modelling.

Study Registration

ISRCTN40334843

Funding Details

This work was supported by the National Institute for Health Research (HTA Project:11/46/03), DDA supported by the UK Medical Research Council (Unit Programme Number U105260566) and by Public Health England.

[495 words]

Contents

List of tables

List of figures

List of abbreviations

CrI	credible interval
ESS	effective sample size.
GOR	Government Office Regions.
GP	general practice.
HDU	high dependency unit.
HPA	Health Protection Agency.
ICC	intra-class correlation coefficient.
ICU	intensive care unit.
ILI	influenza-like illness.
KL	Küllback-Leibler.
MCMC	Markov chain-Monte Carlo.
MR	meta-region.
NHS	National Healthcare Service.
NPFS	National Pandemic Flu Service.
ONS	UK Office for National Statistics.
PHE	Public Health England.
PR	parallel-region.
RCGP	Royal College of General Practitioners.
RMN	Regional Microbiology Network.
SHA	Strategic Health Authority.
SMC	sequential Monte Carlo.
UK	United Kingdom.
USISS	UK Severe Influenza Surveillance System.

Scientific summary

Background

The United Kingdom (UK) National Risk Register lists an outbreak of pandemic influenza as the largest risk faced by the UK outside of acts of terrorism. A prompt public health response to a pandemic, therefore, is vital if the effects of the pandemic are to be mitigated. A business-critical component of such a response is real-time epidemic modelling of the outbreak. As an epidemic progresses, real-time modelling should gradually refine our knowledge about the epidemic and, in particular, the burden it will place on healthcare services. The model should also be able to act as a simulator, allowing for the prospective examination and evaluation of the impact of any proposed epidemic interventions.

Real-time modelling of an ongoing pandemic is, however, not a straightforward task. The model is a tool for statistical analysis of epidemic surveillance data. Model parameters should be progressively informed as data become available over time providing increasingly accurate assessments and prediction of the epidemic evolution. However, as experienced in responding to the 2009 "swine flu" (A/H1N1pdm virus) outbreak, real life epidemic surveillance data are frequently far messier than anticipated, containing contamination, noise and biases that are not always straightforward to foresee. Additionally, epidemics are rarely left to play out naturally. Public health interventions designed either to contain transmission or to relieve the burden placed on healthcare services have the potential to alter drastically our perception of a pandemic or, as in 2009, to interrupt the patterns observed in key data streams. These problems are addressed by using data from multiple sources, so that the different data types can compliment each other to eliminate noise and/or bias.

A modelling framework must, therefore, be able to accommodate this wealth of different data types (typically arriving at daily intervals) and to do so in a timely fashion. It must also be robust to interventions and be able to provide analyses stratified by age groups and geographical location as required by policymakers. Any such model is likely to be highly complex and computationally challenging.

After 2009, a model was developed that could incorporate a number of different types of data but did not allow for spatial stratification. The model was also relatively costly to run in real-time, as it was implemented using a technique for statistical computation called Markov Chain Monte Carlo (MCMC). To this point, the model developed had only ever been used in the retrospective analysis of an entire epidemic, not for real-time purposes. While

MCMC is useful in epidemic reconstruction, it is inefficient for real-time use, as, when new data arrive, it involves the re-analysis of the entire dataset. The ideal method needs, instead, to be sequential so that only the incoming data are used to update the analyses. Sequential Monte Carlo (SMC) techniques provide such an alternative approach.

Objectives

The central objective of this study is to advance the state of the art of real-time modelling of influenza epidemics and to provide a tool to monitor and predict the development of an ongoing pandemic outbreak. These advancements involve:

- Investigating spatial modelling of epidemics to understand how best to account for regionally varying epidemic activity. Two candidate approaches are proposed and examined through the analysis of pandemic surveillance data from 2009. In particular, the strengths and weaknesses of each approach are examined to assess prospectively their potential utility in a future outbreak.
- Building capacity in terms of the different types and increasing volume of data that can be used for real-time modelling.
- Improving the efficiency with which real-time statistical inference can be made.

In light of the above, a suite of software is produced, to achieve the above objectives, designed to provide support to national public health bodies in the event of a pandemic and tailored to the specific requirements of Public Health England (PHE), the responsible public health body in England.

Methods

Two candidate extensions to the existing epidemic model to track accurately a spatially diverse epidemic are proposed. The first approach, labelled the parallel-region (PR) approach, assumes that the epidemic is already established in each region by the time the modelling is initiated. From this point on, inter-region transmission can be considered to be negligible and the epidemics evolve in each region independently, though they will still share some characteristics. Spatially varying parameters, estimated separately for each region, are those depending on population behaviour and composition: the parameters that govern healthcare seeking behaviour; the initial seeding; the reproductive number R_0, the average number of secondary infections caused by a single infection in a fully susceptible population.

The second approach, labelled the meta-region (MR) approach, uses census data on commuting behaviour to generate rates of movement between the different regions. Here, the

country is thought of as a single meta-region, with a population stratified by both age and location. The commuting data inform the relative rates of contact between individuals of each of the strata. There are a number of competing sub-hypotheses within this approach. These regard the effects of density dependence on contact rates, whether it is appropriate to assume individuals commute at random, and the importance of the initial seeding of infection. There will be a preliminary phase of model choice that will identify the relative performance of the model under each combination of these assumptions.

These two methods are implemented to reconstruct the 2009 pandemic. Data include a time series of general practice (GP) consultation data, a short time series of virologically confirmed cases (found through initial attempts at tracing contacts of early infections), virological swabbing data and serological sampling data. The GP consultation data, derived from syndromic surveillance, are counts of individuals reporting a collection of symptoms, known as influenza-like illness (ILI). These data therefore contain a significant proportion of individuals who are consulting for non-pandemic illnesses, contaminating the data. The degree of contamination is identified by the virological swabbing data, where a small sub-sample of individuals consulting their GP give swab samples for testing. The tests will indicate the presence/absence of the pandemic infection. Together, the GP consultation data and the virological swabbing data give a time series that is linked to the pattern of pandemic infection. The serological data come from the testing of blood sera samples taken during the pandemic and tested for the presence of immunity-conferring antibodies. This informs the levels of cumulative incidence, giving scale to the pattern of infection estimated from the GP data. Each data component is therefore vital in disentangling the underlying epidemic dynamics. In future pandemics, it is anticipated that surveillance data of each of these types will be enriched. Additionally, hospitalisation data will be available from all NHS trust hospitals, recording all admissions and, in particular, admissions requiring intensive care, of patients with the pandemic infection.

As an epidemic reconstruction, the spatial analysis is implemented using MCMC.

In a real-time monitoring context, however, it is necessary to develop and test an alternative algorithm to allow for the computationally efficient iteration of epidemic analyses. Starting with a "basic" SMC approach taken from literature, a number of algorithmic developments are proposed to cope with data simulated to correspond to a reasonable "worst-case" epidemic scenario. Here, it is assumed that a public health intervention drastically disrupts the temporal pattern of the observed data and tests the ability of this iterative procedure to respond and adapt appropriately, providing reliable assessment of the epidemic

dynamics. The performance of the SMC algorithm is contrasted against the "gold standard" MCMC in terms of both the quality of the resulting estimates and computational efficiency.

Results

Spatial Modelling

Results show that both the PR and MR approaches are able to reconstruct the epidemic dynamics well, with the PR model providing a better fit to the data than all variants of the MR model. Additionally, the PR approach offers computational benefits, as it can exploit parallel computing, and be implemented in a fraction of the time of the MR approach. However, the PR approach does require significant epidemic activity to have occurred in each region for reliable estimation. If epidemic prediction is required in the early stages of a pandemic, the MR approach might be more useful. Within the MR approach, strong density dependent effects are found in contact rates between individuals, i.e. the chance of two people meeting is inversely proportional to the population size of the region in which they interact. Key model parameters that influence transmission are consistent across both PR and MR approaches, with, in particular R_0 consistently being estimated to be around 1.8. We are also able to estimate the drop in contact due to the over-summer school holidays. Contact rates among 5-14 year-olds fell to 0.6%-0.7% of their term time value, leading to a drop in R_0 of 43%-50%.

Computational Efficiency

In terms of the SMC algorithm to sequentially update estimates as new data become available, in the majority of cases, a fairly straightforward SMC implementation taken from literature would suffice. However, sequential estimation becomes more problematic when the newly observed data are highly informative, as in the situation where a "shock" to the surveillance data is introduced to mimic the effect of a public health intervention. In this case, the SMC algorithm requires very careful construction to ensure it retains the capacity to accurately track the epidemic. An algorithm is pieced together that is semi-automated to minimise the computational effort required to update the analyses. In a moderately complex example, in the immediate aftermath of the intervention, the SMC method even outperforms the MCMC providing more reliable estimates and, with only moderate parallelisation (of the kind that most modern desktops are more than capable), SMC will prove faster to implement.

All of the above models and algorithms have been incorporated into software that is now available for use by PHE in the event of a pandemic. Key PHE personnel are receiving training in its use, which, for the foreseeable future will continue to be offered on an ongoing basis.

Conclusions

The project divides neatly into two components: developing the modelling methodology to provide information to policymakers at the spatial resolution they require; and developing the statistical computing methodology to make robust and timely inference. In terms of the spatial modelling, the PR approach is the most suited approach for real-time epidemic monitoring, even though it has little predictive power early in the epidemic. This shows that either the effects of inter-region transmission are transient, or the available commuting data do not characterise the movement of individuals between regions particularly well. As far as the computational methodology is concerned, in the real-time context, sequential methods for analysis have been shown to be equally adept at providing inference as the more established MCMC, but with a considerable computational advantage. All of the findings have been encoded into software.

The research recommendations arising out of this work are

1. To understand the impact of public health interventions. Here interventions involving school closure and the provision of a service to relieve the burden placed upon GPs have been considered and the model has been accordingly adapted. However, investigation of the modelling adaptations required to incorporate vaccination uptake and effects of antivirals (amongst others) would allow the assessment of these policies, both prospectively and retrospectively. It is anticipated that adaptations to the transmission component of the model to account for such measures would be reasonably straightforward, but to accommodate the data that would inform these adaptations would be more complex.
2. To investigate the utility of alternative sources of epidemic surveillance. The statistical analysis of the epidemic is reliant upon surveillance data on the uptake of healthcare services (GP consultations, hospitalisations). In a widespread and/or severe pandemic, these resources could be severely stretched; hospital beds may not be available and GPs appointment books may be full. At this point, the data generated from these sources may become unreliable. As mentioned in point 1, data on vaccine uptake and antiviral prescriptions administered could potentially help to fill the

knowledge gap, but there are alternative influenza surveillance mechanisms that could be exploited and these should be investigated. Could Internet searches for key influenza terms be useful? Or sales of thermometers? Simplistic modelling studies to investigate these exist, but there is a real gap in the use of this kind of data in a real-world example.

3. To account properly for the uncertainty in serological results. It has been shown that serological data are particularly informative to the kind of modelling effort undertaken in this work. However, the sensitivity and specificity of the testing process is rarely considered, due to uncertainty as to precisely what level of antibodies constitutes long-term immunity to a virus. Proper handling of the uncertainty in these data, as well as their timely provision during a pandemic is essential.

[2122 words]

Plain English summary

In the event of an outbreak of pandemic influenza in England, Public Health England (PHE) has the role of providing up-to-the-moment epidemic assessments to policymakers. To do this, PHE has to make sense of epidemic surveillance data, which are typically incomplete, biased and/or contaminated, and use them to make statements about the present epidemic situation and its likely future path, including estimation of the burden placed on the NHS, and the assessment of the efficacy of proposed interventions.

This is the role of real-time epidemic modelling. A mathematical representation of the ongoing epidemic is developed and used in combination with available epidemic data to produce estimates of key epidemic features and the epidemic trajectory.

The work in this project has enhanced PHE's capacity for carrying out real-time modelling by:

- Adapting an existing epidemic model to produce region-specific epidemic forecasts, increasing its utility to policymakers. Hypotheses regarding how to most appropriately encapsulate transmission of disease within and between regions were assessed both on their fit to data and on their ease of implementation.
- Developing algorithms, building on the latest developments in statistical computation to allow epidemic analyses to be updated in a timely fashion as the epidemic unfolds.
- Establishing a system for analysis of a future pandemic in accord with data scheduled to be available under PHEs strategy for pandemic surveillance, incorporating software and training to key PHE staff.

1 Background

Each year, the United Kingdom (UK) government publishes a document entitled the 'National Risk Register of Civil Emergencies'. The latest edition of the register lists the outbreak of a pandemic influenza virus to be the highest priority non-terrorism risk faced by the UK population.[1] This highlights the importance of the country being in a high state of preparedness for such an outbreak. A key component of any protocol governing the public health response to an outbreak is a plan to monitor and predict the progress of a pandemic in real-time.

During the 2009 outbreak of pandemic A/H1N1 influenza much attention was devoted to the problem of capturing the epidemic dynamics through real-time modelling. The aim of such modelling was to provide up-to-the-moment assessments of the state of the epidemic, as well as to make predictions of its future course, all based upon continually updating streams of information. The models used are mathematical constructs: systems of equations designed to approximate epidemic dynamics, describing the changes over time in the numbers of people within a population who are susceptible to infection, the number currently infected and the number who are presently immune. These equations are governed by a few key (hitherto unknown) quantities known as parameters that usually represent some physical characteristic of the epidemic (e.g., the average duration of infection, or relative rates of contact between members of relevant population groups). To enable assessment of the current state of the epidemic and its future evolution, values for these parameters need to be identified that are consistent with epidemic data. In addition, the uncertainty in the parameter values needs to be properly reflected in such assessments. To make formal, statistical, estimation of model parameters, models can often be simplified to ensure that estimates can be derived from the available data, computational resources and expertise. More generally, as seen in research focused on the evaluation of in-pandemic mitigation strategies,[2, 3] parameter estimates have been obtained on a more ad hoc basis, by using the models and assumed ranges of parameter values to simulate epidemic scenarios. A selection of these parameter values is then retained based on some informal comparison between the corresponding simulated epidemics and the observed data. This type of approach is common to the literature on real-time modelling prior to 2009, in which the proposed methodologies are either heavily reliant upon an idealised set of circumstances and/or ad hoc estimation methods.[4]

Bayesian statistical epidemic models provide a natural, rigorous, framework for the incorporation of relevant contemporaneous surveillance data into the modelling process, alongside collateral information that may be available from other sources. These have been used in the context of real-time monitoring for other infectious diseases. For SARS, models have been proposed and applied for real-time estimation where the focus is on the reproductive number,[5-7] a key epidemic characteristic defined to be the average number of secondary infections caused by a single infection within a fully susceptible population, often denoted R_0. A more complex Bayesian approach is utilised in an application to data stemming from the avian influenza epidemic in the UK poultry industry.[8] Here the availability of individual-level data and the use of computationally intensive Bayesian techniques make it possible to carry out inference on the transmission dynamics, rather than merely the reproductive number. A similar model has been formulated within the Bayesian statistical paradigm to provide real-time estimates of the time-evolving effective reproductive number $R_0(t)$ for a generic emerging disease,[9] an approach that has since been applied to an A/H7N9 outbreak in China and subsequently extended.[10, 11]

However, the modelling approaches above have typically used a single data stream providing direct data on the number of new cases of an infectious disease over time. This is also the case in the context of the 2009 outbreak in Singapore,[12] where a real time reporting system for influenza-like illness (ILI) in sentinel general practices was established, and the resulting data were used to predict the epidemic in real time. In practice, as illustrated by the 2009 outbreak in the UK, direct data are seldom available and, more likely, multiple sources of data exist, each indirectly informing the epidemic development, each subject to possible sources of bias. This calls for more involved complex epidemic modelling that can synthesise the information held within a range of data sources to compensate for the lack of direct observation of the infection process. As a result of this, real-time modelling in the UK in the face of the 2009 A/H1N1pdm outbreak proved to be more demanding and more intricate than had been anticipated.[13]

In response to the 2009 pandemic in England, two approaches to real-time modelling were developed.[14, 15] In the first,[14] the authors present a framework for the real-time assessment of the effectiveness and cost-effectiveness of vaccination strategies, considering the whole of England. Embedded inside a cost-effectiveness model is an age and risk group structured deterministic mass-action SEIR (Susceptible, Exposed, Infectious, Recovered) transmission model, parameterised in terms of an age-specific force of infection i.e. the rate at which susceptible individuals acquire infection. Model parameters are estimated by a hybrid

of ad hoc approaches using as data a scaled version of the estimates of the number of symptomatic cases routinely produced by the Health Protection Agency (from 2011, Public Health England) (HPA) during the outbreak.[16] Uncertainty in key parameters (e.g. R_0) is generated by sampling values of each parameter from a range or distribution to form a 'scenario' from each combination of parameter values. Results from each scenario are compared to the scaled estimates and only the best-fitting 1% of the 60,000 realisations are retained to simulate future incidence and evaluate, with epidemiological uncertainty, the impact of different vaccination strategies on severe outcomes.

In the second approach,[15] data were more directly utilised within a Bayesian statistical framework. The basic modelling features resemble those used to measure the effects of school closure as a strategy for epidemic mitigation in Hong Kong.[17] The primary difference is in the data used, where, instead of using counts of case confirmations alone, an array of different datasets were combined: age- and region-stratified data on GP consultations for ILI;[18] virological positivity data from individuals reporting symptoms of ILI available through the Royal College of General Practitioners (RCGP) surveillance network and the Regional Microbiology Network (RMN) of the HPA; virological case confirmations from the early part of the epidemic; data on the seropositivity of sera samples taken before and during the 2009 pandemic and held by the Weekly Returns Service of the RCGP (see Section 3.1.1 and Appendix 1 for more details).[19] This work, however, considered only the London region. After the 2009 experience, two main issues were left unresolved. The first is the development of a spatial characterisation of the epidemic. This would need to be carried out at a geographical level fine enough to ensure homogeneous epidemic activity within each geographical unit, yet coarse enough to guarantee that the available data in each unit has a sufficiently informative sample size. The second issue is the need to accommodate the greater wealth of epidemic surveillance data supposedly available in future pandemics.[20]
Both developments pose a challenge to existing modelling approaches. In terms of the Bayesian approach,[15] the challenge is to extend the model structure and increase the volume of data to be assimilated into an already complex model in a sufficiently timely fashion for analysis to be feasible in real-time. This requires the development of more computationally efficient methods for Bayesian inference.

1.1 Computational Methods

The Bayesian approach is based on a computational technique known as Markov chain-Monte Carlo (MCMC).[21] MCMC can be computationally burdensome when estimation

and prediction of an evolving epidemic are needed in real time. Every time new data become available, MCMC re-analyses the data in its entirety, requiring possibly millions of evaluations of the model. This is computationally costly and limiting to the speed at which results can be obtained.

Methods to approximate the estimation procedure exist, either by replacing the model with a more-readily evaluated proxy,[22] or by approximating the Bayesian approach.[23] A more appropriate alternative to MCMC are Sequential Monte Carlo (SMC) methods.[24-26] The use of such methods to analyse epidemic data is relatively common,[11,27] yet analyses with a real-time focus are rare,[12] and those using a synthesis of numerous types of data do not exist.
A further complication, not considered in the existing literature, is the need to accommodate the impact that public health interventions might have on the surveillance data underpinning the real-time analysis. As infection becomes more widespread, healthcare facilities become harder to access with those in need of healthcare channelled elsewhere. Any effective real-time computational approach has to cope with the sudden shocks, unforeseen in some cases, that interventions might generate on the time course of the surveillance data.

1.2 Outline

The work reported here will expand upon an existing framework for statistical epidemic modelling,[15] increasing its complexity to allow for spatial heterogeneity in transmission. Two competing spatial modelling approaches are examined (see Sections 3.2 and 4.1), to investigate how epidemic activity in different regions can be most efficiently and accurately estimated. This increased complexity and the extra dimension added to each of the epidemic datasets, add to the computational burden. A general algorithm for Bayesian statistical inference in such a scenario is developed and tested on a suite of synthetic pandemic data, incorporating the presence of ‘shocks’ in surveillance data arising from public health interventions (see Sections 3.6 and 4.2). The report concludes with a discussion (Section 5) and recommendations for future research (Section 6).

2 Study Objectives

The central objective of this study has been to advance the state of the art of real-time modelling of influenza epidemics and to provide a useful tool that can be used to monitor and predict the development of an ongoing pandemic outbreak. This advancement involves:

- Accounting for spatial heterogeneity in transmission. This may be done through the modelling of separate, non-interacting but parametrically linked epidemics in spatially disjoint regions of a country, or through further stratification of the population according to location.
- Building capacity in terms of the different types and increasing volume of data that can be used for real-time modelling.
- Improving the efficiency with which real-time statistical inference can be made.
- Developing a real-time inferential system that is robust to likely pandemic mitigation or treatment interventions.

A suite of software has been produced, to achieve the above objectives, to provide support to national public health bodies in the event of a pandemic and tailored to the specific requirements of PHE, the responsible public health body in England.

Initially, there was also a component of this research promising support to the HPA (now PHE) in the event of a pandemic in their real-time production of estimates and projections of the healthcare burden attributable to the pandemic. Such an outbreak did not occur over the duration of the study and this component of the project has thus been disregarded in the report.

3 Methods

We shall begin by describing (Sections 3.1 and 3.2) the modelling approaches used in this work. In Section 3.3 we will examine the data types that PHE currently envisage being available, at some stage during a pandemic, for inclusion in the analysis. This section also discusses how the structure of the available data from 2009 helped determine the precise parameterisation of the real-time model. Together these will inform the spatial modelling study.

Section 3.5 provides an introduction to Bayesian inference and Section 3.6 discusses the MCMC and SMC methods. Section 3.6 in particular contains a significant amount of technical detail, including the tuning of a number of algorithmic components necessary to achieve timely inference, and may be omitted by the reader not interested in such detail.

3.1 Modelling Methodology: Single-Region Model

The starting point for the investigation in this study is the model and analysis of Birrell et al.[15] Here, information from multiple sources is integrated into a composite model including:

- An age-structured dynamic transmission component;
- A disease component;
- A component describing the mechanism of symptom reporting to healthcare facilities.

A schematic model representation is given in Figure 1. Transmission in the SEIR model is governed by a time-and-age varying force of infection that is dependent upon the population structure, the transmissibility of the virus, the mixing patterns between population strata and the expected time spent in the E and I states. In the disease model layer, a proportion, ϕ, of the newly exposed individuals develop febrile symptoms. In the reporting model layer, further proportions of these symptomatic individuals consult their GP and/or have their symptoms officially confirmed through a virologically positive swab result.

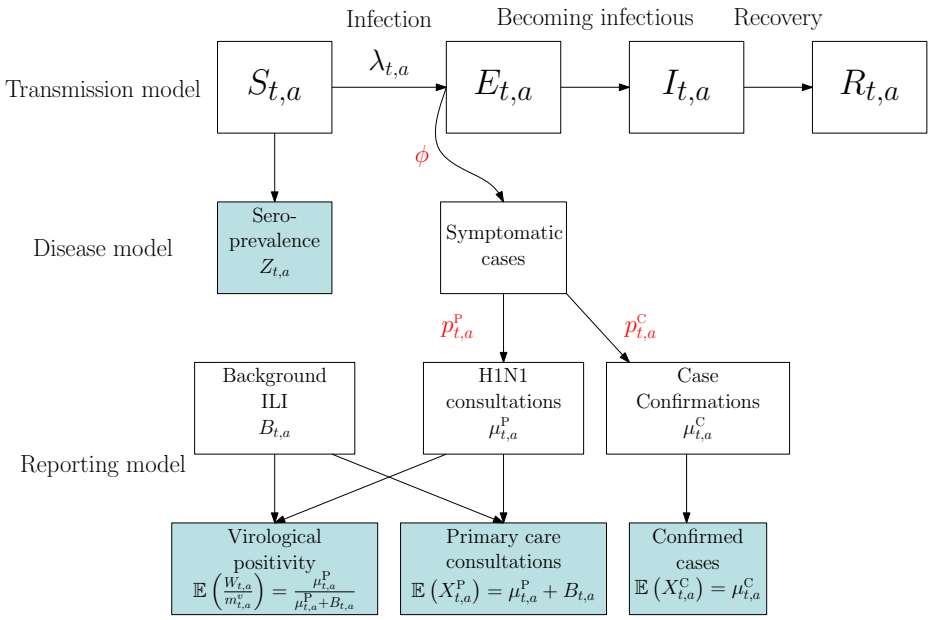

Figure 1 A schematic diagram of the model. Adapted from Birrell et al. (2011).[15]

Ideally, direct data on the number of new infections would be available, and in studies of modelling methodology this type of data are often assumed.[11] However, more realistically, surveillance datasets are noisy and record events (such as GP consultations, see Section 3.3.1) that occur some time later than the infection. Specifically for influenza, disease reporting is frequently through syndromic surveillance, where non-disease specific symptoms are reported. Instead of reporting influenza infections, the reporting is of patients suffering with ILI. Therefore, data from such sources include contamination from patients carrying infections other than the pathogen of interest. This adds greater complexity to the task of disentangling the underlying disease incidence from the available information, particularly as this contamination is likely to vary substantially during the pandemic. To identify the disease incidence, these noisy consultation data are combined with information on virological positivity from complimentary surveillance systems (see Figure 1 and Section 3.3). When multiple time series datasets are available, data on events occurring as close as possible to the time of infection should be preferred as they will be more informative. Alternatively, data arising as a result of severe symptoms are also valuable: severity is a property of the virus and so the proportion of cases that appear in data will be more stable over time.

The equations governing the epidemic dynamics are found in Appendix 1: Single-Region model dynamics.

3.2 Modelling Methodology: Multi-region models

The single-region transmission model of Figure 1 and Appendix 1 is extended to accommodate the evident spatial heterogeneity in the 2009 A/H1N1 pandemic data in two ways: by using a parallel-region (PR) approach or by using a meta-region (MR) approach.[28] These will be introduced in the two following sub-sections.

3.2.1 The parallel-region (PR) model

In the PR modelling approach, the spatially heterogeneous epidemic is assumed to be composed of a number of smaller epidemics occurring in parallel within each spatial unit, with no direct interaction (specifically no transmission) between regions. The rationale here is that the purpose of the real-time model is to monitor the pandemic once infection is widespread. By such a time, it is reasonable to assume that long-range inter-region transmission will be negligible in comparison to that occurring within each region.

The parallel epidemics are still jointly modelled, however, as there is sharing of information on a number of model parameters set to be the same in each region. These parameters are typically those representing biological characteristics of the virus (mean infectious period, proportion symptomatic etc). Additionally, the mixing patterns are assumed to exhibit no regional variation. Appendix 1 presents a system of equations governing single-region dynamics. This system is driven by two key quantities: the reproductive number, R_0; and the initial state of the system, defined by a parameter giving the initial number of infective individuals, I_0. These are region specific parameters ($R_{0,r}$, and $I_{0,r}, r = 1, \dots, R$) as they are functions of both the regional population and the virus. Together, these parameters account for the different timing of the pandemic activity in each region.

The system of dynamic equations given in Equation (8) of Appendix 1 applies within each spatial unit, and so needs little modification.

3.2.2 The meta-region (MR) model

In the MR modelling approach, regions are assumed connected such that transmission is possible between individuals resident in different regions. Here, we look at the country as a whole and treat it as a metapopulation of R regions. Therefore, we can generalise the notation in the system of Equations (8) in Appendix 1 so that the index a now takes values over the range $1, \dots, RA$. It is therefore necessary to define $(RA \times RA)$ contact matrices, $\mathbf{\Pi}(t_k), k = 1, \dots, K$, that describe the rates at which individuals of the various (region- and age-defined) strata come into contact. In the single-region and PR models, this matrix describes the rates at

which individuals of the various age groups interact, and was informed by UK data collected as part of the POLYMOD study (see Equation (10) and relevant text).[29] In this expanded matrix, the entries that correspond to within-region contacts resemble the POLYMOD-based matrices. However, entries corresponding to inter-region interactions are typically of a lower order of magnitude, as people interact less frequently with people living in a different geographic region. These rates of contact are derived from census data on daily commuter movements between regions. The details of how, at the k^{th} timepoint, t_k, the POLYMOD matrices and the commuter data combine to produce contact matrices $\mathbf{\Pi}(t_k)$ are given in Appendix 1. Figure 2 shows a heat map of the elements of $\mathbf{\Pi}(t_1)$ as contact intensities on the absolute and log-scales. The strata are organised within regions, giving the matrix the appearance of an array of sub-matrix blocks within which the POLYMOD patterns of contact are repeated. The blocks on the diagonal give rates of within-region contact and therefore show much higher contact rates.

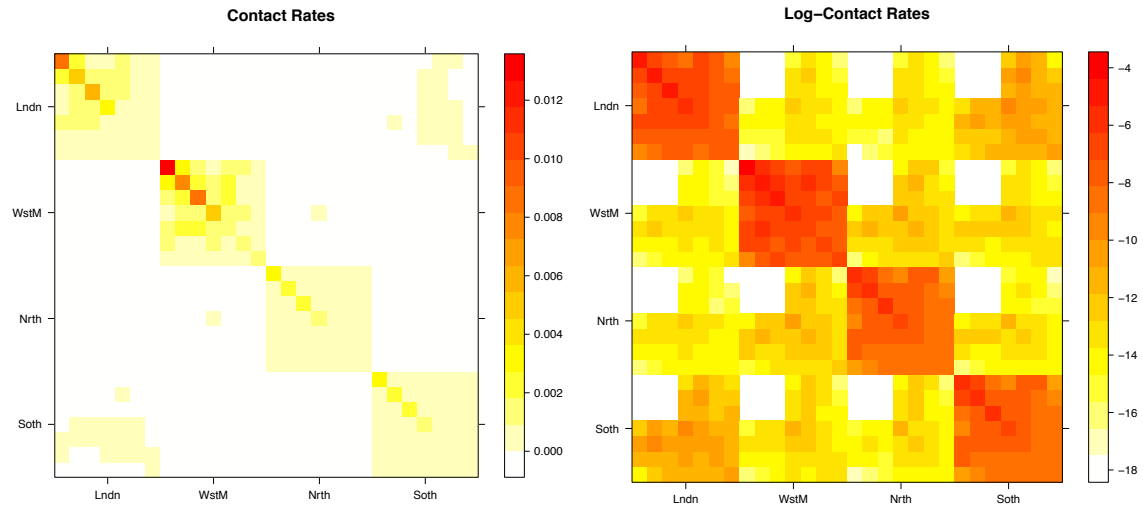

Figure 2: Heat maps for the contact matrices used in the MR model (with regional density dependence) based on contact and log-contact rates respectively. The matrices show a strong block diagonal structure with red areas indicating higher rates of contact. The blocks are ordered such that contacts involving residents of London are in the top row and in the first column, with the ordering of London, West Midlands, North and, in the final row and column, the South. Within blocks, age is increasing top-to-bottom and left-to-right. Red squares indicate interactions of high frequency, white squares interactions of zero frequency. Reproduced from Birrell et al.[28]

The MR model only has one system of dynamic equations of the type in Equation (8) of Appendix 1, removing the flexibility of having region-specific values for R_0 and the initial seeding of infectives, I_0. However, there are a number of modelling considerations to be made when using the MR model, considerations that are not relevant to the PR modelling approach.

Density Dependence: In a single-region model, the POLYMOD-based contact matrices give the relative frequency of contact between pairs of individuals of the different age groups. When these matrices are inserted into the block diagonal of the MR contact matrix (Figure 2), an assumption of frequency-dependent contact is made. This implies that individuals are equally likely to have contact with any other individual in the region irrespective of the region in which they live. Therefore, individuals that live in regions with a higher population will make proportionately more contacts. This may not seem to be a reasonable assumption, as the total population of a region does not necessarily indicate a high population density. The alternative considered here, are density-dependent contacts. In this case, the likelihood of a contact is scaled down by the population size, either of the regional population or of the strata (region and age) population. Both types of density dependence are considered.

Initial Seeding of Infectives: At the beginning of an epidemic, the transmission process is kick-started by a number of initial infectives. The POLYMOD-based contact matrices used for a single region model (and in the PR model) will lead to rapid convergence towards a stable pattern of infection. In other words, for most reasonable choices of initial seeding, it takes only a short time for this seeding to be 'forgotten' by the transmission dynamics.

This is not the case in the MR model, with its block-structured matrix. Infection spreads very slowly between the regions and so the initial seeding is not quickly forgotten. Epidemics seeded with infectives in different regions can lead to very different outcomes. Therefore, while the choice of the seeding in the PR model is not important, it is a significant modelling choice for the MR approach. Therefore, a number of different seedings are considered: the single-region equilibrium distribution (i.e. the stable pattern of infection observed in the PR model); the initial empirical age and region specific distribution of the initial confirmed cases; a hybrid approach using a within-region equilibrium distribution scaled by the empirical distribution of the initial confirmed cases over regions.

Commuting-at-Random: Each model makes the assumption of homogeneous mixing within each stratum. This means that all individuals within a population stratum are as likely to acquire or spread infection as any other individual of the same infection status. In the MR model, where infection is transmitted between regions through the routine movements of commuters, this assumption of homogeneity implies that on any day each individual is equally likely to commute. This is an unrealistic representation, however, as it is likely to be only a subset of people commuting on a regular basis, and a (larger) subset who stay within

their home region. To account for this, adult age groups in each region are further sub-divided into commuters and non-commuters, so that the commuters are a fixed group of people who move each day. This results in the effects of commuting upon transmission being more transient, as there is a smaller, more rapidly exhaustible, supply of susceptible individuals available to transfer infection from one region to another.
The downside to this further level of stratification is that it places an increased computational burden on the model, greatly slowing down the estimation process.

3.3 Data

During the 2009 pandemic, HPA provided A/H1N1pdm incidence estimates for each of the then ten Strategic Health Authorities (SHAs) across England. Two of the SHAs, Greater London and the West Midlands, were believed to have experienced a significant, pre-summer, wave of infection. Ideally we would adopt the same geographical partition. However, the volume of data available from each of the SHAs is insufficient to do so. A reasonable compromise solution is to divide the country into four spatial units: London, West Midlands (the two regions that had significant first waves of infection), North and South. The North region comprises four SHAs: North East, North West, Yorkshire and Humberside and East Midlands. The South region comprises four SHAs: East of England, South Central, South East Coast and South West.

The age categorisation favoured by HPA, now PHE, is to break up the population into the age groups: < 1 years, 1−4 years, 5−14 years, 15-24 years, 25−44 years, 45−64 years, ≥ 65 years. For the rest of this section, a will denote a general population stratum, whether defined by age alone, or by both region and age.

Section 3.3.1 itemises the pandemic data streams that the real-time modelling framework is set-up to work with, detailing, where applicable, how these data were used in modelling the 2009 pandemic and discussing how the various surveillance schemes have evolved over the intervening period. Section 3.3.2 introduces some statistical technical details, showing how these data streams link into the *SEIR* transmission model, covering the distributional assumptions that are required to allow formal statistical inference to be made.

3.3.1 Pandemic Data

GP consultation data PHE carry out syndromic surveillance to monitor influenza activity in the population by routinely collecting data on individuals presenting ILI at GPs. In 2009 such data were provided from two sources. The first source was the Weekly Returns Service of the

Royal College of General Practitioners (RCGP), a sentinel GP network covering a weekly population of approximately 900,000.[30] The second source was the HPA/QSurveillance national surveillance system which covers a much larger population of ≈23 million people.[31] ILI data from both schemes were available stratified by both age group and SHA. In the end, daily ILI reports from the QSurveillance system were used to guide the public health response to the pandemic.[31]

The GP data are reported counts of consultations for non-pandemic specific symptoms and include cases not infected with the pandemic pathogen. Therefore information is required on the proportion of the reported counts that are truly of interest when tracking the levels of transmission of the pandemic infection. RCGP augment their primary care surveillance with virological monitoring.[32] This monitoring involves taking respiratory swabs from a subset (chosen at random) of patients consulting for ILI at participating GPs. A polymerase chain reaction (PCR) assay is then employed to test the swabs for the presence of influenza strains as well as other respiratory virus infections. Similar data are obtained and made available by PHEs Regional Microbiology Network (RMN), covering an additional 400,000 patients in England.[33] A complete account of the virological monitoring undertaken by the two schemes through 2009 can be found elsewhere,[34] but together they provide data on the positivity of the swabs taken by GPs together with the epidemiological information attached to each sample. To ensure high sensitivity of the testing process, swabs were only included in any analysis presented here if the time between symptom onset and the swab being taken was at most five days. Combining this swabbing information with the GP consultation data, the number of consultations that are actually directly due to the pandemic can be estimated.

Since 2009, PHE has expanded its primary care surveillance portfolio, now additionally working with The Phoenix Partnership to access anonymous GP records through their SystmOne computer system.[35] These data could either be combined with the QSurveillance data or could provide an additional sample of data used for model validation. When infection becomes widespread, the National Pandemic Flu Service (NPFS) will be activated. The NPFS is an internet and telephone service designed to expedite the administration of antiviral drugs, alleviating the burden placed upon GP surgeries. In 2009 this service launched and, after a short bedding-in period, was observed to be subject to the same trends as the GP-based data. Those using the service were also swabbed, to understand the underlying pandemic incidence. Due to an anticipated fall in the consultation numbers that would arise as a result of a NPFS launch, these data could easily be added (if the degree

of overlap between the two datasets is understood) or used to replace the GP consultation data to build a picture of the numbers accessing primary care services for ILI.

Virologically confirmed cases Management strategies over the initial stages of the 2009 pandemic were primarily concerned with the containment of the spread of the epidemic, prior to moving into a treatment phase. The initial containment phase was a period of enhanced surveillance during which contacts of known infected individuals were traced and laboratory confirmations of the infection were obtained whenever possible. This work resulted in the generation of the FF100 and the FluZone databases.[34] Routine laboratory confirmations were discontinued on 25th June, but we use here the data only up to 19th June to allow for the gradual cessation in the collection of this type of data. In practice, any real-time modelling is likely not to start within the first five weeks from the start of the outbreak, due to the anticipated difficulty in detecting any signal from the epidemic data at such early stages. Instead it is anticipated that the information on confirmed cases in this period will inform the model construction and provide some prior information (see Sections 3.4.1-2) for various model parameters. In the analysis of the 2009 pandemic data, these data contributed to the analysis in the same way it is proposed (see below and Section 3.3.2) hospitalisation data will in future pandemics.

Hospitalisation data – UK Severe Influenza Surveillance System (USISS) Prior to the 2009 pandemic there was a gap in the surveillance of severe respiratory infections in the UK with regard to hospitalised cases of influenza. During the pandemic a web-based hospital reporting system was established to meet this need. The data were available relatively late in the pandemic and, even now, the biases and weaknesses of the data derived from this reporting system are not well understood. This motivated the development of a more robust, well-tested surveillance scheme for the reporting and handling of such important data. As a result, USISS was initiated during the 2010/11 influenza season, becoming routine for each subsequent season.[36] Data collected during influenza seasons prior to any pandemic outbreak are anticipated to provide baseline information that may prove useful in identifying a pandemic 'signal'.

Outside of a pandemic USISS is a two-stream surveillance system. All National Healthcare Service (NHS) hospital trusts carry out mandatory weekly reporting of admissions of severe influenza cases (i.e. admitted to a high dependency unit (HDU) or an intensive care

unit (ICU) together with laboratory confirmation of infection). USISS also provides sentinel influenza surveillance, through an annually selected random sample of trusts, where testing for the presence of influenza in all patients presenting ILI is mandatory and results are reported together with an array of epidemiological information. In the event of a pandemic being declared, all trusts will switch to this sentinel level of reporting.

In a pandemic, therefore, USISS should provide a time series of reported cases that have two distinct advantages over the GP consultation data:

- They are counts of laboratory confirmed cases, so there is no contamination from non-pandemic ILI.
- The proportion of cases that are reported to USISS should be less volatile over time as it is a function of the severity of the virus and access to hospital services.

However, hospital resources are finite and, in a rapidly developing pandemic, may quickly become exhausted and patients may well be turned away where previously they would have been hospitalised. In such a case, the proportion of cases that are hospitalized will decrease as a function of the increasing incidence. This decrease may be difficult to characterise, potentially limiting the period of time for which the hospitalisation data can be reliably informative.

Serological Data Serological data are the only surveillance data source informing directly the transmission component of the real-time modelling framework. As the prevalence of immunity-conferring antibodies increases, the number of susceptibles decreases. In modelling the 2009 pandemic, the inclusion of serological data has been shown to be crucial to the reconstruction of the underlying epidemic curve.[15]

Initially, the serological data used in the analysis of the 2009 data came from the HPA's annual collection of residual blood serum samples submitted to microbiological laboratories for the purpose of carrying out cross-sectional antibody prevalence studies.[37] Later in the pandemic, it became clear that a more rapid, more representative approach to the collection of serum samples was required. Chemical pathology laboratories were therefore approached at hospitals in each of the RMN regions. This ensured a regular supply of age-stratified serum samples, obtained in a timely fashion with good geographical coverage.[38] In all samples, a haemagglutin-inhibiting antibody titer of 32 was assumed to be sufficient to indicate protection against A/H1N1pdm influenza.[39-41] It is further assumed that there is a two-week delay between infection and seroconversion. Each sample was, therefore, treated to

be representative of the level of cumulative infection among the population 14 days prior to the sampling date. Testing also took place of some residual sera samples collected in 2008 to provide age-specific estimates of baseline antibody prevalence.

Currently, ahead of each winter influenza season, researchers at PHE carry out stratified sampling from the population to select potential participants for a telephone survey regarding the public's attitudes towards influenza vaccination.[36] At the end of the survey, respondents are asked if they would be willing to submit a blood serum sample. Those that agree to take part will submit two samples, one at the start of the season and one at the end of the season. In the event of a pandemic outbreak that does not overlap with the winter flu season, the telephone surveys will be initiated as rapidly as possible.

There is some uncertainty inherent in these data as to precisely what titre value will confer immunity. It is also possible that a different titre level may be required to indicate long-standing immunity to that which indicates recent infection. The real-time modelling system does allow for this potential difference in the titre thresholds but it does not yet account for any uncertainty in these values.

Commuting Data Commuting data were extracted from the UK 2001 census.[42] For individuals aged 16 years and older, these data are in the form of counts of the number of surveyed individuals in each age group and within each Government Office Regions (GOR) who, on the day of the census, travelled into another GOR and how many stayed within their home region. Data were then aggregated so that they conformed to the regional split chosen for modelling the 2009 pandemic - London, West Midlands, North and South. Denote the number of people in age-group a who moved on the day of the census from region r to region s by $C^*_{r,s}(a)$, these numbers were standardised to give

$$C_{r,s}(a) = \frac{C^*_{r,s}(a)}{\sum_{u=1}^{R} C^*_{r,u}(a)}.$$

Equation (13) in Appendix 1 illustrates how these data are combined with information from the UK component of the POLYMOD study to generate contact matrices suitable for use in the MR approach to handling spatially heterogeneous epidemics.

Population totals stratified by age-group and GOR were also derived from UK Office for National Statistics (ONS) data, using the 2008 mid-year estimates.[43]

3.3.2 Distributional Assumptions

Count Data GP consultation data, virological confirmations and hospital admissions are all examples of count data that the real-time modelling framework has been designed to accommodate. We assume these are realisations of either Poisson or Negative Binomial distributions. The expectations of these distributions have derivations that share some common features accounting for:

- A delay from infection to the healthcare event being recorded.
- The fact that these are a proportion of the symptomatic cases: those having a sufficiently severe illness or who make a particular healthcare choice. For data on hospitalisations this proportion is the case-hospitalisation or case-ICU risk, for GP consultations it is a time-evolving propensity for individuals to seek consultation in the presence of symptoms.

Therefore, the expected number of daily reports of hospitalisations, denoted $\mu_a^{\mathrm{h}}(t_k)$ for day t_k within strata a is linked to the daily number of new infections through an expression of the type:

$$\mathbb{E}\left[X_{k,a}^h\right] = \mu_a^{\mathrm{h}}(t_k) = p_a^h(t_k) \sum_{l=0}^{k-1} q_l^h \Delta_a^{(\mathrm{infec})}(t_{k-l}), \tag{1}$$

where $p_a^h(t_k)$ is the relevant case-severity risk, q_l^h is the probability that the time taken from infection to being reported in data as having been hospitalised spans l time intervals, and $\Delta_a^{(\mathrm{infec})}(t_k)$ is the number of new infections at time t_k as found from Equation (11) in Appendix 1.

For GP surveillance data the expected number of consultations arising from the pandemic, $\mu_a^{\mathrm{g}}(t_k)$, is calculated via a similar expression to Equation (1)

$$\mu_a^{\mathrm{g}}(t_k) = p_a^g(t_k) \sum_{l=0}^{k} q_l^g \Delta_a^{(\mathrm{infec})}(t_{k-l}). \tag{2}$$

Before this quantity can be related to data, however, there are a couple of extra considerations:

- The non-pandemic consultations need to be added. Denote by $B_a(t_k)$ the expected values of these at time t_k.
- The within-week pattern of consultations has to be accounted for. Typically no data are reported on weekends and on bank holidays. This leads to a strong artefactual peak in the number of consultations each week on Mondays.

- Although the population coverage of PHE's combined surveillance schemes in England is very high, it is still incomplete and this needs to be accounted for. The expected number of consultations needs to be scaled to allow for this incomplete coverage. Surveillance schemes will report daily coverage figures as a proportion of the total population in each strata, which we denote $D_a(t_k)$.

The expected daily counts of consultations in a general strata a on day t_k are $\mathbb{E}\left[X_{k,a}^g\right]$ such that

$$\mathbb{E}\left[X_{k,a}^g\right] = D_a(t_k)\kappa_{d(t_k)}\left(\mu_a^g(t_k) + B_a(t_k)\right), \tag{3}$$

where $d(t_k)$ indicates the day of the week on which time t_k falls and $\kappa_{d(t_k)}$ is the adjustment factor accounting for the within-week effects on reporting. These factors should be estimated subject to the constraint that $\prod_{d=1}^{7}\kappa_d = 1$.

In particular, the GP data are likely to be highly volatile due to the sensitivity of the population's healthcare seeking behaviour to governmental advice and media reporting. If it is decided that, as in 2009, the most appropriate distribution for the consultation data is the negative binomial, then the real-time model will include dispersion parameters, $\eta_{k,a}^g$, such that the variance is given by:

$$\mathrm{Var}\left[X_{k,a}^g\right] = \left(1 + \eta_{k,a}^g\right)\mathbb{E}\left[X_{k,a}^g\right].$$

Sampling data Both the virological and serological data represent a number of positive readings in a sample of fixed size.

Denote the virological data $\left(m_{k,a}^v, W_{k,a}\right)$, where $m_{k,a}^v$ gives the number of swabs tested within five days of symptom onset and $W_{k,a}$ is the number of those swabs that test positive for the presence of the pandemic pathogen. If we assume the PCR test has test sensitivity k_{sens} and test specificity k_{spec} then these data are binomially distributed with expected value

$$\mathbb{E}\left[W_{k,a}\right] = m_{k,a}^v\left(k_{\text{sens}}\frac{\mu_a^g(t_k)}{\mu_a^g(t_k) + B_a(t_k)} + \left(1 - k_{\text{spec}}\right)\frac{B_a(t_k)}{\mu_a^g(t_k) + B_a(t_k)}\right).$$

In all the analyses presented in this report, the virological testing procedure is assumed to be perfect with $k_{\text{sens}} = k_{\text{spec}} = 1$.

Similarly, we denote the number of blood sera samples that test positive for the presence of antibodies to be $Z_{k,a}$ amongst a total of $m_{k,a}^s$ samples. The expected number of positive samples is linked to the level of susceptibility in the population, $S_a(t_k)$, via:

$$\mathbb{E}[Z_{k,a}] = m_{k,a}^{s}\left(1 - \frac{S_a(t_{k-k_0})}{N_a}\right),$$

where k_0 is a time lag representing the number of time-steps required for the development of antibodies. In 2009 this was taken to correspond to 14 days. If pre-season sampling occurs prior to the chosen t_1, for modelling purposes, these samples can be assumed to be informative about the population prevalence of antibodies on Day 1 of the outbreak and can be added as data at this time.

3.4 Model Parameterisation

Apart from parameters that describe some initial condition of the transmission model, parameters are permitted to vary over time, region and age. Appendix 2 details all the model parameters that can, in principle, be estimated within the real-time model framework. In reality, depending on the availability of relevant data, a subset of parameters is pragmatically chosen for estimation. Table 1 presents a list of the parameters estimated in the spatial analysis of the 2009 pandemic data, indicating whether each parameter varies across regions (denoted 'Spatial') or not (denoted 'Global').

Table 1 Model parameters classified in the PR and MR models as either spatially varying or globally varying. Adapted from Birrell et al.[28]

Parameter	Description	Model	
		PR	MR
$\boldsymbol{\eta}$	Dispersion parameters for GP consultation	Spatial	Spatial
d_I	Average duration of infectious period	Global	Global
ϕ	Proportion of infections that lead to ILI symptoms	Global	Global
$m_k, k = 1, \ldots, 5$	Parameters of the contact matrices*	Global	Global
ψ	Exponential growth rates	Spatial	Global
υ	Initial number of infectives, log-transformed	Spatial	Global
$\boldsymbol{p}^g$	Propensity of ILI patients to consult with their GP	Spatial	Spatial
$\boldsymbol{p}^h$	Propensity of ILI patients to receive case confirmation	Spatial	Spatial
$\boldsymbol{\beta}_B$	Regression parameters determining the rates of background ILI consultation	Spatial	Spatial
κ_d	Day of the week effects on the reporting of GP consultations	Global	Global

*These parameters act as multipliers to elements of the POLYMOD contact matrices[29]: m_1 is the factor by which contact

rates involving adults are down-weighted; m_2, m_3 are reductions in contact rates among children aged 1-4 and 5-14 respectively in the over-summer school holiday; and m_4, m_5 are the corresponding reductions in contact rates for all other school holidays.

Parameters d_I and ϕ are deemed to be properties of the virus and therefore are treated to be constant over region, time and age. Parameters ψ_r and ν_r describe initial conditions (see Appendix 2 for their interpretation) and therefore they have a region-specific value in the PR model and a global value for the MR model. As virological case confirmation data was used in the absence of consistent data on hospitalisations, the proportion of cases that received virological confirmation of their infection, is here set to be $\boldsymbol{p}^h$, is an observation model parameter, relevant only for the first 50 days of the epidemic, while this type of data were still being collected. Therefore, no temporal or age-specific variation is considered, although variation over regions is included on account of the very different levels of pandemic activity in each region over the early period. The specification of the parameter κ_d has already been discussed in the text following Equation (2).

Parameter vector $\boldsymbol{m} = (m_1, \dots, m_5)$ consists of mulitpliers to specified elements of the contact matrices. These parameters are used to measure the impact of school holidays on contacts among 1-4 and 5-14 year-olds, and to down-weight the contribution of all contacts involving at least one adult.

Both parameter vectors $\boldsymbol{\eta}$ and $\boldsymbol{p}^g$ are properties of the reporting model for the GP consultations. Therefore, they both have a temporal changepoint at time $t_k = 83$ days, the time of the NPFS launch. Additionally, $\boldsymbol{p}^g$ differs across ages (different values for children and adults), as well as changing value at two points later in the epidemic to account for the gradual reversion in the public's healthcare-seeking behaviour to pre-NPFS habits. Thus $\boldsymbol{\eta}$ is an 8-dimensional parameter component and the $\boldsymbol{p}^g$ parameters are 32-dimensional.

The parameters describing the rates of non-pandemic ILI consultations, known as the background rates of consultation, have the most complex specification. Regional variation in these rates is specified through a log-linear regression model, allowing information on trends and age effects to be shared across regions. As the background rates of consultation are quite likely to be volatile over time, approximately fortnightly breakpoints are chosen, dividing the 245 days under study into 17 distinct time segments.

The modelling process begins with a first-phase of model choice within the PR modelling framework to specify the precise form of this regression. Letting $\tau(t_k)$ denote the fortnightly

interval into which time t_k falls, and explicitly denoting the strata by (r, a), a saturated model for the consultation rates, $B_{r,a}(t_k)$, takes the form:

$$\log\left(B_{r,a}(t_k)\right) = \mu + \alpha_r + \beta_\tau + \gamma_a + \delta_{r\tau} + \epsilon_{ra} + \varsigma_{\tau a} + \tau_{r\tau a}; \quad (4)$$
$$r \in \{L, W, N, S\}, \tau = 1, \dots, T_X, a = 1, \dots, A.$$

$$\log\left(B_{r,a}(t_k)\right) = \mu^* + \alpha_r^* + \beta_\tau^* + \gamma_a^* + \delta_{r\tau}^* + \epsilon_{ra}^* + \varsigma_{\tau a}^* + \tau_{r\tau a}^*; \quad (5)$$
$$r \in \{L, W, N, S\};\ \tau = T_X + 1, \dots, T;\ a = 1, \dots, A.$$

Parameters in (4) and (5) represent the main effects of region (r), time period (τ) and age group (a) and their interactions; T_X indicates the fortnightly time interval that concludes at the same time as the launch of the NPFS; and the regions $\{L, W, N, S\}$ correspond to London, West Midlands, North and South respectively. This specification implies that there are separate and non-interacting models for the periods pre- and post-NPFS launch. In a preliminary phase of modelling, regression terms from (4) and (5) are sequentially removed until there is an appreciable loss of fit to the data, to obtain simplified versions of the regression equations (see Section 4.1.4).

3.5 Bayesian Inference

In the Bayesian framework, statistical inference about an unknown parameter of interest, θ, proceeds by combining *a priori* information about θ with data from a current study. The initial information on θ is expressed in terms of a probability distribution, $p(\theta)$, known as a prior. This distribution encapsulates all that is known (or not known) about the parameter (e.g. from expert opinion or historical data) before the current study is carried out. After carrying out the study and observing data y, the knowledge about parameter θ is updated to give a probability distribution, known as the posterior, $p(\theta|y)$. This posterior distribution is found through Bayes' formula

$$p(\theta|y) \propto L(y;\theta)\, p(\theta), \quad (6)$$

where $L(y;\theta)$ is the likelihood function, expressing the likelihood of observing data y conditional on the parameter taking value θ. The likelihood for the spatial study and a summary of the chosen prior distributions for the parameters are given in Sections 3.5.1 and 3.5.2.

3.5.1 Likelihood

Denoting K to be the number of days over which we have epidemic data, and using bold to denote (possible) vector quantities, we write that the epidemic dataset is $\boldsymbol{y}_{1:K} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_K)$.

Each data vector $\boldsymbol{y}_k$ contains components $(\boldsymbol{w}_k, \boldsymbol{x}_k^g, \boldsymbol{x}_k^h, \boldsymbol{z}_k)$, consisting of the data types discussed in Section 3.3.2 with each data component containing strata-specific data reported at time t_k.

These data contribute to inference through the likelihood function. Conditional on all the model parameters, using $\boldsymbol{\theta}$ to denote the list of parameters in Table 2, it is assumed that all observations can be considered independent. The likelihood is then expressed as:

$$L(\boldsymbol{y}_{1:K}; \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{a=1}^{RA} \{L(w_{k,a}|m_{k,a}^v, \boldsymbol{\theta}) L(x_{k,a}^g|\boldsymbol{\theta}) L(x_{k,a}^h|\boldsymbol{\theta}) L(z_{k,a}|m_{k,a}^s, \boldsymbol{\theta})\}.$$

The terms inside the product correspond to the likelihood of virological swabbing data, GP consultation data, USISS hospitalisation data and serological data, all reported at time t_k and for each stratum a (assuming here that this encompasses both age group and region).

3.5.2 Priors

Table 2 provides a summary list of model parameters and, in the spatial analysis, their assumed prior distributions, or fixed (known) values as applicable. In some rows of the table, dependence on region has been made explicit through the use of a subscript r.

Table 2 Prior information on model parameters. For each parameter grouping, the table specifies the prior distribution used, or, where the parameter is not to be estimated by the model, its fixed value. Adapted from Birrell et al.[28]

Transmission model parameter	Symbol	Prior/Fixed Value
Exponential growth rates	ψ_r	$\sim\Gamma(6.3, 57)$
Initial log-hazards of GP consultation	υ_r	$\sim N(-19.15, 16.44)$
Mean Infectious Period	d_I	$2 + Z, Z \sim \Gamma(518, 357)$
Mean Latent Period	d_L	2
Contact matrix parameters	m_i	$\sim U[0, 1], \forall i$
Initial proportion susceptible in age group *a*	ρ_a	1 (< 1 years), 0.980 (1-4), 0.969 (5-14), 0.845 (15-24), 0.920 (25-44), 0.865 (45-64), 0.762(65+)
Disease and Reporting model parameters		**Prior/Fixed Value**
Mean (s.d.) of gamma distributed incubation times		1.6(1.8)
Proportion of infections symptomatic	ϕ	$\sim\beta(32.5, 18.5)$
Proportion of cases who consult a GP, varying by age, time and region [Note *i* =1,3,5,7 depending on time interval for	$p_{ra}^g(t_k)$	$p_{ra}^g(t_k) = \log(p_{r,i}/(1 - p_{r,i}))$

child age classes and i = 2, 4, 6, 8 otherwise.		$p_{r,i} \sim \begin{cases} N(-0.187,0.166)\ i = 1,2\ (t_k \leq 83) \\ N(0.426,0.929)\quad i = 3,4\ (t_k \leq 130) \\ N(-0.319,0.263)\ i = 5,6\ (t_k \leq 178) \\ N(-0.284,0.264)\ i = 7,8\ (t_k > 178) \end{cases}$
Proportion of cases lab-confirmed	p_r^h	$\sim\beta(1.03, 2.69)$
Mean (s.d.) of gamma distributed waiting time from symptoms to GP consultation		2.0 (1.2)
Mean (s.d.) of gamma distributed waiting time from symptoms to lab-confirmation		6.6 (3.7)
Mean (s.d.) of gamma distributed reporting delay of GP consultations		0.5 (0.5)
Reporting delay of Lab confirmations		0
GP consultation data dispersion parameters	$\eta_{r,i}$	$\sim\Gamma(0.01, 0.01)$
Regression parameters for the background consultation rates	$\boldsymbol{\beta}_B$	$N_{61}(\mathbf{0}, \boldsymbol{V}^B)$
Day of the week effects on the reporting of ILI cases, log-transformed	$\log \kappa_d$	$N_6(\mathbf{0}, \boldsymbol{V}^\kappa)$

The justifications for the majority of the choices in the table have been given elsewhere[15] and Section 3.4 outlines which parameter components have been considered for the additional spatial variation. Where regional variation exists, parameters are identically distributed in each region.

In short, parameters that are hard to estimate from this type of model and surveillance data, such as d_I and ϕ, have informative prior distributions based on historical studies and analyses of early epidemic data[15]; the prior for $\boldsymbol{p}^g$ uses information from FluSurvey;[44] the priors on the parameters ψ_r and ν_r are given prior distributions that can be considered to be relatively uninformative, with the prior for the components of $\boldsymbol{\eta}$ being particularly diffuse. The bottom two rows of the table are for parameters used to calculate the background consultation rates and the day of the week effects. These are given zero-mean multivariate normal distributions with covariance matrices $\boldsymbol{V}^B$ and $\boldsymbol{V}^\kappa$. The covariance matrices are designed so that the background quantities $B_{r,a}(t_k)$ and $\kappa_d, d = 1, \ldots, 7$, are uncorrelated and identically distributed wherever possible.

3.6 Monte Carlo methods

Typically, the posterior distribution of Equation (6) is only known up to a constant of proportionality and, as a result, is seldom possible to derive analytically, particularly so when

working with a model as complex as that of Figure 1. However, it is possible to obtain a sample from such a distribution. The class of methods used to produce such a sample are called Monte Carlo methods and two of the most common methods are discussed below.

3.6.1 Markov Chain Monte Carlo (MCMC)

MCMC, a widespread and popular algorithm for Bayesian computation, is used to derive estimates of the posterior distribution of the model parameters in the spatial analysis of 2009 pandemic data. More detailed introductions to MCMC can be found elsewhere.[21] However, in short, MCMC techniques are used when it is necessary to sample from a distribution where this sampling cannot be done directly. In any complex modelling scenario, the posterior distribution in Equation (6) represents such a distribution. MCMC works by generating a sequence of values, known as a Markov chain. If allowed to run for long time, this chain will eventually constitute a dependent sample from the desired distribution. Typically, one would run a small number of such chains (say, 2-5), starting each chain at dispersed values: the chains run for a burn-in period until samples derived from each are statistically similar. At this point it can be said that the chains have converged and then the chains run for a sufficient length of time to derive a sample of the desired quality. This can often require many iterations of the chain (in applications of dimension comparable to the dimension of the parameter vector in our example, often $10^4 - 10^6$ iterations may be required) and can often be a time-consuming process as a result.

To see how to generate a sample from a posterior distribution, some technical detail is required. Formally suppose that, at time t_k we are trying to derive a sample from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y}_{1:k})$ where $\boldsymbol{y}_{1:k}$ denotes all the data observed up to the present time. Suppose the parameter value at the n^{th} iteration of the chain is $\boldsymbol{\theta}^n$. From a carefully chosen probability distribution known as the proposal distribution, a new state for the chain is proposed, $\boldsymbol{\theta}^* \sim q_k(\cdot\,|\boldsymbol{\theta}^n)$. This value is then accepted as the next state of the chain with probability

$$\min\left(1, \frac{p(\boldsymbol{\theta}^*|\boldsymbol{y}_{1:k})q_k(\boldsymbol{\theta}^n|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^n|\boldsymbol{y}_{1:k})q_k(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n)}\right). \tag{7}$$

If the proposed parameter value is not accepted then the chain stays where it is and $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n$.

The performance of a MCMC algorithm crucially rests on the choice of the proposal distributions $q_k(\cdot\,|\,\cdot)$. However, regardless of this choice, the algorithm remains highly linear, with minimal scope for taking advantage of the benefits offered by parallel computing. Therefore, this algorithm will struggle to reap any of the benefits of cluster computing. More

importantly, in an iteration of the algorithm, the suitability of proposed values is evaluated using knowledge of the full data likelihood (from time t_0 to time t_k). This will require the evaluation of the system of equations in (8) in Appendix 1 and of Equations (1) and (2). When repeated 10^5 or (orders of magnitude) more times, this can compromise the capacity for timely, real-time inference.

This motivates a more readily parallelisable algorithm, and one that is sequential in nature, demanding only the evaluation of the likelihood of the incoming batch of data, rather than the full data history.

3.6.2 Sequential Monte Carlo (SMC)

In general terms, SMC provides a prescription to sample from a target probability distribution, denoted $\pi(\cdot)$ by sequentially moving through a number of, say L, intermediate distributions $\pi_0(\cdot), \ldots, \pi_L(\cdot) = \pi(\cdot)$. By setting $K = L$ and $\pi_L(\cdot) = p(\cdot\,|\boldsymbol{y}_{1:L})$ it can be seen how this algorithm may lend itself to the problem of online inference. At the k^{th} stage of the sequence of target distributions, a weighted sample of size n_k from $p(\cdot\,|\boldsymbol{y}_{1:k})$ is obtained, denoted:

$$\left\{\left(\omega_k^{(1)}, \boldsymbol{\theta}_k^{(1)}\right), \ldots, \left(\omega_k^{(n_k)}, \boldsymbol{\theta}_k^{(n_k)}\right)\right\}.$$

Here, the weight $\omega_k^{(j)}$ attached to a parameter value $\boldsymbol{\theta}_k^{(j)}$, known in this context as a particle, indicates the relative importance of the j^{th} particle to the sample (known as the particle set). This means that if we have a function of the parameter, such as the epidemic trajectory, which we denote $f(\boldsymbol{\theta})$, then we would estimate it by its weighted mean

$$\frac{1}{n_k}\sum_{l=1}^{n_k} \omega_k^{(j)} f\left(\boldsymbol{\theta}_k^{(j)}\right).$$

The basic idea is that the SMC algorithm proceeds by, upon observing a $(k+1)^{\text{th}}$ batch of data, reweighting the sample according to the likelihood of the new data. This reweighted sample is theoretically representative of the next target distribution $\pi_{k+1}(\cdot) = p(\cdot\,|\boldsymbol{y}_{1:(k+1)})$. Therefore, we can base inferences at time t_{k+1} on the previous sample of parameter values and the new set of weights, which only require the likelihood of the *new* data to be calculated. This represents a significantly reduced computational burden, and as the reweighting for each of the particles can be calculated in parallel, it is a highly parallelisable computation too.

Unfortunately, such a process swiftly suffers from a phenomenon called particle degeneracy. This happens gradually over time as the particle weights scale in such a way that only a very small handful of particles have non-negligible weight. When this degeneracy

occurs, although the weighted sample is of size n_k, the low weight attached to the majority effectively removes them from the sample and estimation and projections are made based on only a handful of particles and are therefore subject to significant error.

To prevent this degeneracy, the sample requires some rejuvenation.[45] The first step of this rejuvenation involves the removal of all those particles of too-low weight. This is done through a process of resampling. Here, a new sample is drawn from the old set of particles according to their weights. The consequence of doing this is that the sample is composed of multiple copies of a much smaller number of identical particles. It is therefore necessary to jitter this sample somehow. To do this, short MCMC implementations are run for each particle, using the current value of the particle as the starting state for the chain.

The SMC algorithm:

A brief overview of this algorithm is below, based on the resample-move algorithm.[46]

1. **Set** $k = 0$. At time t_0, draw a sample $\left\{\boldsymbol{\theta}_0^{(1)}, \ldots, \boldsymbol{\theta}_0^{(n_0)}\right\}$ from the prior distribution, $\pi_0(\boldsymbol{\theta})$, set the weights $\omega_0^{(j)} = 1/n_0$ for all j.
2. **Set** $k = k + 1$. Observe a new batch of data $\boldsymbol{y}_k$. The particles are reweighted according to the likelihood of the incoming data
$$\widetilde{\omega}_k^{(j)} \propto \omega_{k-1}^{(j)} L\left(\boldsymbol{y}_k; \boldsymbol{\theta}_{k-1}^{(j)}\right).$$
3. **Has the particle set become degenerate**? If not, set $\boldsymbol{\theta}_k^{(j)} = \boldsymbol{\theta}_{k-1}^{(j)}, \omega_k^{(j)} = \widetilde{\omega}_k^{(j)}, n_k = n_{k-1}$ and return to step (2).
4. **Resample** Choose n_k and sample $\left\{\widetilde{\boldsymbol{\theta}}_k^{(1)}, \ldots, \widetilde{\boldsymbol{\theta}}_k^{(n_k)}\right\}$ from the set of particles $\left\{\boldsymbol{\theta}_{k-1}^{(1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(n_{k-1})}\right\}$ with probabilities proportional to $\left\{\widetilde{\omega}_k^{(1)}, \ldots, \widetilde{\omega}_k^{(n_{k-1})}\right\}$. Re-set $\omega_k^{(j)} = 1/n_k$.
5. **Move** For all j, move $\widetilde{\boldsymbol{\theta}}_k^{(j)}$ to $\boldsymbol{\theta}_k^{(j)}$ via a short MCMC chain. If $k < K$ return to step 2, otherwise **end**.

Despite the presence of the short MCMC runs in step 5, this still presents a significant improvement over the plain MCMC algorithm because:

- The computationally intensive steps of the algorithm (steps 2 and 5) both allow for calculations on each particle to be made in parallel.
- At most times the particle set will not be degenerate and hence only the likelihood of the new data needs to be calculated to reweight the sample.

- Each of the many parallel MCMC chains can be assumed to start from a point that is sampled from the target distribution. There is therefore no need to allow the chain time to reach convergence and only very low numbers of iterations will be required.
- Before the MCMC phase starts, we already have an estimated sample from the target distribution by using the weighted sample achieved in step 2. This estimate can be used to construct good proposal distributions for the MCMC, improving its efficiency.

A number of algorithmic tweaks have been required to make the algorithm robust to the vagaries of epidemic data and characteristics of the real-time model. The technical detail involved has been presented elsewhere and only a brief overview is given here.[47]

For how long should the MCMC run? The MCMC chains should be run for long enough to have a rich sample of parameter values, but for no longer than is strictly necessary to maintain real-time efficiency. At the start of the MCMC phase, there may be many particles with the same parameter value. These can be defined to be a cluster. The MCMC should be run for long enough so particles from different clusters have fully intermingled and the original clusters are no longer identifiable.

To measure formally the dispersal of the particles, the intra-class correlation coefficient (ICC) is used.[48] This measures the clustering in the value of a summary quantity calculated for each particle. The chosen summary was the projected epidemic "attack rate" (the total cumulative incidence measured as a proportion of the total population).[47] At the start of the MCMC phase, the ICC = 1. As the iterations progress this value will gradually fall and the MCMC iterations will stop once the ICC falls below a pre-defined limit. It has been shown elsewhere that values in the range 0.1-0.2 should be adequate.[47]

Choosing good MCMC proposals Expression (7) gives the acceptance probability for a proposed value for the next state of the chain. Within SMC, it is sought to diversify rapidly the set of particles, without running the chain for too long. To do this, we want to propose values for $\boldsymbol{\theta}^*$ that are not too close to the current values, and that are likely to be accepted. By setting $q_k(\boldsymbol{\theta}|\boldsymbol{\theta}_k^{(n)}) = p_k(\boldsymbol{\theta}|\boldsymbol{y}_{1:k})$, the acceptance ratio would always be 1 (so the proposal will definitely be accepted). This has the added advantage that the proposal is independent of the current state of the chain, so immediately the set of particles would be intermingled.

Unfortunately, we cannot sample directly from $p_k(\boldsymbol{\theta}|\boldsymbol{y}_{1:k})$ so easily. But, after step 2 of the algorithm, we have a weighted sample that should approximate a sample from this

distribution. Therefore, by choosing $q_k(\cdot \mid \cdot)$ to be a multivariate normal distribution centred on the weighted mean and weighted covariance of the particle set calculated at the end of step 2, we have a distribution that approximates the target density and should ensure reasonable rates of acceptance, while rapidly replenishing the particle set. This works well, provided the particle set at the end of step 2 has not become impoverished to the degree it cannot provide a reasonable approximation to the target distribution.[47]

When to rejuvenate? When are the particles degenerate? The standard approach is to rejuvenate the particle set when the effective sample size (ESS) falls below a specific level.[49] The ESS is a measure of the number of independent, equally weighted observations from the target distribution that are as informative as the weighted particle set. At the end of step 2 of the algorithm, the ESS is calculated as:

$$\mathrm{ESS}\left(\left\{\tilde{\omega}_k^{(1)}, \ldots, \tilde{\omega}_k^{(n_{k-1})}\right\}\right) = \frac{\left(\sum_{l=1}^{n_{k-1}} \tilde{\omega}_k^{(l)}\right)^2}{\sum_{j=1}^{n_{k-1}} \left(\tilde{\omega}_k^{(j)}\right)^2}$$

Values of the ESS that are close to n_{k-1} indicate a sample that contains plenty of information about our posterior distribution. A typical level above which the ESS is deemed to be acceptable (and there is no need to rejuvenate the sample) is if $\mathrm{ESS} \geq n_{k-1}/2$.

In some of the examples, such as those considered in Section 4.2, there are times when the addition of the next batch of data in the sequence can lead to a sudden drop in the ESS to very low values. In such cases the MCMC algorithms have too much work to do to adjust the sample and timely inference would not be possible. Therefore, to limit the depletion in the ESS, we introduce rejuvenation steps at intermediate times, between t_k and t_{k+1}, by adding in the data fractionally. If we add a fraction of the data, α, such that $0 < \alpha \leq 1$, then a value α_0 can be identified such that the ESS only falls to approximately $n_k/2$. The next batch of data to arrive is either the remaining portion of the time t_{k+1} data, or a further portion of it, sufficient to once again bring the ESS down to the threshold value. This is the 'real-time' algorithm presented in the technical publication reporting this work.[47]

4. Results

4.1. Spatial modelling

This section presents the statistical results obtained when applying the PR and MR modelling frameworks to reconstruct the 2009 A/H1N1pdm outbreak in England and, in particular, to characterise the impact of inter-region transmission. As discussed in Section 3.2.2, there are a number of competing hypotheses regarding the precise formulation of the MR model and initially we shall present results that assumed the 'best fitting' MR model, before discussing in Section 4.1.5 the exact composition of this model.

4.1.1 Reconstructing the epidemic

Both of the PR and MR models are sufficiently flexible to be able to reproduce the two epidemic waves of the 2009 pandemic. The estimated incidence curves are reproduced in Figure 3. The estimated epidemic in the North is consistent across both models. London and the West Midlands are characterised by bigger first waves of infection (and subsequently smaller second waves) under the PR model, the opposite being true for the South. This is apparent from the peaks in Figure 3 and the given population-level attack rates in Table 3 (age-specific attack rates can be found in Appendix 5). Peak timings in both waves of infection are the same under both modelling approaches and coincide with the start of school holidays, with the exception of the second wave in the West Midlands. Here, a sufficient supply of susceptible individuals remains in the population after the holiday to allow transmission to increase once more (albeit briefly). This may well, however, be a phenomenon of different school term dates in this region to those that predominate elsewhere in the country.

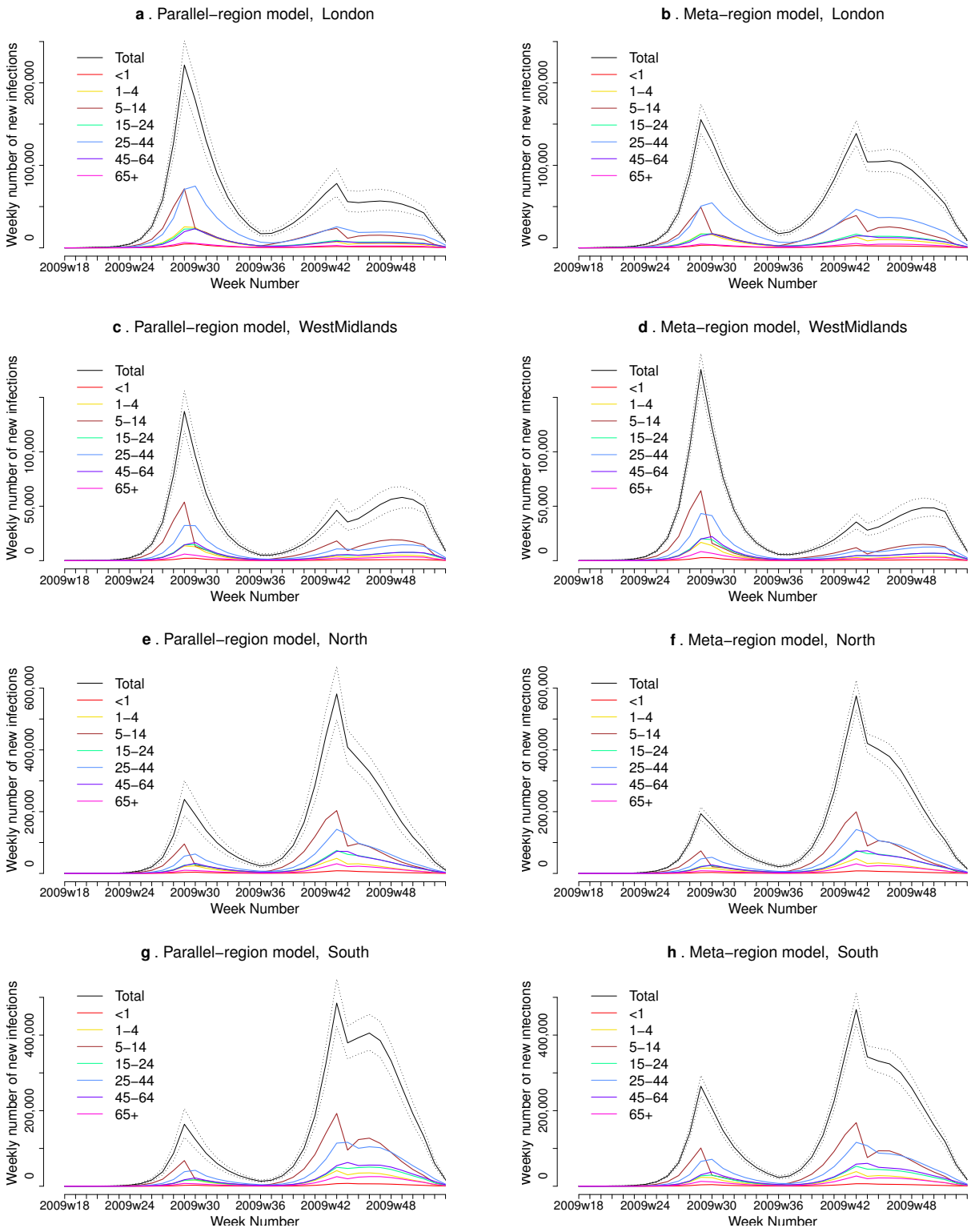

Figure 3 Estimated weekly number of new A/H1N1pdm infections by region (by row) under the PR model (left column) and the MR model (right column). Solid black lines represent incidence summed over age groups with an associated 95% CrI (dashed lines). Figure reproduced from earlier publication.[28]

Table 3 Posterior median and 95% CrIs for cumulative incidence of infection, number of cases (both given in thousands) and attack rates, by region and by pandemic wave (May-August or September-December). Adapted from Birrell et al.[28]

Parallel-region model				
May-August	**London**	**West Midlands**	**North**	**South**
Infections	988 (958, 1124)	525 (456, 600)	1058 (839, 1316)	692 (554, 854)
Cases	152 (123, 184)	80 (65, 98)	161 (121, 215)	105 (80, 139)
Attack rate (%)	13.2 (11.4, 14.9)	9.8 (8.5, 11.2)	5.6 (4.4, 6.9)	3.6 (2.9, 4.5)
September-December				
Infections	764 (641, 901)	571 (483, 656)	3671 (3379, 3987)	3750 (3508, 4021)
Cases	117 (91, 153)	87 (64, 115)	563 (462. 689)	576 (471, 697)
Attack rate (%)	10.1 (8.5, 11.9)	10.6 (9.0, 12.2)	19.3 (17.8, 21.0)	19.6 (18.3, 21.0)
Meta-region model				
May-August	**London**	**West Midlands**	**North**	**South**
Infections	751 (674, 832)	669 (621, 718)	886 (792, 986)	1150 (1036, 1270)
Cases	85 (74, 98)	76 (66, 88)	100 (87,117)	130 (113, 151)
Attack rate (%)	9.9 (8.9,11.0)	12.4 (11.5,13.3)	4.7 (4.2,5.2)	6.0 (5.4, 6.6)
September-December				
Infections	1228 (1129, 1331)	477 (405, 559)	3923 (3721, 4129)	3450 (3256, 3658)
Cases	140 (114, 172)	54 (42, 69)	447 (377, 532)	393 (329, 472)
Attack rate (%)	16.2 (14.9, 17.6)	8.9 (7.5, 10.4)	20.6 (19.6, 21.7)	18.0 (17.0, 19.1)

4.1.2 Estimated epidemic characteristics

Table 4 presents estimates of some key transmission parameters under both models. There is a pleasing consistency across the modelling approaches in the parameter estimates. For example, estimates for the (initial) reproductive number (R_0^{init}), derived from the exponential growth rates, are centred on 1.8, with the region-specific estimates of the PR model being tightly distributed around this value. This is in broad agreement with other estimates for R_0 obtained from a review of 2009 pandemic transmission parameter,[50] and a slight increase on what had been estimated for the single region version of the model.[15] In a similar (single-region) modelling study, much higher estimates for the R_0 associated with the A/H1N1pdm virus have been derived, though this was over the course of a later third wave of pandemic infection occurring in the winter season 2010-11.[51] Similarly, the estimates for the other transmission parameters are robust to the model specification (note the overlapping nature of

the CrIs in Table 4). In particular, parameter m_1 that gives the down-weighting applied to all contacts involving adults, is estimated consistently to be in the range 0.57-0.62. Estimates for m_3 indicate that the summer school holiday period reduced the rate of effective infectious contacts among the 5-14 year-old age group to below 3% of the school term-time figure. However, when averaged over all age groups, this represents a drop in R_0^{init} of between 43% (in London) to 50% (in the South). To compare, a Canadian study recorded a 28% drop in transmissibility during a similar school holiday period.[52] The reduction in the effective contact rates in the other school holidays, as measured by parameters m_4 and m_5 were neither as well estimated (note the width of the credible interval attached to the estimates for parameter m_4) nor did they indicate a similar reduction in the contact rates, the shorter duration of these holidays evidently causing a milder disruption to routine contact patterns. Estimates for the proportion symptomatic, ϕ, do appear to be rather low, although consistent across approaches and with an estimate of 11% based on a closely observed outbreak.[53]

Table 4 Posterior median and 95% CrI for key parameters by modelling approach. Estimates of the reproductive number (R_0) from the PR model are 1.79 (1.74, 1.83), 1.80 (1.76, 1.85), 1.82 (1.78, 1.87), 1.77 (1.73, 1.80) for London, West Midlands, North and South respectively. Adapted from Birrell et al.[28]

Parameter	PR model	MR model
R_0	-	1.81 (1.77, 1.84)
d_I	3.47 (3.35, 3.59)	3.46 (3.34, 3.58)
ϕ	0.154 (0.126, 0.186)	0.114 (0.098, 0.134)
m_1	0.569 (0.536, 0.605)	0.618 (0.584, 0.651)
m_2	0.901 (0.610, 0.996)	0.666 (0.265, 0.740)
m_3	0.007 (0.000, 0.032)	0.006 (0.000, 0.032)
m_4	0.167 (0.008, 0.669)	0.214 (0.004, 0.909)
m_5	0.446 (0.341, 0.557)	0.411 (0.291, 0.528)

4.1.3 Comparison between MR and PR modelling

In Table 5, the posterior mean deviance is used to discriminate between different formulations of the MR model (to be discussed further in Section 4.1.5), comparing each formulation relative to the comparable PR model. If, in fitting these models, MCMC provides a sample of parameter values $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}\}$, then the posterior mean deviance is defined to be $D_m = -(2/n)\sum_{j=1}^{n} \log\left(L_m\left(\boldsymbol{y}_{1:K}; \boldsymbol{\theta}^{(j)}\right)\right)$, where $L_m(\cdot\,;\cdot)$ indicates the likelihood under a specific model, m. It stands to reason, therefore, that lower values of D_m are preferred. The

discrepancy between the PR model and the best performing MR model is 57.89. Due to the regional variation permitted by the PR model in the estimation of R_0^{init} and I_0, the PR model has six more parameters than the MR model. This improvement in deviance for such a small number of parameters suggests that the PR represents a significantly better fit to the data. This compounds the practical benefit of the PR model being markedly faster to implement - it is more suited to parallel computation and the calculation of R_0^* in Equation (10) of Appendix 1, requires the calculation of eigenvalues of (7×7) matrices rather than the (28×28) or (44×44) matrices required by the MR model.

Table 5 Posterior mean (and standard deviation, s.d.) deviances for some candidate parameterisations of the MR model, expressed as a discrepancy from the deviance of the PR model. The smaller values of the posterior mean deviance represent models providing a better fit to the data. Adapted from Birrell et al.[28]

α	Density type	Seed type	Commuting	$\Delta D(\theta)$ (s.d)
0.0	By strata	nextgen	random	3,890 (34.23)
0.0	By strata	nextgen	fixed	4,376 (33.83)
0.0	By strata	empirical	random	4,548 (31.53)
0.0	By strata	empirical	fixed	4,949 (32.21)
0.5	By strata	nextgen	random	3,025 (34.21)
0.5	By strata	nextgen	fixed	2,241 (32.13)
0.5	By strata	empirical	random	3,191 (36.65)
0.5	By strata	empirical	fixed	2,269 (31.90)
1.0	By strata	nextgen	random	2,770 (30.23)
1.0	By strata	nextgen	fixed	2,466 (30.05)
1.0	By strata	empirical	random	2,578 (29.43)
1.0	By strata	empirical	fixed	2,359 (29.39)
1.0	By region	nextgen	random	449.2 (27.60)
1.0	By region	nextgen	fixed	437.9 (27.21)
1.0	By region	empirical	random	170.1 (28.47)
1.0	By region	empirical	fixed	166.4 (29.76)
1.0	By region	hybrid	random	57.89 (27.20)
PR model				0

4.1.4 Finding an optimal parameterisation

Inferences drawn from either the PR or the MR modelling approach are found to be sensitive to the precise form of the regression for the background rates of GP consultation. Because of this, it was important to implement sub-models of Equations (4) and (5) in order to most appropriately characterise the changes in consultation behaviour over the pandemic period. Again, the posterior mean deviance was used to identify a most appropriate model. The real-time PR model was repeatedly implemented with the higher order interactions systematically removed from equations (4) and (5) in the hope of finding simplified regression models without incurring any significant loss of fit to the data. Additionally, some age groups and regions were paired together, to cover gaps where data was too sparse to warrant the additional age/region effects. Under the PR model, the seemingly optimal choice for the regression model, and the one that has been used in the generation of all the results presented in this Section is:

$$\log\left(B_{r,a}(t_k)\right) = \mu + \alpha_r + \beta_\tau + \gamma_a + \delta_{r\tau} + \epsilon_{ra};\ r \in \{L, W, S\};\ \tau = 1, \dots T_X; a = 2, \dots, A$$

$$\log\left(B_{r,a}(t_n)\right) = \mu^* + \alpha_r^* + \beta_\tau^* + \gamma_a^* + \epsilon_{ra}^*;\ r \in \{L, W, S\};\ \tau = T_X + 1, \dots, T; a = 2, \dots, A.$$

Here, the rates for the North and South have been equated $B_{N,a}(t_k) = B_{S,a}(t_k)$, an unsurprising finding given that there is sparse virological data in the North to accurately estimate the non-pandemic consultation rates here. Also, the rates in the two youngest age groups have been set to be equal (note the sum over the a index omits a=1), $B_{r,1}(t_k) = B_{r,2}(t_k)$, again a not unreasonable finding given that the only the virological swabbing and not the QSurveillance GP datasets provide data with sufficient granularity to distinguish between the first two age groups (the < 1 and 1-4 year olds).

When the same model refinement process was undertaken using the MR model, the same regression equations were again preferred.

Having established the form of the background consultation rate regression, the next stage of model building in the MR approach was to consider the alternative model formulations of Section 3.2.2, governing how the model handles density dependence, random commuting and the choice of the initial seeding. Examination of the posterior mean deviances presented in Table 5 shows that density dependence is best accounted for by scaling entries of the contact matrices by the population of the region, not the population of the relevant strata, i.e. by replacing $N_{r,a}$ and $N_{v,a}$ in Appendix 1 Equation (13) with N_r and N_v, the sum of the regional populations over age-groups. Furthermore, it was found that within-region

transmission that is density dependent (corresponding to the case $\alpha = 1$ in Equation (13) of Appendix 1) gave better model fit than either frequency dependent transmission ($\alpha = 0$) or a mixture of the two ($\alpha = 0.5$).

MR model performance is highly sensitive to the choice of initial seeding of infectivity, with the hybrid seed performing most strongly. When the number of strata is expanded to partition between non-commuting and commuting adults, there was no consistent improvement in model performance (nor any particular worsening). However, the extra complexity and computation required to evaluate the model with the expanded number of strata indicates that the reduced stratification would be preferred in a real-time context. This suggests that the effects of inter-region transmission are either highly transient, sufficiently so that its effects are swallowed up by the choice of seeding, or the movement of individuals between regions is poorly characterised by the commuting data. The formulation of the contact matrices assumes that infected individuals move as freely as uninfected individuals and this may well be unrealistic. However, accounting for this would further reduce any difference in model performance between the two approaches, leading to no material adjustment of the conclusions.

All results presented that quote the MR model will refer to the best-performing variant with $\alpha = 1$, density dependence governed by the regional population size, using the hybrid seed and assuming commuting at random.

4.1.5 Goodness-of-fit

Appendices C and D give goodness-of-fit plots for three of the data types (GP consultation data, virological positivity and serological) under the PR and MR models respectively. There is no apparent lack of fit under either model, with most data points lying within the 95% predictive intervals. There are a couple of instances where the model predicted seropositivity is too high (Greater London, ≥ 65) and others where it is too low (Greater London 5-14). It would seem that this arrives due to poor estimation of the initial proportion of susceptible individuals (this is *a priori* estimation, it was not carried out as part of the real-time model effort). Even in these cases, the PR model gives the better fit to these outlying data points, with fewer points missing their predictive intervals. Elsewhere, the performance of the PR model is evidently superior, in accordance with findings already presented.

4.2 Comparison of the real-time performance of the Monte Carlo methods

The results of Section 4.1 were all obtained using MCMC. A posterior sample from the PR model could be derived in about 13-15 hours using MCMC, with some parallelisation of the likelihood. The MR approaches took considerably longer, particularly under the fixed commuter assumption when there were 44 population strata. At this point, the run-time stretches into days. Even at 13 hours, however, this is longer than a typical working day and eliminates the possibility of providing real-time analysis. Furthermore, it is not yet considered that in any future pandemic, there should be a sufficient wealth of data to allow a greater sub-division of England, at least into the nine GORs, or that the improved quality of surveillance data might lead to an expansion in the number and type of parameters that can be estimated. Alternatively, the next pandemic to occur might be longer lasting, giving longer time series of data. All of these factors can greatly increase the level of computation required to draw the required statistical inference. It is not desirable that estimates are rendered obsolete by new data before they can be produced.

This highlights the importance of developing a good statistical algorithm for analysing the data in a timely fashion. The algorithm has also to be able to be expressed in sufficient generality that it can be encoded within software for future use by an infectious disease epidemiologist whose knowledge of computational techniques in statistics may be minimal. As alluded to in Section 1, there is reason to believe that SMC algorithms may permit the iteration of analysis in a much more timely fashion. In this section, the efforts to tailor a suitable yet reasonably general SMC algorithm are discussed, testing the approach against simulated data, the generation of which is described in what follows. The central idea is that the data should be realistic, yet have features that are challenging to track, a reasonable "worst-case" scenario.

4.2.1 Simulated data

It was decided to copy many of the features of the 2009 pandemic. The simulated outbreak starts with an initial burst of infections in the spring, so that the epidemic is in full exponential growth by the time of an over-summer school holiday. The school holiday acts as a break on transmission, partitioning the outbreak into two distinct waves of infection. Although we only consider this one underlying epidemic, we consider two different data scenarios. In the first scenario, it is assumed that there is direct information on confirmed cases, such as might occur in the surveillance of severe disease (e.g. hospitalisation or ICU data from USISS). In the second scenario, ILI consultations, contaminated by non-pandemic

infections replace the confirmed case data. Both data streams are assumed to exist alongside serological data. In the second scenario it is necessary to have companion virological swabbing data to identify the degree of contamination in the ILI data.

To ensure that the task of epidemic tracking is realistic, in the same vein as the NPFS introduction in 2009, it is assumed that there is a "shock" in the data provided by the surveillance schemes. Such a shock would be provided by a public health intervention designed to alleviate overcrowding in primary healthcare services or to reduce the demand on hospital beds. The net result of the intervention is that a much-reduced proportion of cases report their symptoms to the respective surveillance schemes, with the timing of the intervention, as in 2009, following shortly on from the over-summer closure of schools for the holiday period.

To illustrate the size of the system shock that is being considered, the synthetic data to be used in the second scenario are presented in Figure 4.

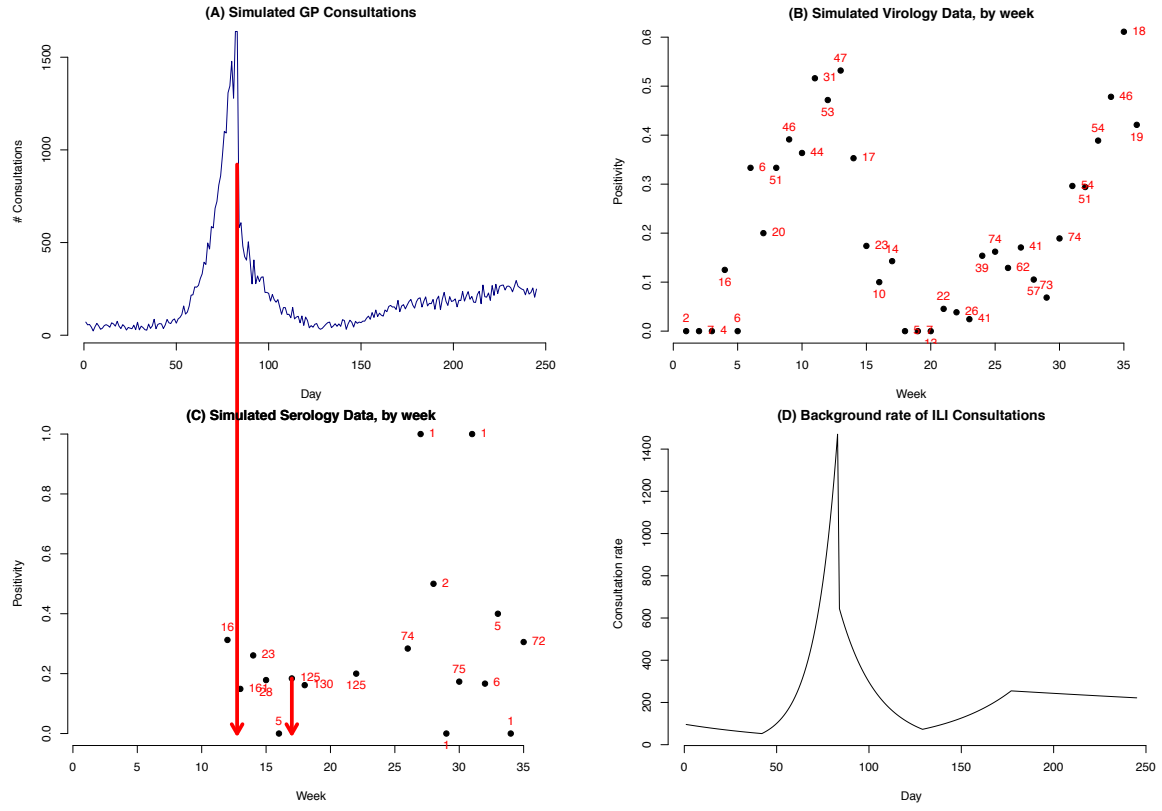

Figure 4 Top row: (A) Observed number of GP consultations; (B) Swab positivity data with numbers representing the size of the weekly denominator. Bottom row: (C) serological data; (D) the pattern of background consultation rates for the GP consultation data aggregated over ages. The red arrows over figures (A) and (C) highlight the timing of some key, informative observations.

4.2.2 Scenario 1: A naïve algorithm

Here we want to compare the relative performance of the SMC algorithm against what we consider to be a gold standard, the MCMC algorithm that was used to derive epidemic inference in earlier analysis of the 2009 data.[15]

To attempt this, MCMC analyses are carried out after 50, 70, 83, 120, 164 and 245 days of data have been observed. The SMC algorithm was then applied starting from the MCMC-derived posterior from 50 days, to see if, after the addition of 20 consecutive days, or batches, of data, the SMC and MCMC derived posteriors are statistically similar. This process was repeated to see if SMC could also bridge the gap between the MCMC analyses at days 70 and 83, days 83 and 120, days 120 and 164, and days 164 and 245.

Initially, a fast, naive SMC algorithm was tried in application to the first data scenario, where we have contamination-free hospitalisation data. Here, the MCMC step embedded within the SMC algorithm would only last for one iteration, and rejuvenation of the particle set would only take place after the assimilation of whole batches of data (none of the fractional addition of data discussed in the 'When to rejuvenate' discussion in Section 3.6.2. Figure 5 shows some of the results of doing this. In the scatter plots, the grey points show the starting MCMC-obtained distribution. Against this, left hand scatter plots are to be compared to the right hand scatter plots. In the right-hand side, all (non-grey) scattered points are of the same colour, because they are all of equal weight, equal importance. In the left-hand column, the darker the point, the greater weight it carries.

The immediate point to notice is that the SMC-obtained posteriors in the top and bottom panels would appear to be comparable to the MCMC-obtained posterior distributions, but the posterior distribution obtained by the naive SMC algorithm at time $t_k = 120$ displays significant degeneracy. The algorithm has not tracked the movement of the posterior density over the interval from 83 days to 120 days. It is this sample impoverishment that makes the naïve SMC inefficient at such a time.

Referring to the plots of the simulated data in Figure 4, the superimposed vertical red arrows identify points in time where there are particularly informative observations. Immediately after time $t_k = 83$ there is a shock to the epidemic system as a public health intervention cuts the proportion of infections that are reported into data. At time $t_k = 110$, there is a particularly large batch of serological data, data that is particularly informative. These two occurrences make it particularly hard for the naive algorithm to track the epidemic. Particularly, after the $t_k = 83$ shock, a number of new parameters become active

(i.e. begin to have influence over the likelihood). As parameters first begin to move away from their prior density, it can cause severe depletion of the particle set. In comparison, the 50 to 70 day and the 164 to 245 day intervals are relatively uneventful and much easier for the algorithm to track.

It is this phenomenon that motivates a focus on the day 83 to day 120 interval moving forward, and also motivated the algorithmic adaptations discussed at the end of Section 3.6.2. We also consider the second data scenario, where we consider the syndromic GP counts, inclusive of non-pandemic noise and virological swabbing data.

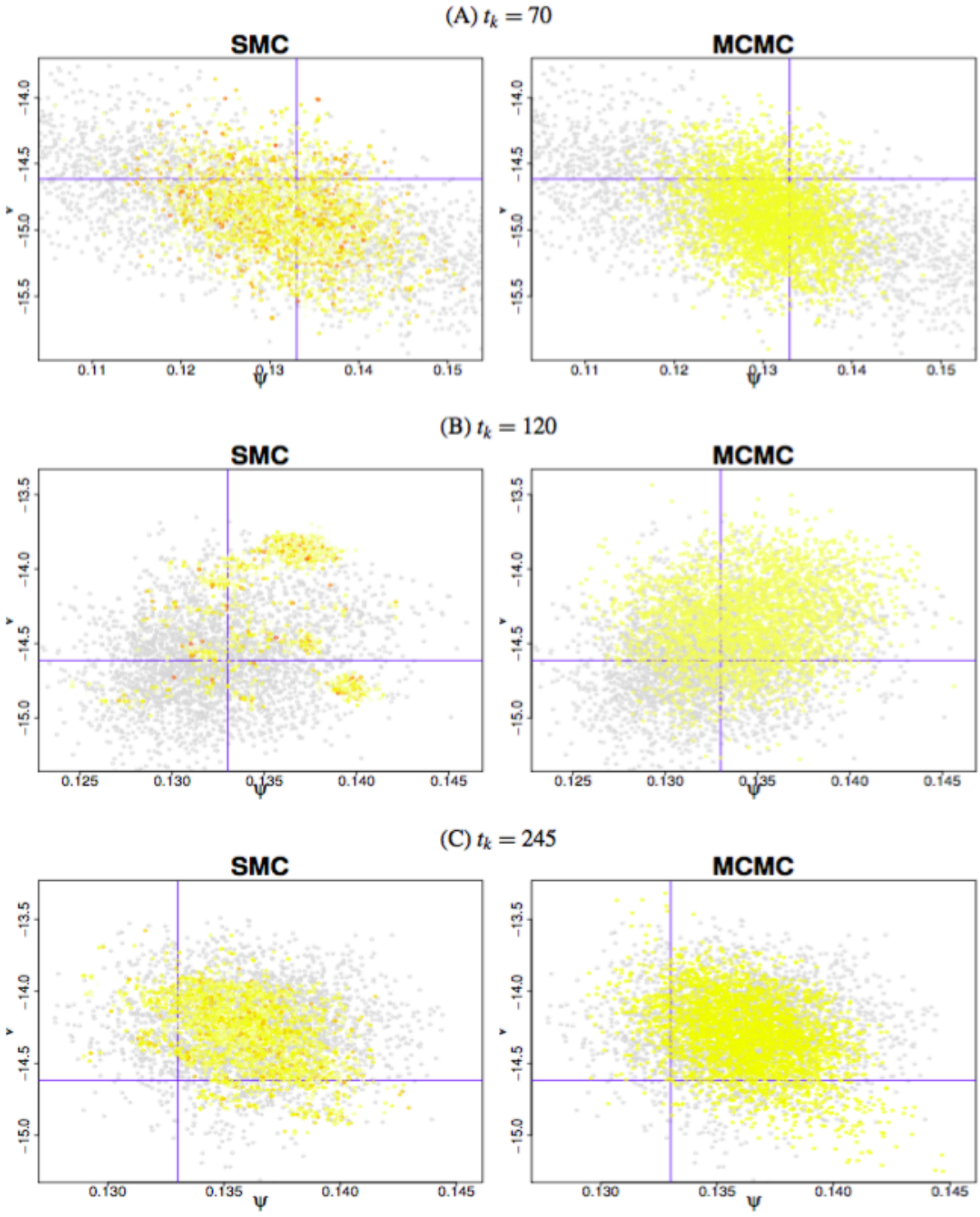

Figure 5 Comparison of naive SMC-obtained posteriors and MCMC-obtained posteriors at $t_k = 70$ (A), $t_k = 120$ (B) and $t_k = 245$ (C) days, via scatter plots for the parameters ψ and υ.

4.2.3 Scenario 2: Heavy-duty SMC

If the MCMC-derived posteriors are to be treated as a (albeit computationally costly) gold-standard, a measure of similarity is needed between the SMC- and the MCMC-derived distributions, and for this we use Küllback-Leibler (KL) divergence.[54] This is a statistic that

gives a measure of how different an estimate for a probability distribution is to its 'true' target. Here, we are presuming that it is the MCMC that represents the truth.

Table 6 gives the KL statistics achieved when the full SMC algorithm was used over the day 83 to day 120 interval, breaking the interval down even further so we can look at KL discrepancies in the immediate aftermath of the shock at $t_k = 83$ days. The table also shows three different levels of ICC threshold used as a stopping criterion for the MCMC phase. At each time, the 'gold standard' MCMC analysis was repeated numerous times and then referred back to a reference analysis. This was done to build up a distribution of KL statistics, so that the SMC analysis could be given a KL 'target', the upper 95 percentile of the KL statistics calculated on the sample of MCMC analyses. If the SMC analysis had a KL statistic that was lower than the target value, then it could be said to be indistinguishable from the MCMC analyses. At this point, we know that the SMC algorithm is responding adequately well.

Table 6 Performance of the adapted SMC algorithm over the interval 83-120 days by ICC threshold.

ICC threshold	0.5	0.2	0.1
84 Days (KL target = 0.732)	2.92	2.87	2.83
85 Days (KL target = 0.135)	3.05	3.00	2.98
86 Days (KL target = 0.365)	3.28	3.24	3.25
87 Days (KL target = 0.276)	2.54	2.45	2.42
90 Days (KL target = 0.159)	1.80	0.353	0.0663
100 Days (KL target = 0.135)	0.157	0.102	0.0890
110 Days (KL target = 0.122)	0.159	0.0774	0.111
120 Days (KL target = 0.119)	0.136	0.0435	0.0708

From Table 6 it can be seen that from about day 87 onwards, the SMC algorithm (for the ICC thresholds 0.1 and 0.2) begins to regularly hit its KL target. The problem, therefore, lies in the days immediate preceding this point in time, where, not only are the KL statistics large, but the MCMC-component of the SMC algorithm was requiring vast numbers of iterations and becoming prohibitively time consuming.

Addressing the speed issue first, it was found that one particular parameter was causing the slow convergence. As discussed elsewhere,[47] it was found that, if proposals for the overdispersion for the negative binomial data $\boldsymbol{\eta}$ were made separately to the rest of the parameter vector $\boldsymbol{\theta}$ (which is updated together in one block), convergence could be achieved

much more rapidly. Figure 6 shows the improvement in the number of iterations required per day from over 400 per particle under the original SMC algorithm to 70 under the tailored version on day 89, with this improved performance evident for a number of days in the aftermath of day 83. Under both schemas, rejuvenations are required at the same times, but don't require the same computational effort. Both versions of the algorithm perform similarly from around day 90-91 onwards.

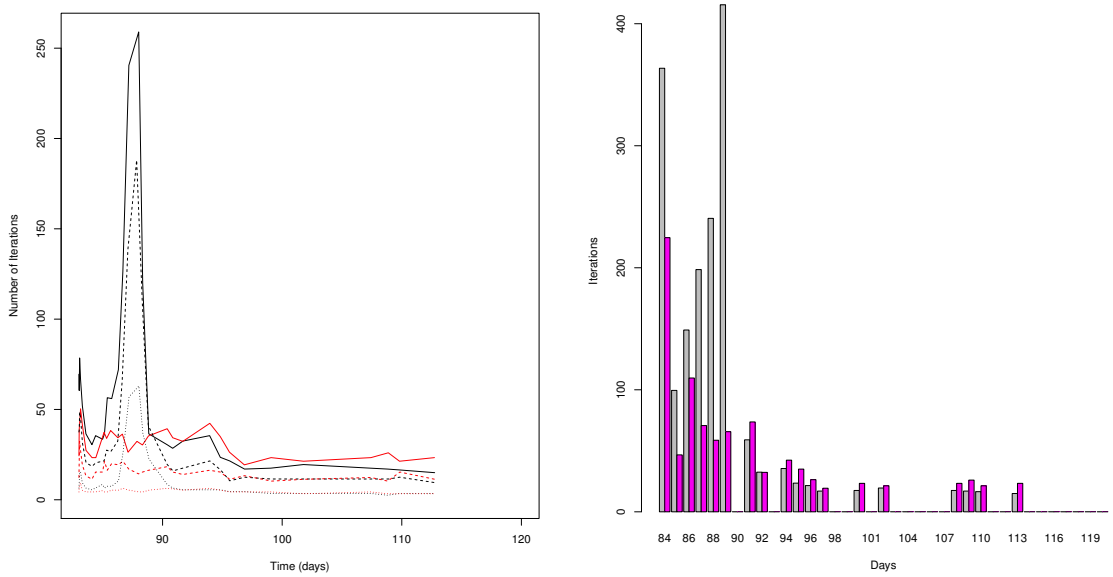

Figure 6 (LHS panel) Number of MH-steps required by the continuous-time SMC algorithm per rejuvenation against the timing of the rejuvenation for both the continuous-time algorithms (black and red correspond to with and without η in the block updates) with the ICC threshold = 0.1 (solid line), 0.2 (dashed line) and 0.5 (dotted line). (RHS panel) Total number of MH-steps required by the continuous-time SMC algorithm per time interval with ICC threshold = 0.1 and with (grey bars) and without (magenta) η in the block updates.

This simple tailoring of the algorithm, only really necessary while a parameter is only weakly informed by the data and has an uninformative prior attached to it, evidently speeds up the algorithm, but it is necessary to show that this causes no degradation in terms of the inference that can be gathered. Table 7 repeats the exercise of Table 6, showing performance that is almost identical over the period 84-87 days (inclusive), the period over which there is a substantial speed up in the implementation of the algorithm. Thereafter, for the thresholds 0.2 and 0.1 the performance is comparable to MCMC, with the exception of what appears to be an anomalous reading for the 0.1 thresholds at 90 days.

Table 7 Performance of the tailored SMC algorithm over the interval 83-120 days by ICC threshold.

ICC threshold	0.5	0.2	0.1
84 Days (KL target = 0.732)	2.97	2.85	2.86
85 Days (KL target = 0.135)	3.06	2.97	2.98
86 Days (KL target = 0.365)	3.27	3.22	3.26

87 Days (KL target = 0.276)	2.51	2.48	2.44
90 Days (KL target = 0.159)	2.10	0.0927	1.42
100 Days (KL target = 0.135)	0.107	0.0835	0.0701
110 Days (KL target = 0.122)	0.197	0.0373	0.0348
120 Days (KL target = 0.119)	0.0999	0.0423	0.0551

But the failure of the algorithm to hit the MCMC thresholds over the interval 85-87 days (and 88, 89 days also, not shown) is a concern, and motivates an examination of scatterplots akin to those of Figure 5. In Figure 7 we look at the some scatterplots of the MCMC- and SMC-derived posterior distributions for two regression parameter components of the non-pandemic ILI consultations, $\boldsymbol{\beta}_B$. The plots headed MCMC are comparable to the SMC-derived plots immediately to their left. What they show is that, most strikingly on the 84 to 85 day and the 85 to 86 day steps, the MCMC algorithm isn't capable of the same coverage of the space that the SMC algorithm achieves. This is a result of poor convergence of the MCMC algorithm. It gets stuck within a smaller range of values. It maybe that the MCMC algorithm needs many more iterations to properly sample the full range of values, but it is already at a considerable computational disadvantage (typical runs are of chains of length 750,000 iterations).

There is, therefore, a strong suggestion that the MCMC analysis, far from being a gold standard, is actually inferior to the SMC analysis. Where Table 7 seemingly shows the supposed inability of the SMC to provide a posterior sample that could be considered representative of the MCMC sample, this may actually be due to the weakness of the MCMC algorithm, and SMC is preferred.

Returning to Figure 6, on the arrival of the data from day 84, each particle can be seen to require 220 iterations to accurately transition to a suitable posterior sample. When considered across 10^4 particles, this represents a number of evaluations of the full likelihood that are less than $4 \times$ what would be required by the 750,000 iterations of the MCMC algorithm typically used to derive a sample. As multiple chains are typically required to correctly diagnose convergence and to provide a sample, then it can be seen that the SMC would only require only very modest benefits from parallelisation to be quicker to compute. When placed on a computing cluster, the SMC is considerably quicker to implement. The SMC algorithm developed here was implemented on a cluster that, depending on availability, permitted simultaneous calculation on 100+ processors. The particles are distributed evenly

across the available processors, so that calculations on many particles are ongoing in parallel. Only at the resampling step and in the calculations of the ESS and the ICC is information shared across the processors. At day 83, we are considering the batch of data for which the greatest computational effort is required for the SMC analysis to derive inference. Elsewhere, the computational benefits of SMC are more clearly observed (for example, by the end of the 83 to 120 day interval, less than 10 iterations are required for rejuvenations). For batches of data that do not lead to a rejuvenation of the particle sample, the SMC updates require a negligible amount of time to compute and are evidently very much quicker than the MCMC analyses, that still have to run very long chains to produce a reasonable posterior sample.

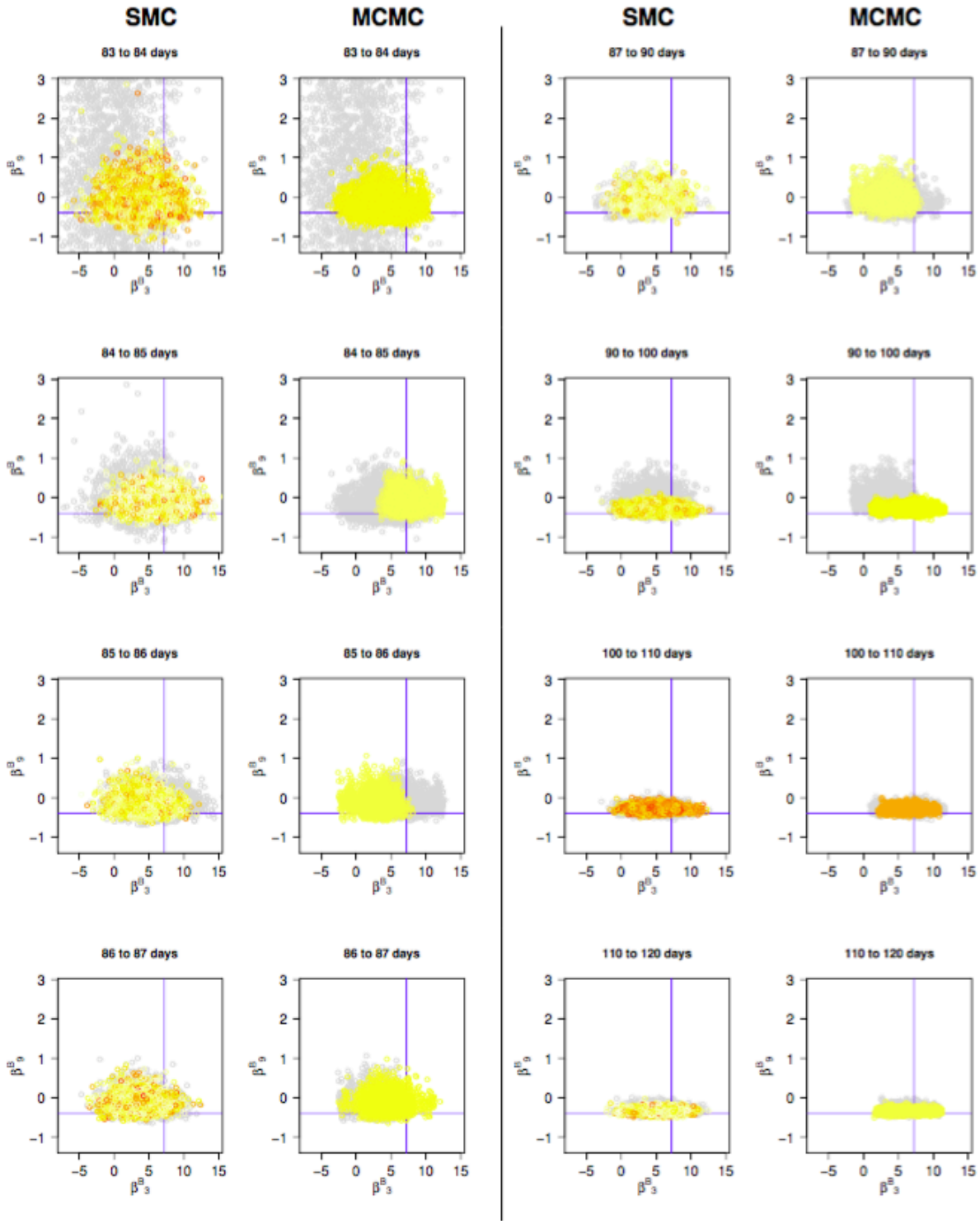

Figure 7: The evolution over time of the marginal joint posterior for two components of the parameter vector β^B, comparing between SMC-obtained and MCMC-obtained posterior distributions. Grey points indicate the distribution at the start of the interval.

5 Discussion

5.1 Achievements and objectives

The objective of this work was to advance the state of the art of real-time modelling of influenza epidemics and to provide a tool that could be used to monitor and predict the development of an ongoing pandemic outbreak.

We have advanced the state of the art by

1. Developing transmission models that account for spatial heterogeneity in the spread of infection;
2. Improving the efficiency with which estimation and prediction of an epidemic can be carried out. This has been achieved through the development of a Sequential Monte Carlo algorithm that will greatly reduce the computational burden for routine data analysis as part of a program of pandemic surveillance;
3. Facilitating the ability of the public health community to provide timely online inference to policymakers through provision of software to implement both of the above. The software has been adapted for the anticipated suite of epidemic data and key PHE scientists are engaged with ongoing training in its use. The computing code (and any related documentation) for the MCMC and SMC implementations of the real-time model are stored in open online repositories.[55,56]

In the initial proposal, there was also a component of this research promising support to the HPA (now PHE) in the event of a pandemic outbreak during the scope of this grant, in their real-time production of estimates and projections of the healthcare burden attributable to the pandemic. Such an outbreak did not occur and this component of the project has thus not yet been activated.

5.2 Strengths and Limitations

5.2.1 Spatial Modelling

This work has led to a coherent, unified, Bayesian statistical analysis of multiple streams of epidemic surveillance data from the 2009 A/H1N1pdm outbreak in England producing age and region stratified epidemic reconstructions (with associated uncertainty) and robust estimates of the parameters of the transmission process. We have explored two modelling approaches: the parallel-region (PR) and the meta-region (MR) models. Both fit adequately

well the various data sources, with highly comparable estimates for both model parameters and epidemic characteristics that are consistent with existing literature.

Each approach has its strengths and limitations.

The PR approach is found to be parsimonious yet sufficiently flexible to capture the underlying dynamics. It is also 'non-parametric', in the sense that the parameters representing the epidemic growth and initial seeding of infectiousness in each region are estimated without being subject to any parametric assumption. However, this lack of linkage means that the spread of infection between regions cannot be forecast and significant epidemic activity has to be observed in all regions to enable estimation of the epidemic burden. This lack of predictive ability is a limitation in the use of the PR approach. However, the greater flexibility becomes an advantage when it comes to epidemic reconstruction and this is observed in a significant improvement in the model fit of the PR approach to the 2009 pandemic data. An additional advantage is that this modelling approach does not rely on the validity of the commuter data to describe the spread of infection, nor does it rely on the assumption that individuals maintain routine commuting behaviour regardless of infection status. Despite the permitted spatial variation in epidemic growth rates, the PR model provides estimates for R_0^{init} that are consistent across regions. Therefore, the spatial heterogeneity in infection is being accounted for through the initial seeding of infectiousness.

The MR model incorporates spatial heterogeneity in transmission, arising from the interaction between regional populations, through commuting flows. This gives the MR model greater power to predict the spatial spread of influenza, enabling the prediction of which will be the next region to experience widespread infection. Early in a pandemic, therefore, the MR approach is more useful in a predictive modelling setting. However, it has been seen elsewhere that long-range interactions have a declining role in the spread of a pandemic once infection is widespread in each region.[57] This is exacerbated for A/H1N1pdm influenza as school-age children, the demographic group most affected, do not contribute to commuter movements. This marginalises, to some degree, the key benefit of this approach. Additionally, the MR approach involves an increased computational burden that limits its use as a tool for timely epidemic tracking as data accumulate over time.

One variant of the MR model investigated here involved the stratification of the population within each region into commuters and non-commuters. This has the effect of assuming each region contains a fixed sub-population of individuals who commute daily. This formulation yields no consistent improvement in model performance, whilst increasing even further the computational cost. Factoring in the 'random' movements of casual and

occasional travellers, which have been quoted to potentially increase the rate of transmission between regions by 25%,[58] would involve further computational burden and is particularly difficult to implement in an inferential setting without appropriate auxiliary information (e.g. if the census data contained information on the purpose of travel).

The MR model could be made more realistic and detailed by assuming that a proportion of those with symptomatic illness may not travel, or that asymptomatic illness is less infectious. However, consideration of such factors would only lessen the contribution of long-range transmission, leaving conclusions unchanged.

To summarise, using a Bayesian statistical framework, the PR model is found to be sufficiently flexible to provide a good fit to data and is quick to implement as it includes lower dimension contact matrices and, particularly, as the non-interacting nature of the regions means that the likelihood calculations can be easily parallelised. Reassuringly, it also provided concurring estimates for the basic reproductive number (R_0^{init}) across the regions, in agreement with the MR approach. However, the PR model can provide little insight on inter-region transmission and the determinants of spatial heterogeneity in the spread of infection because of its simple structure. In a situation where school-age children are the main agents of transmission and baseline transmissibility is not high, spatial models that concentrate on local transmission, like the PR model, provide a powerful and timely tool for use by public health services, helping to inform effective control and containment measures.

5.2.2 Efficient estimation and prediction

We have proposed addressing the substantive problem of real-time tracking of an emergent and realistic epidemic, assimilating multiple sources of information through the development of a suitable Sequential Monte Carlo (SMC) algorithm. When incoming data are stable, this process can be automated using standard algorithms in line with approaches already in the literature.[12,27] However, in the presence of interventions or any other event that may artificially interrupt the epidemic's trajectory or even result in a shock to the epidemic system, it is necessary to adapt the algorithm appropriately. How the algorithm is adapted will depend on the scale of the disruption to the surveillance data. The end result will be a semi-automated SMC algorithm that can be tailored to the nature of the shock to limit the required computation time.

This hybrid SMC can be seen to greatly outperform MCMC when it comes to successively iterating analyses, as will be required in a pandemic scenario. Throughout, we

have compared the divergence between SMC posteriors from posteriors generated by the "gold standard" MCMC. However, this may be an unfair comparison as the MCMC algorithm is based on "plain vanilla" Metropolis updates and could benefit from an in-depth tuning process itself. More sophisticated MCMC algorithms could be used, e.g. differential geometric MCMC or parallelised MCMC.[59, 60]. These could assist with improving MCMC run times. On the other hand, as MCMC steps are the main computational overhead of the SMC algorithm, any development of the MCMC algorithm may also lead to similar improvement of the SMC algorithm. It is also worth adding that the benefits of SMC for real-time analysis have been demonstrated. For a single, one-off, analysis aimed at reconstructing the epidemic dynamics, the SMC algorithm would be implemented differently and may not hold any significant advantage over MCMC.

Finally, the analyses carried out in this work have neglected the first fifty days of the epidemic, concentrating on a period when there is substantial transmission in the population and appropriate data are becoming available. As a result, a deterministic system can adequately describe the future evolution of the pandemic. Stochastic effects are significant and need to be incorporated into the model if monitoring is needed in the earlier stages. A prescription exists for what is known as "particle learning" in the presence of `shocks' in such a setting.[61] Alternatively, to improve the robustness of the inferences, the piecewise linear quantities describing population reporting behaviour could be described by linked stochastic noise processes. This has the potential to reduce the sensitivity of estimates to the presence of changepoints that are not, for whatever reason, foreseeable.

Over the course of this project, the state of the art of statistical computing in epidemic models has advanced in many directions, motivated by influenza and also recent Ebola outbreaks.[11,27,62-64] Each approach, however, uses direct observations of cases or estimates of cases to fit models. It is believed that our approach to tackling a realistic, messy, suite of epidemic data is both novel and critically important.

5.2.3 Pandemic Data

The capacity to provide real-time estimation and prediction of an epidemic is not only dependent on the existence of models and software. Crucial to this ability is the richness of the available public health surveillance data and its timely availability. The UK is well served in terms of the depth and completeness of its influenza surveillance mechanisms and the timely availability of data can almost be guaranteed wherever it arises as a result of routine collection and reporting. This is not quite the case for the serological information, however,

which requires suitable tests to be developed, and samples to be collected and analysed. The role of serological data is shown in Figure 10 of Birrell et al.,[47] where epidemic projections have been sequentially made using only noisy primary care consultation data in the absence of serology data. A reliable picture of the epidemic is not available until the epidemic is almost over. This poses some key questions: are serological samples going to be available in a timely manner, in sufficient quantity and quality, and in the right format? The availability of this data is a potentially limiting factor.

Routine influenza surveillance data may be reliable in terms of its timely provision; it may become unreliable as the demands placed upon healthcare services increase as infection becomes more widespread. Hospital beds and GP appointments are finite and the healthcare system may be operating at capacity for a period. Services such as NPFS are designed to alleviate this burden in primary care, but, in particular, we have to entertain the possibility that the proportion of cases that lead to hospitalisation might diminish. The model will permit time variation in $\boldsymbol{p}^h$ to account for these density-dependent effects, but how to diagnose and characterise this decline are still open problems. As will be discussed in Section 6.1.1, this motivates an exploration of the relationship between primary care ILI surveillance and community ILI surveillance. The aim is to find alternative data sources whose interpretation and collection is robust to high levels of influenza activity and can therefore constitute a valuable addition to the array of data already under consideration.

6 Conclusions

6.1 Research Recommendations

There are a number of very evident areas for future research and we summarise them briefly in what follows.

6.1.1 Alternative ILI surveillance

The majority of ILI presentations to health services will be via primary care, at least up to a point where a service such as the National Pandemic inFluenza Service (NPFS) is initiated. There are a number of potential additional surveillance systems that capture ILI occurring in the community, e.g. phone calls to the NHS 111 service, GP in-hours, GP out of hours, and RCGP spotter practices.[65] Once NPFS is activated, most ILI presentations will be diverted to this service, with integrated self-sampling providing data on virological positivity. The model currently works with GP ILI diagnoses reported via GP in-hours. There is the potential to use an amalgamation of the available surveillance data. If this were to be considered desirable then further research to understand the overlap between the systems would be required as well as how to measure virological positivity under each of the schemes.

As discussed in Section 5.2.3, the reliability of the data from traditional ILI surveillance schemes diminishes as resources become more stretched. This may motivate the potential use of alternative surveillance mechanisms that are not reliant upon the allocation of healthcare resources. These could be through internet searches,[66] social media data,[67] community reporting of symptoms,[44] or even sales of thermometers.[68] How to integrate these types of data into mechanistic models for disease transmission is still little understood and warrants further investigation.

6.1.2 Incorporating Interventions

To improve their utility as an epidemic response tool, the models presented here should be developed further to account for any interventions and mitigation strategies that may be employed during an influenza pandemic. Any interventions deployed and the strategies used will depend upon the nature and severity of the outbreak,[69] though they fall into two categories: pharmaceutical interventions (e.g. antivirals, antibiotics and vaccinations); and non-pharmaceutical interventions (e.g. school closures and hygiene campaigns).

The use of a pandemic specific vaccine would occur a few months after the start of the pandemic once a vaccine had been developed and manufactured. Vaccination has the

effect of changing the state of a substantial proportion of those vaccinated, the size of which would be dictated by the vaccine efficacy, from "susceptible" to "recovered", making the vaccinated individuals bypass other states within the model. Accounting for vaccination would then require a modification to the transmission model structure to enable the flow of vaccinated individuals directly between "susceptible" and "recovered" model states. This flow would be informed by data on vaccine coverage and vaccine efficacy.

A mitigation strategy that has been extensively studied is that of closures of schools. This strategy can be incorporated directly into the current model in manner similar to how the model currently handles school vacations i.e. by manipulating entries of the contact matrix corresponding to school age children. There are several studies that have estimated the changes in contact patterns during school vacations.[70] However, prolonged school closures may lead to wider differences in contact patterns that, to our knowledge, have not been rigorously studied. Further research may well be required to inform how the contact matrix would differ from that of school holiday periods.

Other non-pharmaceutical interventions can be used to slow the spread of infection and buy time for the development of biomedical interventions. These interventions may include hand hygiene campaigns and advice to avoid large gatherings, and they go hand-in-hand with other events that may influence the public's response to the pandemic, such as high-profile illnesses and sensationalist media reporting. These can impact upon transmission, but perhaps more significantly from a modelling perspective, they can influence how people interact with healthcare services and report their illness, affecting directly the data used in the modelling. Therefore, developing an understanding of these behavioural changes in real-time is vital to be able to estimate accurately the scale and spread of infection and the attached uncertainties.

6.1.3 Stochastic Model Adaptations

As commented in Section 5.2.2, the existing modelling approach only has utility once the epidemic is well established and infection is widespread. A consequence of this is that it is difficult to say anything about the early stages of the pandemic with any degree of certainty and parameters describing initial conditions of the dynamic transmission model have no real interpretation. Recent work by Shubin et al.[71] implements a similar model to Birrell et al.[15] to reconstruct the Finnish pandemic in 2009, with the exception that stochastic dynamics are used. Introducing similar stochastic dynamics in our spatial models would be of some interest, particularly in the MR modelling where the pattern of inter-region transmission

would be less rigidly defined by the commuting data. A further aspect that could be improved through the injection of stochasticity is the robustness of inferences. For example, parameters depending on the healthcare-seeking behaviour of populations could be replaced with stochastic noise processes. Such processes could take the form of the endemic/epidemic model,[72] for example, and would remove the sensitivity of the estimated epidemic trajectories derived by the pandemic model to the choice of change points discussed in Section 5.2.2. Furthermore, these processes are sufficiently flexible to incorporate any information on external factors that may influence the healthcare-seeking behaviour of a population.

6.1.4 Timely provision and understanding of serological data

The role of timely serological data has been demonstrated for the 2009 pandemic. As discussed in Section 6.2 below, it is vital that a system should be in place that can be exploited promptly in the event of a pandemic to give reliable information on susceptibility to the pandemic strain both at the beginning and throughout an epidemic.

It is possible that too much weight has been attributed to these data, however. Immunity is assumed to be present in all individuals with HI titre values above a given threshold. In reality, things may not be quite so well defined and it would be ideal if the uncertainty in determining the presence of immunity could be propagated into the analysis of the real-time model. This has been done in an analysis of Dutch A/H1N1pdm data, where protein micro arrays were used as a diagnostic assay to investigate antibody responses in cross-sectional serological samples.[73] This provided a probabilistic statement of whether the individual was susceptible, recently infected or held long-standing immunity. Being able to incorporate this type of information into the analysis would improve the handling of uncertainty in the model and give a clearer picture of initial levels of susceptibility.

6.1.5 Routine Operation

Each autumn/winter there is an annual influenza season over which rates of infection are heightened. Each season is characterised by a unique blend of circulating influenza viruses, leading to illness of varying severity. Typically, the well of immunity in the population to the circulating viruses prevents an escalation to pandemic levels. However, these seasonal infections exert a healthcare burden and models could inform the estimation and prediction of such burden. Routine operation of the real-time modelling system during these seasonal outbreaks ahead of a possible pandemic would provide invaluable insight into algorithmic performance, would allow PHE end users to familiarise themselves with the software and epidemiologists to become familiar with the model's data requirements. Both could then be

formally involved in the change process whenever the model requires adaptation in the light of any unforeseen operational hitches that may be identified and as the relevant surveillance schemes evolve.

6.2 Implications for healthcare

1. Continued funding to support the taking and storing of blood sera samples either via the existing telephone surveys, or through some other mechanism is essential. Without these data, primary care surveillance data are only weakly informative on the levels of underlying infection in the population. In many cases the peak of the epidemic has to be observed before it can be predicted with any confidence. These samples should not only be limited to being taken in-pandemic or post-pandemic. Storing samples from the beginning and the end of traditional flu seasons will provide a supply of blood sera samples to test (once an assay has been developed) that should be representative of the population's initial level of susceptibility to the pandemic influenza strain at the start of a pandemic, a vital ingredient of the transmission model. Furthermore, due to the time required from the start of an outbreak to the development of a suitable assay it is imperative that there are no further delays in the provision of these data to the real-time modelling effort, such as the academic need to publish the data.
2. It is essential that there is attendant virological surveillance accompanying any primary care data (such as GP consultations, or NPFS interactions), to separate out those with the pandemic infection from those with other respiratory viruses as well as the "worried well". On examining the USISS hospitalisation data, it appears that there are many confirmed influenza admissions with missing information on the virus sub-type. As the number of hospitalisations can often be relatively small, mis-labelling at this level could influence by orders of magnitude estimates for incidence. In the event of a pandemic, great effort should be taken to ensure that this does not occur.
3. It is trivial to adjust the transmission model to accommodate the effects of vaccination and so the real-time model can provide a tool to assess the impact of any vaccination strategies that the government may be considering. Other pandemic interventions can be prospectively assessed before their implementation using inferences drawn up to the current time and some reasonable assumptions about the impact of the intervention. However, accommodating data that may arise as a result of a pandemic

intervention (such as data on vaccine uptake) into the real-time modelling process is not so straightforward and this needs to be given some serious future thought.

4. Retention of statistical modelling expertise is essential within the responsible public health body (currently PHE). The statistical procedures described in this report are described as semi-automated, in that their performance can be improved greatly with user input, particularly in the presence of unforeseen epidemic events or "shocks" in surveillance data. Furthermore, a great deal of statistical literacy is required to be able to specify the model to adapt to the unique characteristics of an outbreak of a novel influenza virus and the relevant surveillance data that could become available in future.

Acknowledgements

This work was supported by the National Institute for Health Research (HTA Project:11/46/03) the UK Medical Research Council (Unit Programme Number U105260566) and Public Health England. The authors thank the University of Nottingham, Egton Medical Information Systems (EMIS), and EMIS practices contributing to the QSurveillance database. We thank colleagues at PHE Respiratory Virus Reference Unit and the Specialist Microbiology Network for the provision of GP swab positivity data and for the use of their 'whiteboard' confirmed case data. We also extend thanks to patients of Royal College of General Practitioners Research and Surveillance Centre (RSC) practices who consented to having a flu swab taken and RSC practices for processing and sharing these data. The authors would also like to thank: Lorenz Wernisch, Brian Tom and Gareth Roberts for their insight in the development of Bayesian sequential methods; members of the project's oversight group, and in particular Ian Hall, Iain Barrass and Anne Presanis; and the members of the Study Steering Committee for their time and interest: Chris Robertson (chair), Peter Diggle, Paddy Farrington, and Philip O'Neill.

Contribution of Authors

Paul Birrell (Senior Investigator Statistician, Medical Research Council Biostatistics Unit) designed and conducted the analyses, reviewed literature and prepared the report for publication.

Richard Pebody (Consultant Epidemiologist, Public Health England) provided data and guidance in its usage, developed pandemic data and preparedness protocols.

André Charlett (Head of Statistical Modelling and Economics, Public Health England) provided supervision, developed research recommendations and developed pandemic protocols.

Xu-Sheng Zhang (Senior Modeller, Public Health England) provided modelling assistance.

Daniela De Angelis (Programme Leader, Medical Research Council, Biostatistics Unit) submitted the grant application, directed the analysis and contributed to drafting the report.

Publications

The following publications are associated to the work presented in this report:

1. Birrell PJ, Zhang X-S, Pebody RG, Gay NJ, De Angelis D. Reconstructing a spatially heterogeneous epidemic: Characterising the geographic spread of 2009 A/H1N1pdm infection in England. Sci. Rep. 2016;6. Available from: http://dx.doi.org/10.1038/srep29004.
2. Birrell PJ, De Angelis D, Wernisch L, Tom BDM, Roberts GO, Pebody RG. Efficient real-time monitoring of an emerging influenza epidemic: how feasible? arXiv. 2016;1608.05292v1:36. Available from: http://arxiv.org/abs/1608.05292.

Other Outputs: Verbal Presentations

1. Birrell PJ, Zhang X-S, Pebody RG, Gay NJ, De Angelis D. Real-Time Modelling of a Pandemic Influenza Outbreak. Department for Infectious Disease Epidemiology, Imperial College London, Date: 11/11/2013 http://www1.imperial.ac.uk/publichealth/departments/ide/events/
2. Birrell PJ, Zhang X-S, Pebody RG, Gay NJ, Wernisch L. Towards Real-Time Modelling of a Pandemic Influenza Outbreak. MRC-BSU Centenary Conference, Cambridge UK, Date: 25/03/2014.
3. De Angelis D. Bayesian Inference in Infectious Disease Models: Current Challenges. Bayesian Biostatistics Conference, Zurich. Date: 15/10/2014. http://www.biostat.uzh.ch/bb2014_en.html
4. De Angelis D, Birrell PJ, Current Challenges in Inference for Infectious Disease Models Using Data from Multiple Sources. Mathematical Challenges for Long Epidemic Time Series, University of Warwick. Date: 17/12/2014 http://www2.warwick.ac.uk/fac/sci/maths/research/events/2014-15/nonsymposium/mclets/

Other Outputs: Poster Presentations

1. Birrell PJ, Wernisch L, Roberts GO, Pebody RG, De Angelis D. Efficient real-time statistical modelling for pandemic influenza. Epidemics^4, Amsterdam, Date: 20/11/2013.
2. Birrell PJ, Zhang X-S, Pebody RG, Gay NJ, De Angelis D. Spatial modelling of 2009 pandemic flu in England: characterisation of the impact of migration and geographic variation on the spatial spread of infection. Epidemics4 Conference, Amsterdam, 19/11/2013.

3. Birrell PJ, De Angelis D, Wernisch L, Roberts GO, Pebody RG. Efficient real-time statistical modelling for pandemic influenza. MCMSki IV Conference, Chamonix, France, Date: 05/01/2014. http://www.pages.drexel.edu/~mwl25/mcmski/index.html

Data Sharing

The data used in this modeling study is not data that was generated as a direct result of this project. Instead it is routinely collected pandemic influenza surveillance data, data that have, in most cases, been collected by third-party database providers and licenced to Public Health England. The sharing and usage of these data are, therefore, subject to some contractual limitations. Requests for access to any non-synthetic data should be directed to Richard Pebody at Public Health England (Richard.Pebody@phe.gov.uk).

[*18834 words*]

References

1. Cabinet Office. National Risk Register of Civil Emergencies; 2015. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/419549/20150331_2015-NRR-WA_Final.pdf (*Accessed 6/1/2017*).

2. Ferguson NM, Cummings DAT, Fraser C, Cajka JC, Cooley PC, Burke DS. Strategies for mitigating an influenza pandemic. Nature. 2006 Jul;442(7101):448–452.

3. Germann TC, Kadau K, Longini IM, Macken CA. Mitigation strategies for pandemic influenza. Proc Natn Acad Sci USA. 2006 April;103(15):5935–5940.

4. Hall IM, Gani R, Hughes HE, Leach S. Real-time epidemic forecasting for pandemic influenza. Epidemiology and Infection. 2007;135:372–385.

5. Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. American Journal of Epidemiology. 2004 September;160(6):509–516.

6. Cauchemez S, Boëlle PY, Thomas G, Valleron AJ. Estimating in Real Time the Efficacy of Measures to Control Emerging Communicable Diseases. American Journal of Epidemiology. 2006;164(6):591–597.

7. Cauchemez S, Boëlle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, et al. Real-time estimates in early detection of SARS. Emerging infectious diseases. 2006;12(1):110–113.

8. Jewell CP, Kypraios T, Christley RM, Roberts GO. A novel approach to real-time risk prediction for emerging infectious diseases: a case study in Avian Influenza H5N1. Preventive veterinary medicine. 2009 September;91(1):19–28.

9. Bettencourt LMA, Ribeiro RM. Real Time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases. PLoS ONE. 2008 3(5):e2185.

10. Chowell G, Simonsen L, Towers S, Miller MA, Viboud C. Transmission potential of influenza A/H7N9, February to May 2013, China. BMC Medicine. 2013 Oct;11:214+.

11. Dureau J, Kalogeropoulos K, Baguelin M. Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. Biostatistics. 2013 Jul;14(3):541–555.

12. Ong JBS, Chen MIC, Cook AR, Chyi H, Lee VJ, Pin RT, et al. Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. PloS one. 2010;5(4):e10036.

13. Ball J. Lessons Learnt; 2010. SPI-M-O Committee document. Available from: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@ab/documents/digitalasset/dh_118907.pdf (*Accessed 6/1/2017*).
14. Baguelin M, Van Hoek AJ, Flasche S, White PJ, Edmunds WJ. Vaccination against pandemic influenza A/H1N1v in England: A real-time economic evaluation. Vaccine. 2010 March;28(12):2370–2384.
15. Birrell PJ, Ketsetzis G, Gay NG, Cooper BS, Presanis AM, Harris RJ, et al. Bayesian modelling to unmask and predict the influenza A/H1N1pdm dynamics in London. Proc Natn Acad Sci USA. 2011 November;108(45):18238–18243.
16. Evans B, Charlett A, Powers C, McLean E, Zhao H, Bermingham A, et al. Has estimation of numbers of cases of pandemic influenza H1N1 in England in 2009 provided a useful measure of the occurrence of disease? Influenza and Other Respiratory Viruses. 2011 Nov 1;5(6):e504-12.
17. Wu JT, Cowling BJ, Lau EHY, Ip DKM, Ho LM, Tsang T, et al. School Closure and Mitigation of Pandemic (H1N1) 2009, Hong Kong. Emerging Infectious Diseases. 2010;16(3):538–541.
18. Hippisley-Cox J, Smith S, Smith G, Porter A, Heaps M, Holland R, et al. QFLU: new influenza monitoring in UK primary care to support pandemic influenza planning. Euro surveillance. 2006;11(6).
19. Miller E, Hoschler K, Hardelid P, Stanford E, Andrews N, Zambon M. Incidence of 2009 pandemic influenza A H1N1 infection in England: a cross-sectional serological study. Lancet. 2010 March;375(9720):1100–1108.
20. Hine D. The 2009 Influenza Pandemic: An independent review of the UK response to the 2009 influenza pandemic. 2010. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/61252/the2009influenzapandemic-review.pdf *(Accessed 6/1/2017*).
21. Gamerman D, Lopes HF. Markov Chain Monte Carlo - Stochastic simulation for Bayesian inference. 2nd ed. Chapman and Hall. London; 2006.
22. Farah M, Birrell PJ, Conti S, De Angelis D. Bayesian Emulation and Calibration of a Dynamic Epidemic Model for A/H1N1 Influenza. J Am Statist Ass. 2014;109(508):1398–1411.

23. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of The Royal Society Interface. 2009 February;6(31):187–202.
24. Chopin N. A sequential particle filter method for static models. Biometrika. 2002 Aug;89(3):539–552.
25. Fearnhead P. MCMC, sufficient statistics and particle filters. Journal of Computational and Graphical Statistics. 2002 December;11(4):848–862.
26. Del Moral P, Doucet A, Jasra A. Sequential Monte Carlo samplers. J R Statist Soc B. 2006 Jun;68(3):411–436.
27. Dukic V, Lopes HF, Polson NG. Tracking Epidemics with Google Flu Trends Data and a State-Space SEIR Model. J Am Statist Ass. 2012;107(500):1410–1426.
28. Birrell PJ, Zhang X-S, Pebody RG, Gay NJ, De Angelis D. Reconstructing a spatially heterogeneous epidemic: Characterising the geographic spread of 2009 A/H1N1pdm infection in England. Sci. Rep. 2016;6.
29. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Disease. PLoS Medicine. 2008 March;5(3):e74.
30. Fleming DM. Weekly Returns Service of the Royal College of General Practitioners. Communicable disease and public health / PHLS. 1999 June;2(2):96–100.
31. Harcourt SE, Smith GE, Elliot AJ, Pebody R, Charlett A, Ibbotson S, et al. 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. 2012;140:100–105.
32. Health Protection Agency. Sources of UK flu data: Influenza Surveillance in the United Kingdom. Available from: http://webarchive.nationalarchives.gov.uk/20140626151846/http://hpa.org.uk/Topics/InfectiousDiseases/InfectionsAZ/SeasonalInfluenza/EpidemiologicalData/30influsSourcesofUKfludata/ (*Accessed 6/1/2017*).
33. McCartney C. Regional microbiology network. British Journal of Infection Control. 2008 January;8(1):28–29.
34. Health Protection Agency. Epidemiological report of pandemic (H1N1) 2009 in the UK; 2010. online. Available from: http://webarchive.nationalarchives.gov.uk/20140714084352/http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1284475321350 (*Accessed 6/1/2017*).

35. The Phoenix Partnership (TPP). Real-time Syndromic Surveillance; 2013. Available from: http://www.researchone.org/public-health-monitoring/ (*Accessed 6/1/2017*).
36. Public Health England. Sources of UK flu data: influenza surveillance in the UK; 2014. Available from: https://www.gov.uk/guidance/sources-of-uk-flu-data-influenza-surveillance-in-the-uk (*Accessed 6/1/2017*).
37. Osborne K, Gay N, Hesketh L, Morgan-Capner P, Miller E. Ten years of serological surveillance in England and Wales: methods, results, implications and action. International Journal of Epidemiology. 2000 April;29(2):362–368.
38. Hardelid P, Andrews NJ, Hoschler K, Stanford E, Baguelin M, Waight PA, et al. Assessment of baseline age-specific antibody prevalence and incidence of infection to novel influenza AH1N1 2009. Health Technology Assessment. 2010 December;14(55):115–192.
39. de Jong JC, Palache AM, Beyer WE, Rimmelzwaan GF, Boon AC, Osterhaus AD. Haemagglutination-inhibiting antibody to influenza virus. Developments in Biologicals. 2003;115:63–73.
40. Hobson D, Curry RL, Beare AS, Ward-Gardner A. The role of serum haemagglutination-inhibiting antibody in protection against challenge infection with influenza A2 and B viruses. The Journal of Hygiene. 1972 December;70(4):767–777.
41. Al-Khayatt R, Jennings R, Potter CW. Interpretation of responses and protective levels of antibody against attenuated influenza A viruses using single radial haemolysis. The Journal of Hygiene. 1984 October;93(2):301–312.
42. Office for National Statistics. 2001 Census: Special Workplace Statistics (England, Wales and Northern Ireland); 2009. . UK Data Service Census Support. Downloaded from: http://www.cids.census.ac.uk.
43. Office for National Statistics. Super Output Area mid-year population estimates for England and Wales (experimental); 2008. Available from: https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/middlesuperoutputareamidyearpopulationestimates/mid2002tomid2010/sape8dtmsoasyoaunformattedmid2002tomid2010.zip (*Accessed 6/1/2017*).
44. FluSurvey. Available from: http://www.flusurvey.org.uk (*Accessed 6/1/2017*).
45. Gordon NJ, Salmond DJ, Smith AFM. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. Radar and Signal Processing, IEE Proceedings F. 1993 Apr;140(2):107–113.
46. Gilks WR, Berzuini C. Following a moving target - Monte Carlo inference for dynamic Bayesian models. J R Statist Soc B. 2001;63(1):127–146.

47. Birrell PJ, De Angelis D, Wernisch L, Tom BDM, Roberts GO, Pebody RG. Efficient real-time monitoring of an emerging influenza epidemic: how feasible? arXiv. 2016;1608.05292v1:36.
48. Donner A, Koval JJ. The estimation of intraclass correlation in the analysis of family data. Biometrics. 1980 Mar;36(1):19–25.
49. Liu JS, Chen R. Blind Deconvolution via Sequential Imputations. J Am Statist Ass. 1995;90(430):567–576.
50. Boëlle PYY, Ansart S, Cori A, Valleron AJJ. Transmission parameters of the A/H1N1(2009) influenza virus pandemic: a review. Influenza Other Respir Viruses. 2011 Sep;5(5):306–316.
51. Dorigatti I, Cauchemez S, Ferguson NM. Increased transmissibility explains the third wave of infection by the 2009 H1N1 pandemic virus in England. Proc. Natl. Acad. Sci. USA 2013 Aug;110(33):13422–13427.
52. He D, Dushoff J, Eftimie R, Earn DJ. Patterns of spread of influenza A in Canada. Proceedings of the Royal Society of London B: Biological Sciences. 2013 Nov;280(1770).
53. Khaokham CB, Selent M, Loustalot FV, Zarecki SM, Harrington D, Hoke E, et al. Seroepidemiologic investigation of an outbreak of pandemic influenza A H1N1 2009 aboard a US Navy Vessel San Diego, 2009. Influenza Other Respir. Viruses. 2013 Mar. 7(5). Available from: http://dx.doi.org/ 10.1111/irv.12100.
54. MacKay DJ. Information theory, inference and learning algorithms. Cambridge university press; 2003.
55. Birrell PJ. Real-time model: MCMC code; 2016. Available from: https://gitlab.com/pjbirrell/real-time-mcmc (*Accessed 6/1/2017*).
56. Birrell PJ. Real-time model: SMC code; 2016. Available from: https://gitlab.com/pjbirrell/real-time-smc (*Accessed 6/1/2017*).
57. Gog JR, Ballesteros S, Viboud C, Simonsen L, Bjornstad ON, Shaman J, et al. Spatial Transmission of 2009 Pandemic Influenza in the US. PLoS Comput Biol. 2014 Jun;10(6):e1003635+.
58. Danon L, House T, Keeling MJ. The role of routine versus random movements on the spread of disease in Great Britain. Epidemics. 2009 Dec;1(4):250–258.
59. Girolami M, Calderhead B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011 Mar;73(2):123–214.

60. Banterle M, Grazian C, Lee A, Robert CP. Accelerating Metropolis-Hastings algorithms by Delayed Acceptance. arXiv. 2015;1503.00996v2(1503.00996v2):27.
61. Nemeth C, Fearnhead P, Mihaylova L. Sequential Monte Carlo Methods for State and Parameter Estimation in Abruptly Changing Environments. IEEE Transactions on Signal Processing. 2014 Mar;62(5):1245–1255.
62. Skvortsov A, Ristic B. Monitoring and prediction of an epidemic outbreak using syndromic observations. Mathematical Biosciences. 2012 Nov;240(1):12–19.
63. Yang W, Lipsitch M, Shaman J. Inference of seasonal and pandemic influenza transmission dynamics. Proceedings of the National Academy of Sciences. 2015 Mar;112(9):2723–2728.
64. Kucharski AJ, Camacho A, Flasche S, Glover RE, Edmunds WJ, Funk S. Measuring the impact of Ebola control measures in Sierra Leone. Proceedings of the National Academy of Sciences. 2015 Nov;112(46):14366–14371.
65. Public Health England. Syndromic surveillance: systems and analyses; 2015. Available from: https://www.gov.uk/government/collections/syndromic-surveillance-systems-and-analyses (*Accesssed 6/1/2017*).
66. Zhang Y, Arab A, Cowling BJ, Stoto MA. Characterizing Influenza surveillance systems performance: application of a Bayesian hierarchical statistical model to Hong Kong surveillance data. BMC public health. 2014 Aug;14:850+.
67. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PloS one. 2013;8(12).
68. Todd S, Diggle PJ, White PJ, Fearne A, Read JM. The spatiotemporal association of non-prescription retail sales with cases during the 2009 influenza pandemic in Great Britain. BMJ Open. 2014 Apr;4(4):e004869+.
69. Department of Health. UK Influenza Pandemic Preparedness Strategy 2011; 2011. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213717/dh_131040.pdf (*Accessed 6/1/2017*).
70. Eames KT, Tilston NL, White PJ, Adams E, Edmunds WJ. The impact of illness and the impact of school closure on social contact patterns. Health Technology Assessment. 2010 Jul;14(34):267–312.
71. Shubin M, Lebedev A, Lyytikäinen O, Auranen K. Revealing the True Incidence of Pandemic A(H1N1)pdm09 Influenza in Finland during the First Two Seasons – An Analysis Based on a Dynamic Transmission Model. PLoS Comput Biol. 2016;12(3):1–3.

72. Held L, Hofmann M, Höhle M, Schmid V. A two-component model for counts of infectious diseases. Biostatistics. 2006;7(3):422–437.
73. te Beest DE, Birrell PJ, Wallinga J, De Angelis D, van Boven M. Joint modelling of serological and hospitalization data reveals that high levels of pre-existing immunity and school holidays shaped the influenza A pandemic of 2009 in the Netherlands. Journal of the Royal Society, Interface. 2015 Feb;12(103):20141244+.

Appendices

Appendix 1: Single-region model dynamics

Assume that a population can be split into A strata (in Birrell et al.[15] these are defined by age groups), denoted by $a = 1, \ldots, A$. The E and I states of the SEIR-model depicted in Figure 1 are split into two sub-states, E_1, E_2 and I_1, I_2, to ensure that both the latent and incubation periods are gamma distributed, rather than exponentially distributed, where the modal time is zero. The infection status of the population within each stratum a at discrete times $t_k = k\delta t$, for $k = 0, \ldots, K$ and for appropriately small δt, and $k \geq 1$ is approximated by the deterministic system of difference equations:

$$\begin{aligned} S_a(t_k) &= S_a(t_{k-1})(1 - \lambda_a(t_{k-1})\delta t), \\ E_{1,a}(t_k) &= E_{1,a}(t_{k-1})(1 - \sigma\delta t) + S_a(t_{k-1})\lambda_a(t_{k-1})\delta t, \\ E_{2,a}(t_k) &= E_{2,a}(t_{k-1})(1 - \sigma\delta t) + E_{1,a}(t_{k-1})\sigma\delta t, \\ I_{1,a}(t_k) &= I_{1,a}(t_{k-1})(1 - \gamma\delta t) + E_{2,a}(t_{k-1})\sigma\delta t, \\ I_{2,a}(t_k) &= I_{2,a}(t_{k-1})(1 - \gamma\delta t) + I_{1,a}(t_{k-1})\gamma\delta t. \end{aligned} \tag{8}$$

Parameters σ and γ are related to the mean durations of latent and infectious infection, d_L and d_I respectively, via $\sigma = 2/\mathrm{d_L}, \gamma = 2/\mathrm{d_I}$, and the force of infection, $\lambda_a(t_k)$ is expressed through the Reed-Frost formulation

$$\lambda_a(t_k) = 1 - \prod_{b=1}^{A} \left[(1 - \beta_{a,b}(t_k))^{I_{1,b}(t_k) + I_{2,b}(t_k)}\right] \tag{9}$$

where the quantity $\beta_{a,b}(t_k)$ is the $(a,b)^{\text{th}}$ entry of the $(A \times A)$ matrix $\boldsymbol{\beta}(t_k)$ and gives the infectious pressure exerted on a susceptible individual in stratum a by a single infectious individual in stratum b. This relates to the epidemic's reproductive number, R_0 via the relation

$$\boldsymbol{\beta}(t_k) = \boldsymbol{M}(t_k)\, {}^{R_0}/_{R_0^*} \tag{10}$$

where $\boldsymbol{M}(t_k) = \{M_{a,b}(t_k)\}$ are matrices of relative infective contact rates between individual of age groups a and b derived from POLYMOD data[29] and contact parameters, $m_j, j = 1, \ldots, 5$ that modify these contact rates to allow for increased transmissibility in contacts involving children and the effects of school closures (as described elsewhere[15]). R_0^* denotes the dominant eigenvalue of the time-0 next generation matrix $\boldsymbol{M}^*$ which has $(a,b)^{\text{th}}$ entry

given by $D_a \times M_{a,b}(t_0) \times d_1$, where D_a is the resident population size of people in age group a.

The number of new infections at each time is then given by

$$\Delta_a^{(\text{infec})}(t_k) = S_a(t_{k-1})\lambda_a(t_{k-1})\delta t. \tag{11}$$

New symptomatic infections (however symptoms are being defined) are then given by $\Delta_a(t_k) = \phi\Delta_a^{(\text{infec})}(t_k)$, where ϕ is the proportion of infections that are symptomatic (see Figure 1).

In all of what follows, δt=0.5 days, a duration sufficiently small relative to the expected waiting times in each of the model states. With this choice, the dynamics mapped out by the difference equations in (8) are a close match to the differential equation system that they are designed to replicate.

Parallel-region (PR) model

The equations in the PR model are exactly the same, except that we repeat the system of Equations in (8) for each region, and denote the region by $r = 1, \ldots, R$:

$$\begin{aligned}
S_{a,r}(t_k) &= S_{a,r}(t_{k-1})\big(1 - \lambda_{a,r}(t_{k-1})\delta t\big), \\
E_{1,a,r}(t_k) &= E_{1,a,r}(t_{k-1})(1 - \sigma\delta t) + S_{a,r}(t_{k-1})\lambda_{a,r}(t_{k-1})\delta t, \\
E_{2,a,r}(t_k) &= E_{2,a,r}(t_{k-1})(1 - \sigma\delta t) + E_{1,a,r}(t_{k-1})\sigma\delta t, \\
I_{1,a,r}(t_k) &= I_{1,a,r}(t_{k-1})(1 - \gamma\delta t) + E_{2,a,r}(t_{k-1})\sigma\delta t, \\
I_{2,a,r}(t_k) &= I_{2,a,r}(t_{k-1})(1 - \gamma\delta t) + I_{1,a,r}(t_{k-1})\gamma\delta t.
\end{aligned} \tag{12}$$

with Equations (9) and (10) being similarly adapted:

$$\lambda_{a,r}(t_k) = 1 - \prod_{b=1}^{A}\left[\left(1 - \beta_{a,b,r}(t_k)\right)^{I_{1,b,r}(t_k)+I_{2,b,r}(t_k)}\right],$$

$$\boldsymbol{\beta}_r(t_k) = \boldsymbol{M}(t_k)^{R_{0,\mathrm{r}}}\Big/_{R^*_{0,r}},$$

where, again the quantity $\beta_{a,b,r}(t_k)$ is the $(a,b)^{\text{th}}$ entry of the $(A \times A)$ matrix $\boldsymbol{\beta}_r(t_k)$.

Meta-region (MR) model

Superficially, the system of equations governing the SEIR dynamics is the same as presented in the equation block (8). The subtle difference is that the stratum indicator a now takes values in the range $1, \ldots, RA$. Similarly to Equation (9), the force of infection is given by

$$\lambda_a(t_k) = 1 - \prod_{b=1}^{RA}\left[\left(1-\beta_{a,b}(t_k)\right)^{I_{1,b}(t_k)+I_{2,b}(t_k)}\right].$$

Each of the RA strata corresponds to a unique pairing of region r and age group a', (r, a'). Therefore, a stratum number can be assigned by setting $a = a' + A(r-1)$. Considering two such strata a and b, the infection rate matrix has entries

$$\beta_{a,b}(t_k) = \Pi_{a,b}(t_k) \times {}^{R_0}/_{R_0^*},$$

where $\Pi_{a,b}(t_k)$ are entries of the meta-region contact matrix $\mathbf{\Pi}(t_k)$. The $(RA \times RA)$ matrices, $\mathbf{\Pi}(t_k)$ have a necessarily different structure to the contact matrices used in the single-region model or the PR model. As discussed in Section 3.2.2, the entries of these matrices that correspond to contacts between individuals resident in different regions have been informed on the basis of commuter data from the ONS' 2001 UK census data collection. These data come in the form of matrices, $\boldsymbol{C}(a'), a' = 1, \ldots, A$ that have (r, s) entry $C_{rs}(a')$, for $r, s = 1, \ldots, R$. These matrix entries represent the proportion of age group a who are resident in region r that commute into region s on any given day (see Table 8). However, it is assumed that individuals younger than 16 years old do not commute, leading to the white sections in Figure 2.

Table 8: Commuter matrices. For each of the four adult age groups, cells give the proportion of individuals resident in each region (by row), who move to (or stay in) each of the four regions (columns). Reproduced from Birrell et al.[28]

Matrix $C_{rs}(a)$	London	WM	North	South	London	WM	North	South
	Ages: 16-24 years				Ages: 25-44 years			
London	93.50%	0.11%	0.30%	6.12%	92.10%	0.11%	0.26%	7.54%
WM	0.43%	95.60%	2.29%	1.64%	0.52%	94.80%	2.95%	1.71%
North	0.36%	0.77%	97.30%	1.57%	0.46%	0.99%	96.90%	1.67%
South	5.09%	0.22%	0.45%	94.20%	9.29%	0.30%	0.53%	89.90%
	Ages: 45-64 years				Ages: 65-74 years			
London	93.60%	0.10%	0.24%	6.10%	85%	0.11%	0.23%	14.60%
WM	0.37%	96%	2.32%	1.31%	0.35%	97.40%	1.41%	0.85%
North	0.35%	0.83%	97.80%	1.06%	0.34%	0.47%	98.50%	0.65%
South	6.87%	0.27%	0.46%	92.40%	4%	0.19%	0.33%	95.50%

Commuter movements are assumed to cover a fraction of the total time equal to ξ, and inter-region transmission is only possible in this proportion of time. Assuming, then, that commuting movements all take place at the same time, an individual belonging to a strata a that corresponds to the region/age-group pairing of (r, a') will be in the same region as an

individual from strata b (corresponding to (s, b')), with probability $\sum_{v=1}^{R} C_{r,v}(a')C_{s,v}(b')$. To get a probability of contact between the two individuals, we might need to scale the terms in the sum according to the density dependence assumptions. It is assumed that once in a region, individuals will associate with members of other age groups according to the POLYMOD-based matrices used in the single-region and PR models. Therefore, the meta-region contact matrices $\mathbf{\Pi}(t_k)$ is constructed to have entries

$$\Pi_{a,b}(t_k) = M_{a',b'}(t_k)\left\{\frac{(1-\xi)\delta_{rs}}{\left(D_{r,a'}^{N}\right)^{\alpha}} + \xi\sum_{v=1}^{R}\frac{C_{rv}(a')C_{sv}(b')}{\left(D_{v,a'}^{D}\right)^{\alpha}}\right\}. \quad (13)$$

The $D_{r,a'}^{N} = D_{r,a'}$ are the population sizes for the (r, a') stratum at night (i.e. the size of the resident population) and $D_{r,a'}^{D}$ are the daytime population sizes, the adjusted population sizes after commuter movements have taken place, $D_{r,a'}^{\mathrm{D}} = \sum_{s=1}^{R} C_{sr}(a')D_{s,a'}$. The proportion of total time that a commuter actually spends in the commuting region is taken to be $\xi = 5/14$ on the basis of a daily average of five working days per week, being away from home for a half day when working. The exponent α takes values in [0,1] with a value of 0 indicating frequency-driven transmission and 1, density-driven transmission; δ_{rs} is merely a Kronecker-delta function. It is furthermore important to note that under this model, the R_0^* is calculated on the basis of the next-generation matrix $\mathbf{\Pi}^*$, with entries $\Pi_{a,b}^{*} = D_{r,a'}\Pi_{a,b}(t_0)d_I$, with the various strata being as defined above. It is easy to see how the block-diagonal structure of these matrices arrives, as within these blocks, $r = s$ and the left-hand term within the bracketed sum contributes to the contact rates.

The denominators of Equation (13) give the prevalent degree of density dependence. There are two things to consider here:

- The value of α. A value of $\alpha = 0$ corresponds to frequency dependent transmission (two people are equally like to interact regardless of the population size of their strata), whereas $\alpha = 1$ corresponds to a density dependent effect upon transmission. Values of $\alpha = 0, 0.5$, and 1 will be considered.
- How density dependence should be incorporated. Equation (13) uses dependence upon the population size of the strata of the uninfected individual in an infectious contact. However, it seems more logical to use the population size of the region to which the infectious individual belongs. This is more comparable to the PR model that assumes no density dependent effects across the age groups. Therefore, as an alternative, it is proposed to replace $D_{r,a}^{N}$ with the region-wide population size $D_{v}^{N} =$

$\sum_{a=1}^{A} D_{v,a}^{N}$ and $D_{r,a}^{D}$ with $D_{v}^{D} = \sum_{a=1}^{A} D_{v,a}^{D}$. This tests whether it is the population size of a region that is important in density dependence, as opposed to the age-group constitution of that population.

Another modelling feature to be tested through the specification of the contact matrix in the MR model is the random commuter model formulation vs. the fixed commuter formulation. However the strata are defined, individuals within each stratum are assumed to be homogeneous of behavior. That is, in any given interval, any individual within a specific stratum is equally likely to be one who commutes. This may well prove to be a gross simplification because:

- The bulk of commuting is done by a much smaller sub-population within a stratum, each of who commutes on a regular basis.
- Infectious individuals are still assumed to carry on with normal commuting behaviour, whereas in reality, they may well be too ill to travel.

The first simplification we term to be an assumption of 'commuting at random'. To consider the possibility that there is a fixed sub-population within each stratum that commute regularly, we further stratify the adult age groups into 'commuters' and 'non-commuters'. Commuters are assumed not to stay in their home region during working hours, whereas non-commuters stay in the region in which they are resident with certainty. So, rather than RA strata (in the 2009 A/H1N1pdm example, this is $7 \times 4 = 28$), we partition further. Four of the seven age groups are of adult-age individuals and can be further divided into commuter and non-commuter classes. Therefore we now consider $(A + 4) \times R = 44$ strata. In this example, the expanded matrix has an identical mathematical expression as before (Equation (13)), once the $C_{rs}(a)$ are replaced by $C'_{rs}(a)$, where

$$C'_{rr}(a) = \begin{cases} 1, & a \text{ belongs to a non-commuter class} \\ 0, & a \text{ belongs to a commuter class} \end{cases}$$

$$C'_{rs}(a) = \begin{cases} 0 & , \quad a \text{ belongs to a non-commuter class} \\ \dfrac{C_{rs}(a)}{1 - C_{rr}(a)}, & a \text{ belongs to a commuter class.} \end{cases}$$

The assumption of commuting at random speeds up the spread of infection across the A regions while the fixed commuting assumption increases the transiency of any commuting effects and results in greater heterogeneity in the times of peak infection across the regions. However, simulations have shown that the peak size and attack rate are insensitive to the commuting assumption.

Appendix 2: Single-region model dynamics

Below is an itemised list of model parameters, giving for each a symbol, a short description of the role of the parameter and the types of heterogeneity in the value of the parameter that the real-time model software can accommodate.

η The over-dispersion inherent in the GP consultation counts. Only relevant if the selected likelihood is negative-binomial and not Poisson. **Region-, time- and strata-specific variation permitted.**

d_L The expected duration of the latent stage of infection. **Region-, time- and strata-specific variation permitted.**

d_I The expected duration of the infectious stage of infection. **Region-, time- and strata-specific variation permitted.**

ε The relative infectiousness of state I_2 to state I_1. If not equal to 1, replace, in Equation (9), $I_{1,b}(t_k) + I_{2,b}(t_k)$ with $I_{1,b}(t_k) + \varepsilon I_{2,b}(t_k)$. **Region-, time- and strata-specific variation permitted.**

ϕ The proportion of infections that develop the specific syndromic symptoms (e.g. ILI symptoms). **Region-, time- and strata-specific variation permitted**.

m Multiplicative modifiers of the contact matrices. Can apply to any times or strata, but must apply equally across each region.

k_A Parameter describing the amplitude of oscillation of R0. The time evolution of R0 is given by

$$R_0(t) = R_0^{\text{init}} + A_{R_0}\left(\cos\left(\frac{2\pi(t + sd - pd)}{365.25}\right) - \cos\left(\frac{2\pi(sd - pd)}{365.25}\right)\right).$$

Here A_{R_0} is the amplitude of oscillation for R_0, sd is the day of the year corresponding to day 1 of the epidemic and pd is the day upon which R_0 is expected to peak. To ensure positive values for R_0, k_A takes values in $[0,1)$ and is such that

$$A_{R_0} = \frac{k_A R_0^{\text{init}}}{1 - \cos(2\pi(sd - pd)/365.25)}$$

This parameter has no permitted regional, temporal or strata-specific variation.

ψ Exponential growth rates for the initial stage of the epidemic. Reparameterisation of R_0^{init}, to a parameter more readily identifiable from data.

$$R_0^{\text{init}} = \psi d_I \frac{\left(\frac{\psi d_L}{2} + 1\right)^2}{1 - \frac{1}{\left(\frac{\psi d_I}{2} + 1\right)^2}}$$

Regional variation only allowed.

υ Reparameterisation of the initial number of infectives, used to seed each region, such that υ and I_0 are related by

$$I_0 = \frac{d_I e^{\upsilon} \sum_a N_a}{p_1^g(t_0) R_0^{\text{init}}}$$

Regional variation only allowed.

ρ Initial proportion of the population susceptible to infection. An initial condition, so **regional- and strata-specific variation allowed**.

p^g Proportion of symptomatic cases that will end up in the GP consultation dataset. **Region-, time- and strata-specific variation permitted.**

p^h Proportion of symptomatic cases that will end up in the USISS hospital dataset. **Region-, time- and strata-specific variation permitted.**

B Parameters describing the rates of non-pandemic ILI GP consultation. **Region-, time- and strata-specific variation permitted.**

k_{sens} Test sensitivity of the virological swabbing process. **No variation presently permitted**.

k_{spec} Test specificity of the virological swabbing process. **No variation presently permitted**.

κ_d Day of the week effects on the reporting of GP consultations. **Region-, time- and strata-specific variation permitted.**

Appendix 3: Goodness-of-fit plots for the PR model

Presented here are plots of the goodness-of-fit of the PR model to the GP consultation data, the serological data and the virological data (in all regions except the North where denominators were frequently too small to get a reasonable comparison).

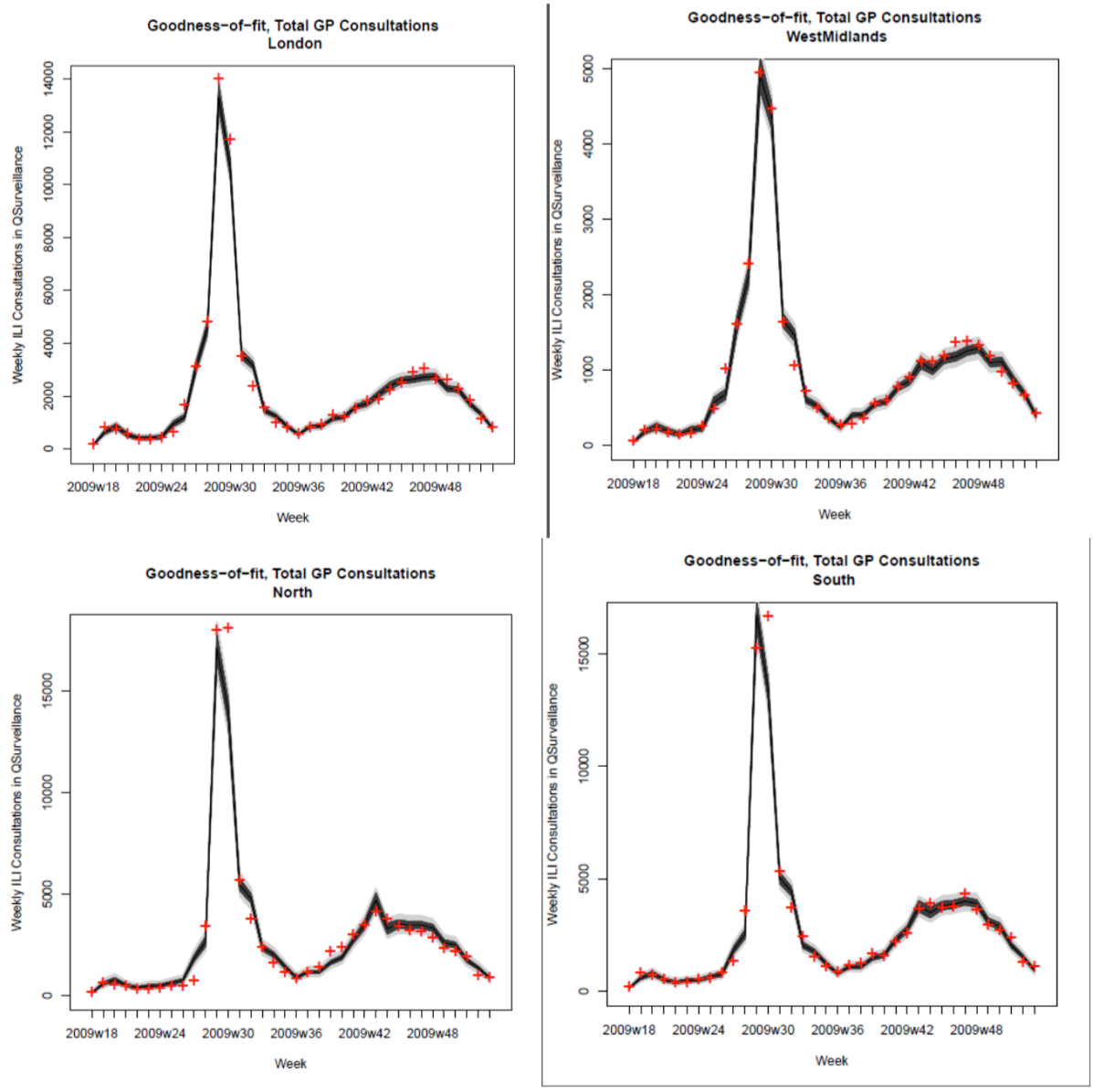

Figure 8 Goodness-of-fit of the PR model to the GP consultation data, aggregated by age. Red +'s indicate the observed numbers. The darker shaded area represent the 95% CrI for the expected number of consultations, the wider, lighter grey interval gives a poster predictive 95% CrI for the observed data – i.e. 95% of the data points should lie within the wider interval over time. Adapted from Birrell et al.[28]

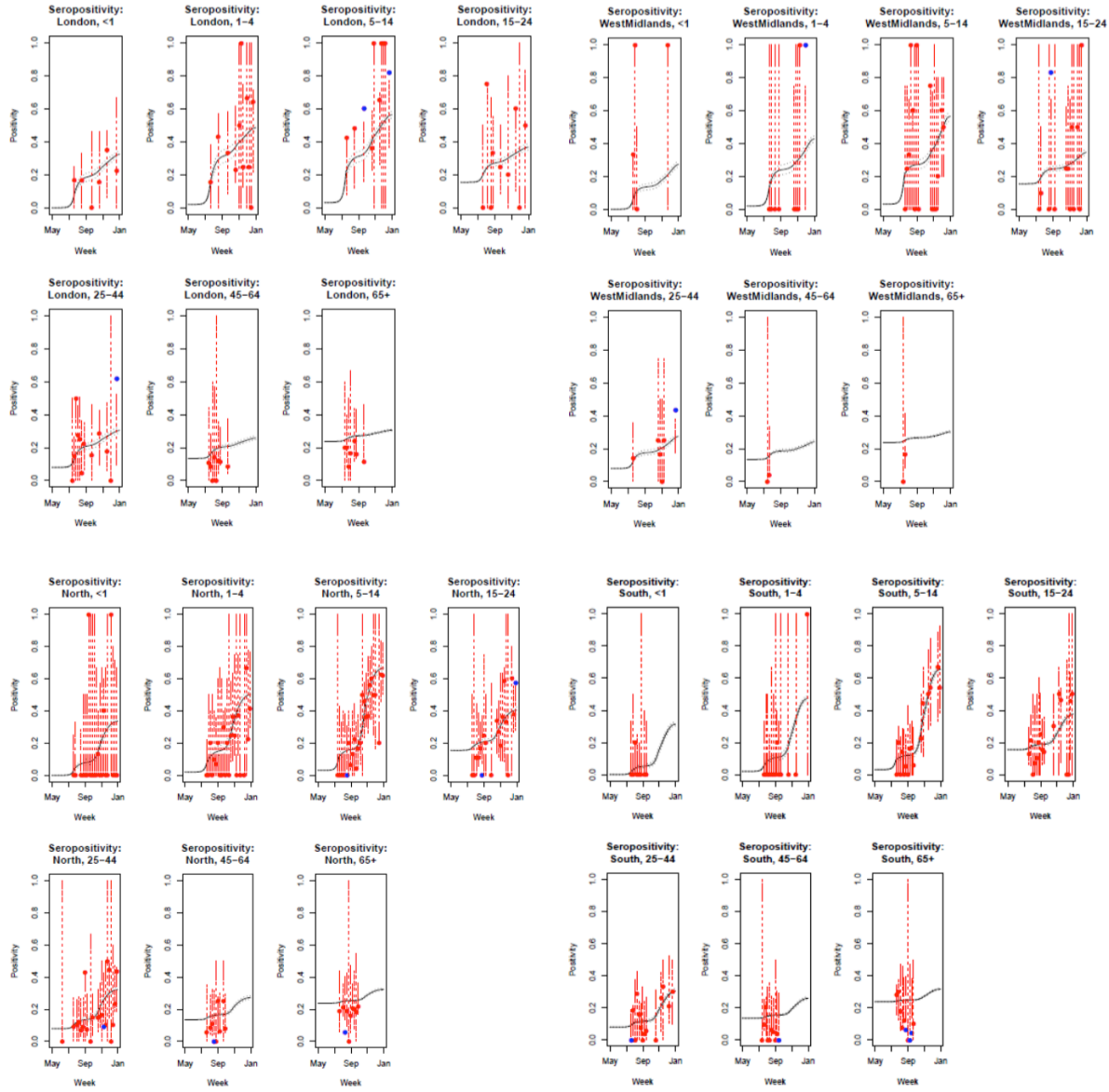

Figure 9 Goodness-of-fit of the PR model to the serological data, stratified by age and region. Data points are marked by the dots, with blue dots being those that are omitted by the model prectied 95% CrIs given by the vertical dashed lines. Adapted from Birrell et al.[28]

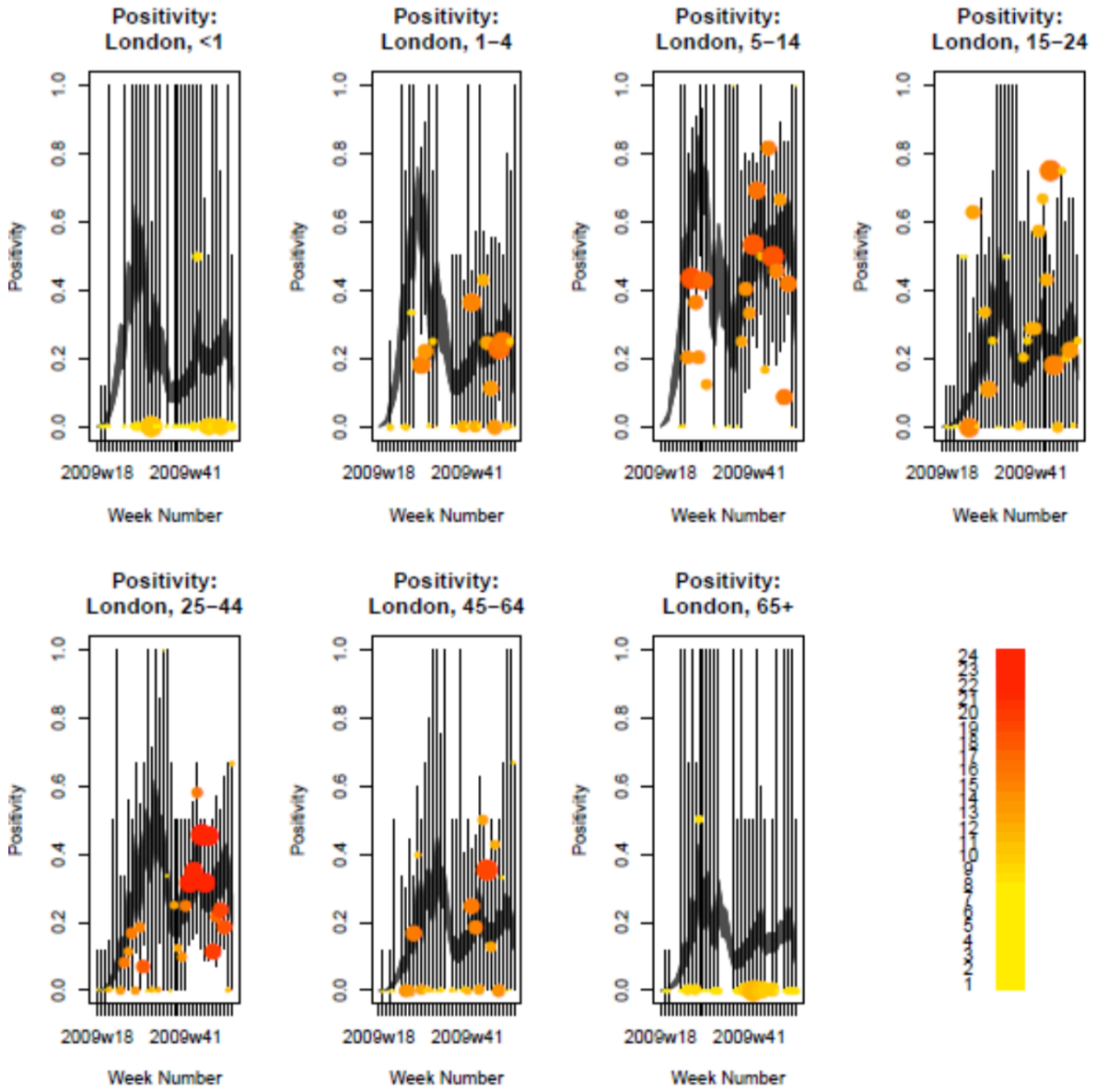

Figure 10: Goodness-of-fit of the PR model to the weekly-aggregated viropositivity data, stratified by age in London. Data points are given by the dots of variable width and colour. The width indicates the size of the denominator relative to other points in the sample plot. The colour of the points indicates the overall size of the denominator, with dark red points being those of largest sample size. The light grey vertical lines give a containing 95% CrI for the data points under the model. Adapted from Birrell et al.[28]

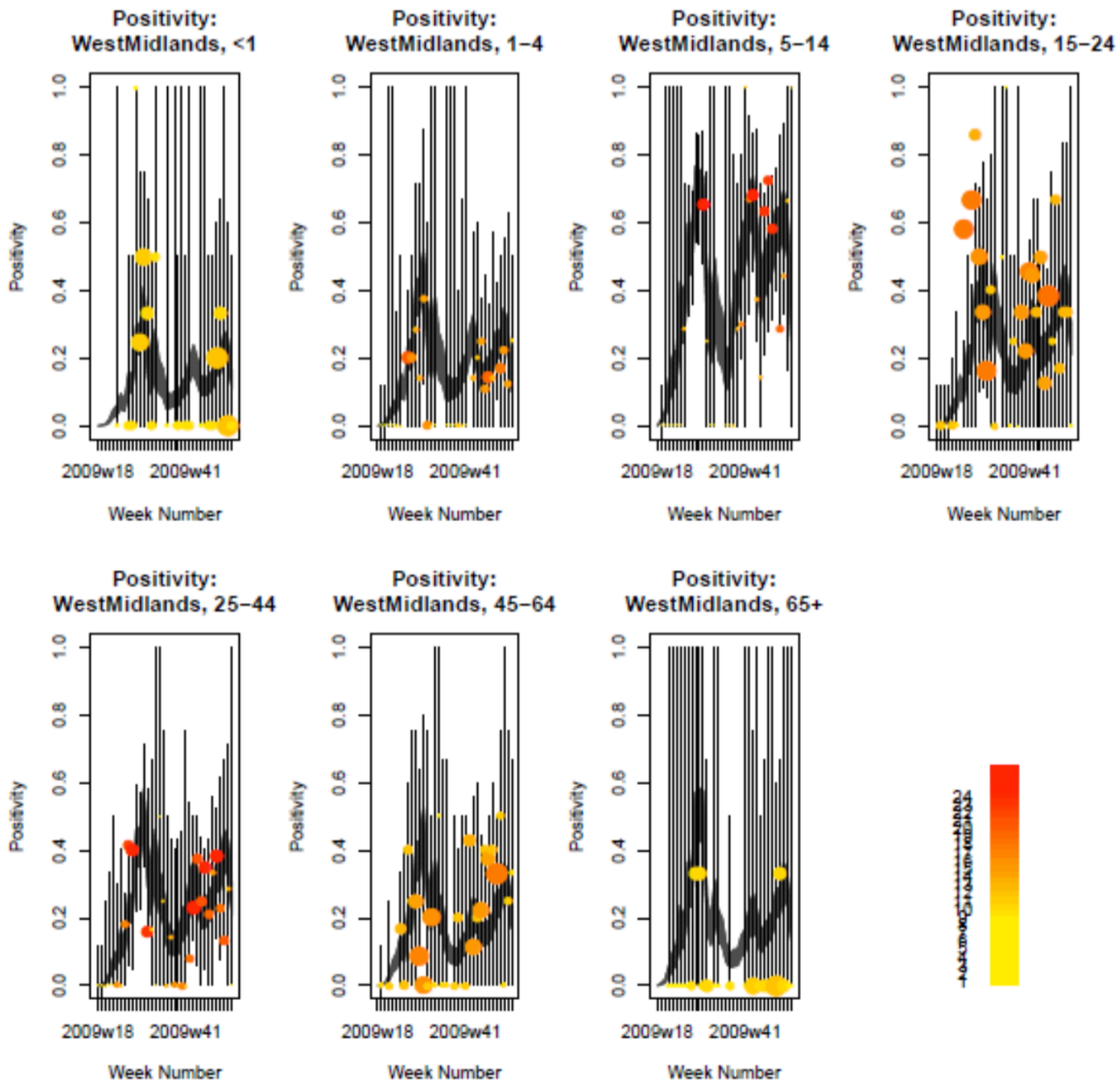

Figure 11 Goodness-of-fit of the PR model to the weekly-aggregated viropositivity data, stratified by age in the West Midlands. See Figure D3 for greater detail. Adapted from Birrell et al.[28]

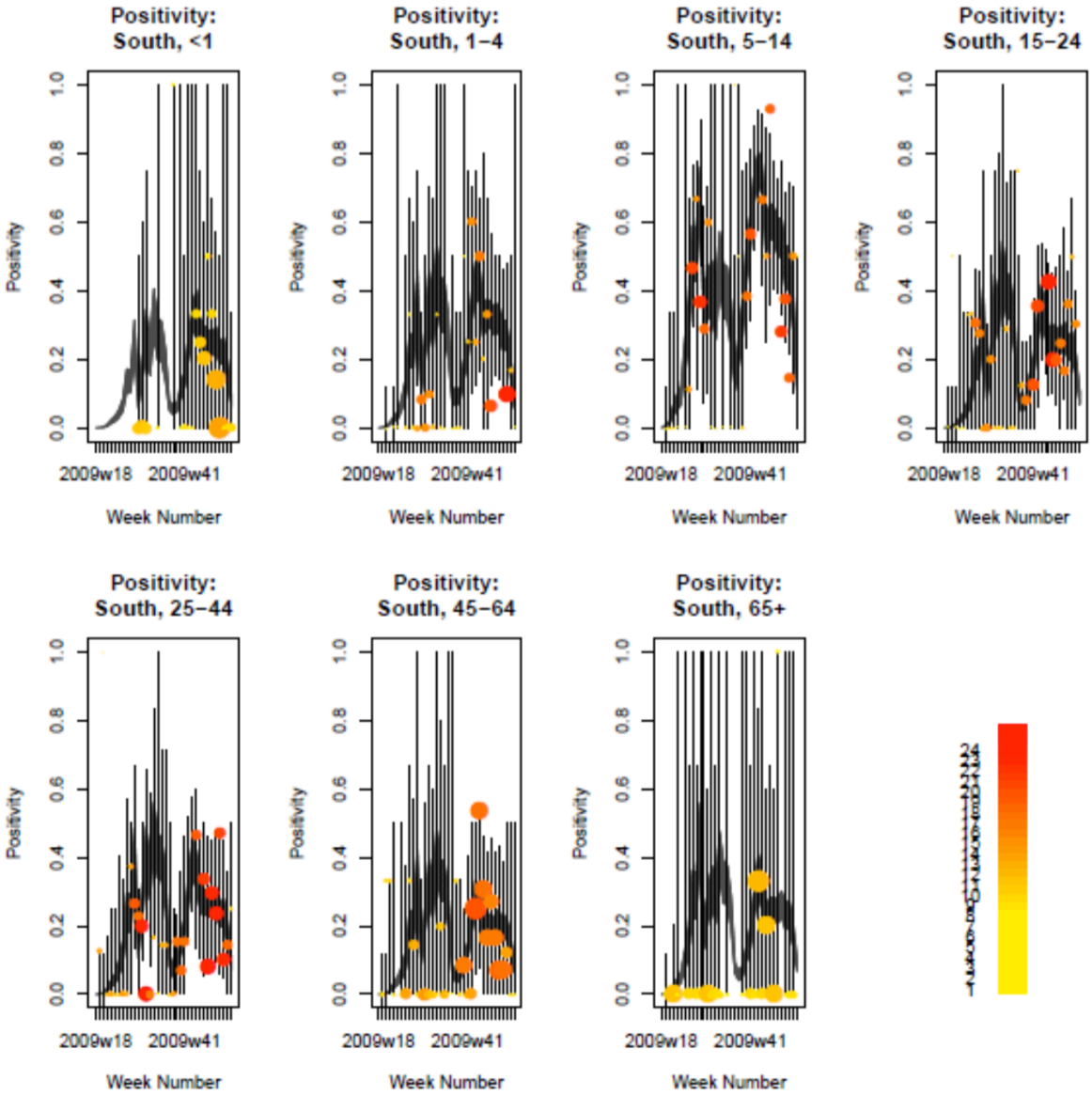

Figure 12 Goodness-of-fit of the PR model to the weekly-aggregated viropositivity data, stratified by age, in the South. See Figure D3 for greater details. Adapted from Birrell et al.[28]

Appendix 4: Goodness-of-fit plots for the MR model

Presented here are plots of the goodness-of-fit of the best-fitting MR model to the GP consultation data, the serological data and the virological data (in all regions except the North where denominators were frequently too small to get a reasonable comparison).

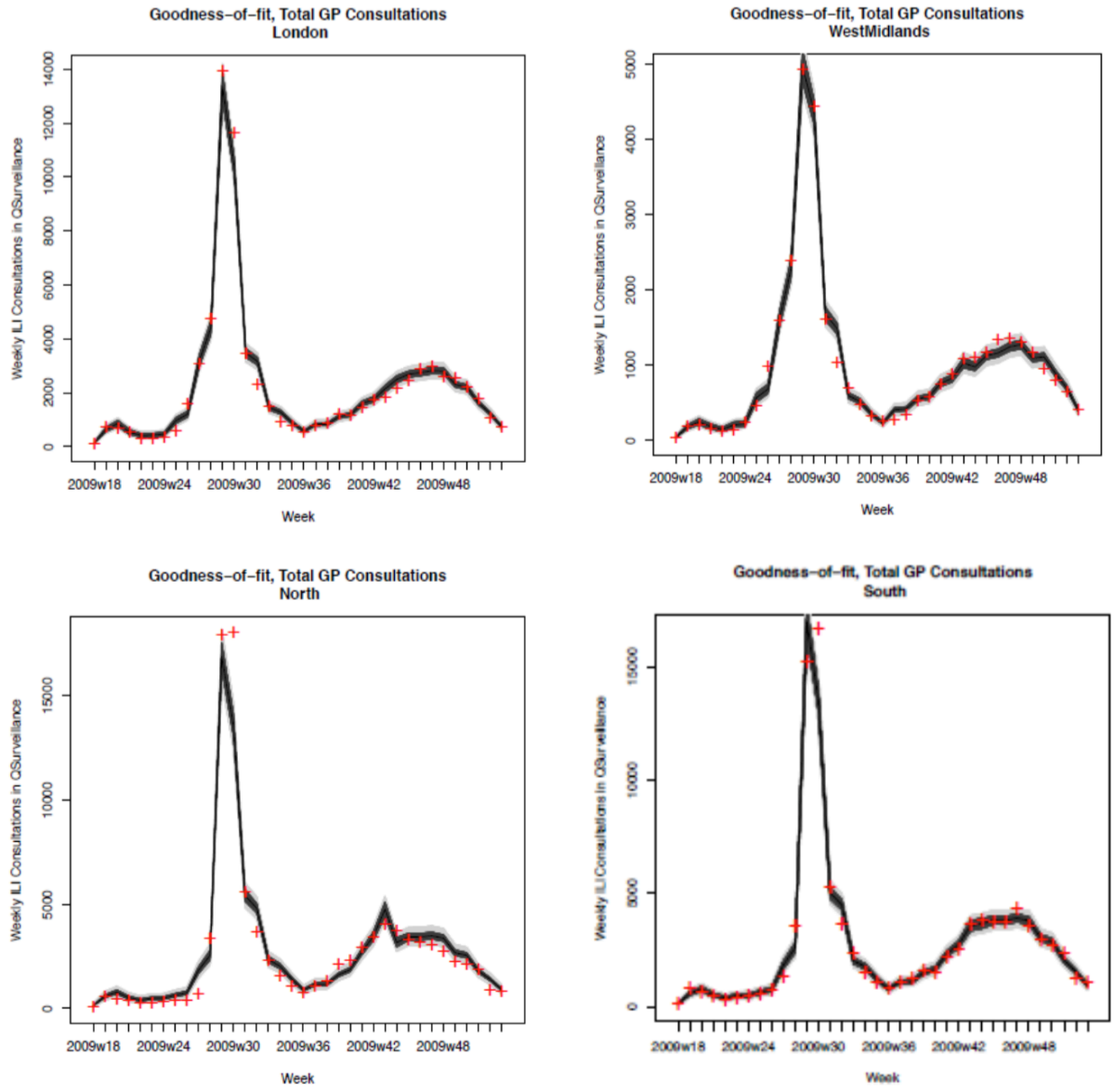

Figure 13 Goodness-of-fit of the MR model to the GP consultation data, aggregated by age. Red +'s indicate the observed number of consultations. The darker shaded area represent the 95% CrI for the expected number of consultations, the wider, lighter grey interval gives a posterior predictive 95% CrI for the observed data – i.e. 95% of the data points should lie within this wider interval. Adapted from Birrell et al.[28]

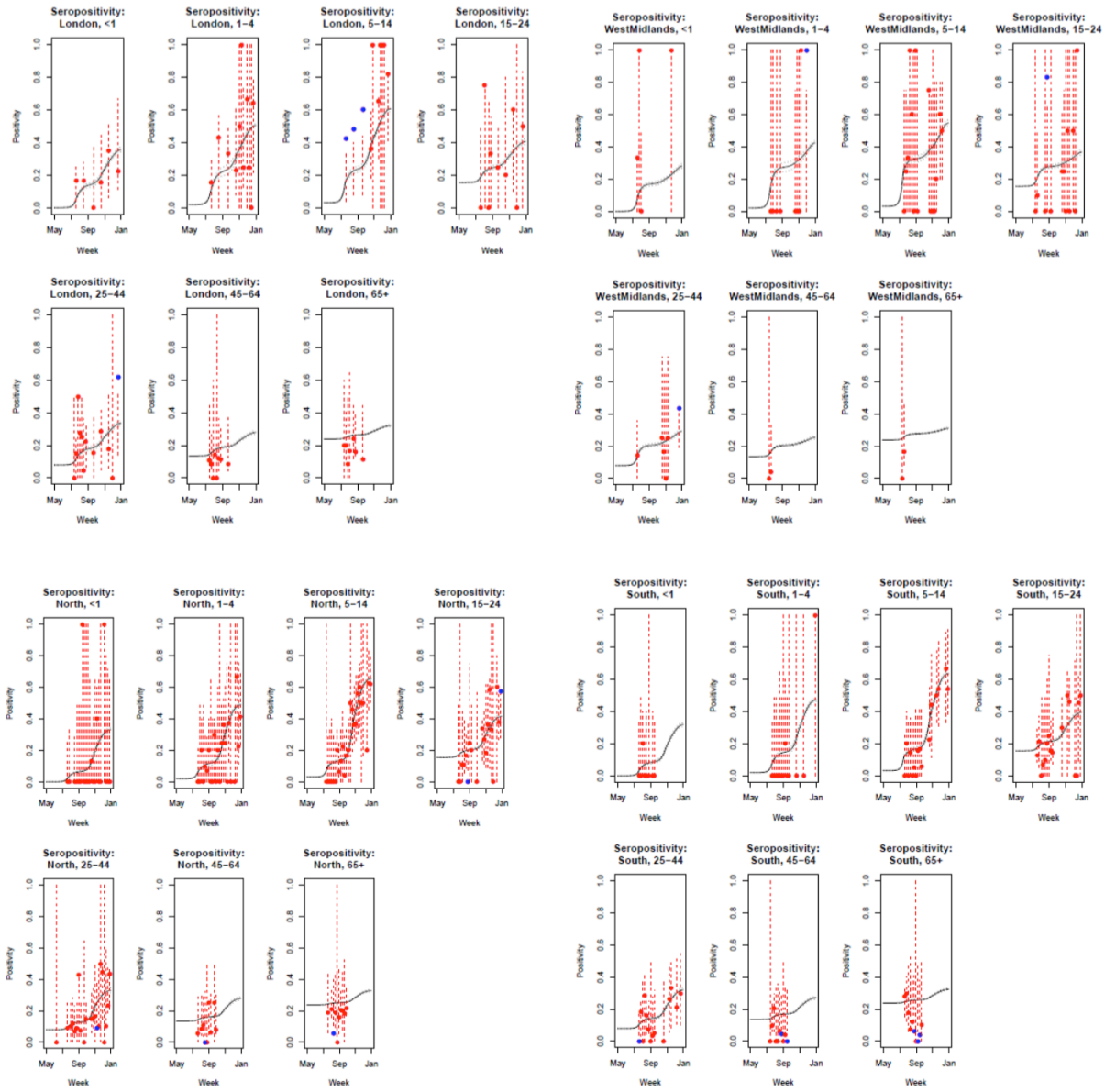

Figure 14 Goodness-of-fit of the MR model to the serological data, stratified by age and region. Data points are marked by dots, with blue dots being those that are omitted by the model predicted 95% CrIs shown by the vertical dashed lines. Adapted from Birrell et al.[28]

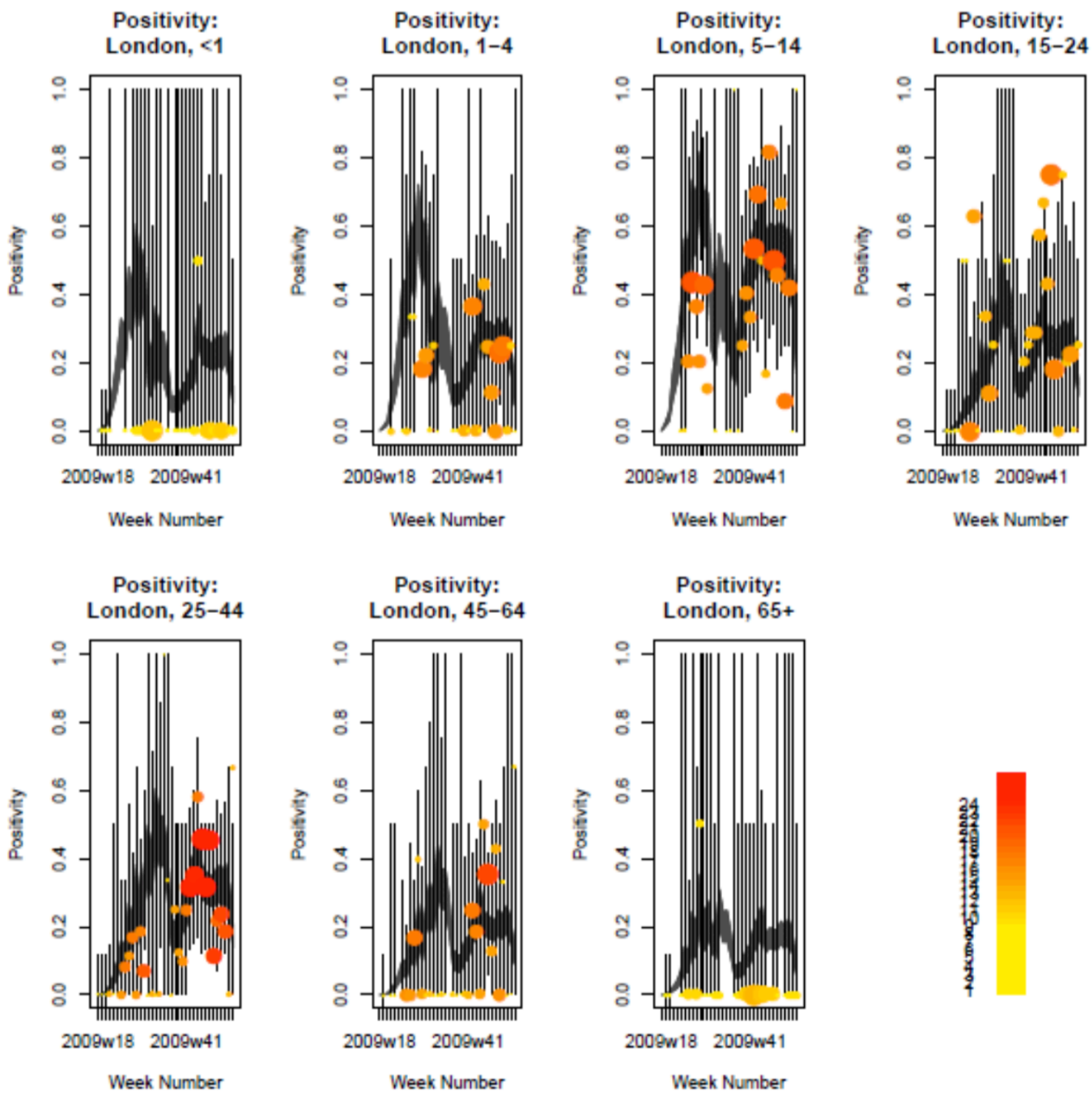

Figure 15: Goodness-of-fit of the MR model to the weekly-aggregated viropositivity data, stratified by age in London. Data points are given by the dots of variable width and colour. The width indicates the size of the denominator relative to other points in the same plot. The colour of the points indicates the overall size of the denominator, with dark red points being those of largest sample size. The light grey vertical lines give a containing 95% CrI for the data points under the model. Adapted from Birrell et al.[28]

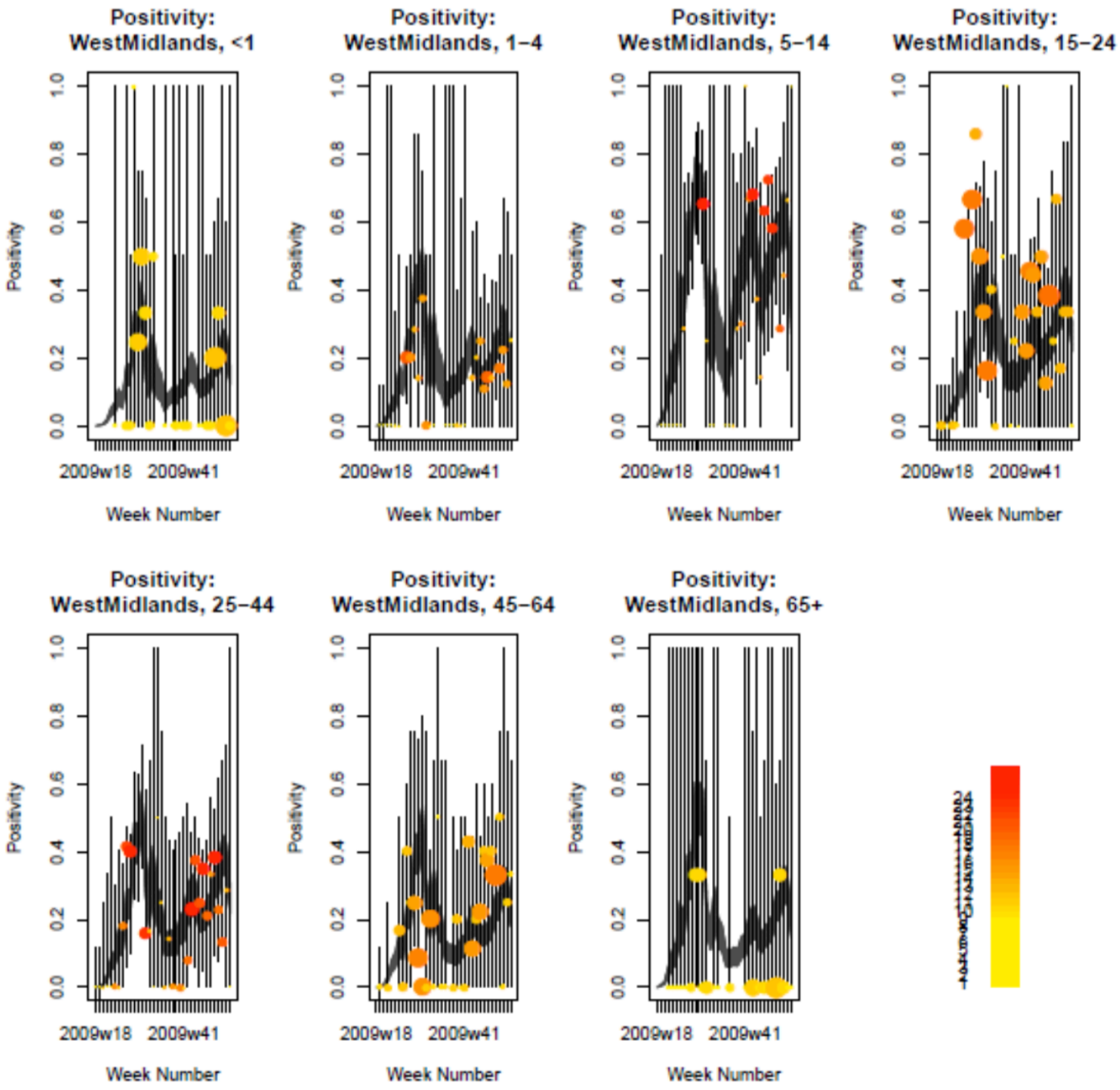

Figure 16 Goodness-of-fit of the MR model to the weekly-aggregated viropositivity data, stratified by age, in the West Midlands. See Figure D3 for greater detail.[28]

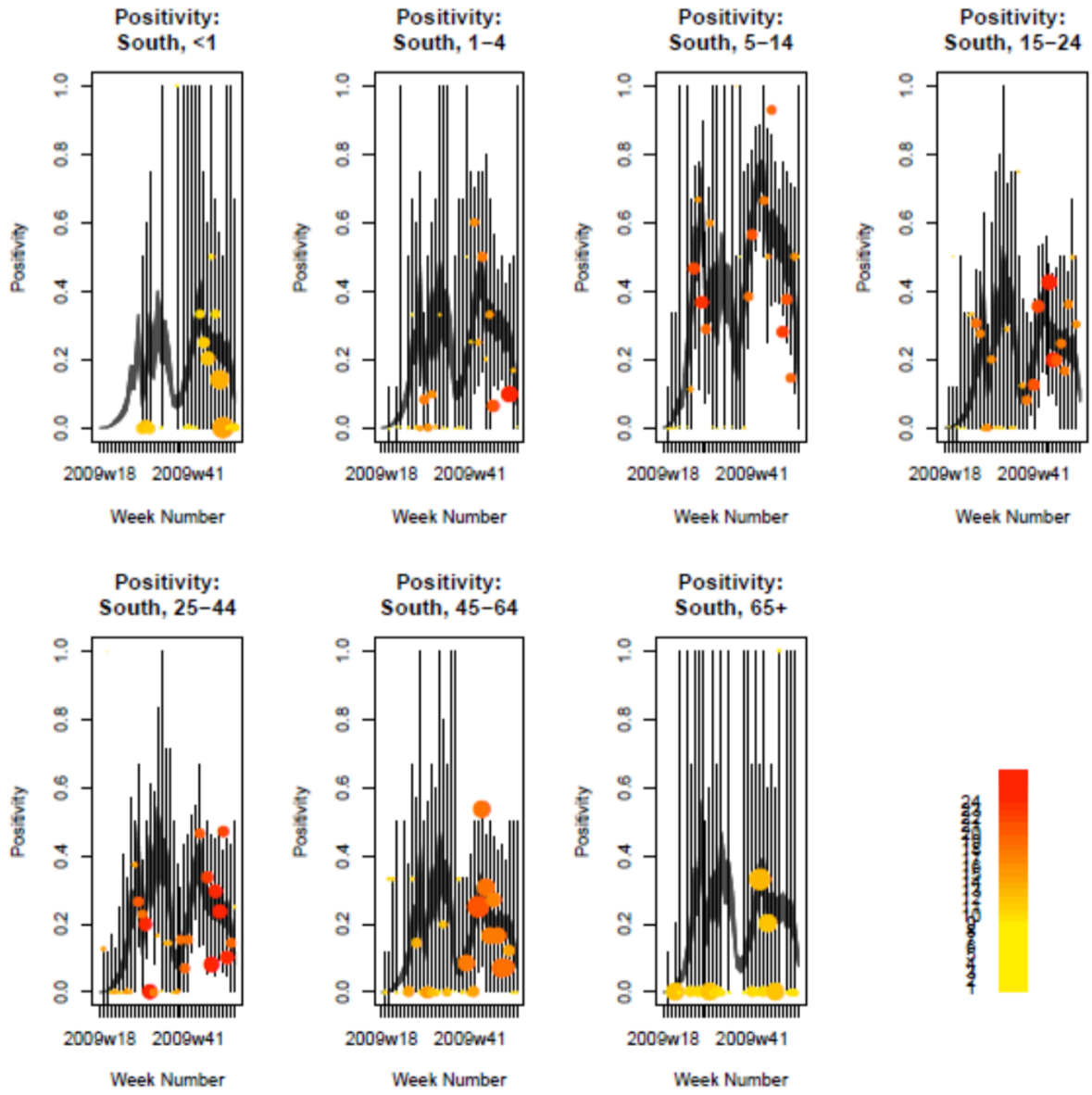

Figure 17: Goodness-of-fit of the MR model to the weekly-aggregated viropositivity data, stratified by age, in the South. See Figure D3 for greater detail.[28]

Appendix 5: Age-specific attack rates

Table 9 Age-specific attack rates. Tables give posterior median (and 95% CrIs) for the age-specific attack rates in each region in both waves of the 2009 pandemic. A) presents the posterior statistics from fitting the parallel-region model; and B) for the meta-region model. The last column gives attack rates averaged over four regions, weighted in accordance with the respective population sizes.[28]

A) Parallel region model

First wave (to end-Aug)	London	West Midlands	North	South	England
Overall	13.2(11.4,14.9)	9.8(8.5,11.2)	5.6(4.4,6.9)	3.6(2.9,4.5)	6.4(5.8,7.1)
<1y	18.7(16.2,21.1)	13.6(11.8,15.4)	8.0(6.4,10.0)	5.2(4.2,6.5)	9.7(8.7,10.6)
1-4 y	29.4(25.4,33.1)	22.1(19.2,25.1)	13.2(10.4,16.5)	8.7(7.0,10.8)	15.3(13.6,17.1)
5-14y	28.4(25.0,31.4)	24.1(21.4,27.0)	13.0(10.4,16.0)	9.0(7.2,11.0)	14.9(13.5,16.4)
15-24y	11.9(10.0,13.6)	9.2(7.8,10.7)	5.6(4.4,7.0)	3.4(2.7,4.2)	6.1(5.4,6.9)
25-44y	13.0(11.1,14.8)	9.6(8.3,11.2)	5.6(4.5,7.1)	3.7(2.9,4.5)	6.7(6.0,7.5)
45-64y	7.0(5.9,8.1)	5.2(4.4,6.1)	3.1(2.4,3.9)	2.0(1.6,2.4)	3.4(3.0,3.8)
65+y	3.5(2.9,4.0)	2.9(2.4,3.4)	1.6(1.3,2.1)	1.1(0.8,1.3)	1.8(1.5,2.0)
Second wave (Sep.-Dec.)					
Overall	10.1(8.5,11.9)	10.6(9.0,12.2)	19.3(17.8,21.0)	19.6(18.3,21.0)	17.1(16.2,18.3)
<1y	13.7(11.5,16.2)	13.7(11.5,15.7)	25.6(23.6,27.6)	26.2(24.6,27.9)	22.3(21.1,23.7)
1-4 y	17.1(14.1,20.7)	18.7(15.6,21.8)	34.8(31.8,37.9)	36.7(34.3,39.2)	30.7(28.9,32.8)
5-14y	24.9(21.0,29.3)	29.2(24.9,33.5)	50.4(46.8,53.9)	53.5(50.6,56.1)	45.6(43.5,47.7)
15-24y	9.5(7.9,11.2)	10.0(8.4,11.6)	19.4(17.6,21.3)	18.8(17.3,20.6)	16.7(15.5,18.2)
25-44y	9.5(7.9,11.2)	9.0(8.4,11.5)	18.6(16.9,20.4)	18.8(17.4,20.4)	16.0(15.0,17.4)
45-64y	5.4(4.5,6.4)	5.5(4.7,6.5)	10.7(9.7,11.8)	10.5(9.7,11.6)	9.4(8.7,10.3)
65+y	3.4(2.8,4.0)	3.6(3.1,4.2)	6.9(6.3,7.6)	6.9(6.3,7.5)	6.1(5.7,6.7)

B) Meta-region model

First wave (to end-Aug)	London	West Midlands	North	South	England
Overall	9.9(8.9,11.0)	12.4(11.5,13.3)	4.7(4.2,5.2)	6.0(5.4,6.6)	6.8(6.1,7.4)
<1y	14.0(12.6,15.4)	16.9(15.7,18.1)	6.5(5.8,7.2)	8.4(7.6,9.3)	9.7(8.8,10.6)
1-4 y	20.6(17.9,23.4)	25.4(22.6,28.0)	9.9(8.4,11.4)	12.7(11.0,14.5)	14.4(12.5,16.3)
5-14y	20.8(19.0,22.7)	29.6(27.6,31.6)	10.4(9.5,11.5)	13.8(12.6,15.1)	15.3(14.0,16.6)
15-24y	9.3(8.1,10.5)	12.2(11.2,13.3)	4.9(4.3,5.6)	6.0(5.4,6.8)	6.7(6.0,7.5)
25-44y	10.0(8.9,11.2)	12.5(11.6,13.6)	4.9(4.3,5.4)	6.4(5.7,7.1)	7.1(6.4,7.9)
45-64y	5.4(4.8,6.1)	7.0(6.4,7.6)	2.7(2.4,3.1)	3.5(3.1,3.9)	3.8(3.4,4.3)
65+y	2.7(2.3,3.0)	4.0(3.6,4.3)	1.4(1.3,1.6)	1.9(1.7,2.2)	2.0(1.8,2.3)
Second wave (Sep.-Dec.)					
Overall	16.2(14.9,17.6)	8.9(7.5,10.4)	20.6(19.6,21.7)	18.0(17.0,19.0)	17.8(16.7,18.9)
<1y	21.7(19.9,23.5)	11.1(9.4,13.0)	26.7(25.4,28.0)	23.5(22.1,24.8)	23.0(21.6,24.4)
1-4 y	28.0(25.2,30.9)	15.1(12.5,18.0)	37.0(35.0,39.2)	32.6(30.4,34.8)	31.5(29.3,33.8)
5-14y	36.7(34.0,39.5)	22.0(18.8,25.5)	51.8(50.2,53.5)	46.7(44.7,48.6)	44.5(42.5,46.6)
15-24y	15.8(14.5,17.3)	9.0(7.6,10.6)	21.2(19.7,22.6)	17.9(16.7,19.3)	17.9(16.6,19.3)
25-44y	15.6(14.3,17.0)	8.7(7.3,10.2)	20.2(18.9,21.4)	17.7(16.5,18.9)	17.3(16.0,18.5)
45-64y	9.1(8.3,10.0)	5.0(4.2,5.9)	11.7(10.8,12.6)	10.2(9.4,11.0)	10.1(9.3,10.9)
65+y	5.6(5.1,6.2)	3.3(2.7,3.9)	7.6(7.0,8.1)	6.7(6.2,7.2)	6.5(6.0,7.1)