LONDON
SCHOOL *of*
HYGIENE
&TROPICAL
MEDICINE

Spatio-Temporal Patterns and Surveillance of Infectious Disease During Emergence and Elimination

Emily Sara Nightingale
BSc, MSc (Med Stats)

Supervisors:
Prof. Graham F Medley
Dr. Oliver J Brady

Thesis submitted in accordance with the requirements for the degree of
Doctor of Philosophy of the University of London

December 2022

Department of Global Health and Development
Faculty of Public Health and Policy
Funded by The Bill and Melinda Gates Foundation

Declaration of Authorship

I, Emily Sara Nightingale, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed [redacted]

Date: 16.12.2022

Abstract

Surveillance of health outcomes in space and time gives insight into the underlying mechanisms driving those outcomes, for example with respect to exposures, risk factors and - in the case of infectious diseases - transmission. There are unique challenges for the analysis and interpretation of such data in the case of a rare and/or declining disease and of a rapidly growing novel epidemic. Policy objectives are also often similar - to target interventions to regions most at risk, both in the present and the near-future, for most efficient and effective use of available resources. This thesis uses the examples of visceral leishmaniasis elimination in north-eastern India and the emergence of the novel SARS-CoV-2 virus in England to explore these ideas, from the perspective of decision-making around policies for disease control.

Spatial variability driven not by transmission of the disease itself but of its observation influences decisions on further intervention - in the case of VL feeding back into surveillance policies to create a cycle of increasing bias - but is rarely quantified. Patterns in observed incidence of both diseases suggest that the impact of these surveillance systems is likely not consistent over space. The potential pattern of this spatially-varying bias is characterised for visceral leishmaniasis with respect to delays to diagnosis - an important indicator not only of the strength of the surveillance system but also of the likelihood of breaking transmission in low incidence areas.

An understanding of the mechanism of surveillance is crucial for appropriate interpretation of the resulting data, and hence for appropriate distribution of intervention efforts. However, in particular in resource-constrained settings of emergence and elimination, the process can be stochastic and difficult to define. This motivates the routine collection of data describing the surveillance process itself, to provide important context to the resulting observations. Ways in which such information could be incorporated going forward are discussed, suggesting directions of future research to more accurately infer the true spatial distribution of disease burden in such settings.

Acknowledgements

Firstly I would like to thank my supervisors, Prof. Graham Medley and Dr. Oliver Brady, for their invaluable support throughout this endeavour - for helping me to develop and refine ideas, providing constructive feedback and building my confidence as an independent researcher. Also thanks to my advisors Dr. Tim Lucas and Dr. Leonardo Bastos for their helpful advice on methodology and implementation.

I thank the Bill and Melinda Gates Foundation for funding the SPEAK India Consortium and in turn myself, giving me the opportunity to undertake this degree alongside my work and offering the flexibility to redirect my time in order to contribute to the COVID-19 pandemic response. I am particularly grateful to colleagues within the consortium at LSHTM, CARE India in Patna and the Vector Control Research Centre (VCRC) in Chennai, without whose generous collaboration this work could never have happened. Joy Bindroo and Sridhar Srikantiah at CARE offered much of their time to help me understand the nuances of the data, and to frame those data within the reality of the situation on the ground. Thank you to the field teams who work tirelessly to collect the surveillance data and to manage and implement the interventions, and to the National Centre for Vector Borne Diseases Control (NCVBDC) for granting me permission to conduct this research around the VL elimination programme.

Thanks to everyone in the CMMID COVID-19 Working Group for giving me the chance to collaborate on many fascinating projects which kept me motivated during very challenging times. Thank you to friends and colleagues at LSHTM and in CMMID for creating a supportive and inspiring space within which to take on a PhD - it's been incredibly valuable to be surrounded by so many friendly faces to both celebrate and commiserate with. Thanks to my family for your endless encouragement, even when it comes in the form of insisting I deserve another break after only ten minutes of work...

Finally thanks to Ben, who has kept me grounded through this wild ride and who is I think still slightly baffled as to why I put myself through it. Thank you for all the times you've quietly brought me coffee and biscuits while I've been lost in my computer screen - it still pains me to think of that time I was too engrossed to notice your handsome horseshoe mustache. I hope one day you'll surprise me again with it.

Contents

COVID-19 Impact Statement **5**

1 Introduction **7**

1.1 Background 7

1.1.1 Spatio-temporal analysis in settings of emergence and elimination . 8

1.1.2 Elimination of visceral leishmaniasis in India 9

1.1.3 Emergence of COVID-19 in England 14

1.2 Aims and Objectives 15

1.3 Outline of thesis 16

2 Methods **17**

2.1 Spatial and temporal structures in disease surveillance data 17

2.2 A frequentist model of spatio-temporal correlation: The endemic-epidemic framework 19

2.3 Bayesian modelling of spatio-temporal correlation 19

2.3.1 The Integrated Nested Laplace Approximation (INLA) 20

2.3.2 Penalised Complexity Priors 20

2.4 The Besag-York-Mollié Model 21

2.5 The Stochastic Partial Differential Equation Model 22

2.6 Disaggregation regression 23

2.7 Application of these methods in the thesis 23

3 A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India **25**

4 Inferring the distribution of visceral leishmaniasis incidence from data at different spatial scales **49**

4.1 Abstract 49

4.2 Background . . . 50
4.3 Materials and methods . . . 53
4.3.1 Data . . . 53
4.3.2 Descriptive analysis . . . 54
4.3.3 Disaggregation model structure . . . 54
4.3.4 Validation . . . 55
4.4 Results . . . 57
4.4.1 Data cleaning . . . 57
4.4.2 Descriptive . . . 57
4.4.3 Disaggregation model fit . . . 58
4.4.4 Model validation and comparison . . . 59
4.5 Discussion . . . 61
4.5.1 Limitations . . . 65
4.5.2 Conclusions . . . 66

5 Spatial variation in diagnosis delay for visceral leishmaniasis in Bihar, India 68

6 The local burden of disease during the first wave of the COVID-19 epidemic in England: Estimation using different data sources from changing surveillance practices 93

7 Discussion 112
7.1 Challenges of monitoring and modelling disease at emergence and elimination 112
7.2 Principle findings of this thesis . . . 113
7.3 Policy Implications . . . 115
7.4 Limitations . . . 117
7.5 Future directions . . . 119

Bibliography 123

Appendices 141
A Supplementary Materials: A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India . . . 141
B SPEAK India Toolkit: Short-term prediction of block-level VL diagnoses for routine surveillance [Draft] . . . 149

C Supplementary Materials: Inferring the distribution of visceral leishmaniasis incidence from data at different spatial scales 155

D Supplementary Materials: Spatial variation in diagnosis delay for visceral leishmaniasis in Bihar, India 160

E Supplementary Materials: The local burden of disease during the first wave of the COVID-19 epidemic in England 171

Note to the reader

Internal hyperlinks have been used throughout this thesis. When reading the thesis on a computer, the reader may click on a hyperlink to refer back to a relevant section. Once finished, the reader can click Alt and ← (at the same time) to return to the original location.

COVID-19 Impact Statement

Between March 2020 to August 2021 I joined the School's pandemic response as a member of the Centre for Mathematical Modelling of Infectious Diseases (CMMID) COVID-19 working group. This work consisted of elements of knowledge generation, internal (within the working group) and external (with outputs shared with government departments and colleagues at other universities/research institutes) contribution. My initial contribution was in assisting with the development and maintenance of a data pipeline, managing multiple streams of data from different sources which were being shared with members of CMMID. I wrote scripts for automated cleaning and processing of the data into a standardised format which could then be used in modelling and analysis tasks.

During this time I contributed to several published outputs from the collaborative work of the group. I initially contributed to early projections of total infections based on observed critical care admissions [1], assisting with finalising and publishing the code and ensuring findings were quickly made available through a summary on the working group website. In addition to the paper described in 6, I published one further paper as sole first author; this explored density-dependence in the transmission of COVID-19 and its potential implications on the impact of lockdown measures in different parts of the country [2]. Specifically, I contributed a regression analysis to assess the association between population density in different functional forms and COVID-19-related deaths at the lower-tier local authority level, while the senior author compared transmission models under those different assumptions of density-dependence. We both then developed the manuscript which combined these two analytical components.

I co-first-authored two publications, the first being a rapid systematic review of length-of-stay for hospitalised COVID-19 patients [3] which consolidated the early available evidence in order to inform the parameterisation of transmission models. I lead the analysis of the extracted estimates after screening, having proposed the initial idea for combining the estimates into a distribution, and my co-first-author and I equally contributed to writing the manuscript. The second investigated trends in NHS 111 notifications in relation to

COVID-19 and how these could be indicative of subsequent case incidence [4]. I helped conceptualise the analysis, co-developed the code, created visualisations and jointly wrote the manuscript. I am also a named author on a subsequent publication which extended this work, developing a self-contained tool for flagging sudden changes in NHS 111 trends for investigation as possible outbreaks [5].

I served as second author on another spatial analysis of COVID-19 disease burden in England which inferred patterns of population movement across the country in response to the changing restrictions from national to local scale [6]. I advised as to how the first author's inferred networks could be linked to the local burden of COVID-19-related deaths that I had been exploring in my other work, and reviewed and edited the manuscript.

From June 2020 I became involved with the OpenSAFELY platform https://www.opensafely.org/, based at the University of Oxford with collaboration from colleagues at LSHTM, NHS clinical researchers, and private healthcare software providers. The original planned analysis had aimed to infer risk of introduction of infection to residential care homes from the disease burden in the population of the surrounding geographic area [7]. This project ultimately proved not to be feasible with the available data and given the early developmental stage of platform. However, being involved from such an early stage allowed me to directly contribute to how the platform evolved, in particular developing the capacity for analyses to be conducted in R as opposed to the default STATA software. I discovered, investigated and defined specific issues with some of the identifying variables which the developers would not otherwise have been aware of. I was closely involved with two other research projects investigating infections within care homes and households, reviewing analysis outputs, advising on navigating these data issues and providing feedback on the manuscripts for publication [8, 9, 10].

Chapter 1

Introduction

1.1 Background

Understanding the distribution of an infectious disease in space can be crucial for anticipating transmission and hence for the successful implementation of control measures, from emergence to elimination. At the emergence of a novel disease, exploration of spatial patterns can give insight into the spread of infection, likely routes of transmission and potential variation in risk according to characteristics of the local population. Prior knowledge of population distribution and usual movement patterns can help anticipate how the disease may disperse spatially from an initial seeding location. However, during the rapid growth phase of a new epidemic, policies and practices evolve in response to accumulating information, which is then reflected in the available data. This can reduce the comparability between data collected at different points in time and place, and hence complicates the interpretation of spatio-temporal patterns. Interpreting different sources of information in combination can provide a clearer perspective into the underlying process. At the other end of the timeline, where an elimination programme has been successful, new cases become increasingly sparse in space and sporadic in time. This makes it increasingly difficult to detect each individual case and hence evaluate true progress. Despite control efforts bringing the overall incidence of a particular disease to low levels across the region as a whole, it is common for transmission to persist within small pockets - often the most socially and economically disadvantaged - of the population. On top of the disproportionate suffering of an already vulnerable group of people, these pools of infection allow resurgence on a large scale to remain a serious threat. Sustaining the achievements of a programme in the long term requires careful surveillance in order to promptly detect and respond to any new evidence of transmission.

From a statistical perspective, there are several challenges in addressing such problems. Any approach taken must acknowledge the lack of independence inherent in spatio-temporal data, potentially explicitly modelling the correlation structure between observations. In both scenarios, the dynamics of the disease are in transition and the smoother, more predictable patterns of an established, endemic disease can break down. Control efforts are also in flux, with policy makers responding to outbreaks as they unfold and adapting their approach as information is accumulated. Analysts should consider how robust their chosen methods may be when applied to low counts and irregular trends, and how they can be evaluated in a way that is relevant for policy and policy-makers, particularly with respect to quantifying uncertainty.

The contexts of disease emergence and elimination have many parallels with respect to the attributes of the data and nature of the analysis problem. In both cases incidence can be low, sporadic and unpredictable. Community awareness can be weak due to relevant information not yet being available (emergence) or due to complacency/waning as a result of extended periods without incidence (elimination). Both are periods of transition, marked by changing disease dynamics in response to changing policy and, vice versa, changing policy in response to changing disease dynamics.

1.1.1 Spatio-temporal analysis in settings of emergence and elimination

Spatial and spatio-temporal analysis are predominantly used in elimination settings to understand the likelihood of equitably attaining elimination goals across affected regions, and to investigate ways in which the resource demands of control programmes may be reduced through more efficient targeting. A wealth of research in mapping and predicting disease burden over space has in particular been motivated by the global elimination goals for malaria [11, 12]. The Malaria Atlas Project https://malariaatlas.org/, for example, makes use of a wide range of data sources from administrative-level surveillance and intervention coverage to local prevalence surveys, vector biology and remote-sensed environmental characteristics, in order to obtain fine-scale maps of burden and risk [13]. Given that transmission is dependent on presence of the competent mosquito, there is a clear geographic component to incidence of malaria; however, the insight that can be gained from spatio-temporal patterns is not limited to vector-borne diseases. Spatio-temporal analysis also contributes to elimination strategies for non-vector-borne diseases such as tuberculosis [14, 15, 16] and HIV [17, 18, 19], helping to identify common risk factors, barriers to care, and co-incidence with other diseases.

Changes in the epidemiology of a disease nearing elimination have been noted in many settings, attributed to an often complex combination of factors such as drug resistance, evolution of new strains, waning immunity, changing risk factors and risk perceptions, and cross-border importation. Examples include malaria [20, 21, 22, 23], tuberculosis [24, 25], yellow fever [26], and meningococcal disease [27]. Particular challenges arise in the case of neglected tropical diseases (NTDs), where data from which to draw inference about elimination tend to be more sparse. Spatial analysis approaches which draw information from nearby observations can improve our understanding based on the limited data available [28], but ideally the surveillance system would be adapted to better capture low and scattered disease burden as elimination is approached and beyond [29]. Across all disease settings, there is clearly a need to reassess policy and potentially redistribute resources in order to navigate the critical time point of near-elimination.

For emerging diseases, spatio-temporal analysis of incident cases has been used to identify new clusters, to assess the geographic scale of outbreaks and to predict future spread [30, 31]. The often limited availability of data at a suitable geographic scale and privacy concerns regarding the first few observed cases of a new disease can, however, restrict what is feasible to analyse [32].

1.1.2 Elimination of visceral leishmaniasis in India

Visceral leishmaniasis (VL) is the acute disease caused by *Leishmania donovani*, a parasite transmitted through infected female *Phlebotomus argentipes* sandflies when they seek a human blood meal to mature their eggs (Figure 1.1.1). Symptomatic infection is largely considered to be fatal if left untreated and, across the remaining endemic regions of South America, East Africa and South Asia, incidence is concentrated within physically and economically vulnerable populations in rural areas. Decaying organic waste around rural human habitations provides ideal breeding sites for the vector, and poor housing conditions and time spent working and sleeping outdoors increases the exposure of these populations to bites. On top of that, malnutrition and high burden of other infectious diseases (e.g. TB, HIV) put people at higher risk of developing severe disease once infected.

In India, the burden of disease is largely contained within the four northeastern states of Bihar, Jharkhand, Uttar Pradesh and West Bengal, with Bihar most broadly affected [34, 35] (Figure 1.1.2). Incidence of VL in India has decreased substantially since the initiation of the regional Kala-Azar Elimination Programme (KEP) in 2005, which aims to tackle the disease across the Indian subcontinent (including the endemic region of

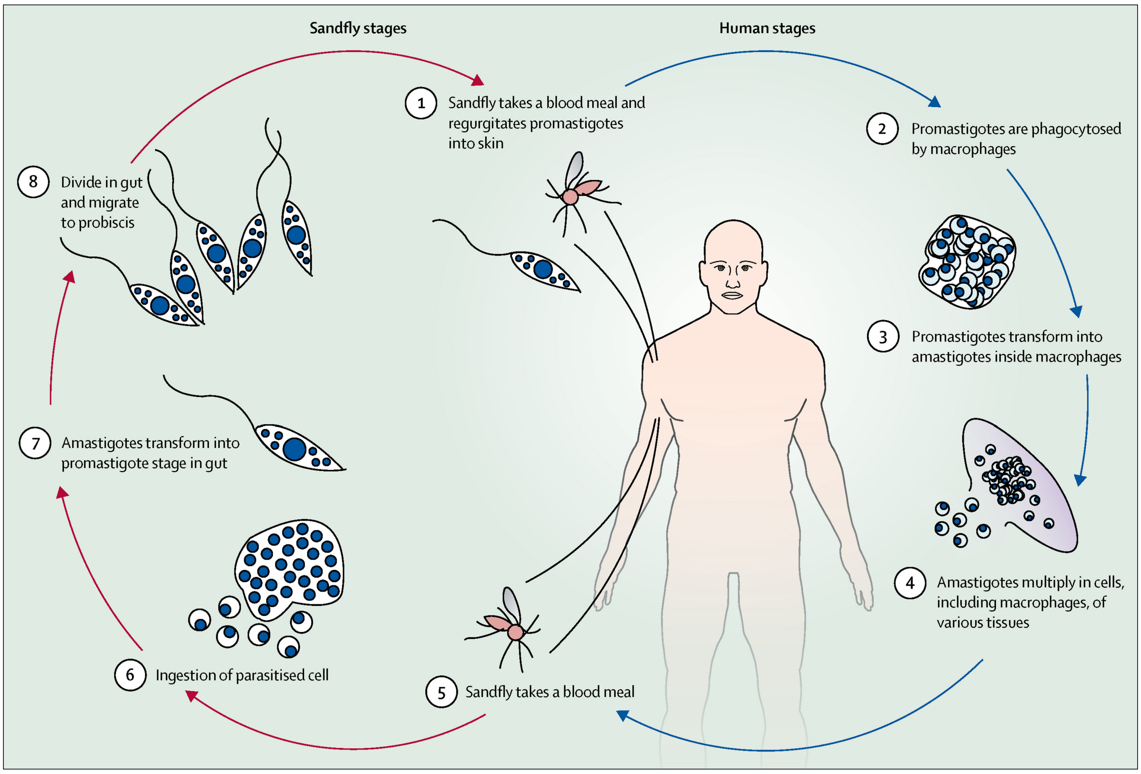
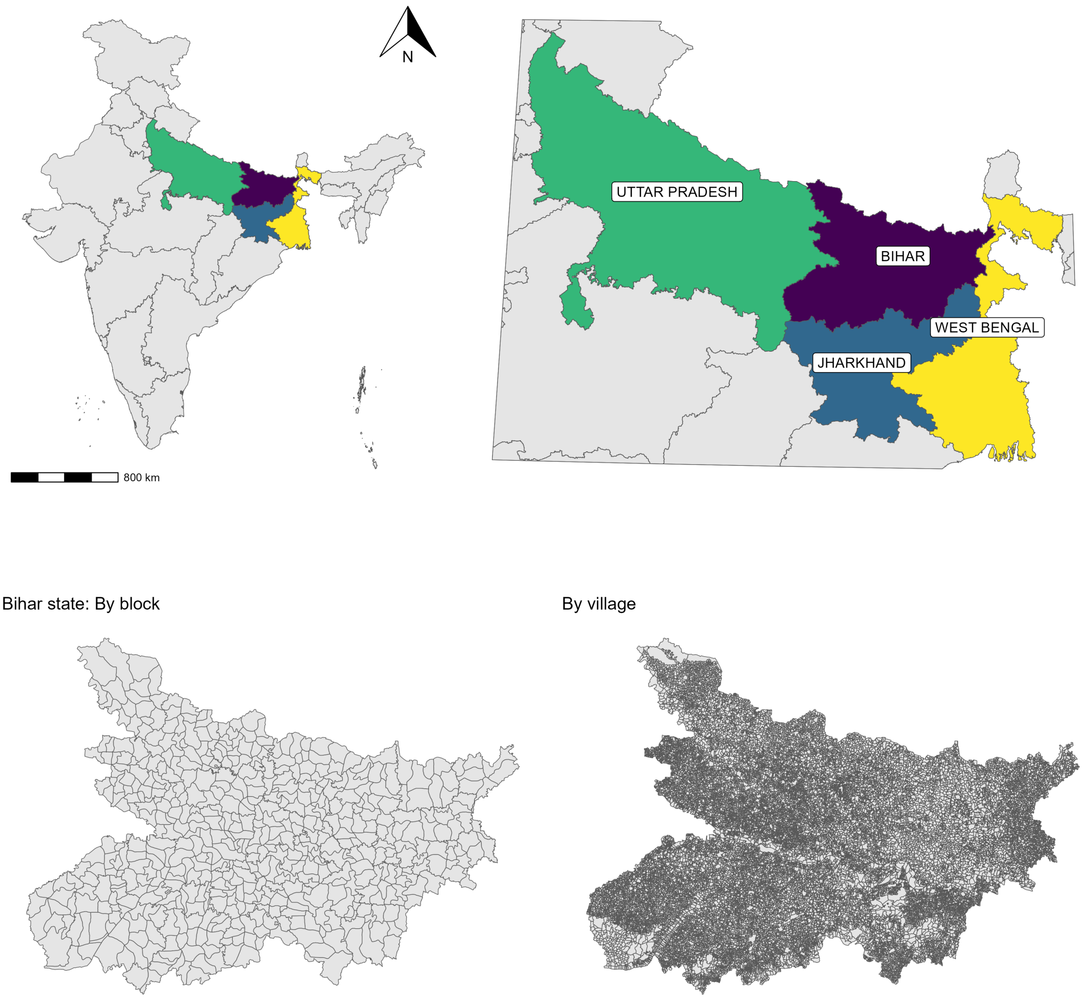

Sandfly stages
Human stages
1 Sandfly takes a blood meal and regurgitates promastigotes into skin
2 Promastigotes are phagocytosed by macrophages
3 Promastigotes transform into amastigotes inside macrophages
4 Amastigotes multiply in cells, including macrophages, of various tissues
5 Sandfly takes a blood meal
6 Ingestion of parasitised cell
7 Amastigotes transform into promastigote stage in gut
8 Divide in gut and migrate to probiscis

Figure 1.1.1: The life cycle of the parasite *L. donovani* that causes visceral leismaniasis disease in South Asia. Reprinted from The Lancet, Vol 392 / Issue 10151, Sakib Burza, Prof Simon L Croft, Prof Marleen Boelaert, Leishmaniasis, Pages No. 951-970, Copyright (2018), with permission from Elsevier. [33]

India, Nepal and Bangladesh) through enhanced case detection, treatment and reduction of vector density [36]. Reported cases in India fell from 29,000 in 2010 to less than 4,000 in 2018 [35, 36]. Historical data, however, illustrate cyclical behaviour in VL incidence with peaks approximately every fifteen years (Figure 1.1.3). It remains to be seen whether the reductions of the past decade reflect a genuine impact of interventions or simply another trough within this cycle.

Indian states are partitioned into administrative units in the form of districts and sub-districts or *blocks* across which the healthcare system is organised; each district has a hospital and each block a primary healthcare centre. The population of each block - varying in size between thirty thousand to several million - is typically clustered into villages of a few thousand people (Figure 1.1.2). The initial target of the VL elimination programme was to achieve “elimination as a public health problem” (EPHP) as part of the World Health Organisation’s roadmap for 2020 [37] - reducing incidence to less than one case per ten thousand per year at the block level. As of the end of 2019, this target was reported to have been achieve in 100% of blocks in Uttar Pradesh and West Bengal, 95% in Bihar and approximately 50% in Jharkhand [38]. It was suggested that



Bihar state: By block

By village

Figure 1.1.2: (Top) The four north-eastern states of India which remain endemic for visceral leishmaniasis. (Bottom) Administrative partitioning of Bihar state by sub-district i.e. "block" (bottom left) and by village (bottom right).

delays between onset of symptoms to treatment, complacency in implementation and poor access to services among the lowest socio-economic classes were among several challenges that had hindered achievement of the target.

The elimination programme in India incorporates two main forms of intervention: vector control and early diagnosis and treatment. Although efforts are coordinated and evaluated at the block level, these interventions are implemented on a village level. Indoor residual spraying (IRS) of insecticides is conducted responsively, with villages being marked for treatment according to whether they have reported any cases within the last three years. However, evidence for the efficacy of the current IRS protocol in practice is unclear, both in terms of sandfly abundance and subsequent disease incidence [39, 40]. Early diagnosis,

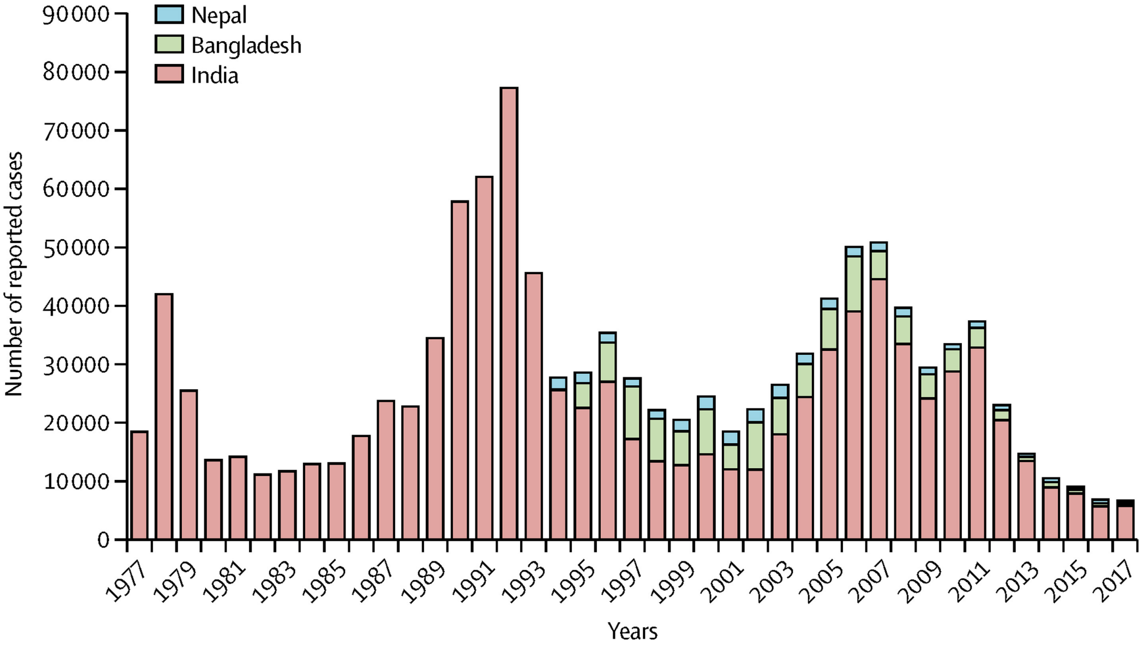


Figure 1.1.3: Incidence of visceral leishmaniasis in the Indian subcontinent from 1977-2017, demonstrating approximately 15-year cycles. Reprinted from The Lancet, Vol 392 / Issue 10151, Sakib Burza, Prof Simon L Croft, Prof Marleen Boelaert, Leishmaniasis, Pages No. 951-970, Copyright (2018), with permission from Elsevier. [33]

isolation from the vector and effective treatment is therefore the only attested method of breaking transmission.

Initial symptoms consist of fever and swelling of the liver and/or spleen, along with other non-specific symptoms common to other diseases endemic to the region, therefore obtaining a correct diagnosis can be challenging and time consuming. Current infection can only be confirmed by splenic or bone marrow aspiration, highly invasive procedures that carry substantial risk. It is therefore standard practice to use a rapid antigen test for diagnosis in combination with clinical diagnosis of symptoms. However, the non-specific and often non-debilitating nature of early symptoms mean that it can take a long time for a person to seek care and be referred for the appropriate test. This makes surveillance of VL particularly difficult as the timing of infection for a diagnosed case is very uncertain.

An estimated 10-20% of VL cases in the South Asian region go on to develop a secondary condition known as Post-Kala Azar Dermal Leishmaniasis (PKDL) up to several years following resolution of the initial infection. The condition manifests as rashes or lesions on the skin which are associated with stigma but the patient otherwise feels well, meaning that care-seeking and surveillance are poorer than for VL. It has been demonstrated that cases of PKDL are infectious to sandflies and hence can contribute to transmission of *L. donovani*, creating a further hurdle to achieving elimination [41, 42, 43].

Despite the overall decrease in incidence of VL, there is considerable heterogeneity between blocks which raises the need for a more targeted approach; the finite resources available must be distributed efficiently in order to continue progress towards elimination. Outbreaks across clusters of villages continue to occur [44] and history has shown these have the potential to develop into large epidemics [45, 46, 47]. Hence, it is important that localised pockets of incidence at a sub-block level are not overlooked. Moreover, there is evidence that symptomatic infection may not develop until several years after initial exposure [48], hence assumptions of transmission potential based purely on recent incidence may be insufficient.

A substantial limitation of the programme is that implementation of these efforts is not uniform across the region but varies according to the perception of some areas as non-endemic or "low-risk". A study in Nepal compared samples of districts included and excluded from the national control programme and found increased delays in care-seeking among patients in non-programme districts [49]. In Bihar there is also widespread use of private and informal health practitioners which, particularly in areas with low awareness due to lack of recent incidence, can cause additional delays in diagnosis and hence extend the period for potential further transmission [50]. A study in Vaishali district, Bihar, [40] demonstrated the strength of a combined, best practice approach to disease control and suggests the need to extend active efforts of vector control, case detection and community engagement to non-endemic but high-risk villages peripheral to hot spot areas. The authors concede, however, that there are substantial logistic and economic barriers to applying this intensive approach across all districts. A solution could be to identify villages which have not been reporting cases themselves but whose environment and demographics would be conducive to transmission if exposed, and specifically target these with increased surveillance.

The COVID-19 pandemic resulted in a disruption of control efforts, upheaval of Bihar's large population of migrant workers and increased economic vulnerability in the state [51]. These factors could have amplified gaps in surveillance and set back the elimination programme substantially [52, 53], potentially allowing escalation of transmission in highly endemic areas and re-establishment of the disease in previously non-endemic areas. Although the 2020 target has been a crucial motivating force for the progress made during the last decade, there is debate as to whether an arbitrary threshold on incidence is a relevant metric for measuring the success of the programme during these final stages.

A revised roadmap published in 2021 now aims for EPHP by 2030 and continues to focus on reducing burden rather than interrupting transmission, with additional targets of less than 1% case-fatality rate and 100% of PKDL cases detected and treated. Despite falling numbers of cases, the past two years have seen an apparent rise in the number of officially-reported deaths, from less than ten per year between 2014 and 2020 to nearly thirty in 2021 and 2022 [36]. This could potentially reflect a larger underlying burden of infection than is being detected through official channels and brings into question the representativeness of the available data. Defining how to *verify* the achievement of these targets therefore poses a substantial challenge that is yet to be addressed.

1.1.3 Emergence of COVID-19 in England

The novel coronavirus SARS-CoV-2 was first formally identified in early January 2020, following a cluster of unexplained cases of pneumonia in Wuhan, China. As a wave of COVID-19 spread through the city and its surrounding provinces, international travel quickly seeded epidemics across the rest of the world and induced the declaration of a global pandemic by the World Health Organisation by the 11th of March. As of November 2022, over 500 million cases and 6.5 million deaths have been reported worldwide [54], of which approximately 22 million and 180,000 occurred in the UK [55].

SARS-CoV-2 is highly infectious through contact and respiration, hence the density and connectivity within most countries' populations allowed rapid escalation of each epidemic. This contributed to strong patterns of spatial correlation in cases of COVID-19 as nearby communities interact with each other, allowing infection to radiate out from initial seeding locations. On top of that, certain common characteristics known to be clustered spatially within the population - such as age, comorbidities and ethnicity - stood out amongst early fatalities and have since been deemed important risk factors for severe symptomatic infection [56]. The result - evident on the scale of England's local authorities - was visible geographic disparity in the burden of disease and motivated a regional, rather than national, approach to interventions. However, the definition of area-specific risk levels was divisive, triggering substantial debate over what regional implementation of restrictions was appropriate for mitigating risk, without disproportionately debilitating certain parts of the population.

A defendable quantification of area-specific risk depends on an understanding of underlying patterns of transmission, which must be ascertained from the available surveillance data. Though useful for real-time analysis, confirmed cases of COVID-19 are not a good indicator of total incidence of infection due to the complex processes behind detection and testing. Similar to the declining completeness of detection in an elimination setting, this can often be an issue during the emergence of a novel disease or resurgence (perhaps in a novel setting), where the necessary infrastructure is not in place to keep up with rapidly accelerating incidence.

Test availability and policy surrounding who has access to tests varied substantially over the course of the epidemic and regionally across the UK [57]. COVID-19-related deaths, including where the virus is recorded as a cause of death or mentioned on the death certificate, could be considered a more consistently measured metric from which to understand patterns of transmission. The question of how best to measure the extent of ongoing transmission became of particular interest as focus moved from emergency mitigation and damage limitation, to sustained control and prevention of further waves of a similarly unmanageable scale. There was - at the time - a need to quantify risk on a regional scale in order for society to return to a 'normal' level of functionality while maintaining manageable levels of infection.

1.2 Aims and Objectives

This thesis explores a range of approaches to infer the underlying spatial and temporal distributions of these two diseases at opposite ends of the timeline, using data collected through routine surveillance. Similarities between the settings of emergence and elimination are drawn, with respect to the challenges of surveillance and of interpreting the resulting data for decision making.

Overall research aims

- To apply statistical modelling techniques to explain the spatial and temporal patterns of disease incidence, in the contexts of emergence and elimination.
- To explore the disconnect between observable indicators of transmission (reported cases) and the true underlying process, considering the implications of this for policy-making at a regional level.

Objectives

1. To apply standard regression methods to produce short-term forecasts of VL diagnoses at the block level, identifying the strengths and limitations of the approach under the particular challenges of this elimination setting.

2. To assess the added value of applying models to high-resolution village-level data compared to down-scaling from the block level predictions for guiding interventions and evaluating progress of the elimination programme.

3. To characterise the spatial distribution of diagnostic delays for VL, assessing individual-level risk factors in order to better understand reasons for delay and considering the contribution of diagnostic delays to risk of resurgence.

4. To compare the spatial distribution of deaths given local population vulnerabilities with that of confirmed cases, to understand how biases in case ascertainment could have influenced perception of risk during the escalation of the COVID-19 epidemic in England.

1.3 Outline of thesis

An outline of the remainder of the thesis is as follows:

- Chapter 2 introduces the methods applied in addressing each of the four stated objectives.

- Chapter 3 presents an analysis of block-level VL incidence between 2013 and 2018, with a view to making short-term forecasts to support logistics planning.

- Chapter 4 explores the distribution of observed disease at a *village* level, and investigates whether inference from block-level data can be extrapolated to this finer scale using a statistical downscaling approach.

- Chapter 5 aims to improve understanding of variation in the strength of surveillance by exploring how individual and village-level characteristics of VL cases diagnosed during 2018 are associated with excessive delays to diagnosis, characterising the structure of residual variation with respect to the GPS location of each case.

- Chapter 6 navigates similar challenges around the imperfect observation of disease incidence but in the contrasting setting of the COVID-19 epidemic in England, making use of multiple data sources to interrogate the biases in observation and how these may have varied between local authorities.

- Chapter 7 finally discusses conclusions that may be drawn across all previous chapters, overall limitations and potential future avenues of research.

Appendices to each of the four analysis chapters are included at the end of the thesis.

Chapter 2

Methods

This chapter introduces some statistical considerations that are relevant to the stated objectives, and explains key methodological concepts which will be applied throughout the remainder of the thesis.

2.1 Spatial and temporal structures in disease surveillance data

Epidemiologists and public health professionals will be familiar with disease surveillance data reported over time. For an infectious disease in particular, the number of cases today will be dependent on the number of cases previously, and epidemic models use this dependence to project the trajectory of cases as an infection spreads through a population. Temporal information about a case (from initial exposure to onset of symptoms, reporting and resolution) can give us insight into the transmission process, underlying risk factors (e.g. relating to seasonality) and the efficiency of the surveillance system.

Disease surveillance data may also be attributed with different types of spatial information. If the data exhibit spatial auto-correlation - in that observations closer in space are more similar than observations further apart - then this spatial information can be exploited to pool information between nearby observations and increase the precision of estimation. Models of spatial processes can be considered to fall into three classes:

(A) Areal

(B) Geostatistical

(C) Point process

Most commonly, routinely-collected disease surveillance data are available as aggregated values (e.g. case counts) at the level of some discrete administrative unit, for example country, state, province or postcode area (Figure 2.1.1 (A)). This is usually referred to as *areal* data, and modelled by a discrete process defining how values in each areas are related to each other. The structure of fixed geographic boundaries between areas can be summarised in the form of an *adjacency* matrix, with non-zero elements if two areas share a boundary and zero otherwise. The simplest case would be to identify direct neighbours with a value of one, but greater detail of the neighbourhood structure may be incorporated by assigning non-zero values to neighbours of higher orders with decreasing weights.

Alternatively, an observed value or individual case may be attributed to a specific point location on the globe, defined by a set of coordinates (Figure 2.1.1 (B) and (C)). A geostatistical process is described by measurements taken at specific locations (e.g. the results of a sero-prevalence survey across sampled villages, or measurements of air quality at monitoring stations), while a point process consists of locations of individual events occurring in space (e.g. cases of disease or occurrence of outbreaks across a region). Both of these types of data reflect the underlying spatial process on a *continuous* scale and hence can allow for highly localised inference.

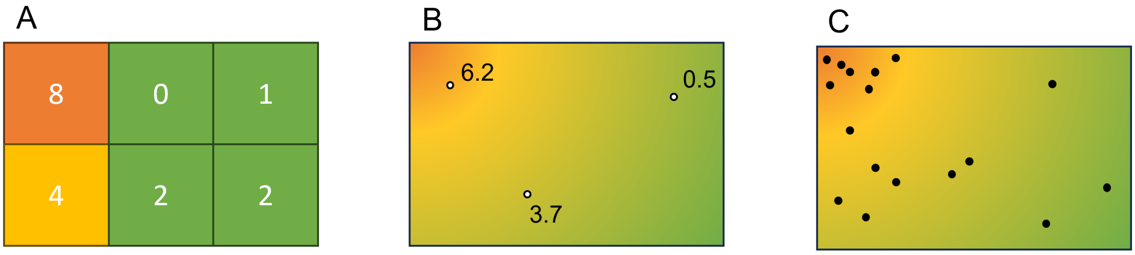


Figure 2.1.1: Types of spatial data. (A) Areal data consisting of counts in discrete, contiguous regions, (B) Geostatistical data consisting of measurements taken at fixed point locations, and (C) point process data consisting of the observed locations of events.

To construct such models, assumptions are usually required about the structure of correlation between observations, such as the distance in space (and/or time) to which the correlation extends (the "lag") and the functional form with which it decays. These choices may be informed by the data, for example through preliminary exploration of (partial) auto-correlation functions at different lags in time [58], or semi-variograms at different distances in space [59]. They may also be estimated during the model fitting process, given some realistic constraints. An optimal specification may not always be identifiable, however, especially when considering correlation in more than one dimension. The potential sensitivity to these assumptions is therefore something that needs consideration when attempting to model complex dependencies.

2.2 A frequentist model of spatio-temporal correlation: The endemic-epidemic framework

The endemic-epidemic model structure described by [60] incorporates information from nearby points in (discrete) space and time by conditioning the mean of a regression-type model on weighted sums of past observations across increasing orders of neighbours. This spatio-temporal structure is designed to capture the epidemic behaviour of infectious disease dynamics, in that the incidence of cases in one place increases the chance of subsequent incidence in nearby places. The author elsewhere demonstrates how it may be derived from traditional compartmental transmission modeling [61]. The model is fit to observed case counts per unit of space and time (a multivariate time series) using maximum likelihood estimation (or penalised maximum likelihood if random effects are included). Chapter 3 will go into greater detail about the model specification and how the distinction of its endemic and epidemic components is defined.

2.3 Bayesian modelling of spatio-temporal correlation

Complex correlation structures are arguably more naturally accommodated within a Bayesian framework, where assumptions regarding the dependence between observations in different space or space-time units may be defined via the priors of random effects. The specified prior distributions are combined with the observed data to reach a posterior distribution, which is often estimated via simulation. A standard simulation approach is Markov Chain Monte Carlo (MCMC), for which a Markov chain is constructed with the posterior as its stationary distribution, and simulations of this are run until the chain is assumed to have converged to that target. When convergence is reached the simulations drawn may be summarised in whatever way is required to describe the posterior distribution, for example with respect to parameters or fitted values.

In theory, a prior correlation structure of arbitrary complexity may be specified and the resulting posterior estimated in this way, but the time taken to reach convergence can easily be prohibitively long (if convergence is reached at all). A large number of parameters (and the potentially complex dependencies between them) will likely create a posterior distribution that is difficult to explore fully using the standard MCMC algorithm. The size of the data to which the model is being fit is also a limiting factor, as the algorithm requires computation of the likelihood for each sample. Adjustments to the computation approach and sampling methodology can be made to speed up the sampling process, such as parallelization and adaptive algorithms that more efficiently cover the parameter space (e.g. Hamiltonian Monte Carlo [62]). However, an alternative is to use an approximation to the posterior which may be calculated deterministically, avoiding the need for time-consuming simulations. A broad class of hierarchical models which are often relevant for spatio-temporal applications may be estimated via a particularly efficient approximation known as the Integrated Nested Laplace Approximation or INLA [63].

2.3.1 The Integrated Nested Laplace Approximation (INLA)

Briefly, the INLA approach can be applied when the model of interest falls into the class of latent Gaussian models. This requires that the unobserved "latent field" of parameters (including for example, linear covariate coefficients and the parameters of functions such as splines or structured/unstructured random effects) jointly follows a multivariate Gaussian distribution, which is conveniently an appropriate assumption for most models relevant to disease surveillance applications. Much of the speed of the approach comes from latent field exhibiting the Markov property of conditional independence, such that the majority of the elements of the covariance matrix are zero (i.e. the matrix is sparse). This property makes it very computationally efficient to invert in order to obtain the precision matrix. The distribution around the mode of the marginal posterior for each parameter is then approximated by a Gaussian, defined from the first three terms of its Taylor expansion. This is known as the simplified Laplace approximation of the distribution (SLA).

The Kullback-Liebler divergence [64] is reported for each marginal posterior to show the difference between the Gaussian and SLA approximations which, if small, indicates that the posterior is well approximated by a Gaussian and therefore the more computationally intensive *full* Laplace approximation is not needed. Although in theory this approach provides an approximation to the posterior, in practice it has been demonstrated to be no less accurate than MCMC with finite sampling [65].

2.3.2 Penalised Complexity Priors

The speed gains provided by the INLA approximation allow us to fit highly complex spatial models that would otherwise be impractical. However, it is nevertheless conservative to give prior weight to the assumption that such complexity is not needed. We can do this by penalising the model's complexity through the prior distributions [66]. Prior weight is distributed to give preference to the value of a parameter which corresponds to the simplest model, rather than to values which would yield more complex models. This means that complex elements only contribute to the final model fit if there is substantial evidence in the data to support it. Penalised complexity priors are often defined according to an upper (or lower) bound C with the form

$$P[\theta > C] = a$$

where a is a small probability. For example, the value of a variance parameter which yields the simplest model would be 0, i.e. the effect is constant. We may assume based on the context and scale of what we're modelling that the effect would be unlikely to have a variance greater than C. A penalised complexity prior on this parameter would give lesser weight to values far from 0, and especially to values higher than C. All Bayesian models presented in this thesis are fit using priors that penalise complexity, where possible.

2.4 The Besag-York-Mollié Model

The Besag-York-Mollie (BYM) model [67] describes *discrete* spatial data through a combination of random effects at the area level, accounting for spatial dependence and purely random residual variation between areas. Specifically, a BYM model across a set of areas i is of the form

$$\eta_i = \beta_0 + \beta \mathbf{x_i} + u_i + v_i$$

where β_0 is the overall mean and $\mathbf{x_i}$ a vector of fixed effects with corresponding coefficients β. The u_i are spatially correlated random effects, normally distributed with mean defined by the average value of its neighbours and variance decreasing with the number of neighbours, d_i. For neighbouring areas j of i (denoted $i \sim j$),

$$u_i \sim N\left(\frac{1}{d_i}\sum_{i \sim j} u_i, \frac{\sigma_u^2}{d_i}\right)$$

The random effects v_i are also normally distributed but independent, with zero mean and variance σ_v^2

$$v_i \sim N(0, \sigma_v^2)$$

The Besag-York-Mollié model may be re-parameterised in order for it to be specified with priors that penalise its complexity [68], creating a default assumption that observations in each area are independent and not correlated with their neighbours. As opposed to a direct sum of $u_i + v_i$, the formula is rearranged to create a *weighted* sum between the spatially-structured and unstructured area-level effects, with a mixing parameter $\phi \in [0, 1]$ dictating the relative contribution of each. Specifically,

$$\left(\sqrt{\phi}u_i + \sqrt{1-\phi}v_i\right)\frac{1}{\sqrt{\tau}}$$

The mixing parameter ϕ can be given a prior which puts more weight on the purely random, unstructured effect, only incorporating strong spatial dependence when the data support it. The remaining parameter is the marginal precision $\tau > 0$ of the combined spatial effect, which is assigned a penalised complexity prior with little weight on values above a specified upper bound. This puts most weight on small values of τ that correspond to large variability between areas.

2.5 The Stochastic Partial Differential Equation Model

When data are indexed *continuously* in space (e.g. by latitude and longitude), the underlying process can instead be modelled with a Gaussian random field (GRF). A GRF is a continuous random process in two dimensions where the value at any finite subset of points follows a multivariate normal (MVN) distribution. The GRF is defined by a range and a variance parameter, dictating the distance to which correlation between two points is evident, and the variance dictates how much (on average) the value of the field differs from its overall mean, respectively. These can both be assigned penalised complexity priors, giving preference for a constant field with infinite range and zero variation.

The INLA approach does not directly apply to a continuously-indexed GRF as the Markov property (on which the approach depends) does not naturally hold. However, if a particular covariance function is assumed (the Matérn covariance function), the process can be expressed as the solution to a set of stochastic partial differential equations (SPDE) [69]. By approximating continuous space with a triangular partition or "mesh", this solution can be represented as a weighted sum across the finite set of vertices (nodes), with piece-wise linear basis functions defined to translate the solution at the nodes to any point within the triangle. The Markov property can then be assumed for this *discretised* process - yielding a GMRF - and hence may be fit using INLA.

2.6 Disaggregation regression

Disaggregation regression exploits associations between disease incidence and environmental and/or climatic conditions which can be measured at a fine spatial resolution (for example via satellite imagery), to infer local-level variation from spatially-aggregated data. The naive application of covariate relationships estimated on an area level to a finer scale is known as "ecological fallacy" and usually leads to incorrect conclusions; relationships measured *across* groups may easily be contradicted or even reversed when assessed *within* groups. Disaggregation regression avoids this fallacy by defining a model on a pixel scale then aggregating the counts across pixels to a likelihood on the area level, weighted by population count. A continuous spatial field across pixels and independent area-level random effects are used to model any spatial structure unexplained by the given covariates. Specifically, for incidence rate r in pixel j in block i with location s_{ij},

$$\log(r_{ij}) = \beta_0 + \beta X_{ij} + u_i + \epsilon_{ij}$$

where X_{ij} are covariate values for pixel j in block i, u_i is a block-level IID random effect with precision τ and ϵ_{ij} is a spatially-correlated noise term, modelled as a Gaussian random field across pixels with Matern covariance structure parameterised by range ρ and scale σ. Assuming case counts per pixel to be Poisson-distributed and conditionally independent given this defined risk surface, the observed total case count in block i also follows a Poisson distribution, with mean obtained by aggregating r_{ij} via a weighted raster a_{ij} (i.e. the population raster) over all pixels j in i

$$y_i \sim Pois\left(\sum_{j=1}^{N_i} a_{ij} r_{ij}\right)$$

The likelihood of the observed sums of pixel counts across areas can be computed according to this distribution, and posterior marginals for the model parameters $\beta_0, \beta, \tau, \rho$ and σ - given the data - are estimated using a Laplace approximation [70, 71].

2.7 Application of these methods in the thesis

Chapter 3 evaluates the endemic-epidemic framework for producing short-term forecasts of VL diagnoses, exploring how the incremental introducton of complexity to the different model components affects fit and out-of-sample predictive power. In Chapter 4, I investigate how well the disaggregation approach can replicate village level disease burden from block level incidence and a set of environmental covariates, comparing it against a robust statistical method commonly used for fine-scale mapping (random forest). Chapter 5 applies the SPDE method to model a continuous pattern of spatial variation in VL diagnosis delays between villages, exploring the relative contribution of individual and village level covariates in explaining this residual spatial pattern. Finally, in Chapter 6, I use the BYM model alongside temporally-dependent random effects to capture variation in reported cases and deaths between UK local authorities over the course of the first COVID-19 epidemic wave. Each chapter will describe in further detail how the specific modelling approach was applied and evaluated, in order to address the particular question of interest.

Chapter 3

A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India

At the time of writing, the WHO target for VL in India was to reduce incidence levels to less than one case per ten thousand population at the block level. This policy was the motivation for this piece of work, with an aim to explore the feasibility of projecting recent trends in incidence forward at short time horizons in order to better monitor the progress of each block towards this target. Routine surveillance monitors incidence rates per block in order to identify high-, moderate- or low-endemicity and determine intervention plans. Block-level incidence was also used to monitor regional elimination status with respect to the number of blocks above or below the target. This approach does not take into account that risk is shared across administrative boundaries and, as such, a block defined as low- or non-endemic does not imply that it is at low risk of transmission or reintroduction in the future. This analysis used an existing framework to model the dependence of monthly, block-level incidence rates on the recent past, both within the same block and across its neighbours.

This paper was submitted to PLOS Neglected Tropical Diseases in October 2019 and published in July 2020.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1600363	**Title**	Ms
First Name(s)	Emily Sara		
Surname/Family Name	Nightingale		
Thesis Title	Spatio-Temporal Patterns and Surveillance of Infectious Disease During Emergence and Elimination		
Primary Supervisor	Prof Graham Medley		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Plos Neglected Tropical Diseases		
When was the work published?	July 2020		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**No**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I extracted and aggregated the data from the routine surveillance database, conducted the analysis, created the visualisations and drafted the manuscript. Graham Medley and Lloyd Chapman supported in conceptualising the analysis and interpreting the results. Other co-authors contributed to my understanding of the data, methodology and policy implications. I created a public repository for the code underlying this published analysis (https://github.com/esnightingale/VL_prediction_paper). I also published a condensed version which only implements the final model to generate predictions from a raw linelist of diagnoses, for use as a more general tool (https://github.com/esnightingale/vl-short-term-prediction).

SECTION E

Student Signature	Emily Nightingale
Date	12 December 2022

Supervisor Signature	Graham Medley
Date	12 December 2022

PLOS NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India

Emily S. Nightingale[1]*, Lloyd A. C. Chapman[1], Sridhar Srikantiah[2], Swaminathan Subramanian[3], Purushothaman Jambulingam[3], Johannes Bracher[4], Mary M. Cameron[5], Graham F. Medley[1]

1 Centre for Mathematical Modelling of Infectious Disease and Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, United Kingdom, **2** CARE India, Patna, Bihar, India, **3** Vector Control Research Centre, Puducherry, Chennai, India, **4** Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland, **5** Department of Disease Control, London School of Hygiene and Tropical Medicine, London, United Kingdom

* Emily.Nightingale@lshtm.ac.uk

OPEN ACCESS

Citation: Nightingale ES, Chapman LAC, Srikantiah S, Subramanian S, Jambulingam P, Bracher J, et al. (2020) A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India. PLoS Negl Trop Dis 14(7): e0008422. https://doi.org/10.1371/journal.pntd.0008422

Editor: Richard Reithinger, RTI International, UNITED STATES

Received: October 24, 2019

Accepted: May 24, 2020

Published: July 9, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pntd.0008422

Data Availability Statement: The data from the Kala-Azar Management Information System (KAMIS) underlying the results in this manuscript cannot be shared publicly because of patient

Abstract

Background

The elimination programme for visceral leishmaniasis (VL) in India has seen great progress, with total cases decreasing by over 80% since 2010 and many blocks now reporting zero cases from year to year. Prompt diagnosis and treatment is critical to continue progress and avoid epidemics in the increasingly susceptible population. Short-term forecasts could be used to highlight anomalies in incidence and support health service logistics. The model which best fits the data is not necessarily most useful for prediction, yet little empirical work has been done to investigate the balance between fit and predictive performance.

Methodology/Principal findings

We developed statistical models of monthly VL case counts at block level. By evaluating a set of randomly-generated models, we found that fit and one-month-ahead prediction were strongly correlated and that rolling updates to model parameters as data accrued were not crucial for accurate prediction. The final model incorporated auto-regression over four months, spatial correlation between neighbouring blocks, and seasonality. Ninety-four percent of 10-90% prediction intervals from this model captured the observed count during a 24-month test period. Comparison of one-, three- and four-month-ahead predictions from the final model fit demonstrated that a longer time horizon yielded only a small sacrifice in predictive power for the vast majority of blocks.

Conclusions/Significance

The model developed is informed by routinely-collected surveillance data as it accumulates, and predictions are sufficiently accurate and precise to be useful. Such forecasts could, for example, be used to guide stock requirements for rapid diagnostic tests and drugs. More

confidentiality and privacy concerns. KA-MIS data are property of the National Vector-Borne Disease Control Programme (NVBDCP, Govt of India), and are managed by CARE India. The data are available from NVBDCP (nvbdcp-mohfw@nic.in) for researchers who meet the criteria for access to confidential data. A simulated version of the dataset used in this manuscript is available at https://github.com/esnightingale/VL_prediction_paper.

Funding: This study was supported by the Bill and Melinda Gates Foundation (https://www.gatesfoundation.org/) through the SPEAK India consortium [OPP1183986] (ESN, LACC, SS, SS, PJ, MMC, GFM). The views, opinions, assumptions or any other information set out in this article are solely those of the authors and should not be attributed to the funders or any person connected with the funders. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

comprehensive data on factors thought to influence geographic variation in VL burden could be incorporated, and might better explain the heterogeneity between blocks and improve uniformity of predictive performance. Integration of the approach in the management of the VL programme would be an important step to ensuring continued successful control.

Author summary

This paper demonstrates a statistical modelling approach for forecasting of monthly visceral leishmaniasis (VL) incidence at block level in India, which could be used to tailor control efforts according to local estimates and monitor deviations from the currently decreasing trend. By fitting a variety of models to four years of historical data and assessing predictions within a further 24-month test period, we found that the model which best fit the observed data also showed the best predictive performance, and predictive accuracy was maintained when making rolling predictions up to four months ahead of the observed data. Since there is a two-month delay between reporting and processing of the data, predictive power more than three months ahead of current data is crucial to make forecasts which can feasibly be acted upon. Some heterogeneity remains in predictive power across the study region which could potentially be improved using unit-specific data on factors believed to be associated with reported VL incidence (e.g. age distribution, socio-economic status and climate).

Introduction

Visceral leishmaniasis in India

The short-term forecasting of diseases targeted for elimination can be a important management tool. Visceral leishmaniasis (VL) is the acute disease caused by *Leishmania donovani*, which is transmitted through infected female *Phlebotomus argentipes* sandflies. In India, the burden of disease is largely contained within the four northeastern states of Bihar, Jharkhand, Uttar Pradesh and West Bengal, with the rural state of Bihar most broadly affected [1–3].

Incidence of VL in India has decreased substantially since the initiation of the regional Kala-Azar Elimination Programme (KEP), which aims to tackle the disease across the Indian subcontinent through enhanced case detection and treatment and reduction of vector density [4]. As a result, reported cases have fallen from 29,000 in 2010 to less than 5,000 in 2018 [3, 4]. The overall target of the programme is to reduce incidence to less than 1 case/10,000 people/year within each "block". Blocks are administrative sub-divisions of a district with population sizes varying from thirty thousand to several million, depending on geographic area and the proportion of urban and rural habitation. As a consequence, the target equates to an absolute total of between three and two hundred cases per year. To support the elimination effort, data are reported to a central repository (Kala-Azar Management Information System, KA-MIS) to construct line lists including the date and location of every diagnosed case.

Despite the overall decrease in incidence, there is considerable heterogeneity between blocks (Fig 1). In some blocks cases are now few and far between, while others remain substantially affected from year to year. The combination of the decrease and the heterogeneity raises the need for a more targeted approach; the finite resources available must be distributed efficiently to continue progress. Additionally, history has shown that VL has the potential to develop into large epidemics [5–7] and hence it is important that localised pockets of incidence

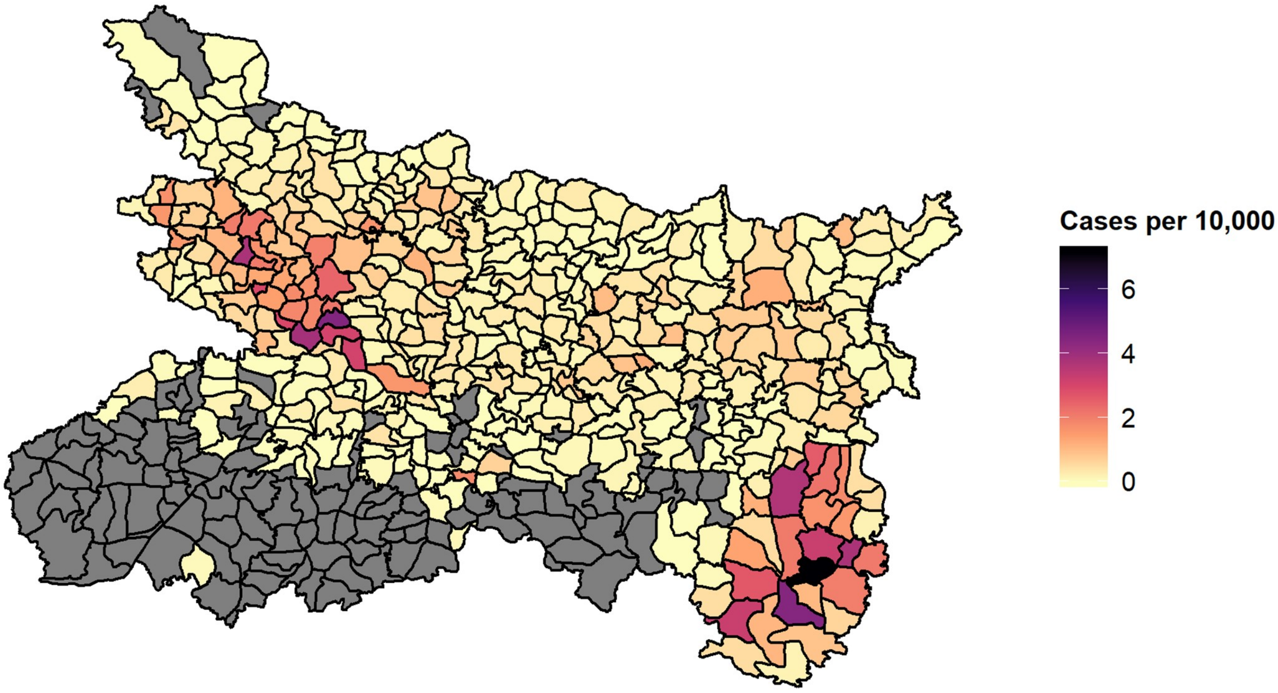


Fig 1. Estimated incidence per 10,000 population per block in 2018, for Bihar and the four endemic districts of Jharkhand (Dumka, Godda, Sahibganj and Pakur). Incidence is estimated according to reported cases in KA-MIS with diagnosis date in between 01/01/2018 and 31/12/2018 and block populations projected from the 2011 census according to decadal, block-level growth rates [9]. Black lines indicate block boundaries. The affected blocks of Jharkhand on average have much higher incidence than Bihar and can be seen in the bottom right of the map. Blocks marked grey had no reported cases during the study period.

https://doi.org/10.1371/journal.pntd.0008422.g001

are not overlooked. Intervention when incidence is low is required to prevent the trajectory from turning upwards again, as cycles of VL incidence appear to occur with a frequency of 10-20 years [8].

The primary aim of this paper is to ascertain the potential utility of predictions based solely on routinely-collected surveillance data, within a ready-made, rapid and relatively easy-to-use framework. Such predictions could serve two purposes; firstly to support logistics, for example in setting minimum stock levels of rapid diagnostic tests and drugs, and secondly to provide an early warning if the number of cases starts to resurge. For this modelling framework to be useful to the elimination programme, it is essential that its predictions are sufficiently accurate. Hence we make predictive accuracy of the forecasting approach the focus of the model selection.

Forecasting and spatio-temporal analysis

There have been many attempts at forecasting the various forms of leishmaniasis across the three affected continents. Lewnard et al. (2014) [10] employ a seasonal ARIMA (Auto-Regressive Integrated Moving Average) model to predict cutaneous leishmaniasis in Brazil, incorporating meterological data and evaluating one, two and three month ahead forecasts. More recently, Li et al. used an extended ARIMA model to predict incidence in Kashgar prefecture, China [11]. However, neither of these attempts to capture spatial variation. Epidemiological data, in particular regarding infectious disease, are often *both* temporally and spatially correlated. That is to say, as well as incidence at one point in time being related to incidence in the

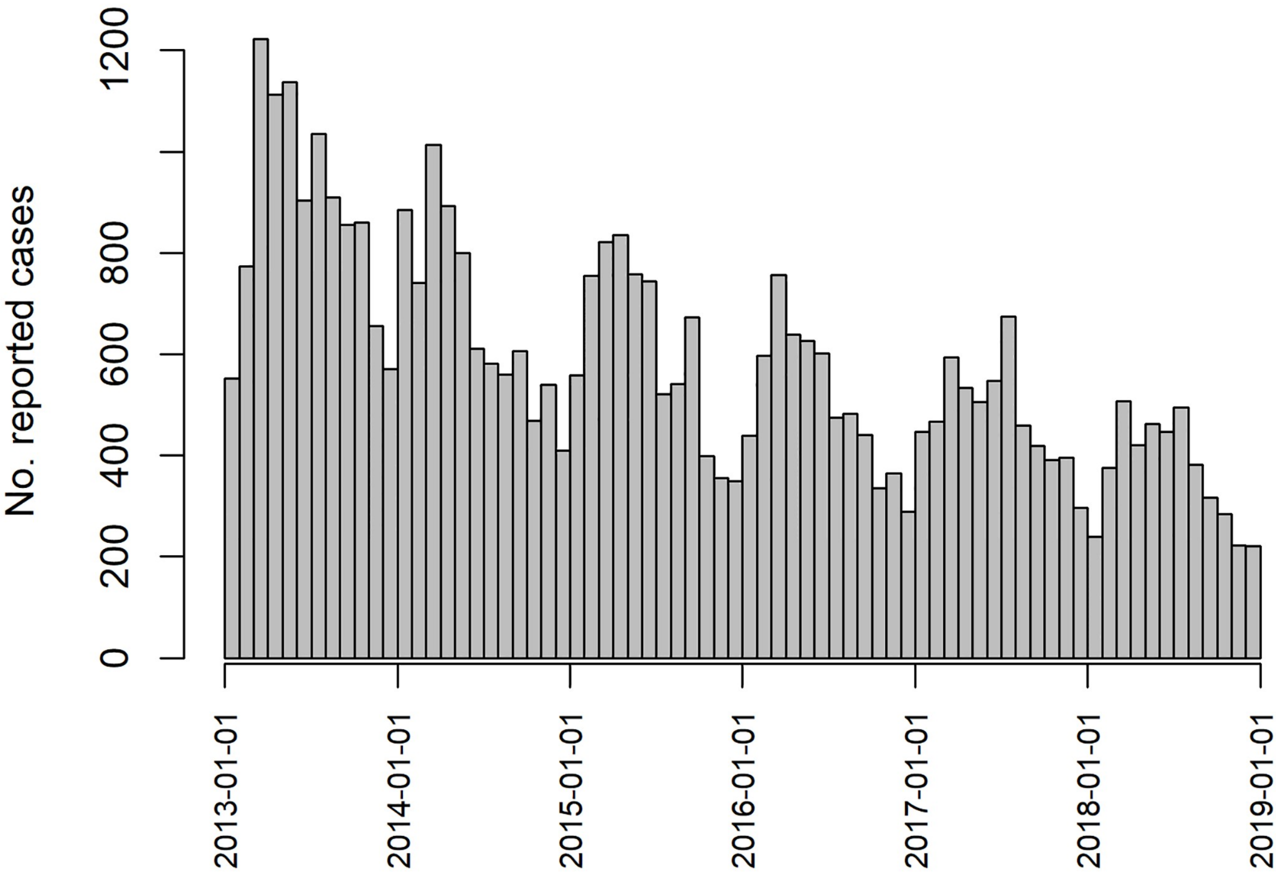


Fig 2. Total monthly reported cases across the study region. The annual cycle (peaking between January and April) and overall decreasing trend are clear at this aggregate level.

https://doi.org/10.1371/journal.pntd.0008422.g002

past, incidence in one area is also related to incidence in nearby areas. Mapping reported VL incidence in India at the block level demonstrates the presence of spatial correlation (Fig 1), with concentrated regions of high incidence appearing in West Bihar and Jharkhand. This could be due to similar geographic and demographic characteristics of neighbouring blocks, or the spread of infection by regular population movement. The seasonal cycle of incidence and overall decreasing trend (Fig 2) are clearly evident in aggregated case counts.

Several statistical approaches have been developed to model count data in space and time. These methods have been largely developed and used for understanding the drivers of patterns, often incorporating additional covariate information describing climate, geography or demography [12, 13] Dewan et al. [14] employ scanning techniques for a regional analysis solely of case data, but do not utilise the approach for prediction. Paixão-Seva et al. (2017) [15] simultaneously model the infected human, vector and dog populations in relation to landscape, climatic and economic factors, and in particular use proximity to a highway and gas pipeline as indicators of human movement. Where aetiology is not the focus, analyses often incorporate GPS locations of cases to identify hotspots and predict disease spread at a local village or household level [16], or across health facilities [17].

In the case of VL on the Indian subcontinent, environmental data are difficult to obtain in real-time at a sufficient spatial and temporal scale for forecasting purposes, and GPS data have not been routinely or uniformly collected across the affected region. As such, statistical approaches to spatio-temporal analysis have been broadly limited to specific study regions within which additional data were collected [18]. Predictions on a regional level have so far been the remit of transmission dynamic modelling [19]. We aim to make use of the reliable and near-complete date and area data within the KA-MIS system, for the whole state of Bihar and the affected region of Jharkhand, to understand how well future cases could be predicted solely from the surveillance data of previous cases. As far as we are aware, no previous attempt has been made to forecast VL at this spatial scale and with this level of coverage for the Indian endemic region.

Often the model which best fits observed data is selected for forecasting, yet goodness of fit does not guarantee predictive power. We therefore also investigate the relationship between the fit and predictive power.

Model framework

A natural modelling approach is to consider the cases in each month in each block as a function of cases in the previous month and in neighbouring blocks. A model framework developed in [20, 21] has been applied previously for modelling cutaneous leishmaniasis in Afghanistan [22]. This framework decomposes the distribution of counts at each point in space and time into three components (auto-regressive, neighbourhood and endemic):

- **Auto-regressive (AR)** *The contribution of previous incidence in the same block to current incidence. A choice must be made about time period of previous incidence considered (i.e. the number of months).*
- **Neighbourhood (NE)** *The contribution of previous incidence in surrounding blocks to current incidence. A choice must be made about both the time period and spatial extent considered (i.e. neighbours, neighbours of neighbours etc.), with indirect neighbours assigned decaying weights, for example, according to a power law.*
- **Endemic (END)** *A function describing the intrinsic incidence related to block factors (such as geography or demography) or seasonality.*

The sum of these components forms the mean structure for a negative binomial distribution used to model the count in each block and month. The epidemic component consists of both auto-regression and spatial/spatio-temporal regression. The maximum distance in space or time at which we assume one block-and-month count affects another is referred to as the maximum spatial or temporal *lag*. The endemic component attempts to explain any remaining variation, potentially due to overall temporal trends, population size and other unit-specific factors.

In addition to the genuine epidemiology of VL, there is an intermediary process of detection and reporting which contributes to the distribution of case counts. A new case in a previously unaffected area triggers active case detection (ACD) which continues for twelve months, therefore contributing to the pattern of temporal correlation. In other words, one case is likely to be promptly followed by more cases—not only because of transmission but also as a result of increased, localised detection effort. We therefore explored a flexible, distributed lag structure [23] which extends the range of spatio-temporal interaction by allowing incidence over multiple previous months to contribute to both the auto-regressive and spatial elements.
The selection of an optimal lag length has been investigated for distributed lag models in one dimension (i.e. time alone) [24], but the impact of introducing a spatial component has not been thoroughly discussed. A strong interdependence between the autoregressive and neighbourhood components is introduced by simultaneously incorporating past information from

the same block and the neighbourhood of that block in a distributed lag model; each block affects subsequent incidence in its neighbours, which in turn affects subsequent incidence in the original block. We apply a semi-systematic approach which attempts to optimise the temporal and spatial lags simultaneously such that one does not mask the effect of the other.

Evaluation of forecasts

The three components described in the previous section (Model Framework) have arbitrary complexity and lead to a large number of candidate models. A key issue is therefore to identify the best-fitting model, or a set of well-fitting models, and to assess to which degree good in-sample (or retrodiction) performance translates to out-of-sample forecasting performance. In-sample performance is widely assessed via the Akaike information criterion (AIC). The AIC balances the model fit and complexity, and has been recommended for model selection for prediction purposes [25]. To assess performance of probabilistic forecasts it is standard to use proper scoring rules [21, 26–29], which offer more detailed scrutiny of the prediction than measures of absolute or squared error (as used, for example, in [30]) by taking into account the whole predicted distribution. In fact, the ranked probability score (RPS) can be considered a generalisation of absolute error, to which it reduces if the forecast distribution consists of a single point. Proper scoring rules measure simultaneously the calibration and sharpness of forecast distributions; they capture the model's ability to predict both accurately and precisely but also to identify its own uncertainty in that prediction [28]. With a well-calibrated model the observed values should appear as having come from the predicted distribution at that point, and we want as precise or sharp a predicted distribution as possible while maintaining that calibration. In contrast, the mean absolute error for example only evaluates how well the central tendency of predictions aligns with the observations. We utilise the ranked probability score (RPS) [26] averaged over all predicted time points (502 blocks * 24 months, so 12048 test predictions), which for a predictive distribution P and an observation x is defined as

$$\overline{\mathrm{RPS}}(P,x) = \sum_{k=0}^{\infty}[F_P(k) - \mathbb{1}(x \leq k)]^2 \tag{1}$$

Here, F_P is the cumulative distribution function of P and 1 is the indicator function. The RPS thus compares the cumulative distribution function of P to that of an "ideal" forecast with all probability mass assigned to the observed outcome x. We use this score rather than the logarithmic score as it is considered more robust [31], and we wish to assign some credit to forecasts near the observed value. The score is negatively oriented, meaning that smaller values are better.

Calibration can in addition be assessed using probability integral transform (PIT) histograms. The PIT histogram shows the empirical distribution of $F_{P;i}(x_i)$ for a set of independent forecasts $i = 1, \ldots, I$. We here use an adapted version for count data suggested by Czado et al [26]. If the forecasts are calibrated, the histogram should be approximately uniform. U and inverse U-shaped PIT histograms indicate that the forecasts imply too little or too much variability, respectively.

A closely-related summary measure which is easy to communicate are empirical coverage probabilities [31]. We will provide coverage probabilities of central 50% and 80% prediction intervals (reaching from the 25% to 75% and 10% to the 90% quantiles of the predictive distribution, respectively). For a calibrated forecast, the empirical coverage probabilities should be close to the nominal levels. However, in the context of sparse, low counts the discreteness of the data often prevents achieving exactly the nominal coverage level. Prediction intervals can then either be slightly conservative (too high coverage), which is usually preferred in practice, or slightly liberal.

Our hypothesis is that models constructed with the *surveillance* framework to accommodate spatio-temporal correlation in disease incidence can provide significantly more accurate predictions (in terms of sharpness and calibration) than a purely parameter-driven (i.e. independent of history and spatial context) model with overall mean and linear time trend. Initially, we examine and discuss the relationship between model complexity, its ability to describe past data (i.e. its fit) and its ability to predict the next month. We then apply this understanding to select an optimal model for prediction with a semi-systematic approach, before comparing its predictive ability for different time horizons.

Materials and methods

Ethical approval

Ethical clearance was granted by the Observational/Interventions Research Ethics Committee at LSHTM (ref: 14674), subject to local approval. Local approval to use this data was granted by Dr Neeraj Dhingra, director of the National Vector Borne Disease Control Programme (GoI). Individual consent was not required as all data were analysed anonymously.

Data

Access to the KA-MIS database of VL cases was provided by the National Vector Borne Disease Control Programme (NVBDCP) and facilitated by CARE India. Individual case records were downloaded for Bihar and Jharkhand, restricted to diagnosis date between 01/01/2013 and 31/12/2018 and then aggregated by block and diagnosis month. This gave reported case counts for 441 blocks. The KA-MIS data were merged with data from the 2011 census [9] (compiled by CARE India) for the two states to produce the final data set, including endemic blocks which had no reported cases during the study period and hence did not appear in KA-MIS. Because we incorporate spatial correlation into the model, it is necessary to not have "holes" of missing data in the map. For individual blocks within the assumed "endemic" region without any reported cases in certain months, case counts were assumed to be "true zeros" since detection efforts should be consistent with the affected neighbouring blocks. The time series for these blocks were imputed with zeros and therefore contributed to the fit of the model. Four entire districts of Bihar, at the edge of the "endemic" region, (Gaya, Jamui, Kaimur and Rohtas) had no reported cases during the period, and were excluded from the analysis.

The final analysis data set included 502 blocks across 38 districts of Bihar and Jharkhand over 72 months.

Model structure

Due to considerable temporal variation in incidence within blocks, as a result of detection effort and cases arising in "clumps", the block-level monthly case counts are widely dispersed. A negative binomial distribution was therefore used to model the block-level case counts throughout.

All models fitted conform to the same negative binomial structure for case counts Y_{it} given previous incidence:

$$Y_{it} \mid \text{past} \sim \text{NegBin}(\mu_{it}, \psi_i) \tag{2}$$

$$\mu_{it} = \underbrace{\lambda_t \sum_{q=1}^{Q} u_q Y_{i,t-q}}_{\text{AR}} + \underbrace{\phi_t \sum_{j \neq 1} \sum_{q=1}^{Q} w_{ij} u_q Y_{j,t-q}}_{\text{NE}} + \underbrace{v_t e_{it}}_{\text{END}} . \tag{3}$$

where Y_{it} denotes the reported case count in block i in month t with population e_{it}, neighbourhood weights w_{ij} for neighbours j of block i, and overdispersion parameter $\psi_i > 0$ such that $\text{Var}(Y_{it}) = \mu_{it}(1+ \psi_i \mu_{it})$. Normalised weights u_q for distributed lags $q = 1, \ldots, Q$ are defined according to a scalar parameter p which is estimated from the data.

$$u_q^0 = p(1-p)^{q-1}, \quad u_q = \frac{u_q^0}{\sum_{q=1}^{Q} u_q^0} \tag{4}$$

The log-transformed parameter of each model component is then defined by a linear regression on any relevant covariates, $\mathbf{X}_{it}$; in this case we consider time with sine and cosine terms to replicate seasonal waves.

$$\log(\lambda_t) = \boldsymbol{\beta}^{\lambda} \mathbf{X}_{it}^{\lambda}, \tag{5}$$

$$\log(\phi_t) = \boldsymbol{\beta}^{\phi} \mathbf{X}_{it}^{\phi}, \tag{6}$$

$$\log(v_t) = \boldsymbol{\beta}^{v} \mathbf{X}_{it}^{v}, \tag{7}$$

where $\boldsymbol{\beta}$ are the regression coefficients.

All models were fit using the R package *surveillance* [32] and its extension *hhh4addon* [33] in R version 3.6.1 (2019-07-05) [34].

Investigating fit and prediction. Thirty random models were drawn from the set of possible formulations (where all three of the endemic-epidemic components are included in some form) and compared on the metrics of interest. This informed the subsequent selection process for the final prediction model.

Code used to produce the results in this paper is available from https://github.com/esnightingale/VL_prediction_paper, along with a simulated version of the dataset from the final selected model.

Model selection

During the selection process, all models were fit to the subset of months 5 to 48 in order to make comparisons between maximum temporal lags up to four months. The remaining 24 months were then predicted sequentially in a "one-step-ahead" (OSA) approach to assess predictive power (as was applied in [10]), either with rolling updates to the fit (incorporating each month's data into parameter estimates to predict the next) or without (using only the training set of data for all predictions) [22, 26]. The average RPS of these predictions served as the primary criteria for model selection, comparing between models of increasing complexity by permutation test with a significance cut-off at 0.001. At the same time, average RPS was compared to AIC from the model's training period fit to assess the relationship between fit to the "observed" data and future prediction.

The following elements were considered for inclusion in the model:

- Log of population density as a covariate in the endemic component, in place of population fraction offset.

- Seasonal variation and linear trend within the coefficients of all three components, serving to vary the relative strength of each component over time.
- Distributed temporal lags up to 4 months, with decaying weights according to a geometric distribution.
- Spatial lags up to maximum of 7th order neighbours, with weights decaying according to a power law ($w_{ij} = o_{ij}^{-d}$, where o_{ij} is the neighbourhood order of blocks *i* and *j*, and the decay exponent *d* is to be estimated).
- Intercept of log population density in the neighbourhood component (*Gravity Law*), to reflect that blocks of high population density may be more strongly influenced by their neighbours due to migration.
- District and state-specific dispersion, allowing the variation in incidence to differ between spatial units.

It was not feasible to allow a block-specific dispersion parameter since many blocks had too few cases to obtain stable estimates.

Finer details of the model selection process are included in S1 Text.

Empirical coverage probabilities. As an alternative measure of prediction utility, we calculated the empirical coverage of prediction intervals produced by each model, with respect to the observed counts. This describes the proportion of points in the test period for which the observed count fell within the middle 50% or 80% of the predicted distribution. For an ideal forecast the empirical coverage will match the nominal level. An empirical coverage probability cannot be considered "strictly proper" [21, 26, 31], as the RPS score is, and hence does not favour sharpness in addition to calibration. However, a high coverage quantile interval may provide useful lower and upper bounds for expected incidence. For more detail see S1 Text.

Longer prediction horizons. For the final model, further predictions were calculated based on a rolling window of three and four months. As with the rolling OSA approach, the model was initially fit to the training set (months 1, . . ., *t*) and this fit used to predict month *t* + 3. The model was then updated with the data from *t* + 1 in order to predict *t* + 4, and so on, in a similar fashion to Lewnard et al. [10]. The RPS of one, three and four month ahead predictions were compared to assess the loss in accuracy with a longer time horizon.

Results

Preliminary analyses of dispersion and exploration of temporal lags are described in S2 Text.

Random model assessment

According to the thirty random models drawn, fit and prediction were found to be strongly correlated (Fig 3A). Predictions were calculated based on either a rolling fit (incorporating each month's data into parameter estimates to predict the next month) or fixed fit (using parameters fit to the training set only for all predictions). The scores for both prediction approaches were very similar for most models, suggesting that the processes defined in these models are consistent over time and hence the quality of prediction does not depend on regular model updates (Fig 3B). This is noteworthy since in practice it may not be possible to update the fits on such a regular basis. Selecting the model based on RPS of predictions from a fixed model fit would best reflect the constraints of reality and be the more conservative approach.

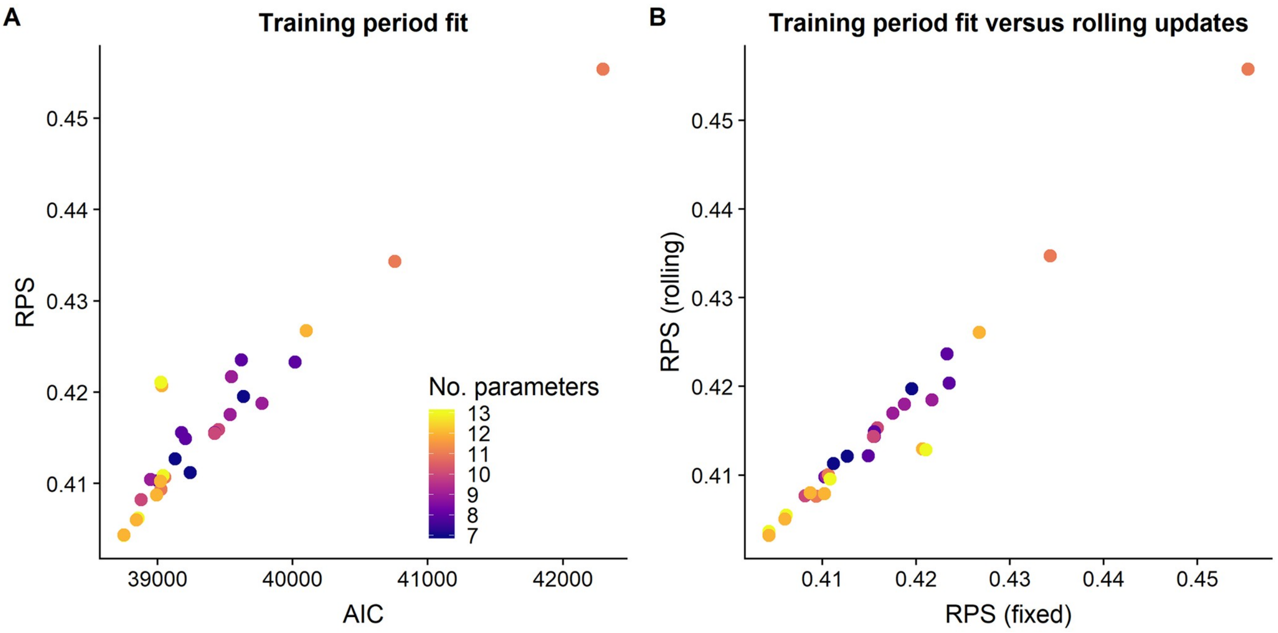


Fig 3. Comparison of predictive performance and model fit, and predictive performance for training period fit and rolling fit updates, for models with randomly selected components. (A) AIC versus RPS for 30 randomly selected models. AIC is calculated from the fit to the training period only (months 13 to 48) and RPS from one-step-ahead predictions (months 49 to 72) based on the same fit. According to this random sample, fit and prediction are strongly correlated; the model which fits best to the observed data produces the best one-step-ahead predictions. (B) RPS of predictions based on the fixed training set fit versus rolling fit updates. Predictive power is very similar between the two prediction approaches.

https://doi.org/10.1371/journal.pntd.0008422.g003

Model selection

As was found with the random model set, the final selected model which demonstrated the highest predictive power as measured by RPS also achieved the closest fit to existing data. Initially, no more than two distributed AR lags could be added to the model without yielding evidence of miscalibration in the predictions. However, once the neighbourhood component was added in the third stage of selection, increasing the AR lags to four months significantly improved both AIC and RPS with no evidence of miscalibration. At this point the endemic linear trend lost significance and therefore was removed in subsequent models. The AIC, RPS and empirical coverage probabilities for all models considered in the selection process are shown in Fig 4. Fit and prediction metrics for all models are given in S1 Table. and PIT histograms for the models selected at each stage are compared in S3 Fig.

We found that as RPS and AIC were improved, the empirical coverage probabilities of prediction intervals were increased far beyond their nominal level. With the final model (Model no. 42), only 5.4% (652/12048) of observations fell outside the 10-90% interval, with an average interval width of just three possible case counts. This predicted distribution is much more conservative in its coverage than a simple linear trend model (coverage 10-90% = 0.905) but attains substantially better fit and RPS, suggesting that more of the improvement comes in the form of calibration. The conservative 90% predicted quantile provides a reliable upper limit for the next month's incidence, to which a management plan could be defined accordingly. The 25-75% prediction interval was found to be of limited use since, with very low counts across the majority of the region, this interval often consists of only a single value. The median would be a more interpretable value to report.

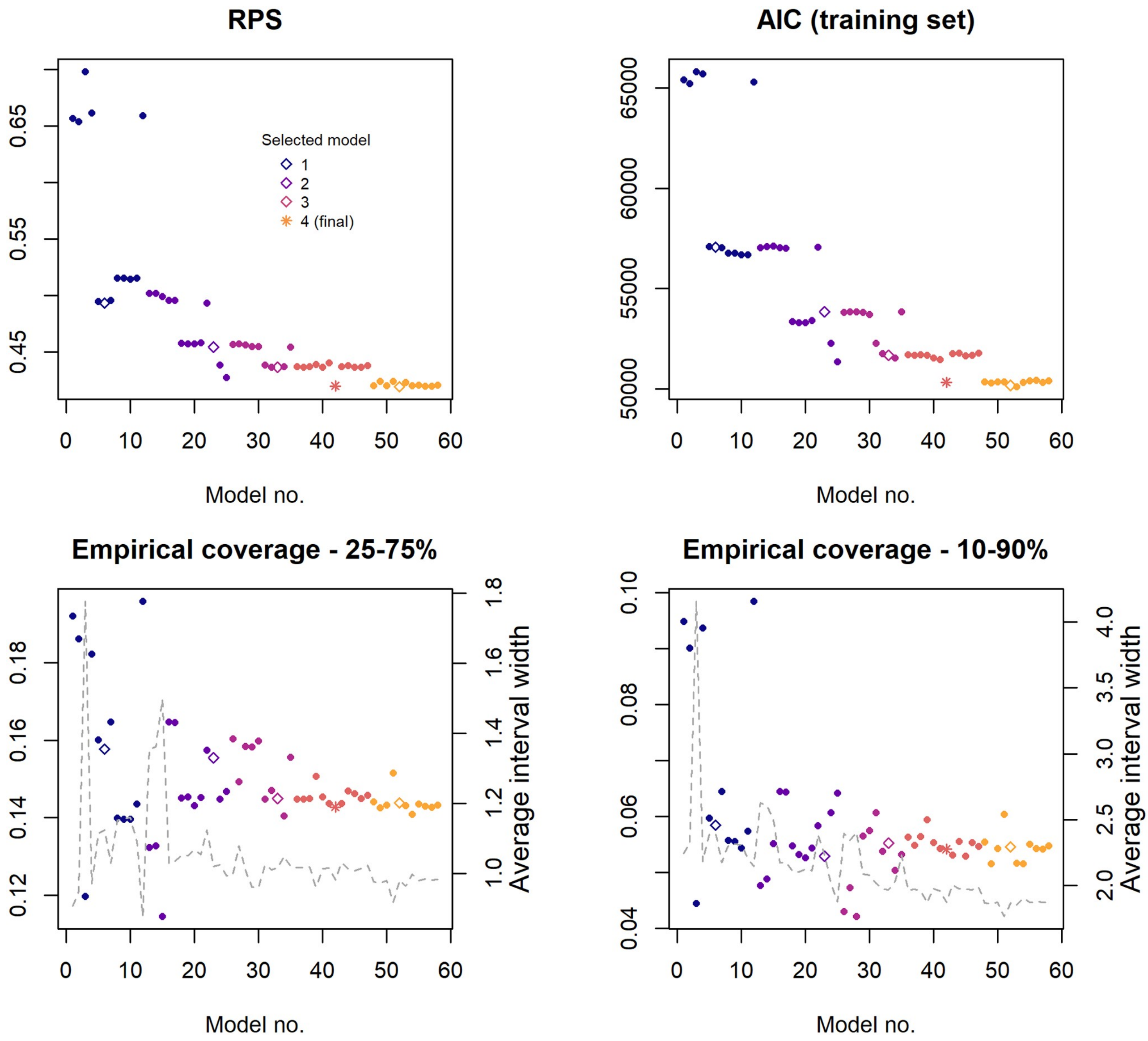


Fig 4. Measures of fit and predictive power throughout the model selection process. Figures illustrate the models tested in chronological order from left to right, with each stage indicated by a different colour. Models were selected at each stage based on the biggest reduction in RPS, subject to calibration; these are identified by hollow points, and the final selected model by a star. For the two variants on the coverage probability, average quantile interval width (representing uncertainty in the predicted case count) is shown on the right axis and by the grey dashed line. Interval width is determined by the count at the upper quantile minus the count at the lower, hence an interval width of two covers three possible count values (e.g. 2, 3, 4).

https://doi.org/10.1371/journal.pntd.0008422.g004

Final model

The final model consists of a negative binomial distribution with a single dispersion parameter and the following mean structure:

$$\mu_{it} = \lambda_{it}\sum_{q=1}^{4} u_q Y_{i,t-q} + \phi_{it}\sum_{j\neq i}\sum_{q=1}^{4} w_{ij} u_q Y_{j,t-q} + e_{it} v_{it} \tag{8}$$

$$\log(v_{it}) = \alpha^{v} \tag{9}$$

$$\log(\lambda_{it}) = \alpha^{\lambda} + \gamma_1^{\lambda} \sin\left(\frac{2\pi}{12}t\right) + \delta_1^{\lambda} \cos\left(\frac{2\pi}{12}t\right) \tag{10}$$

$$\log(\phi_{it}) = \alpha^{\phi} + \gamma_1^{\phi} \sin\left(\frac{2\pi}{12}t\right) + \delta_1^{\phi} \cos\left(\frac{2\pi}{12}t\right) \tag{11}$$

The model fit is dominated by auto-regression; the majority of information with which to predict the current month comes from incidence in the previous four months, with seasonally-varying strength. Since the contribution of each component is modelled on a log scale these parameters have a multiplicative effect, hence the range of the seasonal AR component (approx. [0.6, 0.8]; see S4 Fig) indicates that each month's count is expected to be a certain fraction of the weighted average of the counts over the last four months. This occurs over all blocks and therefore amounts to an overall decreasing trend. After accounting for auto-regression, it was found that the neighbourhood effect did not extend beyond directly bordering blocks with respect to prediction. Seasonality within this component also serves to vary the magnitude of the effect throughout the year.

The contribution of an endemic trend was found to be negligible, reflecting the lack of homogeneity across blocks, and was therefore not included; the reduction in total incidence comes entirely from each block's autoregressive pattern. Block-specific covariate data (e.g. relating to socio-economic or geographic features of the area) would contribute to this component and potentially reveal associations which are consistent across blocks. Random intercepts were tested in the endemic component to capture unexplained block variation, yet did not improve predictive power in a basic model and caused convergence issues in more complex, distributed-lag models.

The relative contributions of the three model components are illustrated for the four blocks with highest average monthly incidence (Gopikandar, Kathikund, Boarijor and Sundarpahari) in Fig 5.

Predictive performance. The final model achieved an overall $\overline{\text{RPS}}$ for one-step-ahead prediction of 0.420, 36% lower than the null (non-spatial and non-autoregressive) model and 8% lower than the best non-spatial model, with individual block-wise averages ranging from 4.3×10^{-5} to 3.47. This equates to a mean absolute error of 0.58, a 30% reduction from the null model. That the RPS is lower than the MAE implies the probabilistic forecast is preferable to a simple point forecast.

Model selection was performed based on the model's *mean* RPS across all blocks and the whole test period but beneath this overall score is a broader distribution of scores for each block-month prediction, influenced by peaks, troughs and otherwise unusual incidence patterns. The histogram in Fig 6 illustrates the distribution over blocks, demonstrating that the final model is able to predict accurately and precisely across the majority of the region, yet there is a small subset of blocks with more widely varying RPS. It should be noted that the overall performance of the model is strongly influenced by blocks with almost no incidence as these yield the very lowest scores. Similarly, there is some correlation between the blocks for which the model performs least well, and the blocks which have historically demonstrated the highest average incidence since higher counts are harder to predict than zeros or single cases. The blocks with the highest RPS also tend to exhibit sporadic patterns or have experienced sudden, sharp changes in incidence (potentially outbreaks) within the test period, which cannot be reproduced by a model primarily informed by an average of past incidence. Examples of these patterns are illustrated in S5 Fig.

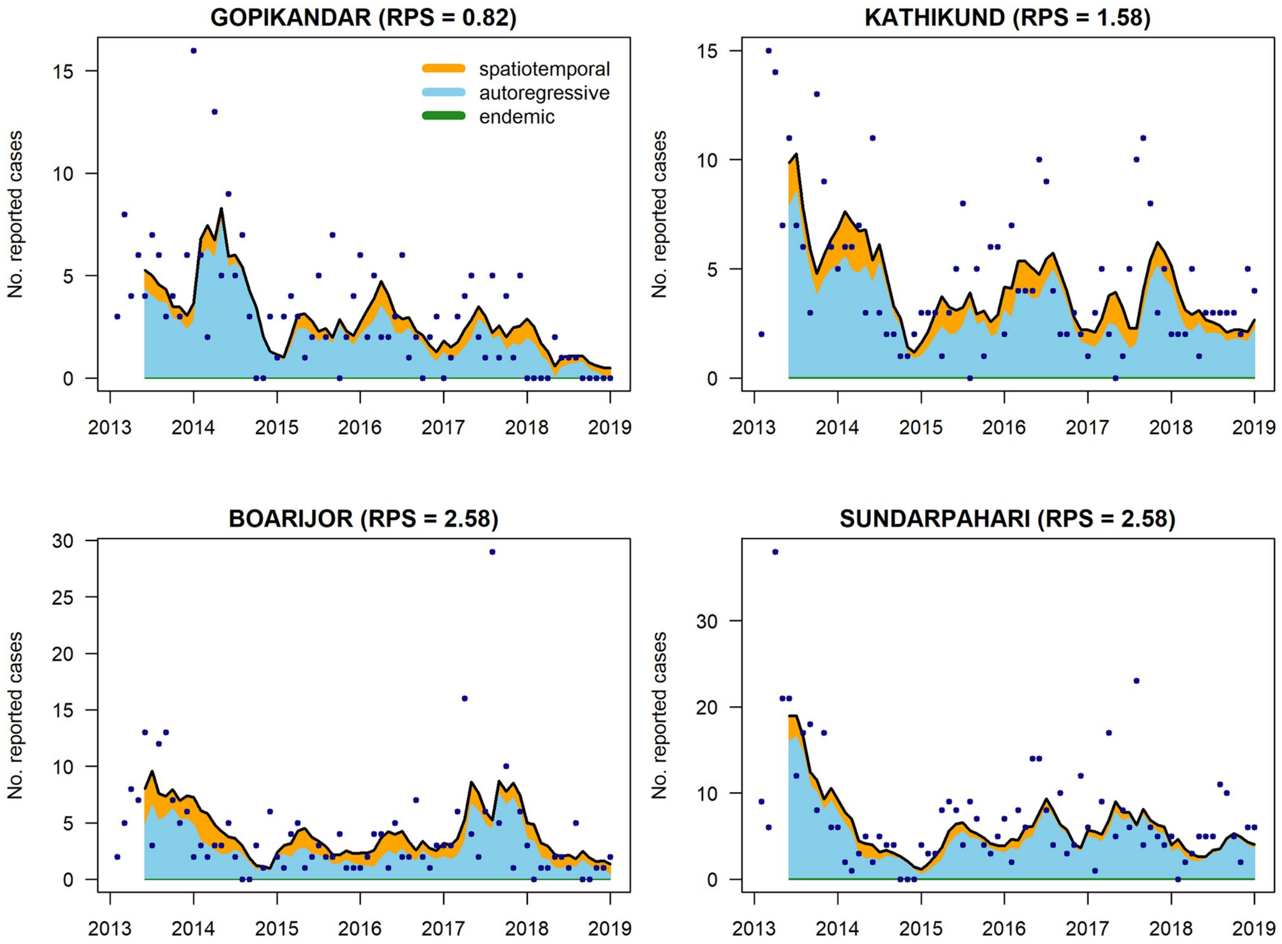


Fig 5. Model fit for the four blocks with highest average monthly incidence (Gopikandar, Kathikund, Boarijor, and Sundarpahari, all in Jharkhand). The observed case counts are indicated by black points and the coloured regions illustrate the relative contribution of the different model components. The contribution of the endemic component is negligible therefore barely visible. The fitted value from the model falls at the upper edge of the coloured region.

https://doi.org/10.1371/journal.pntd.0008422.g005

Pakur, Maheshpur, Boarijor and Sundarpahari in Jharkhand ($\overline{\text{RPS}} = 3.47,\ 2.70,\ 2.58$ and 2.58, resp.) experienced substantial jumps in incidence between May and July 2017, constituting differences of up to 27 cases from one month to the next. Paroo ($\overline{\text{RPS}} = 3.07$) showed a particularly erratic pattern of cases within the test period, with spikes of 21 and 19 cases separated by a few months of $\sim$5 cases and a subsequent fall to just one case by December 2018. Incidence in Garkha has also been inconsistent and appeared to have been on the rise in recent years, until a similar fall at the end of 2018. It should be noted that additional case detection efforts in Jharkhand at the start of 2017 will likely have contributed substantially to the observed spikes at this time.

Three- and four-month-ahead prediction. For the final model, further predictions were calculated based on rolling windows of three and four months. Fig 7 illustrates that the longer time window did not result in a substantial loss in predictive power, with block-wise RPS very similar for the majority of blocks. When compared over the same predicted months, the differences in $\overline{\text{RPS}}$ between one-month-ahead prediction and three-/four-month-ahead were found to be small but statistically significant (-0.024 and -0.028, resp.; $p < 0.0001$ for both). In terms of the empirical coverage, 85.4% of test period observations were captured in the middle 50%

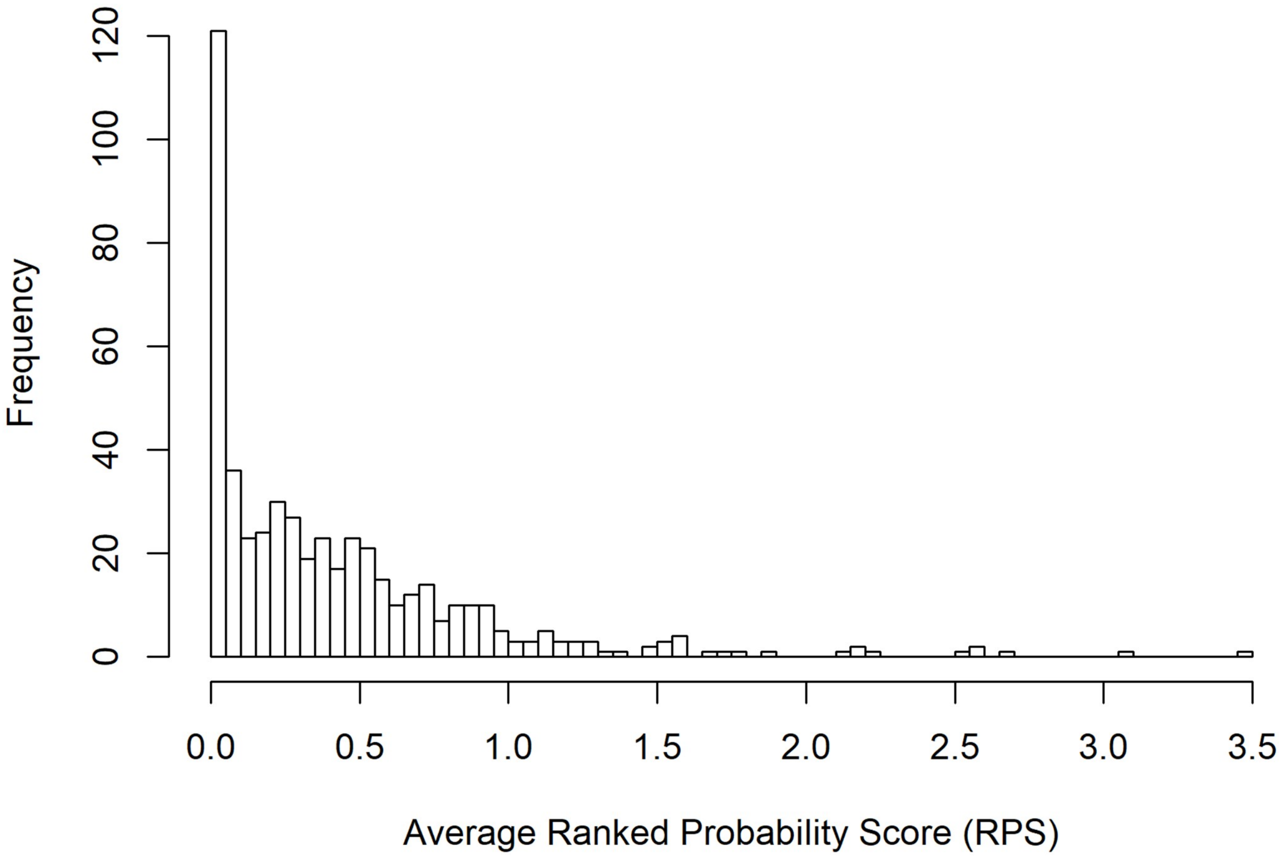


Fig 6. Distribution of time-averaged ranked probability scores across all 502 blocks. Low values reflect accurate and precise prediction. The majority of blocks fall below 1 with a subset for which predictive power varies widely.

https://doi.org/10.1371/journal.pntd.0008422.g006

of the predicted distribution based on a three month window, and 85.7% with a four month window.

Figs 8 and 9 illustrate the coverage of 45-55%, 25-75% and 10-90% prediction intervals for the block with the highest $\overline{\text{RPS}}$ of 3.47 (Pakur, Jharkhand) and a block with $\overline{\text{RPS}}$ of 1 (Bhagwanpur, Bihar). For Pakur, RPS is strongly influenced by the model's inability to match the spike in 2017, yet the incidence in surrounding months is well represented.

Discussion

We have presented the evaluation of a predictive model of VL in Bihar and four endemic districts in Jharkhand, demonstrating a substantial (36% lower RPS) benefit from incorporating spatial and historical case information when compared to a non-spatial, linear trend model. To the best of our knowledge, this is the first time the spatio-temporal correlation of incidence at block level across all the endemic districts of Bihar and Jharkhand has been quantified. We have empirically investigated the performance of different models on prediction performance rather than model fit and produced a statistical model that is capable of accurate forecasting. Such a framework can be used as an important tool for management of endemic diseases.

Given the lack of an effective vaccine and evidence that indoor residual spraying of insecticide fails to significantly reduce sandfly densities and VL incidence in sprayed villages [35, 36],

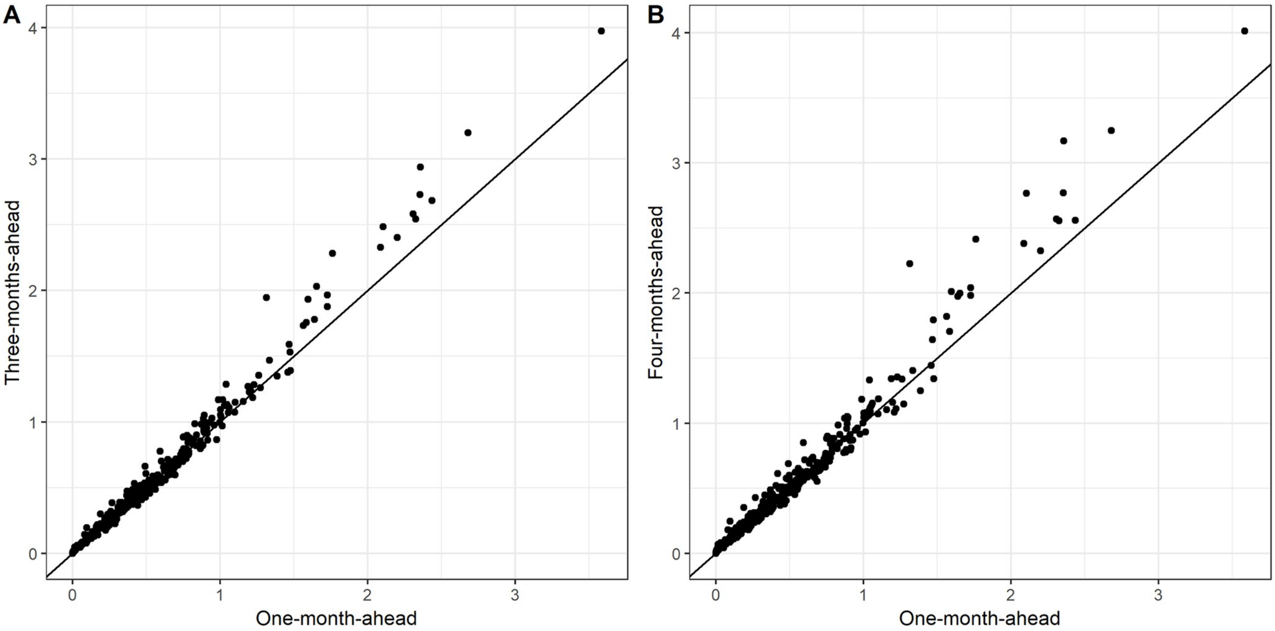


Fig 7. Time-averaged (over months 52-72 for comparability) RPS for three- (A) and four-month-ahead (B) predictions versus one-month-ahead. Scores are closely matched for the majority of blocks (where $\overline{\text{RPS}} < 1.5$) but the differences increase for blocks which are harder to predict.

https://doi.org/10.1371/journal.pntd.0008422.g007

rapid diagnosis and treatment is currently the best method of control. With a block-level estimate of the likely number of cases to arise over the next few months, local management teams could take steps to ensure they are prepared. For example, the 90% quantile of the predicted distribution could be used to inform block-specific minimum stock levels for rapid diagnostic tests and drugs.

In practice, the prediction interval is constrained by the efficiency of the reporting process; the time taken to process diagnosis reports and input the information into the database sets a minimum horizon at which predictions would be genuinely prospective and therefore of practical use. In this paper we have assumed a delay of two months until a month's data can be considered complete, which would necessitate making predictions at least three months ahead of

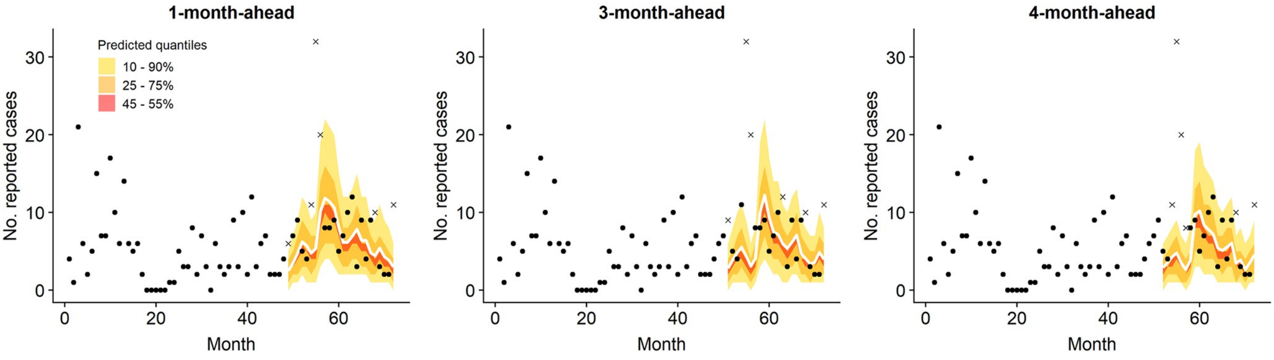


Fig 8. One-, three- and four-step-ahead predictions (solid white line) with 10-90%, 25-75% and 45-55% quantile intervals, for Pakur block in Jharkhand ($\overline{\text{RPS}}$ = 3.47 for one-step-ahead over months 49-72). Observations which fall outside the outer prediction interval are indicated by a cross.

https://doi.org/10.1371/journal.pntd.0008422.g008

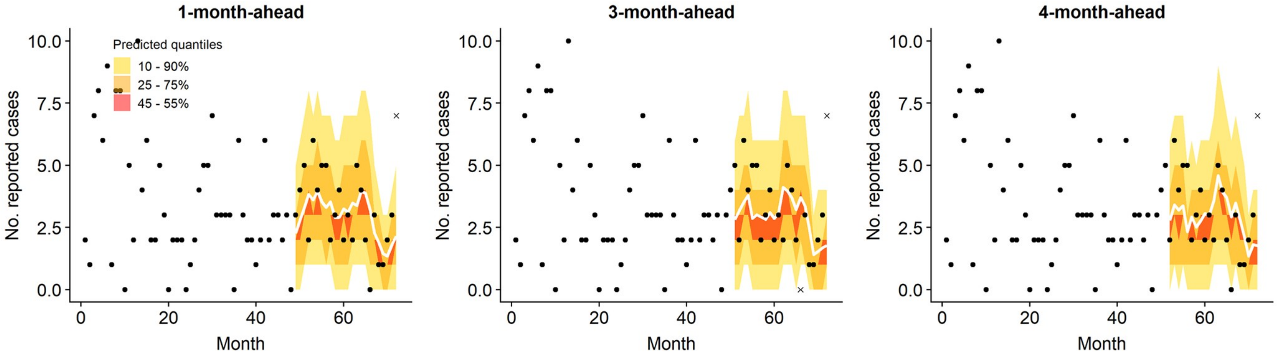


Fig 9. Corresponding predictions for Bhagwanpur block in Bihar ($\overline{\text{RPS}} = 1.00$).

https://doi.org/10.1371/journal.pntd.0008422.g009

that point. However, conservative predictions based on preliminary month totals would still likely be of use to the national control programme.

We have demonstrated here that rolling three-month-ahead predictions are a reasonable approximation to one-month-ahead, but confidence is sacrificed for a minority of blocks as the time horizon is increased. There is a need for discussion with local disease management teams to determine the optimal balance between practicality and uncertainty with respect to predictions. Moreover, the way in which we quantify the accuracy and utility of predictions would benefit from some public health insight; it is highly likely that over- and under-estimation would need to be weighted differently, which may alter which model is deemed preferable. Ideally, the model structure would have been optimised according to predictive power on this slightly longer time horizon, but this is not a trivial task and was deemed beyond the scope of this paper.

There are also potential issues with movement of VL cases across international borders; in particular, the international boundary with Nepal cuts through a VL endemic area, artificially removing some aspects of spatial correlation. Ideally, we would take a regional perspective and also include areas in neighbouring states that have more sporadic reported VL incidence.

It could be argued that the block-level is too coarse a spatial scale for modelling the spread of an infectious disease. Outbreaks of VL occur on a smaller spatial and temporal scale than has been applied here, therefore cannot be anticipated by this model. The transmission dynamic models which are usually employed for this type of problem can be defined on a village, household or even individual level [37], yet this more detailed picture demands many more assumptions which are difficult to justify in this context. The sparseness of cases at this point in the elimination process also means that aggregation at a finer temporal scale might lead to issues with parameter estimation. The block is the unit at which control efforts are co-ordinated, disease burden is monitored, and control targets are set, therefore predictions at this level could prove to be a worthwhile compromise while more realistic transmission models are developed. With more detailed location data, the spread of disease can be modelled as a point process at the village or household level, potentially giving insight into the size and movement of disease clusters or "hot-spots" over time. This technique has previously been applied to the case of VL [38] and may be possible to extend to a larger study region in the near future, following a recent effort to collect GPS co-ordinates of affected villages across Bihar.

In this case the best-fitting model was found to be the best-predicting model. The similarity of prediction and fitting results perhaps reflects the continuity of the processes creating the data. However, consideration of predictive power across the whole range of possible values was key to determining an optimal temporal lag length for short-term prediction. Fit and overall predictive power favoured a high number of lags in order to best capture the spatio-temporal correlation between neighbouring block counts, which appears to contribute to prediction of sudden changes in incidence. However, auto-regression is the dominant model component and appears to be captured by lags up to four months. It would be preferable to specify a different lag length for the auto-regressive and spatial components but this is not currently implemented in the *surveillance* framework. By inspection of PIT histograms, we were able to select the lag length which balanced overall predictive power with capacity to predict at the upper end of the range.

The model selection approach taken in this analysis is semi-systematic; it was not feasible to assess every possible combination of model components. Therefore we aimed to home in on a suitable model by adding components which gave the biggest improvement in predictive performance out of a range of likely options. It was found that once the major components were included in some form, further adjustment largely had the effect of redistributing the variation attributed to each component and did not substantially alter fit or prediction. There is only so much information within the time series of cases to feed the model, so predictive power quickly reaches a limit.

The analysis presented here aims to demonstrate the best that can be done with the minimal information routinely collected by the current programme, but there is evidence that this model still cannot fully account for the heterogeneity in incidence across the region. The lack of geographic and/or demographic covariates beyond population size means that the endemic component in this model is negligible; almost all our information comes from the spatio-temporal correlations, underlining the need for up-to-date data in order to make accurate predictions. Associations between VL incidence and, for example, age and socio-economic quintiles have been demonstrated [18, 39], which may give rise to varied endemic patterns at the block level. This unknown variation could in theory be quantified by random effects within this model framework, but convergence issues (likely due to the large number of zero-counts) made this infeasible in practice.

There is clearly a limitation of fitting such a model over a large number of highly heterogeneous units with minimal unit-specific information. Model selection was performed based on an average score over all blocks and time points for which predictions were made; a model is therefore chosen which predicts well overall, but in doing so sacrifices predictive power for a minority of blocks which do not follow the general trend. Zero counts dominate over all time and space, and the variance of the negative binomial distribution with a universal dispersion parameter is still too restrictive to account for blocks with the highest counts. It is in these areas where additional information on potential predictors of incidence could prove most valuable.

The variation in case counts may be better explained by a zero-inflated process, and the extent of zero-inflation will likely become more prominent as elimination is approached. Bayesian hierarchical models can be used to distinguish sources of variation at different levels and have the benefit of accommodating any informal or incomplete understanding of the transmission process within prior distributions for model parameters. These models have until recently been commonly implemented using Markov Chain Monte Carlo (MCMC) [40], which is computationally intensive for data rich in both space and time. They are however becoming increasingly accessible as a tool for inference and prediction, thanks to user-friendly wrappers which take advantage of fast computation using Integrated Nested Laplace Approximations (INLA) [41]. We will explore this approach in future work.

Conclusion

We have demonstrated a framework for forecasting VL incidence at subdistrict level in India which achieves good predictive performance based on the available routinely collected surveillance data. This framework could be used to make short-term forecasts to provide an early indication of where case numbers are higher (or lower) than expected and to support the logistics of the elimination programme.

Supporting information

S1 Text. Model selection.
(PDF)

S2 Text. Preliminary analyses.
(PDF)

S1 Fig. Districts with unusual incidence patterns resulting in inflated dispersion estimates.
(TIF)

S2 Fig. Probability integral transform (PIT) histograms for models with increasing orders of geometric lags from 1 to 12 months (left to right, top to bottom) in the auto-regressive component. The final model selection process considered up to four lags.
(TIF)

S3 Fig. PIT histograms for selected models at each stage. Model 42 is the final model. Model 52 offered minor improvement in RPS with additional complexity.
(TIF)

S4 Fig. Fitted seasonal waves in auto-regressive (AR) and neighbourhood (NE) model components. Both reflect the first-quarter peak in reported cases but the magnitude of the waves differs, with the contribution of the AR component varying more than that of the NE.
(TIF)

S5 Fig. Blocks with average RPS greater than 2.5 over the test period (Jan 2017—Dec 2018).
(TIF)

S1 Table. Fit and prediction metrics for selected models at each stage. The value of *S* within the formula indicates the number of seasonal waves included. The reported AIC is for the fit to training data only, and RPS is of predictions made without updating this fit (i.e. fixed instead of rolling). C2575 and C1090 refer to the coverage of 50% and 80% quantile intervals, respectively, alongside the average interval width in cases. Model no. 42 is the final model.
(PDF)

Acknowledgments

The authors gratefully acknowledge the support and permission of Dr Dhingra, Director of the National Vector-Borne Disease Control Programme (NVBDCP), Government of India, to use the KA-MIS data.

We thank Joy Bindroo and the staff of CARE India, Patna, for their help with data access and queries. We are also very grateful to Tim Pollington of the University of Warwick for helpful input on the *surveillance* package and model selection.

Author Contributions

Conceptualization: Lloyd A. C. Chapman, Sridhar Srikantiah, Mary M. Cameron, Graham F. Medley.

Data curation: Emily S. Nightingale, Sridhar Srikantiah.

Formal analysis: Emily S. Nightingale, Lloyd A. C. Chapman, Graham F. Medley.

Funding acquisition: Mary M. Cameron, Graham F. Medley.

Investigation: Sridhar Srikantiah.

Methodology: Emily S. Nightingale, Lloyd A. C. Chapman, Johannes Bracher, Graham F. Medley.

Project administration: Graham F. Medley.

Resources: Mary M. Cameron, Graham F. Medley.

Software: Emily S. Nightingale, Johannes Bracher.

Supervision: Lloyd A. C. Chapman, Graham F. Medley.

Validation: Emily S. Nightingale.

Visualization: Emily S. Nightingale.

Writing – original draft: Emily S. Nightingale.

Writing – review & editing: Emily S. Nightingale, Lloyd A. C. Chapman, Swaminathan Subramanian, Purushothaman Jambulingam, Johannes Bracher, Graham F. Medley.

References

1. Alvar J, Vélez ID, Bern C, Herrero M, Desjeux P, Cano J, et al. Leishmaniasis Worldwide and Global Estimates of Its Incidence. PLoS ONE. 2012; 7(5):e35671. https://doi.org/10.1371/journal.pone.0035671 PMID: 22693548
2. Ready P. Epidemiology of visceral leishmaniasis. Clinical Epidemiology. 2014; 6(1):147–154.
3. Singh NS, Singh DP. A Review on Major Risk Factors and Current Status of Visceral Leishmaniasis in North India. American Journal of Entomology. 2019; 3(1):6–14. https://doi.org/10.11648/j.aje.20190301.12
4. National Vector Borne Disease Control Programme. Kala-Azar Situation in India; 2018. Available from: https://www.nvbdcp.gov.in/index4.php?lang=1{&}level=0{&}linkid=467{&}lid=3750.
5. Dye C, Wolpert DM. Earthquakes, influenza and cycles of Indian kala-azar. Transactions of the Royal Society of Tropical Medicine and Hygiene. 1988; 82:843–850.
6. Bora D. Epidemiology of visceral leishmaniasis in India. The National Medical Journal of India. 1999; 12 (2):62–68.
7. Courtenay O, Peters NC, Rogers ME, Bern C. Combining epidemiology with basic biology of sand flies, parasites, and hosts to inform leishmaniasis transmission dynamics and control. PLoS Pathogens. 2017; 13(10):e1006571. https://doi.org/10.1371/journal.ppat.1006571
8. Rijal S, Sundar S, Mondal D, Das P, Alvar J, Boelaert M. Eliminating visceral leishmaniasis in South Asia: the road ahead. BMJ (Clinical research ed). 2019; 364:k5224.
9. Ministry of Home Affairs, Government of India. C.D. Block Wise Primary Census Abstract Data; 2011. Available from: http://censusindia.gov.in/pca/cdb_pca_census/cd_block.html.
10. Lewnard JA, Jirmanus L, Júnior NN, Machado PR, Glesby MJ, Ko AI, et al. Forecasting Temporal Dynamics of Cutaneous Leishmaniasis in Northeast Brazil. PLoS Neglected Tropical Diseases. 2014; 8 (10):e3283. https://doi.org/10.1371/journal.pntd.0003283 PMID: 25356734
11. Li HL, Zheng RJ, Zheng Q, Jiang W, Zhang XL, Wang WM, et al. Predicting the number of visceral leishmaniasis cases in Kashgar, Xinjiang, China using the ARIMA-EGARCH model. Asian Pacific Journal of Tropical Medicine. 2020; 13(2):81–90. https://doi.org/10.4103/1995-7645.275416

12. Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CAS, Sá Carvalho M, et al. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. Computers & Geosciences. 2011; 37(3):371–381. https://doi.org/10.1016/j.cageo.2010.01.008

13. Amro A. Epidemiology and spatiotemporal analysis of visceral leishmaniasis in Palestine from 1990 to 2017. International Journal of Infectious Diseases. 2020; 90:206–212.

14. Dewan A, Abdullah AYM, Shogib MRI, Karim R, Rahman MM. Exploring spatial and temporal patterns of visceral leishmaniasis in endemic areas of Bangladesh. Tropical Medicine and Health. 2017; 45 (1):29. https://doi.org/10.1186/s41182-017-0069-2

15. Sevá AdP, Mao L, Galvis-Ovallos F, Tucker Lima JM, Valle D. Risk analysis and prediction of visceral leishmaniasis dispersion in São Paulo State, Brazil. PLoS Neglected Tropical Diseases. 2017; 11(2): e0005353. https://doi.org/10.1371/journal.pntd.0005353

16. Bhunia GS, Kesari S, Chatterjee N, Kumar V, Das P. Spatial and temporal variation and hotspot detection of kala-azar disease in Vaishali district (Bihar), India. BMC Infectious Diseases. 2013; 13(1):64. https://doi.org/10.1186/1471-2334-13-64

17. Godana AA, Mwalili SM, Orwa GO. Dynamic spatiotemporal modeling of the infected rate of visceral leishmaniasis in human in an endemic area of Amhara regional state, Ethiopia. PLoS ONE. 2019; 14(3): e0212934. https://doi.org/10.1371/journal.pone.0212934

18. Bulstra CA, Le Rutte EA, Malaviya P, Hasker EC, Coffeng LE, Picado A, et al. Visceral leishmaniasis: spatiotemporal heterogeneity and drivers underlying the hotspots in Muzaffarpur, Bihar, India. PLoS Neglected Tropical Diseases. 2018; 12(12):e0006888. https://doi.org/10.1371/journal.pntd.0006888 PMID: 30521529

19. Le Rutte EA, Chapman LAC, Coffeng LE, Jervis S, Hasker EC, Dwivedi S, et al. Elimination of visceral leishmaniasis in the Indian subcontinent: a comparison of predictions from three transmission models. Epidemics. 2017; 18:67–80. https://doi.org/10.1016/j.epidem.2017.01.002 PMID: 28279458

20. Meyer S, Held L, Höhle M. Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. Journal of Statistical Software. 2017; 77(11). https://doi.org/10.18637/jss.v077.i11

21. Held L, Meyer S, Bracher J. Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. Statistics in Medicine. 2017; 36(22):3443–3460. https://doi.org/10.1002/sim.7363

22. Adegboye OA, Adegboye M. Spatially correlated time series and ecological niche analysis of cutaneous leishmaniasis in Afghanistan. International Journal of Environmental Research and Public Health. 2017; 14(3):309. https://doi.org/10.3390/ijerph14030309

23. Bracher J, Held L. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. arXiv e-prints. 2019; p. arXiv:1901.03090.

24. Furlan CPR, Diniz CAR, Franco M. Estimation of lag length in distributed lag models: A comparative study. Advanced Applied Statistics. 2010; 17(2):127–142.

25. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 2nd ed. OTexts: Melbourne, Australia; 2018. Available from: OTexts.com/fpp2.

26. Czado C, Gneiting T, Held L. Predictive model assessment for count data. Biometrics. 2009; https://doi.org/10.1111/j.1541-0420.2009.01191.x PMID: 19432783

27. Gneiting T, Katzfuss M. Probabilistic Forecasting. Annual Review of Statistics and Its Application. 2014; 1(1):125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

28. Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. Assessing the performance of real-time epidemic forecasts: A case study of ebola in the Western area region of sierra leone, 2014-15. PLoS Computational Biology. 2019; 15(2):e1006785. https://doi.org/10.1371/journal.pcbi.1006785

29. Lu J, Meyer S. Forecasting Flu Activity in the United States: Benchmarking an Endemic-Epidemic Beta Model. International Journal of Environmental Research and Public Health. 2020; 17(4):1381. https://doi.org/10.3390/ijerph17041381

30. Chaves LF, Pascual M. Comparing models for early warning systems of neglected tropical diseases. PLoS Neglected Tropical Diseases. 2007; 1(1):e33. https://doi.org/10.1371/journal.pntd.0000033

31. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2007; 69(2):243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

32. Höhle M, Meyer S, Paul M. Surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena; 2016. Available from: https://cran.r-project.org/package=surveillance.

33. Bracher J. hhh4addon: Extensions to endemic-epidemic timeseries modeling from package surveillance; 2018. Available from: https://github.com/jbracher/hhh4addon.

34. R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: https://www.r-project.org/.

35. Poche DM, Garlapati RB, Mukherjee S, Torres-Poche Z, Hasker E, Rahman T, et al. Bionomics of Phlebotomus argentipes in villages in Bihar, India with insights into efficacy of IRS-based control measures. PLoS Neglected Tropical Diseases. 2018; 12(1):1–20. https://doi.org/10.1371/journal.pntd.0006168

36. Picado A, Dash AP, Bhattacharya S, Boelaert M. Vector control interventions for Visceral Leishmaniasis elimination initiative in South Asia, 2005-2010. Indian Journal of Medical Research. 2012; 136(1):22–31.

37. Chapman LAC, Jewell CP, Spencer SEF, Pellis L, Datta S, Chowdhury R, et al. The role of case proximity in transmission of visceral leishmaniasis in a highly endemic village in Bangladesh. PLOS Neglected Tropical Diseases. 2018; 12(10). https://doi.org/10.1371/journal.pntd.0006453

38. Mandal R, Kesari S, Kumar V, Das P. Trends in spatio-temporal dynamics of visceral leishmaniasis cases in a highly-endemic focus of Bihar, India: an investigation based on GIS tools. Parasites & vectors. 2018; 11(1):220. https://doi.org/10.1186/s13071-018-2707-x

39. Chapman LAC, Morgan ALK, Adams ER, Bern C, Medley GF, Hollingsworth TD. Age trends in asymptomatic and symptomatic Leishmania donovani infection in the Indian subcontinent: A review and analysis of data from diagnostic and epidemiological studies. PLOS Neglected Tropical Diseases. 2018; https://doi.org/10.1371/journal.pntd.0006803

40. Jewell CP, Kypraios T, Neal P, Roberts GO. Bayesian analysis for emerging infectious diseases. Bayesian Analysis. 2009; 4(3):465–496. https://doi.org/10.1214/09-BA417

41. Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and spatio-temporal models with R-INLA; 2013. Available from: https://www.sciencedirect.com/science/article/pii/S1877584512000846.

Chapter 4

Inferring the distribution of visceral leishmaniasis incidence from data at different spatial scales

4.1 Abstract

As discussed in Chapter 3, there is justification for moving towards a finer scale than the block level for monitoring the progress of VL elimination in India as case numbers dwindle. Projections of low incidence are difficult to act upon across broad regions of unevenly-distributed communities, yet blocks with low but non-zero incidence cannot be overlooked if the goal is to avoid resurgence and sustain elimination of the public health problem. Conversely, an elimination setting inevitably demands a reduction in the intensity of surveillance for sustainability in the long-term. Bihar's population is spread across 45-50,000 villages, with the vast majority unaffected by VL. Identifying spatial patterns in incidence between nearby villages could help guide more efficient distribution of interventions.

This chapter analyses village level incidence across the whole of Bihar. Spatial autocorrelation in observed incidence and associations with local environmental conditions are explored, and a statistical approach to infer village-level variation from more easily obtained block-level data is evaluated. The disaggregation approach does not estimate village-level incidence more accurately than a baseline prediction of block-homogenous incidence. A robust village-level model also does not yield more accurate prediction than

baseline, suggesting that the limitations of disaggregation are not due to non-linear or interacting covariate effects. Spatial autocorrelation is evident on a global scale but appears weak between neighbouring villages within individual blocks, suggesting that an important transmission mechanism may act stochastically and at a longer range, for example due to migration.

Increasing the range of reactive interventions to neighbouring villages may not improve efficacy. However, village-level surveillance allows rapid detection of further within-village incidence in response to a single reported case. Increased routine surveillance among more mobile population groups may also reduce the risk of reintroduction into previously unaffected villages.

A version of this manuscript will be submitted for publication following completion of the thesis. The code and the underlying data already permitted to be shared have been made available in the following public repository https://github.com/esnightingale/vl-village-level. The village level case data against which the models were validated will be added once approval has been granted by CARE India. I conceptualised the project, cleaned and constructed the spatially-referenced village level data, conducted the analysis and drafted the manuscript, with supervision from Graham Medley and Oliver Brady and methodological input from Tim Lucas. Ashley Schwartzer conducted initial exploration and analysis of the village level case data for her Master's thesis in 2019 [72], which I co-supervised with Dr Lloyd Chapman.

4.2 Background

A key issue raised in previous work forecasting Visceral Leishmaniasis (VL) incidence at the block level is the appropriateness of this geographic scale of inference for drawing actionable conclusions. Spatial correlation in block-level incidence was observed in [73] and it was demonstrated that exploiting these correlations had value for improving short-term temporal predictions. The block, however, is too large of a scale for targeting low levels of transmission and incidence; predictions at a higher spatial resolution are needed. It has been demonstrated previously that the choice of spatial scale/units of analysis can have an unintended influence on conclusions (known as the Modifiable Areal Unit Problem, or MAUP), therefore careful consideration is required as to what partition is appropriate given how the data have been collected [74].

On the other hand, an elimination setting inevitably demands a reduction in the intensity of surveillance in order to be sustainable in the long term, and monitoring incidence at a fine scale is incredibly resource-intensive. The population of Bihar state is spread across approximately 45,000 "villages" - from densely populated wards of the capital city to remote, rural hamlets. The vast majority of these villages are not affected by VL, in particular those south of the Ganges river, while others persistently observe cases and suffer outbreaks.

Case counts of VL are currently monitored at the village level to inform vector control and case detection activities [75], yet villages are treated almost entirely independently; recent observation of any case in a village dictates subsequent years' interventions in that village alone, but not neighbouring villages. This approach to the deployment of active case detection appears to capture a majority of future cases [76] yet inevitably cannot account for sporadic cases in previously unaffected villages. Interventions could be applied more efficiently if sporadically-affected villages were covered within the range of a nearby persistently-affected village, rather than waiting for a response to be triggered within each independently. This chapter therefore also aims to evaluate the evidence for correlation between incidence observed in neighbouring villages, to ascertain whether the efficacy and efficiency of this intervention could be improved by broadening its spatial range.

Clustering of cases within villages has been previously demonstrated [77, 78, 79], while correlation between villages has primarily been explored with respect to climatic and environmental conditions suitable to the sandfly vector [77, 80, 81, 82]. Transmission of VL occurs when adult female sandflies seek human blood to mature their eggs, therefore conditions for sandfly breeding will influence the exposure of the human population. In particular, the type of vegetation, temperature, moisture and living conditions of the human population have been suggested as potentially related to transmission risk [83, 84]. Such studies have however been limited in spatial scale to one or two example districts, usually chosen due to high disease burden (or low, to serve as a control).

This analysis aims to draw inference from the same data source of reported VL diagnoses at the block level, but combining this with remotely-sensed covariate data and exploiting a disaggregation approach [70] to infer the potential distribution of those cases at a more local level. Also sometimes referred to as "downscaling", methods for inferring fine scale variation from spatially aggregated data have progressed substantially in recent years, alongside computational developments in the field of spatial statistics more broadly [85, 86, 87, 71, 88, 89, 90, 13, 91].

It is, however, rarely possible to validate disaggregation approaches against data actually observed at the finer scale. Python et al. [88] were able to exploit two sources of data available at the district level to validate their disaggregation of province-level COVID-19 incidence, but neither had complete country-wide coverage. Previous validation of this particular implementation had only been conducted by simulation [71]. For the case described here, acquisition of GPS coordinates for VL-affected villages in Bihar has made it possible to attribute observed cases to a precise location in space and to infer the locations of unaffected villages through linkage with village boundary polygons. This provides a unique opportunity to evaluate whether disaggregation can accurately replicate the distribution of a block's case count across its constituent villages, for the entire state. It is increasingly inefficient to implement uniform interventions across broad geographic units as incidence continues to decline and transmission may be limited to a few small pockets of the population. Identifying and enumerating each unique village in the state of Bihar is a complex and resource-intensive exercise, yet it is now routine to collect a GPS location for each newly-diagnosed case. This creates an opportunity to take into account the similarities between nearby villages - with respect to both population and transmission risk - to inform the use of targeted interventions, but is more challenging analytically than current practice. This work therefore evaluates an approach which does not depend on the collection and maintenance of surveillance data at the village level, with an aim to assess the added value of this information for our understanding of the spatial distribution of observed disease burden, and potentially of underlying transmission.

Aims and objectives

The overall aim of this analysis is to assess the added value of applying models to high-resolution village-level data compared to down-scaling from the block level predictions for guiding interventions and evaluating progress of the elimination programme. This will be addressed through the following objectives:

1. To construct a data set of village-level VL incidence for 2018 based on GPS coordinates of affected villages and village-level shapefiles, with which to validate village-level predictions.

2. To generate predictions of village-level incidence based on a disaggregation model fit to block-level incidence and pixel-level covariate data

3. To generate predictions of village-level incidence based on a random forest model fit to village-level incidence and village-level covariate data

4. To evaluate the accuracy of each set of predictions against the validation data, relative to a baseline prediction of uniform incidence within blocks

4.3 Materials and methods

4.3.1 Data

Counts of VL cases diagnosed in 2018 per village in Bihar were compiled and shared by CARE India, along with coordinates for the centroid of each village affected. These data were first linked to corresponding village polygons by overlaying the affected village point locations. Where multiple points fell within the same polygon, case counts were aggregated. Polygons in which no points fell were defined as unaffected and attributed with a case count of zero.

Populations were estimated by first extracting and summing 100m pixel values from WorldPop population count raster [92] for the set of village polygons. The counts for constituent villages were then aggregated, alongside case counts and polygon geometries, to yield a block-level (administrative level 3) analysis dataset with which to fit the disaggregation model. Pixel-level predictions from the disaggregation model could then be aggregated according to the same village polygons and validated against the original village level counts (Figure 4.3.1).

Pixel-level covariates

Raster data for elevation (metres above sea level) and distance to inland water bodies (metres) were obtained from WorldPop at a resolution of 100m for the region of Bihar state [93]. Estimated travel time to the nearest urban centre (minutes) was obtained from the Malaria Atlas Project (MAP) at a resolution of 1km [94]. Authors of [94] defined an urban centre as a "contiguous area with 1,500 or more inhabitants per square kilometre, or a majority of built-up land cover coincident with a population centre of at least 50,000 inhabitants". Land surface temperature (LST) and normalised difference vegetation index (NDVI) at a 1km resolution were extracted for the same region from MODIS/Terra satellite data, accessed via the AppEEARS platform [95, 96, 97]. The latter two were initially extracted on a monthly scale and subsequently aggregated to an annual mean and standard deviation. All rasters were resampled to the lowest resolution (1km) for inclusion in the disaggregation model. No uncertainty in these inferred covariate values was incorporated into the analysis.

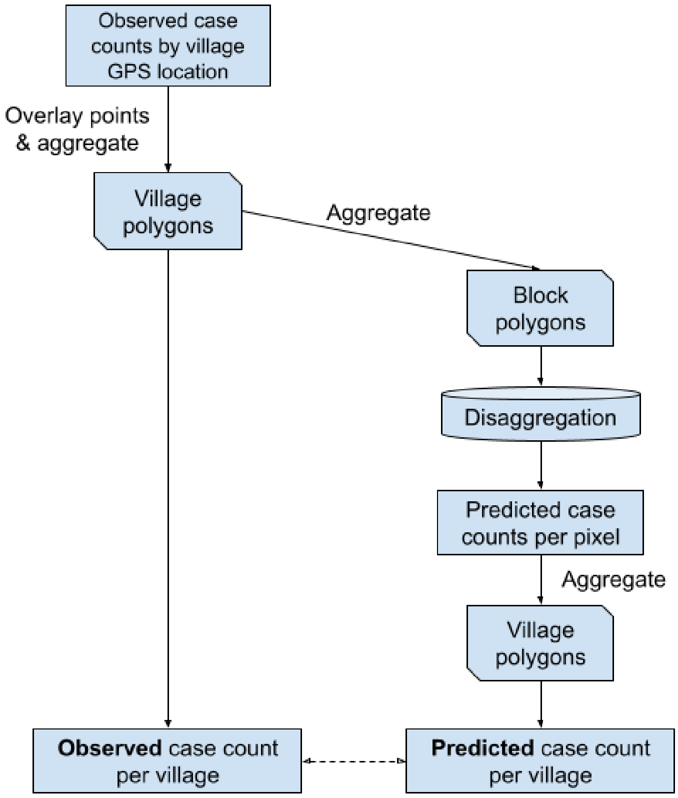

Observed case counts by village GPS location
Overlay points & aggregate
Village polygons
Aggregate
Block polygons
Disaggregation
Predicted case counts per pixel
Aggregate
Village polygons
Observed case count per village
Predicted case count per village

Figure 4.3.1: Data processing steps to perform and validate block-level disaggregation, based on the available geotagged village case counts.

4.3.2 Descriptive analysis

Preliminary analyses assessed the evidence for (global) spatial auto-correlation in incidence between neighbouring blocks and villages for the year 2018, by calculation of Moran's I statistic. The strength of evidence for auto-correlation was interpreted by comparison of the observed statistic value to the distribution of values calculated from 999 permutations of the data under the assumption of spatial independence [98].

4.3.3 Disaggregation model structure

Disaggregation regression combines observed block-level case counts with these finer-scale population and covariate data to predict the potential within-block distribution of incidence [70].

The model is specified as a Poisson regression on the pixel level, with covariate values predicting case counts per pixel. However, the case counts per pixel are not known, and the model parameters are instead optimised relative to the sum of pixel counts across areas (in this case blocks).

For incidence rate r in pixel j in block i with location s_{ij},

$$\log(r_{ij}) = \beta_0 + \beta X_{ij} + GRF(s_{ij}) + u_i$$

Where X_{ij} are covariate values for pixel j in block i, GRF is a Gaussian random field across blocks i and u_i is a block-level IID random effect. The case count in block i is then obtained by aggregating r_{ij} via a weighted raster a_{ij} (i.e. the population raster)

$$cases_i = \sum_{j=1}^{N_i} a_{ij} r_{ij}$$

This is finally linked to the observed case count in block i through a Poisson likelihood

$$y_i \sim Pois(cases_i)$$

The posterior distribution is estimated using a Laplace Approximation, implemented through Template Model Builder (TMB) [99] which offers the necessary flexibility to specify this modified GAMM structure. The Laplace approximation is based on an assumption that the posterior distribution is multivariate Gaussian, therefore estimates are presented with 95% Gaussian confidence intervals calculated from the estimated standard errors.

Pixel level estimates of incidence from the disaggregation model were aggregated with weighting from the 1km population raster to obtain estimated case counts over each village polygon, and then rescaled by the estimated village polygon populations for comparison with observed incidence rates. For polygons in which the estimated village population was exactly zero, the incidence rate was defined as zero.

4.3.4 Validation

The strength of the disaggregation approach for predicting village level incidence will be interpreted relative to two alternative “benchmarks”. First, a baseline prediction will be defined such that all villages are uniformly predicted with the block level incidence rate (i.e. assuming homogeneity of incidence within blocks). This reflects the accuracy of assuming village-level incidence based on crude block-level surveillance.

Secondly, a random forest model [100, 101] will be fit directly to the village level data to serve as a ”gold-standard” for predicting village-level incidence in the hypothetical scenario that village-level data could be routinely available to guide decision-making. This non-parametric approach is commonly used to map spatially-varying phenomena due to its ability to accommodate complex non-linearities and interactions among given predictors (for example, it is the methodological basis for WorldPop’s global population estimates

[86]). This flexibility, in theory, maximises the information that can be gleaned about the pattern of village level incidence from the given data, hence is intended to represent the best that can be done with respect to predicting one village's incidence from nearby or otherwise similar villages.

A random forest model is formed of an ensemble of decision trees, within each of which the training data are partitioned according to splits defined on the given predictors in order to minimise the variation in the outcome for observations within each partition. The predicted value for a new observation is defined as the average outcome of all training observations in the partition within which it falls. Spatial structure will be incorporated by including the latitude and longitude of each village centroid as predictors. To increase robustness against over-fitting, a random subset of predictors are considered when determining each split. The sensitivity of the fit to the size of this subset will be assessed by comparison of four alternatives, considering two, three and six predictors out of the total ten.

The baseline and disaggregation models may both be evaluated against village-level data which were not used for fitting. The random forest model, however, is directly fit to the village-level data therefore a cross-validated measure of predictive power is required. This can be evaluated across "out-of-bag" (OOB) observations; each decision tree within the model is trained with a random subset of observations, therefore predictions can be defined for each observation by averaging the predictions of every tree from which the point was excluded. In this case, 200 trees were trained therefore each observation is predicted out-of-bag 200 times. The correlation (Spearman's rank correlation) and root mean squared error (RMSE) are used to compare between observed village-level incidence and the baseline, OOB random forest and disaggregation-based predictions. Approximate confidence intervals for the correlations are calculated as suggested by [102].

Sensitivity to population estimates

For the majority of villages affected with VL between the years 2013 and 2018, CARE India has estimated population sizes based on their own enumeration during routine visits. The WorldPop raster data yield village populations of broadly similar magnitude to these more accurate, locally-informed estimates, but with substantial noise (Supplementary figure C.1). The robustness of the model validation and comparison to this estimated denominator will therefore be investigated by repeating the comparisons using the CARE estimates to calculate incidence, across villages for which an estimate is available.

4.4 Results

4.4.1 Data cleaning

The raw village incidence data included sixty villages out of 2,186 affected during 2018 which were missing GPS coordinates and hence could not be directly linked to a village polygon, 40 of which fell within only three blocks (Barauli, Bhorey and Kuchaikote in Gopalganj district). As far as possible, these villages were manually matched to polygons according to district, block, gram panchayat (a local unit of usually multiple villages) and village names; five villages (7 cases) were unidentifiable and hence excluded.

Two villages had GPS coordinates which placed them substantially outside the state boundary; data errors were identified in the latitude variable and corrected. A further 8 villages had coordinates which fell marginally outside the boundary; only two of these had reported cases which were attributed to the nearest village polygon, with a tolerance of 500m (Supplementary figure C.2).

When aggregating the population count raster to these polygons, 83 village shapes (0.2%) were calculated to have zero population. None of these were attributed with any reported cases and were therefore ignored in comparison of incidence rates.

4.4.2 Descriptive

The primary analysis data consist of a total of 3,609 new cases of VL diagnosed throughout 2018, across 1,900 villages in 332 blocks. Based on estimated village population counts, block level incidence ranged from zero to just under 6 cases per 10,000 residents (Figure 4.4.1A). On average, village cover an area of one to two square kilometers, while blocks are on a scale of several hundred. The vast majority of incidence was observed across a cluster of blocks in the north-west of the state (Figure 4.4.1B), historically a persistent focal area for VL. The south of the state, partitioned by the Ganges river, observed little to no incidence across the year.

Assuming homogeneity of incidence within blocks implies substantially different expected village level case counts than were observed (Figure 4.4.2). In particular, observed incidence is more sparse and clustered, with many more villages observing either zero or greater than two cases than expected from block level incidence rates. This is supported through calculation of Moran's I statistic, which demonstrates substantial evidence of spatial auto-correlation in observed incidence on the scale of both villages and blocks (Supplementary figure C.3). When evaluated by each block individually, the strength

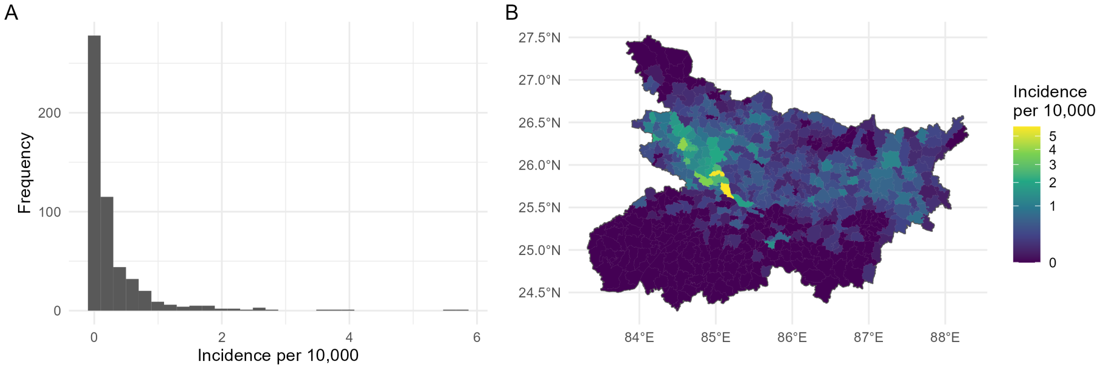


Figure 4.4.1: Observed block level incidence of reported VL for 2018. The vast majority of blocks saw zero or very low levels of incidence, in particular across the south of the state. A cluster of blocks in the west and (to a lesser extent) the east experienced moderate to high levels of incidence.

Table 4.4.1: Summary of estimated population size and standardised covariates included in the disaggregation model, averaged across village polygons and further stratified by village VL status for 2018 (affected/unaffected). Summary values are median [IQR].

Variable		**Overall (N = 44,794)**	**Affected (N = 1,900)**	**Unaffected (N = 42,894)**
Population size		1398.8 [595.1, 3077.1]	3934.8 [2064.4, 8096.2]	1339.1 [571.3, 2916.7]
Elevation (metres above sea level)		62.92 [51.4, 82.84]	57.41 [50.37, 64.59]	63.48 [51.48, 84.22]
Distance to nearest water body (kilometres)		0.86 [0.49, 1.61]	0.58 [0.37, 0.9]	0.88 [0.5, 1.65]
Travel time to nearest urban centre (minutes)		11.75 [5.63, 19.59]	9.73 [4.94, 15.88]	11.87 [5.66, 19.78]
Land surface temperature	Mean	30.27 [29.25, 31.2]	29.73 [28.84, 30.39]	30.3 [29.28, 31.23]
(degrees celsius)	SD	5.45 [4.68, 6.34]	4.85 [4.41, 5.27]	5.5 [4.7, 6.38]
Normalised difference vegetation index	Mean	0.48 [0.45, 0.52]	0.48 [0.44, 0.51]	0.48 [0.45, 0.52]
(range 0-1)	SD	0.16 [0.14, 0.18]	0.16 [0.14, 0.18]	0.16 [0.14, 0.18]

of correlation between constituent villages did not appear to correlate with overall block incidence (Supplementary figure C.4) and in fact very few blocks gave an indication of correlation between their constituent villages.

Covariates

When averaged across village polygons, no clear differences were apparent between affected and unaffected villages with respect to the included covariates, from a univariate perspective (Table 4.4.1). See Supplementary figure C.5 for the raw spatial distribution of all included covariates.

4.4.3 Disaggregation model fit

The smooth spatial field contributes substantially to the overall model fit, attenuating much of the effect of the covariates and rendering the corresponding coefficients as insignificant (Figure 4.4.3A). A fit based only on covariates and the block-level IID effect

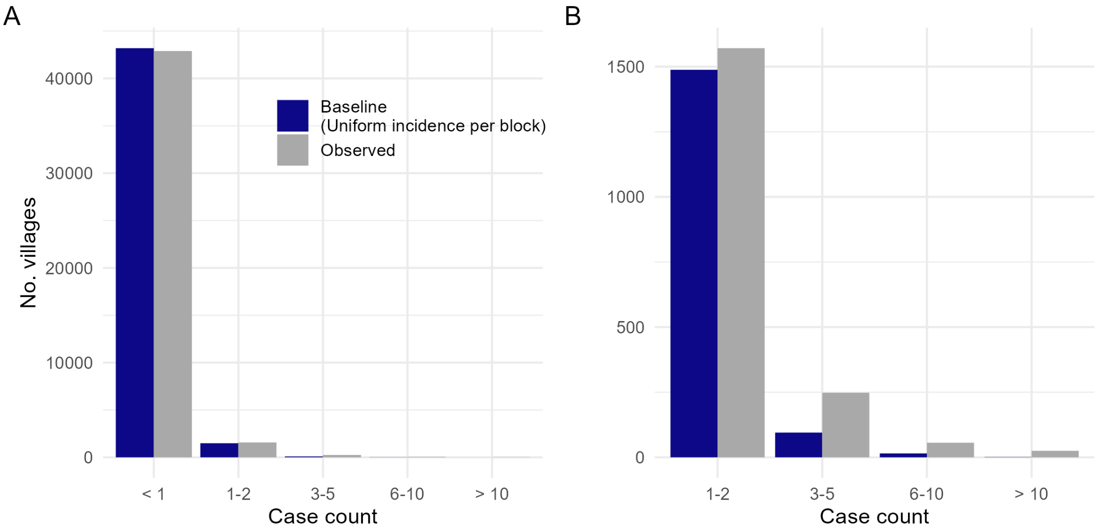


Figure 4.4.2: Observed village incidence for 2018 compared to that which would be expected assuming uniformity of incidence across each block. Panel A illustrates the full distribution and panel B shows detail of the distribution excluding zero-case villages

Table 4.4.2: Summary of agreement between observed village level incidence and predicted, from baseline, disaggregation and village-level models.

Model	Correlation (Spearman's rho [95% CI])	RMSE
Baseline	0.25 [0.245, 0.263]	1.60
Disaggregation: covariates + block IID	0.23 [0.224, 0.242]	3.90
Disaggregation: Full model	0.24 [0.231, 0.250]	3.39
Village-level random forest (OOB predictions)	0.19 [0.176, 0.194]	1.50

suggests that greater VL incidence at the village level is associated with closer proximity to water, lower annual variation in temperature, and lower annual average and greater variation in the vegetation index. The only association for which significance persists in the full model is that with annual variation in NDVI, with greater variation being associated with greater village incidence. Figures 4.4.3B and C illustrate the predicted per-pixel case count from the full disaggregation model and the fitted spatial field.

4.4.4 Model validation and comparison

Upon aggregating these predictions and comparing to observed village level incidence, neither version of the disaggregation model improved on a baseline prediction applying the block-level incidence rate (Table 4.4.2). Out-of-bag predictions from the village-level random forest model attained a slightly lower RMSE than the baseline, but overall had the weakest correlation with observed incidence.

This seemingly contradictory result appears to arise from a negative correlation between

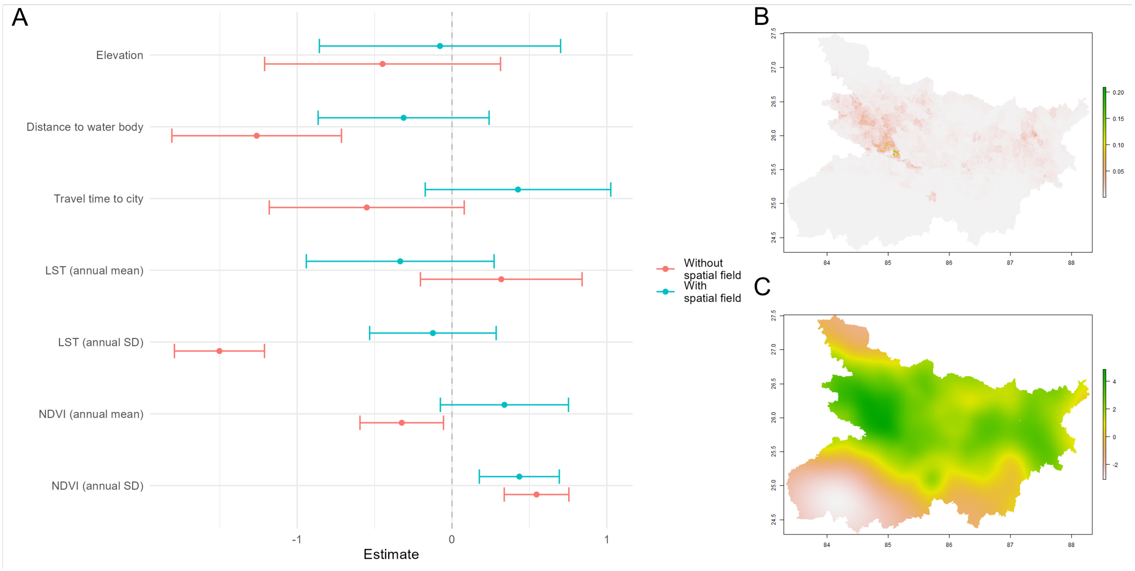


Figure 4.4.3: (A) Estimated covariate coefficients (log-scale) from disaggregation model fits, with and without the smooth spatial field. Point estimates are presented with 95% confidence intervals.(B) Predicted village level incidence and (C) fitted spatial field from the full disaggregation model.

observed and predicted values for villages within moderate-incidence blocks (between 0.5 and 1.5 cases per 10,000; Figure 4.4.4A). For blocks within this category, it seems that the random forest model predicts higher incidence in villages for which lower incidence was observed. However, correlations for all models are very weak when calculated within categories of block endemicity, reaching only as high as 0.15-0.2 in the low category. Much of this correlation will also likely come from accurately predicting zero cases for villages in blocks with zero cases.

Assessing predictions individually, the observed and predicted magnitude of incidence in non-zero villages showed some linear correspondence but with substantial noise (Figure 4.4.5). Disaggregation visually appeared to yield somewhat greater discrimination between affected and unaffected villages than the baseline. Supplementary comparison based on CARE's population estimates demonstrates even weaker correlation between observed and disaggregation-predicted values (Supplementary figure C.6).

The random forest fit was only estimated to have explained around 6.6% of the variation in observed incidence, which decreased with the number of variables tried at each split (to a minimum of 2% when all ten variables were used). The distribution of out-of-bag predictions more closely replicated the observed than the disaggregation-based predictions (Figure 4.4.6), but still under-predicted overall.

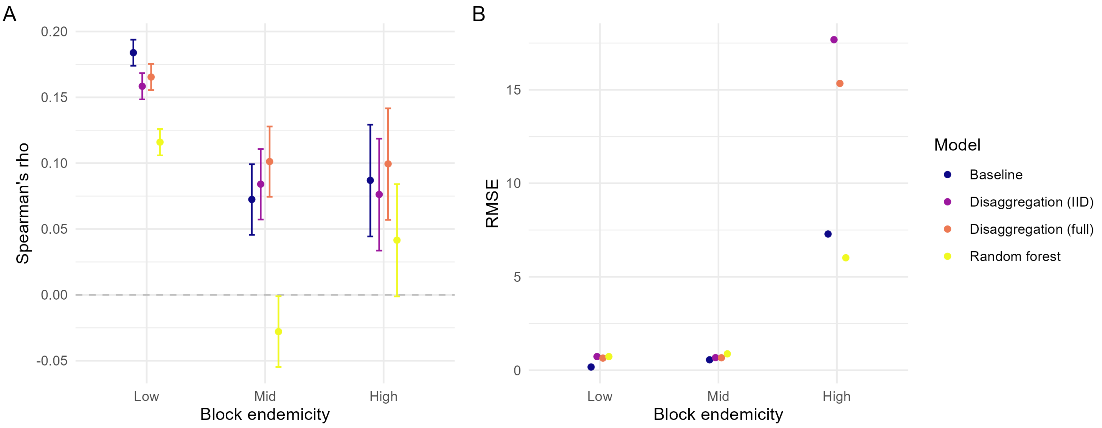


Figure 4.4.4: Comparison of predictive accuracy with respect to Spearman's rho (panel A) and RMSE (panel B) between models, stratified by block endemicity. Approximate 95% confidence intervals are illustrated for the former. Block endemicity categories are defined by observed incidence of less than 0.5 cases per 10,000 (n = 37,354), greater than 0.5 but less than 1.5 cases per 10,000 (n = 5,321) and greater than 1.5 cases per 10,000 (n = 2,119).

4.5 Discussion

Analysis of village level VL incidence on this scale has not previously been feasible. The work of CARE India's field teams to enumerate the villages of Bihar, defining a master list to which every diagnosed case may be linked, has allowed incidence to be calculated and investigated at the village level, where previously this was only possible by block. Further recent efforts to geo-locate all VL-affected villages opens the possibility of exploring spatial patterns at this scale, linking cases by geographic proximity and to the local environment. This is, as far as we are aware, the first state-wide analysis of village-level VL incidence in Bihar.

Disaggregation regression provides an opportunity to interrogate fine scale variation from the type of administrative level surveillance data which is routinely available in many endemic / elimination settings. In this example, the approach was not found to be effective for estimating village level burden of VL from block level surveillance data. Moreover, even when fitting a model directly to village-level data and allowing for more complex, non-linear relationships with the local environmental conditions, it was still not possible to accurately predict incidence at withheld villages.

Evidence of heterogeneity is observed within blocks at the village level; however, a simple assumption of uniformity within blocks crudely captures the broader spatial patterns across the state and therefore still provides somewhat reasonable predictions of village level incidence overall. Preliminary investigation of spatial auto-correlation at the two scales supports the idea that patterns of correlation are evident on the broader, block-level

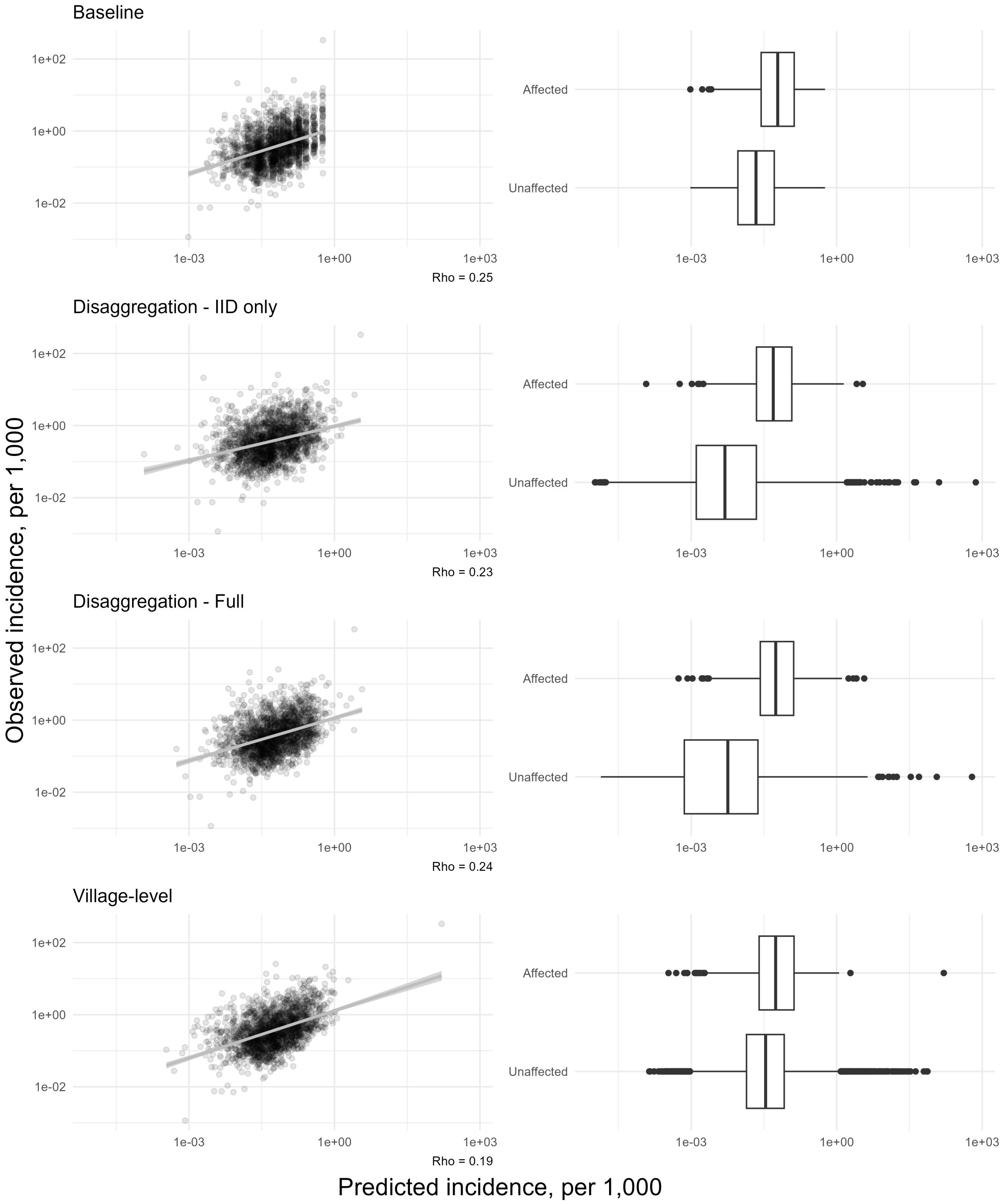


Figure 4.4.5: Comparison of predicted to observed village incidence rates, with respect to magnitude and presence/absence. Scatter plots only include *affected* villages, with non-zero observed and predicted incidence. Grey lines illustrate a simple linear trend of observed against predicted. The x-axes in both columns are limited between 1e-5 and 750 per 1,000.

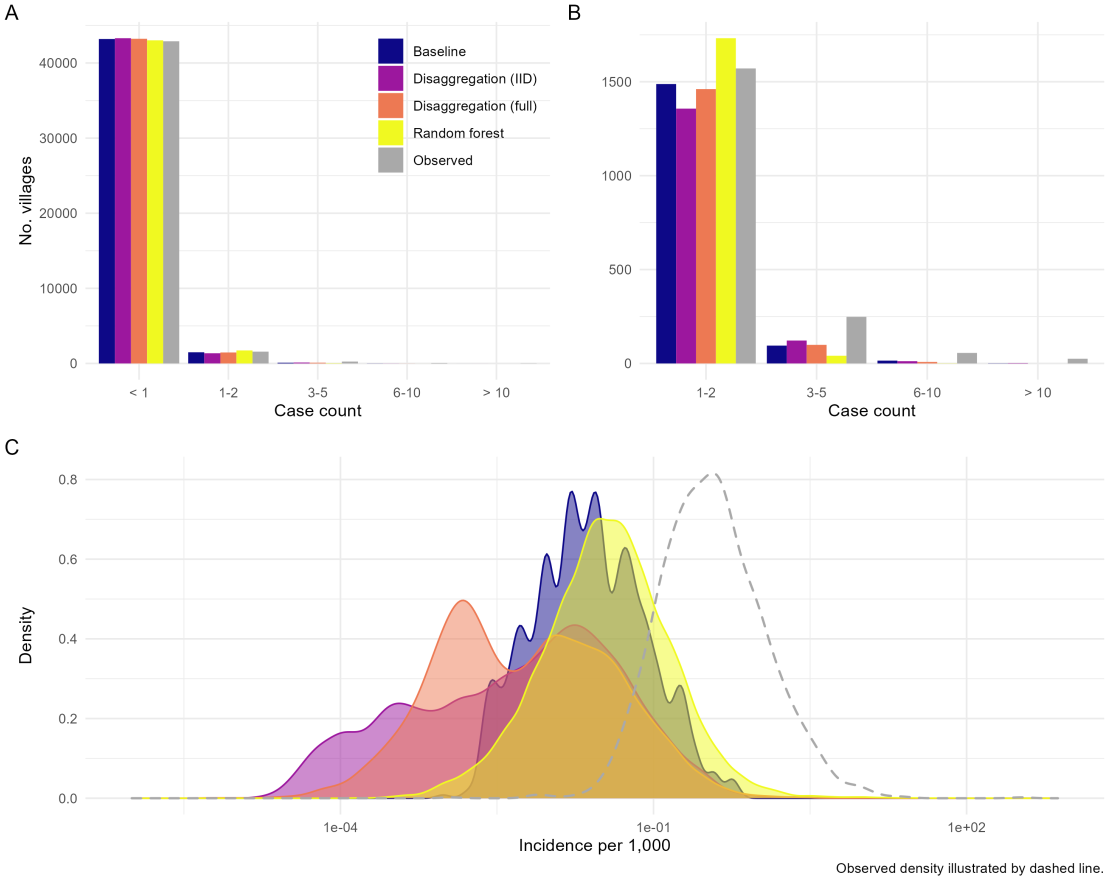


Figure 4.4.6: Distribution of observed versus predicted case counts from each candidate model. Panel (A) illustrates the full distribution and panel (B) shows detail of the distribution excluding zero-case villages. All models slightly overestimate the total number of unaffected villages (defined here as villages in which the expected case count is less than 0.5), and underestimate villages with a higher case count. (C) Overall densities of predicted village incidence rates from each modelling approach compared to the observed (dashed line).

but not necessarily between neighbouring villages. As has been demonstrated previously [83, 103], we observed that cases were clustered within villages, with three or more cases observed in substantially more villages than would be expected from uniform within-block incidence. This did not, however, appear to be informative of incidence in the surrounding villages.

Previous work estimating the spatial range of sandfly movement and of human-to-human transmission supports the same conclusion that a long range of direct transmission is unlikely [79, 103, 104]. Bihar has a highly mobile population with many migrant workers [105], which has been linked to increased VL risk [106, 44] and other health concerns for the worker and their accompanying family [107, 108]. It may be the case that VL outbreaks are more often triggered by importation from infected humans from more highly endemic regions, as opposed to infected sandflies, making longer distance movement of people a critical mechanism in the persistence of transmission at this stage of elimination.

There are a number of potential explanations for the poor performance of disaggregation against the simpler model. Firstly, the strength of the approach depends on associations with spatial covariates from which to infer that local variation. Despite the biological link between VL transmission and the environment via the sandfly vector, the environmental characteristics considered here were not found to have strong relationships with observed VL incidence. This is explicitly demonstrated by the poor performance of the village-level random forest model. Previous geostatistical and ecological analyses of environmental risk factors have demonstrated some evidence of association across a range of variables, but within a much more limited set of locations, and in some cases only indirectly with respect to sandfly abundance rather than VL incidence [80, 84, 109]. In this analysis, only annual variation in the vegetation index had an association with incidence that was robust to the addition of the spatial field. This could be linked to differences in agricultural practices between higher and lower incidence regions; however, likely correlation between covariates means that individual effects should not be over-interpreted.

Socio-demographic factors will also play a role in facilitating transmission - either through increased exposure or decreased access to care - but are not usually feasible to measure or estimate on a fine and continuous spatial scale. For example, sleeping and defecating outdoors increases exposure and is more common in less affluent, rural areas where VL burden is high [78, 106, 110]. Such mechanisms could have been captured in part by travel time to the nearest urban centre, yet this was not found to be informative in either the disaggregation or village-level models. A more relevant - and still continuously-measurable

- factor may be travel time to a health facility which offers VL diagnosis and/or treatment, since not all public health facilities in the state are equipped to offer this. For vulnerable populations living in poverty, there will be a large financial barrier associated with this distance that could delay intervention and extend opportunity for onward transmission. It has also been suggested that cases of VL-HIV co-infection and post-kala-azar dermal leishmaniasis (PKDL) likely make an increasingly important contribution to the persistence of transmission [79, 111]. These are not as thoroughly recorded through routine channels as primary VL infection is, therefore the spatial distribution of them is even less well understood.

There are examples in which area level data are combined with data collected at specific point locations (for example from prevalence surveys) within a joint model that draws on the information of both spatial scales [87, 13, 89, 112]. Wilson and Wakefield [112] found deterioration of accuracy when fitting only to areal census data versus point and areal, which worsened when cases were split across larger areas. Incorporating some village-level data into the disaggregation via a joint model, even if limited to a few focal locations, may improve the accuracy of prediction on this scale.

4.5.1 Limitations

The fitting of the disaggregation models assumed simple linear relationships with log-transformed incidence, whereas the true underlying dynamics may be highly non-linear. However, non-linearity was not evident in preliminary scatter plots of incidence versus village-averaged covariates, and the more flexible random forest approach did not yield any improvement on prediction. Inferring the appropriate functional form of covariates within a spatially-indexed model is complex, since spatial patterns in covariates which drive the outcome may be easily absorbed by spatially-correlated random effects [113]. This was evident here from the change in model coefficients when a spatial field was included. There may be scope for developing the current implementation of disaggregation regression to employ restricted spatial regression as suggested in [113], fitting the spatial random effects only within the residual space after adjustment for the specified fixed effects. Lucas et al. [87] demonstrated the use of machine learning techniques to first identify relevant non-linear relationships with covariates from point-prevalence data to then feed into a disaggregation model, and found that this improved accuracy relative to a baseline using only the raw covariates.

The definition of village-level incidence applied here is also limited. The list of villages defined in KAMIS does not uniquely link to official population estimates from the national census and recent, locally-informed estimates were only available for villages that have been visited by CARE's field teams for VL surveillance purposes. The population denominators estimated instead from WorldPop's 2015 global estimates do not closely align with CARE's estimates where available (median and IQR of 1400 [600 - 3080] and 3160 [1720 - 5800], respectively), and include some unrealistically extreme population sizes. In particular, the WorldPop data appear to overestimate the population in the state capital Patna, perhaps due to differences in the definition of "village" shapes which constitute wards of the city. Supplementary comparison based on CARE's population estimates in fact resulted in weaker correlation between predicted and observed incidence, although this may be in part due to the smaller number of observations for which the alternative estimate was possible.

Finally, the case counts with which the models were fit and against which they are validated are only those which have been observed and reported; if observation and/of reporting are spatially biased then this could be a poor validation set. Village-level targeting of active case detection means that cases may be more likely to be observed in and around historically affected villages, which may induce or exaggerate patterns of spatial auto-correlation. Such patterns driven by the observation process and not by underlying transmission would not necessarily be explained by the environmental covariates that were considered.

4.5.2 Conclusions

The possibility of inferring fine scale variation in disease burden from large-scale routine data through disaggregation regression would be incredibly valuable to policy makers, in particular in resource-constrained elimination settings. This analysis, however, highlights practical limitations that commonly arise with surveillance in such settings. At this stage of near-elimination of VL in Bihar, incidence appears largely stochastic. It's possible that relationships between the environment and transmission that naturally arise from the underlying biological mechanisms have been broken down by intensified control efforts, their patterns becoming increasingly fragmented as incidence has fallen to very low levels. Cases continue to arise in within-village clusters, yet this incidence does not appear to be informative for incidence in neighbouring villages.

We conclude that local level VL surveillance most likely is necessary for effective targeted interventions, but that the value of this information is largely in the ability to rapidly respond and detect secondary cases village by village, rather than in the ability to then anticipate incidence in the surrounding area. A fully geographically-targeted approach does not seem feasible given the stochastic nature of incidence that we have observed. Broader patterns of spatial correlation across the state as a whole may show the observable impact of varying interventions between endemic and non-endemic regions. An alternative (or complementary) approach that specifically targets surveillance among mobile population groups at risk of (re-)introducing infection to previously unaffected villages (for example through migration for work or marriage) may instead be justified.

Chapter 5

Spatial variation in diagnosis delay for visceral leishmaniasis in Bihar, India

We conclude from chapter 4 that the observed distribution of VL burden across villages appears to be highly heterogeneous and more complex than can be explained by simple biological links between local environmental conditions and transmission. It further raises the question of the extent to which surveillance *effort* biases the observed distribution of incidence, an issue which would likely be amplified on a finer and more sparsely burdened geographic scale.

This paper investigates potential variability in surveillance in different parts of the state, considering the length of delay to diagnosis experienced by observed cases as a potential indicator for the strength of surveillance in the patient's village. One particular focus is the hypothesis that the presumption of a block being 'endemic" or 'non-endemic" for VL (based on historical incidence) may contribute to such variability, since effort and resources tend to be more concentrated within areas assumed to be most highly-endemic. When this assumption of risk is only informed by historical patterns of incidence, it could establish a self-fulfilling prophecy in that the disease is only observed within areas it is expected to be observed. This analysis primarily focusses on diagnosis delay as a proxy measure of surveillance strength, however these delays also have direct implications for elimination with respect to the risk of onward transmission, particularly in low-endemic areas.

This version of the manuscript was accepted for publication by The Lancet Regional Health - Southeast Asia in April 2023.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1600363	**Title**	Ms
First Name(s)	Emily Sara		
Surname/Family Name	Nightingale		
Thesis Title	Spatio-Temporal Patterns and Surveillance of Infectious Disease During Emergence and Elimination		
Primary Supervisor	Prof Graham Medley		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	The Lancet Regional Health: Southeast Asia
Please list the paper's authors in the intended authorship order:	Emily S Nightingale, Joy Bindroo, Pushkar Dubey, Khushbu Priyamvada, Aritra Das, Caryn Bern, Sridhar Srikantiah, Nupur Roy, Tanu Jain, Naresh Gill, Mary M Cameron, Tim C D Lucas, Graham F Medley, Oliver J Brady
Stage of publication	**Submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I conceptualised and conducted the analysis based on data collected for a previous analysis by Pushkar Dubey and co-authors, who provided guidance on the field context and approved the final manuscript. I created the visualisations, drafted and revised the manuscript and wrote and published the underlying code (https://github.com/esnightingale/vl-spatial-diagnosis-delay). Tim Lucas provided input from a methodological perspective, and reviewed the results and manuscript along with Graham Medley and Oliver Brady.

SECTION E

Student Signature	Emily Nightingale
Date	12 December 2022

Supervisor Signature	Graham Medley
Date	12 December 2022

Spatial variation in delayed diagnosis of visceral leishmaniasis in Bihar, India.

Emily S Nightingale[1], Joy Bindroo[2], Pushkar Dubey[2], Khushbu Priyamvada[2], Aritra Das[2], Caryn Bern[3], Sridhar Srikantiah[2], Nupur Roy[4], Tanu Jain[4], Naresh Gill[4], Mary M Cameron[5], Tim C D Lucas[6], Graham F Medley[7], Oliver J Brady[1]

[1]Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK.
[2] Bihar Technical Support Program, CARE-India Solutions for Sustainable Development, Patna, India.
[3] Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, United States.
[4] National Centre for Vector Borne Diseases Control, Government of India, Delhi, India.
[5] Department of Disease Control, London School of Hygiene and Tropical Medicine, London, UK.
[6] Department of Health Sciences, University of Leicester, Leicester, UK.
[7] Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, UK.

Abstract

Background

Visceral leishmaniasis (VL) is a debilitating and - without treatment - fatal disease which burdens the most impoverished communities in northeastern India. Control and, ultimately, elimination of VL depends heavily on prompt case detection, yet a proportion of cases remain undiagnosed many months after symptom onset. Delay to diagnosis increases the chance of onward transmission, and poses a risk of resurgence in populations with waning immunity.

Methods

The spatial distribution of diagnostic delays was explored using a Bayesian model fit to geo-located cases using INLA, assuming days of delay as Poisson-distributed and adjusting for individual- (age, sex, HIV) and local-level (recent incidence, vector control, health facility access) characteristics. Residual variance was modelled with an explicit spatial structure. Cumulative delays were estimated under different scenarios of active case detection coverage.

Findings

The 4,270 cases analysed were prone to excessive delays outside existing endemic “hot spots”, beyond the focus of interventions. Cases diagnosed within recently-affected blocks and villages experienced shorter delays on average (by 13% 95% CrI [2.9% - 21.7%] and 7% [1.3% - 13.1%], respectively) than those in non-recently-affected areas.

Interpretation

Delays to VL diagnosis when incidence is low could influence whether transmission is interrupted or resurges. Narrowing surveillance to priority, high-burden areas may increase the likelihood of excessive delays in peripheral areas. Active surveillance driven by observed incidence may miss the risk posed by as-yet-undiagnosed cases in low-endemic areas, and be insufficient for achieving and sustaining elimination.

Funding

This work was funded by BMGF.

Research in context

Evidence before this study: We searched PubMed using the terms “leishmaniasis, visceral”, (“diagnosis” or “seeking”), (“delay*” or “duration”) and (“India” or “Nepal” or “Bangladesh”) for all articles published up to 11th April 2022, yielding 40 results. Three modelling studies demonstrate the important role of diagnosis delays in both transmission and evaluation of control efforts and three studies found delay to be associated with higher mortality risk. Four studies of sampled VL patients suggest that risk factors for delay include age, sex, HIV status, socio-economic and cultural factors, awareness, misdiagnosis, physical access and availability of diagnosis. Preference for private practitioners was found to be a key driver of delay in five studies and two discussed delays in relation to patient costs. One study concluded a benefit of active case detection in reducing length of delay among all reported cases in the 18 months following its implementation, having accounted for important variability by age, sex, and HIV status. Two studies simply summarised delays observed among sampled populations, one in an outbreak setting. From a spatial perspective, five studies considered variation in promptness of diagnosis between large administrative units such as districts or countries. Some related this to control activity in the area but did not investigate whether differences could be explained by the presence of other risk factors in the population.

Evidence suggests that delays to diagnosis should be a key concern for the elimination programme. The geographic distribution of VL cases experiencing excessive delays has, however, not been explored beyond a coarse scale. Moreover, few studies estimating average delays or investigating risk factors were conducted after the initiation of intensified case detection in 2017, which has likely had a strong influence on the efficiency of diagnosis. Many are now over a decade old and therefore not representative of the current, near-elimination context.

Added value of this study:

As observed in some previous studies, older age and HIV positivity were associated with longer delays to diagnosis. In addition, cases resident in villages which have experienced recent incidence, and more broadly in blocks officially classified as VL endemic, were found to on average report shorter delays. It was also found that areas in which cases appear prone to excessive diagnosis delays, unexplained by included covariates, do not coincide with areas with highest incidence burden. Further investigation of model-predicted delays suggested that the return on active case detection efforts with respect to days of delay avoided may vary between endemic and non-endemic blocks.

Implications of all the available evidence: There are specific subgroups of the population who are at risk of excessive diagnosis delays, whether due to their individual characteristics, their geographic location relative to interventions, or both. These individuals could be reservoirs of infection in their community for many months, allowing transmission to persist and potentially triggering outbreaks. There is evidence to suggest that incidence-based targeting of case detection may not capture all areas of concern for transmission. *Adaptive* targeting of active case finding could reduce diagnostic delay by considering a range of individual, spatial and historical factors - for example not only past incidence but also reported delays among recently detected cases. These findings are relevant not only for the specific case of VL in Bihar, but also for any elimination setting in which prompt case detection is an important pillar of the elimination strategy.

Introduction

Control of visceral leishmaniasis (VL) on the Indian sub-continent depends on prompt detection and treatment of cases through recognition of clinical symptoms or screening in affected areas. Early symptoms of VL are non-specific (including fever, fatigue and weight loss) and, especially where VL awareness is low, misdiagnosis is common. As a result, those afflicted may go undetected for several months or - in extreme cases - years, despite the presence of active detection measures. Evidence suggests that longer time to diagnosis is associated with increased mortality risk (1) and undetected cases serve as reservoirs of infection in their community, allowing transmission to persist and forming a barrier to achieving and sustaining elimination (2).

An improved programme of active case detection (ACD) (3) was initiated in 2016/7 and its efficacy in reducing overall time to diagnosis has been demonstrated (4). However, a non-negligible proportion of cases are diagnosed several months after onset of symptoms. Dubey et al. (4) report that during the first 19 months of improved ACD in Bihar, 66% of diagnosed cases reported symptom onset greater than 30 days prior to diagnosis, and 10.5% greater than 90 days prior. Le Rutte et al. (5) estimated in 2017 that elimination of VL in the Indian sub-continent could be achieved by 2020 with sufficient coverage of vector control, "*provided that the average onset-to-treatment (OT) time does not exceed 40 days".* The persistence of this minority of cases with long delays to diagnosis is therefore deserving of further investigation.

Barriers to diagnosis of VL have been investigated in several previous studies. VL burden is broadly associated with the most socially and economically disadvantaged communities in India (6) and, despite government compensation for expenditure to access VL diagnosis and treatment, patient costs remain an important barrier (7). Mondal et al. (8) screened households in sampled villages and found a high proportion of undiagnosed cases in districts not well-served by health care facilities, and a lower proportion in districts with greater availability of VL care (i.e. districts considered affected/endemic in which the programme is active).

Rahman et al. (9) interviewed VL patients in Bangladesh and found logistical barriers to prompt diagnosis such as remoteness of the health centre, wet season transport limitations, restricted ability to travel due to and limited availability of RK39 rapid diagnostic tests in the area. This was combined with lack of understanding due to illiteracy, lack of recent incidence and preference for first consulting more local traditional healers. In Bihar there is also widespread use of private and informal health practitioners which can cause additional delays (10).

Dubey et al. (4) explored patient characteristics associated with longer delays between symptom onset and diagnosis among all cases of VL diagnosed between January 2018 and June 2019. It was concluded that younger age and detection via active surveillance were associated with shorter delays, while male sex and HIV positivity were associated with longer delays.

What has not been considered is where *geographically* individuals are experiencing excessive delays, in relation to each other and in relation to the activities of the control programme. Control and surveillance of VL in Bihar is targeted according to recently observed incidence (11), resulting in a spatially-varying intensity of intervention. This work aims to investigate the spatial distribution of delays and understand some of its potential driving factors.

Materials and Methods

Data sources

This work is based on secondary analysis of data collated for a previous study (4) evaluating active case detection measures. Case reports of individuals diagnosed between 01/01/2018 and 31/07/2019 (N = 5,030) were cross-referenced with suspect case registers over the same period in order to identify the route of detection for each patient as active (via targeted surveillance) or passive (self-referral).

Low specificity of the recommended rapid diagnostic test (RK39) requires that only cases suffering at least 14 days of fever are suspected for VL and eligible for testing (3). The primary outcome was therefore defined as the reported duration of fever prior to diagnosis *beyond* the standard criteria of 14 days, hereafter referred to as “diagnosis delay”. This was considered theoretically avoidable delay within the current guidelines. Cases diagnosed *within* 14 days of fever onset were not considered comparable to the rest of the population and excluded.

Village locations

Location data are available for every village with at least one case reported to the Kala-Azar Management Information System (KAMIS) from 2013-18. Cases described above were linked to their resident village and corresponding GPS location via a unique ID. Location data were predominantly unavailable for villages not affected between 2013-18.

Health facility access

Capacity for diagnosis and treatment of VL is not consistent across all health facilities in Bihar, as treatment centres were originally established to be near the most affected villages (12) (*Supplementary figure S1 (A))*. A tool developed by The Malaria Atlas Project (MAP) was used to estimate minimal travel time between villages and the available diagnosis and treatment facilities by relative “accessibility” (13), accounting for distance and ease of travel (*Supplementary figure S1 (B)*).

Methods

Baseline model structure

Reported diagnosis delay (in days) for each case, Y_i, is assumed to be Poisson-distributed with mean λ_i, with independent and identically distributed observation-level random effects (OLRE) to account for overdispersion (14). The model is fitted within a Bayesian framework using the Integrated Nested Laplace Approximation (INLA) approach.

Formally,

$$Y_i \sim Po(\lambda_i)$$
$$\log \lambda_i = \beta_0 + x_i$$

where

$$x_i \sim N(0, \sigma)$$

with a penalised-complexity hyperprior (20) set on the standard deviation σ, such that $P[\sigma > 1] = 0.01$. This penalises deviation from the simplest case in which the standard deviation is equal to 0 (i.e. constant) and specifies that the variance of these random effects is not expected to be greater than 1.

It is common for self-reported duration data to exhibit “heaping”, in which individuals show a preference for certain (usually rounded) intervals of time, and there has been suggestion that this behaviour may bias parameter estimates (15). The final model was therefore refitted with a binary outcome of delay exceeding 30 days, to assess the robustness of inferred covariate effects.

Covariates

Covariates at both individual and village level were considered within three domains: patient, village risk awareness and village accessibility.

Characteristics of the *patient* included age (standardised), sex (male / female), HIV status (positive / negative at diagnosis), marginalised caste status (scheduled caste or tribe / other), previous treatment for VL or post-kala azar dermal leishmaniasis (PKDL) (yes / no), occupation (none / unskilled / skilled / self-employed or salaried) and route of detection (ACD or passive/self-reported)

Village characteristics were defined under two domains. Block endemicity (endemic / non-endemic), targeting of indoor-residual spraying (IRS) (yes / no) and village incidence of VL (non-zero / zero) in the previous year (2017) were considered indicators of *risk awareness* in the local population. Estimated travel time (minutes) to the nearest diagnostic or treatment facility and diagnosis during the rainy season (June - September) were defined under the domain of *accessibility*.

Both ACD and IRS are incidence-targeted interventions, triggered by incidence during the last three years. As such, these variables are expected to be to some extent correlated with 2017 village incidence.

Estimated covariate effects are presented as risk ratios (RRs) with 95% credible intervals (CrI).

Variable selection

The association of each covariate with observed delay was explored through univariately within the baseline model structure. Multivariate models were then fit for each domain in turn, and significant covariates selected based on the adjusted coefficients' 95% CrI. A full model was then fit to include the selected covariates in all three domains.

Spatial analysis

The correlation between delays experienced in nearby villages was modelled with a spatially-structured random field over the GPS locations for all villages, using the INLA-SPDE approach for estimation (17). This approach approximates a spatially-continuous field via stochastic partial differential equations (SPDE) across a triangular mesh. A prior structure which penalises complexity was also assumed for the hyperparameters of this component (range and standard deviation). A range of prior specifications for the SPDE model were explored to assess sensitivity to this choice and are illustrated in *Supplementary figure S7*.

A spatial field was initially added to the baseline, OLRE-only model, to characterise the spatial pattern in the absence of the explanatory power of the covariates. Each covariate domain was then reintroduced in turn and finally in combination, resulting in the following structure:

$$\log \lambda(i) = \sum_j \beta_j c_j(i) + \sum_k \beta_k c_k(v_i) + s(v_i) + x_i$$

where $c_j(i)$ are individual-level covariate values for case i, $c_k(v_i)$ are village level covariate values for the village v_i of case i, $s(v_i)$ is the spatial random field and x_i the OLRE.

The contribution of each domain of covariates in explaining the spatial pattern of delays was explored via the percentage change in mean absolute value (MAV) across the fitted spatial field when each covariate domain was reintroduced. The percentage change in MAV of the OLRE was also calculated to assess the contribution of each in explaining the non-spatially-structured residual variation.

Model assessment

The value of including both covariates and an explicit spatial structure was assessed via Widely Applicable (also known as Watanabe-Akaike) Information Criterion (WAIC) and leave-one-out (LOO) cross-validation, relative to the baseline OLRE-only model. Model predictions were compared on the logarithmic score (logs) (18) and on the Brier score (19) for classification of delays greater than 30 days. Spatial and non-spatial cross-validation approaches were compared to assess the contribution of the spatial random field to prediction (see *Supplementary Materials B)*.

Final model prediction

The expected extent of excessive delays from the selected model were mapped over all affected districts. Predictions were calculated for a fine grid of points across the area, reflecting the expected delay for an arbitrary individual at that location, otherwise comparable on all covariates. The posterior distribution is summarised by a mean and an exceedance probability with a threshold of 30 days and plotted to form a smooth map. In particular, regions in which the predicted exceedance probability is above 0.5 (i.e. where delay longer than 30 days is more probable than delay within 30 days) are highlighted.

Impact of ACD

To explore the potential impact of extending or restricting ACD across endemic and non-endemic regions of Bihar, hypothetical delays were predicted under two scenarios of ACD coverage among the individuals in this study (0% and 100%). Predicted days of delay where either no or all cases were detected via ACD were compared to the expected delays with ACD as originally observed. The difference in terms of total person-days of delay was stratified by the endemicity of the block and summarised over 10,000 posterior samples to capture uncertainty.

Data statement

All analyses were performed in R version 4.1.2 (2021-11-01). The written code has been made available at https://github.com/esnightingale/vl-spatial-diagnosis-delay. The full analysis dataset cannot be publicly shared as it contains both sensitive (HIV infection) and identifiable (age, sex and GPS of resident village) information on individual patients.

Results

Data cleaning

Of 5030 patients diagnosed with VL between 01/01/2018 and 31/07/2019, 649 residents of villages with no known GPS location and one with an assumed erroneous GPS location substantially (>10km) beyond the state boundary were excluded. Two patients had been removed from KAMIS due to recognition of an error therefore were also excluded. A further 84 were excluded due to missing HIV status, caste status, occupation or VL/PKDL treatment history. HIV status had the greatest proportion of missingness at 1.3%. Excluding incomplete observations had negligible impact on the distribution of delays, with equal means (31 days) and quartile ranges (11-44 days) before and after exclusion (*Supplementary table S1*).

24 (0.5%) cases reported fever duration less than 14 days. Overall, patients diagnosed with less than 14 days of fever were younger, less likely to be female, more likely to reside in VL-endemic blocks and closer in travel time to diagnostic and treatment facilities (*Supplementary table S2*).

A summary of the data cleaning process is illustrated in *Supplementary figure S2* and a comparison of included and excluded cases presented in Supplementary tables *S1* and *S2*.

Descriptive

4,270 VL patients diagnosed within the study period and with complete covariate information and linked to a GPS-located village were included for analysis. These had reported duration of fever ranging from 14 to 510 days at the point of diagnosis. The geographic spread and distribution of diagnosis delay for included patients is illustrated in *Figure 1B*.

[Fig1]

A descriptive summary of characteristics of included patients is presented in Table 1, and an illustration of the full correlation matrix between all considered covariates is shown in Supplementary figure S3.

[Table 1]

Variable selection

Among patient-specific covariates, age, HIV status and detection by ACD were found to be associated with length of delay (estimated RRs and 95% CrI of 1.14 [1.13,1.15], 1.54 [1.31, 1.81] and 0.74 [0.69, 0.79] in univariate analyses, respectively; *Figure* 2). No clear association was found for caste status or VL/PKDL treatment history, with the direction of effect switching between univariate and multivariate analyses.

Within the “risk awareness” domain, block endemicity and non-zero village incidence in the previous year were associated with shorter delays. Estimated RRs for these two covariates were very similar in univariate analyses (0.85 [0.80, 0.91] and 0.86 [0.80, 0.91], respectively), suggesting that they may capture some of the same variation. Although IRS targeting had a negative effect in univariate analysis, this was lost when accounting for the other covariates in the domain (adjusted RR 0.99 [0.91, 1.07]).

Within the “access” domain, no clear univariate associations were found. When travel time was combined with season in multivariate analyses, time to treatment facility (in minutes) had a borderline positive association with delay (1.02 [0.99, 1.05]). For completeness, this covariate was selected for comparison of all three domains in later analyses.

[Fig2]

Spatial analysis and final model

Incorporating an explicit spatial structure in the residuals alongside the chosen covariates yields the lowest WAIC out of all models compared (*Table 1*). Gains on out of sample prediction are also evident, with respect to both log score and Brier score on predicting exceedance of 30 days.

[Tab2]

The estimated covariate effects from the final, spatial model were consistent with those from the non-spatial model (*Supplementary Figure S4*). Being aged one standard deviation above the mean (28 years) and being HIV positive were associated with a 13% (95% CrI [9.3% - 16.0%]) and 28% [9.2% - 49.4%] increase in delay, respectively. Diagnosis via active rather than passive case detection was associated with a 22% [17.9% - 26.8%] reduction in delay. In terms of local awareness of VL, patients residing in blocks considered endemic and villages with non-zero incidence in the year prior to diagnosis experienced 13% [2.9% - 21.7%] and 7% [1.3% - 13.1%] shorter delays, respectively, after adjusting for the sources of individual level variation described above. The final model gave some indication of an increase in delay with longer travel time to a treatment facility however the evidence for this remained weak.

The fitted spatial effect had a posterior range (the approximate distance beyond which correlation falls below 0.1) of 47km (95% CrI [26km - 84km]), and a standard deviation of 0.32 [0.23 - 0.42]. The standard deviation of the OLRE decreased from 0.99 [0.97 - 1.01] in the null model to 0.96 [0.94 - 0.99] in the non-spatial model, and finally to 0.93 [0.90 - 0.95] in the final, spatial model, as more of the residual variance could be explained by other components. Converting to a binomial likelihood to compensate for heaping did not substantially alter the inferred relative effects of the covariates (*Supplementary figure S5*).

Figure 3 (A) illustrates the spatial pattern of diagnosis delays estimated from the final model, assuming diagnosed cases are comparable on all factors apart from location. *Figure 3 (B)* translates these projections to exceedance probabilities, mapping the estimated probability of observing delay greater than 30 days at any location. Less opaque areas indicate where the probability is close to 0.5 and hence exceedance of 30 days is least certain. The pattern highlights regions in the north west (across Siwan, Gopalganj and Paschim Champaran districts), north east at the Nepal border (Supaul and Araria districts), and further south (Patna, Vaishali and Munger) across which delays are on average expected to be longer than 30 days. It also flags more focal regions of possible concern around Saraiya (Muzaffarpur district), Kalyanpur (Samastipur) and Sonbarsa (Sitamarhi) blocks. This pattern differs from that observed in total incidence (*Figure 3 (C)*); the cluster of higher incidence blocks between the Ghaghara and Gandak rivers north west of Patna is not reflected by a comparable cluster in the distribution of diagnosis delays.

[Fig3]

The map of predicted exceedance probabilities is illustrated in *Supplementary figure S6,* alongside an alternative to Figure 3B using a higher cut-off of 0.75.

Impact of ACD

In total over all observations, predicted total person-days of delay was reduced by just under 15% when ACD coverage was increased to 100% of cases, equating to a reduction of 7.7 (98% CrI [5.5 - 9.8]) days per case among those originally detected by passive case detection (PCD) (*Figure 4*). This reflects a reduction of 8.7 [6.2 - 11.0] days per reassigned case in non-endemic blocks, compared to only 6.7 [4.8 - 8.6] days in endemic blocks. By increasing ACD detection from its current value of 40.1% to 100%, the overall average estimated delay decreased by 4.6 days, from 31.5 to 26.9. See *Supplementary table S4* for a full table of estimates.

Conversely, in the complete absence of ACD (0% of cases), estimated total person-days of delay *increased* by around 9% - an average difference of 7 [4.6 - 9.6] days per case among those originally detected by ACD. The difference between endemic and non-endemic blocks is also clear in this scenario, with a greater increase observed in endemic blocks (7 [4.6 - 9.6] days per reassigned case) than in non-endemic blocks (6.3 days per reassigned case). In the absence of any ACD, the average estimated delay for all VL cases increased by 2.8 days.

[Fig4]

Discussion

In any disease elimination setting, a new set of challenges arises as incidence is suppressed to very low numbers. The effort required to detect each individual case grows rapidly, yet it is at this stage - when immunity and attention are potentially waning - that prompt detection is crucial to avoid resurgence. Sparsity of incidence across a broad geographic area prompts focussing attention and resources on specific areas considered to be most at risk based on recent observed data. However, this reactive approach may have unintended consequences for the observation of incidence, biasing surveillance as a result of feedback between case detection and detection effort.

This work highlights a geographic pattern (with a range of around 50km) to the villages in which cases experience the longest delays and motivates further investigation to understand what drives this pattern. Currently, areas of concern are identified for intervention according to recent observed incidence. However, ACD could be more effective if guided not only by incidence but by where delays are longest or most problematic for transmission. In all model fits, ACD was found to be strongly related to the time taken to obtain a diagnosis, associated with greater than 20% shorter delay on average than PCD. We estimated that if all cases in this study had been detected actively, the total person-days of delay accumulated during this period may have been reduced by nearly 15%. This translated to a reduction of 5 days per case in recently endemic blocks versus 4.2 in recently non-endemic, suggesting that gains from active detection in terms of person-days of delay avoided may be greater across recently non-endemic than endemic blocks. Characterising this spatial variation offers guidance to areas in which there is greatest scope to reduce delays - and hence transmission risk - through increased active surveillance coverage.

The inferred relationship between the length of delay and recent incidence in the region could reflect the impact of waning awareness and detection effort in areas which have not been recently affected. This concurs with previous work investigating variation in seeking of and access to VL diagnosis. A study in Nepal compared samples of districts included and excluded from the national control programme and found increased delays in care-seeking among patients in non-programme districts (20). Awareness and attitudes around VL have been evaluated in various settings, with one study concluding that this may affect the likelihood of treatment-seeking through appropriate channels (21) and another finding understanding to be lacking even among individuals having experienced VL in their household (22). The possibility should be considered that both the benefit of ACD and promptness of independent care-seeking may wane as we move closer to elimination.

Focusing attention on areas considered “high risk” from recently observed incidence may risk delaying diagnosis and treatment among the few cases which arise in low-endemic areas. This could be a concern since recent evidence has arisen of increasing, sporadic incidence of VL beyond the main endemic regions (23)**.** A study in Vaishali district (24) suggested the need to extend active efforts of vector control, case detection and community engagement to non-endemic but high-risk villages peripheral to hotspot areas; however, they conceded that there are substantial economic barriers to applying this intensive approach.

ACD is laborious and the cost severely limits its viability in areas with no recently reported cases. Yet, Dial et al. (12) make the case that bolstering efforts in meso- and low-endemic districts may prove to be cost-effective in the long term. We found evidence that ACD may have greater scope for reducing delay in less-endemic areas - perhaps since these communities lack awareness to promptly recognise symptoms and self-refer - providing further justification for maintaining robust surveillance here. However, it should be considered whether there is a more economical approach to active surveillance than its current form. A cost assessment could, for example, be made of a strategy not to intensively *detect cases* but to intensively *increase awareness* (of the disease, its diagnosis/treatment, and of PKDL) outside the assumed endemic areas.

If the past decade of efforts continue to be successful and incidence declines to near-negligible levels in many districts, our findings suggest that this may result in longer delays for the few remaining cases. Medley et al. (25) suggest that prompt diagnosis may be key for India to follow the examples of Nepal and Bangladesh in achieving elimination as a public health problem, but there is scope for further investigation of the consequence of delays among few cases on risk of outbreaks and resurgence. It is clear that key epidemiological features ought to be carefully and regularly monitored as programme objectives are achieved, generating feedback with which to periodically update procedures.

Limitations

Self-reported symptom durations are prone to bias; the raw data exhibit heaping at rounded time intervals and literature suggests that this behaviour can bias parameter estimates (15). However, refitting the final model for a binary outcome only reduced the precision of estimates rather than altering the estimated effects. The subset of observations not linkable to GPS locations or with other missing characteristics could also bias the observed spatial pattern of delays or estimated covariate effects. Moreover, the grouping of individual observations by village could mask or dilute important associations. It is the intention of KAMIS data managers that, going forward, each patient's data would be linked to an individual household location as opposed to only the village centre. The increased identifiability of these data would, however, need to be carefully navigated in order to take advantage of this finer information for the purposes of surveillance and analysis.

Our interpretation of ACD impact assumes no unobserved confounding in estimation of the intervention's effect. ACD is triggered by incidence in the last 12 months therefore this is a strong candidate for confounding, but is adjusted for with block and village level indicators in the model. A more rigorous analysis of ACD specifically, which considered assumed causal relationships between covariates in more detail, may better pinpoint where and in which populations its benefit might be greatest relative to the cost.

This analysis only describes the behaviour of symptomatic infection among detected cases. The observed delay data may under-represent the upper tail of the distribution since presence in the dataset is conditional on having recognisable symptoms and obtaining a diagnosis at all. The majority of infections with *L. donovani* are asymptomatic and resolve without intervention (26), yet xenodiagnostic evidence suggests that asymptomatics do not contribute substantially to transmission (27). If poorer detection of symptomatic cases overall corresponds with less prompt diagnosis as observed here, the absence of as-yet-undetected cases from the analysis could render our results conservative and suggest that inferred areas of longer delay could reflect an even greater problem in practice.

Also excluded are cases of post-kala-azar dermal leishmaniasis (PKDL), a more poorly-reported secondary form of leishmania infection (28) which may contribute increasingly to transmission as VL incidence declines (28). Delays to diagnosis are usually longer than for VL yet - since detection of PKDL can be a by-product of VL surveillance - may exhibit similar spatial patterns.

Conclusions

Reduction of avoidable delays to diagnosis and treatment is a key objective in the pursuit of visceral leishmaniasis elimination across the Indian subcontinent. Previous work has identified some groups at risk of delayed care-seeking, but we demonstrate that heterogeneity remains in the promptness of diagnosis across the state. This spatial variation may in part be explained by differences in risk awareness as a result of recent VL incidence in the community. Evidence suggests that returns on active detection may vary between regions at different stages of elimination, and we suggest that further mathematical modelling may clarify how delays could perpetuate transmission in low incidence areas. The efficacy of active case detection in reducing delays is clear, yet its intensity and geographic extent may need to be reassessed as the region approaches elimination.

References

1. Das A, Karthick M, Dwivedi S, Banerjee I, Mahapatra T, Srikantiah S, et al. Epidemiologic Correlates of Mortality among Symptomatic Visceral Leishmaniasis Cases: Findings from Situation Assessment in High Endemic Foci in India. PLoS Negl Trop Dis. 2016 Nov 21;10(11):e0005150.
2. Chapman LAC, Spencer SEF, Pollington TM, Jewell CP, Mondal D, Alvar J, et al. Inferring transmission trees to guide targeting of interventions against visceral leishmaniasis and post–kala-azar dermal leishmaniasis. Proc Natl Acad Sci. 2020 Oct 13;117(41):25742–50.
3. India National Vector Borne Diseases Control Programme. Accelerated Plan for Kala-azar Elimination 2017. 2017;1–94.
4. Dubey P, Das A, Priyamvada K, Bindroo J, Mahapatra T, Mishra PK, et al. Development and Evaluation of Active Case Detection Methods to Support Visceral Leishmaniasis Elimination in India. Front Cell Infect Microbiol. 2021 Mar 24;11.
5. Le Rutte EA, Chapman LAC, Coffeng LE, Jervis S, Hasker EC, Dwivedi S, et al. Elimination of visceral leishmaniasis in the Indian subcontinent: a comparison of predictions from three transmission models. Epidemics. 2017 Mar;18:67–80.
6. Boelaert M, Meheus F, Sanchez A, Singh SP, Vanlerberghe V, Picado A, et al. The poorest of the poor: a poverty appraisal of households affected by visceral leishmaniasis in Bihar, India. Trop Med Int Health. 2009 Jun;14(6):639–44.
7. Okwor I, Uzonna J. Social and Economic Burden of Human Leishmaniasis. Am J Trop Med Hyg. 2016 Mar 2;94(3):489–93.
8. Mondal D, Singh SP, Kumar N, Joshi A, Sundar S, Das P, et al. Visceral leishmaniasis elimination programme in India, Bangladesh, and Nepal: reshaping the case finding/case management strategy. PLoS Negl Trop Dis. 2009;3(1):e355.
9. Rahman KM, Olsen A, Harley D, Samarawickrema IVM, Butler CD, Zahid K, et al. Early diagnosis of kala-azar in Bangladesh: Findings from a population based mixed methods research informing the post-elimination era. Parasitol Int. 2021 Dec 1;85:102421.
10. Boettcher JP, Siwakoti Y, Milojkovic A, Siddiqui NA, Gurung CK, Rijal S, et al. Visceral leishmaniasis diagnosis and reporting delays as an obstacle to timely response actions in Nepal and India. BMC Infect Dis. 2015 Dec;15(1):43.
11. Srinivasan R, Ahmad T, Raghavan V, Kaushik M, Pathak R. Positive Influence of Behavior Change Communication on Knowledge, Attitudes, and Practices for Visceral Leishmaniasis/Kala-azar in India. Glob Health Sci Pract. 2018 Mar 21;6(1):192–209.
12. National Vector Borne Disease Control Programme. Standard Operating Procedure for Kala-Azar and Post-Kala-Azar Dermal Leishmaniasis Case Search [Internet]. 2020 [cited 2022 Mar 18]. Available from: https://nvbdcp.gov.in/Doc/SOP_Kala-azar_PKDL_Aug_2020.pdf
13. Dial NJ, Medley GF, Croft SL, Mahapatra T, Priyamvada K, Sinha B, et al. Costs and outcomes of active and passive case detection for visceral leishmaniasis (Kala-Azar) to inform elimination strategies in Bihar, India. PLoS Negl Trop Dis. 2021 Feb 3;15(2):e0009129.

14. Weiss DJ, Nelson A, Gibson HS, Temperley W, Peedell S, Lieber A, et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature. 2018 Jan 18;553(7688):333–6.
15. Harrison XA. Using observation-level random effects to model overdispersion in count data in ecology and evolution. PeerJ. 2014 Oct 9;2:e616.
16. Heitjan DF, Rubin DB. Ignorability and Coarse Data. Ann Stat. 1991;19(4):2244–53.
17. Moran PAP. Notes on Continuous Stochastic Phenomena. Biometrika. 1950;37(1/2):17–23.
18. Lindgren F, Rue H. Bayesian Spatial Modelling with R-INLA. J Stat Softw. 2015 Feb 16;63(1):1–25.
19. Czado C, Gneiting T, Held L. Predictive Model Assessment for Count Data. Biometrics. 2009;65(4):1254–61.
20. Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Mon Weather Rev. 1950 Jan 1;78(1):1–3.
21. Lim DJ, Banjara MR, Singh VK, Joshi AB, Gurung CK, Das ML, et al. Barriers of Visceral Leishmaniasis reporting and surveillance in Nepal: comparison of governmental VL-program districts with non-program districts. Trop Med Int Health. 2019 Feb;24(2):192–204.
22. Govil D, Sahoo H, Pedgaonkar SP, Chandra Das K, Lhungdim H. Assessing Knowledge, Attitudes, and Preventive Practices Related to Kala-A: A Study of Rural Madhepura, Bihar, India. Am J Trop Med Hyg. 2018 Mar;98(3):857–63.
23. Kumar Bhat N, Ahuja V, Dhar M, Ahmad S, Pandita N, Gupta V, et al. Changing Epidemiology: A New Focus of Kala-azar at High-Altitude Garhwal Region of North India. J Trop Pediatr. 2017 Apr 1;63(2):104–8.
24. Kumar V, Mandal R, Das S, Kesari S, Dinesh DS, Pandey K, et al. Kala-azar elimination in a highly-endemic district of Bihar, India: A success story. PLoS Negl Trop Dis [Internet]. 2020 May 4 [cited 2020 Sep 29];14(5). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7224556/
25. Medley GF, Hollingsworth TD, Olliaro PL, Adams ER. Health-seeking behaviour, diagnostics and transmission dynamics in the control of visceral leishmaniasis in the Indian subcontinent. Nature. 2015 Dec 3;528(7580):S102-108.
26. Hirve S, Boelaert M, Matlashewski G, Mondal D, Arana B, Kroeger A, et al. Transmission Dynamics of Visceral Leishmaniasis in the Indian Subcontinent - A Systematic Literature Review. PLoS Negl Trop Dis. 2016 Aug;10(8):e0004896.
27. Singh OP, Tiwary P, Kushwaha AK, Singh SK, Singh DK, Lawyer P, et al. Xenodiagnosis to evaluate the infectiousness of humans to sandflies in an area endemic for visceral leishmaniasis in Bihar, India: a transmission-dynamics study. Lancet Microbe. 2021 Jan;2(1):e23–31.
28. Zijlstra EE, Alves F, Rijal S, Arana B, Alvar J. Post-kala-azar dermal leishmaniasis in the Indian subcontinent: A threat to the South-East Asia Region Kala-azar Elimination Programme. Vol. 11, PLoS Neglected Tropical Diseases. Public Library of Science; 2017.

Author contributions

ESN, OJB and GFM conceptualised the study. ESN performed the analyses, wrote the manuscript and produced the remote code repository. OJB and GFM provided supervision, advised on the study concept and provided feedback on the manuscript. TL provided feedback on the methodology and presentation of results. JB, PD and KB were responsible for collection, cleaning and maintenance of the data. CB and SS supervised the data collection. NR and NG provide direction on VL research objectives on behalf of the National Vector Borne Disease

Control Programme. MMC supervised as PI of the SPEAK India consortium. All co-authors provided feedback on the manuscript and approved the final submitted version.

Acknowledgments

The authors offer their profound gratitude to the field teams at CARE India who work tirelessly to collect and organise the case details and spatial data without which this work would not have been possible, and the National Centre for Vector Borne Diseases Control for their permission to use it.

Funding

This work was funded by the Bill and Melinda Gates Foundation (ESN, MCC: OPP1183986, GFM: OPP1184344). OJB was funded by a UK Medical Research Council Career Development Award (MR/V031112/1). TCDL was supported by the NIHR Applied Research Collaboration East Midlands (ARC EM). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. CB, JB, KP, AD and SS were funded by the Bill and Melinda Gates Foundation (OPP1196454).

Declaration of interests

The authors declare that they have no competing interests.

Ethical statement

Ethical approval was obtained from the London School of Hygiene and Tropical Medicine ethics committee for this study (ref:26841) and from the National Vector Borne Disease Control Programme in India (NVBDCP) for the work of the broader SPEAK India research consortium. The ethics committee of the All India Institute of Medical Sciences-Patna approved the ACD effectiveness evaluation protocol for analysis of data from the Kala-Azar Management Information System (KAMIS); no new data were collected under the research protocol.

Figures

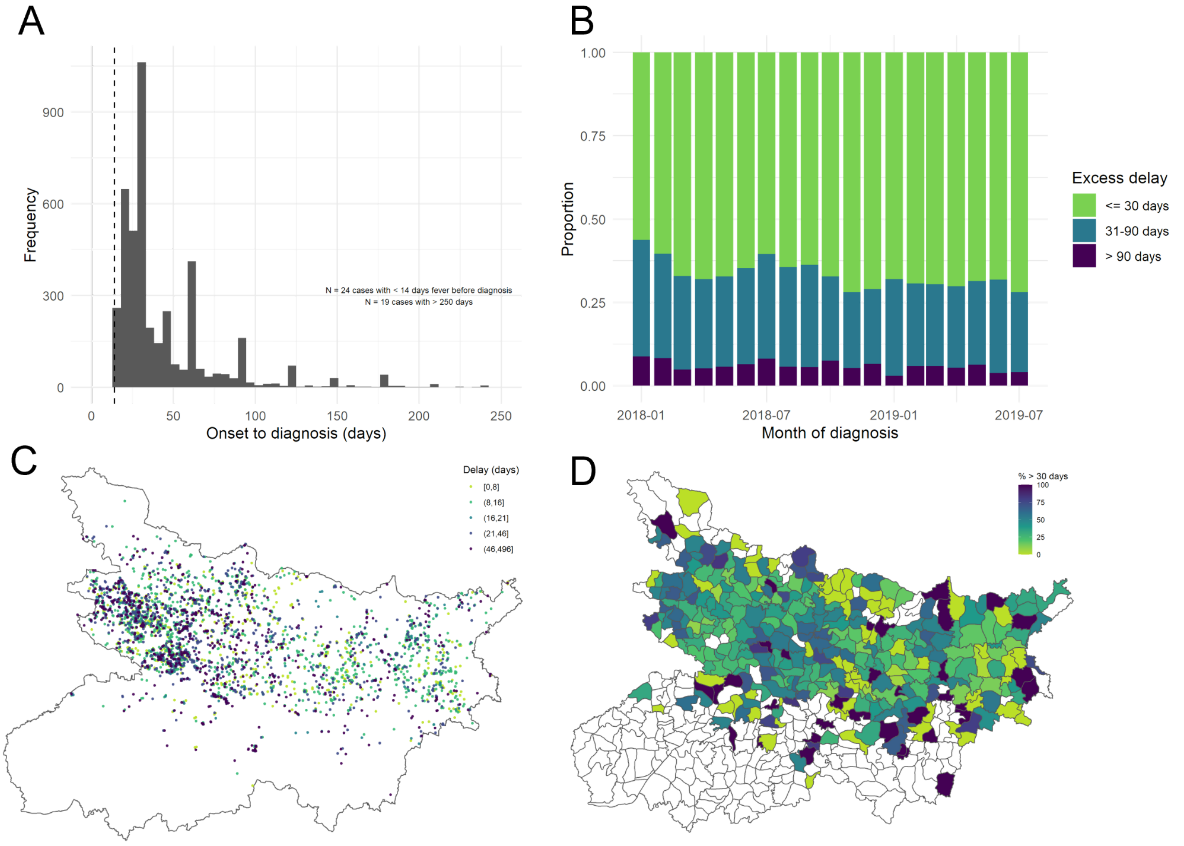


Figure 1:* (A)** *Distribution of reported days from onset of fever to diagnosis for all initially included cases. The dashed line marks the 14 day criteria for diagnosis. Note the visible heaping in reported duration, indicating a preference for 30 day intervals. Panel* ***(B) *illustrates how the proportion of cases experiencing excessive delays varies by month of diagnosis. Panels* ***(C)*** *and* ***(D)*** *illustrate the geographic distribution of reported diagnosis delays, according to GPS location of resident village (after exclusion of < 14 day durations) and by resident block.*

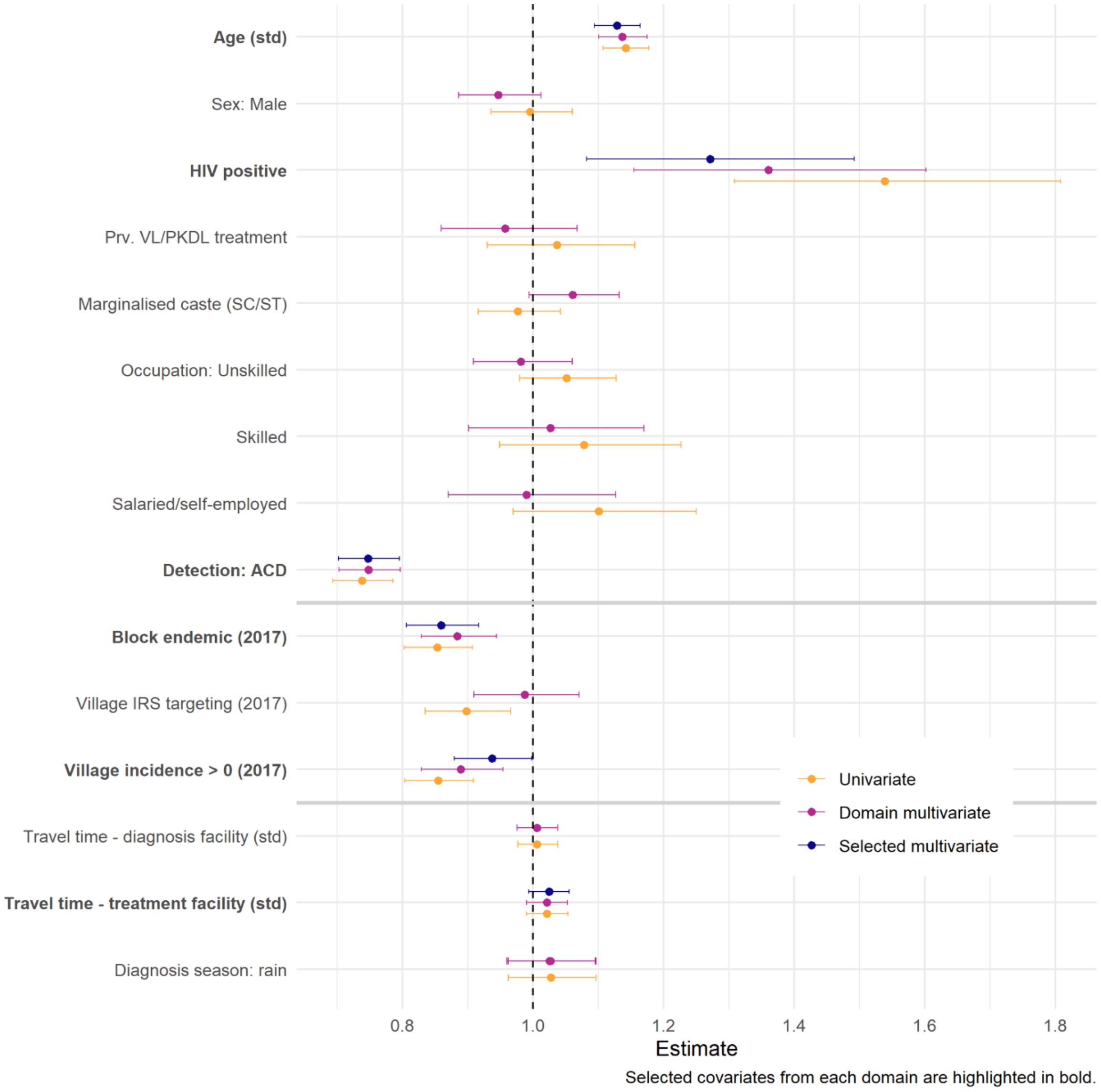


Figure 2: *Coefficient estimates (with 95% CrI) obtained from non-spatial model fits: univariate, multivariate within each covariate domain and multivariate with selected covariates from all domains. Selected covariates are also highlighted in bold on the y-axis. Note that domain models were fit to include* either *travel time to diagnosis or to treatment facility - but* not both *- due to collinearity in these covariates (closest diagnosis facility may also be closest treatment facility), therefore the domain coefficient for diagnosis season is estimated twice.*

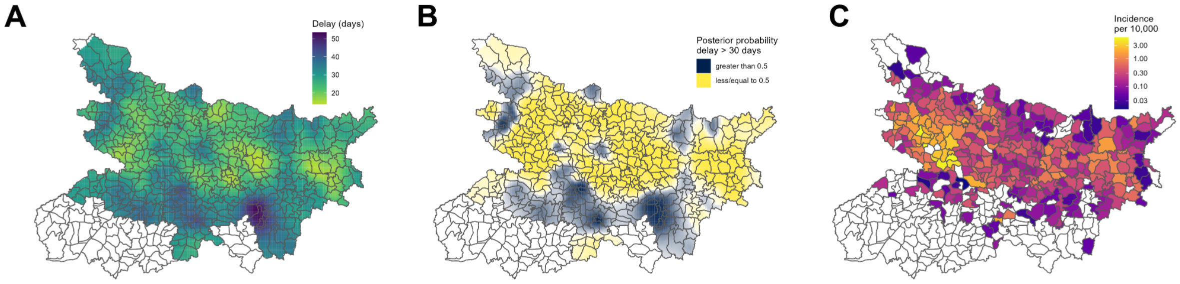


Figure 3:* *(A) *Model-estimated spatial variation in delay,assuming that cases are comparable on all factors except location.* ***(B)*** *Probability of these predicted delays exceeding 30 days, categorised to highlight where probability is greater than (yellow) or less than (black) 0.5. The opacity of colour reflects distance of the estimate from 0.5 i.e. the strength of the classification.* ***(C)*** *Observed total block-level incidence per 10,000. Note: Estimates are not mapped for districts within which no cases were observed during the period of the study.*

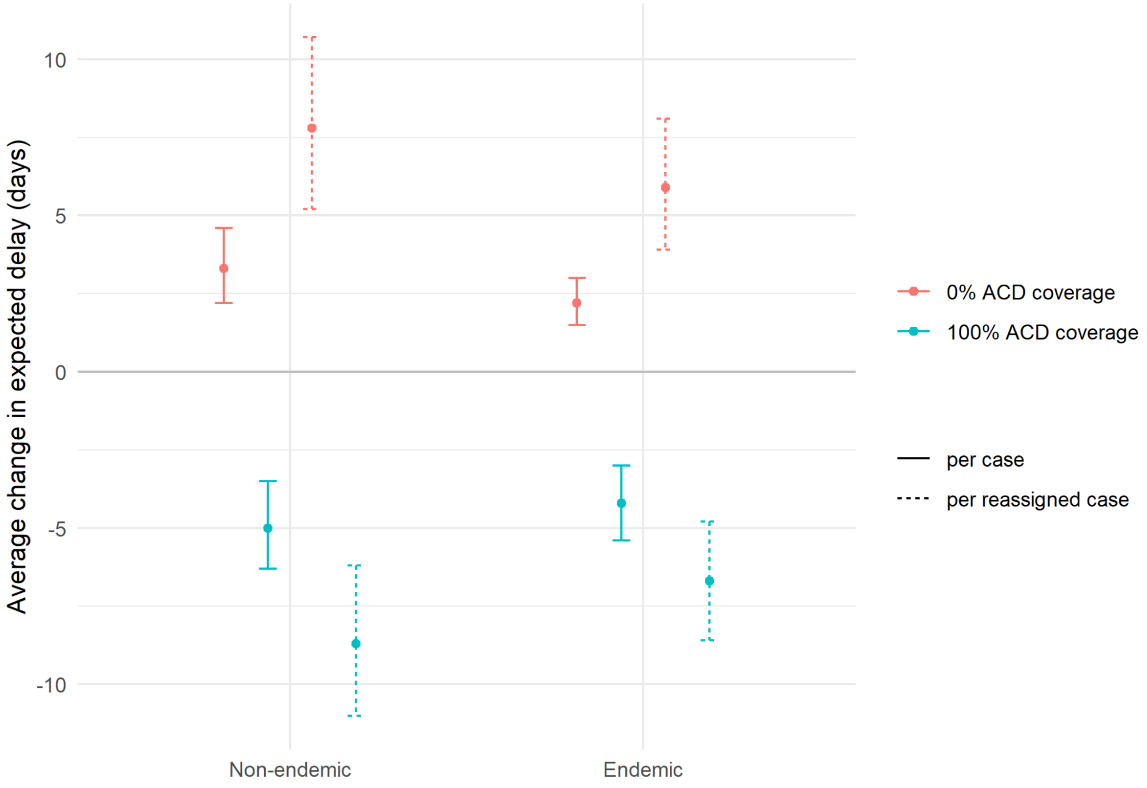


Figure 4: *Change in expected diagnosis delay under different ACD coverage scenarios, stratified by recent block endemicity. Baseline is taken as the expected delay under the actual coverage observed in this population. Estimates are shown as average days per case in total and average days per case for which detection route was reassigned under the scenario (i.e. those originally ACD in the 0% scenario, and those originally PCD in the 100% scenario). Point estimates are medians and intervals are 98% credible intervals over 10,000 posterior samples.*

Tables

***Table 1:** Descriptive summary of characteristics of 4,270 VL patients included in the analysis.*

Variable		**N**	**Delay, mean (SD)**	**Delay > 30 days, N (%)**
Sex	Female	1829	30 (34.4)	623 (34)
	Male	2441	31.9 (41.2)	814 (33)
Age (years)	< 13 years	1154	25.9 (30.4)	327 (28)
	13-25 years	1056	27 (32)	303 (29)
	26-42 years	1031	35.4 (43.1)	392 (38)
	> 42 years	1029	36.6 (45.8)	415 (40)
Scheduled caste or tribe	No	2778	32.1 (40.7)	958 (34)
	Yes	1492	29.2 (33.9)	479 (32)
Occupation	Unemployed	2506	29.9 (35.8)	818 (33)
	Unskilled	1213	32 (41.1)	421 (35)
	Skilled	272	33.5 (37.4)	99 (36)
	Self-employed/salaried	279	34.7 (48.8)	99 (35)
HIV status	Negative	4112	30.1 (36.1)	1356 (33)
	Positive	158	55.1 (73.7)	81 (51)
Previous VL/PKDL treatment	No	3896	30.8 (37.3)	1308 (34)
	Yes	374	34.2 (48.7)	129 (34)
Detection route	Passive (self-report)	2557	35.2 (41.6)	1004 (39)
	Active	1713	24.8 (32.3)	433 (25)
Block endemic in 2017	No	2332	34.4 (43.8)	847 (36)
	Yes	1938	27 (30.4)	590 (30)
Village IRS targeted in 2017	No	993	33.8 (42.1)	359 (36)
	Yes	3277	30.2 (37.3)	1078 (33)
Village incidence > 0 in 2017	No	1929	34.5 (42.9)	726 (38)
	Yes	2341	28.2 (34.2)	711 (30)
Travel time to nearest diagnosis facility	< 15 minutes	2662	31.1 (37.3)	910 (34)
	15-30 minutes	1299	31.2 (41.6)	421 (32)
	> 30 minutes	309	30.7 (34)	106 (34)

Travel time to nearest treatment facility	< 15 minutes	1619	29.6 (34.9)	545 (34)
	15-30 minutes	1812	32.2 (41.4)	620 (34)
	> 30 minutes	839	31.5 (38.4)	272 (32)

Table 2: *Model comparison on within-sample and out-of-sample fit. The minimum of each metric is shaded in grey. The difference in WAIC (*ΔWAIC*) between each model value and the minimum is presented as opposed to the absolute value.*

Model		**Within-sample**	**Out-of-sample (random CV)**		**Out-of-sample (spatial CV)**	
		ΔWAIC	**Brier score**	**Log score**	**Brier score**	**Log score**
A (Baseline)	Non-spatial, no covariates	8.8	0.2102	3.9780	0.3544	4.1384
B	Non-spatial, all covariates	4.4	0.2181	3.9361	0.3090	4.0596
C	Spatial, no covariates	2.9	0.1943	3.9333	0.3223	4.1559
D (Final)	Spatial, all covariates	0	0.2065	3.9051	0.2370	3.9058

Chapter 6

The local burden of disease during the first wave of the COVID-19 epidemic in England: Estimation using different data sources from changing surveillance practices

The emergence of COVID-19 in England presents a contrasting example in which *multiple* data sources could be exploited in order to identify and address surveillance biases. Each source offered a different perspective of the common underlying process, and biological and mechanistic links between them meant that information could be drawn across all of them, balancing the strengths and limitations of each. This paper describes an analysis which combines data on reported test positives, COVID-19-related deaths and a nation-wide infection survey to better understand the progression of the epidemic across local authorities of England, during the early months when testing was limited to certain settings. This sheds light on the impact of those initial testing constraints on our observation of the scale of the epidemic, and how this impact potentially varied across different parts of the country.

This paper was submitted to BMC Public Health in November 2021 and published in April 2022.

Please note that the original publication contained an error in the acknowledgment of three primary authors, which has since been corrected.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1600363	**Title**	Ms
First Name(s)	Emily Sara		
Surname/Family Name	Nightingale		
Thesis Title	Spatio-Temporal Patterns and Surveillance of Infectious Disease During Emergence and Elimination		
Primary Supervisor	Prof Graham Medley		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	BMC Public Health		
When was the work published?	April 2020		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**No**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I initially conceptualised the analysis with input from Graham Medley and he and Oliver Brady provided supervision. I conducted the analysis, created visualisations, wrote the manuscript and published the code (https://github.com/esnightingale/covid_deaths_spatial). Sam Abbott, Timothy Russell and Rachel Lowe provided methodological input and reviewed the manuscript. The CMMID Covid-19 working group supported data access, data cleaning and review of the final manuscript.

SECTION E

Student Signature	Emily Nightingale
Date	12 December 2022

Supervisor Signature	Graham Medley
Date	12 December 2022

BMC Public Health

RESEARCH

Open Access

The local burden of disease during the first wave of the COVID-19 epidemic in England: estimation using different data sources from changing surveillance practices

Emily S. Nightingale[1,2*], Sam Abbott[2,3], Timothy W. Russell[2,3] and CMMID Covid-19 Working Group

Abstract

Background: The COVID-19 epidemic has differentially impacted communities across England, with regional variation in rates of confirmed cases, hospitalisations and deaths. Measurement of this burden changed substantially over the first months, as surveillance was expanded to accommodate the escalating epidemic. Laboratory confirmation was initially restricted to clinical need ("pillar 1") before expanding to community-wide symptomatics ("pillar 2"). This study aimed to ascertain whether inconsistent measurement of case data resulting from varying testing coverage could be reconciled by drawing inference from COVID-19-related deaths.

Methods: We fit a Bayesian spatio-temporal model to weekly COVID-19-related deaths per local authority (LTLA) throughout the first wave (1 January 2020–30 June 2020), adjusting for the local epidemic timing and the age, deprivation and ethnic composition of its population. We combined predictions from this model with case data under community-wide, symptomatic testing and infection prevalence estimates from the ONS infection survey, to infer the likely trajectory of infections implied by the deaths in each LTLA.

Results: A model including temporally- and spatially-correlated random effects was found to best accommodate the observed variation in COVID-19-related deaths, after accounting for local population characteristics. Predicted case counts under community-wide symptomatic testing suggest a total of 275,000–420,000 cases over the first wave - a median of over 100,000 additional to the total confirmed in practice under varying testing coverage. This translates to a peak incidence of around 200,000 total infections per week across England. The extent to which estimated total infections are reflected in confirmed case counts was found to vary substantially across LTLAs, ranging from 7% in Leicester to 96% in Gloucester with a median of 23%.

Conclusions: Limitations in testing capacity biased the observed trajectory of COVID-19 infections throughout the first wave. Basing inference on COVID-19-related mortality and higher-coverage testing later in the time period, we could explore the extent of this bias more explicitly. Evidence points towards substantial under-representation of initial growth and peak magnitude of infections nationally, to which different parts of the country contribute unequally.

Introduction

The COVID-19 epidemic has impacted communities heterogeneously across England since evidence first emerged of local transmission in March 2020 [1]. Spatio-temporal patterns in transmission - driven, for example, by connectivity between regions, timing of

*Correspondence: emily.nightingale@lshtm.ac.uk
[1] Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, UK
Full list of author information is available at the end of the article

initial exposure and impact of control measures - offer valuable insight into the development of the epidemic. However, such patterns are difficult to observe and interpret from raw reported case data alone, due to uneven vulnerabilities in the local population and changes in surveillance policy over time [2].

In particular, laboratory confirmation of cases was initially restricted to urgent clinical need of patients and healthcare staff ("pillar 1") before being expanded to encompass all symptomatic cases in the wider community ("pillar 2") from 18 May 2020 [3]. These data therefore reflect different subsets of total infections at different points of the epidemic. Deaths - in particular the broad class of COVID-19-*related* deaths, including both test-confirmed and clinically suspected cases (where the disease is considered to be the primary cause of death or a contributing factor) - can be considered more consistently recorded over time. We seek to exploit the biological link between the two sources of data to obtain a clearer picture of the burden of COVID-19 during the first wave, and to quantify the extent of under-ascertainment - by which we mean the gap between reported, confirmed cases and total infections - during scale up of testing.

Observed variation in the rate of COVID-19-related deaths can be considered a result of two spatially varying components: variation in incidence of infection and variation in fatality risk among those infected. Several individual-level characteristics have been highlighted as risk factors for COVID-19 case fatality - including age, deprivation and belonging to certain ethnic groups - all of which are themselves geographically clustered (Fig. S1). The influence of this on a population level is evident in local summaries of mortality rates in England and Wales [4], and will lead to substantial variation in the number of infections which give rise to observed deaths among the populations of different local areas. These factors should therefore be taken into account in order to understand how the relative number of deaths to infections varies over space and time. Changes in surveillance affect the probability of an infection being reported as a case and are therefore also important to account for when observing changes in the relative number of deaths to cases over time.

Previous studies have demonstrated several different methods for estimating the number of cases from reported deaths. Jombart et al. [5] offered an early attempt to infer symptomatic cases from the occurrence of a single death, concluding that there would have been in the region of several hundreds of cases by the time the first death was recorded. Russell et al. [6] proposed an approach based on published estimates of baseline case-fatality rates to estimate the proportion of unreported cases over time directly, at national and regional levels for a range of early-affected countries. Nicholson et al. [7] further discuss the impact of ascertainment bias in the UK's surveillance systems and present an approach to quantify it through a joint analysis of targeted symptomatic and randomised testing data.

There have also been a number of studies exploring the spatial dynamics of COVID-19, within various country settings. Castro et al. [8] considered the timing of deaths and cases to understand the detected and undetected movement of the epidemic across Brazil. Cuadros et al. [9] explored differences in temporal trends in incidence rates between rural and urban counties in the US, but did not consider the proximity of counties in space. Amdaoud et al. [10] evaluated spatial autocorrelation statistics to analyse the early spread of COVID-19 across Western Europe, and explored how death rates related to demographic characteristics and measures of wealth, health care and social trust.

Other work exploring variation in mortality between local geographies of the UK has not accounted for the lack of independence between the units of interest, implicitly assuming that geographical regions can be considered independent after adjusting for a set of population covariates [11]. However, small and frequently zero counts in death data at a local level can limit precision of estimates when analysed independently. Sartorius et al. [12] do explicitly account for this dependence, adding a data-driven spatial structure in the form of correlated random effects within a mechanistic SEIR model, but fit to pillar one case counts only, assuming these represent a fitted proportion of total infections constrained between 5 and 40%, as informed by two systematic reviews of the asymptomatic proportion. This does not account for the proportion of individuals who are symptomatic but do not obtain a confirmed diagnosis.

This analysis aims to extend the concept of inferring infections and cases from deaths down to a local level while accounting for varying population characteristics, timing of first exposure and other unexplained sources of spatial correlation. With this approach we pursue a clearer understanding of the relative burden of disease across the country and how each locality contributes to the national picture.

Materials and methods

Data sources

Anonymised line lists of reported COVID-19-related deaths between 1 January and 30 June 2020 were provided by Public Health England (PHE). COVID-19-related deaths were considered to include those with COVID-19 recorded as an underlying cause, or where COVID-19 was mentioned as a contributing factor but not specified as cause of death. These two categories included a total of 52,560 reported deaths in England which occurred between 5 January and 30 June 2020, of which 39,332 had laboratory-confirmed infections and 13,228 non-confirmed but suspected. Counts were then aggregated by lower-tier local authority (LTLA), week of death (counted from Wednesday 1 January 2020) and 10-year age group. Aggregation by week was chosen in order to avoid excessive zero or low counts and potential day-of-week reporting effects, and to obtain a smoother representation of the epidemic curve. Records which did not have a LTLA provided ($n = 74$) were excluded. Local authority shapefiles and single-age population estimates were obtained from the Office for National Statistics (ONS) [13, 14] and matched to the aggregated death data. For descriptive purposes, the distributions of rates of deaths and confirmed cases across LTLAs are summarised by median and inter-quartile range (IQR).

LTLAs can be classified into one of four geographical categories: London borough (10.3% of total LTLAs), metropolitan district (11.5%), non-metropolitan district (60.3%) and unitary authority (17.9%). The former two categories capture the major urban areas of the country (including Birmingham, Liverpool, Manchester, Sheffield, Leeds and Newcastle) with high connectivity both nationally and internationally, while the latter capture predominantly rural areas and smaller towns or cities.

PCR-confirmed cases (i.e. COVID-19 infections identified through both pillar 1 and pillar 2 surveillance) were obtained from the same source and aggregated to the same spatial and temporal resolution. Finally, estimates of infection prevalence in England were obtained from the ONS COVID-19 infection survey pilot (15) which was initiated in May 2020. These are presented as an estimated percentage (plus 95% confidence interval based on the survey sample size) of the population who would test positive via PCR for COVID-19 during rolling fortnightly intervals.

Case definitions

For the remainder of the paper, infections confirmed with a positive PCR test and recorded in official case data *prior* to the expansion of symptomatic community testing on 18 May 2020 will be referred to as *pre-P2 cases,* and infections confirmed *following* expansion of testing will be referred to as *post-P2 cases.* It is noted that, due to piloting of pillar 2 testing among high-risk groups, a proportion of pre-P2 cases will have been detected via the pillar 2 route. We conservatively define the surveillance policy change from the point at which pillar 2 was fully available to all symptomatic individuals - assuming that case data from this point most accurately reflect the increased coverage of the expanded system - and define the terminology according to this distinction. We will also introduce the concept of *predicted-P1+P2 cases*, meaning the predicted infections which would have been PCR-confirmed in the hypothetical scenario in which symptomatic community testing had been in place since the beginning of the epidemic (January 2020). These predicted-P1+P2 cases form a subset of total symptomatic cases, conditional on the additional criteria that the case must be both symptomatic and seek and obtain a confirmatory positive test result. Lateral flow devices were not introduced for asymptomatic testing until later in the year [3] and therefore are not considered here. All references to deaths imply *COVID-19-related deaths*, i.e., those where either PCR-confirmed or clinically diagnosed COVID-19 infection is recorded on the death certificate.

Model structure

Bayesian mixed effects models for deaths per week and per LTLA were fitted using integrated nested Laplace approximation (INLA), implemented via the R package *R-INLA* [15, 16].

To facilitate comparison in observed deaths across local authorities with different population age distributions, age-adjusted expected deaths, E, were calculated for each LTLA to serve as an offset in models. Expected counts were based on age-specific weekly mortality rates averaged over the observed time period and over the country as a whole. See Additional file 1 for details. Weekly reported deaths in LTLA i were assumed to follow a negative binomial (NB) distribution, with log link function, offset by $\log(E_i)$.

In addition to age, two population level characteristics were considered as risk factors for case fatality: level of deprivation and distribution of ethnicity. The Index of Multiple Deprivation (IMD) score is defined as a relative measure of deprivation between Lower Super Output Areas (LSOAs) and incorporates a range of social, economic and health factors [17]. LSOAs are defined such that each belongs to a unique LTLA, therefore IMD scores could be aggregated to the median across all LSOAs in each LTLA and categorised by quintiles. To account for the heterogeneous distribution of ethnicity across the country, the percentage of minority ethnic groups in each LTLA population (relative to white as the national majority) was calculated according to estimates from the most recent (2011) census of England and Wales (specifically table DC2101EW “Ethnic group by sex by age”, all persons and all age categories) [18]. The number of residents self-identifying as non-white was aggregated from a five-category classification (White, Mixed/multiple ethnic, Asian/Asian British, Black/African/Caribbean/

Black British, and Other) and calculated as a proportion of the total LTLA population.

The temporal dependence in the data was modelled using a combination of random effects with random walk (RW) correlation structures [19]. A second-order random walk (RW2) on the number of weeks since the first observed death (the "epidemic" week) was intended to capture the shifted epidemic curve in each LA. Additionally, a first-order random walk (RW1) on calendar week was included to capture any overall deviations from these epidemic trends (potentially as a result of policy and behavioural change). As such, the number of deaths in any one LTLA during 1 week are a priori assumed to be correlated with the number of deaths across the prior 2 weeks. Models in which the second-order RW on epidemic week was fitted separately within each of the four geography categories were also considered.

Models without any specified spatial structure were compared to those with independent and identically-distributed (IID) random effects per LTLA, and with a combination of IID and structured, conditional autoregressive effects (as described by Besag, York and Mollié [20], hereafter referred to as BYM), parameterised with a mixing parameter between the two [21]. The latter allowed assessment of the contribution of local spatial correlation to the fit of the model, relative to purely random (IID) variation.

Six models were fitted and compared:

(A) *Baseline* Observed deaths ~ log(E) (offset) + Overall epidemic trend (RW2) + calendar week trend (RW1) + covariates (IMD, % minority); no spatial structure.
(B) A + geography-dependent epidemic trends
(C) A + IID spatial structure.
(D) B + IID spatial structure.
(E) A + BYM spatial structure.
(F) B + BYM spatial structure.

The distributions of structured random effects (spatial and temporal) were fit with penalised complexity priors on the precision and BYM mixing parameters [22], and fixed effects fit with weakly-informative gaussian priors centred at zero. A more detailed specification of all models can be found in the Additional file 1.

All analyses were performed in R version 3.6.3 (2020-02-29). All code used to run these analyses have been made available at https://doi.org/10.5281/zenodo.5763664.

Model comparison

Models were compared using the Widely Applicable Information Criterion (WAIC) [23] and log score [24]. Pearson residuals between fitted values and observed were averaged per LTLA and mapped as a visualisation of the spatial structure unexplained by each model. Posterior samples ($n = 1000$) were drawn to explore the uncertainty in predictions and aggregated over LTLAs to give total trajectories over time.

Comparison to post-P2 cases

It was assumed that post-P2 cases (swabbed from 18 May 2020 onwards) were reflective of the higher coverage surveillance and less obscured by capacity constraints. A fixed lag of 1 week between date of swabbing and date of death was applied to infer *predicted-P1 + P2* cases from modelled deaths in the primary analysis, while a sensitivity analysis was conducted assuming two- and three-week lags. This choice was informed by the swab-death delay distribution observed in this dataset (median 6 days, IQR 8 days), while also considering an external report from the COVID Clinical Information Network (CO-CIN) [25] which suggested an overall longer and more varied distribution (median 13 days, IQR 14 days) between onset of symptoms and death. The possibility was considered that the lag between testing and death may have been shorter early in the epidemic, with cases predominantly being tested in a hospital setting when symptoms were already severe. However, the available data on swabbing and death dates did not suggest a difference between pre- and post-P2 cases (median 6 days pre-P2 and 7 days post-P2, with equal quartiles of 3–11 days), and therefore one fixed lag was assumed for the entire period. It was assumed that variation over this period of time in the ratio of post-P2 cases to deaths would be predominantly a result of varying completeness of observation of cases, rather than of a difference in underlying case-fatality risk.

The approach taken to infer predicted-P1 + P2 cases from reported deaths consisted of three steps. First, smoothed trajectories of deaths per week and per LTLA, corrected for spatial heterogeneity in case-fatality risk factors, were obtained from the fitted model (1000 posterior samples predicted at averaged covariate values with non-age-stratified population offset). An LTLA-level ratio of cases per covid-related death (post-P2 case per death ratio, CPDR) was then estimated for every week beyond 18 May 2020 and for each posterior sample, lagging the modelled death counts by 1 week (two and three weeks for the sensitivity analysis) and comparing to post-P2 cases. CPDRs were summarised over all post-P2 weeks to obtain a median and IQR for each LTLA, which were then used to scale up the posterior samples over the whole time period. This yielded an estimate of the magnitude of cases giving rise to those deaths, which would have been detected under expanded surveillance. The distributions across posterior samples are summarised into 1, 25, 75 and 99%

quantiles - yielding 50 and 98% Credible Intervals (CrI) - for presentation.

Inferring infection and rate of detection

The previous steps yield local trajectories of COVID-19 cases which would have been detected through combined hospital and community-based symptomatic testing, had such capacity been available throughout the first epidemic wave. However, post-P2 cases detected under expanded surveillance remain a subset of the total number of infections, which also include those that are asymptomatic or otherwise undetected. The ONS COVID-19 infection survey pilot [26] suggested that per fortnight between 27 April and 24 May 2020 around 0.25% of the population of England would have tested positive for COVID-19, with this percentage steadily decreasing to 0.03% by the beginning of July. To investigate the gap between total infection incidence and detected cases, these data were combined with post-P2 case counts over the same period to infer a rate of detection under expanded surveillance (see Additional file 1). This rate of detection was then applied to the entire trajectories of predicted-P+P2 cases to estimate the number of infections represented by those detected cases. Observed pre- and post-P2 counts could then be compared to these estimated infections to infer the percentage of infections detected over time and within each LTLA.

Results

A summary of the observed incidence of covid-related deaths and pre—/post-P2 confirmed cases is shown in Fig. 1. Over time, the early exposure of London is clear in both deaths and confirmed cases, with the two epidemic curves following a similar shape and peaking prior to the other geographies. Outside of London, confirmed

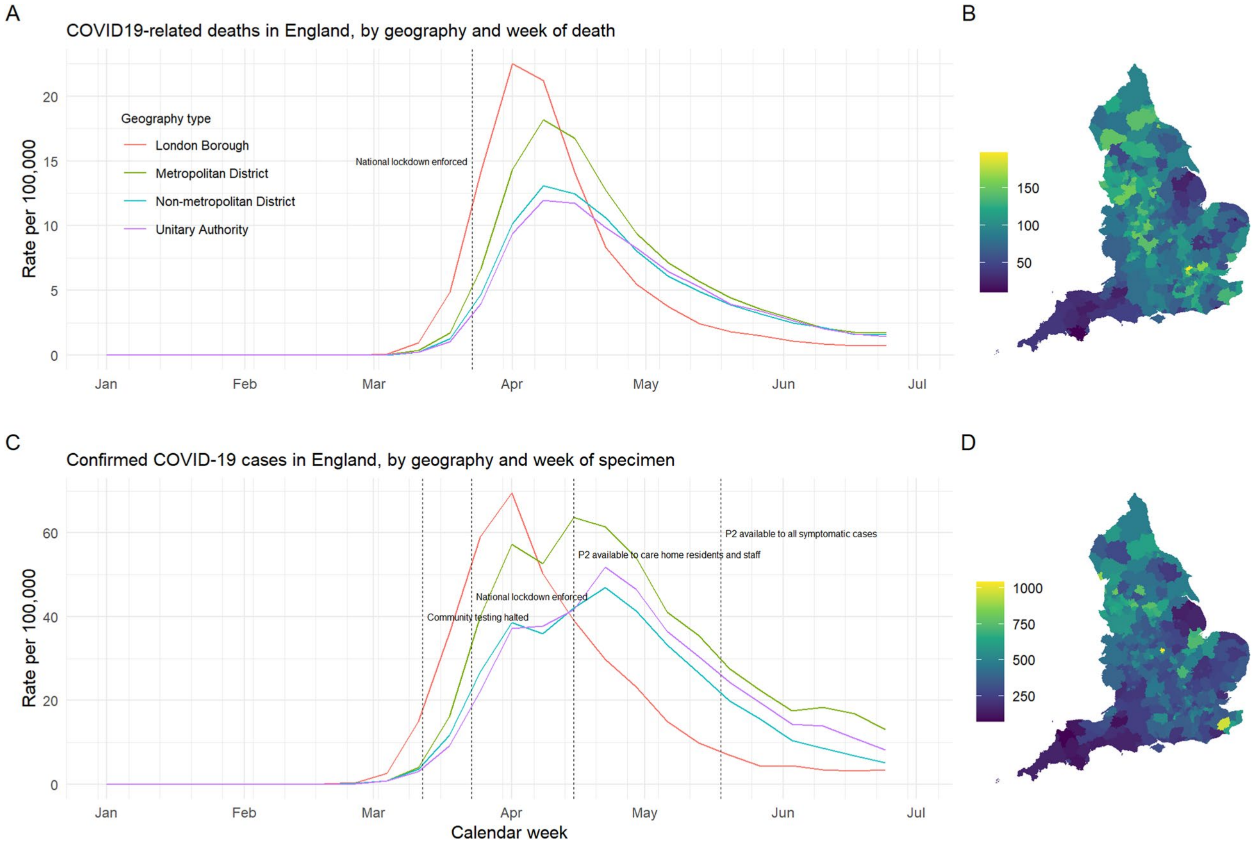


Fig. 1 Rates of COVID-19-related deaths and confirmed cases in England, by geography and week of death, and by lower-tier local authority (LTLA). **(A, B):** Weekly rates per 100,000 population of COVID-19-related deaths and confirmed cases, respectively, by geography type. Trajectories of reported deaths follow a smooth epidemic curve while the peak in case counts appears to be truncated across geographies outside of the early-affected London region, potentially as a result of national lockdown measures but also of testing constraints. Dashed vertical lines mark dates of significant policy changes with respect to confirmatory testing of suspect cases. **(C, D):** The same data instead presented as total rates per 100,000 per LTLA, across the entire first wave (1 January 2020 to 30 June 2020). Time periods are set according to the date of specimen and date of death, respectively

case counts appear to be truncated between the end of March and the end of April, approximately coinciding with the implementation of the national lockdown on 23 March 2020.

Overall, COVID-19-related mortality rates ranged from 10 per 100,000 in South Hampshire to 196 per 100,000 in Hertsmere (median [IQR]: 90.6 [71.4, 112.1]). Cumulative incidence of confirmed cases was more varied between LTLAs, ranging from 71 per 100,000 in Torridge in North Devon, to 1040 per 100,000 in the East Midlands city of Leicester (median [IQR]: 379.2 [298.3, 491.5]). Supplementary Fig. S1 illustrates the substantial variation in the population characteristics assumed to contribute to case-fatality risk across the country.

Model selection

By comparison of information criteria (WAIC) and cross-validated log score, it is clear that adjustment for epidemic timing and the specified fatality risk covariates (model A) were insufficient alone to explain the spatial distribution of deaths across England. Out of the six candidate models, the BYM spatial model and temporal trends specific to the geography of the LTLA was selected as offering the lowest WAIC and best cross-validated fit (model F) (Table 1). Models with unstructured, IID random effects per LTLA performed comparably to the BYM model and the overall magnitude of error appeared to be reduced, but spatial structure in the residuals was still evident (Supplementary Fig. S2).

Final model

The final model suggested strong associations between weekly rates of COVID-19-related deaths in a LTLA, quintiles of deprivation score and proportion of minority ethnicities in the population (RR = 1.27 with 95% CrI [1.10–1.47] between the 1st and 4th quintiles of IMD; RR = 1.01 [1.006–1.015] per percentage increase in minority ethnic population), after adjusting for the size and age distribution of the local population (Supplementary Table S2). Despite a clear monotonic trend through the first four quintiles of deprivation score, the difference between the 1st and 5th (most deprived) quintiles dropped slightly and was estimated with a wider CrI (RR = 1.21 [0.97, 1.49]), perhaps due to the smaller number of LTLAs which fall into this category. Differences in the shape of the epidemic between each geography type were best captured by four separately fitted trends as opposed to one overall trend, and residual heterogeneity between LTLAs (i.e., not captured by covariates) was explained by a combination of spatial correlation and random noise.

Posterior samples drawn from the selected model illustrated a close fit to the epidemic trajectories overall and within each specific geography (Fig. 2). Fits for a random sample of individual LTLAs are illustrated in Supplementary Fig. S3.

The fitted posterior for the BYM mixing parameter, Φ, implies that at least 86% (posterior mean 95, 95% CrI: [86–99%]) of the residual spatial variation (accounting for the specified covariates and temporal trends) could be explained by correlation between neighbouring LTLAs as opposed to random noise. This suggests that there is correlation in observed mortality in neighbouring areas which is not explained by similarities in the size, age distribution, ethnic composition or deprivation level of their populations. A decomposition of the fitted spatial random effects for each LTLA is illustrated in Supplementary Fig. S4.

Comparison to post-P2 cases

Prior to the expansion of pillar 2 surveillance, the median CPDR per LTLA was 4.1 confirmed cases per covid-related death (IQR [3.4,5.0]). From 18 May 2020 onwards, this increased to a median of 5.2 with more variation between LTLAs (IQR [3.3,8.6]). Further detail of the spatial heterogeneity in CPDR across the country is illustrated in Supplementary Fig. S5.

Figure 3 illustrates the trajectories of predicted-P1+P2 cases inferred from the model-predicted deaths per LTLA, aggregated overall and by geography. Although the more comprehensive surveillance was assumed to be in place by mid-May, the trajectories of inferred and actual cases appear similar from the end of April to early May,

Table 1 Overall model comparison by WAIC and log score

Model	WAIC	Log score	Diff WAIC	Diff log score
B + BYM spatial	**24,750**	**2.601**	–	–
B + IID spatial	24,801	2.606	51	0.005
A + BYM spatial	25,602	2.690	851	0.089
A + IID spatial	25,665	2.697	914	0.096
Geog-specific temporal (**B**)	26,344	2.768	1593	0.167
Temporal only (**A**)	26,865	2.823	2115	0.222

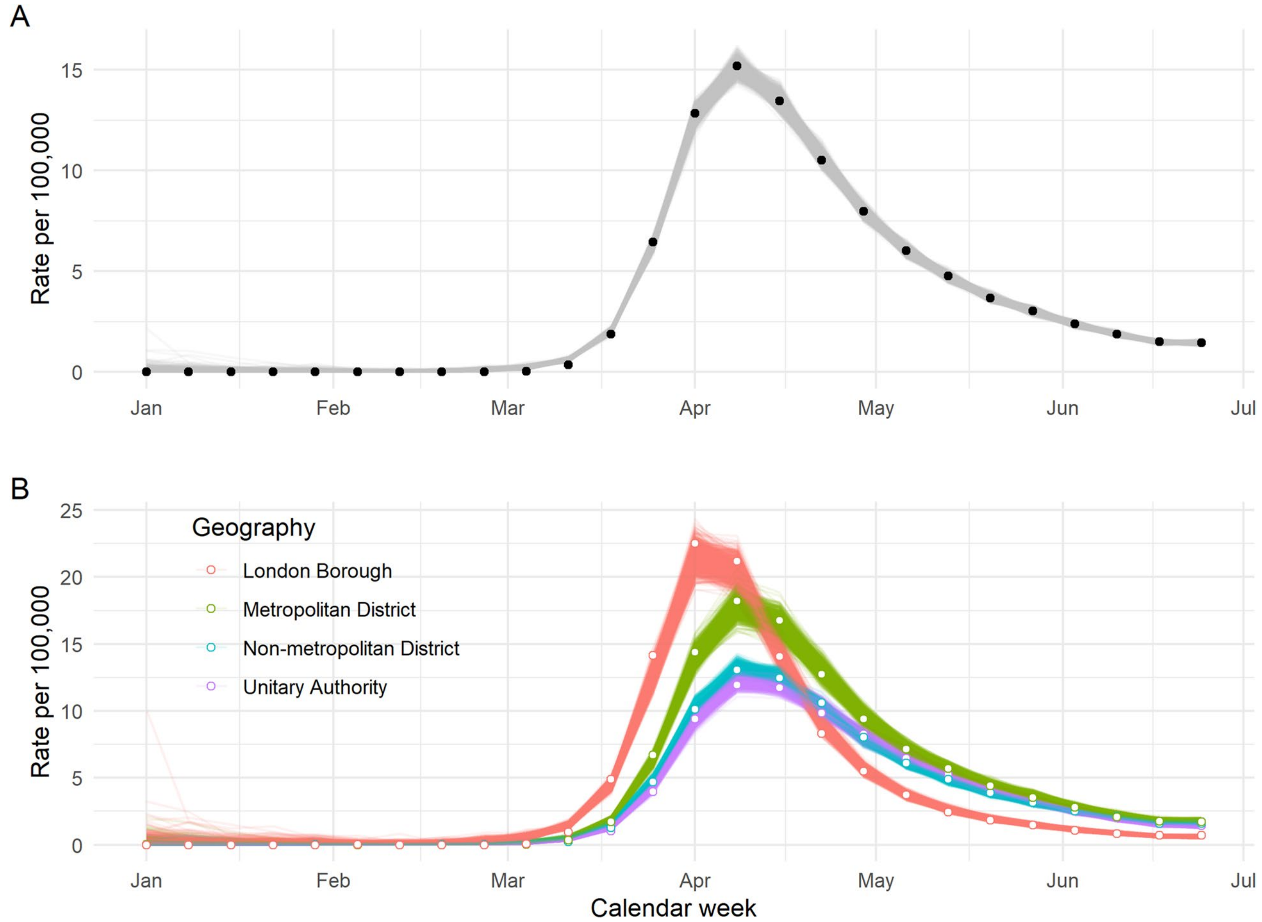


Fig. 2 Final model fit (1000 posterior samples) over time, as a national total and by geography type. The final model describes observed weekly COVID-19-related deaths per LTLA in terms of the size, age, ethnicity and deprivation level of the population, temporal trend and spatial correlation between neighbouring LTLAs. Observed rates of covid-related death per 100,000 population are shown in black (A) and white (B). Each grey/coloured line represents one sampled trajectory from the fitted model, and variation between these reflects uncertainty in the fit

when testing was accessible to care home residents and staff, over 65s and key workers [19]. Overall, the reconstructed epidemic curve of predicted-P1+P2 cases yields a median of over 100,000 additional cases - an increase of 45% - over the course of the first wave (Table 2).

The four geography types contribute unevenly to this difference. In London, the reconstructed counts suggest a relative under-representation of the peak incidence in observed confirmed cases, which narrows relatively rapidly from April onwards as numbers decline and testing capacity increases. The implied under-ascertainment in London is of a much smaller magnitude than the other three geographies; in particular for metropolitan districts and unitary authorities, results suggest that, at the height of the epidemic, confirmed cases potentially constituted less than half of the symptomatic cases which would have been detected under the expanded system. For the predominantly rural non-metropolitan districts the difference at the peak is less substantial, though still greater than that of London. From late-April, total confirmed case incidence across these LTLAs actually exceeds the reconstructed counts, by a small margin which diminishes towards the beginning of the summer (see Fig. 3A).

Assuming a longer two-week lag between testing and death yields a much larger difference of 86%, and for 3 weeks this increases further to 135%. A comparison of reconstructed national totals based on three different lags is included in Supplementary Table S3 and illustrated in Fig. 4. It is clearly shown that assuming a longer lag between case confirmation and death yields a higher and earlier peak in the reconstructed trajectory of cases.

Estimation of total infections

Figure 5 illustrates the estimated national incidence of infection according to the ONS infection survey pilot,

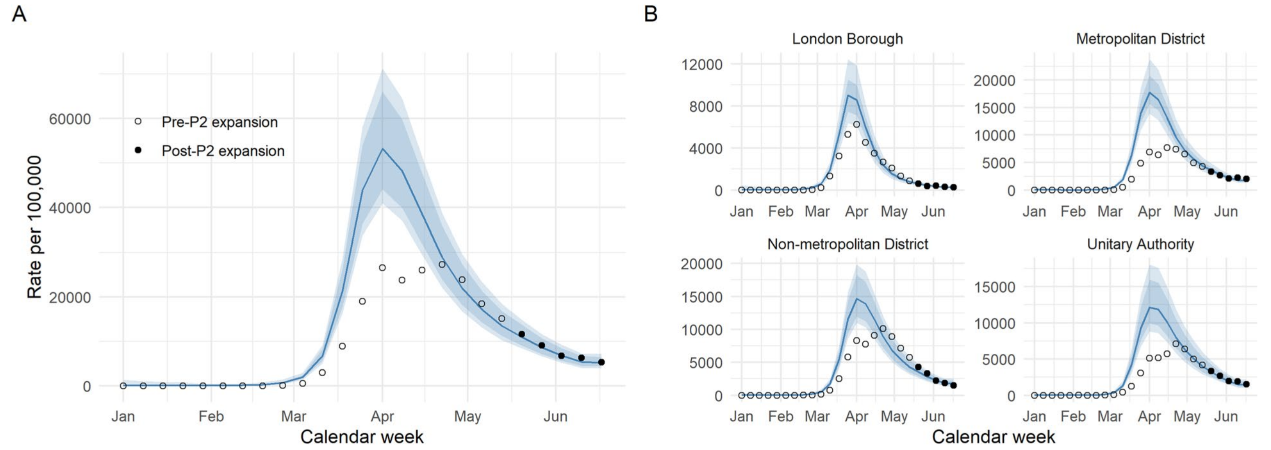


Fig. 3 Predicted-P1+P2 cases, according to lagged and scaled-up predictions from the selected model for COVID-19-related deaths, in total (A and aggregated by geography type (B). 50–98% credible intervals are shown by the blue shaded areas. Observed totals of confirmed cases per week are indicated by black points - unfilled prior to P2-expansion and filled post-P2 expansion. Predicted-P1+P2 cases suggest the potential shape and magnitude of the first wave peak if community symptomatic testing (pillar 2) - in addition to hospital-based testing (pillar 1) - had been available from the beginning of the epidemic

Table 2 Summary of observed and predicted-P1+P2 case counts over the first wave, nationally and by geography

	Observed, test- confirmed cases (*up to week starting 2020-06-17*)	**Predicted (*median [98% CrI]*)**	**Percentage difference**
England total	231,817	335,083 [275,482 - 418,847]	44.5 [18.8–80.7]
London Borough	33,399	43,664 [35,881 - 51,337]	30.7 [7.4–53.7]
Metropolitan District	64,007	109,717 [95,734 - 129,216]	71.4 [49.6–101.9]
Non-metropolitan District	79,441	97,786 [78,764 - 122,095]	23.1 [−0.9–53.7]
Unitary Authority	54,970	83,723 [67,673 - 110,968]	52.3 [23.1–101.9]

alongside total observed and reconstructed cases across the country. Comparison of observed cases from weeks starting 18 May to 15 June 2020 with these estimated total infections suggested an overall rate of detection of 25% (95% CI propagated from infection prevalence estimates: 13–58%). The total wave of infections over the entire period implied by this rate of detection is indicated by the grey curve.

This yields a cumulative total of 1.3 million infections (98% CrI 1.04 to 1.74 million) throughout the first wave, of which the observed confirmed cases (n =231,817) constitute 17.5% (98% CrI 13.3 to 22.3%).

Within each LTLA, cumulative incidence of confirmed cases constituted a median of 23% of estimate total infections (Fig. 6A). The highest rates of detection were found in Gloucester and Teignbridge in the south-west, both with estimates of over 96% (98% CrIs [87, 110%] and [81, 121%], respectively), while less than 7% detection was estimated in Leicester, Tunbridge Wells and Bradford (98% CrIs [3, 11%], [6, 15%] and [4, 10%], respectively). See Supplementary Fig. S6 for predicted trajectories in these LTLAs. Figure 6B presents the estimated detection rate for each LTLA compared to the total observed incidence, grouped by region. In most regions, greater observed incidence coincides with poorer detection of total infections. However, in London and the North, the trend leans more into the opposite direction. Supplementary Table S4 reports cumulative estimates of total infections nationally, by geography type and by region. By week, the level of under-ascertainment decreases in magnitude from February to April and settles between 25 and 30% from late-April to June (Fig. 6C).

The final predicted-P1+P2 and total infections for the entire time series in each LTLA are included in Additional file 2.

Discussion

This analysis has demonstrated that it is possible to generate plausible case burden estimates from COVID-19-related death data and, in doing so, investigate the impact of changes in surveillance practices over the first months of the epidemic.

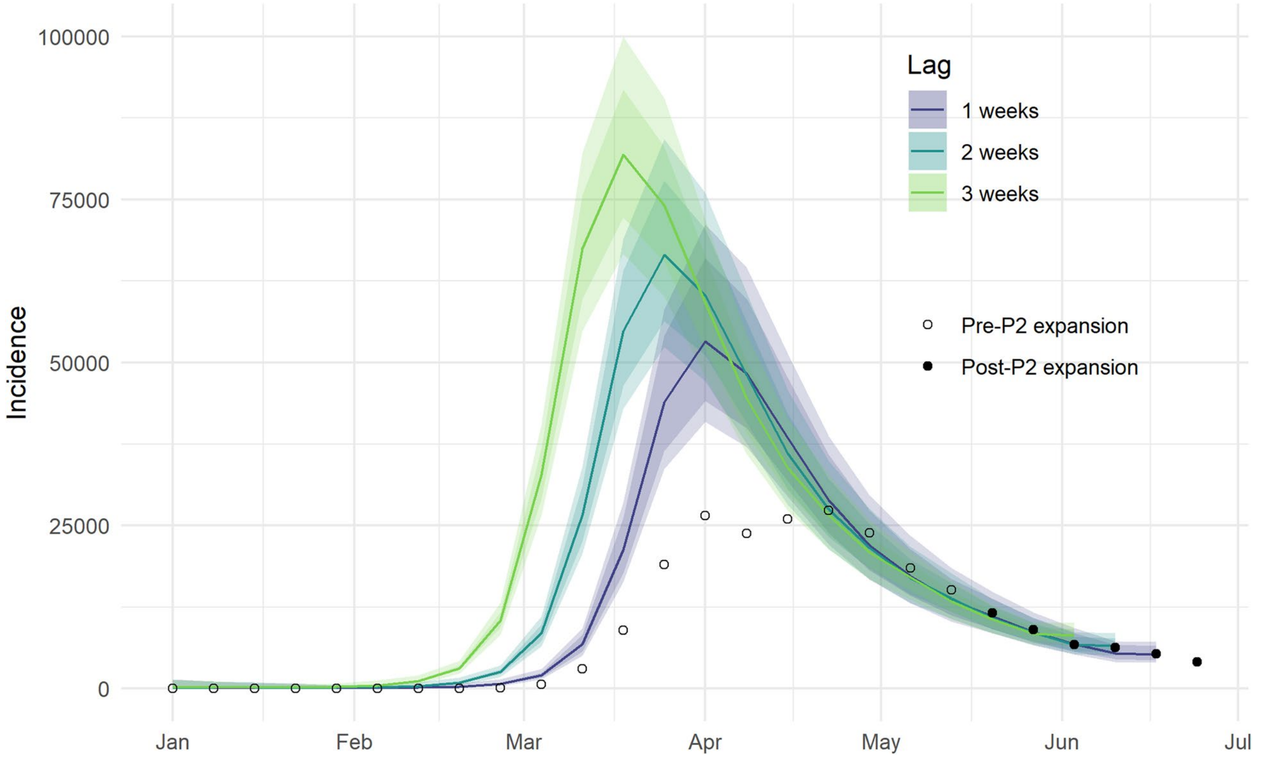


Fig. 4 Comparison of predicted-P1+P2 cases assuming one-, two- and three-week lags between date of swabbing and date of death. Shaded intervals represent 50–98% credible intervals

The model development process highlights a clear spatial structure to the incidence of covid-related deaths at a sub-national level, which is not explained by variation in the timing of initial exposure (epidemic week), or well-documented risk factors of COVID-19 death (age, deprivation and ethnic distribution in the local population). Similarities in risk between neighbouring LTLAs can be an important factor to consider for the design of local mitigation strategies, particularly in response to the detection of new variants.

Assuming that the epidemic curve of deaths represents a specific subset of total symptomatic - i.e., detectable under Pillar 1 and 2 testing strategies - infections, this analysis suggests that over 100,000 additional cases may have been counted across the country in the absence of the initial constraints on testing capacity. The uncertainty around this estimate is relatively broad (98% CrI [44,000 - 250,000]), predominantly as a result of uncertainty in the case-per-death ratio used to translate between the two measures. The increased heterogeneity between LTLAs in estimated CPDR following pillar 2 expansion may to some extent be attributed to much lower counts of both cases and deaths as the epidemic waned, and the occurrence of local outbreaks. Overall, we estimate around four post-P2 confirmed cases per covid-related death across all LTLAs, or equivalently a rate of 0.25 deaths per case. This is higher than estimates of the case-fatality rate i.e. the rate of deaths among confirmed cases, due to our broader definition of covid-*related* deaths as opposed to deaths directly attributed to COVID-19 *among* confirmed cases.

Cases ascertained, even under the expanded system, remain a subset of total infections. The case estimates obtained here were therefore combined with estimates of infection incidence from the ONS's pilot survey in order to explore the rate of detection over time and between LTLAs. This investigation suggested that, following the roll-out of symptomatic community testing, around a quarter of infections in England were being detected - a value consistent with estimates obtained by Colman et al. [27] for the period of June to November 2020 - compared to only around 10% during the first months of the pandemic.

The extent of this under-ascertainment was found to vary not only over time alongside the expansion of testing capacity, but also between LTLAs. Comparing the final model fit to the observed deaths suggested relatively little deviation of each LTLA from the fitted geography-specific trends, yet the reconstructed infections differ from observed test positives with much

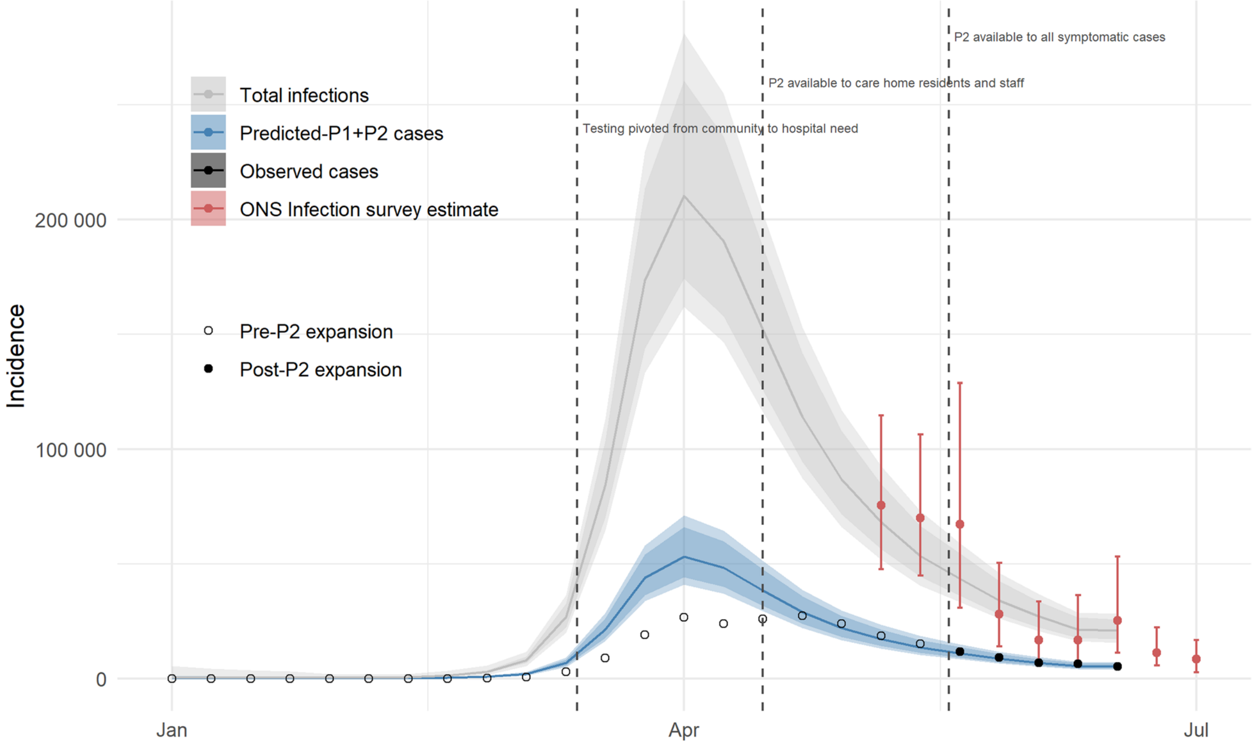


Fig. 5 Estimated weekly incidence of infections in England (grey), inferred from predicted-P1+P2 cases (blue) and an assumed detection rate of 25% under expanded surveillance. Rate of detection is estimated by comparison of incidence estimates from the ONS infection survey (shown in red) and observed case counts (shown in black) between weeks starting 18 May to 15 June 2020. This rate is then applied to predicted-P1+P2 cases to obtain the estimated trajectory of total infections

more variation between LTLAs. Greater observed incidence rates appeared to coincide with poorer detection of total infections - perhaps reflecting the impact of reaching testing capacity - yet this trend was found to be inconsistent between regions. This demonstrates the sporadic nature of case observation over time and space and highlights certain areas of the country in which surveillance was perhaps more strongly impacted by testing constraints. These differences may in part be attributed to local variation in the relationship between cases and deaths which isn't sufficiently captured by the assumed case-fatality covariates. We therefore advise that LTLA-specific estimates should be interpreted with consideration of the local context.

Early projections based on critical care admissions by Jit et al. [28] suggested an incidence of over 8000 infections per day in the UK by mid-March 2020. Assuming detection of 25% of infections under hospital and community symptomatic testing, from this study we estimate a total of around 111,000 infections during the two middle weeks of March, equating to an average of just under 8000 per day in England alone. On the other hand, via a mechanistic modelling approach, another study estimated daily total infections in the UK to have reached in the region of several hundred thousand by late-March [29]. Genomic analysis suggests that importations into the UK alone peaked mid-March with up to 1000 per day [30].

Russell et al. [6] took a data-driven approach in estimating that the peak incidence of *symptomatic* infections across the UK had occurred by mid-April 2020 with a magnitude of around 100,000 per day, and concluded that during March only 3–10% of such cases were being detected. This suggests a substantially higher peak than our estimate for England alone of just over 200,000 *total* infections per week - on average 28,000 per day - even accounting for the distribution of population between the constituent countries. Our estimates suggest that the percentage of these total infections reflected in confirmed case counts varied from around 7% at the start of March to 11% at the end, slightly higher than the detection rate Russell et al. estimated among symptomatic infections. Overall, estimates of infection incidence appear to be variable across studies, at least in part due differences in case definitions and aggregation over space and time.

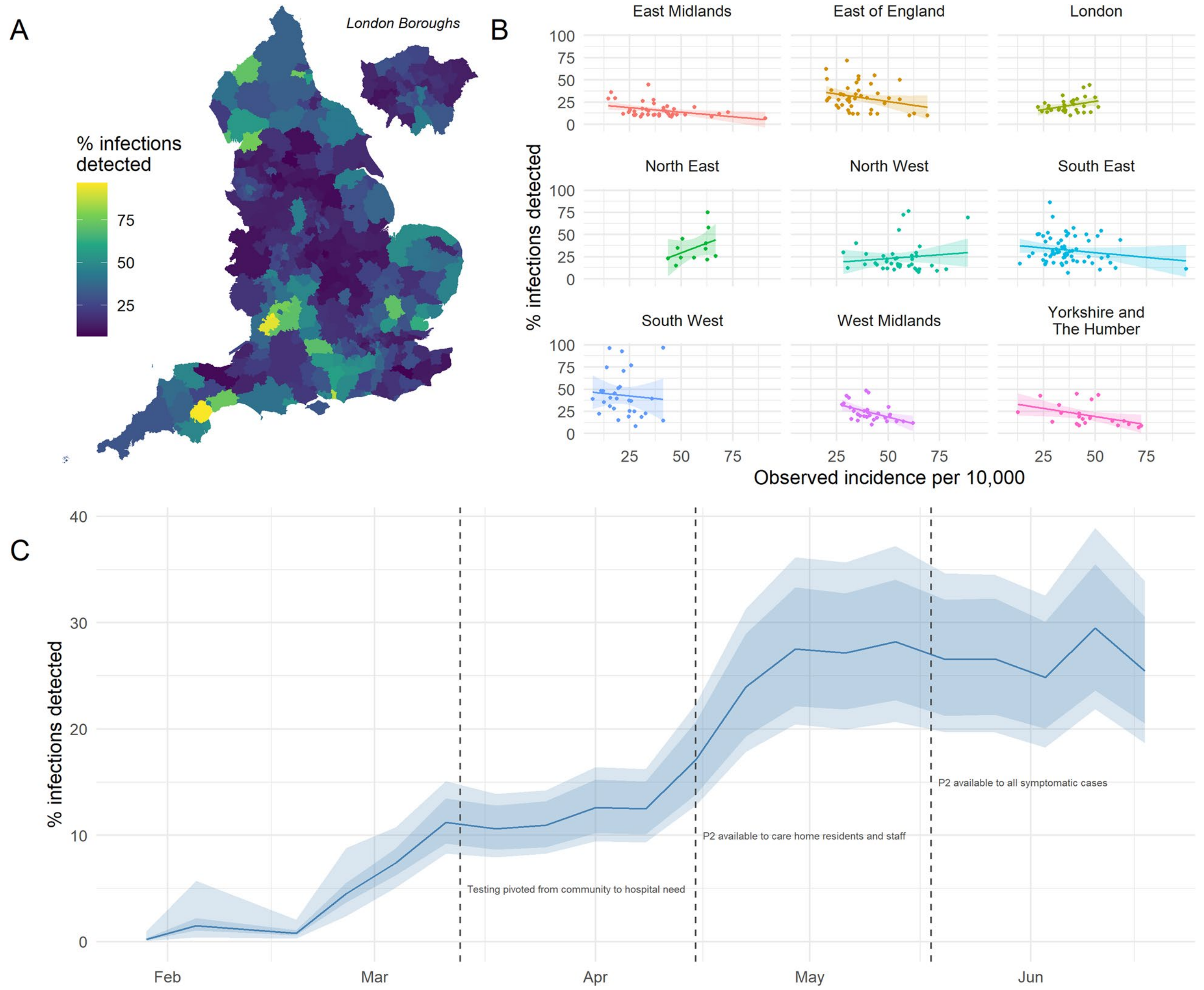


Fig. 6 Estimated percentage of total infections represented in observed case counts, per LTLA (panels **A** and **B**) and per week (panel **C**), between 2020 and 01-01 and 2020-06-17. LTLAs of Gloucester and Teignbridge stand out as having the highest percentage of detected infections, with estimates of over 96%. Panel **B** illustrates the same estimates in panel **A** but grouped by region, against the total observed incidence per 100,000. Total infections over the time period are estimated based on the predicted-P1+P2 cases and an assumed infection detection rate of 25% under expanded surveillance. All panels reflect median predictions over 1000 posterior samples, with panel **C** additionally showing 50–98% credible intervals

Limitations

The interpretation of these findings depends on several key assumptions, most importantly that variability over time in the ratio of confirmed cases to COVID-19-related deaths is predominantly the result of varying accessibility of testing. It is however plausible that fatality risk would have varied over time, potentially increasing towards the peak of the wave due to strain on hospitals forcing re-prioritisation of care or decreasing later on as treatment options improved. Also, it was assumed that variability in the delay from swabbing to death on the individual level would be diluted by aggregation, hence a fixed-value lag (with its influence explored in sensitivity analysis) would suffice. Summarising observed delays between swabbing and death within the available data gave no reason to suggest a difference between the time periods pre- and post-pillar 2 expansion, therefore the same fixed lag was assumed throughout the epidemic wave. A more exact approach, however, would have been to incorporate the full distribution of swab-death delays and redistribute the observed deaths in time according to an imputed point of detection.

Only three broad characteristics were considered as case-fatality risk factors, which essentially serve as proxy measures for complex combinations of underlying comorbidities and health indicators across the population. Dichotomising self-identified ethnicity in a population into “majority” and “minority” groups is crude, given that risk has been found to differ between ethnic groups in different ways [31]. We implicitly assume that these estimates from the national census are representative of the population. Several studies report case-fatality risk as being overall higher among biological males [32–34], yet there is also debate as to how the effect may interact with other key risk factors such as age and deprivation [35]. Here it was found that the ratio of males to females varied only marginally between LTLA populations and was uninformative for the observed mortality rate.

In individual level analysis, comorbidities such as cardiovascular disease, diabetes, and cancer were shown to have an association with mortality after adjusting for both ethnicity and deprivation level [36]. The local prevalence of such conditions is however a component in the calculation of the deprivation score used here. There are likely nuances and complex interactions between granular risk factors for mortality [37, 38] which are not yet understood in sufficient detail to be explicitly defined in such a model. The measures used here are therefore intended to capture high-level differences, and further work exploring additional covariates associated with mortality on both an individual and environmental/contextual level might improve population-level risk estimates.

Finally, this approach does not account for the dynamics of transmission in and around long-term residential care facilities during the early months of the pandemic, within which many deaths during the first wave occurred [39]. The nature of infections in these settings - with respect to mortality, testing and management - is different to that which occurred in the wider community, yet both care home and community deaths were treated equally in this analysis. For LTLAs with a particularly large care home population, the estimated case-per-death ratio may be higher than it would have been excluding these particularly vulnerable individuals. However, adjusting for the age distribution of the LTLA population in the underlying deaths model should at least in part attenuate this source of variation. This study aimed to explore broad, population-level patterns in incidence of deaths and detection of cases, whereas characterising the contribution of incidence within residential care settings would require a more fine-scaled, context-specific analysis. There was further substantial transmission within healthcare settings which we have not included separately [40].

Conclusions

Effective and efficient control of an infectious disease epidemic relies on appropriate quantification of risk at a local level from available surveillance data. However, there are many reasons for which such data may not be equally representative of disease burden across different regions and populations. In the case of the COVID-19 epidemic in England, it is known that limitations in testing capacity distorted the observed trajectory of cases during the first wave. In this analysis, by combining more consistently reported data on deaths and more representative case data from later in the epidemic, it was possible to reconstruct a plausible trajectory of symptomatic cases which could have been detected in the absence of the early testing constraints, and further to infer the total number of infections these reported cases would represent. This facilitated a comparison between the two testing policies and highlighted heterogeneity in case ascertainment across different regions of the country.

The burden of disease and impact of the response to this pandemic will be evaluated in detail for years to come. Considering how changes in surveillance policy can obscure the spread of an epidemic - using methods such as those demonstrated here - will be essential, in particular for understanding the consequences of the country’s initial level of pandemic preparedness.

Abbreviations

BYM: Besag York Mollié; CPDR: Cases Per Death Ratio; CO-CIN: COVID Clinical Information Network; CrI: Credible Interval; IID: Independent and Identically Distributed; IMD: Index of Multiple Deprivation; INLA: Integrated Nested Laplace Approximation; IQR: Inter-Quartile Range; LSOA: Lower Super Output Area; LTLA: Lower Tier Local Authority; ONS: Office for National Statistics; PCR: Polymerase chain reaction; RW: Random Walk; WAIC: Widely Applicable Information Criterion.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12889-022-13069-0.

Additional file 1. Supplementary Materials.

Additional file 2. Final predicted-P1+P2 and total infections for the entire time series in each LTLA.

Acknowledgments

The authors would like to thank Christopher Jewel and Jonathon Reed for valuable discussion during the early stages of this work.
The following authors were part of the Centre for Mathematical Modelling of Infectious Disease COVID-19 working group: Eleanor M. Rees, Rosalind M. Eggo, Matthew Quaife, Fiona Yueqian Sun, Carl A. B. Pearson, Kiesha Prem, James D. Munday, Sophie R. Meakin, Graham Medley, Kevin van Zandvoort, W. John Edmunds, Alicia Rosello, Sebastian Funk, Kathleen O’Reilly, Billy J. Quilty, Simon R. Procter, Amy Gimma, Adam J. Kucharski, Arminder K. Deol, Jon C. Emery, Nikos I. Bosse, Hamish P. Gibbs, David Simons, Stéphane Hué, Christopher I. Jarvis, Petra Klepac, Yang Liu, Anna M. Foss, Charlie Diamond, C. Julian Villabona-Arenas, Akira Endo, Rein M. G. J. Houben, Stefan Flasche, Timothy W. Russell, Samuel Clifford, Gwenan M. Knight, Joel Hellewell, Nicholas

G. Davies, Katherine E. Atkins, Damien C. Tully, Megan Auzenbergs, Mark Jit & Rachel Lowe.
The following funding sources are acknowledged as providing funding for the CMMID Covid-19 working group:
This research was partly funded by the Bill & Melinda Gates Foundation (OPP1183986: ESN). This research was partly funded by the Global Challenges Research Fund (GCRF) project 'RECAP' managed through RCUK and ESRC (ES/P010873/1: TJ). This research was partly funded by the National Institute for Health Research (NIHR) using UK aid from the UK Government to support global health research. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care (Health Protection Research Unit for Modelling Methodology HPRU-2012-10096: TJ). UK MRC (LID DTP MR/N013638/1: QJL; MC_PC_19065: TJ). Authors of this research receive funding from UK Public Health Rapid Support Team funded by the United Kingdom Department of Health and Social Care (TJ). Wellcome Trust (210758/Z/18/Z: SA). The following funding sources are acknowledged as providing funding for the working group authors. Alan Turing Institute (AE). BBSRC LIDP (BB/M009513/1: DS). This research was partly funded by the Bill & Melinda Gates Foundation (INV-003174: KP, MJ, YL; NTD Modelling Consortium OPP1184344: CABP, GM; OPP1180644: SRP; OPP1191821: KO'R, MA). DFID/Wellcome Trust (Epidemic Preparedness Coronavirus research programme 221303/Z/20/Z: CABP, KvZ). Elrha R2HC/UK DFID/Wellcome Trust/This research was partly funded by the National Institute for Health Research (NIHR) using UK aid from the UK Government to support global health research. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care (KvZ). ERC Starting Grant (#757688: CJVA, KEA; #757699: JCE, MQ, RMGJH). This project has received funding from the European Union's Horizon 2020 research and innovation programme—project EpiPose (101003688: KP, MJ, PK, WJE, YL). This research was partly funded by the Global Challenges Research Fund (GCRF) project 'RECAP' managed through RCUK and ESRC (ES/P010873/1: AG, CIJ). HDR UK (MR/S003975/1: RME). Nakajima Foundation (AE). NIHR (16/137/109: BJQ, CD, FYS, MJ, YL; Health Protection Research Unit for Modelling Methodology HPRU-2012-10096: NGD; PR-OD-1017-20002: AR). Royal Society (Dorothy Hodgkin Fellowship: RL; RP\EA\180004: PK). UK DHSC/UK Aid/NIHR (ITCRZ 03010: HPG). UK MRC (LID DTP MR/N013638/1: EMR; MC_PC 19065: RME; MR/P014658/1: GMK). Wellcome Trust (206250/Z/17/Z: AJK, TWR; 208812/Z/17/Z: SC, SFlasche; 210758/Z/18/Z: JDM, JH, NIB, SFunk, SRM). No funding (AKD, AMF, DCT, SH).

Authors' contributions

ESN, GFM and OJB conceptualised the analysis and GFM and OJB supervised the work. ESN performed the analyses and prepared all Figs. SA, TWR and RL gave feedback on methodology and presentation of results. ESN wrote the manuscript. SA, RL, GFM and OJB reviewed and edited the manuscript. ESN produced the remote code repository. All co-authors provided feedback on the analyses and the manuscript, and agreed with the final submitted version. Authors who were part of the Centre for Mathematical Modelling of Infectious Disease COVID-19 working group each contributed in processing, cleaning and interpretation of data, interpreted findings, contributed to the manuscript, and approved the work for publication.

Funding

ESN and GFM are funded by the Bill and Melinda Gates Foundation (ESN: OPP1183986, GFM: OPP1184344). OJB and SA are funded by the Wellcome Trust (OJB: 206471/Z/17/Z, SA: 210758/Z/18/Z, TWR: 206250/Z/17/Z). RL is funded by the Royal Society (Dorothy Hodgkin Fellowship).

Availability of data and materials

The datasets supporting the conclusions of this article are available at https://doi.org/10.5281/zenodo.5763664.

Declarations

Ethics approval and consent to participate

Approval for the use of anonymised line list data was granted by Public Health England and the Department for Health and Social Care. Consent of individuals was not required as no patient identifiable information was used.

All methods were performed in accordance with the relevant guidelines and regulations regarding analysis of human data, including SAMPL guidelines for reporting of statistical analyses.

Consent for publication

Consent to publish was not required as this manuscript contains no individually identifiable details or images.

Competing interests

The author(s) declare(s) that they have no competing interests.

Author details

[1]Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, UK. [2]Centre for Mathematical Modelling of Infectious Disease (CMMID), London School of Hygiene & Tropical Medicine, London, UK. [3]Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [4]Centre on Climate Change and Planetary Health, London School of Hygiene & Tropical Medicine, London, UK. [5]Barcelona Supercomputing Centre (BSC), Barcelona, Spain. [6]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.

Received: 25 November 2021 Accepted: 14 March 2022
Published online: 11 April 2022

References

1. Holden B, Quinney A, Padfield S, Morton W, Coles S, Manley P, et al. COVID-19: public health management of the first two confirmed cases identified in the UK. Epidemiol Infect. 2020;148 Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7484301/.
2. Sherratt K, Abbott S, Meakin SR, Hellewell J, Munday JD, Bosse N, et al. Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of COVID-19 in England. Philos Trans R Soc B Biol Sci. 2021;376(1829):20200283.
3. Coronavirus (COVID-19): scaling up testing programmes [Internet]. GOV.UK. Available from: https://www.gov.uk/government/publications/coronavirus-covid-19-scaling-up-testing-programmes. Accessed 14 May 2021.
4. Deaths involving COVID-19 by local area and socioeconomic deprivation - Office for National Statistics [Internet]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19bylocalareasanddeprivation/deathsoccurringbetween1marchand17april. Accessed 17 Mar 2021.
5. Jombart T, van Zandvoort K, Russell TW, Jarvis CI, Gimma A, Abbott S, et al. Inferring the number of COVID-19 cases from recently reported deaths. Wellcome Open Res. 2020;5:78.
6. Russell TW, Golding N, Hellewell J, Abbott S, Wright L, Pearson CAB, et al. Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. BMC Med. 2020;18(1):332.
7. Nicholson G, Lehmann B, Padellini T, Pouwels KB, Jersakova R, Lomax J, et al. Improving local prevalence estimates of SARS-CoV-2 infections using a causal debiasing framework. Nat Microbiol. 2022;7(1):97–107.
8. Castro MC, Kim S, Barberia L, Ribeiro AF, Gurzenda S, Ribeiro KB, et al. Spatiotemporal pattern of COVID-19 spread in Brazil. Science. 2021 Apr 14;372(6544):821–6.
9. Cuadros DF, Branscum AJ, Mukandavire Z, Miller FD, MacKinnon N. Dynamics of the COVID-19 epidemic in urban and rural areas in the United States. Ann Epidemiol. 2021;59:16–20.
10. Amdaoud M, Arcuri G, Levratto N. Are regions equal in adversity? A spatial analysis of spread and dynamics of COVID-19 in Europe. Eur J Health Econ. 2021;22:1–14.
11. Verhagen MD, Brazel DM, Dowd JB, Kashnitsky I, Mills MC. Forecasting spatial, socioeconomic and demographic variation in COVID-19 health care demand in England and Wales. BMC Med. 2020;18(1):203.
12. Sartorius B, Lawson AB, Pullan RL. Modelling and predicting the spatio-temporal spread of COVID-19, associated deaths and impact of key risk factors in England. Sci Rep. 2021;11(1):5378.
13. Local Authority Districts (April 2019) Names and Codes in the United Kingdom [Internet]. Available from: https://geoportal.statistics.gov.uk/

datasets/c3ddcd23a15c4d7985d8b36f1344b1db_0. Accessed 17 Mar 2021.

14. Population estimates for the UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics [Internet]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2019estimates. Accessed 17 Mar 2021.
15. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser B Stat Methodol. 2009;71(2):319–92.
16. Martins TG, Simpson D, Lindgren F, Rue H. Bayesian computing with INLA: New features. Comput Stat Data Anal. 2013;67:68–83.
17. English indices of deprivation 2019 [Internet]. GOV.UK. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019. Accessed 1 Feb 2021.
18. DC2101EW (Ethnic group by sex by age) - Nomis - Official Labour Market Statistics [Internet]. Available from: https://www.nomisweb.co.uk/census/2011/dc2101ew. Accessed 1 Feb 2021.
19. Codling EA, Plank MJ, Benhamou S. Random walk models in biology. J R Soc Interface. 2008;5(25):813–34.
20. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math. 1991;43(1):1–20.
21. Riebler A, Sørbye SH, Simpson D, Rue H. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Stat Methods Med Res. 2016;25(4):1145–65.
22. Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH. Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. Stat Sci. 2017;32(1):1–28.
23. Watanabe S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. J Mach Learn Res. 2010;11:3571–94.
24. Gneiting T, Aftery AER. Strictly Proper Scoring Rules , Prediction , and Estimation. J Am Stat Assoc. 2007;102:477.
25. CO-CIN: COVID-19 - Time from symptom onset until death in UK hospitalised patients, 7 October 2020 [Internet]. GOV.UK. Available from: https://www.gov.uk/government/publications/co-cin-covid-19-time-from-symptom-onset-until-death-in-uk-hospitalised-patients-7-october-2020. Accessed 20 Nov 2020.
26. Coronavirus (COVID-19) Infection Survey: England - Office for National Statistics [Internet]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/coronaviruscovid19infectionsurveydata. Accessed 14 Oct 2021.
27. Colman E, Enright J, Puspitarani G, Kao R. Estimating the proportion of SARS-CoV-2 infections reported through diagnostic testing. medRxiv. 2021. https://doi.org/10.1101/2021.02.09.21251411.
28. Jit M, Jombart T, Nightingale ES, Endo A, Abbott S, Group LC for MM of IDC-19 W, et al. Estimating number of cases and spread of coronavirus disease (COVID-19) using critical care admissions, United Kingdom, February to March 2020. Eurosurveillance. 2020;25(18):2000632.
29. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020;584(7820):257–61.
30. du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. 2021;371(6530):708–12.
31. Collaborative TO, Mathur R, Rentsch CT, Morton CE, Hulme WJ, Schultze A, et al. Ethnic differences in COVID-19 infection, hospitalisation, and mortality: an OpenSAFELY analysis of 17 million adults in England. medRxiv. 2020; https://doi.org/10.1016/S0140-6736(21)00634-6.
32. Peckham H, de Gruijter NM, Raine C, Radziszewska A, Ciurtin C, Wedderburn LR, et al. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. Nat Commun. 2020;11(1):6317.
33. Bienvenu LA, Noonan J, Wang X, Peter K. Higher mortality of COVID-19 in males: sex differences in immune response and cardiovascular comorbidities. Cardiovasc Res. 2020; https://doi.org/10.1093/cvr/cvaa284.
34. Gebhard C, Regitz-Zagrosek V, Neuhauser HK, Morgan R, Klein SL. Impact of sex and gender on COVID-19 outcomes in Europe. Biol Sex Differ. 2020;11(1):29.
35. Dehingia N, Raj A. Sex differences in COVID-19 case fatality: do we know enough? Lancet Glob Health. 2021;9(1):e14–5.
36. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584(7821):430–6.
37. Lassale C, Gaye B, Hamer M, Gale CR, Batty GD. Ethnic disparities in hospitalisation for COVID-19 in England: The role of socioeconomic factors, mental health, and inflammatory and pro-inflammatory factors in a community-based cohort study. Brain Behav Immun. 2020;88:44–9.
38. Townsend MJ, Kyle TK, Stanford FC. Outcomes of COVID-19: disparities in obesity and by ethnicity/race. Int J Obes. 2020;44(9):1807–9.
39. Gordon AL, Goodman C, Achterberg W, Barker RO, Burns E, Hanratty B, et al. Commentary: COVID in care homes—challenges and dilemmas in healthcare delivery. Age Ageing. 2020;49(5):701–5.
40. Knight G, Pham TM, Stimson J, Funk S, Jafari Y, Pople D, et al. The contribution of hospital-acquired infections to the COVID-19 epidemic in England in the first half of 2020. Res Sq. 2022. https://doi.org/10.21203/rs.3.rs-1140332/v1.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

BMC Public Health

CORRECTION Open Access

Correction: The local burden of disease during the first wave of the COVID-19 epidemic in England: estimation using different data sources from changing surveillance practices

Emily S. Nightingale[1,2*], Sam Abbott[2,3], Timothy W. Russell[2,3], CMMID Covid-19 Working Group, Rachel Lowe[2,4,5,6], Graham F. Medley[1,2] and Oliver J. Brady[2,3]

Correction to: BMC Public Health 22, 716 (2022) https://doi.org/10.1186/s12889-022-13069-0

The original publication of this article [1] contained 2 errors:

1. 3 main authors were shown as collaborators, this affected Rachel Lowe, Graham F. Medley and Oliver J. Brady
2. The full collaborator list was not available

The original article has been updated with the correct information.

Author details
[1]Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, UK. [2]Centre for Mathematical Modelling of Infectious Disease (CMMID), London School of Hygiene & Tropical Medicine, London, UK. [3]Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [4]Centre on Climate Change and Planetary Health, London School of Hygiene & Tropical Medicine, London, UK. [5]Barcelona Supercomputing Centre (BSC), Barcelona, Spain. [6]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.

Published online: 07 June 2022

Reference
1. Nightingale ES, et al. The local burden of disease during the first wave of the COVID-19 epidemic in England: estimation using different data sources from changing surveillance practices. BMC Public Health. 2022;22:716. https://doi.org/10.1186/s12889-022-13069-0.

The original article can be found online at https://doi.org/10.1186/s12889-022-13069-0.

*Correspondence: emily.nightingale@lshtm.ac.uk

[2] Centre for Mathematical Modelling of Infectious Disease (CMMID), London School of Hygiene & Tropical Medicine, London, UK
Full list of author information is available at the end of the article

Chapter 7

Discussion

7.1 Challenges of monitoring and modelling disease at emergence and elimination

In settings of emergence and elimination when observed case incidence is highly heterogeneous, it is crucial to consider how cases arise over space and how patterns observed in the past may give insight into the future. Spatial targeting of interventions is desirable in both settings, where constraints on resources often make blanket approaches infeasible. However, this targeting must inevitably be informed by what is observed through the biased lens of surveillance. Spatial variability in surveillance can have a strong influence on how disease burden is understood to be distributed and on the policy decisions made as a result.

Under-prepared systems often mean that the rapid escalation of emerging epidemics cannot be fully observed. This was clearly the case during the COVID-19 pandemic and early, important decisions were made based on an unrepresentative picture of disease burden. In elimination settings, predominantly passive surveillance systems established during a period of endemicity may be insufficient to achieve the necessary completeness of observation to verify and maintain extremely low levels of transmission. This can result in an incorrect perception of the control or even elimination of a disease from a region. Evaluating the strength of surveillance systems should therefore be a crucial point of focus in settings where any form of elimination is the goal.

7.2 Principle findings of this thesis

This thesis set out to explore and address these challenges within the contexts of two different diseases and settings. Firstly, in chapter 3, a commonly-asked programmatic question - regarding the short-term forecasting of disease for the purpose of resource targeting - is addressed for the case of visceral leishmaniasis elimination in Bihar. In this case a practically acceptable level of predictive accuracy was attained. However, it was highlighted that, despite the policy relevance of sub-district ('block') level with respect to incidence targets and implementation of interventions, short-term forecasts at this spatial scale are much less useful when cases are very sparse. Moreover, patterns of incidence were observed in some blocks which appeared to reflect the sudden detection of a cluster of cases at one point in time, rather than the natural incidence of individual cases. This drew attention to the challenges of drawing inference from routinely reported diagnoses in the presence of active case detection measures. These patterns of sporadically high incidence on a block-month scale are evidence of surveillance biases in the data, which alter how the observed spatial and temporal patterns can be interpreted in terms of true disease burden.

Chapter 4 follows on from these conclusions by using a disaggregation approach to draw inference at a finer scale that is more practical to act upon in terms of intervention, and more relevant to the actual vector-borne transmission of VL. Spatial patterns of environmental conditions suitable for the sandfly vector are not found to be informative of spatial patterns in incidence between villages. The reason behind this may be linked to low levels of incidence across the majority of the state at this stage of elimination. Spatial patterns arising from the underlying biological mechanism of vector-borne transmission may have become increasingly fragmented at very low levels of incidence, to the point where they are no longer identifiable. At this stage there may also be a greater contribution of transmission at a longer range through population movement, rather than driven by vector movement and abundance.

The inaccuracy of the disaggregated predictions may give further evidence for routine surveillance data not being representative of true incidence; spatial patterns at the village level may not be predictable based on environmental covariates and spatial autocorrelation if much of the pattern in fact arises from locally-targeted surveillance effort. On the other hand, inaccuracy may arise from the methodological approach itself as opposed to the underlying data. Previous applications of this disaggregation regression methodology have not had access to data at the disaggregated scale with which to formally validate. This analysis is therefore unique in testing how the method performs against real-world data and gives an insight into its limitations, for example in not accommodating non-linear associations with the specified covariates.

Chapter 5 pushes further into the idea of varying surveillance effort by investigating excessive delay to diagnosis experienced by reported cases, in relation to the geographic location of their village and its history of VL. After adjusting for individual level factors which could affect symptom presentation (e.g. HIV infection) and the detection of each case through passive (self-reported) or active (i.e. targeted) channels, substantial residual variation was found to remain at the village level. This appeared to be associated with previous VL burden and interventions in the village and by further, unexplained, correlation between nearby villages. Areas in which cases on average experienced longer delays were identified in the south and north-east of the state, beyond the high burden areas at the focus of the elimination programme. Even having adjusted for each individual's route of detection, there may be an indirect effect of active case detection (ACD) on the time taken for cases to report passively; the presence of these activities in reaction to recent incidence could more generally raise awareness and responsiveness among the local population.

Diagnosis delay is used here as an indirect indicator, having been previously shown to be associated with detection effort [114]. The fact that observed heterogeneity in delay cannot be fully attributed to ACD could suggest that multiple factors would also play a role in the probability of diagnosis overall, across different locations and populations. In the absence of another quantifiable measure, this provides an initial insight into how the non-uniform distribution of ACD may be biasing the observed distribution of disease across the state, and what the pattern of this bias may look like. Delays to diagnosis and treatment of infectious cases also directly impede the interruption of transmission in areas of low incidence, potentially creating the conditions to trigger an outbreak and unravel the progress of recent years. In the absence of effective prophylactic measures, ensuring consistently prompt diagnosis across the entirety of the state will be critical to avoid resurgence.

Finally, chapter 5 presents the emergence of COVID-19 as an example of how the spatial pattern of surveillance bias can be more thoroughly interrogated when multiple data sources may be drawn upon. In the case of VL, substantial changes in disease burden occur on the scale of several years, therefore questions around burden can reasonably be explored from a purely spatial perspective, within a fixed time window. The COVID-19 pandemic, on the other hand, was dynamic and rapidly evolving. It was therefore important to model the epidemic behaviour with temporal as well as spatial trends. Variation was found between England's local authorities in terms of the extent to which confirmed cases of COVID-19 were representative of total infection during the first epidemic wave. Although testing was constrained across the country during the first few months, it appeared to have a differing impact on the proportion of symptomatic cases that were detected in different parts of the country, in particular between urban and rural locations. Similarly to the case of VL in Bihar, the limitations of the surveillance system did not appear to impact observation uniformly over space, and potentially led to an inaccurate view of where the heaviest burden lay.

7.3 Policy Implications

Although a common request of modellers from policy-makers, validated short-term forecasts are rare. Since the work in chapter 3 was published, several other attempts have been made to forecast VL in different settings [115, 116, 117]. These, however, apply univariate time series methods to predict for a region as a whole without also accounting for variation over space. Hussein-Alkhateeb et al. recently reviewed early warning systems for several outbreak-prone diseases [118] and found that, although a multitude of work has been published on the topic, prediction validity was often not presented. Moreover, the practicality of integrating highly complex statistical models into routine surveillance is a key limitation that is rarely addressed. In a recent qualitative analysis of interviews with key stakeholders and policy-makers, Dial et al. [119] state this as one of the key barriers to the use of modelling outputs in policy.

This work demonstrates that short-term forecasting of VL case burden at a sub-national level is feasible and practical, and there is an ongoing dialogue with partners in-country about integration of this into programmatic use (Appendix A) via the new, online KA-MIS dashboard. This would allow programme managers and policy-makers to review both the current situation and what is expected over the next few months, and plan the distribution of resources accordingly. It would also allow policy-makers to obtain near-real time estimates of elimination target attainment if recent observed trends were to continue. The relevant spatial resolution for inference may, however, differ between an epidemiological and policy perspective. For VL, the policy-relevant partition was initially the sub-district or 'block' - the scale at which the threshold for elimination as a public health problem was set and monitored. However, from an epidemiological perspective, the smaller scale of village was more relevant since risk factors, exposures and conditions for transmission can be highly heterogeneous across these discrete communities. The ideal scale on which to answer a given epidemiological or policy question must be balanced against the practical limitations of data collection at that scale and the computational effort to analyse it. Chapter 4 suggests that prediction of fine-scale incidence from block-level data, i.e. prediction over space rather than ahead in time, is more challenging. This highlights the value of on-the-ground surveillance at a local level, and justifies the investment of time and resources in this.

In addition to case incidence, key indicators of the performance of the elimination programme (treatment delay, duration of fever, testing for HIV co-infection) are now monitored through the KA-MIS dashboard, but only to the level of district. The work in this thesis suggests that monitoring at this scale may miss important focal areas of poor performance within broadly well-performing districts. Chapter 5 in particular shows the value of analysing characteristics of individual cases on a finer spatial scale, yielding the important insight that the success of the control programme on a regional level has not been felt equitably across the population at risk.

Combined with the current targeted ACD in endemic villages, additional spot-checks within *non-endemic* villages - particularly in areas found to be prone to longer delays - would give crucial reassurance that the lack of observation does indeed reflect a lack of incidence. The definition of ”high-risk” may be extended to encompass not only risk inferred from recent incidence, but also risk due to uncertainty as a result of lack of recent surveillance activity in the area. Cases identified through such additional checks could be isolated and treated much more quickly than if they had waited to self-report, reducing the risk of reestablishing transmission in the area. It is also vital that the timing and location of ACD activities is recorded in addition to the diagnosis of cases. Such data would give an indication of the likely sensitivity of the surveillance system in different areas, from which we could better understand our uncertainty in observed case counts. This could be communicated to stakeholders alongside reported incidence rates to convey a level of confidence in the attainment of incidence targets.

Well-designed prevalence surveys can provide a representative estimate of true disease burden which may be compared against routine notifications to quantify the extent of under-detection. Yet, in the case of a disease nearing elimination, such a survey would need to be vast to capture the low levels of prevalence. For a novel emerging disease, there may be insufficient knowledge of its dynamics to design an appropriate survey with respect to scale and sampling approach. In either case, even the largest prevalence surveys are rarely powered to estimate on a sub-national level. The approach described in chapter 4 offers a possible alternative, but we find that it cannot replace information gathered directly at a local scale in this elimination setting. Chapter 6 demonstrates a retrospective approach to infer the relative extent of under-detection, which captures the impact of testing strategies on detection of *symptomatic* cases but still depends on survey data to infer the magnitude of *asymptomatic* infections.

It is clear that, despite their cost, surveys of this kind play an important role in emergence and elimination settings. In the example of COVID-19 emergence, we circumvented we the lack of sub-national infection survey estimates by making the assumption that it was only severity and case detection which varied sub-nationally, and not infection overall. This is a crude assumption but perhaps justifiable given the high transmissibility of the virus and the overall connectivity of England's population; it is highly likely that, even before the start of the data, the epidemic had been seeded and was circulating undetected in most parts of the country. In elimination settings and in particular in a rural and less connected region such as Bihar, it is plausible for broad areas to have not been exposed to infection for some time. In these cases, surveys should be carefully targeted to certain populations and locations in which the routine data is weakest and which pose the greatest threat in terms of recrudescence, adding important information and precision where it is needed most.

7.4 Limitations

These two disease settings demonstrate the challenges of disentangling the mechanisms of incidence and surveillance. Chapter 3 finds that the practical interpretation of forecasts is severely limited when no attempt is made to distinguish the two. Chapters 4 and 6 attempt to exploit the nature of the underlying biological mechanisms as a way to differentiate between variation in true incidence and variation in surveillance. Chapter 5 explores an indirect indicator of the strength of surveillance in the form of diagnosis delays, but how representative this indicator is cannot be easily verified.

The methods applied here all make the assumption of similarity or dependence between observations according to spatial proximity. Although the mechanism of transmission for both COVID-19 and VL is broadly conducive to this simple assumption, there are other factors at play which in reality result in complex connections between both geographically close and distant cases. Distant regions can be strongly inter-dependent in terms of transmission as a result of population movement, for example across travel networks between major towns or cities [120]. The dominance of distance-based spatial structures in models of vector-borne disease incidence has been raised as a limitation elsewhere [121].

The regions at risk of VL in Bihar are in particular connected from a non-geographic perspective via temporary migration for work [105, 108]. This movement likely increased as a result of the COVID-19 pandemic, at the same time that access to care was disrupted and impoverished populations left even more vulnerable [122, 123, 124, 125]. Future analyses of routine VL surveillance data will need to consider how this can be accounted for more realistically than with simple proximity-based assumptions.

The assumed strength of correlation may be weighted according to the "flow" of population between areas (for example inferred from mobile data [6]) or some composite measure of ease of travel based not only on distance but also road condition, public transport connections and cost. The latter was considered within chapter 5 as a covariate capturing accessibility of diagnosis facilities, but not to quantify the correlation of observations to each other. Locally adaptive methods can accommodate less smooth changes between neighbouring units than the standard conditional auto-regressive structure [126], and even greater flexibility can be achieved using penalised smoothing splines [127]. However, identifying data with which to inform hypothesised connections poses a challenge; Lee et al. [127], for example, define non-Euclidean 'proximity' using census data on cross-municipality commuting, but suggest that longer-range air travel data may be more appropriate.

Of course any arbitrarily complex correlation structure may in theory be specified, but adding complexity to the assumed spatial structure inevitably results in greater (potentially prohibitively so) computational cost. The efficiency of the INLA algorithm relies heavily on the sparse precision matrices implied by simple, nearest-neighbour assumptions of spatial dependence. More complex models may therefore demand a return to a more traditional sampling-based approach, for example taking advantage of the improved efficiencies offered by the Hamiltonian Monte-Carlo algorithm as implemented in Stan [128].

The question has been raised as to how much of the spatial correlation we are trying to model is attributable to common risk factors, exposures and connectivity within the population, and how much is an artifact of surveillance. For the case of VL, it was not possible to directly interrogate this question since the implementation of ACD is not recorded and does not follow a fixed structure that could be modeled. The dataset described in chapter 5 defines the detection route of each case - inferred through data on referral source and linkage to suspect case registers - therefore providing some evidence for when ACD has been implemented. These data include cases defined as detected via ACD within villages which had not had recently reported cases and hence should not in theory have been subject to ACD efforts. This could suggest that ACD is in practice implemented on a more ad-hoc basis than the official protocol states, or could be a consequence of the imperfect linkage of patients to a village. Either way, it demonstrates that the local impact of ACD is difficult to quantify from the available information.

Efforts to identify and record referral route for newly diagnosed cases (e.g. through a local informant designated to identify suspect cases or otherwise) have recently improved and relative proportions are now presented on the KA-MIS dashboard, albeit aggregated over the whole state. Further analysis of case data which distinguish actively and passively detected cases could give more insight into how ACD is deployed and how its yield may be distributed differently to passive, self-reporting across the state.

7.5 Future directions

The digitisation of routine surveillance data with the advent of KA-MIS in 2017 has created the first opportunities to draw inference of VL burden on a state-wide scale in near-real time. Definition of how the population of Bihar is divided into enumerated village communities, and the initiation of a plan to geolocate all new cases as they are diagnosed, opens the door to finer scale inference than has been possible before. These new opportunities for inference may have the potential to fill gaps in knowledge of the epidemiology of VL where mechanistic modelling approaches have in the past fallen short. At present, however, there remain limitations in the available data which prevent them being used to their best advantage. In particular, a record of when and where village-level interventions (vector control and case detection) are implemented, uniquely linked to the master list of villages used in case surveillance, could be incorporated into routine data collection. The evidence from controlled studies as to the efficacy of these targeted interventions is mixed, therefore these data could facilitate valuable investigation into their practical, real-world impact on case detection.

Avilov et al. [129] propose an approach to estimate a case detection rate for TB by considering a competition model between disease progression, death, spontaneous self-cure and detection. Their approach, however, depends upon the ability to stratify observed cases by disease stage at the point of detection, whereas routinely-collected data regarding disease progression for VL are limited. Time from symptom onset to diagnosis is in practice not considered to be reliably reported and is often missing, recorded deaths are rare and the durations between these events very uncertain. Shaweno et al. [130] extend the idea and present a geo-spatial hidden Markov model to address a similar question, exploiting inherent spatio-temporal correlation and predictor-driven mechanisms to estimate the underlying incidence process separately from the binomial detection process. However, their model is fit across only 66 small geographic units or *kebeles*. Fitting such a model across the incredibly sparse process of village-level VL may present computational challenges. These approaches demonstrate the kind of inference around the detection process that may be possible but would need to be adapted to the specific setting of VL, and potentially would depend on adjustments to the kind of data which are routinely recorded.

Surveillance bias is common in many fields from which parallels with infectious disease epidemiology may be drawn. In particular within ecology, a common goal is to estimate a spatially continuous process of occurrence or abundance from observations at different locations. This process can be subject to bias resulting from non-random chance of the target (for example, an animal) being detected from the observer's location. This could be interpreted as a similar process to that of varying disease surveillance coverage discussed within this thesis. The mechanism of the bias may be due to distance sampling, wherein the chance of detection decreases with increasing distance from the observer, or preferential sampling, where observation is more common at locations where the animal is expected to be. The reactive process of ACD, along with the finding from chapter 5 that diagnosis delays appear shorter in areas with recent incidence, could reflect a process of the latter type. In this case, observation is dependent on the underlying incidence process itself.

Distance sampling may be counteracted by incorporating into the model a function of decaying detection probability with distance from the known observer location [131]. Often, inverse-care laws apply in disease surveillance contexts - people at greater "distance" from health care (be that by physical distance or as a result of economic/social barriers) have greater difficulty obtaining appropriate care and as a result have poorer health outcomes [132]. Simplistically, this could be described as a distance sampling process in which cases in more remote locations are less likely to be detected than those with easier access to a health facility, with this probability decreasing with geographic distance. However, the analysis in chapter 5 suggested that distance with respect to travel time to a facility equipped for VL care was not strongly associated with the length of time taken to obtain a correct diagnosis. There are likely a number of complex, interacting factors at play which would not be possible to capture in a smoothly decaying function over only one dimension. The reactive triggering of ACD could possibly be modelled within a similar framework to that used with distance sampling, for example as a step-change of increased detection probability for a period of time following a new passively-detected case. The concept of assuming an increased rate following observation of an event is analogous to a "self-exciting" or "Hawkes" processes, the potential value of which has been demonstrated for epidemiological applications [133, 134, 135, 136]. These processes - initially defined to model earthquakes and their aftershocks - distinguish the observation of events over time into two contributing factors: a background intensity and a "triggering" function. This defines an intensity which is conditional on the history of the process, such that the occurrence of one event temporarily increases the intensity for a period of time before eventually decaying back to the background rate. Equivalently, a triggering function in space increases the intensity for the surrounding area, decaying with distance. Estimation of such models using the INLA method has been proposed [137].

Preferential sampling can be addressed using a joint model between the locations at which the process is observed and the measurement at that location, with the two components sharing a common spatial field [138, 139]. Not accounting for this dependence has been shown to result in biased inference [139], however this bias may be attenuated with adjustment for appropriate confounding variables that link the process to the observation locations. Investigation into whether preferential detection of VL cases is evident in historically affected areas would mean that decisions as to how targeted surveillance is used in future can be made with a better understanding of its limitations.

The types of models described here can be classified as *semi-mechanistic*. These are statistical, data-driven models which also incorporate known structures or mechanisms of the data generating process - such as the links between observations at different disease stages and the processes of detection - as in classical transmission models. Spatio-temporal statistical models can in theory be incredibly flexible, but require substantial amounts of data to fit the complex patterns that arise from infectious disease surveillance. Yet, even with the emergence of a novel disease of which there is little prior knowledge, we are never completely agnostic about the processes by which the disease presents and is recorded. In settings of emergence and elimination where the available data may be limited or biased, semi-mechanistic models can allow our broader understanding of the underlying physical and biological process to inform and constrain estimation.

It is not possible to know how the first COVID-19 wave may have played out if a more robust system had been in place to inform those early policy decisions; however, an understanding of how it fell short could contribute to improving the country's response to an emerging epidemic in the future. The elimination of VL in India, on the other hand, is at a crucial turning point where decisions are being made about the use of resources to monitor and maintain suppression of transmission going forward. Insights gained within this thesis highlight how the current system of targeted active surveillance may leave blind spots where transmission could resurge - a crucial consideration when we are no longer faced with the task of measuring burden of disease, but of validating its absence. Finally, this thesis identifies opportunities to adapt the way in which surveillance may be conducted and interpreted, so that the country can continue to move towards truly equitable elimination of the disease.

Bibliography

[1] M. Jit, T. Jombart, E. S. Nightingale, A. Endo, S. Abbott, L. C. f. M. M. o. I. D. C.-. W. Group, and W. J. Edmunds, “Estimating number of cases and spread of coronavirus disease (COVID-19) using critical care admissions, United Kingdom, February to March 2020,” *Eurosurveillance*, vol. 25, p. 2000632, May 2020. Publisher: European Centre for Disease Prevention and Control.

[2] E. S. Nightingale, O. J. Brady, C. C.-. w. Group, and L. Yakob, “The importance of saturating density dependence for population-level predictions of SARS-CoV-2 resurgence compared with density-independent or linearly density-dependent models, England, 23 March to 31 July 2020,” *Eurosurveillance*, vol. 26, p. 2001809, Dec. 2021. Publisher: European Centre for Disease Prevention and Control.

[3] E. M. Rees, E. S. Nightingale, Y. Jafari, N. R. Waterlow, S. Clifford, C. A. B. Pearson, C. W. Group, T. Jombart, S. R. Procter, and G. M. Knight, “COVID-19 length of hospital stay: a systematic review and data synthesis,” *BMC Medicine*, vol. 18, p. 270, Sept. 2020.

[4] Q. J. Leclerc, E. S. Nightingale, S. Abbott, and T. Jombart, “Analysis of temporal trends in potential COVID-19 cases reported through NHS Pathways England,” *Scientific Reports*, vol. 11, p. 7106, Mar. 2021. Number: 1 Publisher: Nature Publishing Group.

[5] T. Jombart, S. Ghozzi, D. Schumacher, T. J. Taylor, Q. J. Leclerc, M. Jit, S. Flasche, F. Greaves, T. Ward, R. M. Eggo, E. Nightingale, S. Meakin, O. J. Brady, n. null, G. F. Medley, M. Höhle, and W. J. Edmunds, “Real-time monitoring of COVID-19 dynamics using automated trend fitting and anomaly detection,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 376, p. 20200266, July 2021. Publisher: Royal Society.

[6] H. Gibbs, E. Nightingale, Y. Liu, J. Cheshire, L. Danon, L. Smeeth, C. A. B. Pearson, C. Grundy, LSHTM CMMID COVID-19 working group, A. J. Kucharski, and R. M. Eggo, "Detecting behavioural changes in human movement to inform the spatial scale of interventions against COVID-19," *PLOS Computational Biology*, vol. 17, p. e1009162, July 2021.

[7] E. Nightingale, A. Schultze, R. M. Eggo, K. Wing, and The OpenSAFELY Collaborative, "Spatiotemporal risk of infection of care homes during the first wave of the COVID-19 pandemic in the UK," 2021.

[8] A. Schultze, C. Bates, J. Cockburn, B. MacKenna, E. Nightingale, H. J. Curtis, W. J. Hulme, C. E. Morton, R. Croker, S. Bacon, H. I. McDonald, C. T. Rentsch, K. Bhaskaran, R. Mathur, L. A. Tomlinson, E. J. Williamson, H. Forbes, J. Tazare, D. J. Grint, A. J. Walker, P. Inglesby, N. J. DeVito, A. Mehrkar, G. Hickman, S. Davy, T. Ward, L. Fisher, D. Evans, K. Wing, A. Y. Wong, R. McManus, J. Parry, F. Hester, S. Harper, S. J. Evans, I. J. Douglas, L. Smeeth, R. M. Eggo, and B. Goldacre, "Identifying Care Home Residents in Electronic Health Records - An OpenSAFELY Short Data Report," *Wellcome Open Research*, vol. 6, p. 90, Apr. 2021.

[9] A. Schultze, E. Nightingale, D. Evans, W. Hulme, A. Rosello, C. Bates, J. Cockburn, B. MacKenna, H. J. Curtis, C. E. Morton, R. Croker, S. Bacon, H. I. McDonald, C. T. Rentsch, K. Bhaskaran, R. Mathur, L. A. Tomlinson, E. J. Williamson, H. Forbes, J. Tazare, D. Grint, A. J. Walker, P. Inglesby, N. J. DeVito, A. Mehrkar, G. Hickman, S. Davy, T. Ward, L. Fisher, A. C. Green, K. Wing, A. Y. Wong, R. McManus, J. Parry, F. Hester, S. Harper, S. J. Evans, I. J. Douglas, L. Smeeth, R. M. Eggo, B. Goldacre, and D. A. Leon, "Mortality among Care Home Residents in England during the first and second waves of the COVID-19 pandemic: an observational study of 4.3 million adults over the age of 65," *The Lancet Regional Health - Europe*, vol. 14, p. 100295, Mar. 2022.

[10] K. Wing, D. J. Grint, R. Mathur, H. P. Gibbs, G. Hickman, E. Nightingale, A. Schultze, H. Forbes, V. Nafilyan, K. Bhaskaran, E. Williamson, T. House, L. Pellis, E. Herrett, N. Gautam, H. J. Curtis, C. T. Rentsch, A. Y. S. Wong, B. MacKenna, A. Mehrkar, S. Bacon, I. J. Douglas, S. J. W. Evans, L. Tomlinson, B. Goldacre, and R. M. Eggo, "Association between household composition and severe COVID-19 outcomes in older people by ethnicity: an observational cohort study using the

OpenSAFELY platform," *International Journal of Epidemiology*, p. dyac158, Aug. 2022.

[11] V. A. Alegana, P. M. Atkinson, C. Lourenço, N. W. Ruktanonchai, C. Bosco, E. Z. Erbach-Schoenberg, B. Didier, D. Pindolia, A. Le Menach, S. Katokele, P. Uusiku, and A. J. Tatem, "Advances in mapping malaria for elimination: fine resolution modelling of Plasmodium falciparum incidence," *Scientific Reports*, vol. 6, p. 29628, July 2016.

[12] J. N. Odhiambo, C. Kalinda, P. M. Macharia, R. W. Snow, and B. Sartorius, "Spatial and spatio-temporal methods for mapping malaria risk: a systematic review," *BMJ Global Health*, vol. 5, p. e002919, Oct. 2020. Publisher: BMJ Specialist Journals Section: Original research.

[13] D. J. Weiss, T. C. D. Lucas, M. Nguyen, A. K. Nandi, D. Bisanzio, K. E. Battle, E. Cameron, K. A. Twohig, D. A. Pfeffer, J. A. Rozier, H. S. Gibson, P. C. Rao, D. Casey, A. Bertozzi-Villa, E. L. Collins, U. Dalrymple, N. Gray, J. R. Harris, R. E. Howes, S. Y. Kang, S. H. Keddie, D. May, S. Rumisha, M. P. Thorn, R. Barber, N. Fullman, C. K. Huynh, X. Kulikoff, M. J. Kutz, A. D. Lopez, A. H. Mokdad, M. Naghavi, G. Nguyen, K. A. Shackelford, T. Vos, H. Wang, D. L. Smith, S. S. Lim, C. J. L. Murray, S. Bhatt, S. I. Hay, and P. W. Gething, "Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000–17: a spatial and temporal modelling study," *Lancet (London, England)*, vol. 394, p. 322, July 2019. Publisher: Elsevier.

[14] P. MacPherson, M. Khundi, M. Nliwasa, A. T. Choko, V. K. Phiri, E. L. Webb, P. J. Dodd, T. Cohen, R. Harris, and E. L. Corbett, "Disparities in access to diagnosis and care in Blantyre, Malawi, identified through enhanced tuberculosis surveillance and spatial analysis," *BMC Medicine*, vol. 17, p. 21, Jan. 2019.

[15] L. Xia, S. Zhu, C. Chen, Z.-Y. Rao, Y. Xia, D.-X. Wang, P.-R. Zhang, J. He, J.-Y. Zhang, and J.-L. Wu, "Spatio-temporal analysis of socio-economic characteristics for pulmonary tuberculosis in Sichuan province of China, 2006–2015," *BMC Infectious Diseases*, vol. 20, p. 433, June 2020.

[16] A. I. McIntosh, H. E. Jenkins, L. F. White, M. Barnard, D. R. Thomson, T. Dolby, J. Simpson, E. M. Streicher, M. B. Kleinman, E. J. Ragan, P. D. v. Helden, M. B. Murray, R. M. Warren, and K. R. Jacobson, "Using routinely collected laboratory

data to identify high rifampicin-resistant tuberculosis burden communities in the Western Cape Province, South Africa: A retrospective spatiotemporal analysis," *PLOS Medicine*, vol. 15, p. e1002638, Aug. 2018. Publisher: Public Library of Science.

[17] B. J. Coburn, J. T. Okano, and S. Blower, "Using geospatial mapping to design HIV elimination strategies for sub-Saharan Africa," *Science Translational Medicine*, vol. 9, p. eaag0019, Mar. 2017. Publisher: American Association for the Advancement of Science.

[18] A. Aturinde, M. Farnaghi, P. Pilesjö, and A. Mansourian, "Spatial analysis of HIV-TB co-clustering in Uganda," *BMC Infectious Diseases*, vol. 19, p. 612, July 2019.

[19] L. Palk, J. T. Okano, L. Dullie, and S. Blower, "Travel time to health-care facilities, mode of transportation, and HIV elimination in Malawi: a geospatial modelling analysis," *The Lancet Global Health*, vol. 8, pp. e1555–e1564, Dec. 2020. Publisher: Elsevier.

[20] I. N. Nkumama, W. P. O'Meara, and F. H. A. Osier, "Changes in Malaria Epidemiology in Africa and New Challenges for Elimination," *Trends in Parasitology*, vol. 33, pp. 128–140, Feb. 2017.

[21] C. Cotter, H. J. Sturrock, M. S. Hsiang, J. Liu, A. A. Phillips, J. Hwang, C. S. Gueye, N. Fullman, R. D. Gosling, and R. G. Feachem, "The changing epidemiology of malaria elimination: new strategies for new challenges," *The Lancet*, vol. 382, pp. 900–911, Sept. 2013.

[22] R. J. Maude, C. Nguon, A. M. Dondorp, L. J. White, and N. J. White, "The diminishing returns of atovaquone-proguanil for elimination of Plasmodium falciparum malaria: modelling mass drug administration and treatment," *Malaria Journal*, vol. 13, p. 380, Sept. 2014.

[23] V. da Cruz Franco, P. C. Peiter, J. J. Carvajal-Cortés, R. dos Santos Pereira, M. d. S. Mendonça Gomes, and M. C. Suárez-Mutis, "Complex malaria epidemiology in an international border area between Brazil and French Guiana: challenges for elimination," *Tropical Medicine and Health*, vol. 47, p. 24, Apr. 2019.

[24] D. Pedrazzoli, I. Abubakar, H. Potts, P. R. Hunter, M. E. Kruijshaar, O. M. Kon, and J. Southern, "Risk factors for the misdiagnosis of tuberculosis in the UK,

2001–2011," *European Respiratory Journal*, vol. 46, pp. 564–567, Aug. 2015. Publisher: European Respiratory Society Section: Agora.

[25] A. Reid, A. D. Grant, R. G. White, C. Dye, E. Vynnycky, K. Fielding, G. Churchyard, and Y. Pillay, "Accelerating progress towards tuberculosis elimination: the need for combination treatment and prevention," *The International Journal of Tuberculosis and Lung Disease*, vol. 19, pp. 5–9, Jan. 2015.

[26] B. E. Bassey, F. Braka, R. Onyibe, O. O. Kolude, M. Oluwadare, A. Oluwabukola, O. Omotunde, O. A. Iyanda, A. A. Tella, and O. S. Olanike, "Changing epidemiology of yellow fever virus in Oyo State, Nigeria," *BMC Public Health*, vol. 22, p. 467, Mar. 2022.

[27] R. Gilca, G. Deceuninck, B. Lefebvre, R. Tsang, R. Amini, V. Gilca, M. Douville-Fradet, F. Markowski, and P. D. Wals, "The Changing Epidemiology of Meningococcal Disease in Quebec, Canada, 1991–2011: Potential Implications of Emergence of New Strains," *PLOS ONE*, vol. 7, p. e50659, Nov. 2012. Publisher: Public Library of Science.

[28] B. Amoah, C. Fronterre, O. Johnson, M. Dejene, F. Seife, N. Negussu, A. Bakhtiari, E. M. Harding-Esch, E. Giorgi, A. W. Solomon, and P. J. Diggle, "Model-based geostatistics enables more precise estimates of neglected tropical-disease prevalence in elimination settings: mapping trachoma prevalence in Ethiopia," *International Journal of Epidemiology*, vol. 51, pp. 468–478, Apr. 2022.

[29] H.-A. Hatherell, H. Simpson, R. F. Baggaley, T. D. Hollingsworth, and R. L. Pullan, "Sustainable Surveillance of Neglected Tropical Diseases for the Post-Elimination Era," *Clinical Infectious Diseases*, vol. 72, pp. S210–S216, June 2021.

[30] C. A. Lippi, A. M. Stewart-Ibarra, M. Romero, R. Lowe, R. Mahon, C. J. V. Meerbeeck, L. Rollock, M. G.-S. Hilaire, A. R. Trotman, D. Holligan, S. Kirton, M. J. Borbor-Cordova, and S. J. Ryan, "Spatiotemporal Tools for Emerging and Endemic Disease Hotspots in Small Areas: An Analysis of Dengue and Chikungunya in Barbados, 2013–2016," *The American Journal of Tropical Medicine and Hygiene*, vol. 103, pp. 149–156, Apr. 2020. Publisher: The American Society of Tropical Medicine and Hygiene Section: The American Journal of Tropical Medicine and Hygiene.

[31] P. R. Naufal Spir, L. E. Prestes-Carneiro, E. S. Fonseca, A. Dayse, R. Giuffrida, and L. A. Z. D'Andrea, "Clinical characteristics and spatial distribution of Visceral leish-

maniasis in children in São Paulo state: an emerging focus of Visceral leishmaniasis in Brazil," *Pathogens and Global Health*, vol. 111, pp. 91–97, Feb. 2017. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/20477724.2017.1289666.

[32] C. Robertson and T. A. Nelson, "An Overview of Spatial Analysis of Emerging Infectious Diseases," *The Professional Geographer*, vol. 66, pp. 579–588, Oct. 2014. Publisher: Routledge _eprint: https://doi.org/10.1080/00330124.2014.907702.

[33] S. Burza, S. L. Croft, and M. Boelaert, "Leishmaniasis," *The Lancet*, vol. 392, pp. 951–970, Sept. 2018. Publisher: Elsevier.

[34] P. Ready, "Epidemiology of visceral leishmaniasis," *Clinical Epidemiology*, vol. 6, no. 1, pp. 147–154, 2014. ISBN: 1179-1349 (Electronic)$\backslash$r1179-1349 (Linking) Publisher: Dove Press.

[35] N. S. Singh and D. P. Singh, "A Review on Major Risk Factors and Current Status of Visceral Leishmaniasis in North India," *American Journal of Entomology*, vol. 3, no. 1, pp. 6–14, 2019.

[36] National Center for Vector Borne Diseases Control (NCVBDC), "KALA-AZAR SITUATION IN INDIA."

[37] WHO, "Accelerating Work To Overcome the Global Impact of Neglected Tropical," *Accelerating work to overcome neglected tropical diseases: a roadmap for implementation.*, 2012.

[38] World Health Organization, Regional Office for South-East Asia, "Independent Assessment of Kala-Azar Eliminaiton Programme in India," tech. rep., World Health Organization, 2020.

[39] D. M. Poche, R. B. Garlapati, S. Mukherjee, Z. Torres-Poche, E. Hasker, T. Rahman, A. Bharti, V. P. Tripathi, S. Prakash, R. Chaubey, and R. M. Poche, "Bionomics of Phlebotomus argentipes in villages in Bihar, India with insights into efficacy of IRS-based control measures," *PLoS Neglected Tropical Diseases*, vol. 12, no. 1, pp. 1–20, 2018.

[40] V. Kumar, R. Mandal, S. Das, S. Kesari, D. S. Dinesh, K. Pandey, V. R. Das, R. K. Topno, M. P. Sharma, R. K. Dasgupta, and P. Das, "Kala-azar elimination in a highly-endemic district of Bihar, India: A success story," *PLoS Neglected Tropical Diseases*, vol. 14, May 2020.

[41] E. E. Zijlstra, A. M. Musa, E. a. G. Khalil, I. E. Hassan, and A. M. El-Hassan, "Post-kala-azar dermal leishmaniasis," *The Lancet Infectious Diseases*, vol. 3, pp. 87–98, Feb. 2003. Publisher: Elsevier.

[42] E. E. Zijlstra, F. Alves, S. Rijal, B. Arana, and J. Alvar, "Post-kala-azar dermal leishmaniasis in the Indian subcontinent: A threat to the South-East Asia Region Kala-azar Elimination Programme.," *PLOS Neglected Tropical Diseases*, vol. 11, p. e0005877, Nov. 2017. Publisher: Public Library of Science.

[43] P. Salotra, H. Kaushal, and V. Ramesh, "Containing Post Kala-Azar Dermal Leishmaniasis (PKDL): Pre-requisite for Sustainable Elimination of Visceral Leishmaniasis (VL) from South Asia," in *Kala Azar in South Asia: Current Status and Sustainable Challenges* (E. Noiri and T. Jha, eds.), pp. 7–21, Cham: Springer International Publishing, 2016.

[44] A. Kumar, S. Saurabh, S. Jamil, and V. Kumar, "Intensely clustered outbreak of visceral leishmaniasis (kala-azar) in a setting of seasonal migration in a village of Bihar, India," *BMC Infectious Diseases*, vol. 20, p. 10, Jan. 2020.

[45] C. Dye, "The epidemiology of canine visceral leishmaniasis in southern France: classical theory offers another explanation of the data," *Parasitology*, vol. 96, p. 19, Feb. 1988. Publisher: Cambridge University Press.

[46] D. Bora, "Epidemiology of visceral leishmaniasis in India," *The National Medical Journal of India*, vol. 12, no. 2, pp. 62–68, 1999. ISBN: 0970-258X (Print).

[47] O. Courtenay, N. C. Peters, M. E. Rogers, and C. Bern, "Combining epidemiology with basic biology of sand flies, parasites, and hosts to inform leishmaniasis transmission dynamics and control," *PLoS Pathogens*, vol. 13, no. 10, p. e1006571, 2017.

[48] P. Saha, S. Ganguly, M. Chatterjee, S. B. Das, P. K. Kundu, S. K. Guha, T. K. Ghosh, D. K. Bera, N. Basu, and A. K. Maji, "Asymptomatic leishmaniasis in kala-azar endemic areas of Malda district, West Bengal, India," *PLoS Neglected Tropical Diseases*, vol. 11, Feb. 2017.

[49] D. J. Lim, M. R. Banjara, V. K. Singh, A. B. Joshi, C. K. Gurung, M. L. Das, G. Matlashewski, P. Olliaro, and A. Kroeger, "Barriers of Visceral Leishmaniasis reporting and surveillance in Nepal: comparison of governmental VL-program

districts with non-program districts," *Tropical Medicine and International Health*, vol. 24, pp. 192–204, Feb. 2019. Publisher: Blackwell Publishing Ltd.

[50] J. P. Boettcher, Y. Siwakoti, A. Milojkovic, N. A. Siddiqui, C. K. Gurung, S. Rijal, P. Das, A. Kroeger, and M. R. Banjara, "Visceral leishmaniasis diagnosis and reporting delays as an obstacle to timely response actions in Nepal and India," *BMC Infectious Diseases*, vol. 15, p. 43, Dec. 2015.

[51] I. c. . U. K. R. Parth MN, "Bihar's migrant workers are returning to cities as rural employment schemes fall short," 2020. Publisher: https://scroll.in.

[52] J. Toor, E. R. Adams, M. Aliee, B. Amoah, R. M. Anderson, D. Ayabina, R. Bailey, M.-G. Basáñez, D. J. Blok, S. Blumberg, A. Borlase, R. C. Rivera, M. S. Castaño, N. Chitnis, L. E. Coffeng, R. E. Crump, A. Das, C. N. Davis, E. L. Davis, M. S. Deiner, P. J. Diggle, C. Fronterre, F. Giardina, E. Giorgi, M. Graham, J. I. D. Hamley, C.-I. Huang, K. Kura, T. M. Lietman, T. C. D. Lucas, V. Malizia, G. F. Medley, A. Meeyai, E. Michael, T. C. Porco, J. M. Prada, K. S. Rock, E. A. Le Rutte, M. E. Smith, S. E. F. Spencer, W. A. Stolk, P. Touloupou, A. Vasconcelos, C. Vegvari, S. J. de Vlas, M. Walker, and T. D. Hollingsworth, "Predicted Impact of COVID-19 on Neglected Tropical Disease Programs and the Opportunity for Innovation," *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, Sept. 2020.

[53] E. A. L. Rutte, L. E. Coffeng, J. Muñoz, and S. J. d. Vlas, "Modelling the impact of COVID-19-related programme interruptions on visceral leishmaniasis in India," *medRxiv*, p. 2020.10.26.20219758, Oct. 2020. Publisher: Cold Spring Harbor Laboratory Press.

[54] World Health Organization, "Coronavirus disease (COVID-19)."

[55] Public Health England, "GOV.UK Coronavirus (COVID-19) in the UK."

[56] E. J. Williamson, A. J. Walker, K. Bhaskaran, S. Bacon, C. Bates, C. E. Morton, H. J. Curtis, A. Mehrkar, D. Evans, P. Inglesby, J. Cockburn, H. I. McDonald, B. MacKenna, L. Tomlinson, I. J. Douglas, C. T. Rentsch, R. Mathur, A. Y. S. Wong, R. Grieve, D. Harrison, H. Forbes, A. Schultze, R. Croker, J. Parry, F. Hester, S. Harper, R. Perera, S. J. W. Evans, L. Smeeth, and B. Goldacre, "Factors associated with COVID-19-related death using OpenSAFELY," *Nature*, vol. 584, pp. 430–436, Aug. 2020. Number: 7821 Publisher: Nature Publishing Group.

[57] NHS UK, "Coronavirus (COVID-19): getting tested."

[58] P. Cowpertwait and A. Metcalfe, *Introductory Time Series With R.* Springer International Publishing, Jan. 2009.

[59] N. Cressie, "Geostatistics," in *Statistics for Spatial Data*, pp. 27–104, John Wiley & Sons, Ltd, 1993. Section: 2 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119115151.ch2.

[60] S. Meyer, L. Held, and M. Höhle, "Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance," *Journal of Statistical Software*, vol. 77, no. 11, 2017. ISBN: 9780128048283 _eprint: 1411.0416.

[61] A. B. Lawson, S. Banerjee, R. P. Haining, and M. D. Ugarte, eds., *Handbook of Spatial Epidemiology.* New York: Chapman and Hall/CRC, Apr. 2016.

[62] M. B. Girolami, Mark, "Hamiltonian Monte Carlo for Hierarchical Models," in *Current Trends in Bayesian Methodology with Applications*, Chapman and Hall/CRC, 2015. Num Pages: 24.

[63] H. Rue, S. Martino, and N. Chopin, "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 2, pp. 319–392, 2009. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2008.00700.x.

[64] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, Mar. 1951. Publisher: Institute of Mathematical Statistics.

[65] T. D. Smedt, K. Simons, A. V. Nieuwenhuyse, and G. Molenberghs, "Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models," *Archives of Public Health*, vol. 73, no. Suppl 1, p. O2, 2015. Publisher: BioMed Central.

[66] D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye, "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors," *Statistical Science*, vol. 32, pp. 1–28, Feb. 2017. Publisher: Institute of Mathematical Statistics.

[67] J. Besag, J. York, and A. Mollié, "Bayesian image restoration, with two applications in spatial statistics," *Annals of the Institute of Statistical Mathematics*, vol. 43, pp. 1–20, Mar. 1991.

[68] A. Riebler, S. H. Sørbye, D. Simpson, and H. Rue, "An intuitive Bayesian spatial model for disease mapping that accounts for scaling," *Statistical Methods in Medical Research*, vol. 25, pp. 1145–1165, Aug. 2016.

[69] F. Lindgren, H. Rue, and J. Lindström, "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 4, pp. 423–498, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2011.00777.x.

[70] A. K. Nandi, T. C. D. Lucas, R. Arambepola, P. Gething, and D. J. Weiss, "disaggregation: An R Package for Bayesian Spatial Disaggregation Modelling," Jan. 2020. arXiv:2001.04847 [stat].

[71] R. Arambepola, T. C. D. Lucas, A. K. Nandi, P. W. Gething, and E. Cameron, "A simulation study of disaggregation regression for spatial disease mapping," May 2020. Number: arXiv:2005.03604 arXiv:2005.03604 [stat].

[72] A. Schwartzer, "Incidence trends and spatiotemporal clustering of visceral leishmaniasis in Bihar, India from 2013-2018," Master's thesis, London School of Hygiene & Tropical Medicine, 2019.

[73] E. S. Nightingale, L. A. C. Chapman, S. Srikantiah, S. Subramanian, P. Jambulingam, J. Bracher, M. M. Cameron, and G. F. Medley, "A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India," *PLOS Neglected Tropical Diseases*, vol. 14, p. e0008422, July 2020. Number: 7 Publisher: Public Library of Science.

[74] R. Bivand, "Revisiting the Boston data set - Changing the units of observation affects estimated willingness to pay for clean air," *REGION*, vol. 4, p. 109, May 2017.

[75] N. V. B. D. C. Programme, "Standard Operating Procedure for Kala-Azar and Post-Kala-Azar Dermal Leishmaniasis Case Search," 2020.

[76] J. Bindroo, K. Priyamvada, L. A. C. Chapman, T. Mahapatra, B. Sinha, I. Banerjee, P. K. Mishra, B. Rooj, K. Kundan, N. Roy, N. K. Gill, A. Hightower, M. P. Sharma, N. Dhingra, C. Bern, and S. Srikantiah, "Optimizing Village-Level Targeting of Active Case Detection to Support Visceral Leishmaniasis Elimination in India," *Frontiers in Cellular and Infection Microbiology*, vol. 11, p. 648847, 2021.

[77] C. A. Bulstra, E. A. Le Rutte, P. Malaviya, E. C. Hasker, L. E. Coffeng, A. Picado, O. P. Singh, M. C. Boelaert, S. J. de Vlas, and S. Sundar, "Visceral leishmaniasis: spatiotemporal heterogeneity and drivers underlying the hotspots in Muzaffarpur, Bihar, India," *PLoS Neglected Tropical Diseases*, vol. 12, no. 12, p. e0006888, 2018. Publisher: Public Library of Science.

[78] D. Perry, K. Dixon, R. Garlapati, A. Gendernalik, D. Poché, and R. Poché, "Visceral leishmaniasis prevalence and associated risk factors in the saran district of Bihar, India, from 2009 to July of 2011," *The American Journal of Tropical Medicine and Hygiene*, vol. 88, pp. 778–784, Apr. 2013.

[79] L. A. C. Chapman, S. E. F. Spencer, T. M. Pollington, C. P. Jewell, D. Mondal, J. Alvar, T. D. Hollingsworth, M. M. Cameron, C. Bern, and G. F. Medley, "Inferring transmission trees to guide targeting of interventions against visceral leishmaniasis and post–kala-azar dermal leishmaniasis," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 25742–25750, Oct. 2020. Publisher: Proceedings of the National Academy of Sciences.

[80] G. Bhunia, N. Chatterjee, V. Kumar, N. Siddiqui, R. Mandal, S. Das, and S. Kesari, "Delimitation of kala-azar risk areas in the district of Vaishali in Bihar (India) using a geo-environmental approach," *Memórias do Instituto Oswaldo Cruz*, vol. 107, pp. 609–20, Aug. 2012.

[81] G. S. Bhunia, S. Kesari, N. Chatterjee, V. Kumar, and P. Das, "Localization of kala-azar in the endemic region of Bihar, India based on land use/land cover assessment at different scales," *Geospatial Health*, vol. 6, pp. 177–193, May 2012.

[82] S. Sudhakar, T. Srinivas, A. Palit, S. K. Kar, and S. K. Battacharya, "Mapping of risk prone areas of kala-azar (Visceral leishmaniasis) in parts of Bihar State, India: an RS and GIS approach," *Journal of Vector Borne Diseases*, vol. 43, pp. 115–122, Sept. 2006.

[83] C. Bern, O. Courtenay, and J. Alvar, “Of Cattle, Sand Flies and Men: A Systematic Review of Risk Factor Analyses for South Asian Visceral Leishmaniasis and Implications for Elimination,” *PLOS Neglected Tropical Diseases*, vol. 4, p. e599, Feb. 2010. Publisher: Public Library of Science.

[84] A. Y. M. Abdullah, A. Dewan, M. R. I. Shogib, M. M. Rahman, and M. F. Hossain, “Environmental factors associated with the distribution of visceral leishmaniasis in endemic areas of Bangladesh: modeling the ecological niche,” *Tropical Medicine and Health*, vol. 45, p. 13, 2017.

[85] C. E. Utazi, J. Thorley, V. A. Alegana, M. J. Ferrari, K. Nilsen, S. Takahashi, C. J. Metcalf, J. Lessler, and A. J. Tatem, “A spatial regression model for the disaggregation of areal unit based data to high-resolution grids with application to vaccination coverage mapping,” *Statistical Methods in Medical Research*, 2018.

[86] F. R. Stevens, A. E. Gaughan, C. Linard, and A. J. Tatem, “Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data,” *PLOS ONE*, vol. 10, p. e0107042, Feb. 2015. Publisher: Public Library of Science.

[87] T. C. D. Lucas, A. K. Nandi, S. H. Keddie, E. G. Chestnutt, R. E. Howes, S. F. Rumisha, R. Arambepola, A. Bertozzi-Villa, A. Python, T. L. Symons, J. J. Millar, P. Amratia, P. Hancock, K. E. Battle, E. Cameron, P. W. Gething, and D. J. Weiss, “Improving disaggregation models of malaria incidence by ensembling non-linear models of prevalence,” *Spatial and Spatio-temporal Epidemiology*, p. 100357, July 2020.

[88] A. Python, A. Bender, M. Blangiardo, J. B. Illian, Y. Lin, B. Liu, T. Lucas, S. Tan, Y. Wen, D. Svanidze, and J. Yin, “A Downscaling Approach to Compare COVID-19 Count Data from Databases Aggregated at Different Spatial Scales,” SSRN Scholarly Paper ID 3627252, Social Science Research Network, Rochester, NY, June 2020.

[89] H. J. Sturrock, J. M. Cohen, P. Keil, A. J. Tatem, A. Le Menach, N. E. Ntshalintshali, M. S. Hsiang, and R. D. Gosling, “Fine-scale malaria risk mapping from routine aggregated case data,” *Malaria Journal*, vol. 13, p. 421, Nov. 2014.

[90] I. C. Hanigan, T. B. Chaston, B. Hinze, M. Dennekamp, B. Jalaludin, Y. Kinfu, and G. G. Morgan, “A statistical downscaling approach for generating high spatial resolution health risk maps: a case study of road noise and ischemic heart disease

mortality in Melbourne, Australia," *International Journal of Health Geographics*, vol. 18, p. 20, Sept. 2019.

[91] S. Majumder, Y. Guan, B. J. Reich, S. O'Neill, and A. G. Rappold, "Statistical downscaling with spatial misalignment: Application to wildland fire PM2.5 concentration forecasting," *Journal of agricultural, biological, and environmental statistics*, vol. 26, pp. 23–44, Mar. 2021.

[92] WorldPop, "Global 100m Population total adjusted to match the corresponding UNPD estimate," 2020.

[93] WorldPop and M. Bondarenko, "Global BSGM outputs 100m," Nov. 2018.

[94] D. J. Weiss, A. Nelson, H. S. Gibson, W. Temperley, S. Peedell, A. Lieber, M. Hancher, E. Poyart, S. Belchior, N. Fullman, B. Mappin, U. Dalrymple, J. Rozier, T. C. D. Lucas, R. E. Howes, L. S. Tusting, S. Y. Kang, E. Cameron, D. Bisanzio, K. E. Battle, S. Bhatt, and P. W. Gething, "A global map of travel time to cities to assess inequalities in accessibility in 2015," *Nature*, vol. 553, pp. 333–336, Jan. 2018.

[95] Z. Wan, S. Hook, and G. Hulley, "MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006," 2015.

[96] K. Didan, "MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V006," 2015.

[97] AppEEARS Team., "Application for Extracting and Exploring Analysis Ready Samples (AppEEARS). Ver. 3.3.1.," 2022.

[98] "moran.mc: Permutation test for Moran's I statistic in spdep: Spatial Dependence: Weighting Schemes, Statistics."

[99] "TMB Documentation: Introduction."

[100] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.

[101] A. Liaw and M. Wiener, "Classification and Regression by randomForest.," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[102] D. G. Bonett and T. A. Wright, "Sample size requirements for estimating pearson, kendall and spearman correlations," *Psychometrika*, vol. 65, pp. 23–28, Mar. 2000.

[103] L. A. C. Chapman, C. P. Jewell, S. E. F. Spencer, L. Pellis, S. Datta, R. Chowdhury, C. Bern, G. F. Medley, and T. D. Hollingsworth, "The role of case proximity in transmission of visceral leishmaniasis in a highly endemic village in Bangladesh," *PLOS Neglected Tropical Diseases*, vol. 12, Oct. 2018. Publisher: Public Library of Science.

[104] D. M. Poché, Z. Torres-Poché, R. Garlapati, T. Clarke, and R. M. Poché, "Short-term movement of Phlebotomus argentipes (Diptera: Psychodidae) in a visceral leishmaniasis-endemic village in Bihar, India," *Journal of Vector Ecology: Journal of the Society for Vector Ecology*, vol. 43, pp. 285–292, Dec. 2018.

[105] A. Datta, "Circular Migration and Precarity: Perspectives from Rural Bihar," *The Indian Journal of Labour Economics: The Quarterly Journal of the Indian Society of Labour Economics*, vol. 63, no. 4, pp. 1143–1163, 2020.

[106] K. Priyamvada, J. Bindroo, M. P. Sharma, L. A. C. Chapman, P. Dubey, T. Mahapatra, A. W. Hightower, C. Bern, and S. Srikantiah, "Visceral leishmaniasis outbreaks in Bihar: community-level investigations in the context of elimination of kala-azar as a public health problem," *Parasites & Vectors*, vol. 14, p. 52, Jan. 2021.

[107] A. Ranjan, T. Bhatnagar, G. R. Babu, and R. Detels, "Sexual Behavior, HIV Prevalence and Awareness Among Wives of Migrant Workers: Results from Cross-sectional Survey in Rural North India," *Indian Journal of Community Medicine: Official Publication of Indian Association of Preventive & Social Medicine*, vol. 42, pp. 24–29, Mar. 2017.

[108] R. P. Roshania, R. Giri, S. A. Cunningham, M. F. Young, A. Webb-Girard, A. Das, G. S. Mala, S. Srikantiah, T. Mahapatra, and U. Ramakrishnan, "Early life migration and undernutrition among circular migrant children: An observational study in the brick kilns of Bihar, India," *Journal of Global Health*, vol. 12, p. 04008, 2022.

[109] G. Bhunia, V. Kumar, J. Kumar, S. Das, and S. Kesari, "The use of remote sensing in the identification of the eco-environmental factors associated with the risk of human visceral leishmaniasis (kala-azar) on the Gangetic plain, in north-eastern India," *Annals of tropical medicine and parasitology*, vol. 104, pp. 35–53, Jan. 2010.

[110] N. N. H. Valero and M. Uriarte, "Environmental and socioeconomic risk factors associated with visceral and cutaneous leishmaniasis: a systematic review," *Parasitology Research*, vol. 119, pp. 365–384, Feb. 2020.

[111] K. Cloots, P. Marino, S. Burza, N. Gill, M. Boelaert, and E. Hasker, "Visceral Leishmaniasis-HIV Coinfection as a Predictor of Increased Leishmania Transmission at the Village Level in Bihar, India," *Frontiers in Cellular and Infection Microbiology*, vol. 11, p. 604117, 2021.

[112] K. Wilson and J. Wakefield, "Pointless spatial modeling," *Biostatistics*, vol. 21, pp. e17–e32, Apr. 2020.

[113] J. S. Hodges and B. J. Reich, "Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love," *The American Statistician*, vol. 64, pp. 325–334, Nov. 2010. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/tast.2010.10052.

[114] L. E. Coffeng, E. A. Le Rutte, J. Muñoz, E. R. Adams, J. M. Prada, S. J. de Vlas, and G. F. Medley, "Impact of Changes in Detection Effort on Control of Visceral Leishmaniasis in the Indian Subcontinent," *The Journal of Infectious Diseases*, vol. 221, pp. S546–S553, June 2020.

[115] H.-l. Li, R.-j. Zheng, Q. Zheng, W. Jiang, X.-l. Zhang, W.-m. Wang, X. Feng, K. Wang, and X.-b. Lu, "Predicting the number of visceral leishmaniasis cases in Kashgar, Xinjiang, China using the ARIMA-EGARCH model," *Asian Pacific Journal of Tropical Medicine*, vol. 13, p. 81, Feb. 2020. Company: Medknow Publications and Media Pvt. Ltd. Distributor: Medknow Publications and Media Pvt. Ltd. Institution: Medknow Publications and Media Pvt. Ltd. Label: Medknow Publications and Media Pvt. Ltd. Publisher: Medknow Publications.

[116] V. Rahmanian, S. Bokaie, A. Haghdoost, and M. Barooni, "Temporal analysis of visceral leishmaniasis between 2000 and 2019 in Ardabil Province, Iran: A time-series study using ARIMA model," *Journal of Family Medicine and Primary Care*, vol. 9, pp. 6061–6067, Dec. 2020.

[117] K. B. A. Pimentel, R. S. Oliveira, C. F. Aragão, J. Aquino Júnior, M. E. S. Moura, A. S. Guimarães-e Silva, V. C. S. Pinheiro, E. G. R. Gonçalves, and A. R. Silva, "Prediction of visceral leishmaniasis incidence using the Seasonal Autoregressive Integrated Moving Average model (SARIMA) in the state of Maranhão, Brazil," *Brazilian Journal of Biology*, vol. 84, Jan. 2022. Publisher: Instituto Internacional de Ecologia.

[118] L. Hussain-Alkhateeb, T. R. Ramírez, A. Kroeger, E. Gozzer, and S. Runge-Ranzinger, "Early warning systems (EWSs) for chikungunya, dengue, malaria, yel-

low fever, and Zika outbreaks: What is the evidence? A scoping review," *PLOS Neglected Tropical Diseases*, vol. 15, p. e0009686, Sept. 2021. Publisher: Public Library of Science.

[119] N. Dial, S. Croft, L. Chapman, F. Terris-Prestholt, and G. Medley, "Challenges of using modelling evidence in the visceral leishmaniasis elimination programme in India," *PLOS Global Public Health*, vol. 2, p. e0001049, Nov. 2022.

[120] J. Aagaard-Hansen, N. Nombela, and J. Alvar, "Population movement: a key factor in the epidemiology of neglected tropical diseases," *Tropical medicine & international health: TM & IH*, vol. 15, pp. 1281–1288, Nov. 2010.

[121] S. A. Lee, C. I. Jarvis, W. J. Edmunds, T. Economou, and R. Lowe, "Spatial connectivity in mosquito-borne disease models: a systematic review of methods and assumptions," *Journal of The Royal Society Interface*, vol. 18, no. 178, 2021. Publisher: Royal Society.

[122] P. J. Hotez, A. Fenwick, and D. Molyneux, "The new COVID-19 poor and the neglected tropical diseases resurgence," *Infectious Diseases of Poverty*, vol. 10, p. 10, Jan. 2021.

[123] S. Irudaya Rajan, P. Sivakumar, and A. Srinivasan, "The COVID-19 Pandemic and Internal Labour Migration in India: A 'Crisis of Mobility'," *The Indian Journal of Labour Economics*, vol. 63, pp. 1021–1039, Dec. 2020.

[124] K. D. Rao, J. Kaur, M. A. Peters, N. Kumar, and P. Nanda, "Pandemic response in pluralistic health systems: a cross-sectional study of COVID-19 knowledge and practices among informal and formal primary care providers in Bihar, India," *BMJ Open*, vol. 11, p. e047334, Apr. 2021. Publisher: British Medical Journal Publishing Group Section: Global health.

[125] R. Suresh, J. James, and B. R. S.j, "Migrant Workers at Crossroads–The Covid-19 Pandemic and the Migrant Experience in India," *Social Work in Public Health*, vol. 35, pp. 633–643, Sept. 2020.

[126] D. Lee and R. Mitchell, "Locally adaptive spatial smoothing using conditional autoregressive models," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 62, no. 4, pp. 593–608, 2013. Publisher: [Wiley, Royal Statistical Society].

[127] S. A. Lee, T. Economou, and R. Lowe, “A Bayesian modelling framework to quantify multiple sources of spatial variation for disease mapping,” *Journal of The Royal Society Interface*, vol. 19, no. 194, 2022. Publisher: Royal Society.

[128] Stan Development Team, “Stan Modelling Language Users Guide and Reference Manual,” 2022.

[129] K. K. Avilov, A. A. Romanyukha, S. E. Borisov, E. M. Belilovsky, O. B. Nechaeva, and A. S. Karkach, “An approach to estimating tuberculosis incidence and case detection rate from routine notification data,” *The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease*, vol. 19, pp. 288–294, i–x, Mar. 2015.

[130] D. Shaweno, J. M. Trauer, J. T. Denholm, and E. S. McBryde, “A novel Bayesian geospatial method for estimating tuberculosis incidence reveals many missed TB cases in Ethiopia,” *BMC Infectious Diseases*, vol. 17, p. 662, Oct. 2017.

[131] D. L. Miller, M. L. Burt, E. A. Rexstad, and L. Thomas, “Spatial models for distance sampling data: recent developments and future directions,” *Methods in Ecology and Evolution*, vol. 4, no. 11, pp. 1001–1010, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12105.

[132] T. Lancet, “50 years of the inverse care law,” *The Lancet*, vol. 397, p. 767, Feb. 2021. Publisher: Elsevier.

[133] S. Meyer, J. Elias, and M. Höhle, “A Space-Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence,” *Biometrics*, vol. 68, no. 2, pp. 607–616, 2012. _eprint: 1508.05740.

[134] M.-A. Rizoiu, S. Mishra, Q. Kong, M. Carman, and L. Xie, “SIR-Hawkes: Linking Epidemic Models and Hawkes Processes to Model Diffusions in Finite Populations,” in *Proceedings of the 2018 World Wide Web Conference*, WWW ’18, (Republic and Canton of Geneva, CHE), pp. 419–428, International World Wide Web Conferences Steering Committee, Apr. 2018.

[135] A. Reinhart, “A Review of Self-Exciting Spatio-Temporal Point Processes and Their Applications,” *Statistical Science*, vol. 33, pp. 299–318, Aug. 2017. Number: 3 Publisher: Institute of Mathematical Statistics _eprint: 1708.02647.

[136] H. J. T. Unwin, I. Routledge, S. Flaxman, M.-A. Rizoiu, S. Lai, J. Cohen, D. J. Weiss, S. Mishra, and S. Bhatt, "Using Hawkes Processes to model imported and local malaria cases in near-elimination settings," *PLOS Computational Biology*, vol. 17, p. e1008830, Apr. 2021. Publisher: Public Library of Science.

[137] F. Serafini, F. Lindgren, and M. Naylor, "Approximation of bayesian Hawkes process models with Inlabru," June 2022. arXiv:2206.13360 [stat].

[138] P. J. Diggle, R. Menezes, and T.-l. Su, "Geostatistical inference under preferential sampling," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 59, no. 2, pp. 191–232, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2009.00701.x.

[139] J. Watson, J. V. Zidek, and G. Shaddick, "A general theory for preferential sampling in environmental networks," *The Annals of Applied Statistics*, vol. 13, pp. 2662–2700, Dec. 2019. Publisher: Institute of Mathematical Statistics.

Appendices

A Supplementary Materials: A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India

Supporting Information

S1 Text

Model Selection

Starting with a basic, endemic-only model (including a population offset and linear trend in time), potential extensions of the three core components were added in turn and measures of fit and predictive power were calculated. The addition which yielded the best improvement in the RPS of OSA predictions, subject to calibration (p not less than 0.1 for test of calibration based on RPS), was selected and then all remaining options tested again. This process was repeated until no further extension of the model made a significant ($p < 0.001$) improvement to predictive power (as determined by a permutation test on the RPS). This stringent criterium was employed in order to prioritise simplicity over complexity. If at any point an individual model parameter lost significance, the element associated with this parameter was removed in subsequent models.

Empirical Coverage Probabilities

Again using a one-step-ahead approach, the 25th and 75th quantiles of the predicted distribution were calculated and a score of 0 or 1 assigned if the observed value fell inside or outside this quantile range respectively. This binary score was assigned for each block and each month in the test set, such that we could subsequently calculate a proportion of prediction intervals which did not capture the true count. Thus, the overall score, C, is given by

$$C = \frac{1}{n_i n_t} \sum_{i,t} \mathbb{1}[y_{it} \leq q_{i,t,0.25} | y_{it} \geq q_{i,t,0.75}] \tag{1}$$

where y_{it} is the observed count for block i at month t, n_i and n_t the total number of blocks and months respectively, and $q_{i,t,p}$ the p^{th} quantile of the predicted distribution. We also investigated such a score using 10th and 90th quantiles, to ascertain whether these could be used as approximate lower and upper bounds for case counts.

S1 Table

Fit and prediction metrics for selected model at each stage. The value aof *S* within the formula indicates the number of seasonal waves included. The reported AIC is for the fit to training data only, and RPS is of predictions made without updating this fit (i.e. fixed instead of rolling). C2575 and C1090 refer to the coverage of 50% and 80% quantile intervals, respectively, alongside the average interval width in cases. Model no. 42 is the final model.

Stage	Model No.	END	AR	NE	Dispersion	No. parameters	AIC	RPS	Calibration (p-value)	C1090	Avg. width
0	1	offset + 1 + t			1	3	65412	0.657	<0.0001	0.095	2.243
1	2	offset + 1 + t + seas(∼1, S=1)			1	5	65227	0.654	<0.0001	0.090	2.330
1	3	offset + 1			1	2	65811	0.698	<0.0001	0.044	4.158
1	4	offset + 1 + t + logpopdens			1	4	65708	0.662	<0.0001	0.094	2.180
1	5	offset + 1 + t	AR(1)		1	4	57100	0.495	0.109	0.060	2.386
1	**6**	**offset + 1 + t**	**AR(1) + seas(∼1, S=1)**		**1**	**6**	**57058**	**0.493**	**0.115**	**0.058**	**2.388**
1	7	offset + 1 + t	AR(1) + seas(∼1 + t, S=1)		1	7	57031	0.496	<0.0001	0.064	2.171
1	8	offset + 1 + t		NE(2)	1	5	56755	0.516	0.003	0.056	2.304
1	9	offset + 1 + t		NE(2) + logpopdens	1	5	56763	0.516	0.002	0.056	2.313
1	10	offset + 1 + t		NE(2) + seas(∼1, S = 1)	1	7	56685	0.515	0.001	0.054	2.308
1	11	offset + 1 + t		NE(2) + seas(∼1 + t, S = 1)	1	8	56680	0.516	0.203	0.057	2.201
1	12	offset + 1 + t			State	4	65310	0.659	<0.0001	0.098	2.145
2	13	offset + 1 + seas(∼1, S=1)	AR(1) + seas(∼1, S=1)		1	7	57024	0.502	<0.0001	0.048	2.627
2	14	offset + 1	AR(1) + seas(∼1, S=1)		1	5	57101	0.502	<0.0001	0.049	2.612
2	15	offset + 1 + t + logpopdens	AR(1) + seas(∼1, S=1)		1	6	57128	0.499	<0.0001	0.055	2.496
2	16	offset + 1 + t	AR(1) + seas(∼1 + t, S=1)		1	7	57031	0.496	<0.0001	0.064	2.171
2	17	offset + 1 + t	AR(1) + seas(∼1 + t, S=2)		1	9	56996	0.496	<0.0001	0.064	2.176
2	18	offset + 1 + t	AR(1) + seas(∼1, S=1)	NE(2)	1	8	53362	0.458	0.210	0.055	2.105
2	19	offset + 1 + t	AR(1) + seas(∼1, S=1)	NE(2) + seas(∼1, S = 1)	1	10	53300	0.457	0.294	0.053	2.101
2	20	offset + 1 + t	AR(1) + seas(∼1, S=1)	NE(2) + seas(∼1 + t, S = 1)	1	11	53301	0.458	0.125	0.053	2.122
2	21	offset + 1 + t	AR(1) + seas(∼1, S=1)	NE(2) + logpopdens	1	8	53398	0.458	0.144	0.054	2.111
2	22	offset + 1 + t	AR(1) + seas(∼1, S=1)		State	7	57059	0.493	0.123	0.058	2.389
2	**23**	**offset + 1 + t**	**AR(2) + seas(∼1, S=1)**		**1**	**6**	**53833**	**0.455**	**0.189**	**0.053**	**2.230**
2	24	offset + 1 + t	AR(3) + seas(∼1, S=1)		1	6	52279	0.439	0.005	0.061	2.017
2	25	offset + 1 + t	AR(4) + seas(∼1, S=1)		1	6	51342	0.428	<0.0001	0.064	1.877
3	26	offset + 1 + seas(∼1, S=1)	AR(2) + seas(∼1, S=1)		1	7	53806	0.457	<0.0001	0.043	2.395
3	27	offset + 1	AR(2) + seas(∼1, S=1)		1	5	53844	0.458	<0.0001	0.047	2.340

Stage	Model No.	END	AR	NE	Dispersion	No. parameters	AIC	RPS	Calibration (p-value)	C1090	Avg. width
3	28	offset + 1 + t + logpopdens	AR(2) + seas(∼1, S=1)		1	6	53835	0.456	<0.0001	0.042	2.404
3	29	offset + 1 + t	AR(2) + seas(∼1 + t, S=1)		1	7	53815	0.455	0.002	0.056	2.087
3	30	offset + 1 + t	AR(2) + seas(∼1 + t, S=2)		1	9	53692	0.455	0.001	0.057	2.079
3	31	offset + 1 + t	AR(3) + seas(∼1, S=1)		1	6	52279	0.439	0.005	0.061	2.017
3	32	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(1)	1	7	51749	0.437	0.181	0.054	1.974
3	**33**	**offset + 1 + t**	**AR(2) + seas(∼1, S=1)**	**NE(1) + seas(∼1, S = 1)**	**1**	**9**	**51675**	**0.437**	**0.122**	**0.055**	**1.966**
3	34	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(3) + seas(∼1 + t, S = 1)	1	11	51543	0.437	0.656	0.050	2.029
3	35	offset + 1 + t	AR(2) + seas(∼1, S=1)		State	7	53831	0.455	0.192	0.053	2.230
4	36	offset + 1 + seas(∼1, S=1)	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	1	10	51701	0.437	0.085	0.056	1.961
4	37	offset + 1	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	1	8	51673	0.437	0.194	0.055	1.969
4	38	offset + 1 + t + logpopdens	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	1	9	51691	0.437	0.153	0.056	1.962
4	39	offset + 1 + t	AR(2) + t	NE(1) + seas(∼1, S = 1)	1	8	51670	0.439	0.001	0.059	1.865
4	40	offset + 1 + t	AR(2) + seas(∼1, S=2)	NE(1) + seas(∼1, S = 1)	1	11	51545	0.437	0.115	0.055	1.973
4	41	offset + 1 + t	AR(2) + seas(∼1 + t, S=2)	NE(1) + seas(∼1 + t, S = 1)	1	15	51446	0.441	0.563	0.054	1.959
4	**42**	**offset + 1**	**AR(4) + seas(∼1, S=1)**	**NE(1) + seas(∼1, S = 1)**	**1**	**8**	**50323**	**0.420**	**0.346**	**0.054**	**1.872**
4	43	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(1) + t	1	8	51749	0.437	0.545	0.053	2.003
4	44	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼logpopdens, S = 1)	1	9	51780	0.438	0.202	0.056	1.975
4	45	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(3)	1	8	51642	0.437	0.333	0.053	1.972
4	46	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	State	10	51676	0.437	0.118	0.055	1.964
4	47	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼logpopdens + t, S = 1)	State	11	51782	0.438	0.314	0.055	1.988
5	48	offset + 1 + seas(∼1, S=1)	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	1	10	50342	0.420	0.297	0.055	1.867
5	49	offset + 1	AR(4) + seas(∼1 + t, S=1)	NE(1) + seas(∼1 + t, S = 1)	1	10	50296	0.424	0.614	0.052	1.864
5	50	offset + 1 + logpopdens	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	1	9	50332	0.420	0.439	0.054	1.870
5	51	offset + 1	AR(4) + t	NE(1) + seas(∼1, S = 1)	1	7	50336	0.424	0.000	0.060	1.763
5	**52**	**offset + 1**	**AR(4) + seas(∼1, S=2)**	**NE(1) + seas(∼1)**	**1**	**10**	**50164**	**0.419**	**0.194**	**0.055**	**1.868**
5	53	offset + 1	AR(4) + seas(∼1 + t, S=1)	NE(1) + seas(∼1 + t, S = 2)	1	14	50097	0.423	0.782	0.052	1.851
5	54	offset + 1	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼1 + t, S = 1)	1	9	50324	0.420	0.620	0.052	1.904
5	55	offset + 1	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼logpopdens, S = 1)	1	8	50401	0.421	0.425	0.055	1.873
5	56	offset + 1	AR(4) + seas(∼1, S=1)	NE(1)	1	6	50416	0.420	0.251	0.054	1.877
5	57	offset + 1	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	State	9	50325	0.420	0.342	0.054	1.873
5	58	offset + 1	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼logpopdens + t, S = 1)	State	10	50405	0.421	0.537	0.055	1.876

Figures

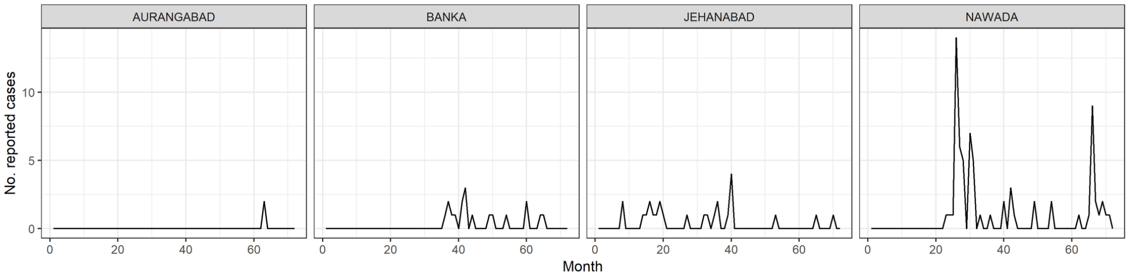


Figure A.1: Districts with unusual incidence patterns resulting in inflated dispersion estimates.

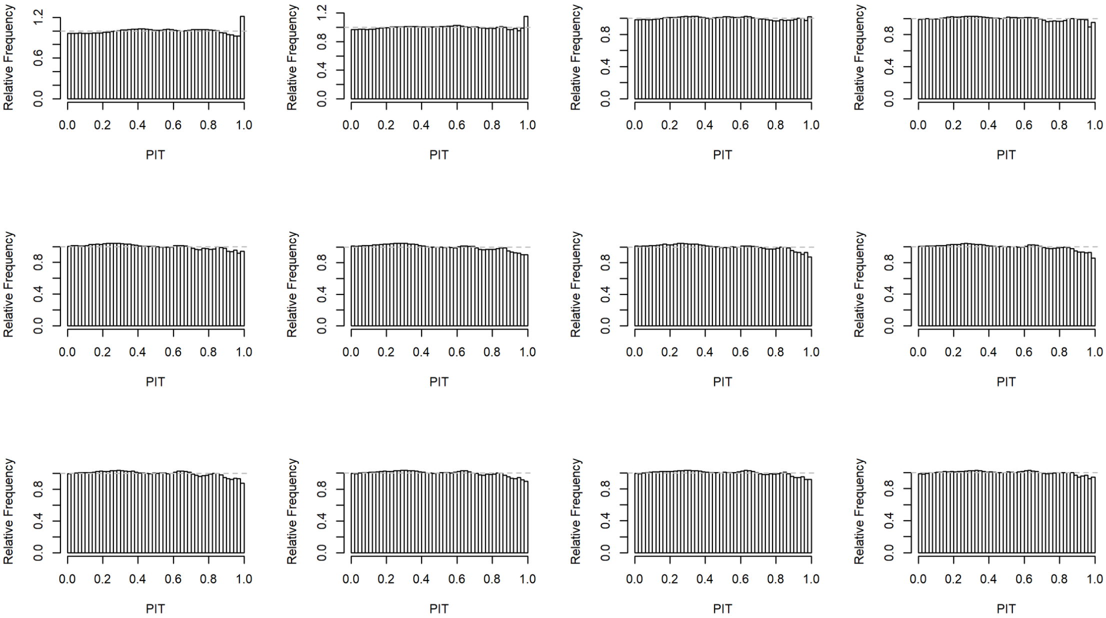


Figure A.2: Probability integral transform (PIT) histograms for models with increasing orders of geometric lags from 1 to 12 months (left to right, top to bottom) in the auto-regressive component. The final model selection process considered up to four lags.

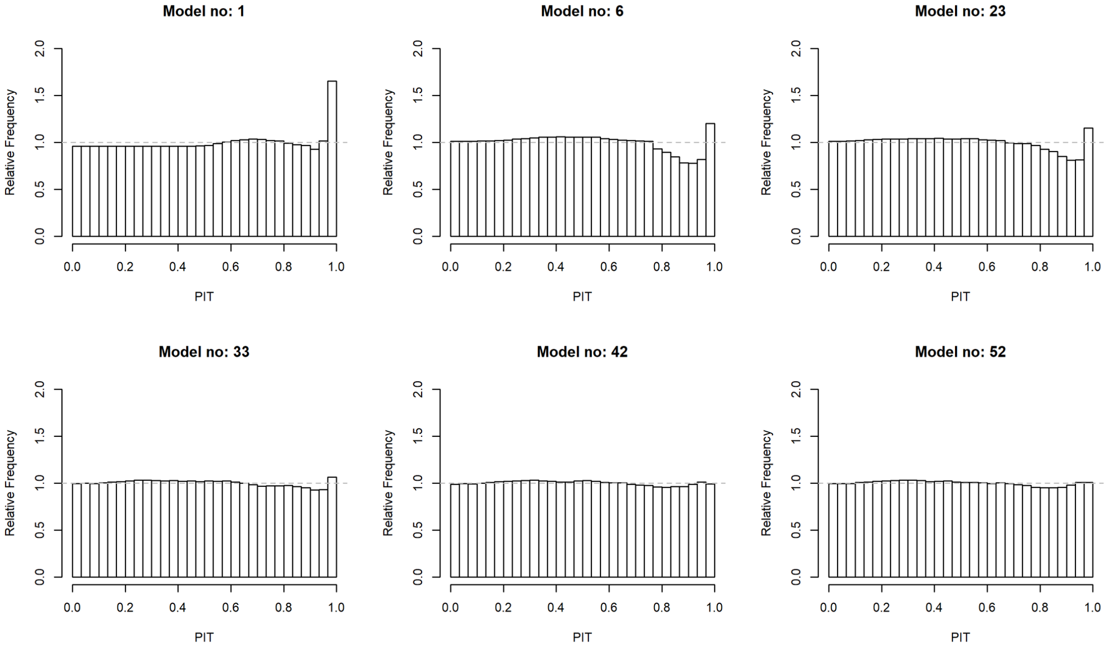


Figure A.3: PIT histograms for selected models at each stage. Model 42 is the final model. Model 52 offered minor improvement in RPS with additional complexity.

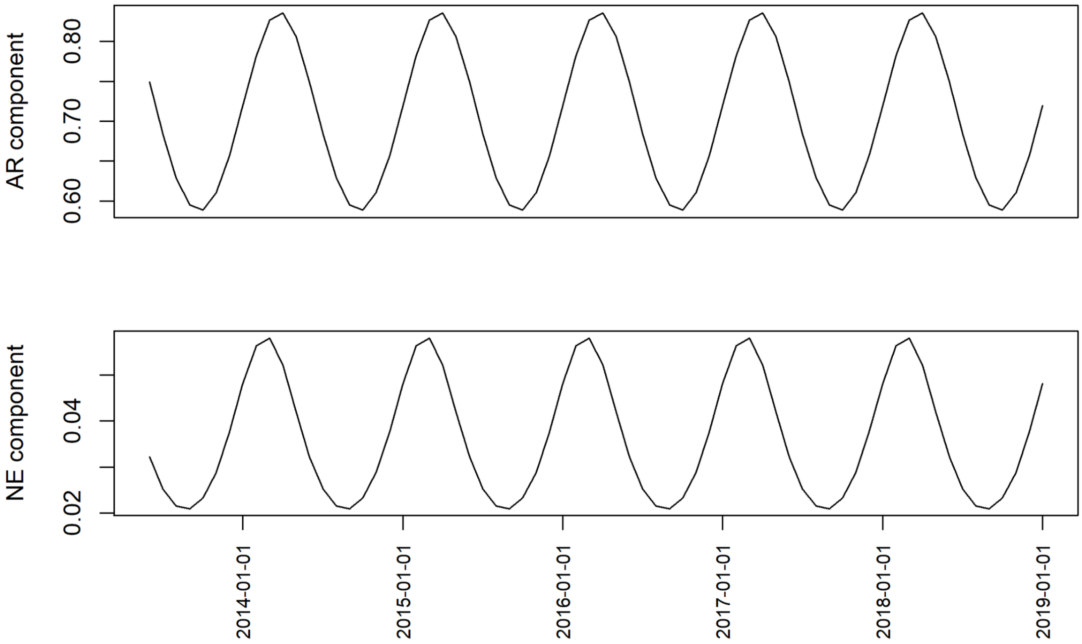


Figure A.4: Fitted seasonal waves in auto-regressive (AR) and neighbourhood (NE) model components. Both reflect the first-quarter peak in reported cases but the magnitude of the waves differs, with the contribution of the AR component varying more than that of the NE.

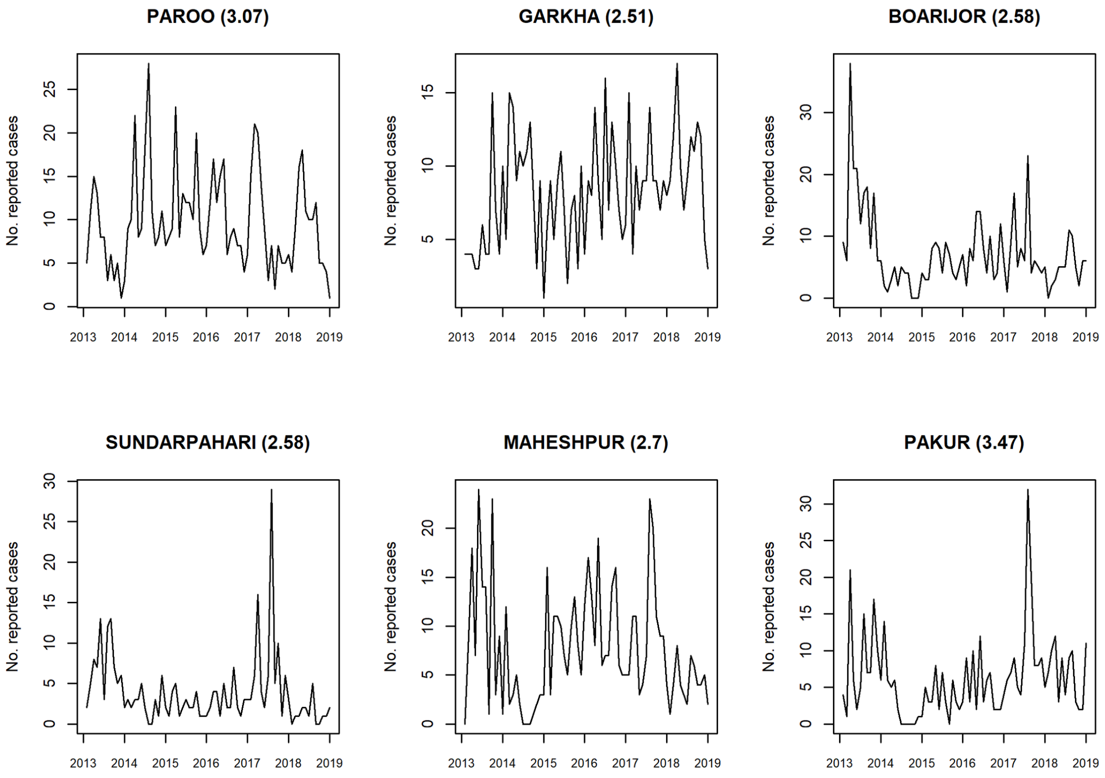


Figure A.5: Blocks with average RPS greater than 2.5 over the test period (Jan 2017—Dec 2018).

B SPEAK India Toolkit: Short-term prediction of block-level VL diagnoses for routine surveillance [Draft]

Short-term prediction of block-level VL diagnoses for routine surveillance

Emily S Nightingale

2022-12-01

Introduction

The WHO target of elimination of VL as a public health problem was the motivation for developing this tool. The aim was to explore the feasibility of projecting recent trends in incidence forward at short time horizons in order to better monitor the progress of each block towards this target.

Routine surveillance monitors incidence rates per block in order to identify high-, moderate- or low-endemicity and hence determine intervention plans. Block-level incidence was also used to monitor regional elimination status with respect to the number of blocks above or below the target. This approach does not take into account that risk is shared across administrative boundaries and, as such, a block defined as low- or non-endemic does not imply that it is at low risk of transmission or reintroduction in the future.

This prediction tool employs an existing statistical framework [1] to model the dependence of monthly, block-level incidence rates on the recent past, both within the same block and across its neighbours.

Use of the tool

This document serves to explain the analysis pipeline written to obtain short-term predictions of visceral leishmaniasis diagnoses at the block level in Bihar, India, employing the model described in [2]. The purpose of this work was to estimate potential incidence over the next 3-4 months based on historical incidence patterns over time and between neighbouring blocks. These may be used to monitor progress towards elimination goals and potentially to identify blocks for closer attention, in which patterns diverge from the regional trends.

All code has been written and tested in R version 3.6.3 (2020-02-29) and requires the following packages: *tidyverse, lubridate, reshape2, surveillance, hhh4addon, rgdal, spdep, here, sf, tictoc.* The code can be accessed through a public repository at https://github.com/esnightingale/vl-short-term-prediction.

Data

Raw linelist data are currently downloaded manually from the KA-MIS web portal (https://ka-mis.org/) via a secure login provided by CARE India.

The dataset of interest is the state-level resident diagnosis table, consisting of individual diagnosis records for both VL and PKDL with date and block of residence. To ensure all relevant records are included, data are downloaded without filter and the time span restricted within a later data cleaning step.

This data table currently may only be downloaded by state, so to include multiple states in the analysis requires downloading the separate files and appending before running the data aggregation step.

The raw data are read in, cleaned, split by case type (VL or PKDL) and aggregated by block and month of diagnosis.

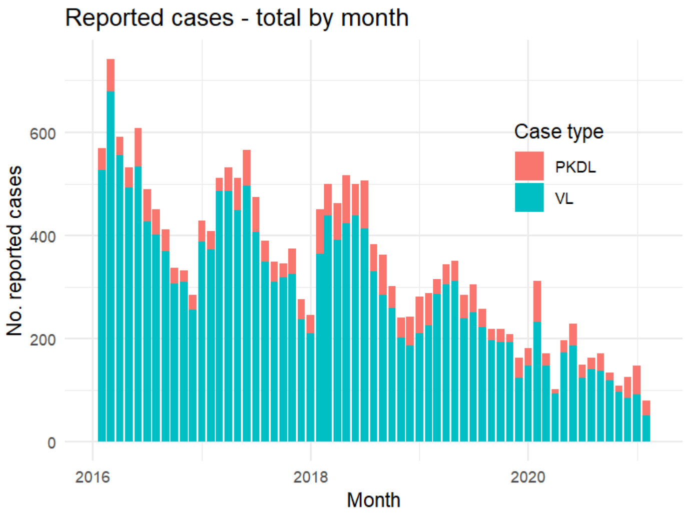


Figure 1: Monthly new reported diagnoses for VL and PKDL across the state of Bihar, from January 2016 to February 2021.

Pipeline

The pipeline consists of the following steps:

1. **Set up populations** – Read block population estimates from the most recent census (2011) and project forwards per month according to estimated decadal growth rates.
2. **Clean and aggregate linelist** – Read in raw data, clean and filter to specified period, then aggregate by block and month for analysis.
3. **Set up data formats for HHH4 model** – The HHH4 model framework requires a certain structure of input data (defined as a “spatial time series” object or *stsObj*), consisting of matrices of case and population counts, a shapefile and neighbourhood matrix for the spatial units, start date and frequency of observations.
4. **Fit the model and predict ahead** – Predictions are obtained by drawing simulations of the trajectory *n* months beyond the fitted data, on a rolling basis. A test period is also specified (currently set to the most recent six months) with which to check predictive accuracy.
5. **Summarise and plot the predictions** – Calculate quantiles over simulated trajectories, by block, state and month, and plot these prediction intervals.

These steps are implemented in sequence via the script *run_all.R.* Parameters which can be adjusted are:

- Start and end dates between which to aggregate data and fit model
- Number of months to include as a test period for prediction
- Number of months to forecast ahead of observed data
- Number of simulations to draw

If the specified end date is beyond the observed range of diagnosis dates in the raw data, the end date is redefined from the last observed diagnosis date rounded down to the start of the last month (in order to exclude diagnoses from the latest, incompletely reported month).

Outputs

The following are example outputs based on a model fit to diagnoses from 2016-01-01 to 2021-02-01, with rolling 3-month-ahead predictions.

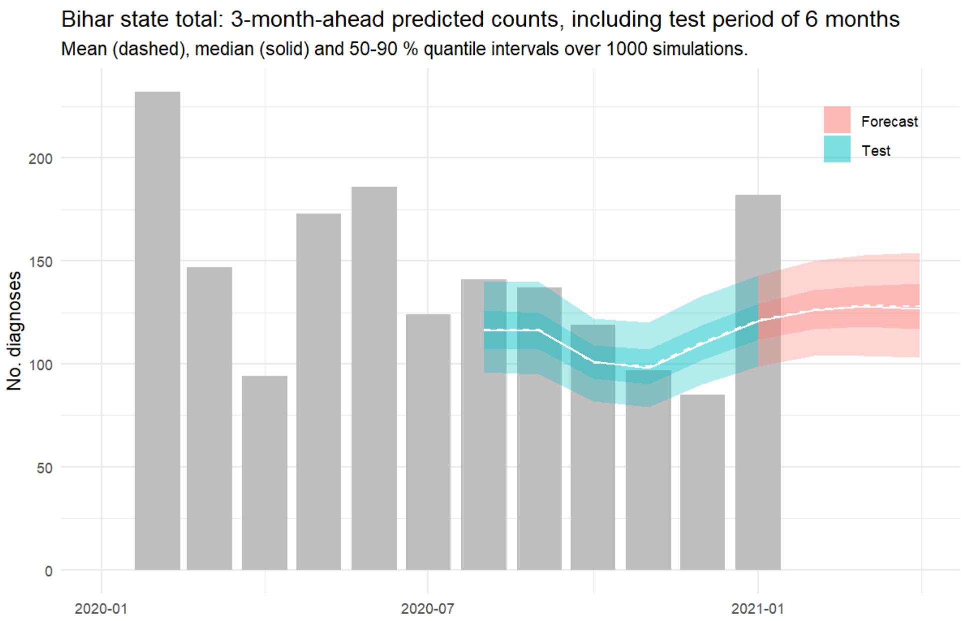


Figure 2: State total counts with three-month-ahead predictions based on 1000 simulations from the model, for a test period of six months and a forecast period of three months. For each predicted month, the model is refit to include the latest observed data.

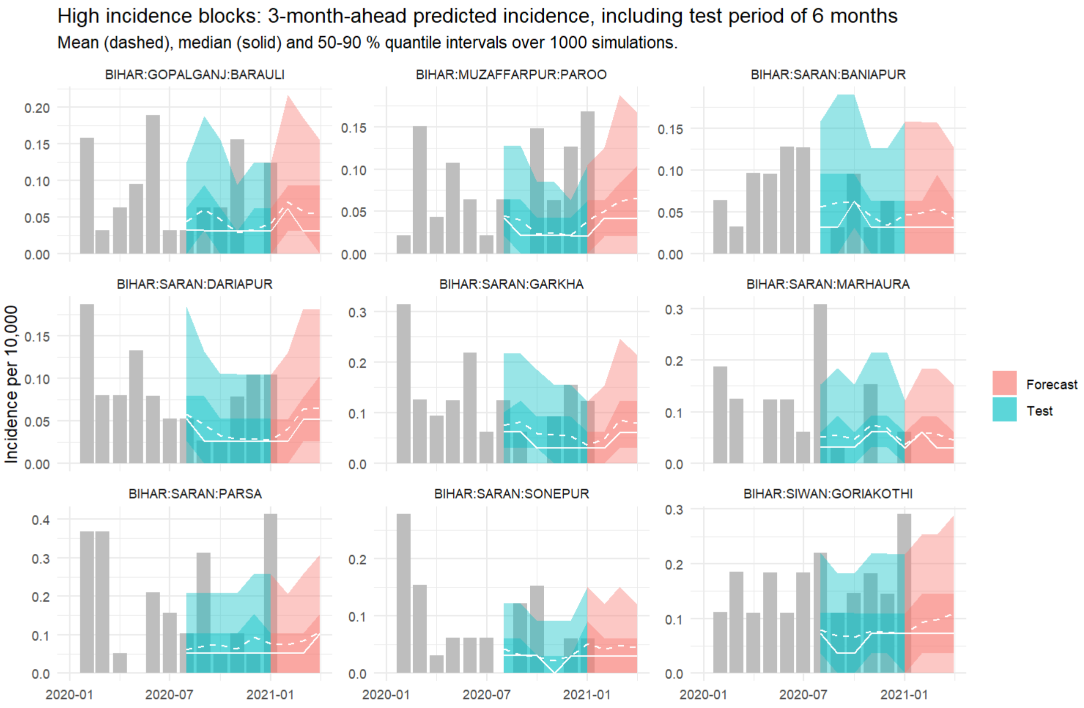


Figure 3: Equivalent predictions at the block level, for blocks with highest total incidence since 2019.

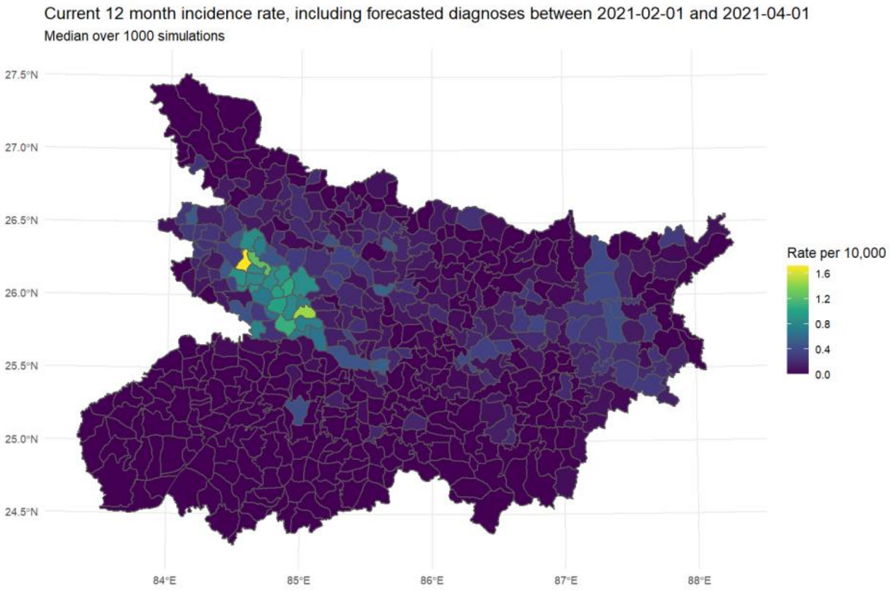

Current 12 month incidence rate, including forecasted diagnoses between 2021-02-01 and 2021-04-01
Median over 1000 simulations
27.5°N
27.0°N
26.5°N
26.0°N
25.5°N
25.0°N
24.5°N
84°E
85°E
86°E
87°E
88°E
Rate per 10,000
1.6
1.2
0.8
0.4
0.0

Figure 4: Estimated block-level incidence rates for the most recent twelve months, including the three month forecasts just obtained.

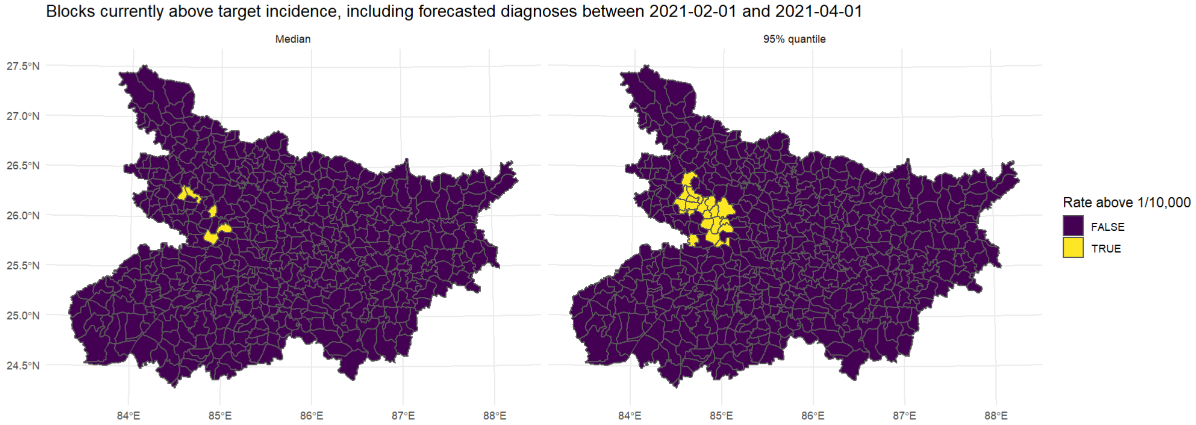


Figure 5: Blocks above the 1 per 10,000 target incidence rate according to estimated 12 month incidence, based on median and upper 95% quantile of simulations.

Other applications

Provided that relevant administrative boundary and population data were available alongside observed case counts, the tool may be adapted to other settings (with respect to geographic region and/or disease). However, two key limitations of the described approach are, firstly, that predictions at this administrative level will have less practical use as case counts fall to very low levels, and secondly, that the predictions only reflect patterns of *reported* cases and not of underlying incidence or transmission. These limitations should be taken into account when considering applying the tool in new contexts.

References

[1] S. Meyer, L. Held and M. Höhle, "Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance," *Journal of Statistical Software,* vol. 77, no. 11, 2017.

[2] E. S. Nightingale, L. A. C. Chapman, S. Srikantiah, P. Jambulingam, J. Bracher, M. M. Cameron and G. F. Medley, "A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India," *PLOS Neglected Tropical Diseases,* vol. 14, no. 7, 2020.

C Supplementary Materials: Inferring the distribution of visceral leishmaniasis incidence from data at different spatial scales

Figures

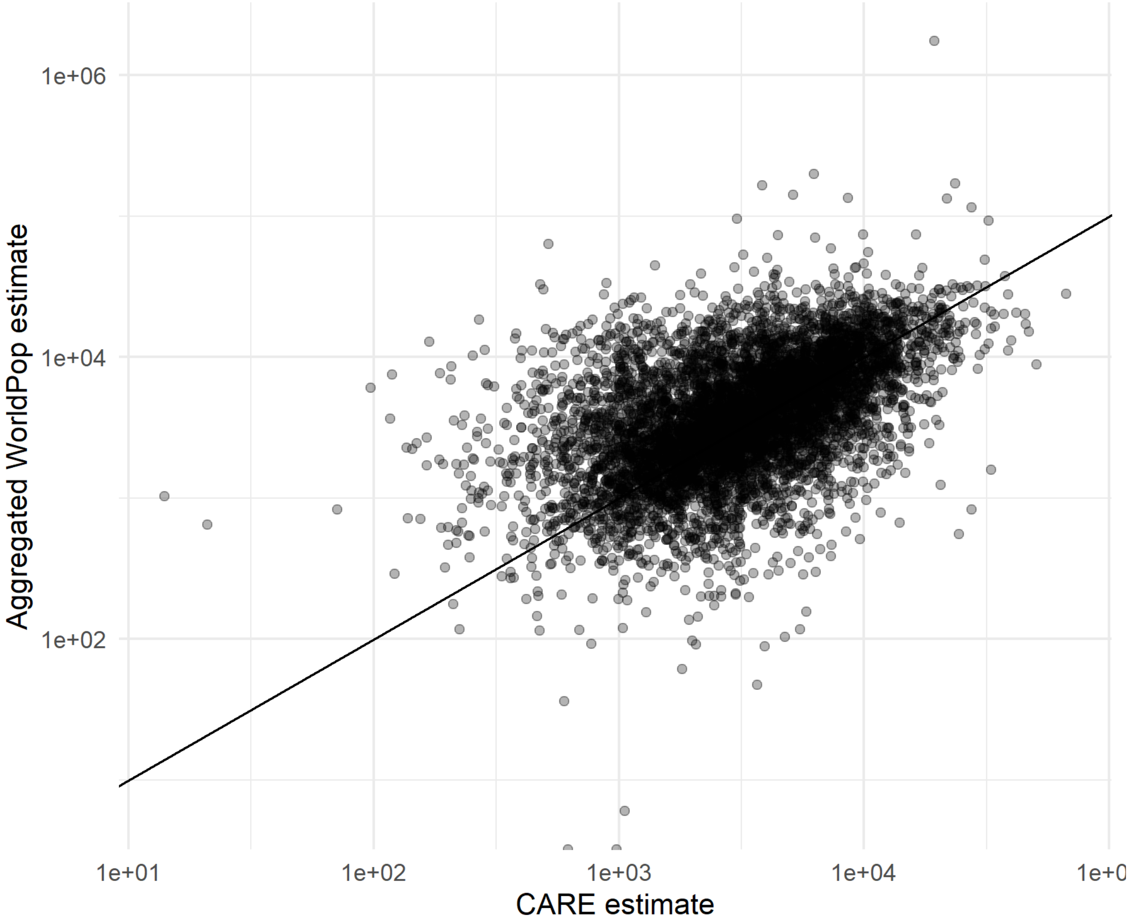


Figure C.1: Village population estimates obtained by CARE field teams through routine surveillance versus estimates for the same villages aggregated from WorldPop raster data.

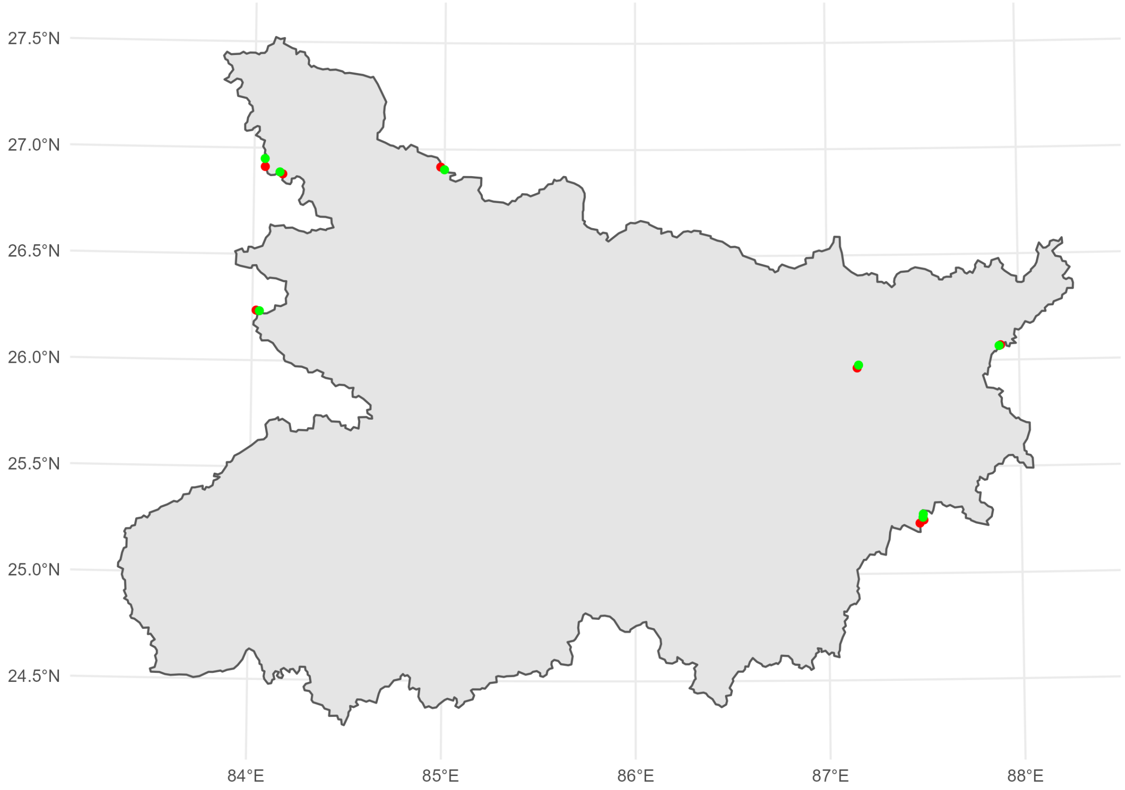


Figure C.2: Village GPS locations which did not initially fall within a village polygon. Original locations are shown in red with snapped locations in green.

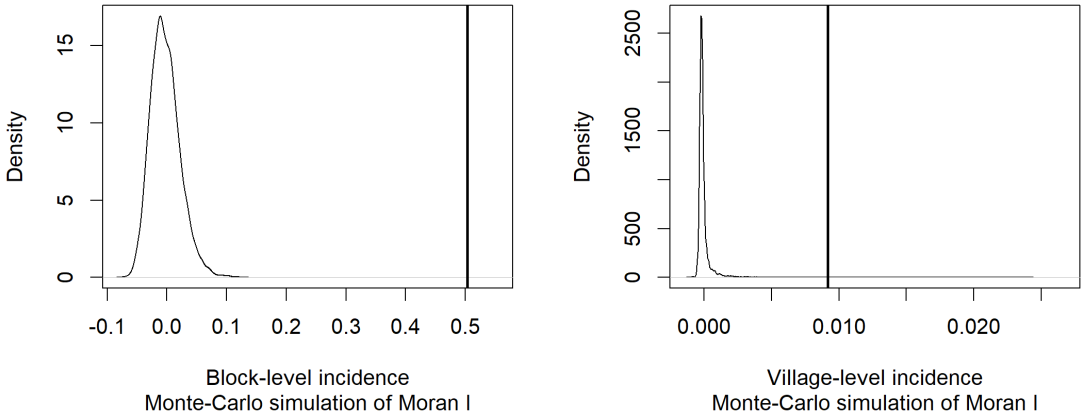


Figure C.3: Evaluation of Moran's I statistic for VL incidence at the block (left) and village (right) level. The solid vertical line illustrates the observed value, alongside the distribution of 999 values simulated under the assumption of complete spatial randomness.

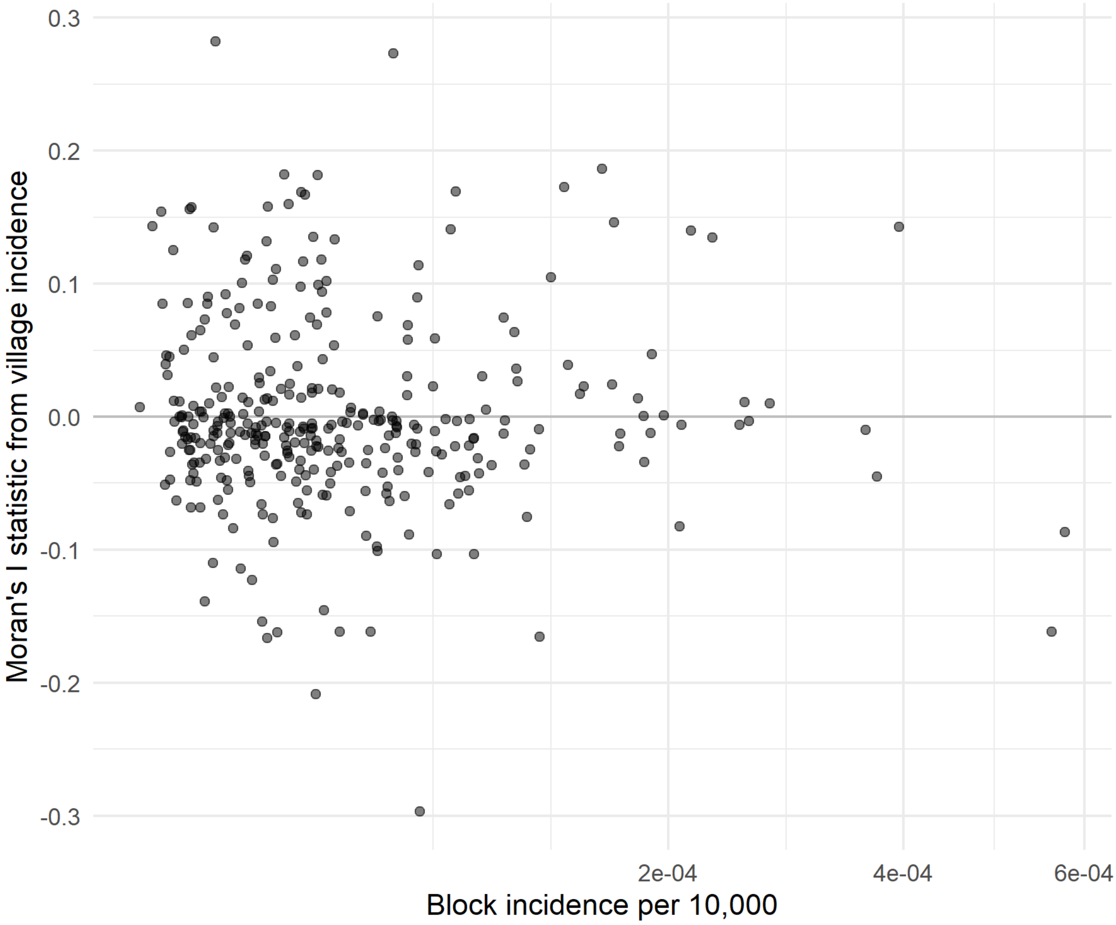


Figure C.4: Moran's I statistic calculated across constituent villages of each non-zero incidence block, and plotted against the block's overall incidence rate. There does not appear to be any trend between the magnitude of the statistic to the level of block endemicity.

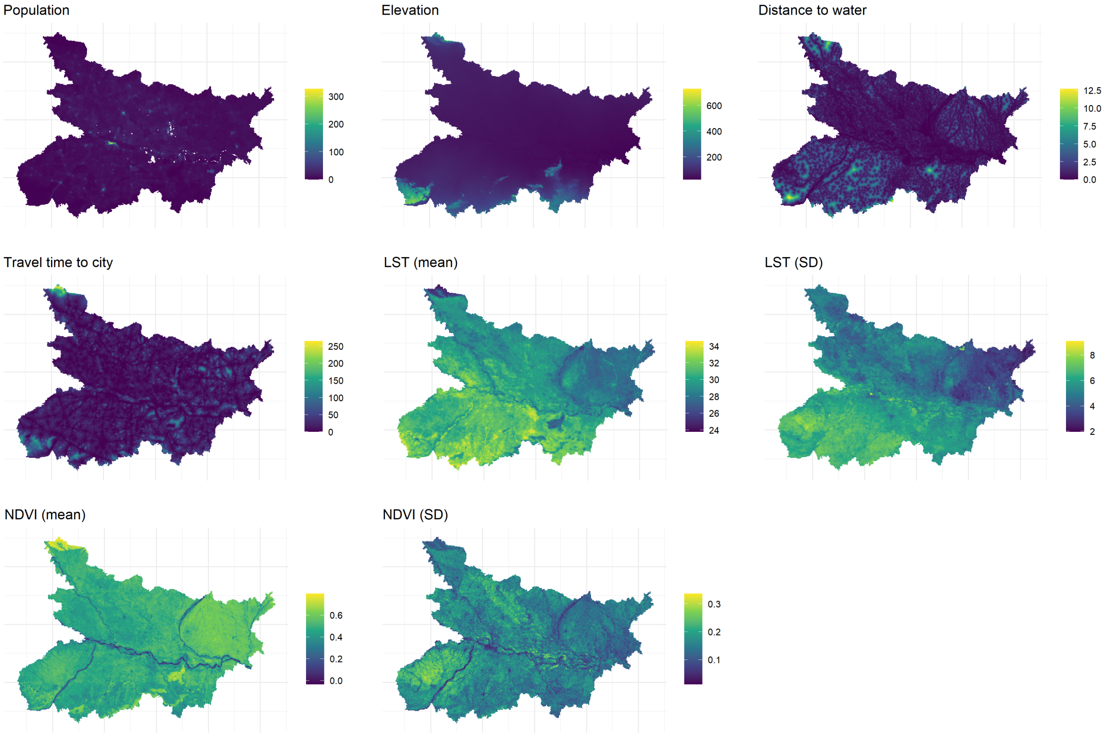


Figure C.5: Population and covariate raster data included in the disaggregation model (resolution 1km).

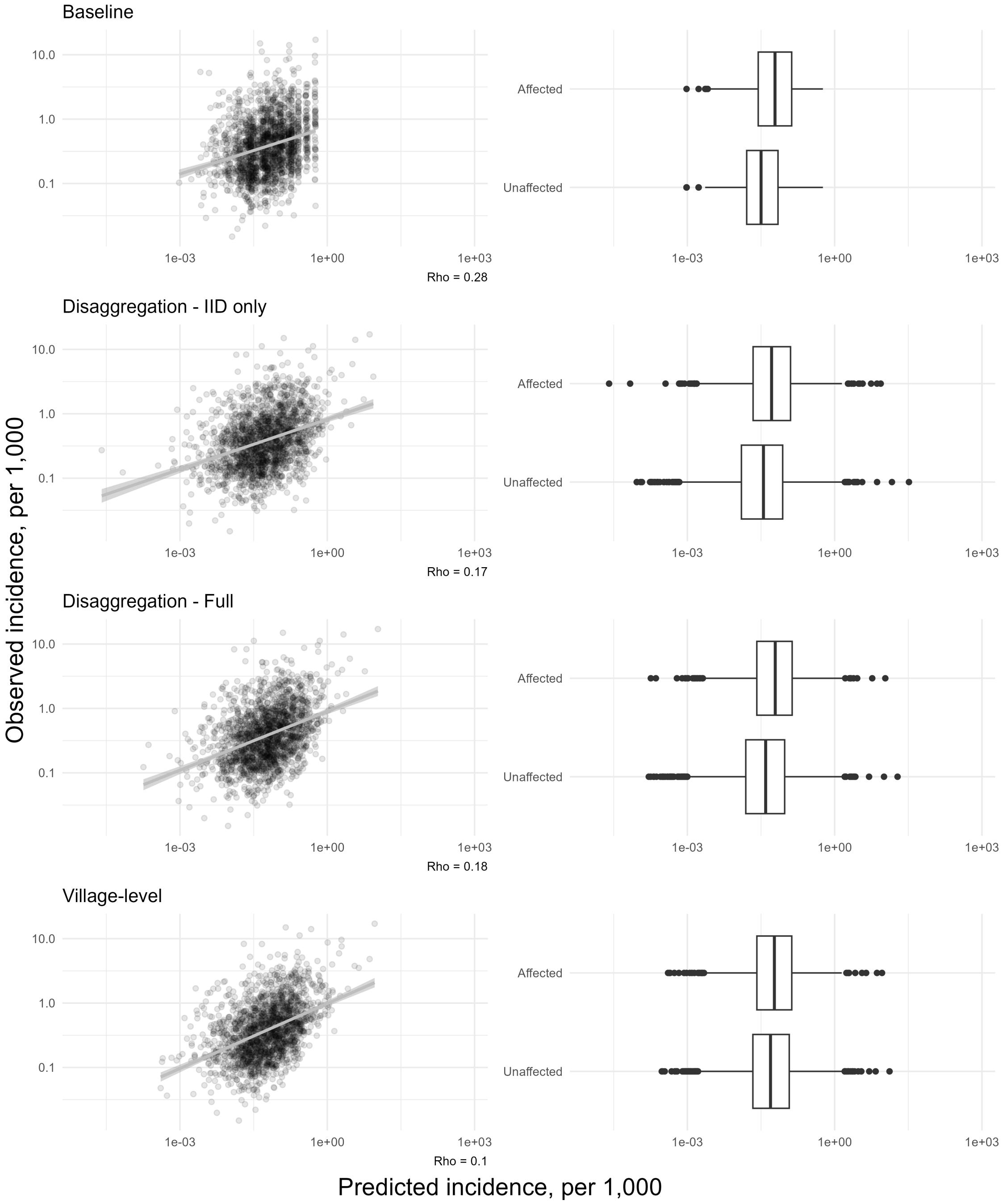


Figure C.6: Comparison of predicted to observed village incidence rates, relative to locally-informed population estimates from CARE India as opposed to aggregation of WorldPop estimates to village polygons. Scatter plots only include affected villages, with non-zero observed and predicted incidence. Box plots include only villages with non-zero predicted incidence, to allow log transformation of the x-axis (note the different x-axis scales between models).

D Supplementary Materials: Spatial variation in diagnosis delay for visceral leishmaniasis in Bihar, India

Supplementary Materials

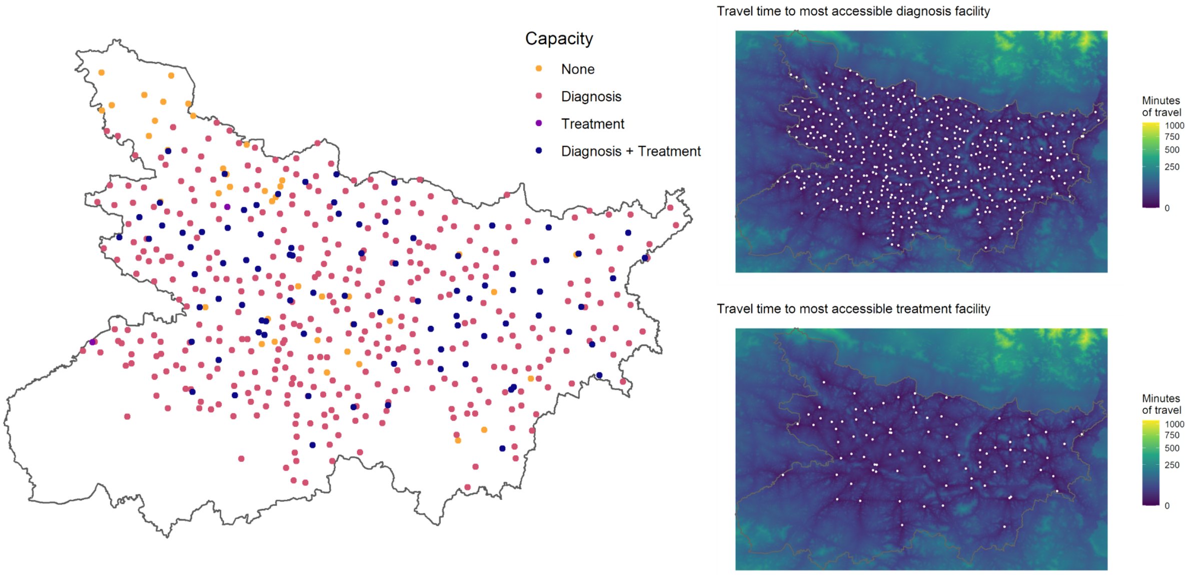
A Mesh construction

A triangular mesh on which to base the spatial SPDE model was constructed such that the distance between nodes was between 2km (the average distance between nearest-neighbour affected villages) and 10km.

B Cross-validation

Fifty iterations of spatial and non-spatial cross-validation were performed, to assess the contribution of the random field to prediction. For the former, all observations within a 50km radius of a randomly sampled point were withheld from model fitting and delay for the sampled point then predicted. For the latter, only the sampled point was withheld and then predicted. A cross-validated logarithmic score (logs) (19) was calculated across all fifty test observations, summarising the log posterior density at the observed value. Classification of delays greater than 30 days was also assessed via the Brier score (20), defined as the mean squared difference between the posterior probability of delay exceeding 30 days (the *exceedance probability*) and the observed (binary) value.

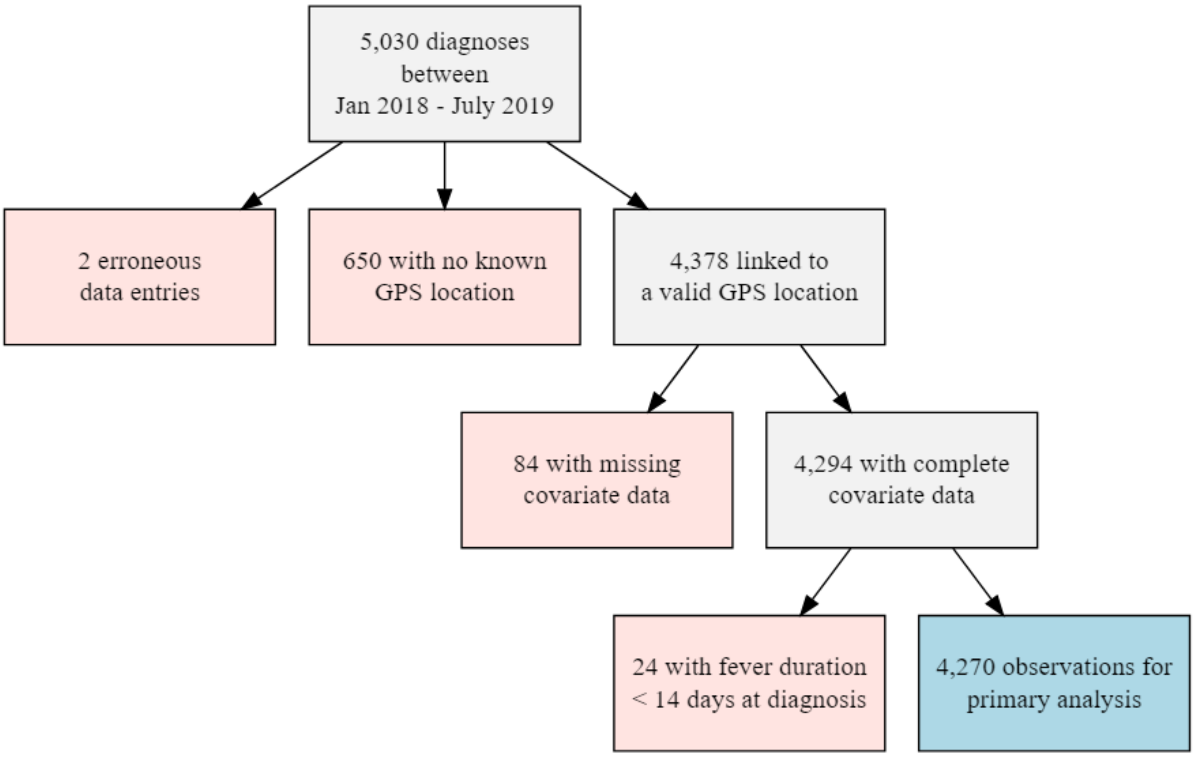
Figures



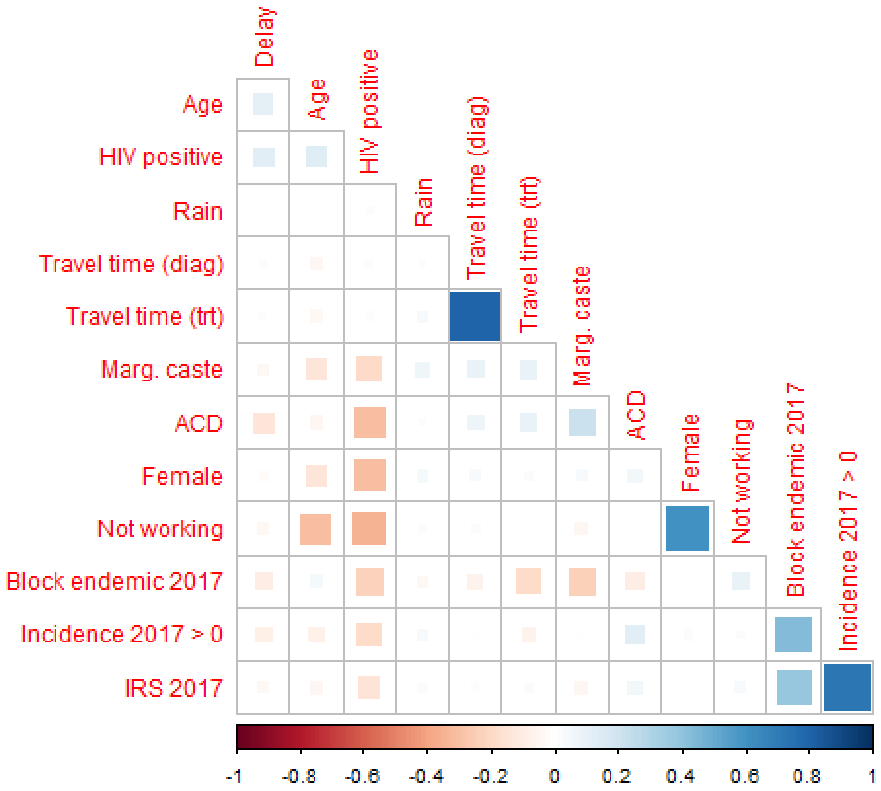
Travel time to most accessible diagnosis facility

Travel time to most accessible treatment facility

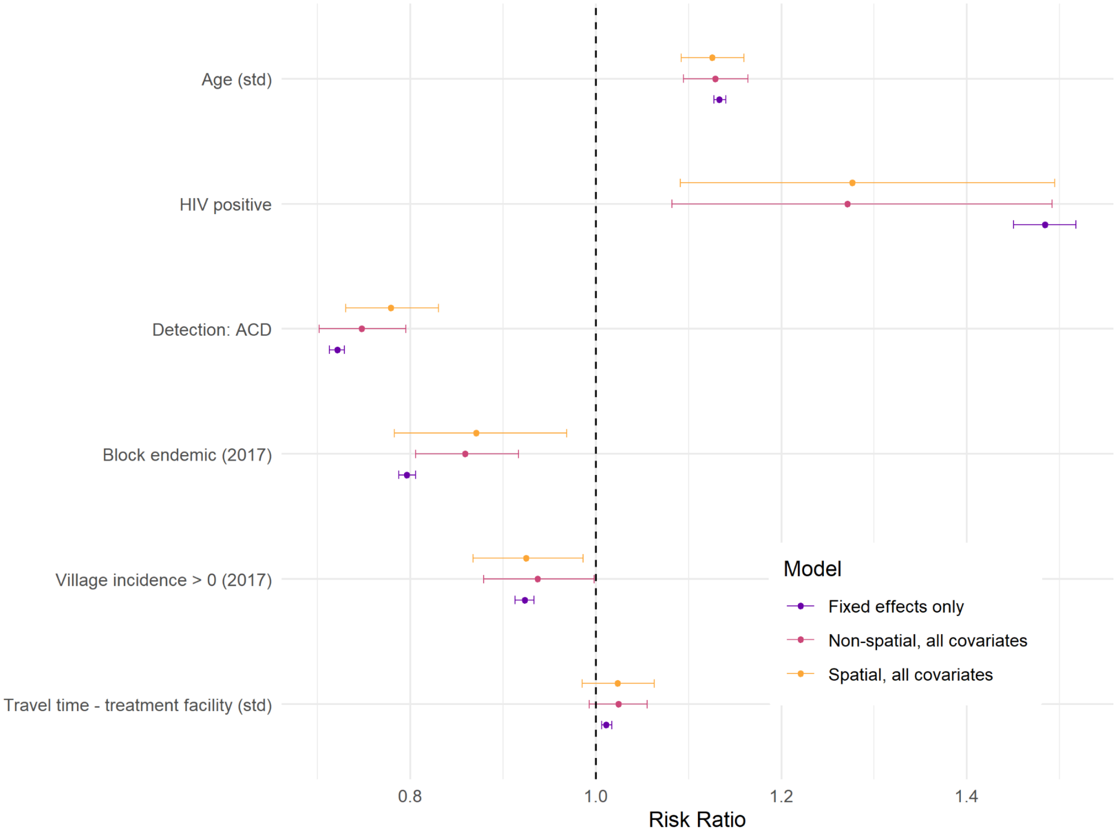
Supplementary figure S1: *Locations of health facilities in 33 endemic districts of Bihar, with capacity for diagnosis and/or treatment of VL. Minimum estimated travel time to one of these facilities from any point in the state is illustrated in the panels on the right.*


5,030 diagnoses between Jan 2018 - July 2019
2 erroneous data entries
650 with no known GPS location
4,378 linked to a valid GPS location
84 with missing covariate data
4,294 with complete covariate data
24 with fever duration < 14 days at diagnosis
4,270 observations for primary analysis

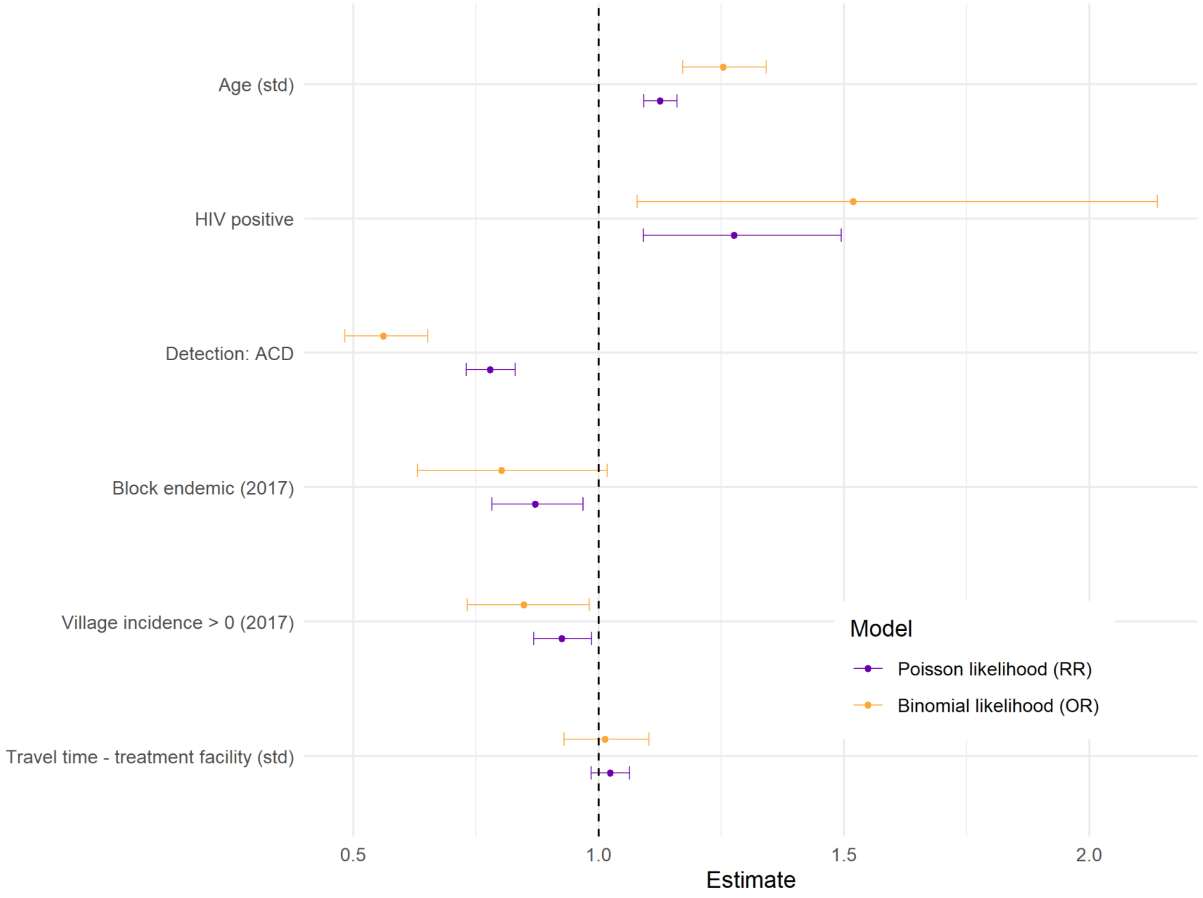
Supplementary figure S2: Flow chart of the data cleaning process, illustrating number of observations excluded under each criteria and the remaining observations included in the primary analysis.



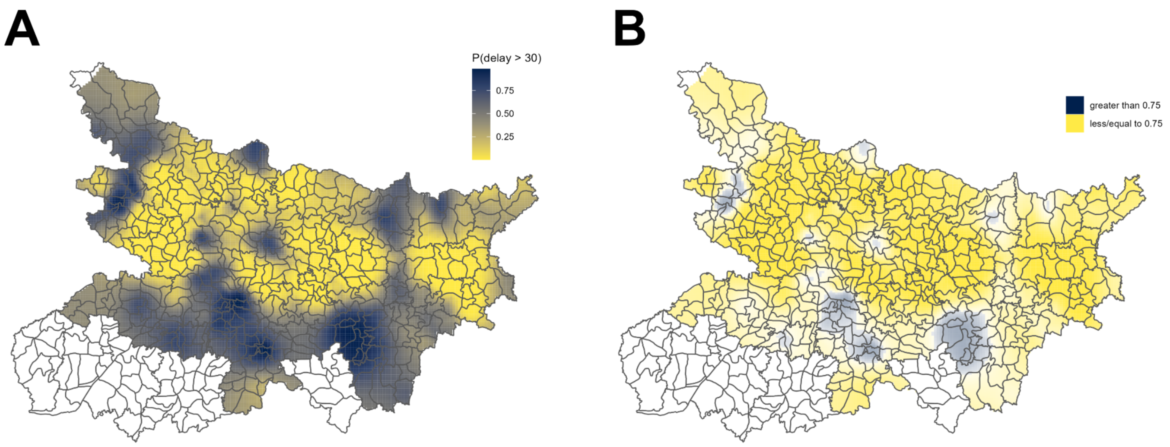
Supplementary figure S3: *Correlations between diagnosis delay and all covariates considered. The point-biserial method is used for correlation between binary and continuous variables, and the Pearson method otherwise. The size of the square for each combination indicates the strength of correlation, while the colour indicates both strength and direction. The strongest correlations are observed between travel time to diagnosis facility and travel time to treatment facility, as to be expected since several facilities provide both services, between village vector control and incidence in 2017, and between employment and sex.*



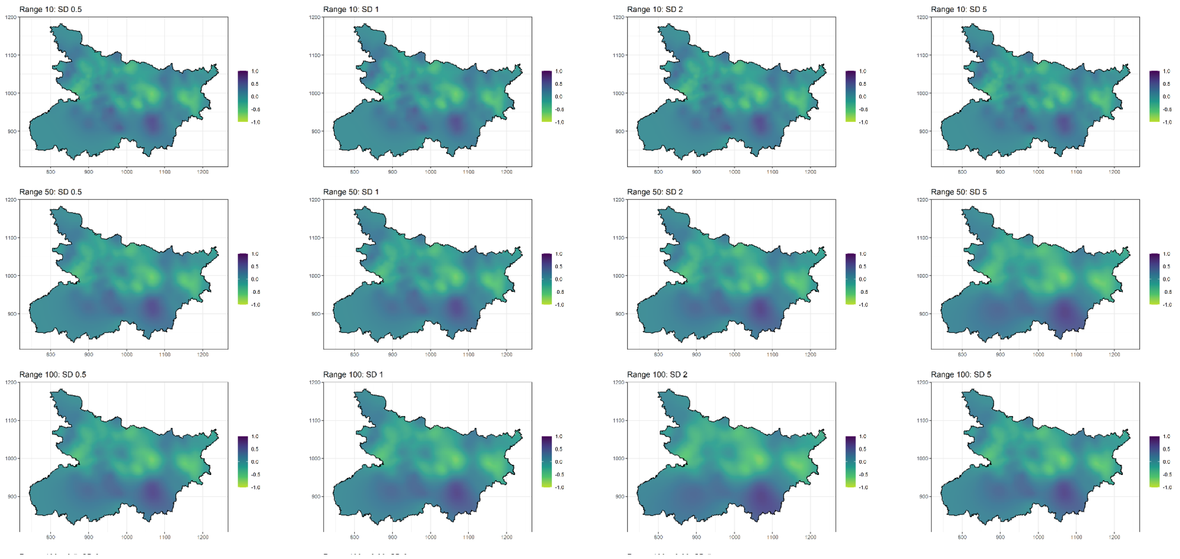
Supplemental figure S4: *Comparison of coefficient estimates between non-spatial and spatial models. Estimates are also shown from a model with only fixed effects. All significant covariates remain so with addition of the spatially-structured random effect. The estimated effect size for detection route and historical block endemicity are marginally reduced, while that of historical village incidence is increased.*



Supplementary figure S5: *Comparison of coefficient estimates (with 95% CrIs) between final model (Poisson likelihood) and an alternative using a binomial likelihood with a cut-off of 30 days' delay. It is important to note that estimates from the final model represent risk ratios while for the alternative binomial model they represent odds ratios. Therefore the magnitude of effects cannot be directly compared, only the direction.*



Supplementary figure S6: (A) *Predicted probabilities of delay exceeding 30 days and* ***(B)*** *an alternative version of Figure 3B with a higher cut-off of 0.75, to illustrate the impact of the choice of cut-off value. Delays greater than 30 days are only expected with high probability in limited regions of the south/south-east and north west.*



Supplementary figure S7: *Sensitivity analysis of the SPDE prior specification, comparing fitted spatial fields from the final model, with varying prior range and standard deviation.*

Tables

Supplementary table S1: *Summary of 736 observations out of 5,030 which were excluded from the dataset prior to analysis, due to either missingness of GPS location for the village or missingness in one or more of the individual-level covariates of interest.*

	Excluded	Included

N	736	4294
Delay, median [IQR]	16 [11-46]	16 [11-44]
Age, median [IQR]	28 [13-45]	25 [12-42]
Female, n (%)	285 (38.7)	1833 (42.7)
HIV positive, n (%)	241 (33.3)	1498 (34.9)
HIV missing, n (%)	12 (1.6)	0 (0)
Previous VL/PKDL treatment, n (%)	35 (5.2)	158 (3.7)
Previous treatment missing, n (%)	69 (9.4)	0 (0)
Marginalised caste, n (%)	59 (8.1)	377 (8.8)
Caste status missing, n (%)	8 (1.1)	0 (0)
Unemployed, n (%)	399 (55)	2525 (58.8)
Missing occupation, n (%)	11 (1.5)	0 (0)
Diagnosed through ACD, n (%)	277 (37.6)	1720 (40.1)

Supplementary table S2: *Summary of patient and village level characteristics of cases reporting less than or greater than 14 days of fever prior to diagnosis. Intervals presented are approximate 95% confidence intervals for mean or percentage.*

	Duration of fever before diagnosis	
	< 14 days	**>= 14 days**
N	24	4270
Age	16 [9-30]	25 [12-42]
Female, n(%)	4 (16.7)	1829 (42.8)
Marginalised caste, n(%)	6 (25)	1492 (34.9)
HIV positive, n (%)	0 (0)	158 (3.7)
Previous VL/PKDL treatment, n (%)	3 (12.5)	374 (8.8)
Unemployed, n (%)	19 (79.2)	2506 (58.7)
Diagnosed through ACD, n (%)	7 (29.2)	1713 (40.1)
Resident of village with non-zero VL incidence in 2017, n (%)	12 (50)	2341 (54.8)
Resident of village targeted for IRS in 2017, n (%)	16 (66.7)	3277 (76.7)
Resident of village in an block classed as endemic in 2017, n (%)	17 (70.8)	1938 (45.4)

Travel time to diagnostic facility, median [IQR]	9 [5-12]	12 [7-18]
Travel time to treatment facility, median [IQR]	16 [11-19]	18 [11-27]

Supplementary table S3: *Changes in magnitude of fitted random effects (non-spatial and spatial) with inclusion of each covariate domain.*

Included domain	Non-spatial effect (OLRE)		Spatial effect (SPDE)	
	Mean absolute value	% change	Mean absolute value	% change
None (*Model C*)	0.7005	-	0.2517	-
Patient (age, HIV, detection)	0.6889	-1.66	0.2374	-5.69
Awareness (block endemicity, village incidence)	0.6978	-0.39	0.2289	-9.06
Access (travel time to treatment facility)	0.7005	0.00	0.2495	-0.87

Supplementary table S4: *Summary of delays associated with active and passive case detection, by recent block endemicity. The relative gains from ACD compared to PCD appear greater in blocks which had not been recently classified as endemic.*

	Non-endemic 2017	**Endemic 2017**
No. blocks	290	44
Total population	76484757	9389809
No. cases detected	2332	1938
via ACD (%)	995 (42.7)	718 (37.0)
Mean delay - overall	34.4	27.0
via ACD	27.8	20.8
via PCD	39.5	30.7
Difference PCD-ACD (Std. err.)	11.7 (1.8)	9.8 (1.4)

Supplementary table S5: *Estimated total person-days of delay (median and 98% credible interval over 10,000 posterior samples) in the scenario of (a) complete ACD coverage, i.e. with all observations redefined as actively detected and (b) no ACD coverage, i.e. with all observations redefined as passively detected. The impact of the detection scenario is summarised with respect to the change in total person-days of delay relative to the original fitted values, as an overall total and split by the endemicity of the block.*

Scenario	Measure	Total	Non-endemic 2017	Endemic 2017
Observed ACD coverage *(Baseline)*	No. blocks	334	290	44
	No. detected cases	4270	2332	1938
	Via ACD (%)	1713 (40.1)	995 (42.7)	718 (37.0)
	Expected total person-days delay	134 631 [133 760, 135 495]	81 345 [80 675, 82 000]	53 283 [52 739, 53 829]
	Per case *(ACD + PCD)*	31.5 [31.3, 31.7]	34.9 [34.6, 35.2]	27.5 [27.2, 27.8]
Modelled complete (100%) ACD coverage	Expected total person-days delay -	114 793 [109 454, 120 650]	69 772 [66 603, 73 230]	45 042 [42 779, 47 478]
	Per case *(originally ACD + PCD)*	26.9 [25.6, 28.3]	29.9 [28.6, 31.4]	23.2 [22.1, 24.5]
	Change from baseline	-19 811 [-25 180, -14 118]	-11 575 [-14 693, -8 256]	-8 234 [-10 479, -5 851]
	Per case *(originally ACD + PCD)*	-4.6 [-5.9, -3.3]	-5 [-6.3, -3.5]	-4.2 [-5.4, -3]
	Per reassigned case *(originally PCD only)*	-7.7 [-9.8, -5.5]	-8.7 [-11, -6.2]	-6.7 [-8.6, -4.8]
Modelled no (0%) ACD coverage	Expected total person-days delay	146 645 [142 519, 151 121]	89 114 [86 421, 92 047]	57 530 [55 974, 59 195]
	Per case *(originally ACD + PCD)*	34.3 [33.4, 35.4]	38.2 [37.1, 39.5]	29.7 [28.9, 30.5]
	Change from baseline	12 009 [7 942, 16 437]	7 761 [5 129, 10 612]	4 246 [2 812, 5 813]
	Per case *(originally ACD + PCD)*	2.8 [1.9, 3.8]	3.3 [2.2, 4.6]	2.2 [1.5, 3]
	Per reassigned case *(originally ACD only)*	7 [4.6, 9.6]	7.8 [5.2, 10.7]	5.9 [3.9, 8.1]

E Supplementary Materials: The local burden of disease during the first wave of the COVID-19 epidemic in England

Supplementary Materials

A Spatial aggregation

Four sub-regions of Buckinghamshire were aggregated in order to match most recent population estimates, since the LTLA was recently sub-divided. The City of London was combined with Westminster due to its very small resident population, and the Isles of Scilly were excluded since no COVID-19-related deaths had been reported there within this time period. Overall, reported deaths were attributed to 312 spatial units across England.

LTLAs can be classified into one of four geographical categories: London borough (10.3 % of total LTLAs), metropolitan district (11.5 %), non-metropolitan district (60.3 %) and unitary authority (17.9 %). The former two categories capture the major urban areas of the country (including Birmingham, Liverpool, Manchester, Sheffield, Leeds and Newcastle) with high connectivity both nationally and internationally, while the latter capture predominantly rural areas and smaller towns or cities.

B Age-adjusted expected deaths

Specifically, if N_m is the total observed deaths in age group m and P_m the estimated total population of England within the same age group, then r_m is defined as the total age-specific mortality rate

$$r_m = N_m / P_m$$

over the whole period. These rates are scaled down to estimated average rates per week by dividing by the number of observed weeks in the study period (~ 25). If P_{im} is the estimated population in age group m within local authority i, then the expected number of deaths per week, E_{im}, for age group m in LA i is calculated as

$$E_{im} = r_m * p_{im}$$

Finally, the expected deaths overall in LA i is

$$E_i = \sum_{m=1}^{M} E_{im}$$

where $M = 10$ denotes the total number of age groups. These expected values form a baseline which assumes all LTLAs exhibit the same age-specific mortality rates, and that these rates are constant over the observed period. We then conduct the analysis on the standardised mortality ratio, SMR, of observed deaths, per week and LA, over expected.

C Model Formulae

The overall structure of fitted models for number of deaths Y and expected count E is as follows:

$$Y_{itG} \sim NB(\mu_{itG} E_{itG}, \psi)$$
$$log(\mu_{itG}) = \beta_0 + \Sigma_{j=1}^{m} \beta_j z_{ij} + \gamma_{(t-t_0)G} + \delta_t + \zeta_i^S$$

for LTLA i in calendar week t, where $\Sigma_{j=1}^{m} \beta_j z_j$ denotes the contribution of fixed covariate effects, γ and δ the temporal random effects on epidemic week (denoted $t - t_{0i}$ for the week of first t_{0i} death) and calendar week respectively, and ζ the spatial random effect. *NB* reflects the chosen negative binomial likelihood.

The temporal random effects are defined with random walk (RW) correlation structures. A random walk of order one (RW1) assumes that the increments $\delta_t - \delta_{t-1}$ between each time step are Gaussian distributed with mean 0 and precision τ. A second order random walk (RW2) assumes the same of the second order increments $\delta_t - 2\delta_{t-1} - \delta_{t-2}$ and hence describes a smoother trend. Specifically, γ is modelled by a second-order random walk with precision τ_γ, fit either across all LTLAs or replicated by geography $G \in$ {*London borough*, *metropolitan district*, *non-metropolitan district*, *unitary authority*}. δ is modelled by a first-order random walk with precision τ_δ. ψ is the size parameter (1/overdispersion) for the negative binomial distribution.

Three candidate structures for the spatial random effect were considered. The index $S \in$ {*Null*, IID, *BYM*} indicates either no spatial model, the completely unstructured IID model or the Besag-York-Mollie spatially-structured model parameterised with precision τ_ζ and mixing parameter ϕ. These are defined as follows:

Null:

$$\zeta_i^{Null} = 0$$

IID:

$$\zeta_i^{IID} = u_i$$

BYM:

$$\zeta_i^{BYM} = \frac{1}{\sqrt(\tau_\zeta)}(\sqrt(1-\phi)v_i + \sqrt(\phi)u_i)$$

Priors

Gaussian priors with mean 0 and precision 0.1 were specified for the fixed covariate effects. Penalised complexity priors were specified for the precisions of the three structured random effects (temporal and spatial) such that $P(1/\sqrt(\tau) > U/0.31) = 0.01$, with the upper limit U defined as the standard deviation of residuals from the null, fixed-effect-only model, averaged over the relevant index (epidemic week, calendar week, LTLA). The BYM mixing parameter ϕ is also given a penalised complexity prior, such that $P(\phi > 0.5) = 2/3$.

D Rate of detection under symptomatic community testing

The ONS COVID-19 infection survey was piloted from April 2020, conducting PCR tests in samples of the population in order to estimate the prevalence of test-positive infections in the country over time. Estimates during this early period are presented as a percentage of the population who would test PCR-positive, by rolling fortnight, and were translated to an approximate weekly incidence by dividing by two, assuming test-positivity duration of one week and simple steady-state dynamics.

Assuming the population of England to be 56 million, the total weekly incidence of test-positives was calculated for weeks starting 18 May to 15 June 2020. The cumulative count of infections over this period was then compared to the cumulative count of confirmed cases to estimate the detection rate of infections under expanded surveillance.

Supplementary Figures

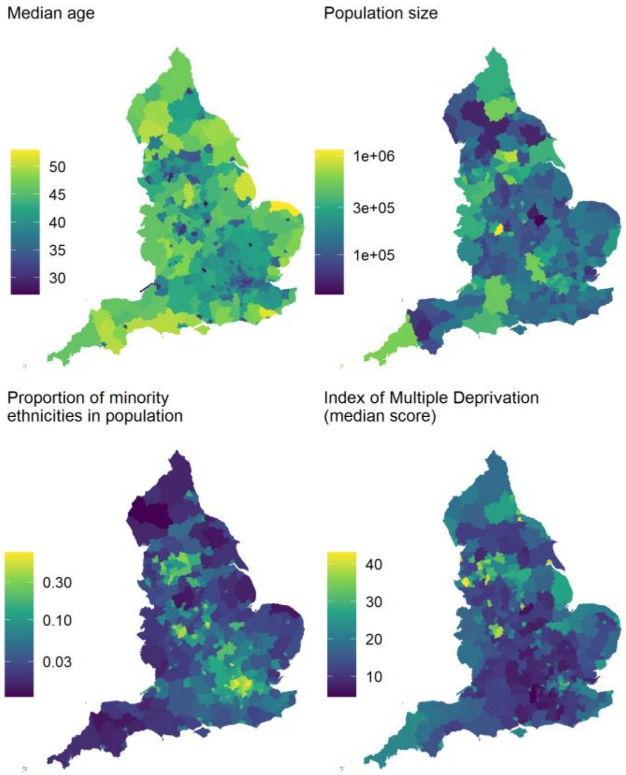

Median age
50
45
40
35
30
Population size
1e+06
3e+05
1e+05
Proportion of minority ethnicities in population
0.30
0.10
0.03
Index of Multiple Deprivation (median score)
40
30
20
10

Figure S1: Distribution of LTLA-level characteristics used in modelling of mortality risk. Younger age and greater minority proportion are characteristic of urban centres, whereas deprivation is more pronounced across northern LTLAs.

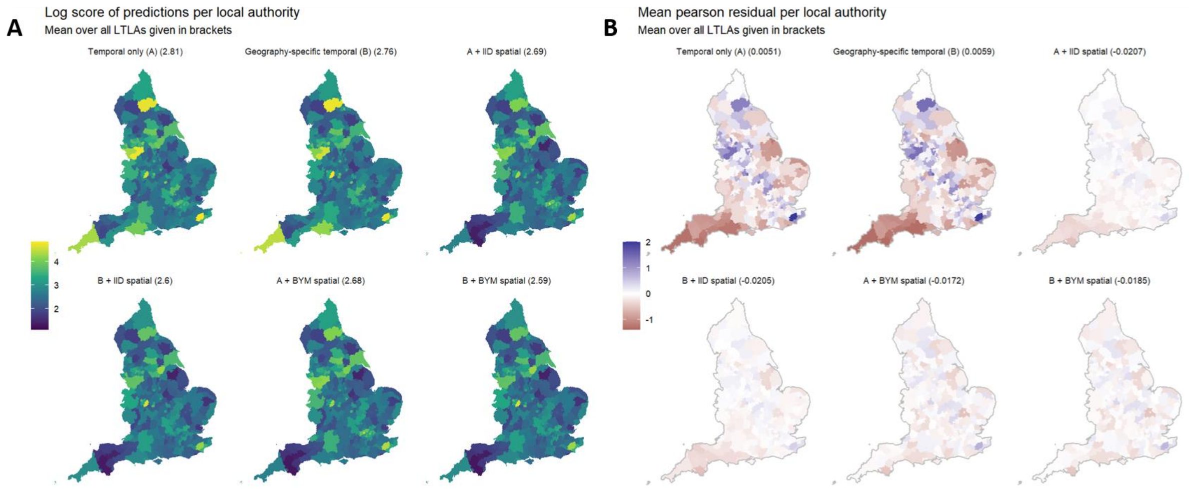

A
Log score of predictions per local authority
Mean over all LTLAs given in brackets
Temporal only (A) (2.81)
Geography-specific temporal (B) (2.76)
A + IID spatial (2.69)
B + IID spatial (2.6)
A + BYM spatial (2.68)
B + BYM spatial (2.59)
4
3
2
B
Mean pearson residual per local authority
Mean over all LTLAs given in brackets
Temporal only (A) (0.0051)
Geography-specific temporal (B) (0.0059)
A + IID spatial (-0.0207)
B + IID spatial (-0.0205)
A + BYM spatial (-0.0172)
B + BYM spatial (-0.0185)
2
1
0
-1

Figure S2: Averaged log scores (top) and Pearson residuals (bottom) from models fitted to weekly deaths per local authority which occurred between 2020-01-01 and 2020-06-30. Adding spatial random effects reduces the magnitude of error overall, with the conditional autoregressive structure from the BYM model providing the best cross-validated fit.

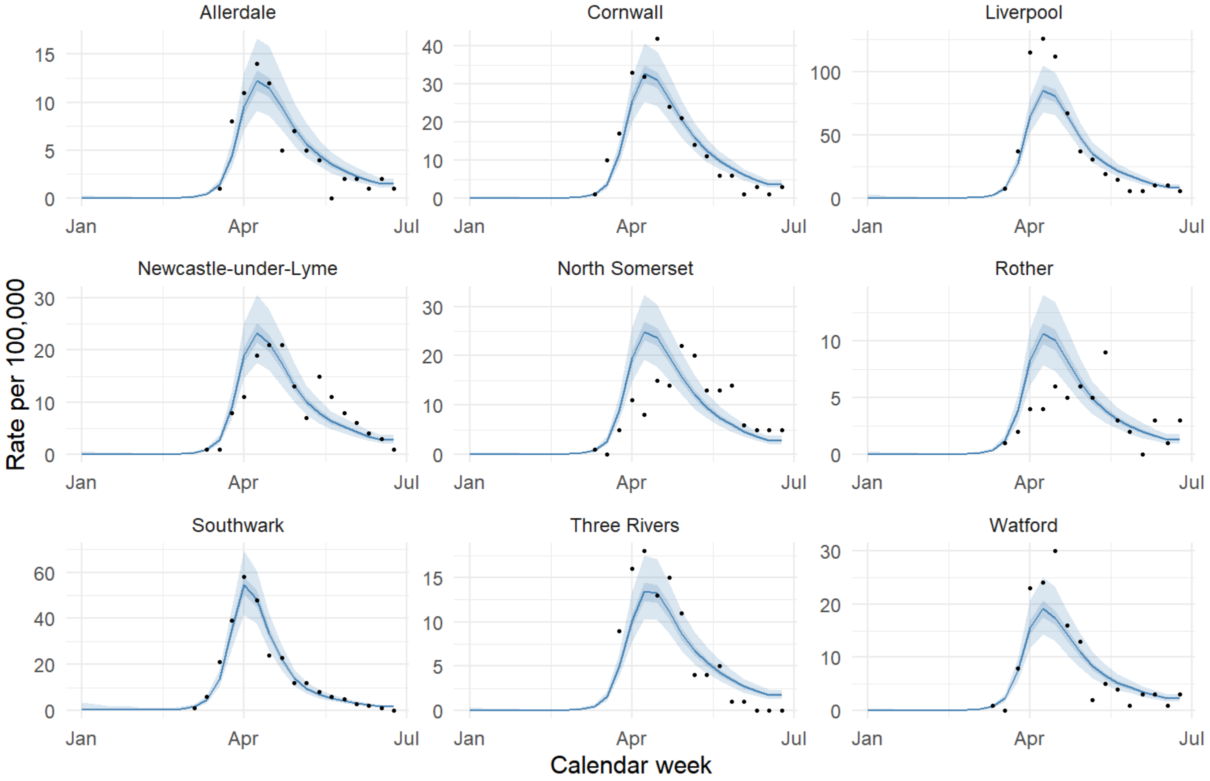


Figure S3: Fit of the selected model for nine randomly sampled LTLAs, over 1,000 posterior samples. For the LTLA fits, observed rates of COVID-19-related death per 100,000 are shown in black, with 50-98% credible intervals.

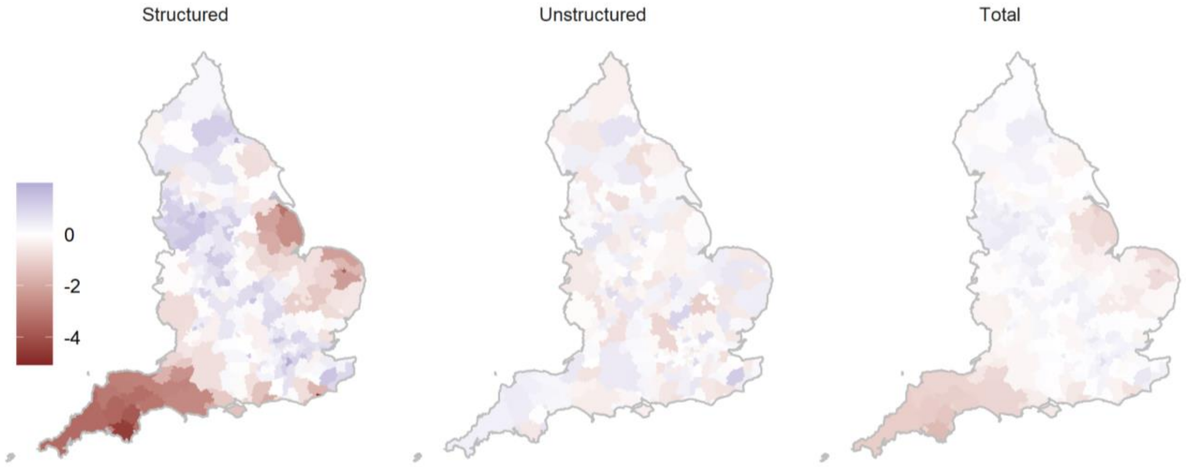


Figure S4: Decomposition of the fitted BYM spatial model into structured and unstructured components. For the selected model, the percentage of residual spatial variation attributable to the local correlation structure was estimated as 95% (95% CrI [86 - 99]).

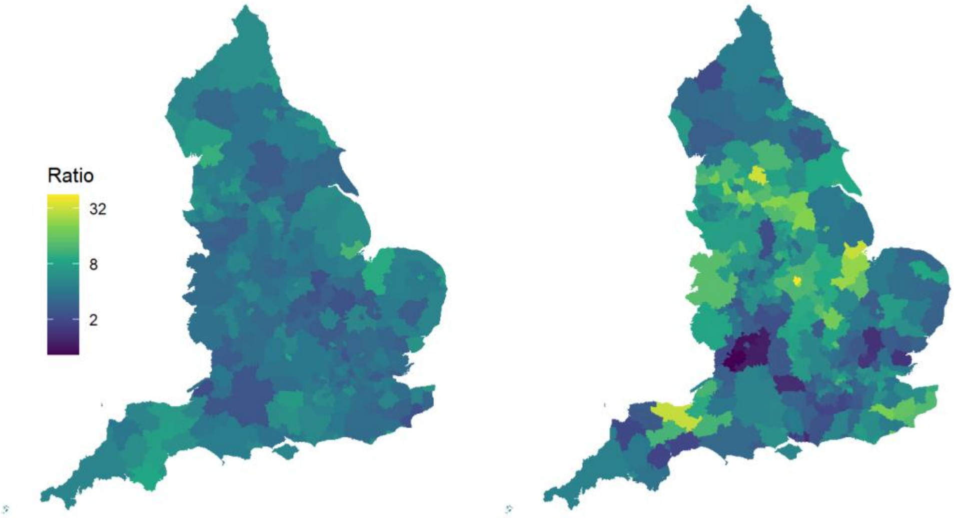


Figure S5: Median ratio between weekly observed cases and one-week-lagged modelled deaths per LTLA (population case-fatality ratio), before (left) and after (right) expansion of pillar 2 testing for all symptomatic individuals from 2020-05-18. Greater variation post-P2 expansion will in part be attributable to overall smaller counts of deaths per LTLA.

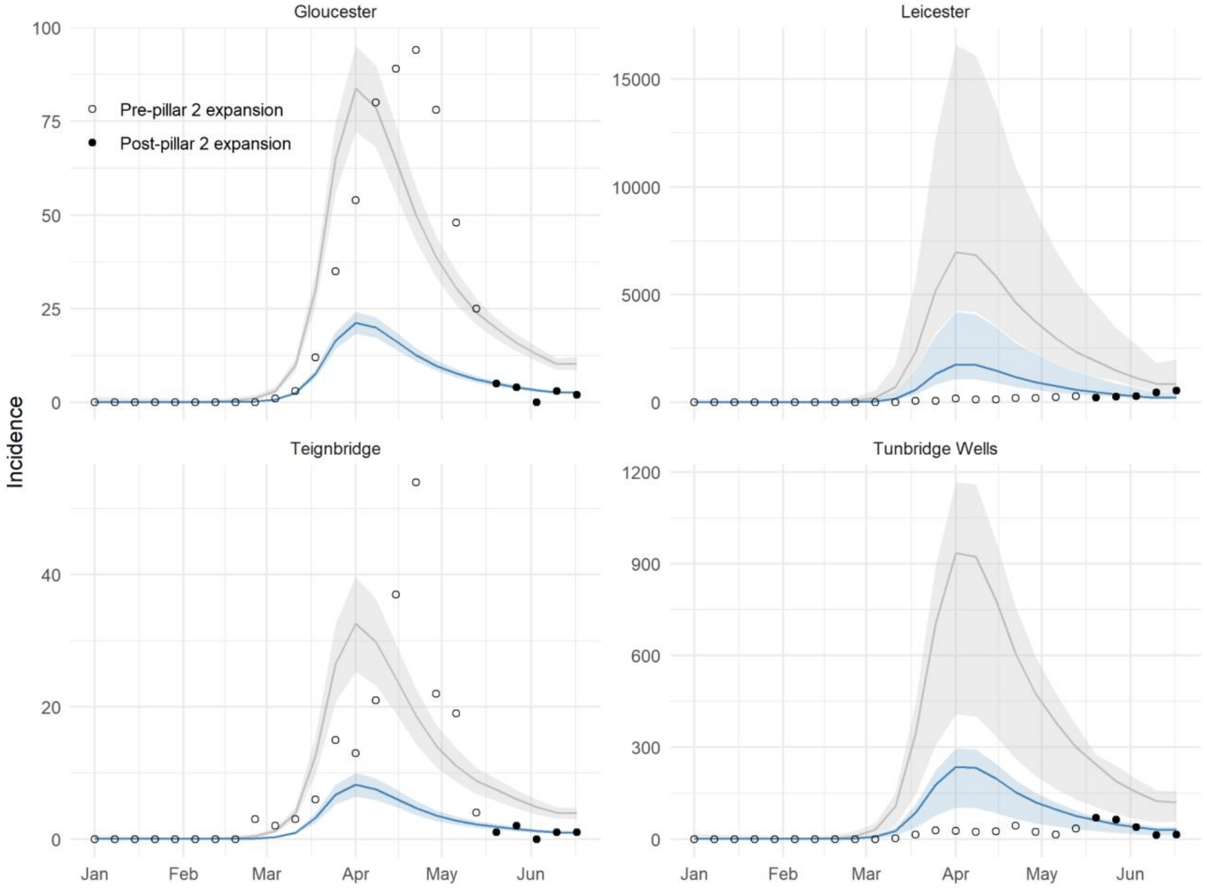


Figure S6: Predicted-P1+P2 cases (blue) and total infections (grey) over time, within LTLAs with the highest and lowest estimated detection rates. Estimates for Gloucester and Teignbridge were 96.6% [87%, 110%] and 96.1% [81%, 121%], while for Leicester and Tunbridge Wells were 6.7% [3%, 11%] and 6.8% [6%, 15%].

Supplementary Tables

Table S1: Summary of LTLA-level characteristics, overall and by geography type (median [IQR]). Age is defined as the estimated median according to age-specific population estimates for each LTLA, IMD as the median score across lower super output areas (the level at which the score is calculated) within each LTLA, and % minority population as the percentage of the LTLA population identifying as non-white according to the most recent census (2011).

		Median age	Median IMD score	% Minority population
Overall		41 [37, 45]	16.1 [11.4, 22.4]	0.05 [0.03, 0.13]
By geography	London Borough	34.5 [33, 36]	20.4 [13.9, 26.5]	0.39 [0.31, 0.47]
	Metropolitan District	39 [35, 41]	27.2 [21.4, 31]	0.11 [0.04, 0.19]
	Non-metropolitan District	43 [40, 46]	13.8 [10.8, 18.4]	0.04 [0.02, 0.07]
	Unitary Authority	39.5 [35.75, 43]	19.1 [13, 23.9]	0.06 [0.03, 0.14]

Table S2: Estimated coefficients for LTLA-level covariates (posterior mean and 95% credible interval), from the final selected model. Estimates are multiplicative due to the log link function, hence a value greater than one would indicate a positive association with COVID-19-related mortality rate and a value less than one a negative effect. A higher percentage of minority ethnicities and higher deprivation quintile in the LTLA were found to be associated with higher rates of COVID-19-related mortality, after accounting for the age and size of the population. Fixed effect estimates (posterior mean and 95% credible interval) from the final selected model.

Covariate		Estimate [95% CrI]
% minority ethnicity		1.01 [1.006, 1.015]
IMD score quintile	1 (least deprived)	1
	2	1.03 [0.96, 1.12]
	3	1.17 [1.06, 1.30]
	4	1.27 [1.10, 1.47]
	5 (most deprived)	1.21 [0.97, 1.49]

Table S3: Sensitivity analysis comparing predicted-P1+P2 cases under assumed lags of two and three weeks between confirmatory testing and death. As in Table 2, counts reflect the hypothetical scenario in which expanded surveillance (hospital- and community-based symptomatic testing) were available from the start of the epidemic. The differences explored here are a result of assuming a longer (either two or three week) average lag between the date a case is initially swabbed for testing and the date of death.

	Observed, test-confirmed cases (*up to week starting 2020-06-10*)	**Two week lag**		**Observed, test-confirmed cases (*up to week starting 2020-06-03*)**	**Three week lag**	
		Predicted (*median [IQR]*)	**Percentage difference**		**Predicted (*median [IQR]*)**	**Percentage difference**
England total	226,522	418,627 [352,699 - 493,737]	84.8 [55.7 - 118]	220,218	515,598 [452,200 - 582,182]	134.1 [105.3 - 164.4]
London Borough	33,118	54,447 [45,905 - 63,585]	64.4 [38.6 - 92]	32,809	67,350 [60,818 - 75,379]	105.3 [85.4 - 129.8]
Metropolitan District	61,976	130,758 [117,755 - 144,591]	111 [90 - 133.3]	59,757	153,856 [141,322 - 165,896]	157.5 [136.5 - 177.6]
Non-metropolitan District	77,965	125,167 [101,393 - 151,902]	60.5 [30 - 94.8]	76,100	159,341 [133,909 - 185,609]	109.4 [76 - 143.9]
Unitary Authority	53,463	107,980 [91,151 - 127,351]	102 [70.5 - 138.2]	51,552	134,568 [120,599 - 149,287]	161 [133.9 - 189.6]

Table S4: Summary of observed confirmed cases and estimated total infections, by geography and region.

	Observed, test-confirmed cases (up to week starting 2020-06-17)	Estimated total infections (median [98% CrI])	Percentage difference
England	231,817	1,323,622 [1,038,213 – 1,737,564]	17.5 [13.3 - 22.3]
London Borough	33,399	172,478 [122,976 – 233,570]	19.4 [14.3 - 27.2]
Metropolitan District	64,007	433,397 [345,610 – 568,269]	14.8 [11.3 - 18.5]
Non-metropolitan District	79,441	386,269 [292,924 – 511,017]	20.6 [15.5 - 27.1]
Unitary Authority	54,970	330,717 [247,024 – 479,398]	16.6 [11.5 - 22.3]
East Midlands	20,053	176,864 [131,958 – 271,843]	11.3 [7.4 - 15.2]
East of England	23,058	116,037 [88,180 – 151,430]	19.9 [15.2 - 26.1]
London	33,399	172,478 [122,976 – 233,570]	19.4 [14.3 - 27.2]
North East	14,981	48,207 [36,572 – 64,796]	31.1 [23.1 - 41]
North West	41,607	258,741 [212,280 – 315,380]	16.1 [13.2 - 19.6]
South East	33,249	138,331 [106,817 – 184,415]	24 [18 - 31.1]
South West	12,623	47,713 [27,010 – 65,821]	26.5 [19.2 - 46.7]
West Midlands	24,874	129,636 [104,508 – 156,690]	19.2 [15.9 - 23.8]
Yorkshire and The Humber	27,973	235,422 [184,770 – 335,078]	11.9 [8.3 - 15.1]