Imperial College London

`

Epidemiology and Forecasting of Influenza and COVID-19

`

Haowei Wang

Department of Infectious Disease Epidemiology

School of Public Health

Imperial College London

Thesis submitted for the degree of Doctor of Philosophy

August 2023

Declaration of Originality

I declare that this thesis and the work described within have been completed solely by myself under the supervision of Professor Steven Riley. Any work within that is not my own have been appropriately referenced in this thesis. Research ideas and plans were conceptualised with the support of Professor Steven Riley. I was a core member of the REal-time Assessment of Community Transmission-1 (REACT-1) study from May 2020 to April 2022, mainly responsible for estimating community prevalence in England and for risk factor analysis. I collaborated with colleagues on many aspects of the REACT-1 analyses and had the lead role for studying age, regional and socio-economic factors. I developed, managed and ran the code to produce the tables, figures and regression results for these variables for most REACT-1 outputs. I conducted all analyses presented in this thesis.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Respiratory infectious diseases (RID) have significant health and social impacts globally, and the speed of transmission and the scale of outbreaks pose significant challenges to public health systems and social stability. As the two major respiratory infectious diseases, influenza and SARS-CoV-2 are both highly contagious and clinically complex. Influenza causes large-scale outbreaks every year, posing a serious threat, especially to vulnerable populations. SARS-CoV-2 has spread rapidly across the world since its emergence in late 2019, leading to the COVID-19 pandemic and triggering public health crises worldwide. Therefore, this PhD thesis focuses on a number of important aspects of influenza and SARS-CoV-2, aiming to gain an in-depth understanding of their transmission status in the community, and prevalence in different subgroups of populations, and to construct reliable forecasting frameworks to support pandemic control and public health decision-making.

The thesis systematically analysed the transmission dynamics of SARS-CoV-2 in England through community surveillance, revealing its epidemiological trends and the prevalence of SARS-CoV-2 infection in different populations with respect to their social, economic and living characteristics. The results of the risk factor analysis demonstrated health inequality existed among ethnic groups, household size, and occupational activity, providing a scientific basis for targeted prevention and control and resource allocation. Furthermore, an efficient forecasting framework based on advanced data analysis and machine learning techniques was developed, which can accurately predict the trends of influenza and SARS-CoV-2 transmission, potentially offering a more practical tool for decision-making by the government and health departments than other currently available frameworks. These research results are expected to provide useful references for global epidemic prevention and control, as well as valuable experience for in-depth research and practice in the fields of public health and data science.

Acknowledgements

I would like to thank my supervisor, Professor Steven Riley, for his guidance and support throughout my PhD. Steven, you have been an amazing supervisor, and I am incredibly fortunate to be your PhD student and have worked with you for the past few years.

I thank the team Flu, Dr. Kylie Ainslie, Dr. Oliver Eales, Dr. Jame Hay, Dr. Caroline Walters, and Dr. Ada Yan. Thank you all for the help and support in providing me with invaluable advice and suggestions during my PhD research. Thank the REACT study investigation team, with whom I have worked closely to achieve excellent results in my PhD research. Also thank Professor Kin O Kwok, not only for the academic collaboration, but also for valuable advice on my career path.

I also would like to my family. Thank my mom Mrs Yong Fang (方咏女士) and dad Mr Huibin Wang (汪徽滨先生) - thank you for your unconditional love and support along the way, being your daughter was the best thing that ever happened to me. Thank my dear grandparents Mrs Quan Wang (汪权女士), Mr Renli Wang (汪任理先生), and Mr Genrong Fang (方根荣) for your unfailing love and support. Thank my boyfriend Shawn Xue (薛帅), for your constant companionship and support, and for always being there for me when I needed it most. I am more than fortunate to meet you and have you on this life journey. Thank you to my two kitties, Rice Cake and Giselle, for keeping me company with your fluffy bodies through the tough times.

Finally, I want to give a special thanks to my grandmother Mrs Lifen Luo (罗丽芬女士) in heaven. You have been my strongest support as I have grown up. Even though you are no longer with me, I still feel your presence. Your wisdom and care will always guide me forward. In this thesis, I want to dedicate my accomplishments to you, hope you will be proud of your granddaughter. Thank you for your companionship and love, which I will always remember.

Table of Contents

Declaration of Originality **2**
Copyright Declaration **3**
Abstract **4**
Acknowledgements **5**
Table of Contents **7**
List of Figures **10**
List of Tables **12**
List of abbreviations and acronyms **13**
Chapter 1 Introduction and Epidemiology of Influenza and SARS-CoV-2 ..... **18**
1.1 Introduction to the Thesis 18
1.2 Influenza 21
1.2.1 Key Biology of Influenza Viruses 21
1.2.1.1 Gene Structure of Influenza Virus 21
1.2.1.2 Viral Replication 22
1.2.1.3 Immune Response 24
1.2.2 Surveillance 26
1.2.2.1 Clinical Surveillance of Influenza 26
1.2.2.2 Global Surveillance of Influenza 27
1.2.3 Epidemiology of Influenza 28
1.2.3.1 Global Burden of Influenza 28
1.2.3.2 Seasonality of Influenza 31
1.3. SARS-CoV-2 33
1.3.1 Biology of SARS-CoV-2 34
1.3.1.1 Gene Structure of SARS-CoV-2 34
1.3.1.2 Viral Replication 35
1.3.1.3 Immune Evasion 37
1.3.2 Surveillance 38
1.3.2.1 Case Definitions and Testing 38
1.3.2.2 Global Surveillance Networks of SARS-CoV-2 40
1.3.3 Epidemiology of SARS-CoV-2 41
1.3.3.1 Transmission 42
1.3.3.2 Global Patterns 43
1.3.3.3 Disease Burden and Mortality of SARS-CoV-2 44
1.4. Community Surveillance 46
Chapter 2 Infectious Diseases Forecasting **48**
2.1 Background 48

2.1.1 Overview....48
2.1.2 Forecast Targets and Time Horizons....49
2.1.3 Nowcasting....50
2.1.4 Interactions between Surveillance and Forecasting....52
2.1.5 Scenario Analysis versus Forecasting....53
2.1.6 Challenges....55
2.1.6.1 Data Quality....55
2.1.6.2 Visualisation of Forecasting Results....56
2.2 Evaluation....58
2.2.1 Model Training and Testing....58
2.2.2 Accuracy Metrics....59
2.3 Models Used for Influenza and COVID-19 Forecasting....61
2.3.1 Time-series Statistical Models....62
2.3.2 Mechanistic Models....63
2.3.2.1 Compartmental models....63
2.3.2.2 Agent-based Model (ABM)....66
2.2.3 Machine Learning (ML) and Deep Learning (DL) Models....67
2.2.3.1 ML/DL in Influenza/influenza-like Illness (ILI)....67
2.2.3.2 ML/DL in COVID-19....68
Chapter 3 Forecasting Influenza Incidence as an Ordinal Variable Using Machine Learning....72
3.1 Introduction....72
3.2 Methods....74
3.2.1 Data....74
3.3.2 Models....75
3.2.3 Forecast and Features....76
3.2.4 Accuracy Metrics....80
3.2.4.1 Macro-averaged Mean Absolute Error....80
3.2.4.2 Mean Zero-one Error....81
3.3 Results....82
3.4 Discussion....96
Chapter 4 The Interaction of Age and Geography in the Transmission Dynamics of SARS-CoV-2 during the 2020/21 UK Winter Wave....101
4.1 Introduction....101
4.2 Methods....103
4.2.1 Data Collection....103
4.2.2 Statistical Analysis....105
4.3 Results....106
4.3.1 Overall Levels of Swab-positivity during the Second Wave in England....106
4.3.2 Higher Prevalence of Test-positivity in Children....109

4.3.3 Heterogeneous Regional Pattern in Prevalence of SARS-CoV-2 Infection.....114
4.3.4 Interaction of Weighted Prevalence between Age Group and Geographic Regions in England.....116
4.4 Discussion.....121
Chapter 5 Health Inequalities in SARS-CoV-2 Infection during the Second Wave in England: REACT-1 Study..... 124
5.1 Introduction.....124
5.2 Methods.....126
5.2.1 Data Processing - Socio-demographic Groups.....126
5.2.2 Statistical Analysis.....128
5.3 Results.....128
5.3.1 Higher Prevalence of Test-positivity in Males.....128
5.3.2. Ethnicity.....136
5.3.3 Household Size and Neighbourhood Deprivation.....143
5.3.4 Occupation.....145
5.3.5 Sensitivity Analysis for Vaccination Status.....150
5.4 Discussion.....152
Chapter 6 Forecasting Regional-level COVID-19 Hospitalisation in England as an Ordinal Variable Using the Machine Learning Method..... 155
6.1 Introduction.....155
6.2 Methods.....157
6.2.1 Data.....157
6.2.2 Primary Forecast Model.....159
6.2.3 Combinations of Predictors.....162
6.2.4 Forecasting.....163
6.2.5 Evaluation.....164
6.3 Results.....165
6.4 Discussion.....190
Chapter 7 Discussion..... 193
7.1 Summary of Findings.....193
7.2 Limitations of the Research and Future Directions.....195
7.3 Implications of the Research.....202
Reference..... 205

List of Figures

Figure 2.1 Susceptible-infected-recovered (SIR) model. 64

Figure 2.2 Susceptible-exposed-infectious-removed (SEIR) model. 64

Figure 3.1 Schematic for time series split cross-validation. 79

Figure 3.2 World map. 83

Figure 3.3 Distribution of categorical incidence levels for weekly influenza-like illness for countries included in the study....84

Figure 3.4 Differences in macro-averaged mean absolute error (mMAE) by model....86

Figure 3.5 Differences in mean-zero error (MZE) by model. 88

Figure 3.6 Macro-averaged mean absolute error (mMAE) of XGBoost models. 90

Figure 3.7 Mean-zero error (MZE) of XGBoost models. 91

Figure 3.8 Macro-averaged mean absolute error (mMAE) for thirty-two countries. 92

Figure 3.9 Mean-zero error (MZE) for 32 countries. 93

Figure 3.10 Comparison of Actual and predicted incidence by country. 94

Figure 3.11 Comparison of actual and predicted incidence. 95

Figure 3.12 Distribution of predictions against each decile of observation for the XGBoost model with an extended window. 99

Figure 4.1 Comparison of overall weighted prevalence as measured by REACT-1 to surveillance case data (Public Health England, PHE). 108

Figure 4.2 Weighted prevalence of swab positivity by age groups. 113

Figure 4.4 Nested weighted prevalence by age and region for rounds 5 to 10. 117

Figure 5.1 Forest plots of sex and household size. 129

Figure 5.2 Forest plot of ethnicity categories and detailed subgroups of ethnicity. 137

Figure 5.3 Forest plot of deprivation. 144

Figure 5.4 Forest plot of occupations for rounds 5 to 10 and an average across rounds weighted inverse to variance (rounds 5 to 10). 147

Figure 6.1 Epidemic curve of weekly hospitalisation by NHS regions in England from 19 March 2020 to 31 December 2022. 166

Figure 6.2 Distribution of ordinal levels for weekly hospitalisations for NHS regions discretised by the uniform method. 167

Figure 6.3 Epidemic curve of weekly hospitalisation by NHS regions in England from 19 March 2020 to 31 December 2022. 169

Figure 6.4 Model performance of predicting n-tile hospitalisation levels with only epidemiological predictors. 171

Figure 6.5 Evaluation metrics comparison of 4-week ahead forecasts between models trained with different combinations of predictors.....173

Figure 6.6 Performance of the XGBoost models incorporated with epidemiological and weather data, measured by macro-averaged Mean Absolute Error (mMAE) and accuracy. 174

Figure 6.7 Performance of the XGBoost models incorporated with epidemiological and mobility data, measured by macro-averaged Mean Absolute Error (mMAE) and accuracy. 176

Figure 6.8 Performance of XGBoost models with different sets of features, measured by macro-averaged Mean Absolute Error (mMAE) and accuracy. 178

Figure 6.9 Comparison of model performance by NHS regions for XGBoost models and baseline models trained with different combinations of predictors when the prediction horizon is 4 week-ahead for the test period 2022. 180

Figure 6.10 The relative importance (Gain) of predictors used by the XGBoost ordered model. 181

Figure 6.11 Comparison of NHS-region-specific performance for the XGBoost ordered model. 183

Figure 6.12 Comparison of actual and predicted hospitalisation levels by NHS regions. 187

Figure 6.13 Comparison of changes in evaluation metrics when using the n-tile method versus the uniform method. 189

List of Tables

Table 2.2 Comparison of advantages and limitations of different forecasting methodologies. 71

Table 3.1 Time ranges of training and test sets for the XGBoost model. 77

Table 3.2 List of input features and value ranges for the XGBoost model. 77

Table 3.3 Summary of values of hyperparameters used for XGBoost models. 80

Table 3.4 Overall macro-averaged mean absolute error (mMAE) of XGboost compared with baseline models (Historical average and Null models). 87

Table 3.5 Overall mean-zero error (MZE) of XGboost compared with baseline models (Historical average and Null models). 89

Table 4.1 Overall unweighted and weighted prevalence with 95% CI for rounds 5 to 10....109

Table 4.2 Unweighted and weighted prevalence by age groups and regions from rounds 5 to 10. 110

Table 4.3 Nested weighted prevalence by region and age groups. 118

Table 5.1 The unweighted and weighted prevalence of swab-positivity by socio-demographic variables for rounds 5 to 10. 130

Table 5.2 Odds ratios for rounds 5 to 10, and an average weighted inversely to variance for core variables. 134

Table 5.3 The unweighted and weighted prevalence of swab-positivity for detailed ethnicity subgroups for rounds 5 to 10. 139

Table 5.4 Odds ratios for rounds 5 to 10, and an average weighted inversely to variance for detailed subgroups of ethnicity. 141

Table 5.5 Odds ratios for rounds 5 to 10, and an average weighted inversely to variance for detailed subgroups of occupations. 148

Table 5.6 Odds ratios for rounds 9 and 10, and an average weighted inversely to variance for linked vaccination data. 151

Table 6.1 Detailed descriptions of predictors. 159

Table 6.2 Summary of predictors incorporated by each model. 163

Table 6.3 Hyperparameter values used by the XGBoost models. 164

Table 6.4 Predictive performance of models when the ordinal level of hospital admissions is defined by the n-tile method. 179

Table 6.5 Predictive performance of models when the ordinal level of hospital admissions is defined by the uniform method. 188

List of abbreviations and acronyms

ABM	Agent-based model
ACE2	angiotensin-converting enzyme 2
ADCC	antibody-dependent cellular cytotoxicity
ANFIS	adaptive neuro-fuzzy inference system
ANN	Artificial Neural Network
AR	autoregressive
ARDS	acute respiratory distress syndrome
ARIMA	AutoRegressive Integrated Moving Average
CDC	Centers for Disease Control and Prevention
CNN	convolutional neural network
cRNA	complementary RNA
cRNP	complementary ribonucleoprotein
CT	cycle threshold
DL	deep learning
DNN	deep neural network
E protein	Envelope protein
ECMWF	European Centre for Medium-Range Weather Forecasts
EMD	Empirical mode of decomposition
ERA5	fifth-generation reanalysis
FluID	Flu Informed Decisions
FNN	feedforward neural networks
GBD	Global Burden of Disease
GFT	Google Flu Trends
GISRS	Global Influenza Surveillance and Response System
GLM	generalised linear model

GOARN	Global Outbreak Alert and Response Network
GOPC	general outpatient clinics
GPR	gaussian process regression
HA	hemagglutinin
IAV	influenza A viruses
IBV	influenza B viruses
ICU	intensive care unit
ICV	influenza C viruses
IDV	influenza D viruses
IFN	type I interferon
ILI	influenza-like illness
ILINet	Outpatient Influenza-like Illness Surveillance Network
LASSO	least absolute shrinkage and selection operator
LRI	lower respiratory tract infection
LSTM	long short-term memory
LTLA	lower-tier local authority area
M	matrix protein
M protein	membrane protein
MA	moving average
MAE	mean absolute error
MAPE	mean absolute percentage error
MARE	mean relative absolute error
MERS-CoV	Middle-East respiratory syndrome coronavirus
MHC	major histocompatibility complex
ML	machine learning
MLP-ICA	multi-layered perceptron-imperialist competitive algorithm

mMAE	macro-averaged mean absolute error
mRNA	messenger RNA
MSE	Mean square error
MZE	mean-zero error
N protein	Nucleocapsid protein
NA	neuraminidase
NCIRD	National Center for Immunization and Respiratory Diseases
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NLRP3	NOD-like receptor
NobBS	Nowcasting by Bayesian Smoothing
NP	nucleoprotein
NPIs	non-pharmaceutical interventions
NS1	non-structural proteins 1
NS2	non-structural proteins 2
OLR	ordered logistic regression
ONS	Office for National Statistics
OR	odds ratio
PA	polymerase acidic protein
PB1	polymerase basic protein 1
PB2	polymerase basic protein 2
PCR	polymerase chain reaction
PHEIC	Public Health Emergency of International Concern
PPE	personal protective equipment
PRRs	pattern recognition receptors
REACT-1	REal-time Assessment of Community Transmission-1

RF	random forest
RID	respiratory infectious disease
RIG-I	retinoic acid-inducible gene I
RIM	Random Iterative Method
rMAE	relative mean absolute error
RMSE	root mean square error
RNN	recurrent neural network
rNTPs	ribonucleoside triphosphates
RTI	respiratory tract infection
RT-PCR	reverse transcription-polymerase chain reaction
S protein	spike protein
SARI	severe acute respiratory infection
sARIMA	Seasonal AutoRegressive Integrated Moving Average
SARS-CoV-1	severe acute respiratory syndrome coronavirus
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SEIR	susceptible-exposed-infectious-removed
SIQR	susceptible-infected-quarantined-recovered
SIR	Susceptible-infectious-recovered
SVM	support vector machine
SVR	support vector regression
TLR	Toll-like receptors
URI	upper respiratory tract infection
VAR	vector autoregression
vRNA	viral RNA
vRNPs	viral ribonucleoprotein complexes
WHO	World Health Organization

XGboost	Extreme Gradient Boosting

Chapter 1 Introduction and Epidemiology of Influenza and SARS-CoV-2

1.1 Introduction to the Thesis

Infectious diseases are caused by germs (viruses, bacteria, or other pathogenic microbes) (Morens, Folkers & Fauci, 2004). Among infectious diseases, respiratory infectious diseases (RIDs) have captured the attention of the public because they cause high levels of morbidity and mortality. Germs that cause infections of the respiratory system are frequently transmitted by mucus and saliva when a person coughs, sneezes, or talks. Respiratory tract infections (RTI) are mainly classified as either upper respiratory tract infections (URIs) or lower respiratory tract infections (LRIs).

Respiratory infectious diseases pose a considerable burden to health every year, despite the fact that cost-effective methods for their prevention are available. According to the latest estimate from the 2019 Global Burden of Disease (GBD) study, RIDs caused ~2.1 billion cases and over 3.6 million deaths worldwide, accounting for 36% of all infectious disease deaths in the non-pandemic year 2019 (Institute for Health Metrics and Evaluation, 2019) The impact of RIDs on health systems can also be measured in economic terms. For example, in the United States, the economic impact of seasonal flu is estimated to be around $10.4 billion and $16.3 billion per year in direct medical costs and indirect productivity costs respectively (Young-Xu et al., 2017).

Respiratory infectious diseases also cause pandemics which can result in even greater negative impacts, for example, the 1918 influenza pandemic, the 1968 influenza pandemic, and the 2009 influenza pandemic. The COVID-19 pandemic, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is the latest disruptive global event caused by RID, leading to substantial mortality and morbidity globally, with over 600 million

confirmed cases reported, resulting in nearly 7 million deaths as of October 2022 (World Health Organization (WHO), 2020). As this thesis is finalised, the BA.2.86 variant is just being described (Centers for Disease Control and Prevention, 2023).The COVID-19 pandemic is not yet over, and potential subsequent waves of infection may occur. These estimates will increase. More specifically, COVID-19 causes significant demand on a number of resources, including oxygen, ward beds, intensive care units (ICU), anaesthetists, intensivists, healthcare staff, and personal protective equipment (PPE). There always remains a need for non-COVID-19 care, lending to an unbalance of healthcare resources between two types of patients.

Understanding the epidemiology of RIDs is essential to develop timely and effective interventions to mitigate their impact on individuals and communities. Epidemiological studies provide valuable information on the transmission dynamics, risk factors, and outcomes of respiratory infectious diseases. By identifying patterns of disease spread, researchers and public health officials can better understand the routes of transmission and risk factors associated with severe outcomes. For instance, epidemiological investigations have revealed that respiratory viruses, such as influenza and SARS-CoV-2, are transmitted through respiratory droplets, emphasizing the importance of preventive measures such as hand hygiene, respiratory etiquette, and social distancing (Zhu et al., 2020; Zhou et al., 2020b). Furthermore, by analysing epidemiological data, researchers can identify vulnerable populations, such as the elderly and individuals with underlying health conditions, who may require targeted interventions to reduce their risk of severe illness (Huang et al., 2020).

Forecasting respiratory infectious diseases is equally crucial for preparedness and response efforts. By employing computational and mathematical models and by using other statistical analyses, researchers can project the future trajectory of disease outbreaks, enabling public health authorities to anticipate potential surges in infections and plan resource allocation accordingly. For example, forecasting models have been instrumental in predicting the

spread of seasonal influenza and guiding the formulation of influenza vaccines each year (Boddington et al., 2021). During the COVID-19 pandemic, forecasting models have been employed to estimate the demand for hospital beds, ventilators, and other medical resources, aiding healthcare systems in capacity planning and preparedness (Chang et al., 2020; Ferguson et al., 2020; Krylova et al., 2020; Johnson et al., 2023; Deschepper et al., 2021). Moreover, accurate forecasting of the spread of RIDs is essential for implementing timely and targeted control measures. For example, during the omicron wave in the UK, 2-week ahead forecasts were used by senior officials to plan for possible surges in healthcare demand (Mellor et al., 2023).

To understand the epidemiology and develop accurate forecasting models for RIDs, the thesis is organised as follows. The remainder of this chapter reviews the epidemiology of the respiratory infectious diseases that are the focus of the thesis - influenza and SARS-CoV-2, providing the biological and epidemiological background required for the construction of the later analytical methods. Chapter 2 gives an overview of the forecasting models that have been developed and applied to influenza and SARS-CoV-2 surveillance data. Chapter 3 proposes a machine learning forecasting framework to predict influenza-like-illness (ILI) as an ordinal outcome and applies it to the influenza surveillance data of multiple countries worldwide prior to the COVID-19 pandemic, during which influenza transmission ceased in many populations. Chapter 4 uses data from the the REal-time Assessment of Community Transmission-1 (REACT-1) study REACT-1 study to investigate the epidemiology of SARS-CoV-2 infection with respect to age and geographical region in England. Building on Chapter 4, Chapter 5 further analyses the epidemiology of SARS-CoV-2 infection and identifies risk factors related to occupation, sociodemographic status and ethnicity in England. Chapter 6 returns to forecasting to extend the machine learning framework developed in Chapter 3 for influenza and applies it to the prediction of COVID-19 hospital admission levels. Finally, in Chapter 7, the public health implications of findings from the previous six chapters are discussed.

1.2 Influenza

1.2.1 Key Biology of Influenza Viruses

This section provides an overview of the key aspects of influenza biology, including the gene structure of influenza viruses, the process of cell replication, and the immune response elicited by the host. By investigating these fundamental aspects, we can gain insights into the mechanisms underlying viral replication, pathogenesis, and host immune interactions, ultimately aiding in the development of interventions and therapeutics.

1.2.1.1 Gene Structure of Influenza Virus

Influenza viruses are members of the Orthomyxoviridae family and are characterised by their envelope, segmented genome, and negative-stranded viral RNA (vRNA) structure. Influenza viruses are categorised into four types, namely influenza A, B, C and D viruses, distinguished by their immunological and biological characteristics. While influenza A viruses (IAV) and influenza B viruses (IBV) both cause seasonal epidemics in humans, IAVs are of particular significance because they are the only known type of influenza viruses able to cause pandemics, such as the 1918 pandemic caused by the H1N1 subtype, and subsequent pandemics caused by H2N2 in 1957, H3N2 in 1968, and another H1N1 in 2009 (Taubenberger & Morens, 2008; Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team et al., 2009; Palese, 2004). Infections caused by influenza C viruses (ICV) typically result in mild symptoms and are not considered capable of triggering widespread outbreaks among humans (Sederdahl & Williams, 2020). Finally, influenza D viruses (IDV) predominantly affect cattle and swine and only occasionally transmit to other animal species. There is currently no evidence to suggest they have the ability to infect humans or induce illness in them (Liu et al., 2020). Since the IDVs are newly discovered, and many basic questions are still being addressed, they will not be discussed further below.

The structure of IAVs and IBVs exhibit a number of similarities, while ICVs display a greater degree of divergence. The genomes of IAVs and IBVs are composed of eight segments of

negative-sense, single-stranded vRNA (Lamb, 2008). Each vRNA segment encodes for different viral proteins essential for viral replication and pathogenesis. The segments for IAVs are designated as: polymerase basic protein 2 (PB2), polymerase basic protein 1 (PB1), polymerase acidic protein (PA), hemagglutinin (HA), nucleoprotein (NP), neuraminidase (NA), matrix protein (M), and non-structural proteins 1 and 2 (NS1 and NS2) (Lamb, 2008). IBVs have similar vRNA segments, with the key differences that segment 6 encodes for the NA and NB proteins, and segment 7 encodes the M1 and BM2 proteins. ICVs possess seven segments of single-stranded vRNA, distinguishing them from other influenza viruses by the absence of the NA gene.

The antigenic properties of the two surface spike glycoproteins, HA and NA, further classify IAVs into subtypes. These subtypes are defined by eighteen distinct HA subtypes and 11 distinct NA subtypes, denoted as H1 through H18 and N1 through N11, respectively. Although over 130 combinations of influenza A subtypes have been observed in the wild, predominantly in avian species, the potential for numerous additional subtype combinations exists due to the phenomenon of virus "reassortment." Reassortment refers to the exchange of gene segments between influenza viruses (Steel & Lowen, 2014). This process occurs when two different influenza viruses infect a single host simultaneously and exchange genetic material (Steel & Lowen, 2014). For instance, the H5N1 viruses and their genotypes emerged through multiple reassortments with other influenza subtypes, and further diversified into distinct clades, which spread to impact more than 60 countries. (Vijaykrishna et al., 2008).

1.2.1.2 Viral Replication

Understanding the intricate processes involved in the cell replication of influenza can provide insights for the development of effective antiviral drugs and vaccines. Targeting key viral proteins involved in replication, such as the polymerase complex or the matrix protein, may

inhibit viral replication and potentially limit disease progression (Pflug et al., 2017). Furthermore, identifying host factors involved in the replication process may uncover potential targets for therapeutic interventions.

The majority of our understanding regarding the molecular mechanisms underlying viral messenger RNA (mRNA) synthesis and vRNA replication for influenza viruses has come from investigations of IAVs. IBVs are inferred to have similar mechanisms as IAVs for transcription and replication processes due to their comparable genomic structure and the presence of polymerase proteins that exhibit extensive sequence homology with those found in IVAs (Krug & Fodor, 2013). Therefore, we will focus on the IAVs in this and the following sections, and all the "influenza viruses" mentioned later in this section refer to IAVs.

The replication of influenza viruses within host cells is a multifaceted process involving intricate interactions between viral proteins and host factors. Following attachment to host cells through the binding of viral HA proteins to sialic acid receptors on the cell surface, the viral envelope fuses with the host cell membrane, facilitating the release of viral ribonucleoprotein complexes (vRNPs) into the cytoplasm. When the virus enters the host cell, the vRNPs are transported into the nucleus, where viral RNA replication and transcription occur. Polymerase proteins, including PB2, PB1, and PA, will get involved in these processes.

Within the nucleus, the transcription and replication of vRNAs are performed by the polymerase (Dou et al., 2018). The replication of the influenza genome occurs in two stages. The first stage is the synthesis of complementary RNA (cRNA) strands from the negative-sense vRNAs through transcription. In the second stage, these cRNAs serve as templates to produce new copies of vRNAs, which can be either incorporated into newly formed virions or used for further replication. cRNA production occurs through an unprimed mechanism that relies on the accurate pairing of free ribonucleoside triphosphates (rNTPs), typically GTP

and ATP, with the 3' end of the vRNA template (York et al., 2013; Newcomb et al., 2009). Following its release from the polymerase, the cRNA associates with newly synthesised NP molecules and a single molecule of the viral polymerase, assembling into a complementary ribonucleoprotein (cRNP) complex (York et al., 2013).

Once vRNAs are synthesised, they are exported from the nucleus back to the cytoplasm, where subsequent stages of the replication cycle take place. The M1 protein binds to the newly synthesised vRNAs and assists in their nuclear export. The vRNAs then associate with the NP to form the vRNP complexes, preparing for viral assembly and packaging into progeny virions. In the final stages of the replication cycle, the viral components, including HA and NA proteins, the vRNPs, and matrix proteins, are assembled at the host cell membrane. The NA protein facilitates the release of mature virus particles from the cell surface by cleaving sialic acid residues, preventing viral aggregation and promoting the spread of infection.

1.2.1.3 Immune Response

The ability to evade and manipulate the host immune system is one of the key features of influenza viruses. Immune responses triggered by influenza virus infections in humans exhibit considerable breadth and durability. However, influenza viruses can elude these immune responses over time due to their high mutation rates and antigenic adaptability (Al Khatib et al., 2019). Vaccines are considered the most effective available defence against infection and disease, but their efficacy in comparison to some other viral vaccines is limited (Krammer, 2019). Understanding the underlying mechanism of immune response to influenza virus infection is important for vaccine development and for understanding the current immune status of the population (Krammer, 2019).

The immune response to influenza involves a complex interplay between the innate and adaptive immune systems, orchestrated by various immune cells and soluble factors. Upon infection, the initial line of defence is the innate immune response, which provides rapid but nonspecific protection against the virus. The innate immune system is intricately structured to identify microbial infections by recognising specific molecular patterns that are distinctively found in microorganisms, including viruses (Iwasaki & Pillai, 2014). The recognition of influenza viruses is achieved by pattern recognition receptors (PRRs). There are at least three different classes of PRRs involved: Toll-like receptors (TLR3, TLR7, and TLR8), retinoic acid-inducible gene I (RIG-I)-like receptors and NOD-like receptors (NLRP3) (Pang & Iwasaki, 2012). Both RIG-I and TLRs trigger the activation of type I interferon (IFN) and cytokine gene expressions, whereas NLRP3 leads to the activation of caspase-1, which cleaves and releases cytokines such as interleukin 1β (IL-1β) and IL-18.

Following the innate response, adaptive immunity is activated. The adaptive immune response to influenza is primarily mediated by B cells and T cells. B cells produce virus-specific antibodies that can neutralise the virus, prevent viral entry into cells, and infected cell clearance by antibody-dependent cellular cytotoxicity (ADCC) (Janssens et al., 2022). T cells, on the other hand, recognise viral peptides presented on major histocompatibility complex (MHC) molecules and can directly kill infected cells or support the activities of B cells and other immune cells (Swain, McKinstry & Strutt, 2012).

The major targets of the adaptive immune response to influenza are the viral surface glycoproteins HA and NA (Krammer, 2019). Antibodies targeting the HA protein can inhibit viral entry into host cells and limit viral replication, while antibodies against the NA protein can prevent the release of newly formed viral particles from infected cells. However, due to the high genetic variability of influenza viruses, the HA and NA proteins can undergo

antigenic drift and recombination, leading to the emergence of new strains that can evade pre-existing immunity (Koel et al., 2013).

1.2.2 Surveillance

Influenza outbreaks pose significant challenges to public health due to their ability to spread rapidly and cause substantial morbidity and mortality. Routine surveillance and outbreak investigations can be both helpful in reducing that impact.

1.2.2.1 Clinical Surveillance of Influenza

Clinical surveillance is instrumental in detecting and tracking influenza cases, providing critical information for public health decision-making and response efforts (Al Khatib et al., 2019). By monitoring clinical symptoms and laboratory-confirmed cases fast and accurately, clinical surveillance provides early warnings, enhances the understanding of influenza epidemiology, and aids in implementing preventive measures (Al Khatib et al., 2019).

One approach to clinical surveillance is syndromic surveillance, which involves monitoring the symptoms reported by patients seeking healthcare prior to the laboratory confirmation (Mandl et al., 2004). Syndromic surveillance systems track data from various sources, such as emergency departments, outpatient clinics, and pharmacies, and analyse them in near real-time to detect trends of influenza activity (Buehler et al., 2007). ILI or influenza-associated hospitalisations captured by syndromic surveillance provide early warning signals of influenza outbreaks and help identify vulnerable populations (Olson et al., 2007).

Laboratory testing is another essential component of clinical surveillance for influenza. Diagnostic tests, including polymerase chain reaction (PCR) assays and viral culture, detect and confirm the presence of influenza viruses in respiratory samples (Uyeki et al., 2019). These tests provide information on the circulating influenza strains, subtypes, and their

antigenic characteristics, which is essential for monitoring influenza virus evolution and guiding strain selection for vaccine campaigns (Krammer & Palese, 2015). Efficient laboratory testing also helps identify the degree to which influenza viruses are contributing to severe cases of respiratory disease, such as those requiring hospitalisation or associated with severe complications, facilitating appropriate clinical management and public health response (Uyeki et al., 2019).

1.2.2.2 Global Surveillance of Influenza

National influenza surveillance systems are important in monitoring the occurrence and spread of influenza epidemics and providing academic researchers with reliable and sufficient data sources to develop forecasting models. At the global level, several methods and strategies are employed to facilitate timely and accurate data collection, reporting, aggregation and analysis. These efforts aim to provide a comprehensive understanding of influenza activity worldwide, inform public health responses, and support pandemic preparedness.

Data collection for global influenza surveillance involves the gathering of information on influenza cases, including their geographic location, demographic characteristics, clinical presentation, and virological characteristics (Global Influenza Surveillance and Response System (GISRS), n.d.; Caini et al., 2015). Multiple sources contribute to the data collection, such as sentinel surveillance systems (Olson et al., 2007; Virtual meeting, 2022), laboratory testing (Al Khatib et al., 2019), and healthcare utilisation data (Murray et al., 2018). Sentinel surveillance systems, commonly implemented in many countries, involve a network of healthcare facilities that report ILI or severe acute respiratory infection (SARI) cases to designated public health authorities (World Health Organization. Regional Office for South-East Asia, 2002).

To ensure standardised reporting of influenza data, global reporting systems have been established. The WHO coordinates and facilitates the Global Influenza Surveillance and Response System (GISRS), which works closely with national influenza centres, collaborating centres, and other reference laboratories worldwide (Global Influenza Surveillance and Response System (GISRS), n.d.). GISRS serves as a platform for exchanging influenza-related information worldwide, including virus isolates, antigenic and genetic characterization data, and epidemiological updates, which enhances data collection and sharing among different countries and regions (Global Influenza Surveillance and Response System (GISRS), n.d.). Additionally, real-time monitoring of influenza viruses and their genetic evolution can help early detection of novel influenza viruses with potential epidemics so as to optimise global preparedness and timing of vaccination campaigns (Bedford et al., 2015).

1.2.3 Epidemiology of Influenza

This section provides an overview of key aspects of influenza epidemiology, including the global burden of influenza, seasonal patterns, the impact on healthcare systems, and the emergence of pandemics and novel influenza strains.

1.2.3.1 Global Burden of Influenza

Influenza is a significant global public health concern, causing substantial morbidity and mortality each year all around the world (Khanna et al., 2008). The burden of influenza is assessed from multiple perspectives, including annual incidence, morbidity, and mortality rates, which vary across regions and populations (Li et al., 2018; Kovács et al., 2014).

The annual global incidence of influenza constitutes a major health hazard, with many millions of cases reported worldwide each year. It is estimated that influenza viruses cause between 3 and 5 million cases of severe illness annually, leading to 290000 to 650000

respiratory-related deaths (World Health Organization, 2023a). The morbidity associated with influenza extends beyond the number of cases and includes various aspects such as healthcare utilisation (Groves et al., 2022; O'Halloran et al., 2021), and hospitalisations (Ortiz et al., 2014). Influenza-related hospitalisations place a substantial strain on healthcare systems globally, particularly during peak seasons. It is estimated that influenza-associated hospitalisations range from 140,000 to 710,000 annually (Iuliano et al., 2018).

In addition to adverse health impacts on individuals, influenza also exerts a considerable societal and economic burden (Young-Xu et al., 2017). It affects productivity and disrupts healthcare services, particularly during peak influenza seasons when healthcare facilities are overwhelmed with increased patient demand (Verikios et al., 2012). The burden on healthcare systems is further compounded by the need for intensive care unit (ICU) admissions and the use of antiviral medications and other medical interventions (Fleming & Elliot, 2008). The economic burden of influenza is significant, encompassing direct medical costs, productivity losses, and the societal impact of illness and premature death (Molinari et al., 2007). In the United States, the combined direct medical expenses have been approximated at 3.2 billion US dollars annually, while the indirect costs reached 8 billion US dollars (de Courville et al., 2022). Similarly, in the European Union, the annual cost of seasonal influenza is estimated to range from 6 billion euros dollars to 14 billion euros dollars (de Courville et al., 2022).

Morbidity

Morbidity refers to the burden of illness, encompassing the number of cases, the severity of illness, and the impact on individuals' health, which is usually represented or estimated using prevalence or incidence (Hernandez & Kim, 2022). Prevalence measures the relative frequency of a specific symptom or characteristic within a population at a particular point in time, whereas incidence describes the rate at which individuals in a specific population

acquire a specific symptom or characteristic over a particular period of time (Noordzij et al., 2010).

Certain groups of the population such as children, the elderly, and individuals with underlying health conditions, are at higher risk of severe illness and complications from influenza. Children, especially those under the age of five, are disproportionately affected by influenza, accounting for a significant proportion of hospitalisations and deaths (Iuliano et al., 2018; Wang et al., 2020b). The elderly, particularly those over the age of 65, are also at increased risk of severe outcomes (Reed et al., 2015). Among children, asthma is the most prevalent underlying risk factor, while chronic obstructive pulmonary disease and other chronic lung conditions are commonly associated with hospitalisation among the elderly (Grohskopf et al., 2022). In addition, elderly individuals, pregnant women, and individuals with chronic medical conditions are particularly vulnerable to severe illness and complications (Jain et al., 2009). The full list of commodities associated with severe influenza-related complications defined by the US Advisory Committee on Immunization Practices can be found at (Grohskopf et al., 2022).

Mortality

Mortality, on the other hand, refers to the number of deaths resulting from the diseases under study, which can be communicated as a rate or an absolute number (Hernandez & Kim, 2022). Mortality is typically expressed as a death rate per 1000 individuals, thus it is calculated as the number of deaths within a specified time period and population by the total population, and then multiplied by 1000 to indicate the "per 1000" rate for the population under investigation (Hernandez & Kim, 2022). Influenza-related mortality varies each year, and the severity of influenza season in terms of mortality can range from mild to severe (Simonsen et al., 1997). Older individuals aged above 65 years have higher influenza-associated mortality rates (Reed et al., 2015). The impact of influenza on mortality is multifaceted and includes direct influenza-related deaths as well as indirect deaths attributed

to complications such as pneumonia, cardiovascular events, and exacerbation of underlying health conditions (Thompson et al., 2003).

Estimating influenza-related mortality is challenging due to the variability in reporting practices, coding inconsistencies, and the presence of comorbidities. Influenza mortality is traditionally estimated by the excess mortality method, primarily focused on developed countries in northern temperate regions, with some notable exceptions in tropical and sub-tropical populations (Yuan et al., 2021; Yang et al., 2015). Epidemiological studies employ statistical methods and surveillance data to estimate excess mortality during influenza seasons, taking into account baseline mortality rates and temporal associations with influenza activity (Simonsen et al., 2005; Thompson, Comanor & Shay, 2006). One commonly used model is the Serfling model, which calculates excess mortality by comparing observed deaths during the influenza season with the expected baseline mortality (Serfling, 1963). Other modelling approaches include time-series models, compartmental models, and statistical regression models that capture the complex dynamics of influenza transmission and its impact on morbidity and mortality (Yang et al., 2009; Viboud et al., 2006; Yang, Lipsitch & Shaman, 2015; Shaman & Karspeck, 2012).

1.2.3.2 Seasonality of Influenza

Influenza exhibits distinct seasonal patterns, with its occurrence varying across different seasons and regions worldwide. The occurrence of influenza follows a consistent pattern in many countries, with most temperate populations experiencing annual epidemics during their winter: influenza transmission in northern temperate regions is observed to typically increase and reach a peak between December and February, and gradually diminish between April and May (Bloom-Feshbach et al., 2013). For temperate regions in the southern hemisphere, the onset of influenza seasonal activity generally commences around June because of the reversed seasons (Kenah et al., 2011). However, the seasonality of influenza becomes

increasingly intricate in tropical and subtropical regions (Soebiyanto et al., 2014). Influenza activity can be more sustained throughout the year, with less pronounced seasonal peaks, exhibiting variability influenced by diverse climate factors and potential altitude (Azziz Baumgartner et al., 2012).

Accurately describing the seasonality of influenza allows decision-makers to respond to outbreaks efficiently, optimise vaccination strategies, and implement effective public health interventions (Axelsen et al., 2014; Chretien et al., 2014). However, the timing of seasonality of influenza varies between years and across different regions and can be affected by multiple factors, which makes it challenging for scientists to develop forecasting models that can be applied to most regions with accurate forecasts of key characteristics of the seasonality, such as the season onset and end times (Tamerius et al., 2011).

Climatic factors affect the transmission of influenza and are partly responsible for its seasonality, with scientific investigations primarily focussing on temperature and humidity. In temperate regions, lower temperatures and humidity during the winter months create favourable conditions for influenza virus survival and transmission (Lowen & Steel, 2014). In addition to climatic factors, indoor crowding, and reduced ventilation, in colder months contribute to the higher transmission of influenza viruses (Shaman et al., 2010). In tropical or subtropical regions, where the temperature is relatively more constant, other factors, such as rainfall patterns and humidity, are more likely to influence the seasonality of influenza (Tamerius et al., 2013).

Population density and human mixing patterns also contribute to the variation in patterns of influenza disease. High-density urban areas with increased social interactions facilitate the spread of influenza viruses, leading to more pronounced seasonal peaks (Lofgren Eric et al., 2007; Tamerius et al., 2013). School terms and holiday periods also influence influenza

transmission, as children are often key drivers of influenza outbreaks due to their higher susceptibility and increased social contact within schools (Cauchemez et al., 2014). Changes in population behaviours, such as increased travel during holidays or migration, can introduce new influenza strains into different regions, affecting the timing and intensity of seasonal outbreaks (Shaman et al., 2010).

1.3. SARS-CoV-2

As a global pandemic pathogen, SARS-CoV-2 not only triggered a public health crisis, but also reshaped global scientific research, prevention and control strategies, and social lives. The sudden emergence of this new pathogen reminds us that in the context of globalisation, we are facing the great challenges that emerging infectious diseases may bring.

In the past few decades, influenza has been the representative of respiratory infectious diseases, causing annual seasonal outbreaks and also requiring continuous strengthening of surveillance and vaccination. However, the unexpected rise of SARS-CoV-2 prompted an urgent re-examination of our understanding of respiratory viruses and their potential to cause a global pandemic. This major shift in the infectious disease landscape caused unprecedented attention from the public and encouraged researchers to re-prioritise research focus on pandemic preparedness.

Therefore, this section will delve into multiple aspects of SARS-CoV-2, covering its genetic characteristics, surveillance, and transmission dynamics, which provide prerequisite background information for the following chapters about regression analysis and forecasting of SARS-CoV-2.

1.3.1 Biology of SARS-CoV-2

1.3.1.1 Gene Structure of SARS-CoV-2

SARS-CoV-2, the virus that causes the disease COVID-19, belonging to the Coronaviridae family, which encompasses a diverse group of enveloped, positive-sense RNA viruses. This family is further classified into four genera: Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus. SARS-CoV-2 falls under the Betacoronavirus genus, along with other notable coronaviruses such as severe acute respiratory syndrome coronavirus (SARS-CoV-1) and Middle-East respiratory syndrome coronavirus (MERS-CoV) (Zhu et al., 2020; Drosten et al., 2003; Zaki et al., 2012).

SARS-CoV-2 virus displays morphological characteristic features of coronaviruses. It has a spherical shape with an average diameter ranging from 65 to 125 nanometres (nm) (Shereen et al., 2020). The viral envelope is derived from the host cell membrane and contains various viral structural proteins, including the spike (S) protein, envelope (E) protein, membrane (M) protein, and nucleocapsid (N) protein (Walls et al., 2020). These proteins are involved in viral entry, viral particle assembly, and immune evasion.

The gene structure of SARS-CoV-2 consists of a single-stranded, positive-sense RNA genome (Wu et al., 2020b). The structure of SARS-CoV-2 viral particles is complex and highly organised. The viral envelope, composed of a lipid bilayer, surrounds the viral nucleocapsid. The nucleocapsid consists of the vRNA genome tightly associated with N proteins, forming a helical structure. The S protein protrudes from the viral envelope and gives coronaviruses their distinctive crown-like appearance when observed under electron microscopy, hence the name "coronavirus" (Walls et al., 2020). The S protein, composed of S1 and S2 subunits, facilitates viral entry into host cells by binding to the angiotensin-converting enzyme 2 (ACE2) receptor. The S protein mediates attachment to the ACE2 receptor on target cells, facilitating viral fusion and entry into the host cell cytoplasm (Walls

et al., 2020). This interaction between the spike protein and the ACE2 receptor is a key determinant of the virus's tropism and pathogenesis.

Apart from the spike protein, other structural proteins, such as the envelope (E) and membrane (M) proteins, are involved in the assembly and release of viral particles (Schoeman & Fielding, 2019). The E protein participates in virus assembly, while the M protein is responsible for maintaining the shape and structure of the viral envelope (Walls et al., 2020). The N protein, which is tightly associated with the vRNA genome, is important for viral replication, packaging of the viral genome, and modulating host immune responses.

1.3.1.2 Viral Replication

The viral replication cycle of SARS-CoV-2 involves a series of intricate steps that enable the virus to enter host cells, replicate its genetic material, produce viral proteins, assemble new viral particles, and release them to infect neighbouring cells.

The replication cycle begins with SARS-CoV-2 attachment to the host cell surface. The S glycoprotein on the viral envelope interacts with the ACE2 receptor present on the host cell membrane (Walls et al., 2020). This attachment facilitates the subsequent fusion of the viral envelope with the host cell membrane. The vRNA genome is released from the viral particle when the virus enters the host cell, and the genome is used as a template for the translation of viral proteins and the replication of more copies of the whole genome. The replication process involves the synthesis of a complementary negative-sense RNA strand, which forms the template for the production of positive-sense RNA genomes. The resulting positive-sense RNA genomes can then be used for both viral protein synthesis and as templates for the replication of additional vRNA.

Transcription and translation are vital steps in the viral replication cycle. Viral mRNAs are produced by host cell machinery and then translated by ribosomes into viral proteins, including replicase proteins and structural proteins. Replicase proteins are essential for vRNA replication, while structural proteins are involved in the assembly of new viral particles. Assembly of new viral particles takes place in the host cell cytoplasm. The structural proteins, such as the S, E, M, and N proteins, interact with the vRNA genome to form the viral nucleocapsid.

The nucleocapsid, along with other viral components, is then assembled at the host cell membrane, forming new viral particles (Trionfetti, 2023). The release of viral progeny from infected host cells is accomplished through the budding process. Newly formed viral particles acquire their envelope by budding from the host cell membrane, incorporating viral structural proteins. These viruses can infect neighbouring cells and initiate subsequent rounds of infection or be expelled from the body along with saliva and mucus (Huang et al., 2021).

SARS-CoV-2 virus and influenza viruses share some similarities in the process of viral replication. Firstly, they both need to bind to the host cell receptor to enable the replication process to start (V'kovski et al., 2021; Skehel & Wiley, 2000). Secondly, both viruses need to release their RNA genome into the host cell, where it replicates and transcribes (Resa-Infante et al., 2011; Wu et al., 2020a).

Apart from similarities, there are also some significant differences between the SARS-CoV-2 virus and influenza viruses. The genome of the SARS-CoV-2 virus is relatively large and consists of a single-stranded positive-strand RNA (Wu et al., 2020b), while the genome of the influenza viruses comprises eight segments of negative-sense, single-stranded RNA (Lamb, 2008). Differences in the size and structure of these genomes affect their replication

mechanisms (Mahy, 2010). In addition, the SARS-CoV-2 virus and influenza viruses rely on different proteins during protein synthesis, therefore building differently structured virtual particles and leading to different assembly and release processes (Arya et al., 2021; Chutinimitkul et al., 2010).

1.3.1.3 Immune Evasion

SARS-CoV-2 adopts various strategies to evade or modulate the host's immune response, allowing it to establish infection and persist within the host before successfully transmitting onwards. This successful system enabled the virus to quickly disseminate across continents and affected numerous countries, leading to a global pandemic. In this section, we will describe the strategies the SARS-CoV-2 virus possessed that allowed it to evade the immune response.

Interferon (IFN) signalling is one evasion strategy employed by SARS-CoV-2. The virus has been shown to suppress the production of type I IFNs, which are crucial for the early antiviral response (Hadjadj et al., 2020). By inhibiting the production or signalling of these essential cytokines, SARS-CoV-2 can dampen the host’s innate immune response, thereby promoting viral replication and spread (Park & Iwasaki, 2020).

SARS-CoV-2 has also been found to modulate both innate and adaptive immune responses. The virus can suppress the activation and function of immune cells, such as dendritic cells and T cells, which play critical roles in initiating and orchestrating immune responses (Freeman et al., 2020). This modulation of immune responses can lead to impaired antigen presentation, reduced T cell activation, and impaired antibody responses, ultimately allowing the virus to evade immune surveillance and promote viral persistence.

Antigenic variation is another key mechanism utilised by SARS-CoV-2 to evade immune detection (Weisblum et al., 2020). The virus possesses a high mutation rate, leading to the emergence of new variants with altered antigenic properties. This antigenic variation allows the virus to escape recognition by pre-existing antibodies generated during previous infections or vaccination, thereby facilitating immune evasion (Starr et al., 2021). The emergence of SARS-CoV-2 variants, such as those of the Omicron variant and its sublineages, has raised concerns about their increased immune escape from prior infection or vaccination (Xia et al., 2022).

Furthermore, SARS-CoV-2 immune evasion may also contribute to disease severity, and more specifically, high variation in disease severity where many people have mild symptoms while some have much more severe symptoms. SARS-CoV-2 has been shown to induce an excessive and dysregulated immune response in some individuals, leading to a phenomenon known as a cytokine storm (Moore & June, 2020). This uncontrolled release of proinflammatory cytokines can cause tissue damage and severe disease progression (Tay et al., 2020). By inducing a dysregulated immune response, SARS-CoV-2 impairs immune cell function and evades immune-mediated clearance.

1.3.2 Surveillance

1.3.2.1 Case Definitions and Testing

The identification of COVID-19 cases is essential for effective disease surveillance, control, and patient management. Case definitions serve as standardised criteria to classify suspected or confirmed COVID-19 cases. The WHO has provided definitions of COVID-19 cases that guide healthcare professionals in identifying and reporting cases based on clinical presentation, exposure history, and laboratory confirmation (World Health Organization, 2020a). Suspected cases are individuals who exhibit symptoms consistent with COVID-19, such as fever, cough, and shortness of breath. During the initial stages of the pandemic, to

be a probable case it was also required that people had a history of potential exposure to the virus through travel to affected areas or contact with confirmed cases. Confirmed cases are individuals with laboratory evidence of SARS-CoV-2 infection, either by detecting viral RNA through molecular tests like reverse transcription-polymerase chain reaction (RT-PCR) or by antigen testing (World Health Organization, 2020a).

The most commonly used diagnostic test to detect SARS-CoV-2 was the reverse transcription-polymerase chain reaction (RT-PCR) assay. RT-PCR is a molecular test used to detect the presence of viral RNA in respiratory specimens (e.g., nasopharyngeal swabs) (Corman et al., 2020). RT-PCR provides highly accurate results among symptomatic individuals and can detect active infections even in asymptomatic individuals, therefore it remained a primary testing method for diagnosing COVID-19 for a substantial time during the acute phase of the pandemic (Centers for Disease Control and Prevention, 2020a).

Lateral flow antigen tests were developed after PCR tests. They are rapid diagnostic tests that detect specific viral proteins (antigens) from respiratory specimens, offering faster turnaround times compared to RT-PCR which can be achieved in point-of-care and home settings without laboratory equipment (Zhou et al., 2021). Antigen tests are valuable for mass screening and quick identification of potentially infectious individuals (Porte et al., 2020). However, they may have lower sensitivity, especially during the early stages of infection when viral loads are relatively low. We note that they are highly sensitive during periods of high viral load when transmission is much more likely to occur (Puhach, Meyer & Eckerle, 2023; Kirby et al., 2023).

In addition, serological tests, also known as antibody tests, detect the presence of antibodies produced by the immune system in response to SARS-CoV-2 infection. These tests are used to identify past infections and assess the prevalence of exposure in populations (Meyer

et al., 2020). However, serological tests are not suitable for diagnosing acute infections, as it takes several days to weeks for detectable antibodies to develop after infection (National Institutes of Health, 2023).

Despite the importance of diagnostic testing, several limitations with respect to the sensitivity and specificity of tests, accessibility and availability of test resources, and emerging variants must be considered. First of all, both molecular and antigen tests may yield false-negative results, especially during the early stages of infection or in individuals with low viral loads (Kucirka et al., 2020). False-positive results can also occur due to cross-reactivity with other coronaviruses or non-specific immune responses. However, the lack of a golden rule of sensitivity and specificity makes assessing the unbiased prevalence in the population difficult (Pouwels et al., 2021). Secondly, the availability of testing resources can vary across regions and countries, leading to disparities in testing capacity and access to testing. Individuals living in developing countries or more deprived regions have limited access to testing, which may result in underreporting of cases and hinder effective disease surveillance and control (Ahmad et al., 2020; Afriyie et al., 2020). Finally, the ongoing emergence of SARS-CoV-2 variants presents challenges for testing accuracy. Some mutations in the virus may affect the performance of diagnostic tests, potentially leading to false results (Yadav et al., 2022).

1.3.2.2 Global Surveillance Networks of SARS-CoV-2

Effective global surveillance networks play a pivotal role in monitoring and forecasting the spread of the SARS-CoV-2 virus. International surveillance networks and institutions including the Global Influenza Surveillance and Response System (GISRS), Global Outbreak Alert and Response Network (GOARN) and Centers for Disease Control and Prevention (CDC) have provided techniques to collect and analyse epidemiological data (World Health Organization, n.d.; Global Outbreak Alert and Response Network, n.d.; Centers for Disease Control and Prevention, 2020b). These networks pivoted from influenza

to facilitate data collection, analysis, and dissemination of information across international borders, enabling a timely and coordinated response to the evolving COVID-19 pandemic.

1.3.3 Epidemiology of SARS-CoV-2

The field of SARS-CoV-2 epidemiology has emerged as a vast and dynamic area of research that coincided with the duration of my PhD studies. The onset of the COVID-19 pandemic marked a transformative moment in global health, prompting an extensive surge of scholarly investigations into the epidemiological characteristics, transmission dynamics, clinical outcomes, and public health responses related to the virus. As a novel respiratory virus with far-reaching implications, SARS-CoV-2 swiftly captured the attention of researchers, policymakers, and healthcare professionals worldwide. This unprecedented interest led to the rapid production of a myriad of papers, spanning a diverse range of topics within the field.

The significance of studying the epidemiology of SARS-CoV-2 cannot be overstated. The virus's rapid transmission and wide-ranging impact on public health systems underscore the urgency of comprehending its patterns of spread, risk factors, and potential mitigation strategies. The evolution of this field has facilitated the development of real-time surveillance systems, predictive modelling tools, and evidence-based interventions to curb transmission and mitigate the consequences of the pandemic. Insights gained from the epidemiological research have informed critical decision-making processes, from public health policy formulation to resource allocation and outbreak response planning.

This surge in research activity is evidenced by a multitude of publications that have emerged during the ongoing pandemic. The diversity of topics explored includes viral transmission dynamics, identification of vulnerable populations, effectiveness of non-pharmaceutical interventions, genomic sequencing to trace viral variants, healthcare capacity planning,

vaccine distribution strategies, and the interplay between socio-demographic factors and disease outcomes. Here, the epidemiology of SARS-CoV-2 will be discussed from three perspectives: transmission, global patterns, and disease burden.

1.3.3.1 Transmission

SARS-CoV-2 spreads through respiratory droplets and aerosols generated when an infected individual talks, coughs, or sneezes. Via these droplets, virions can be directly inhaled by individuals in close proximity, leading to person-to-person transmission. In addition to respiratory droplets, SARS-CoV-2 can also spread via aerosols, which are smaller particles that can remain suspended in the air for extended periods, especially in poorly ventilated indoor spaces (Morawska & Milton, 2020). The apparent higher frequency of aerosols for SARS-CoV-2 compared with influenza means that airborne transmission is more important, particularly in enclosed settings such as crowded indoor spaces, healthcare facilities, and public transportation (Stadnytskyi et al., 2020; Bourouiba, 2020).

Contact with contaminated surfaces, also known as fomites, is another mode of transmission for SARS-CoV-2. The virus can survive on various surfaces, such as plastic, metal, and glass, for several hours to days, depending on environmental conditions (van Doremalen et al., 2020). When individuals touch these contaminated surfaces and subsequently touch their face, mouth, or eyes, they may introduce the virus into their respiratory system, leading to infection. While the primary mode of transmission is through respiratory droplets and aerosols, the potential for fomite transmission underscores the importance of hand hygiene and surface disinfection in mitigating the spread of the virus.

The virus can spread in various settings, each with unique implications for transmission dynamics. In households, close and prolonged contact between family members increases the risk of transmission. Household transmission has been shown to be a significant driver of

COVID-19 outbreaks, and individuals living in larger-sized households are at higher risk of acquiring the virus (Chadeau-Hyam et al., 2022c). In healthcare facilities, nosocomial transmission can occur due to the proximity of patients, healthcare workers, and visitors (Nishiura et al., 2020; Guo et al., 2020), which poses a challenge for infection control, especially in settings where personal protective equipment (PPE) and infection prevention measures may be inadequate (Conly et al., 2020; Hossain et al., 2021; Martin-Delgado et al., 2020; Mahmood et al., 2020).

Community gatherings have also been identified as sources of SARS-CoV-2 transmission. Echoing a key characteristic of SARS-CoV-1 transmission (Riley et al., 2003), large gatherings, such as weddings, religious ceremonies, and parties, can lead to super-spreader events, where a single infected individual can transmit the virus to a large number of people (Adam et al., 2020). The virus can also spread in workplaces, schools, and public transportation settings, where individuals interact closely and in confined spaces.

1.3.3.2 Global Patterns

SARS-CoV-2 has rapidly spread across the globe, causing a pandemic with highly significant health, social, and economic consequences. As of the present date, the global distribution and prevalence of SARS-CoV-2 infections continue to ebb and flow in a relatively asynchronous way, with high levels of cases, hospital admissions and deaths across different countries and regions.

The timeline of the COVID-19 cases reveals the rapidity with which SARS-CoV-2 spread globally. The first cases of COVID-19 were reported in December 2019 in Wuhan, China (Zhu et al., 2020). By the end of January 2020, confirmed COVID-19 cases were also reported in other countries, including the United States, Singapore, South Korea, Australia, France, and Germany, leading the World Health Organization (WHO) to declare a Public

Health Emergency of International Concern (PHEIC) on January 30, 2020 (World Health Organization, 2020b). Subsequently, the pandemic's geographical spread accelerated, with new cases emerging in Europe, the Americas, Africa, and other regions.

Up to the date of writing this chapter, COVID-19 has affected almost every country in the world. The United States, India, and Brazil have reported some of the highest numbers of confirmed cases and deaths (Johns Hopkins University & Medicine, 2020). However, when comparing the impact of the pandemic on different populations, it is important to consider the variations in testing capacities and reporting practices across different countries, as these factors can affect the accuracy of surveillance data. Middle- and low-income countries with limited testing capabilities may have higher underreporting rates, leading to a potential underestimation of the true prevalence of SARS-CoV-2 infections. The distribution and dynamics of the COVID-19 pandemic have also varied over time. Multiple waves of infections have occurred in different countries and regions, driven by factors such as the timing of the emergence of new variants, changes in public health interventions, and the levels of population immunity. Some countries, such as China, experienced early surges in cases followed by declines, while others (e.g., the UK) faced multiple waves with varying intensities (Yamey et al., 2020; Sutherland, Headicar & Delong, 2021).

1.3.3.3 Disease Burden and Mortality of SARS-CoV-2

The COVID-19 pandemic has placed an immense burden on healthcare systems worldwide. The surge in hospitalisations placed a significant strain on healthcare facilities. Studies have shown that advanced age is a critical risk factor for hospitalisation. Older adults, especially those aged 65 and above, are more likely to experience severe illness and require hospitalisation compared to younger age groups (Richardson et al., 2020; Cummings et al., 2020). Furthermore, individuals with underlying health conditions such as cardiovascular

diseases, diabetes, respiratory disorders, and immunosuppression have a higher likelihood of hospital admission (Guan et al., 2020a; Williamson et al., 2020).

Intensive care has been crucial for managing severe cases of COVID-19, particularly patients with acute respiratory distress syndrome (ARDS) and multi-organ failure. The demand for intensive care unit (ICU) beds and ventilators has been a major challenge during the pandemic. Studies have reported that ICU admissions are most common among older adults and individuals with comorbidities (Grasselli et al., 2020; Bhatraju et al., 2020). Moreover, critically ill patients often require prolonged stays in the ICU, leading to added strain on healthcare resources. Reducing the demand for intensive care to ensure that fatality rates did not increase was a major challenge for many healthcare systems during the early stages of the pandemic.

The high mortality rate caused by COVID-19, especially compared to influenza, is a key feature of its epidemiology. It varies across countries and regions due to differences in healthcare infrastructure, public health measures, and population demographics. Older age remains a prominent risk factor for death from COVID-19. Data from various studies show that mortality rates increase significantly with advancing age, with older individuals facing the highest risk of death (Verity et al., 2020; Levin et al., 2020). Comorbidities significantly contribute to the mortality risk associated with COVID-19. Individuals with pre-existing conditions such as hypertension, diabetes, obesity, chronic respiratory diseases, and cardiovascular diseases have experienced higher death rates (Zhou et al., 2020a; Docherty et al., 2020). The presence of multiple comorbidities further increases the risk of severe outcomes and mortality in COVID-19 patients (Ssentongo et al., 2020).

1.4. Community Surveillance

The surveillance systems discussed in previous sections were specifically influenza or SARS-CoV-2. Community surveillance for respiratory infectious disease is a key part of monitoring and controlling transmission within populations. Community surveillance involves the systematic collection, analysis and interpretation of epidemiological and clinical data from community sources to assess disease prevalence, incidence, trends and the impact of interventions It provides important insights into the dynamics of respiratory infectious diseases, helping to detect outbreaks early, guiding resource allocation and supporting public health decision-making. It is different to clinical surveillance because in community surveillance observations are not made only on individuals who present to a clinical setting.

A key feature of community surveillance is its ability to capture a broader and more representative picture of the burden of disease than traditional healthcare facility-based surveillance (Guerra, Acharya & Barnadas, 2019). By including individuals who may not seek medical care because of mild respiratory symptoms, community surveillance provides a more comprehensive understanding of the dynamics of disease transmission, as well as variation in symptom severity across populations. This approach is particularly applicable to respiratory infections with a range of clinical presentations, such as influenza, SARS-CoV-2 and other emerging pathogens.

A variety of data collection methods can be adopted for community surveillance to provide a comprehensive assessment of disease prevalence and transmission patterns. These data sources include sentinel surveillance systems, integrated surveillance systems, healthcare utilisation data and laboratory tests (Lombardo et al., 2003; Thacker & Stroup, 1994).
In addition, community-based surveys and studies provide valuable insights into prevalence, risk factors and transmission dynamics (Riley et al., 2020b). These surveys are particularly important in the case of new sources of infection, such as COVID-19, as they reveal the

extent of undetected infections and the impact that immunity may have on future transmission.

The advent of digital technology and social media has also helped integrate non-traditional data sources into community surveillance efforts (Blazquez & Domenech, 2018). Monitoring Internet search trends, social media conversations and other digital signals can provide additional information on disease awareness, public perceptions and symptom profiles, leading to a more complete understanding of disease dynamics (Aiello, Renson & Zivich, 2020; Fung, Tse & Fu, 2015).

Chapter 2 Infectious Diseases Forecasting

In this chapter, the relevant background literature for infectious disease forecasting will be provided, with a particular focus on influenza and COVID-19. Firstly, the background information of infectious disease forecasting in general with examples from influenza and COVID-19 will be described, followed by a brief discussion on specific challenges faced by those who forecast infectious diseases. In the next section, the development of forecasting models for influenza and COVID-19 will be reviewed, respectively. This involves classifying the main research categories and reviewing mathematical and statistical models and machine learning techniques. Finally, validation strategies and evaluation metrics commonly used for model performance will be discussed.

2.1 Background

2.1.1 Overview

A forecast refers to quantitative estimates about future unobserved events, conditional on events that have been observed in the past and present (Lauer, Brown & Reich, 2020). The most familiar forecasts are probably for daily weather and severe natural disasters, such as storms and floods, where accurate forecasts are crucial to saving lives and properties (Moller, 2001; Yu, Yang & Li, 2018; Morss, Demuth & Lazo, 2008). Likewise, infectious disease forecasting can provide significant support for decision-makers to take timely measures to prevent the spread of infectious diseases and allocate healthcare resources, as well as to reduce economic costs during epidemics.

When responding to a public health emergency, decision-makers often have to make reactive decisions rapidly under fast-changing circumstances with insufficient information, which may lead to bias (Glasser et al., 2011; Fischer et al., 2016; Wernstedt et al., 2019). Data updates from traditional surveillance systems are usually retrospective and time-

lagged, making them of only limited utility for planning and decision-making in real-time. In order to fill these gaps, forecasting -- estimating present events or those to be expected in the near future -- can be conducted to provide guidance on the utility, scale, and timing of prevention strategies (Fischer et al., 2016). For instance, during the COVID-19 pandemic, communication and implementation of prevention measures, including improving test capability and vaccinations, have been critical for deploying and managing public health resources (Mustafa et al., 2021).

2.1.2 Forecast Targets and Time Horizons

Forecast targets are the not-yet-known outcomes that we want to predict. Ideally, forecast targets should include quantitative definitions and be selected to meet specific public health objectives (Shaman, Yang & Kandula, 2014). These targets for COVID-19 forecasting are primarily the number of incident cases, hospitalisations and deaths in the near future. Similar metrics are used in influenza forecasting in addition to key seasonal features such as peak incidence, time of peak and time of season commencement (Biggerstaff et al., 2016; Brooks et al., 2015; Gao et al., 2020; Lin et al., 2021; Verity et al., 2020; Chen et al., 2020). For influenza, they aim to help stakeholders understand the characteristics of the current season relative to previous ones but also the prospective changes of the current season (Yang, Lipsitch & Shaman, 2015; Yang et al., 2015; Zhang et al., 2017; Biggerstaff et al., 2016). A number of different targets are used. For instance, the FluSight forecast program defines them based on the Outpatient Influenza-like Illness Surveillance Network (ILINet) percentage of visits for influenza-like illness (ILI), weighted by state population (CDC Epidemic Prediction Initiative, 2016). Baseline ILI was established by computing the mean percentage of ILI visits during non-influenza weeks for the past three seasons and using two standard deviations (Lutz et al., 2019). However, since the COVID-19 pandemic, these periods have needed to be adjusted to the most recent three seasons excluding the period of the COVID-

19 pandemic (Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD), 2022).

The ability to predict a broader range of targets, especially if some of these targets will allow a greater level of forecast accuracy and reliability, will enhance the public's confidence in modelling and better meet the needs of stakeholders. Forecast targets are usually expressed and predicted as continuous variables (Hyndman & Athanasopoulos, 2021). However, simpler and more interpretable targets that still be able to deliver useful information should be considered as an alternative. The ordinal target is one example that still keeps the ordering relations of indicators such as cases, deaths or hospitalisations such that it can effectively predict the trend will increase rapidly, grow moderately, remain constant, decline moderately, or decline rapidly in the future period (Buri et al., 2020; Ten Have, 1996; Armstrong & Sloan, 1989; McCullagh, 1980).

Another consideration for infectious disease forecasting is the forecast horizon, the time into the future about which predictions are being made, which can vary from 1-week ahead to various longer periods (Soyiri & Reidpath, 2013; Myers et al., 2000). For example, if our current week is week 40 (t = 40), then a (t+1)-step ahead forecast refers to the targets (e, g., incidence) in week 41, which is equivalent to 1-week ahead forecast. For FluSight, short-term targets are forecasts of the weighted ILI percentage 1- to 4-week ahead of its publication (Lutz et al., 2019). Regardless of the model type, multiple recent infectious disease forecasting studies have focused on short time scales (e.g., weekly or monthly) (Meakin et al., 2022; Birrell et al., 2011; Reich et al., 2016a; Shaman et al., 2013).

2.1.3 Nowcasting

Nowcasting, or “predicting the present”, a term originally coined in meteorology, has gained significant traction in various fields, including epidemiology (Mass, 2012). This concept refers

to the estimation of the current state of an event or a phenomenon based on real-time data, as opposed to forecasting which predicts future events (Mass, 2012). While forecasting relies on historical data to project future occurrences, nowcasting focuses on the immediate present, often using the most recent data to assess current conditions (McGough et al., 2020).

In the field of infectious diseases epidemiology, nowcasting is essential for accurately comprehending the present status of disease transmission and the immediate demands of healthcare systems, particularly in the context of quickly changing scenarios such as the COVID-19 pandemic. Unlike traditional forecasting, which might estimate the number of cases or the spread of infectious diseases in the coming weeks or months, nowcasting provides an up-to-date picture of epidemiological situations potentially including infection rates, hospitalisation needs, and the effectiveness of public health interventions, which is a pivotal aspect of effective public health response and resource allocation (McGough et al., 2020; Du et al., 2023; Shaman et al., 2013). In particular, nowcasting is often used to estimate the total number of people who are currently infectious, which is a measure of individual risk. For example, the number of infectious individuals during the exponential phase of a rapidly growing epidemic can be far higher than the current case count once reporting lags and testing processes are accounted for (UKHSA, n.d.; Elliott et al., 2022a).

Nowcasting also complements traditional forecasting by filling in the gap between the occurrence of an event and the availability of consolidated data. For instance, a study on nowcasting by Bayesian Smoothing (NobBS) demonstrates the application of this method in real-time epidemic tracking (McGough et al., 2020). It has been implemented in different disease surveillance contexts, such as dengue in Puerto Rico and ILI in the United States. This approach helps in estimating total case numbers using incomplete, time-stamped reported case data. The model learns from historical information on cases reported at multiple delays and uses this relationship to predict the number of not-yet-reported cases in

the present. Another application of nowcasting in epidemiology is in the monitoring of seasonal diseases like influenza. Traditional surveillance systems often report data with a lag, which can delay public health responses. Nowcasting, using real-time health reports and digital data sources like internet search trends, can provide a more current picture of disease activity, allowing for quicker responses to outbreaks (Shaman et al., 2013).

The methodology behind nowcasting involves the use of statistical models and algorithms that can handle and interpret incomplete or near-real-time data. Techniques such as Bayesian inference models and machine learning algorithms have been employed to develop nowcasting models in epidemiology (Babii, Ghysels & Striaukas, 2022; McGough et al., 2020; Sahai et al., 2022). These models can quickly assimilate new data as it becomes available, updating their estimates accordingly.

In conclusion, nowcasting constitutes a significant advancement in epidemiological research. It delivers a real-time overview of disease proliferation and current healthcare demands, thus furnishing prompt and crucial insights that enhance and augment the capabilities of traditional forecasting methods. As real-time data collection and processing technologies continue to improve, the role of nowcasting in public health is likely to become even more prominent, providing an essential tool for rapid response in the face of emerging health threats.

2.1.4 Interactions between Surveillance and Forecasting

Influenza surveillance and influenza forecasting are closely interconnected and mutually beneficial in the realm of infectious disease monitoring and control. Surveillance systems provide historical and real-time data on influenza activity, including the number of reported or laboratory-confirmed cases, hospitalisations, or deaths, geographic distribution, and circulating strains. Forecasting systems are usually designed to predict the outputs from

those same surveillance systems, employing statistical, mathematical and machine learning techniques to build influenza forecasting models to predict future trends in influenza transmission (Shaman & Karspeck, 2012).

Different types of surveillance data are used as forecast targets and can also be used as predictors for different other targets. Forecasting models can incorporate clinical surveillance data (e.g., the number of influenza cases, hospital admissions and deaths) along with external predictors to identify patterns and predict the future trends of infectious disease activity. Summary variables from diagnostic data, particularly information on dominant circulating influenza strains and their characteristics, can also help inform the modelling of virus transmission dynamics and the potential impact of antiviral resistance (Biggerstaff et al., 2018). Additionally, virological surveillance data, including genetic sequencing and antigenic characterization of circulating influenza strains, help identify emerging variants and monitor antigenic drift, which are important inputs for forecasting models that incorporate strain-specific information (Reich et al., 2019a).

2.1.5 Scenario Analysis versus Forecasting

Forecasting is sometimes confused with another method that often uses mechanistic models of disease transmission in the decision-making and planning process called scenario analysis. There is sometimes a lack of distinction between them as they both involve predicting future outcomes.

Scenario analysis aims to assess the effects of potential future events on the performance of a system, often where events are alternative policy choices (van der Heijden, 2005). It involves considering various alternative outcomes, referred to as scenarios, and presenting the different achieved outcomes (van der Heijden, 2005). Scenario analysis encompasses the prediction of the anticipated value of a performance indicator within a specified time

frame, considering diverse situations and associated modifications in system parameter values under an uncertain environment (Balaman, 2019; Wright & Goodwin, 2009). Scenario analysis can be applied to estimate the changes in infectious disease dynamics under different theoretical best-case (optimistic) or worst-case (pessimistic) scenarios, which is useful for governments to adjust control measures accordingly and mitigate transmissions (Padmanabhan et al., 2021; van Genugten, Heijnen & Jager, 2003; Napoli et al., 2015). During the COVID-19 pandemic, for instance, scenario analysis was employed to understand the potential impacts of different policy decisions or public health interventions (Howerton et al., 2023; Cheng et al., 2023; Martin-Moreno et al., 2022).

In contrast, forecasting is designed to be the best estimate of a specific quantitative outcome and should have well-characterised uncertainty. It usually employs quantitative inputs and methodologies to predict future events or trends, primarily through the analysis and interpretation of historical data (Makridakis, Spiliotis & Assimakopoulos, 2018; Hyndman & Athanasopoulos, 2021). It focuses on generating estimates of the future based on patterns and relationships observed in past data (Scott Armstrong, 2011). For example, in the context of infectious disease epidemiology, forecasting aims to anticipate the intensity and magnitude of the spread of a disease by using statistical and mathematical models to analyse past trends and project these into the future (Lutz et al., 2019). However, the focus of forecasting is typically on determining the most probable or expected outcome, rather than exploring a range of possibilities (Chatfield, 2013).

Given that both are projections of the future, there is some overlap between forecasting and scenario analysis: forecasting can estimate future outcomes under various scenarios. However, while forecasting often feeds into scenario analysis as a component, providing estimates for individual scenarios, it does not encompass the broad range of possibilities that scenario analysis does. In this respect, forecasting could be seen as a specific type of scenario within scenario modelling - often representing the 'most likely' scenario. Ultimately,

scientists cannot completely redefine two commonly used words with finely nuanced differences: the public will always consider them both to be a form of prediction, and both will be assessed for accuracy to some degree.

2.1.6 Challenges

In the past decades, a growing amount of academic research has been focused on forecasting infectious diseases, and models have proven successful in making predictions for pathogens such as influenza (Lutz et al., 2019; Chretien et al., 2014), Zika (Kobres et al., 2019), and Ebola (Chretien, Riley & George, 2015). The development of various modelling methods, including mathematical and statistical models, machine learning and deep learning techniques, provides opportunities to reduce the impacts of infectious diseases by capturing early warning signs and shortening the time between an outbreak or wave's onset and its control. However, the practical and statistical challenges posed by infectious disease forecasting remain to be addressed.

2.1.6.1 Data Quality

Effective collection and preservation of reliable and up-to-date data are critical to analytics and efficient management during public health emergencies (Jain et al., 2020). However, in many cases, data are incomplete, noisy, or subject to reporting biases, which remain significant obstacles to improving the accuracy of model forecasts (Kesper, Wenz & Taentzer, 2020; Galhardas & Oliveira, 2005).

Although researchers have focused on improving the quality of models, there has sometimes been limited progress in improving the quality of data. Unlike the thousands of weather stations all over the world that collect continuous large-scale data for weather forecasting (Lynch, 2008), this rich and highly accurate data stream is not present in the field of infectious diseases (Moran et al., 2016). In epidemiological surveillance, data is collected

initially from varied sources, including patient self-reporting and interviews, surveys and questionnaires, and laboratory tests, and then aggregated to be available to researchers (Jajosky & Groseclose, 2004). Even for official systems whose published data are recognised as benchmark data, they often are unable to capture and report all cases without significant delays between initial reporting and aggregation, which prevents researchers from providing timely and accurate forecasting support. Therefore, there is a need for novel survey methods and their data streams to facilitate the development and implementation of increasingly refined infectious disease forecasting models.

2.1.6.2 Visualisation of Forecasting Results

For forecasting results to be of value, they need to be communicated effectively to allow stakeholders (such as policymakers in the public health sector, healthcare suppliers and the media) to react quickly upon correct understanding and interpretations (Lutz et al., 2019). Applying infectious disease forecasting to public health is challenging since most of the forecasts are not standardised, which makes communication with audiences outside the field of epidemiology difficult.

One method to convey information more intuitively to the relevant audience is to use interactive visualisation tools (Viziflu, 2018). Interactive elements can also be valuable, allowing users to engage with the data actively, explore different aspects, and gain a deeper understanding of the material (Weissgerber et al., 2016; Perkel, 2018). While numerous data visualisation tools have been developed,, they often require custom coding and specialised knowledge, posing a challenge for model developers who may lack the time or design expertise (Wilke, 2019; Midway, 2020)(Wilke, 2019). This gap can result in visualisations that are not user-friendly, leading to potential misinterpretations in policy and decision-making contexts. To address this, interdisciplinary collaboration becomes essential. Collaborative efforts between epidemiologists, data scientists, and visualisation experts can

lead to the development of visual tools that are not only scientifically accurate but also intuitively understandable, thereby facilitating better decision-making by policymakers and public health officials (Park et al., 2022).

Recent forecasting studies have shown some consistency in the visualisation of their results. Figures providing clear and coherent visual comparisons of historical data with predictive outcomes are normally included (Chowell et al., 2022; Tabataba et al., 2017; Souza et al., 2023; Bracher et al., 2021; Reich et al., 2019b). Another key commonplace is the communication of uncertainty. Elements such as probabilistic models and confidence intervals are shown via visualisations, providing a more accurate picture of the forecasting results to help stakeholders understand the range of potential outcomes (Joslyn & Savelli, 2021; Osthus et al., 2019; Generous et al., 2014). Furthermore, within individual forecasting studies, consistent reporting styles are adhered to by researchers to ensure coherence and clarity. The layout of the visualisation is organised to guide the viewer's attention naturally through the data, prioritising important information and maintaining a logical flow.

Notable efforts in this area include the CDC's initiatives to refine communication around seasonal influenza forecasts, which have involved improving the clarity and accessibility of visual information (CDC, 2023; Viziflu, 2018). In addition, a network of Forecasting Centres of Excellence is being established to improve the accuracy and communication of seasonal and pandemic influenza forecasts in real-time (Gov, 2019). These efforts may contribute to the identification of optimal practices for visualising forecasting results as well as provide a framework for developing communications and visualisation in the future.

2.2 Evaluation

2.2.1 Model Training and Testing

For a predictive model to have practical value on real-world problems for policymakers, it is required to demonstrate its generalisability to data that were not used during the process of model fitting. Models that achieve excellent forecasts on in-sample data may still not perform well on out-of-sample data (Pant & Starbuck, 1990; Bartolomei & Sweet, 1989). In-sample data refers to a training set of data used to fit a model, while out-of-sample data is known a test set of data that have not previously been used with the model. Although it is not possible to be sure how well a model will work on each new dataset, careful use of out-of-sample data ensures a higher likelihood of generalisability.

The essential challenge in forecasting is to train a model in a manner that minimises error when the model is applied to the test set. In the case of real-time forecasts, the test set data remains unobserved at the time of model fitting. However, when researchers develop predictive models, historical data are normally used, which means that test set data actually have been observed. Although waiting and using the real-time assessment is the gold standard way to ascertain the model's performance, it takes time to wait for reliable results which could be a significant limitation in emergencies. Therefore, tests based on already-gathered but out-of-sample data are widely used to ensure the credibility and applicability of findings in forecasting research.

An out-of-sample study starts with data partitioning, dividing historical data series into a training set which is used to estimate and select a model and a test set which is retained to evaluate the model's predictive performance. According to (Tashman, 2000), two procedures can be chosen for the out-of-sample test for time series data - fixed-origin evaluation and rolling-origin evaluation. A detailed schematic representation of the rolling-origin procedure is

illustrated by (Schnaars, 1986). Here I will briefly describe their definition. Let T denote the final time point of the training set, and N denote the forecasting horizon N-step ahead.

Fixed-origin evaluation is also known as single forecasting origin evaluation. Commencing from the origin time point T, we generate forecasts for subsequent time periods T+1, …, T+N. By deducting each of these forecasts from the observed data values of the test period, we derive the forecast errors. These errors can be aggregated using different methods to obtain overall summary statistics for the model. This method will be used in Chapter 3.

While using rolling-origin evaluation, we iteratively update the forecasting origin and generate forecasts from each updated origin. With each update of the forecasting origin, a new observation is incorporated into the training period. In some studies, researchers have opted to maintain a consistent-sized training period by removing the earliest observation during each update (Callen et al., 1996; Swanson & White, 1997). Rolling-origin evaluation is more frequently recommended than fixed-origin evaluation as it enables more efficient rules for series splitting, and mitigates the impact of exceptional events occurring at any particular origin, making the errors less sensitive to such occurrences (Tashman, 2000). Rolling-origin evaluation will be used in both Chapters 3 and 6 later.

The comprehensive review and examination of out-of-sample tests to assess the accuracy of forecasting can be found in (Tashman, 2000).

2.2.2 Accuracy Metrics

As with any forecast, it is critical to measure the accuracy of infectious disease forecasts so that users have sufficient context to interpret results. As Out-of-sample tests have been discussed in the previous section, here I will discuss different types of accuracy metrics that could be used in the test. Extensive literature exists on how to score and evaluate forecasts

of all types, and the evaluation for point forecasts will be the focus of this section. There is no doubt that the choice of appropriate metrics depends on how the forecast is going to be used.

Error is a common focus of metrics when assessing accuracy, which is the difference between the predicted value and the actual value (Koutsandreas et al., 2022). Let y_t denote the actual value observed at time t, and y_t is the predicted value for the same time, then the forecast error et is defined as

$$e_t = y_t - \hat{y}_t$$

Generally, accuracy metrics are classified into five categories, scale-dependent measures, measures based on percentage errors, measures based on relative errors, relative measures and scaled errors (Hyndman & Koehler, 2006). Table 3.2 summarises the representative statistical accuracy measures for each category:

Table 2.1 Summary of statistical accuracy metrics.

Acronyms	Name	Category
MSE	Mean Square Error	Scale-dependent
RMSE	Root Mean Square Error	Scale-dependent
MAE	Mean Absolute Error	Scale-dependent
MAPE	Mean Absolute Percentage Error	Based on percentage errors
RMSPE	Root Mean Square Percentage Error	Based on percentage errors
MRAE	Mean Relative Absolute Error	Based on relative errors
rMAE	Relative Mean Absolute Error	Relative measures
MASE	Mean Absolute Scaled Error	Scaled errors

Mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) are scale-dependent metrics typically used to compare and evaluate different models applied to the same dataset. Nevertheless, these scale-dependent metrics are not suitable for comparing the forecast performance across different datasets. The scale-independent metrics based on percentage errors and based on relative errors such as mean absolute

percentage error (MAPE) and mean relative absolute error (MARE) are therefore used when comparing model performance across different data sets. However, they have the limitation of being undefined or infinite if $y_t = 0$ for any time t, since the percentage error is given by $percentage\ error_t = \frac{e_t}{y_t} \times 100\%$ (Hyndman & Koehler, 2006).

To facilitate a comparative evaluation of point forecasts in practical applications, relative metrics that scale the forecasting error relative to that of a benchmark model are suggested by many researchers (Reich et al., 2016b; Hyndman & Koehler, 2006; Gneiting & Raftery, 2007). Relative mean absolute error (rMAE) is one widely used example, which involves dividing the MAE of one forecasting model by the benchmark MAE:

$$rMAE = \frac{MAE}{MAE^*}$$

Where MAE* denotes the MAE obtained from the benchmark model. Similarly, RMSE, MAPE, RMSPE, etc can be formulated into relative measures in this way.

2.3 Models Used for Influenza and COVID-19 Forecasting

Mathematical models have been widely applied to infectious disease data to gain insights into transmission dynamics, with the aim of identifying strategies for prevention and control of the spread of diseases. Many of these approaches have been adopted to develop forecasting models for influenza outbreaks and, more recently during the COVID-19 pandemic.

Major modelling approaches mainly fall into three categories:1) time-series statistical models, 2) mechanistic models, and 3) machine learning (ML) / deep learning (DL) models. In this section, I aim to discuss and assess the difference between the major modelling approaches in these three categories that have been used to generate forecasts.

2.3.1 Time-series Statistical Models

Statistical models have also played a critical role in forecasting the spread and severity of infectious diseases. One type of statistical model that has become increasingly popular for influenza and COVID-19 forecasting is time series models which assume that future values can be predicted based on past observations, due to their ability to capture temporal dependencies in the data. Among time-series models, the AutoRegressive Integrated Moving Average (ARIMA) model is typically used. ARIMA is a linear combination of the differenced autoregressive (AR) model with a moving average (MA) model, where AR represents a stochastic variable that is regressed against its own lagged values, while MA refers to a weighted moving average of lagged white noise (Box et al., 2015). For non-stationary time-series data, the component "Integrated" (I) indicates the number of times that the original values are differenced (Box et al., 2015). ARIMA is normally expressed as:

$$y'_t = c + \emptyset_1 y'_{t-1} + \emptyset_2 y'_{t-2} + \ \ldots + \emptyset_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

, where y'_t is the differenced series and the differencing process can be repeated multiple times.

The above equation of ARIMA can be written as ARIMA (*p, d, q*), where:

p: is the number of lags for past values in the AR part,

d: is the number of differences needed to achieve stationary, and

q: is the number of lags for the forecast errors in the MA part.

Some special cases of ARIMA models arise with different choices of parameters, which are summarised in Table 2.1. When there is reason to believe that a seasonal effect exists, as is often the case with influenza, a seasonal ARIMA (sARIMA) model (Box et al., 2015) may be utilised to account for the seasonality in a time series.

2.3.2 Mechanistic Models

Most mechanistic models can be classified as either compartmental or agent-based models, which will be introduced in this section respectively.

2.3.2.1 Compartmental models

Compartmental models are a frequently used type of mechanistic model of infectious disease that can be used for forecasting (Tang et al., 2020). Individuals within a population can be allocated to distinct compartments, with each compartment representing a specific stage of the epidemic (Perra & Gonçalves, 2015). Compartmental models can be analysed using a deterministic framework or a sotchastic framework. If simulation is used to generate a random sequence of events that update the compartments, then that is considered to be a stochastic framework. For the deterministic framework, the quantity of newly infected individuals is calculated deterministically for a specific count of susceptible and infectious individuals. Conversely, in a stochastic model, the number of new infections varies, as it is determined by a random distribution. This variability reflects the inherent uncertainty and randomness in the process of disease transmission (Chang et al., 2014).

As an example, the Susceptible-infectious-recovered (SIR) model proposed by W. O. Kermack and A. G. McKendrick (Kermack, McKendrick & Walker, 1927) nearly a hundred years ago, has been considered one of the fundamental compartmental models in epidemiology. The SIR model describes the dynamic of individuals in the population transit from the susceptible stage to the infectious and finally to recovery (or deaths) over the process of an epidemic (Figure 2.1) (Tang et al., 2020). It holds significant importance in infectious disease modelling since its flexibility can be extended to incorporate more complicated scenarios. One variation of the SIR model is the Susceptible-exposed-infectious-removed (SEIR) model, in which an additional stage "Exposed" is introduced for individuals who may have been exposed to the virus but cannot yet be confirmed as infected

(Figure 2.2) (Tang et al., 2020). The SEIR can be further expanded by incorporating an intervention factor to consider the impact of external interventions on infection rates.



Figure 2.1 Susceptible-infected-recovered (SIR) model.

Individuals are divided into three compartments according to their disease status: susceptible to being infected, infected, infectious or have recovered from the infectious stage. β and γ are transition rates between compartments, where β represents the infection rate of individuals moving from the susceptible stage to infected stage, and γ represents recovery rate that infected individuals move to the recovered stage.

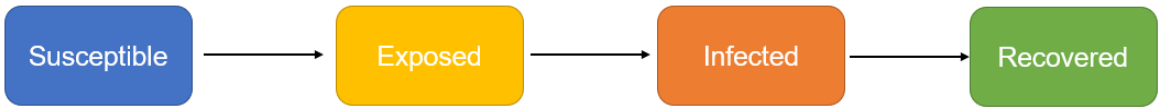


Figure 2.2 Susceptible-exposed-infectious-removed (SEIR) model.

The exposure compartment is added in the SEIR model (Figure 2.1), which refers to an incubation period during which individuals have contracted the disease but are not infectious themselves.

Compartmental models in epidemiology, traditionally depicted through continuous frameworks, can also be effectively utilised in a discrete form, both in value/state and time. This discrete approach is particularly relevant in situations where epidemiological data is collected at regular intervals or when disease dynamics exhibit distinct stages. For instance, in the study of seasonal influenza, data is often reported on a weekly basis, making the discrete discrete-time SIR model more applicable (Lin, 2008). Discrete SIR models calculate transitions between states at specific time points, aligning closely with the rhythm of data collection and the natural progression of the disease or with clear incubation periods or when interventions, such as vaccinations, are implemented at specific times (Brauer, Castillo-

Chavez & Feng, n.d.). This is notably different from continuous models, which assume a constant, uninterrupted flow of changes.

Deterministic compartment models are frequently used in studying infectious diseases within large populations, where the randomness inherent in individual differences tends to average out, thus the robustness of approximation for observed data remains. Although deterministic compartment models may fail to capture the different contact patterns in subpopulations that exist in reality, they can mimic the behaviour of an outbreak and flexibility in combining with other approaches (e.g., particle filtering) to improve forecast accuracy is attractive to modellers (Varghese et al., 2021a; Sturniolo et al., 2021; Safarishahrbijari et al., 2017). They have provided many successful applications in the analysis and prediction of key epidemic features for influenza outbreaks (Birrell et al., 2011; Shaman et al., 2013; Osthus et al., 2017) and the COVID-19 pandemic (Wu, Leung & Leung, 2020; Giordano et al., 2020; Samui, Mondal & Khajanchi, 2020; Gatto et al., 2020; Kucharski et al., 2020b) in different countries.

Nonetheless, in smaller populations, this randomness has a more significant impact on the transmission process, rendering deterministic compartment models less appropriate (Zhang et al., 2021; Daley & Gani, 1999; Lin, 2008). Thus, the significance of stochastic compartmental epidemic models which account for randomness in analysing infectious disease dynamics is widely acknowledged (Ullah et al., 2023). Stochastic compartment models integrate random variations to mirror the unpredictability observed in real-world infectious disease spread. They can simulate a range of possible scenarios, each with a certain probability, which is useful for analysing the dynamics of disease transmission under varying conditions and for planning responsive public health strategies. This stochastic approach allows for the integration of random variations and probabilities, reflecting the unpredictability and variability inherent in the spread of infectious diseases. For instance, in a stochastic model, the chance of infection from an infectious individual to a susceptible one

is not a fixed rate but a probability that can vary, accounting for factors like individual interactions and environmental influences (Ming et al., 2016).

Overall, compartmental models are useful approaches for understanding and interpreting epidemiological information, especially when it is required to take into account external factors, such as governmental control measures, of the epidemic being investigated to reflect the impacts.

2.3.2.2 Agent-based Model (ABM)

Agent-based model (ABM) is an alternative to compartmental models, representing individuals explicitly and simulating the dynamic of disease transmission based on individual interactions (Panovska-Griffiths et al., 2021). ABMs also require disease stages to be defined. However, unlike compartmental models in which a group of individuals in a population is assigned to one compartment at each stage, compartments used in ABM only represent disease states for an individual at each time (Dorjee et al., 2013). ABMs consider all individuals and their interactions in each disease state for the step of the simulation. Hence, ABMs have the advantage of accounting for heterogeneous individual behaviours, such as the presence of "super spreading" who exhibit a higher-than-average contact rate and contribute significantly to disease transmission (Dorjee et al., 2013; Lau et al., 2020), and assessing the impact of interventions like contact tracing and social distancing (Aleta et al., 2020; Kretzschmar et al., 2020; Kucharski et al., 2020). ABMs remain a popular method among researchers in forecasting the transmission of influenza and COVID-19 (Halloran et al., 2008; Rao, Chernyakhovsky & Rao, 2009; Cooley et al., 2011; Cocucci et al., 2022; Kumaresan et al., 2023; Castro, Reis & Salles, 2022; Reveil & Chen, 2022). However, usually, ABMs are far more computationally intensive than compartmental models.

2.2.3 Machine Learning (ML) and Deep Learning (DL) Models

Machine learning (ML) and deep learning (DL) algorithms are increasingly being employed for forecasting purposes in public health and medical research (Santosh & Gaur, 2021; Delpino et al., 2022). They have the potential to aid in the control of infectious diseases by facilitating the spatial and temporal prediction of the development and spread of the virus (World Health Organization, 2021). Models based on ML/DL algorithms are capable of detecting complex patterns within data of high dimensionality (LeCun, Bengio & Hinton, 2015). Compared to mechanistic and statistical models, models based on ML/DL are driven by data and do not make prior assumptions on data or parameters. They potentially provide greater flexibility when it comes to capturing the impact of potential external features. This makes them highly suitable for infectious disease forecasting, which frequently encompasses diverse factors such as meteorological factors and population behaviours.

Various ML/DL algorithms including support vector machine (SVM), random forest (RF) and long short-term memory (LSTM) networks have been applied to different types of data sets (e.g., surveillance data (Al-Omari et al., 2019), genomic sequencing data (Chen & Ishwaran, 2012), and social media (Athanasiou et al., 2023) and internet data (Ayyoubzadeh et al., 2020; Dai & Han, 2023)) to predict patterns for infectious disease data (Galasso, Cao & Hochberg, 2022; Rajaraman et al., 2018; Soebiyanto, Adimi & Kiang, 2010; Guefrechi et al., 2021). The findings of these studies have demonstrated that ML/DL algorithms can offer accurate predictions and achieve comparable or even higher accuracy compared to traditional epidemic models and/or statistical models.

2.2.3.1 ML/DL in Influenza/influenza-like Illness (ILI)

Multiple studies have incorporated ML/DL algorithms to track influenza//influenza-like illness (ILI) activity and predict their future trends. Support vector machines (SVM) are employed (Kim & Ahn, 2019) to analyse news articles to predict weekly ILI patient ratio, and (Nsoesie

et al., 2021) have demonstrated that SVM and random forest (RF) outperform ARIMA and multivariable linear regression in forecasting ILI trends based on digital data (Google search data in this study) at the city level. Google search data was also utilised to predict ILI in general outpatient clinics (GOPC) in Hong Kong (Xu et al., 2017). The authors compared the predictive performances of Deep learning (DL) with feedforward neural networks (FNN) (DL with FNN) with other three models (generalised linear model (GLM), least absolute shrinkage and selection operator (LASSO), and ARIMA), and DL with FNN the lowest RMSE, MAPE and MAE. The long short-term memory (LSTM) model was applied to Google Flu Trends (GFT) and CDC influenza data, and improved accuracy in predicting trends of influenza epidemics up to 5 weeks ahead compared to the traditional time-series models (Venna et al., 2019). Additionally, a new DL model framework called 'CNNRNN-Res' was proposed by (Wu, 2018), which combines convolutional neural network (CNN), recurrent neural network (RNN) and residual links and applies it to real-world ILI datasets. Their results showed consistent improvements in the accuracy of CNNRNN-Res compared to representative linear (autoregression (AR) and vector autoregression (VAR)) and non-linear techniques (gaussian process regression (GPR)).

2.2.3.2 ML/DL in COVID-19

Since the emergence of COVID-19 was recognised as a global threat, various ML/DL algorithm-based predictive models have been developed to forecast the spread of the pandemic from different perspectives. Accurate predictions of the spread of COVID-19 including new cases, mortality and hospital admissions can provide critical information to improve the effective coordination and administration of patient care across the entire healthcare system. Therefore, understanding the temporal trends of COVID-19 and improving models' predictivity in the form of a time series is an important topic that I discuss here.

A variety of ML/DL algorithms have been employed to develop forecast systems to predict COVID-19 cases in the near future. A real-time prediction model based on Artificial Neural Network (ANN) was proposed by (Wieczorek et al., 2020), with geo-location and numerical data globally from the preceding 2 weeks as input training data to predict daily new COVID-19 cases. The model showed high accuracy in 20 countries or regions, even reaching 100% for some geographies (Wieczorek et al., 2020). eVision is an epidemic forecaster that integrates LSTM along with search engine statistics, which was designed to improve the prediction accuracy of the weekly number of cases of communicable diseases (Shaghaghi et al., 2021). It achieved 89% accuracy when it was applied to predict the weekly number of COVID-19 cases in the United States. Adaptive neuro-fuzzy inference system (ANFIS), a type of ANN which is capable of modelling complex time and space patterns, is popular among academic researchers and has been used to predict daily confirmed COVID-19 cases in China, Italy, USA, Iran, Korea and India (Al-Qaness et al., 2020a, 2020b; Saif, Das & Biswas, 2021; Ardabili et al., 2020). Another ANN model called Modified Auto-Encoder was adopted by (Distante et al., 2020) to predict the epidemic curve of daily confirmed COVID-19 cases in 21 regions in Italy. Four ML and DL algorithms, support vector regression (SVR), polynomial regression, standard deep neural network (DNN), and LSTM were compared on the prediction of the number of COVID-19 cases in the future 10 days across the world, and the results demonstrate that PR produced the lowest root mean square error (RMSE) compared to other methods for predicting the transmission of COVID-19 (Punn, Sonbhadra & Agarwal, 2020).

Some studies proposed hybrid prediction models to improve prediction accuracy by incorporating DL algorithms and traditional epidemiological or time-series models. For example, (Zhang & Liu, 2021) embedded LSTM into the susceptible-infected-quarantined-recovered (SIQR) model to predict confirmed and death cases in Brazil. While (Hasan, 2020) combined the Empirical mode of decomposition (EMD) and ANN to form a new hybrid model EEMD-ANN to predict confirmed deaths globally. Compared to traditional statistical analysis,

EEMD-ANN outperformed with the lowest MSE and the highest R-squared (Hasan, 2020). In a separate study, ANFIS and multi-layered perceptron-imperialist competitive algorithm (MLP-ICA) were hybridised to predict the time series of daily cases and mortality rate in Hungary (Pinter et al., 2020).

There are many other research areas related to the application of ML/DL algorithms beyond incidence forecasting. For example, identifying risk factors associated with COVID-19 infection or deaths allows policymakers to make better decisions on implementing proper intervention strategies (Balogh, Harman & Kreuter, 2022; Sarkar & Chakrabarti, 2020; de Moraes Batista et al., 2020; Chen et al., 2018). Also, predicting the mortality outcome for COVID-19 patients can efficiently support clinical decision-making for COVID-19 patients who are at critical risk of mortality (Wong et al., 2021; Alballa & Al-Turaiki, 2021). In addition, SARS-CoV-2 diagnosis and severity, anticipating intensive care unit (ICU) admission and the outcome of ICU patients, and length of hospital stay, are all crucial for optimizing decision-making and effectively allocating scarce resources during peak periods (Alballa & Al-Turaiki, 2021; Wynants et al., 2020; Albahri et al., 2020; Alabool et al., 2020; Shi et al., 2021; Islam et al., 2021).

A brief comparison of methodologies that have been adopted in infectious disease forecasting is summarised below in Table 2.2.

Table 2.2 Comparison of advantages and limitations of different forecasting methodologies.

Model	Advantage	Limitation
Compartmental models (SIR, SEIR, etc.)	They are flexible to incorporate additional conditions such as interventions or behavioral changes; They also can be combined with other approaches such as particle filtering to improve performance.	Sometimes fail to capture the differences in contact patterns for different subgroups (e.g., age groups)and environment because of the assumption of fully mixed and homogeneous population
Agent-based models (ABM)	They capture the heterogeneity of individual behaviours, which introduces realism the simulated process.	ABMs can be computationally expensive; Determining the appropriate values for model parameters can be challenging.
Time-series models (ARIMA, sARIMA, etc.)	They are easy to implement and interpret.	Unable to capture non-linearity.
Machine learning / Deep learning models	They are capable of analysing large and complex data sets to identify patterns or trends that may be difficult for humans to detect; They do not reuqire parametric assumptions; Faster training process.	They are 'black box' algorithms without explicit model specifications, which add the difficulty of interpretation; Require high-quality, comprehensive data to achieve optimal performance while data are often incomplete or biased in the real world.

Chapter 3 Forecasting Influenza Incidence as an Ordinal Variable Using Machine Learning

This chapter has been published on Wellcome Open Research (Wang, Kwok & Riley, 2024), awaiting peer review. The preprint is available on Medrxiv (Wang, Kwok & Riley, 2023).

3.1 Introduction

Influenza forecasting plays a critical role in helping healthcare planners and policymakers to improve their response to large seasonal epidemics and to mitigate their impact in terms of morbidity and mortality as well as social and economic impacts. Before the COVID-19 pandemic, a contagious respiratory illness caused by influenza viruses posed continual threats globally, causing an estimated 1 billion cases and up to 650000 deaths annually (World Health Organization, 2023b, 2019). Accurate and timely forecasting of influenza epidemics in terms of the start of the epidemic, the time and size of the peak, and the duration of the epidemic enable policymakers to take effective interventions and optimise the allocation of healthcare resources. For example, to conduct the maximum number of elective surgical procedures prior to opening up space in intensive care wards for community-acquired pneumonia patients. However, reliable influenza forecasts remain a substantial challenge due to the variation of dominant influenza strains and environmental factors affecting the outbreak intensity (Viboud & Vespignani, 2019; Ali & Cowling, 2021; Chretien et al., 2014).

Different analytical methods have been used as the basis for forecast models of influenza disease. These methods can be classified into two broad categories: mechanistic models and statistical methods (Chretien et al., 2014; Nsoesie et al., 2014; Brooks et al., 2018). Mechanistic models attempt to reproduce key features of the underlying mechanism of transmission. Typical examples include classic compartmental models (Yang, Lau &

Cowling, 2020; Kramer & Shaman, 2019; Shaman et al., 2013; Kermack, McKendrick & Walker, 1927) and agent-based models (ABMs) (Laskowski et al., 2011; Arduin et al., 2017; Dugas et al., 2013; Spaeder, Stroud & Song, 2012). Statistical approaches are phenomenological and do not attempt to reproduce the transmission mechanism, such as autoregressive integrated moving average (ARIMA) (Achrekar et al., 2011; Perrotta, Tizzoni & Paolotti, 2017) and Gaussian Process Regression (GPR) (Chen et al., 2019).

Machine learning (ML) models have gained much attention in recent years and have gradually become the third category of models. Given that machine learning is sometimes differentiated from traditional statistics for its focus on prediction, it is not surprising that these methods are being applied to forecasting influenza. Traditional statistical models have a longstanding emphasis on inference, which is achieved through creating and fitting project-specific probability models. They are often less well-suited to data with large sample sizes and input variables (Bzdok, Altman & Krzywinski, 2018). By contrast, ML methods are data-driven and avoid making prior assumptions about underlying correlations and rather employ algorithms to identify patterns in the data (Bzdok, 2017; Bzdok, Krzywinski & Altman, 2017). In addition, ML methods are flexible in taking different types of input variables and a huge number of observations into consideration to improve predictive performance. The usage of some popular machine learning and deep learning methods in influenza forecasting has been discussed, such as Long Short Term Memory (LSTM) (Venna et al., 2019), Support Vector Machine (SVM) (Nsoesie et al., 2021; Liang et al., 2018; Singh & Kaur, 2021), and neural networks (Wu, 2018; Aiken et al., 2021; Li et al., 2020; Hu et al., 2018).

Many natural extreme phenomena, such as floods and earthquakes, are often described on an ordinal scale, but infectious disease incidence is usually described as a count or a percentage of a population. The use of the moving epidemic method is a notable exception, which is now routinely used to compare and communicate the current state of epidemics across nearby connected populations (Vega et al., 2013). Ordinal time-series classification

and forecasting have been applied in some non-epidemical areas such as air quality index and infant sleep status, and they have demonstrated high accuracy of predictions with ML models (Chen & Chiu, 2021). Ordinal indicators have been used to quantify ILI activity in the US (FluView, 2012; CDC Epidemic Prediction Initiative, 2021), but influenza forecasting has remained focused on non-ordinal target observations (Reich et al., 2019a), limiting the range of analytical methods that can be applied.

Extreme Gradient Boosting (XGboost) is a decision-tree-based ensemble machine learning method employing the gradient boosting algorithm (Chen & Guestrin, 2016a). It has demonstrated good prediction performance on a wide range of problems in different industries, including finance, physics and clinical research (e.g. patient diagnosis) (Wang, Cheng & Dong, 2022; Qin, 2022; Zhang, Deng & Jia, 2020; Ogunleye & Wang, 2020).
In this paper, we use XGBoost to create a categorical forecast model for weekly numbers of influenza cases, ranging from 1- to 4-weeks ahead of the current time. The model uses only data from prior weeks and time of year as potentially informative features. We evaluate the performance of XGBoost by accuracy metrics such as mean-zero error (MZE) and macro-averaged mean absolute error (mMAE) and compare it with baseline models and show that this approach has merit as a generic forecasting approach for data from many different countries.

3.2 Methods

3.2.1 Data

We used a global web-based collection tool for epidemiological indicators and data on influenza, Flu Informed Decisions (FluID) dataset from the World Health Organization (WHO) (World Health Organization, n.d.). This platform collects weekly Influenza-Like Illness (ILI) incidence data from WHO member countries and regions, which is either submitted on a

weekly basis or obtained by the WHO from regional networks such as EUROFlu (World Health Organization, 2019).

The FluID data was available from ISO year 2010 week 1 to 2017 week 52 and included data from 146 countries initially. To ensure sufficient data for model training, countries with less than 50% of the data were excluded. Further countries that did not have at least 10 weeks of data for each year between 2010 and 2017 were also excluded. After applying these criteria were applied to the dataset, 32 countries remained, primarily in Europe and North America.

The key field in the data was ILI incidence, which was an integer ranging from 0 to 44965 (in week 52 of 2017 in the USA). This field was transformed into an ordinal variable by discretizing into N different bins with equal intervals, using N = 10 as the default. The highest incidence data for each country was used to add 10% as the upper boundary. The range from 0 to the upper limit was then divided into ten equal ordinal intervals, each of which was mapped to an ordinal value from 1 to 10, 1 represented the lowest incidence level, while 10 represented the highest level. Note that each country's classification was based on its own influenza incidence level, so the same category in two different countries corresponded to different absolute levels of incidence.

3.3.2 Models

We used XGBoost to establish a model to predict short-term weekly influenza incidence levels (Chen & Guestrin, 2016a). XGBoost is an implementation of the gradient-boosted decision trees and has been developed to improve computation speed and predictive performance for a variety of problems, including classification and regression (Chen & Guestrin, 2016b). Gradient boosting is an algorithm where new models are added in an adaptive way based on the residuals or errors of predictions from prior models and then

combined to make the final prediction. Boosting is an ensemble method that creates a strong prediction model by iteratively combining a number of weak classifiers. New weak classifiers are added to correct the errors made by existing models, and every new model is added sequentially until no further improvements can be achieved or until a maximum number of models is added. Gradient boosting uses a gradient descent algorithm to minimise the loss when adding new models, i.e., at every optimisation step, only models that reduce the residual or errors are added. XGBoost uses a second-order Taylor expansion to minimise the loss function and added regularization terms to prevent overfitting.

To further validate the effectiveness of XGBoost for influenza forecasting, we constructed two additional baseline models for comparison: 1) null model, in which the prediction of the target week is the same as the most recent available observation week, and 2) historical average model in which the prediction of the target week is the average of observations of the same weeks in other years. The usual definition for historical average models is to mean value over prior observations for that week of the year (Hyndman & Athanasopoulos, 2021). Here, we use a slightly different categorical definition of the incidence level with the highest frequency across the same week in the other years in the study. The aim of the historical average model is similar to that of the null model, which is to provide a baseline for comparison with the accuracy of the XGBoost model, reflecting how users might informally forecast the incidence levels for the next 1 to 4 weeks.

3.2.3 Forecast and Features

Our target was to predict ordinal categories for each week at the country level. We first did regular machine learning forecasting, in which we trained the model with a training set of fixed durations and then assessed the model's performance on a testing set. We designed our analysis to use 5 years of training data to predict one year of outcomes. The first training

set used data from 2010 to 2014, the second training set used data from 2011 to 2015, and so on (Table 3.1).

Table 3.1 Time ranges of training and test sets for the XGBoost model.

Training set		Test set
Start Year	End Year	Year
2010	2014	2015
2011	2015	2016
2012	2016	2017

For predictors in the short-term forecasting model, we used: categorical incidence levels of the prior n week and prior (*n* + *1*) weeks, the month of the year (from January to December) and the season (spring, summer, autumn, and winter), where n represents the n-week ahead forecast (Table 3.2). In other words, the prior n and (*n* + *1*) weeks depend on which week to be predicted. For example, if we aim to predict the ILI incidence level for week 7, and we have the most recent data up to week 6, then week 7 is predicted as a 1-week ahead forecast. Thus, the prior *n* (*n* = 1) week of week 7 is week 6 and the prior (*n* + *1*) week is week 5 will be used to help improve the accuracy. Similarly for the longer forecasting horizons, if week 7 is predicted as a 2-week ahead forecast, that means the most recent data is only available up to week 5, and then the prior 2 and 3 weeks are week 5 and week 4 respectively.

Table 3.2 List of input features and value ranges for the XGBoost model.

Feature	Value
ILI incidence level of prior n week	Intergers between 1 and 10
ILI incidence level of prior (n+1) week	Intergers between 1 and 10
Season of prior n week	Spring, Summer, Autumn, Winter.
Month of prior n week	January, February, March, April, May, June, July, August, September, October, November and December.

*N represents the number of weeks ahead to be predicted, and n can take values from 1, 2, 3 and 4.

We adopted two types of training methods to train the model and then make predictions. The first method is the fixed training window approach, which is the fixed-origin evaluation

introduced in Chapter 2. In the fixed window approach, we train the XGBoost model only once with the fixed training set and then predict the test set. Since the data were collected on a weekly basis, we are able to update our model by including the new coming data and retraining it to see if model performance can be improved. Thus, the other method is to train the model with an extended training set for each week, which is called the extending window approach in this paper, and it was described in Chapter 2 as a rolling-origin evaluation. For example, if we are going to predict the week *i* (*i* > 2) of the year 2015, our training set used data from 2010 until 2014 for all weeks and data from 2015 for weeks 1 to week (*i-1*). Our test data was week *i* in 2015. The baseline models were trained with the fixed window approach only. Weeks with missing values for predictors were removed from both the training set and the test set.

XGBoost provides a large number of hyperparameters to help achieve optimal performance. They are mainly divided into three types, general parameters, booster parameters, and learning task parameters (Distributed (Deep) Machine Learning Community, n.d.). Booster parameters are closely related to the performance of the model, which is the focus of hyperparameter tuning. Grid search was used to final optimal values for hyperparameters max_depth, min_child_weight, subsample, colsample_bytree, learning rate and gamma. Grid search can be challenging and time-consuming due to the many parameters to optimise, even with XGBoost's rapid convergence. Our gird search for hyperparameters was carried out as below:

1. Firstly, find the optimal gamma and eta (learning rate) at the same time since they have an impact on the performance of the model. The values searched for gamma are 0.1, 0.2, 0.5, 1, 1.5, 2, and 10, while those for eta are 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, and 0.3. We ran all possible combinations of these two hyperparameter values to tune, and the one with the best performance was selected as the optimal value for gamma and eta, respectively.

2. With the optimal values of gamma and eta obtained in the previous step, a grid search was conducted for max_depth, and min_child_weight range from 0.1 to 1.
3. Made a grid-search over subsample and colsample_bytree simultaneously range from 0.1 to 1.

K-fold cross-validation is used during the tuning process to assess the model's performance with different combinations of hyperparameter values. Since there is a dependency between weekly ILI incidence and future values that cannot be used to forecast past values, the traditional K-fold method that randomly splits the training set into K folds is not applicable to our data. Instead, we used cross-validation on an extending basis (Figure 3.1). We selected data from 2010 to 2014 as the overall training set for the cross-validation, but during the process of rolling cross-validation, we started with data from 2010 as the first fold of the training set and checked the accuracy of prediction for the data from 2011. Data from 2010 and 2011 then formed the second fold of the training set, and data from 2012 became the test for accuracy check.

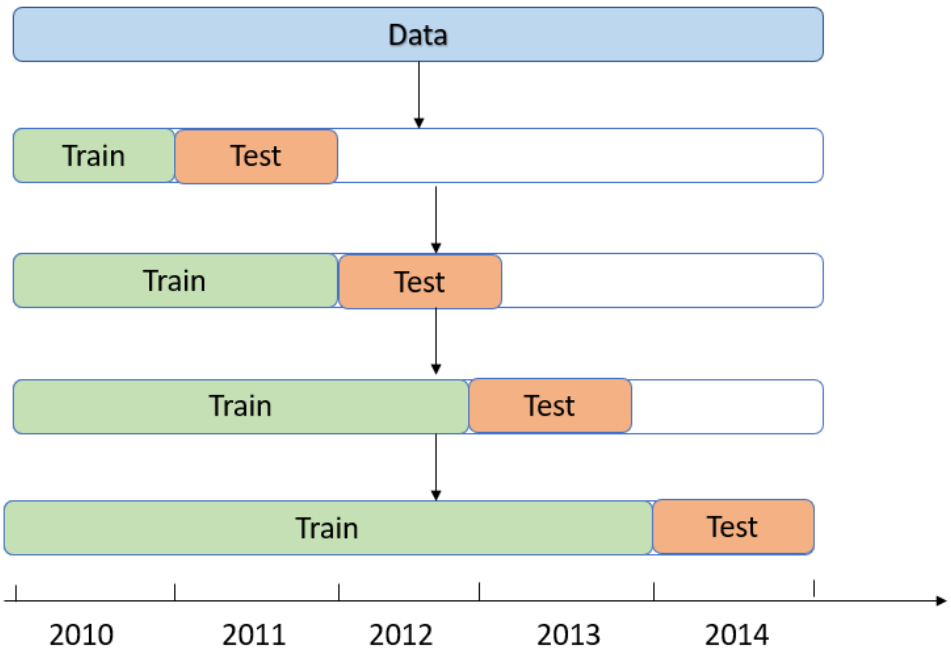


Figure 3.1 Schematic for time series split cross-validation.

The raw output contained the predicted probability of each data point belonging to each ILI incidence level and we selected the level with the highest probability as the final prediction. The accuracy metric used in the tuning process is the macro-averaged mean absolute error (mMAE) which will be defined in detail later. This process was repeated until 2014 and became the last test set. The final optimal values of the hyperparameter are given in Table 3.3.

Table 3.3 Summary of values of hyperparameters used for XGBoost models.

Hyperparameter	Value
maximum depth of a tree (max_depth)	4
minimum sum of sample weight (hessian) needed in a leaf node (min_child_weight)	1
subsample ratio of the training set (subsample)	1
subsample ratio of conlums used to construct each tree (colsample_bytree)	0.8
minimum loss reduction (gamma)	0
learning rate (eta)	0.02

3.2.4 Accuracy Metrics

We used two of the most commonly used metrics in ordinal classification problems (Baccianella, Esuli & Sebastiani, 2009), macro-averaged mean absolute error (mMAE) and mean zero-one error (MZE) to evaluate the model performance. Both metrics are defined as negatively oriented penalties that we aim to minimise: the lower the score, the better the forecast is considered.

3.2.4.1 Macro-averaged Mean Absolute Error

Macro-averaged mean absolute error (mMAE) measure is adapted from the traditional class-based metric mean absolute error (MAE) which assesses the average deviation of the predicted class from the actual class, and it is defined as,

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i^* - y_i|$$

where y_i and y_i^* denotes the true class and the predicted class respectively. However, MAE averages effectiveness across individual observations, which does not reflect the imbalanced distribution of classes in our datasets if many observations are in the lowest class and very few are in higher classes. Instead, we used macro-averaged mean absolute error (mMAE) as one of our metrics to assess the performance of our models in terms of predicting ordered outcomes. mMAE is obtained by first computing the MAE on a per-class basis and then averaging the results across the classes so that mMAE is insensitive to class imbalance. Let y_i be the true class and y_i^* be the predicted class of *i-th* week in the test set. Let N_k be the number of true cases with the class *k* where *k* ∈ 1, 2, ... , 9, 10. There are 10 classes in our classification problem, i.e., $N_k = \sum_{i=1}^{N} I_{y_i = k}$ and $N_k = \sum_{i=1}^{N} I_{y_i = k}$. The Macro-averaged mean absolute error is calculated with the following formula:

$$mMAE = \frac{1}{10}\sum_{k=1}^{10}\frac{1}{N_k}\sum_{i=1}^{N}|y_i^* - y_i|I_{y_i = k}$$

, where

$$I_{y_i = k} = \begin{cases} 1, & if\ y_i = k \\ 0, & if\ y_i \neq k \end{cases}.$$

3.2.4.2 Mean Zero-one Error

Our second metric is the mean zero-one error (MZE) which is more frequently known as the error rate of classifiers and is, essentially, the proportion of time that a correct prediction is made

$$MZE = \frac{1}{N}\sum_{i=1}^{N} I_{y_i^* = y_i}$$

$$= 1 - Acc$$

where

$$I_{y_i^* = y_i} = \begin{cases} 1, & if\ y_i^* = y_i \\ 0, & if\ y_i^* \neq y_i \end{cases}.$$

y_i and y_i^* denotes the true class and the predicted class respectively. *Acc* is the accuracy of the model, i.e., the number of correctly predicted categories divided by the total number of predictions. Therefore, MZE ranges between 0 and 1, and the lower the MZE, the higher accuracy indicates better predictive performance.

3.3 Results

Country-level forecasts were performed using ILI data from 32 countries, located in northern temperate regions (Figure 3.2). Due to the incompleteness of data, countries located outside of temperate regions were excluded from the analysis. The countries in the northern hemisphere's temperate regions demonstrated a similar trend in influenza activity, with ILI incidence typically rising at the end of the year and reaching a peak level at the beginning of the following year (Figure 3.3B). During the 2010 - 2017 surveillance, after excluding weeks with missing values, the data availability ranged between 229 and 417 weeks, with each country contributing an average of 337 weeks of data (median = 391 weeks). Of these weeks, over 60% fell into level 1, while 0.76% of the total weeks fell into levels 9 and 10 (Figure 3.3A).

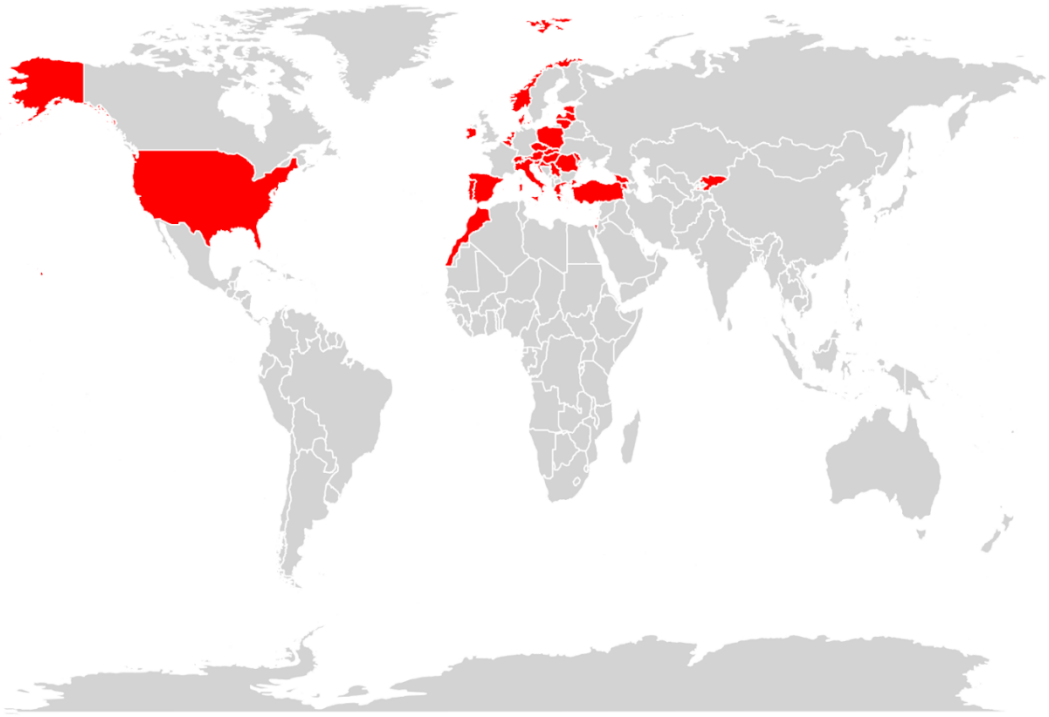


Figure 3.2 World map.

Thirty-two countries included in this study are marked in red. Samples are mainly from North America and Europe.

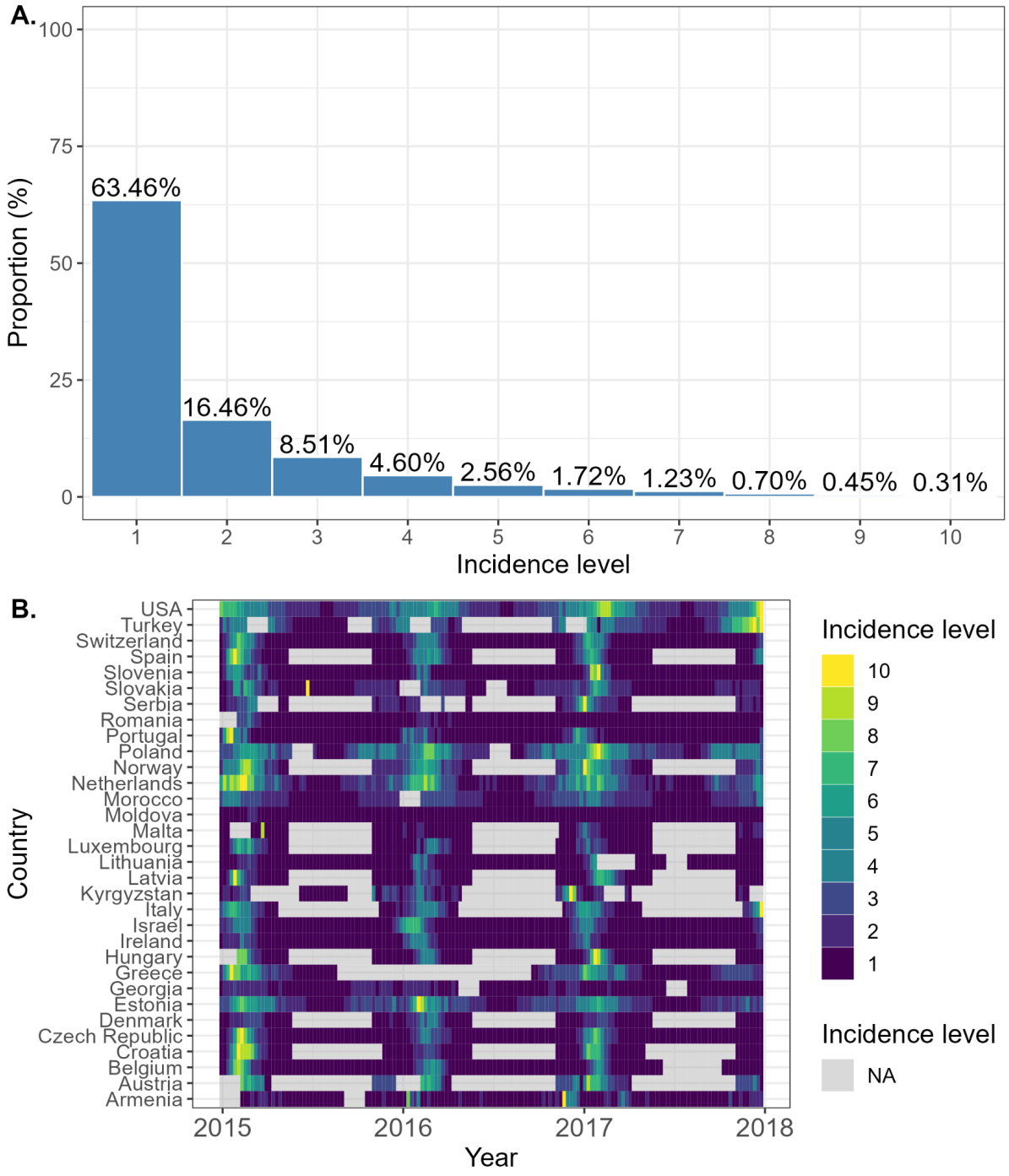


Figure 3.3 Distribution of categorical incidence levels for weekly influenza-like illness for countries included in the study.

A. Overall frequency of every level appearing in 32 countries. **B**. Incidence level in each country by time. Grey-shaded areas indicate that no available were data for those weeks.

Overall, the XGBoost model with an extended window generated fewer errors and demonstrated more stable predictive performance than the XGBoost model with a fixed training window or the baseline models. The short-term prediction errors of different models are summarised in Table 3.4, by averaging the mMAE across countries and presented for each prediction horizon (1-week ahead, 2-week ahead, etc.). We separately compare the mMAEs of three test periods (2015 to 2017). From Table 3.4, for these three test periods, we show that prediction accuracy measured by mMAEs decreases with the increasing forecast length (i.e. one to four-week in advance), but both XGBoost models (with extending window and fixed window) uniformly outperform the two baseline models (Figure 3.4, Table 3.4). In particular, the mMAEs of the XGBoost model with an extended window remain below 1, on average across all countries, even as the forecast length increases, indicating that, on average, the predicted classes are either correct or only one class away from the true class. This XGBoost model has the lowest average MZEs for all prediction horizons compared to the XGBoost model with a fixed window and baseline models (Figure 3.5, Table 3.5).

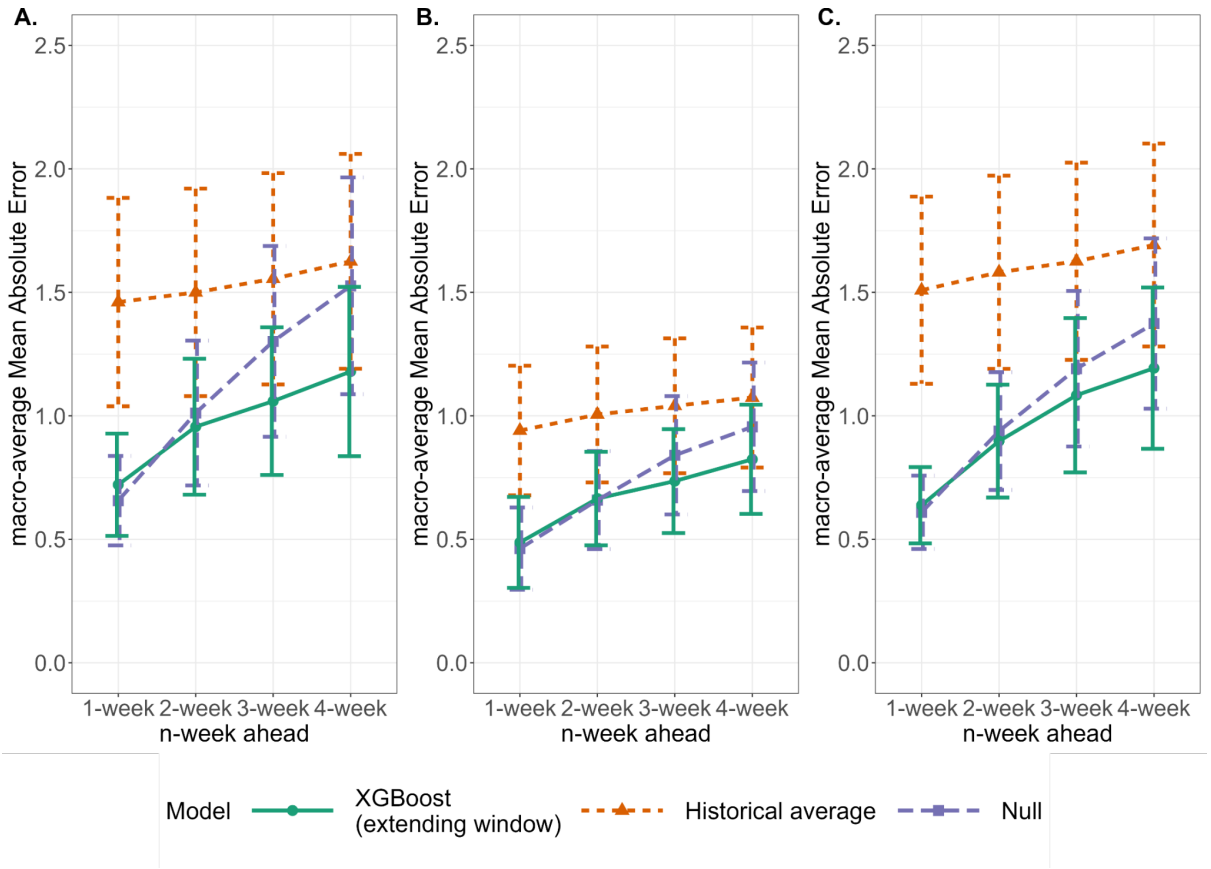


Figure 3.4 Differences in macro-averaged mean absolute error (mMAE) by model.

Comparison of mMAE for XGBoost (with an extended window) and baseline models (Historical average and Null models) while forecasting 1 to 4 weeks ahead. **A**. 2015, **B**. 2016, **C**. 2017.

Table 3.4 Overall macro-averaged mean absolute error (mMAE) of XGboost compared with baseline models (Historical average and Null models).

Year	Models	mMAE* of n-week ahead			
		n = 1	n = 2	n = 3	n = 4
2015	XGBoost with extending window	0.721	0.956	1.060	1.179
	XGBoost with fixed window	0.822	0.997	1.102	1.221
	Historical average model	3.316	3.449	3.554	3.705
	Null model	2.908	1.848	2.447	2.908
2016	XGBoost with extending window	0.488	0.665	0.736	0.824
	XGBoost with fixed window	0.520	0.675	0.789	0.836
	Historical average model	3.338	3.200	3.270	3.247
	Null model	2.751	1.826	2.429	2.751
2017	XGBoost with extending window	0.638	0.898	1.084	1.193
	XGBoost with fixed window	0.645	0.918	1.148	1.268
	Historical average model	3.484	3.604	3.734	3.878
	Null model	2.663	1.720	2.221	2.663

*Overall mMAEs of 1- to 4-week ahead forecasts for each model are calculated as the average of 32 countries' mMAEs by year.

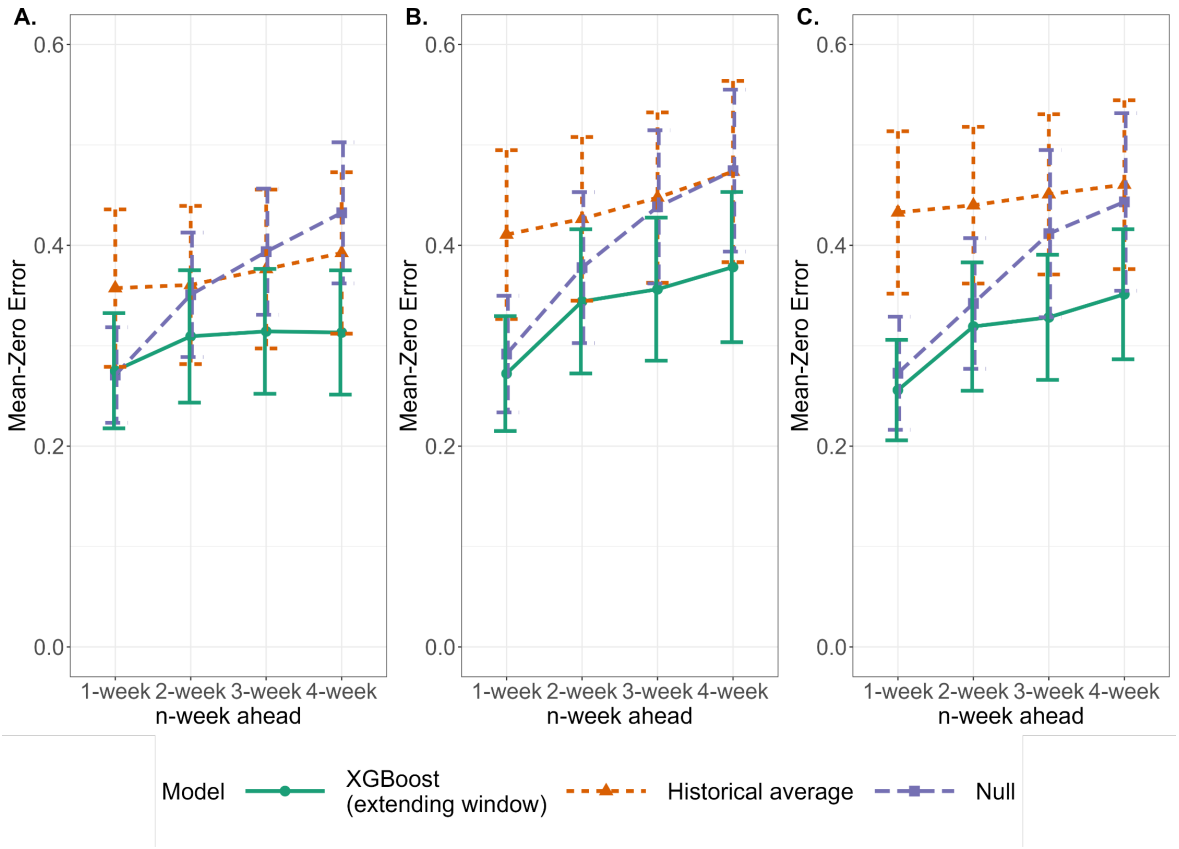


Figure 3.5 Differences in mean-zero error (MZE) by model.

Comparison of mean-zero error (MZE) for XGBoost (with an extended window approach) and baseline models (Historical average and Null models) while forecasting 1 to 4 weeks ahead for the test period year 2015 to 2017. **A**. 2015, **B**. 2016, **C.** 2017.

Table 3.5 Overall mean-zero error (MZE) of XGboost compared with baseline models (Historical average and Null models).

Year	Models	MZE* of n-week ahead			
		1	2	3	4
2015	XGBoost with extending window	0.275	0.309	0.314	0.313
	XGBoost with fixed window	0.277	0.311	0.317	0.320
	Historical average model	0.323	0.322	0.332	0.349
	Null model	0.406	0.328	0.373	0.406
2016	XGBoost with extending window	0.272	0.344	0.356	0.378
	XGBoost with fixed window	0.282	0.348	0.357	0.380
	Historical average model	0.358	0.368	0.382	0.405
	Null model	0.431	0.334	0.397	0.431
2017	XGBoost with extending window	0.256	0.319	0.328	0.351
	XGBoost with fixed window	0.268	0.339	0.356	0.370
	Historical average model	0.396	0.408	0.415	0.427
	Null model	0.416	0.317	0.38	0.416

*Overall MZEs of 1- to 4-week ahead forecasts for each model are calculated as the average of 32 countries' MZEs by year.

The mMAEs for the extending window approach were consistently lower compared to the XGBoost model with a fixed window (Figure 3.6, Table 3.4). The mMAEs for the fixed window approach only remained below 1 for 1- to 4-week forecasts in 2016, but exceeded 1 for 3- and 4-week ahead forecasts in 2015 and 2017. Although the two XGBoost models had similar MZEs, the extending window approach still resulted in a lower MZE for each prediction horizon (Figure 3.7, Table 3.5). The better performance of the extending window approach can be attributed to its continuous addition of the predicted weeks' observations to the training set, allowing the model to learn more data, resulting in smaller prediction errors.

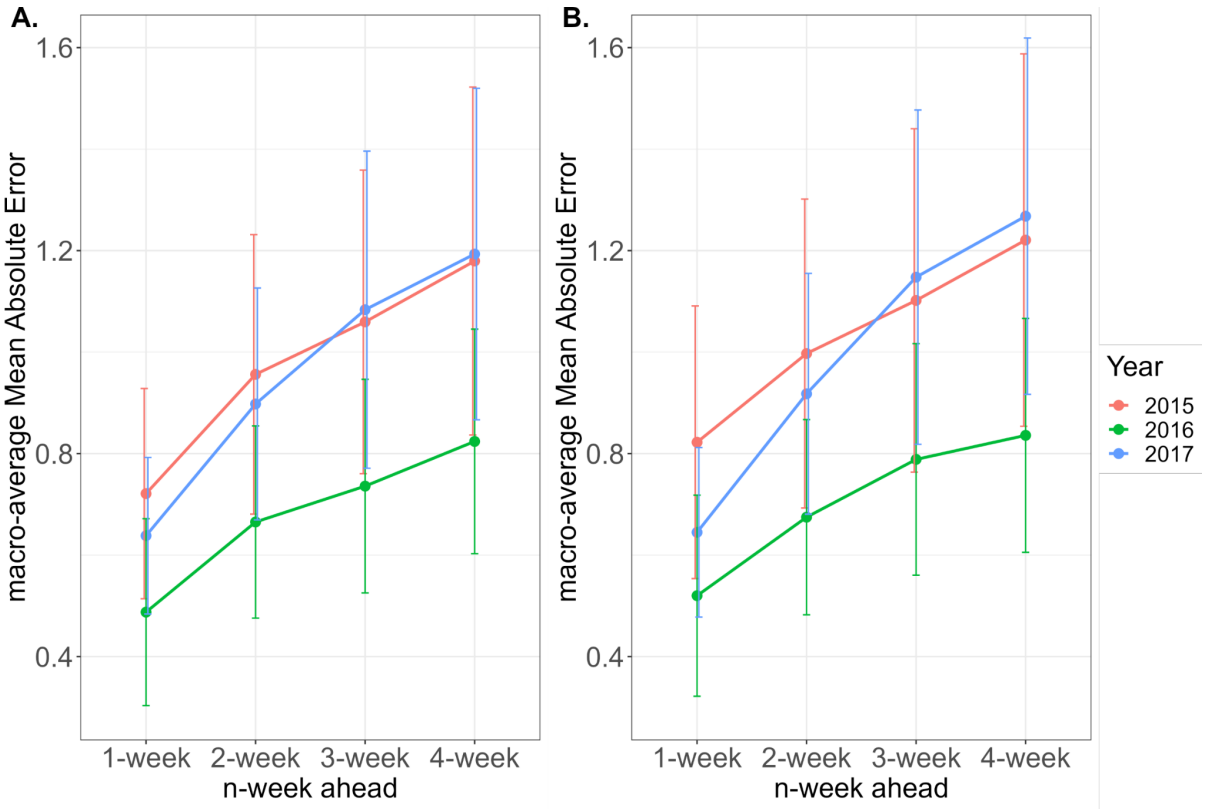


Figure 3.6 Macro-averaged mean absolute error (mMAE) of XGBoost models.

Comparison of mMAEs for weekly forecasting ranging from 1-to 4-week ahead for the years 2015 to 2017. **A**. The XGBoost model with an extended window approach. **B**. The XGBoost model with a fixed window approach.

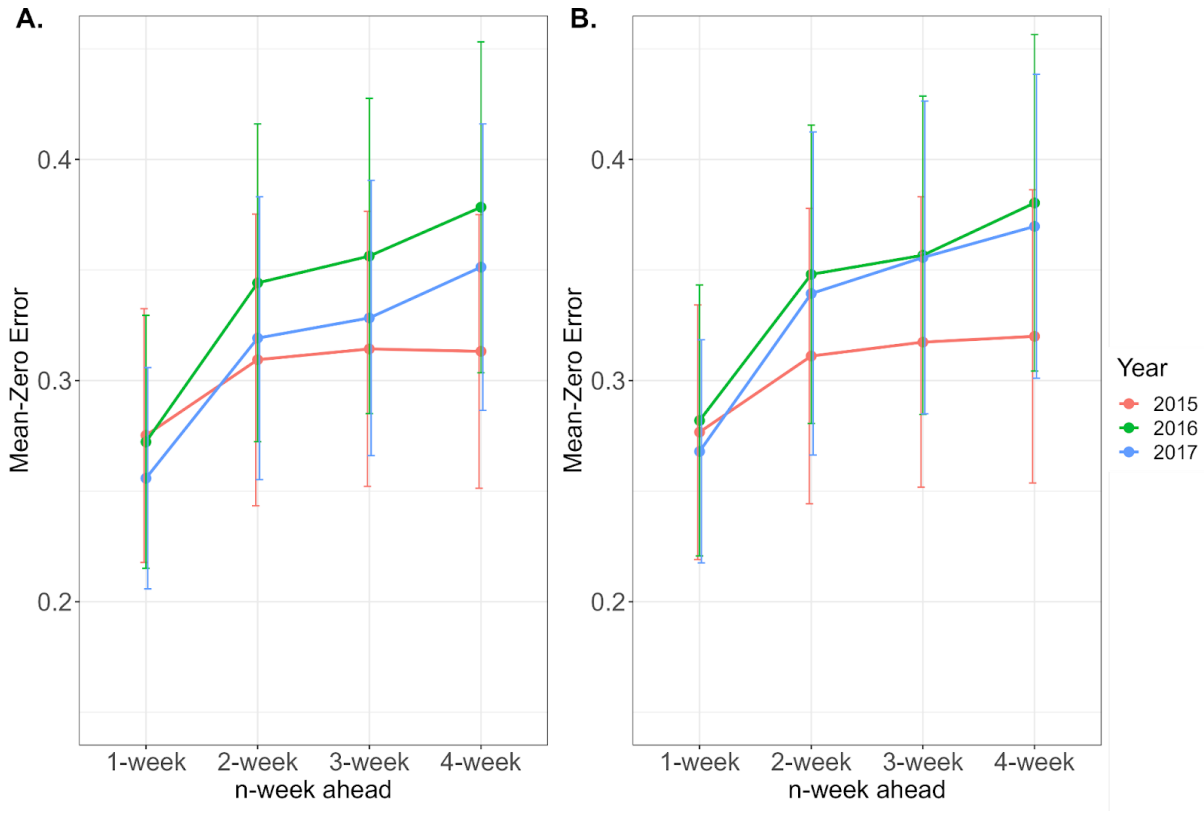


Figure 3.7 Mean-zero error (MZE) of XGBoost models.

Comparison of MZEs for weekly forecasting ranging from 1- to 4-week ahead in 2015 to 2017. **A**. The XGBoost model with an extended approach. **B**. The XGBoost model with a fixed window approach.

There is substantial variation in the prediction accuracy, as indicated by mMAE, among the countries (Figure 3.8). For instance, Moldova, had remarkably low mMAEs from 2015 to 2017, even achieving 0 in 2016 and 2017. Conversely, certain countries, such as Hungary and Norway, consistently exhibit much higher mMAEs compared to other countries. Similar results were found for MZEs (Figure 3.9). The performance of XGBoost on the test set can be affected mainly by the quality of training data. By checking the missing data for countries, we discover that countries with high mMAEs often have more incomplete data than better-performing countries, implying that they have a smaller sample size of training sets, failing to capture complex epidemic patterns of influenza (Figure 3.3B).

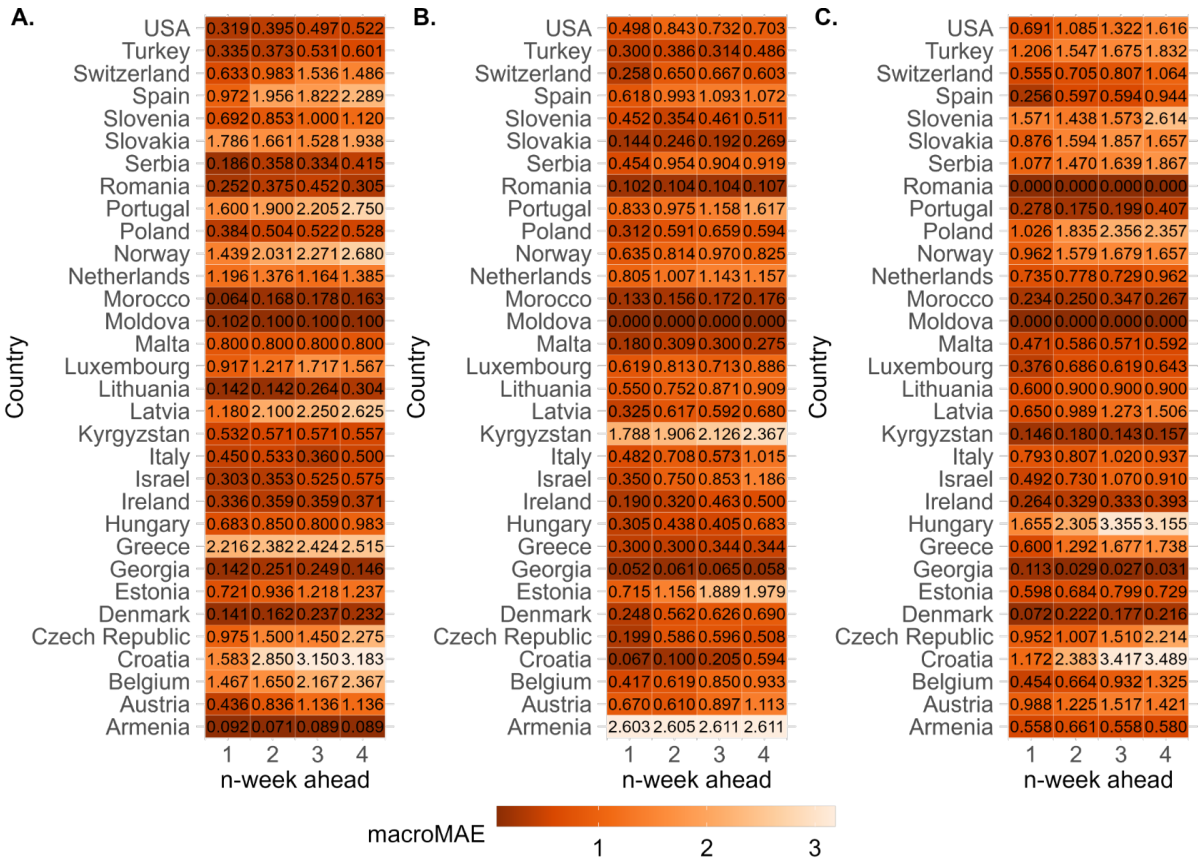


Figure 3.8 Macro-averaged mean absolute error (mMAE) for thirty-two countries.

mMAEs of predictions by the XGBoost model with an extended window for the year **A**. 2015. **B**. 2016. and **C**. 2017. mMAE values were indicated in each shaded box. The higher the mMAE, the darker the shaded box.

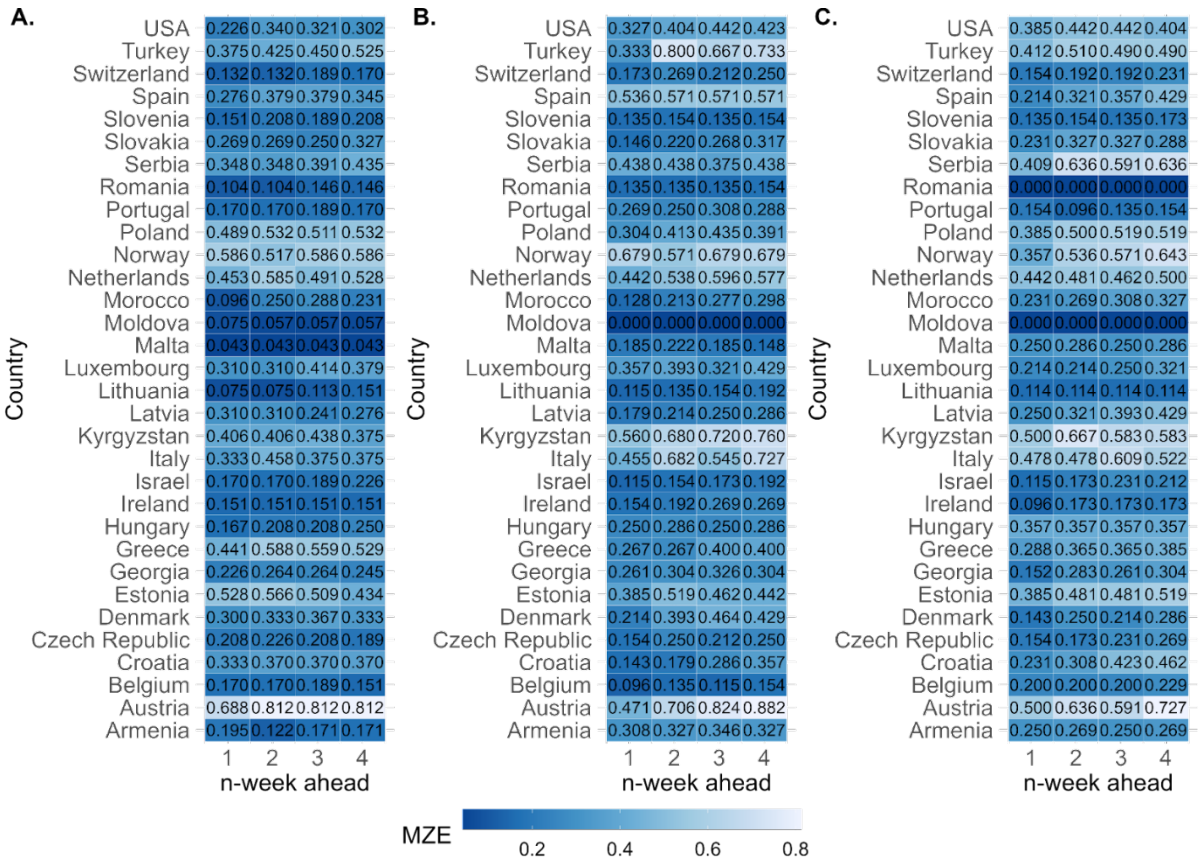


Figure 3.9 Mean-zero error (MZE) for 32 countries.

MZEs of predictions by the XGBoost model with an extended window for the year **A**. 2015. **B**. 2016. and **C**. 2017. MZE values were indicated in each shaded box. The higher the MZE, the darker the shaded box.

We investigated the influenza distribution and forecasts over a 3-year period in four countries selected based on their mMAEs and data completeness and found different qualitative drivers of accuracy. The plots in Figure 3.10 compare the predicted values from four models with actual data for 1 to 4-week-ahead forecasts in Moldova, Switzerland, Estonia and Hungary separately. Moldova had the lowest mMAE, but its influenza activity barely fluctuated with 97.5% (390 out of 400 weeks) of the weeks remaining at level 1. Conversely, Switzerland and Estonia had 25.3% (82 out of 406 weeks) and 72.5% (290 out of 400 weeks), respectively, above the baseline level. All four models successfully predicted peak influenza season for 1-week and 2-week ahead forecasts in 2015 and 2016 in Switzerland and Estonia, but the historical average model failed to identify peaks for 3-week and 4-week ahead forecasts (Figure 3.10). The other three models could

identify peaks up to 4-week ahead of forecasts, with the XGboost with extending approach being the most accurate (Figure 3.10). The deviation between actual observation and prediction is typically high when there is a sudden increase or decrease in flu incidence levels (Figure 3.11). In Hungary, due to the limited sample size of the training set and the complexities of influenza activity, a large deviation between actual observation and prediction even for the 1-week forward forecast. The prediction of XGboost with an extending window provided the closest prediction to the observed peak (Figure 3.11).

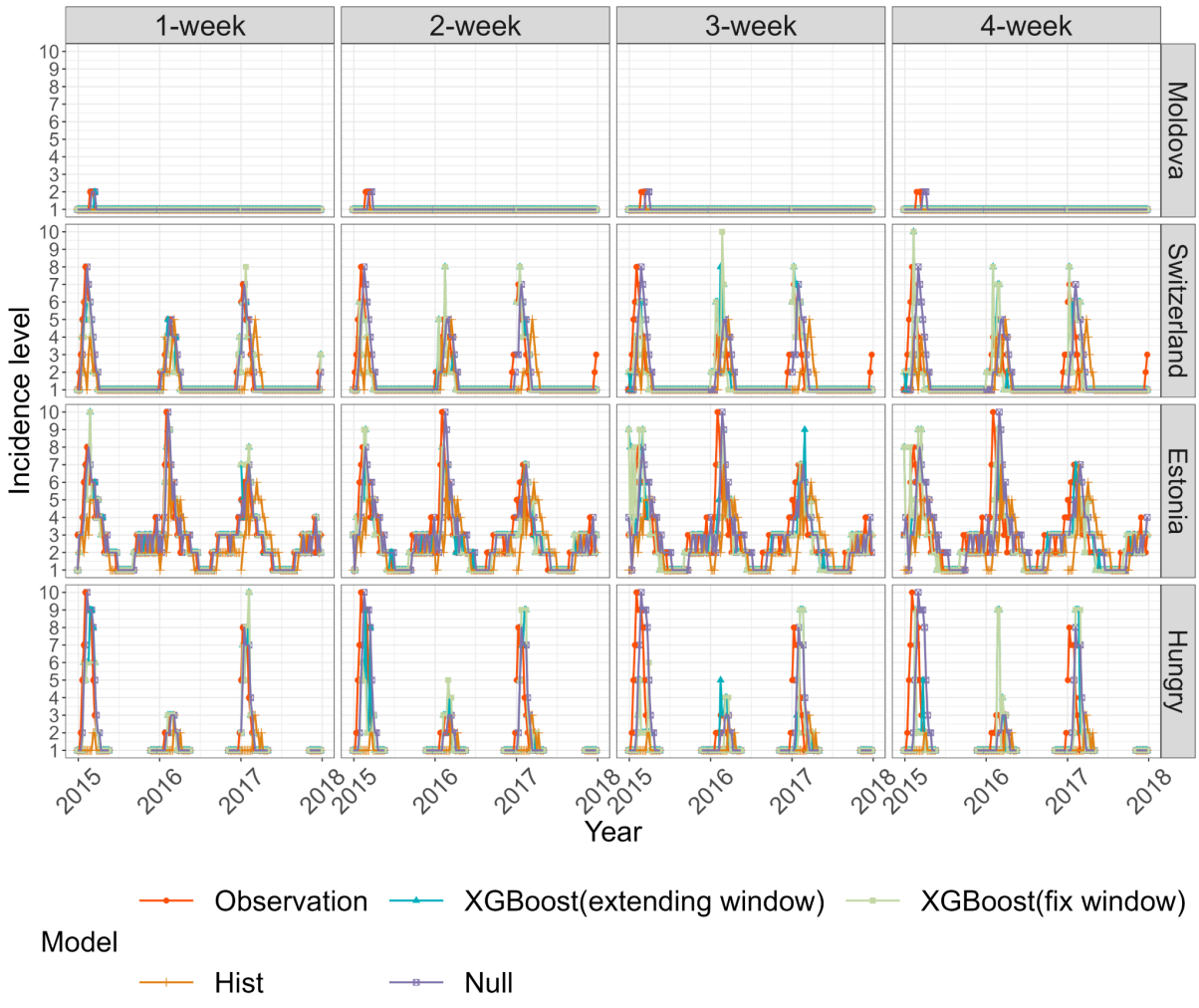


Figure 3.10 Comparison of Actual and predicted incidence by country.

Countries Moldova, Switzerland, Estonia and Hungary are selected. Models include the two XGBoost models, the historical average model and the null model while forecasting 1- to 4-week ahead for the test period year 2015 to 2017.

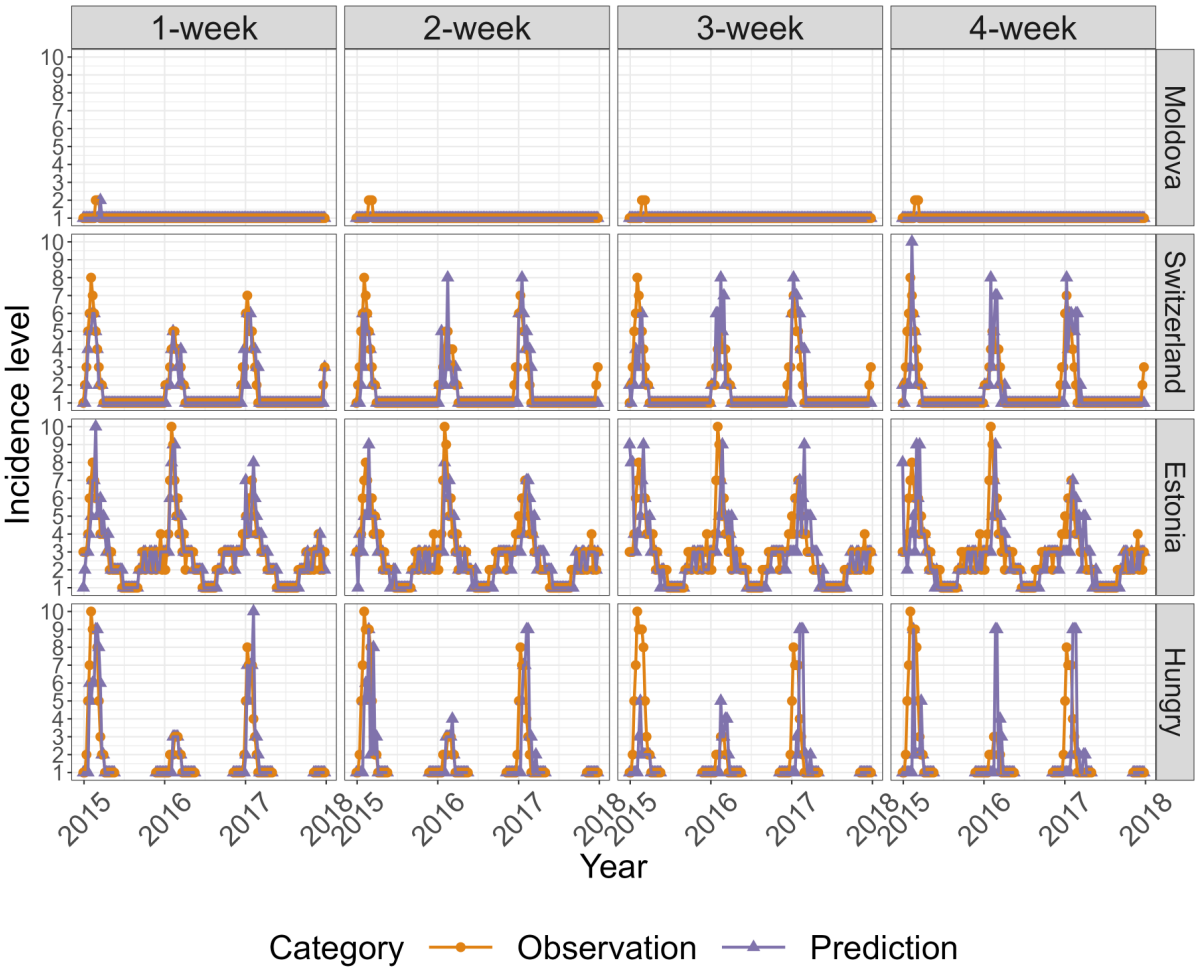


Figure 3.11 Comparison of actual and predicted incidence.

Predicted incidences are generated by the XGBoost model with an extended window for Moldova, Switzerland, Estonia and Hungry; for forecasts 1 to 4 weeks ahead for the test period year 2015 to 2017.

3.4 Discussion

In this work, we have proposed a potential machine learning model for influenza forecasting using a 10-unit ordinal variable to describe the case incidence. We evaluated two XGBoost machine learning models, one with a fixed window training set and the other with an extended window training set and compared them to baseline models. We measured their accuracy using the mean zero-one error (MZE) and macro-averaged mean absolute error (mMAE). Our results showed that both XGBoost models outperformed the baseline models, with the extended window XGBoost model consistently achieving the highest accuracy.

We defined the forecast target on an ordinal scale, which is often how influenza incidence is communicated (Vega et al., 2013). This contrasts with other studies predicting time-series ILI activity as a continuous outcome based on the XGBoost algorithm. The XGBoost algorithm has been used as a base model in the ensemble approach to predict influenza activity at the regional level and demonstrated the contribution of XGBoost to improving accuracy (Cheng et al., 2020; Choi, Kim & Ahn, 2019). However, our results are not directly comparable to these studies since our work is different in the following ways. First of all, we defined the outcome as an ordinal variable rather than continuous and conducted categorical forecasting for ILI activity levels. Furthermore, the proposed framework in this study is solely based on the XGBoost therefore we are able to assess the applicability and predictivity of the XGBoost in influenza forecasting, which has not been verified by other studies before.

The use of a categorical outcome enabled us to apply machine learning algorithms. Although the proposed XGBoost model has advantages, it has not been tested to predict ILI incidence as an ordinal variable against other influenza surveillance data sources. Machine learning algorithms' performance can be limited by the quality of the training and testing data, and inconsistencies between different databases may reduce the generalisability of the XGBoost prediction model (Jain et al., 2020). In addition, it has not been applied to predict other

infectious diseases such as COVID-19. Further testing is needed to determine the applicability of the in real-world public health practice.

Our baseline models provide a basic framework for assessing the potential of machine learning forecasts. We have chosen relatively simple NULL models for categorical outcomes without prior assumption or hypothesis because the vast majority of public health officials do not access current forecasting methods and those methods haven't been rigorously compared against these null hypotheses. The baseline models are intended to reflect the most basic of implicit forecast assumptions: they assume that incidence will remain unchanged and revert to its historical average. Our methods offer improvements over these assumptions, for example, the 4-week forecast horizon XGBoost model with an extended window being more accurate than assuming the incidence would follow the historical average and more accurate than the default 1-week ahead model with no change assumption. In the future, we will compare the proposed XGBoost model in this work to more sophisticated and informative models than the current baseline models, such as the historical average based on observed features (Angers, Biswas & Maiti, 2017).

This work has several limitations. One of the challenges in working with infectious disease datasets is the limited period of data available for model estimation and evaluation, compared to other datasets used in machine learning. For each country in the study, only 5 years of data were used for model training, leaving only 3 years for testing. Cross-validation was used to find the optimal hyperparameters instead of using an independent validation set. Additionally, countries with fewer weeks of data in the training set tend to have higher prediction errors, which suggests that accuracy will improve as more data becomes available. Furthermore, the results should be interpreted with caution, given the disruption to influenza transmission during the COVID-19 pandemic, as patterns may change as the transmission is re-established.

We used ILI datasets and generated forecasts at the country level which made including the time-varying external variables difficult. For instance, meteorologic factors, such as temperature and humidity, have been recognised to have impacts on the transmission of the virus and they normally are reported by regions rather than by countries. To obtain country-level information for such factors, it is required to process regional data such as taking the average over regions, which may cause information loss and lead to worse predictions. However, this generic framework can be used by any country that has data in a similar format.

Our forecasts showed reduced accuracy in predicting turning points of the epidemic (Figure 3.12). When there is a sudden increase in incidence levels, they are often underestimated. On the other hand, a sudden decrease in incidence levels results in them being easily overestimated. This is especially important in a public health context where accurately predicting rare cases of much higher than usual incidence levels or earlier or later ends of the epidemic is crucial. Thus, improving the prediction stability at the extreme points of season patterns remains a priority in ongoing forecasting work. One possible direction for future exploration is to investigate the impact of incorporating alternative data sources to capture the dynamics of extreme points in seasonal patterns. For instance, exploring the integration of additional environmental or socio-economic variables that could affect the behaviours of these extreme points may provide valuable insights and improve prediction accuracy.

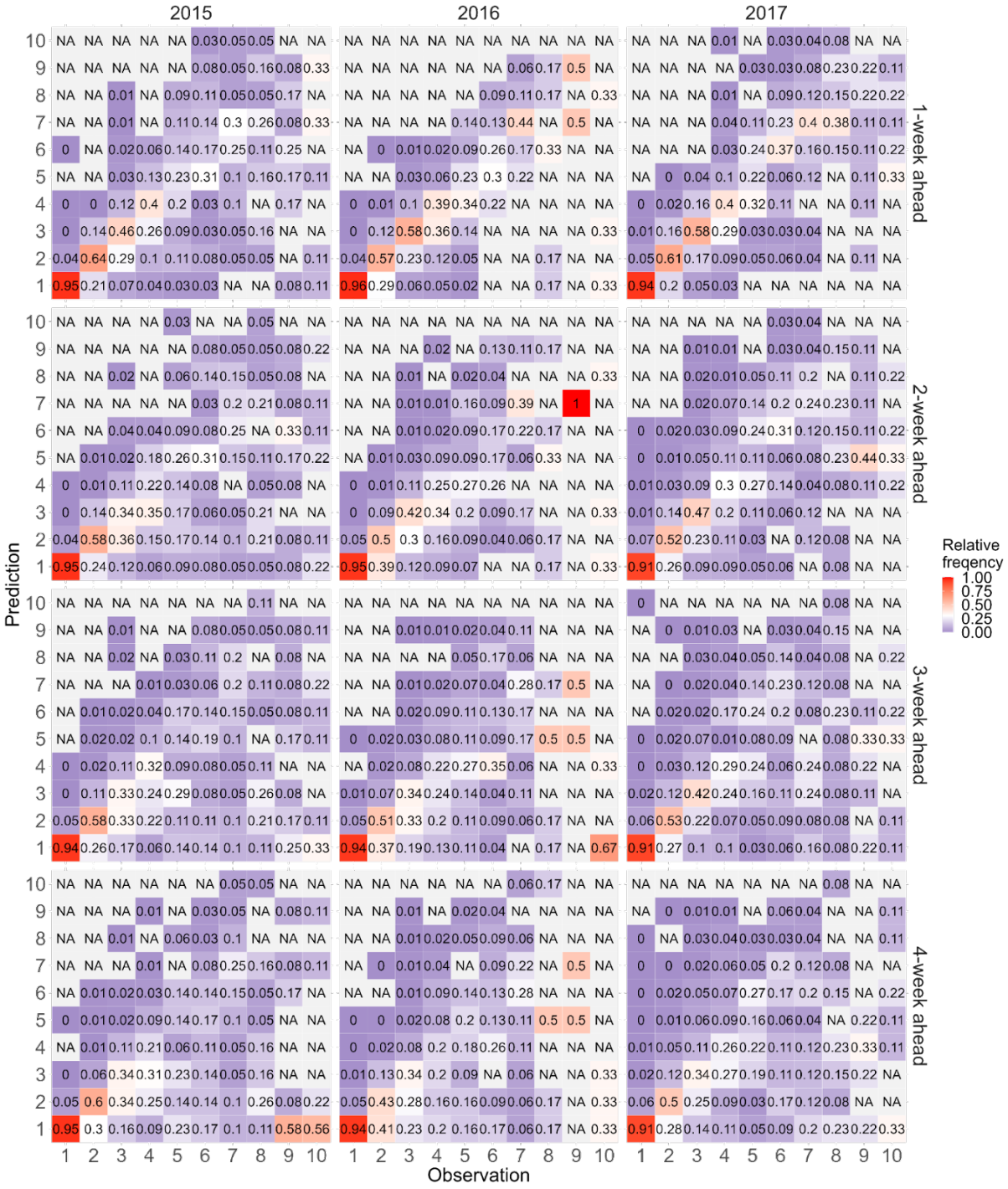


Figure 3.12 Distribution of predictions against each decile of observation for the XGBoost model with an extended window.

Rows of heat maps are ordered from 1 week ahead (top) to 4 weeks ahead (bottom). Columns of heat maps are ordered from 2015 (left) to 2017 (right). Values on the diagonal are the accuracy of forecasts, while other values represent the frequency at each level that was incorrectly predicted to be the other level.

Forecasting influenza as an ordinal outcome is a feasible task for machine learning. The widely used XGBoost model, even with a limited set of features, provides significantly more accurate predictions than the standard baseline models. With datasets that have a longer history and comprehensive spatial coverage, it is possible to achieve more accurate forecasts. Similar to other epidemiological models, the framework can easily be expanded to include population serology and population mobility (Liang et al., 2022) or other relevant features as more data become available (Kwok et al., 2018).

Chapter 4 The Interaction of Age and Geography in the Transmission Dynamics of SARS-CoV-2 during the 2020/21 UK Winter Wave

4.1 Introduction

The outbreak of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that emerged in late 2019 has led to the ongoing COVID-19 pandemic and has rapidly evolved into a global health crisis, challenging healthcare systems, economies, and social structures worldwide. Characterised by person-to-person transmission, the virus has spread rapidly, leading to a range of clinical outcomes, from asymptomatic infections to severe respiratory distress and fatalities (Wu & McGoogan, 2020; Guan et al., 2020b). The World Health Organization (WHO) declared COVID-19 as a pandemic on March 11, 2020, marking a turning point in the fight against the virus (World Health Organization, 2020c). Governments and public health authorities implemented a series of non-pharmaceutical interventions, such as social distancing measures, mask mandates, and lockdowns, to mitigate transmission and protect vulnerable populations (Anderson et al., 2020; Hale et al., 2021). However, as the variants of the virus progressed, different countries and territories experienced multiple waves of surges in SARS-CoV-2 infection at different times, while the United Kingdom is one of the countries most affected by the COVID-19 pandemic.

During the initial wave of the COVID-19 pandemic in early 2020, the UK faced significant challenges in managing the outbreak, with healthcare systems under immense strain and the implementation of strict public health measures to curb transmission. The first wave subsided in June 2020, with a period of declining cases and relaxed restrictions (Prime Minister's Office, 2020a; Sutherland, Headicar & Delong, 2021). However, the second wave of COVID-19 infections occurred later in the year, which began in the fall of 2020 and

extended into early 2021 (Prime Minister's Office, 2020b). This second wave was associated with the emergence of new variants of the SARS-CoV-2 virus, the Alpha variant (B.1.1.7), which was identified as more transmissible than previous strains (Davies et al., 2021). Vaccination was not available to the public in the UK until mid-way through the second wave, with the majority only being vaccinated towards the end of the weave.

The UK government developed a COVID-19 surveillance dashboard, based on community testing of symptomatic individuals, aiming to track the progress of the pandemic in the country (UK Health Security Agency, 2020b). However, variations in testing availability and accessibility and varied definitions of COVID-19 cases adopted by different UK nations may have caused the bias to reflect the actual trend (UK Health Security Agency, 2020c). To reduce such bias and improve the representativeness of the data stream, a large-scale surveillance study program called REal-time Assessment of Community Transmission-1 (REACT-1) was launched in May 2020 (Riley et al., 2020b).

I was a member of the Imperial College REACT-1 analytical team from May 2020 to April 2022 (see also Declaration of Originality). I was deeply involved in various aspects of the project, including questionnaire design and data analysis. My responsibilities included processing and analysing raw data obtained from Ipsos (Ipsos, n.d.), leading on analysis of community prevalence by age groups and geographic regions, and investigating health inequalities among different socio-demographic groups. I helped design, then led the implementation of the analytical pipelines for these results and I worked closely with senior colleagues to agree the interpretation that was published contemporaneously in REACT-1 reports.

REACT-1 study assesses the community prevalence in England by cross-sectional repeated random sampling based on the population of England (Riley et al., 2020b). In addition to the

swab results, the REACT-1 study also collects sociodemographic information (e.g., age, geographic location, occupation, etc.), which allows us to investigate disparities of prevalence among individuals with different backgrounds and risk factors of SARS-CoV-2 infection (Imperial College London, 2020).

The results of the REACT-1 study from the end of the first wave showed variations in the prevalence of COVID-19 cases across different age groups and geographic regions in England (Riley et al., 2021a). The first half of the second wave from September 2020 to December 2020 was a critical phase in the UK because this was the period with relatively high social mixing but during which vaccine coverage was low therefore vaccination had not yet had a substantial impact. Understanding how age and geography interact in driving SARS-CoV-2 transmission during this period is vital to inform targeted public health interventions and develop effective strategies for mitigating the impact of the next time the UK experiences a similar stage with a severe wave of respiratory infection. The second half of the second wave from January 2021 to April 2021 was a period of lower social mixing and increasing levels of vaccination. In this chapter, we aim to present a detailed epidemiological analysis of the trends of prevalence of COVID-19 cases over time during wave two, focusing on how the disparities across age and geographical regions in England changed from the first to the second half and investigating the interaction between them.

4.2 Methods

4.2.1 Data Collection

The REACT-1 study protocol has been described thoroughly elsewhere (Riley et al., 2020b). In brief, we randomly select individuals aged 5 years and above from the National Health Service (NHS) patient list to send personalised invitations, among 315 lower-tier local authority areas (LTLAs) in England. For those participants who indicated consent to take part, they received test kits and instructions delivered by post to obtain self-administrated

nose and throat swabs. The REACT-1 study was conducted on a round basis with the start of each round approximately one month apart. In each round, we send out about 720000 to 760000 swab tests and questionnaires for their demographic and socio-economic information (REal-time Assessment of Community Transmission (REACT) Study, n.d.). There were nineteen rounds in total between May 2020 and March 2022 with over 2.5 million individuals tested. Here, the second wave of COVID-19 cases we focus on is between September 2020 and March 2021, which corresponds to rounds 5 to 10.

The sample size of the study ranged from between 140000 and 180000 people across rounds 5 to 10 (REal-time Assessment of Community Transmission (REACT) Study, n.d.). These samples were collected using dry swabs and then transported to a commercial laboratory using the cold chain to ensure sample integrity. Reverse-transcription-polymerase-chain-reaction (RT-PCR) test with E gene and N gene targets was used to obtain binary outcome (swab positive versus swab negative) of samples. We used cycle threshold (CT) values as a proxy to assess the intensity of the viral load and CT values are inversely correlated with the amount of RNA present in the sample. If both E and N genes were detected, the result of a swab will be defined as positive regardless of CT values. When the E gene was detected, the N gene was also detected. However, in some samples, the N gene was detected but the E gene was not. In this case, samples are considered positive with a CT value for the N gene less than 37. Similarly, samples are considered negative if neither the E gene nor the N gene is detected or if a CT value for the N gene is equal to or greater than 37.

The age of participants was logged as a continuous variable between 5 and 106 years. We converted continuous age into eight categorical age groups: 5 to 12, 13 to 17, 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64 years, and 65 years or older. Individuals were divided into nine geographical regions in England: they are North West, North East, Yorkshire and The

Humble, West Midlands, East Midlands, East of England, South West, London and South East.

4.2.2 Statistical Analysis

The unadjusted prevalence was defined as the proportion of positive samples of total samples, and it will be referred to as unweighted prevalence in the following text. To address non-response bias and unequal selection probabilities inherent in the sampling process, we adjusted unweighted prevalence with pre-obtained weights to estimate the representative prevalence of the population in England, which is referred to as weighted prevalence. The weights were calculated via Random Iterative Method (RIM) weighting (Sharot, 1986), instead of interlacing multiple variables, RIM weighting introduced variables in a sequence. It can start to calculate weights with certain variables, and then this set of weights is corrected by the following variables. This process is repeated until the last variable has been weighted appropriately. We adjusted prevalence to the population of age by sex; LTLA counts; ethnic group; and IMD deciles. Data on age by gender and LTLA counts used here were extracted from the Office for National Statistics (ONS) mid-2019 population estimates (Park, 2020), and the ethnic group from ONS Labour Force Survey (Office for National Statistics, n.d.). The samples were weighted to age by sex and LTLA counts first and re-weighted with all four variables to obtain the final weights. The overall weighted prevalence for age groups and regions was calculated with the survey package in R (Lumley, 2004, 2023).

We explored the interaction between age groups and geographical regions by estimating the weighted nested prevalence using Bayesian multilevel regression model (Shor et al., 2007). The better ability of Bayesian multilevel regression model than the classic survey weighted approach at national and regional levels, particularly when sample sizes are small, has been demonstrated by many simulated and empirical studies (Gelman et al., 2016; Christofoletti et al., 2021; Downes & Carlin, 2020; Warshaw & Rodden, 2012; Si et al., 2017). The

incorporation of random effects in the multilevel model, facilitating partial pooling, ensures the stability of estimates for subnational levels, which could otherwise pose challenges with the classic survey weighted strategy.

Here, a regression model was run to predict the outcome of swab positivity as a function of region and age groups nested within a region, defined below:

$$y_{ijk} \sim Bernoulli\left(n,\ p_{ij}\right)$$

$$logit\left(p_{ij}\right) = \ \alpha + \ weight * \alpha_{region[i]} + weight * \alpha_{region:age[ij]}$$

$$\alpha_{region} \sim Normal\ (0,\ \sigma_{region})$$

$$\alpha_{region:age} \sim Normal\left(0,\ \sigma_{region:age}\right)$$

$$\alpha \sim Normal\ (0,\ 1)$$

$$\sigma_{region} \sim HalfCauchy\ (1)$$

$$\sigma_{region:age} \sim HalfCauchy\ (1).$$

Geographic region (9 levels) was modelled as a random effect, and age groups were nested within each region, expressed as an interaction term (72 levels). Swab positivity is a binary outcome and the probability of being positive was modelled by the multilevel logistic regression.

The model was fitted using the brms package developed in R (Nalborczyk et al., 2019). All the analyses in this chapter were completed under R version 4.0.5.

4.3 Results

4.3.1 Overall Levels of Swab-positivity during the Second Wave in England

England experienced a large second wave of SARS-CoV-2 infection between September 2020 and March 2021, with rising overall swab positivity from September 2020 to January

2021 followed by a decline to March 2021 (Figure 4.1, Table 4.1). Between 18 September 2020 and 30 March 2021, 10756 samples were positive overall from 1470736 swabs.

The resurgence detected during mid-August and early September 2020 (Riley et al., 2021a) (round 4) continued with weighted prevalence reaching 0.60% (0.55%, 0.71%) in round 5 (18 September to 5 October 2020) (Figure 4.1). Infections increased further across round 6 (16 October to 2 November 2020), where we found 1,732 positives out of 160,175 swabs giving a weighted prevalence of 1.30% (1.21%, 1.39%), more than twice that observed in the previous round 5. Compared with round 6, there was a 28% decrease in prevalence in round 7 (13 November to 3 December 2020), likely due to the implementation of a second national lockdown (5 November to 2 December 2020) (Prime Minister's Office & Street, 2020).

Rounds 8 to 10 covered the period from 6 January to 30 March 2021, immediately following the start of a third national lockdown in England on 6 January 2021. In round 8, at the very start of the third national lockdown, we found 2,282 positives out of 167,642 swabs giving a weighted prevalence of 1.57% (1.49%, 1.66%), which was more than 50% higher than the 0.94% (0.87%, 1.01%) recorded in round 7 (13 November to 3 December 2020). This was the highest prevalence observed by the REACT-1 study up to that point. Round 9 (4 to 23 February 2021) included 165,456 valid swabs of whom 689 tested positive, with a weighted prevalence of 0.49% (0.44%, 0.55%), representing a fall of over 60% from round 8. Prevalence continued to fall to round 10 (11 to 30 March 2021), when we found 227 positives out of 140,844 swabs giving a weighted prevalence of 0.20% (0.17%, 0.23%), an approximately 60% reduction from round 9.

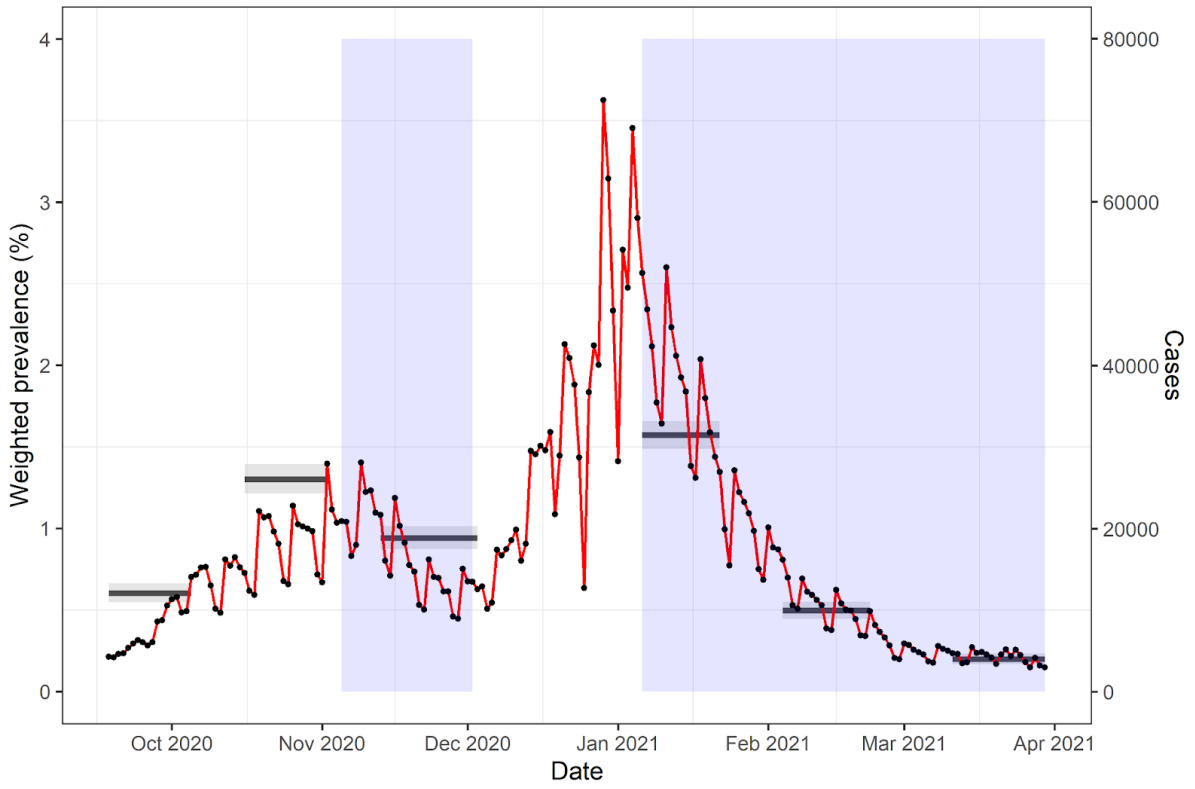


Figure 4.1 Comparison of overall weighted prevalence as measured by REACT-1 to surveillance case data (Public Health England, PHE).

Overall weighted prevalence for rounds 5 to 10 of the REACT-1 study is in black solid lines (left-hand side y-axis) with a 95% confidence interval (grey shaded boxes). The temporal dynamic of surveillance data is shown by black points and the red solid line (right-hand side y-axis). The right y-axis is scaled by 20,000 per case. Blue-shaded areas indicate the time period of the second and the third lockdowns in England.

Table 4.1 Overall unweighted and weighted prevalence with 95% CI for rounds 5 to 10.

Round	Tested swabs	Positive swabs	Unweighted prevalence (95% CI)	Weighted prevalence (95% CI)	First sample	Last sample
5	174,949	824	0.47% (0.44%, 0.50%)	0.60% (0.55%, 0.71%)	18/09/2020	05/10/2020
6	160,173	1,731	1.08% (1.03%, 1.13%)	1.30% (1.21%, 1.39%)	16/10/2020	02/11/2020
7	168,181	1,299	0.77% (0.73%, 0.82%)	0.94% (0.87%, 1.01%)	13/11/2020	03/12/2020
8	167,642	2,282	1.36% (1.31%, 1.42%)	1.57% (1.49%, 1.66%)	06/01/2021*	22/01/2021
9	165,456	689	0.42% (0.39%, 0.45%)	0.49% (0.44%, 0.55%)	04/02/2021	23/02/2021
10	140,844	227	0.16% (0.14%, 0.18%)	0.20% (0.17%, 0.23%)	11/03/2021	30/03/2021

* Includes small number of samples from previous days.

4.3.2 Higher Prevalence of Test-positivity in Children

Periods of high prevalence appeared to be driven by younger age groups (Figure 4.2, Table 4.2). Although swab positivity increased in all age groups between rounds 5 and 6 it fell slightly on average between rounds 6 and 7. However, during this period prevalence among primary (5-12 years old) and secondary (13 to 17 years old), aged children continued to increase, which may partly be explained by the fact that primary and secondary schools remained open during the second lockdown between November and December 2020 (Prime Minister's Office & Street, 2020). All age groups reached their highest prevalence in round 8, before dropping from round 9 to round 10. Across the second wave, participants aged below 35 years had a higher prevalence than those aged 35 years and above years for all six rounds, further supporting the notion that younger people contributed substantially to transmission during the second wave prior to the widespread uptake of vaccination in this age group, which was first offered on 20 May 2021 (England, 2020).

Table 4.2 Unweighted and weighted prevalence by age groups and regions from rounds 5 to 10.

A. Rounds 5 to 6.

Variable	Category	Round 5				Round 6			
		Positive	Total	Unweighted Prevalence	Weighted Prevalence	Positive	Total	Unweighted Prevalence	Weighted Prevalence
Age	05-12	59	13,771	0.43% (0.33% , 0.55%)	0.58% (0.47% , 0.73%)	113	11,245	1.00% (0.83% , 1.21%)	1.27% (0.96% , 1.68%)
	13-17	59	10,480	0.56% (0.43% , 0.73%)	0.60% (0.43% , 0.84%)	119	8,616	1.38% (1.15% , 1.65%)	1.72% (1.33% , 2.21%)
	18-24	83	6,852	1.21% (0.97% , 1.50%)	0.70% (0.49% , 1.00%)	120	6,308	1.90% (1.58% , 2.27%)	2.38% (1.79% , 3.16%)
	25-34	86	15,683	0.55% (0.44% , 0.68%)	1.59% (1.18% , 2.13%)	206	15,242	1.35% (1.17% , 1.55%)	1.56% (1.30% , 1.86%)
	35-44	121	22,954	0.53% (0.44% , 0.63%)	0.69% (0.53% , 0.90%)	214	21,003	1.02% (0.89% , 1.16%)	1.25% (1.05% , 1.48%)
	45-54	157	29,380	0.53% (0.45% , 0.62%)	0.58% (0.48% , 0.70%)	313	26,897	1.16% (1.04% , 1.30%)	1.26% (1.10% , 1.44%)
	55-64	116	31,884	0.36% (0.30% , 0.44%)	0.37% (0.30% , 0.46%)	326	29,752	1.10% (0.98% , 1.22%)	1.20% (1.05% , 1.38%)
	65+	143	43,944	0.33% (0.27% , 0.38%)	0.35% (0.28% , 0.43%)	320	41,110	0.78% (0.70% , 0.87%)	0.82% (0.71% , 0.94%)
Region	South East	104	39,496	0.26% (0.22% , 0.32%)	0.29% (0.23% , 0.37%)	244	36,164	0.67% (0.59% , 0.76%)	0.69% (0.60% , 0.80%)
	North East	61	6,702	0.91% (0.70% , 1.17%)	1.06% (0.80% , 1.42%)	89	5,846	1.52% (1.22% , 1.87%)	1.49% (1.17% , 1.89%)
	North West	199	19,468	1.02% (0.89% , 1.17%)	1.21% (1.01% , 1.46%)	372	17,866	2.08% (1.88% , 2.30%)	2.39% (2.10% , 2.73%)
	Yorkshire and The Humber	74	11,658	0.63% (0.50% , 0.80%)	0.84% (0.60% , 1.17%)	189	10,697	1.77% (1.53% , 2.03%)	2.32% (1.90% , 2.81%)
	East Midlands	101	21,897	0.46% (0.38% , 0.56%)	0.56% (0.43% , 0.72%)	233	20,458	1.14% (1.00% , 1.29%)	1.25% (1.07% , 1.45%)
	West Midlands	70	16,343	0.43% (0.33% , 0.54%)	0.60% (0.40% , 0.89%)	196	14,860	1.32% (1.14% , 1.52%)	1.59% (1.18% , 2.13%)
	East of England	81	25,625	0.32% (0.25% , 0.39%)	0.30% (0.23% , 0.38%)	150	23,231	0.65% (0.55% , 0.76%)	0.69% (0.57% , 0.83%)
	London	79	17,499	0.45% (0.36% , 0.56%)	0.54% (0.42% , 0.70%)	141	15,757	0.89% (0.75% , 1.05%)	0.97% (0.79% , 1.18%)
	South West	55	16,260	0.34% (0.25% , 0.44%)	0.33% (0.23% , 0.46%)	117	15,294	0.77% (0.63% , 0.92%)	0.80% (0.64% , 0.99%)

B. Rounds 7 to 8.

Variable	Category	Round 7				Round 8			
		Positive	Total	Unweighted Prevalence	Weighted Prevalence	Positive	Total	Unweighted Prevalence	Weighted Prevalence
Age	05-12	147	11,873	1.24% (1.05% , 1.45%)	1.44% (1.16% , 1.77%)	170	11,727	1.45% (1.25% , 1.68%)	1.59% (1.32% , 1.93%)
	13-17	172	9,861	1.74% (1.50% , 2.02%)	2.04% (1.69% , 2.46%)	172	8,947	1.92% (1.66% , 2.23%)	2.25% (1.85% , 2.73%)
	18-24	86	6,475	1.33% (1.06% , 1.64%)	1.41% (1.04% , 1.91%)	156	6,753	2.31% (1.98% , 2.70%)	2.44% (1.96% , 3.03%)
	25-34	130	14,758	0.88% (0.74% , 1.05%)	1.00% (0.80% , 1.23%)	263	15,031	1.75% (1.55% , 1.97%)	1.95% (1.68% , 2.26%)
	35-44	160	20,846	0.77% (0.65% , 0.90%)	0.89% (0.73% , 1.09%)	314	21,764	1.44% (1.29% , 1.61%)	1.62% (1.41% , 1.85%)
	45-54	220	27,338	0.80% (0.70% , 0.92%)	0.86% (0.73% , 1.01%)	412	27,982	1.47% (1.34% , 1.62%)	1.61% (1.43% , 1.81%)
	55-64	188	31,396	0.60% (0.52% , 0.69%)	0.64% (0.53% , 0.76%)	413	31,989	1.29% (1.17% , 1.42%)	1.29% (1.15% , 1.45%)
	65+	196	45,634	0.43% (0.37% , 0.49%)	0.50% (0.41% , 0.61%)	382	43,449	0.88% (0.80% , 0.97%)	0.93% (0.82% , 1.05%)
Region	South East	254	38,296	0.66% (0.58% , 0.75%)	0.73% (0.63% , 0.84%)	578	39,156	1.48% (1.36% , 1.60%)	1.61% (1.46% , 1.77%)
	North East	53	5,673	0.93% (0.70% , 1.22%)	0.94% (0.66% , 1.35%)	62	5,731	1.08% (0.84% , 1.38%)	1.22% (0.87% , 1.69%)
	North West	157	18,337	0.86% (0.73% , 1.00%)	1.02% (0.83% , 1.24%)	220	17,917	1.23% (1.08% , 1.40%)	1.38% (1.16% , 1.64%)
	Yorkshire and The Humber	107	11,288	0.95% (0.78% , 1.14%)	1.25% (0.99% , 1.58%)	93	10,665	0.87% (0.71% , 1.07%)	0.80% (0.62% , 1.03%)
	East Midlands	221	22,001	1.00% (0.88% , 1.15%)	1.19% (1.00% , 1.41%)	225	21,698	1.04% (0.91% , 1.18%)	1.16% (0.99% , 1.36%)
	West Midlands	156	15,757	0.99% (0.84% , 1.16%)	1.24% (0.95% , 1.61%)	198	15,088	1.31% (1.14% , 1.51%)	1.66% (1.32% , 2.10%)
	East of England	125	24,627	0.51% (0.42% , 0.60%)	0.58% (0.47% , 0.71%)	391	25,178	1.55% (1.41% , 1.71%)	1.78% (1.57% , 2.02%)
	London	136	15,566	0.87% (0.73% , 1.03%)	1.09% (0.89% , 1.32%)	390	15,641	2.49% (2.26% , 2.75%)	2.83% (2.53% , 3.16%)
	South West	90	16,636	0.54% (0.44% , 0.66%)	0.58% (0.45% , 0.76%)	125	16,568	0.75% (0.63% , 0.90%)	0.87% (0.71% , 1.08%)

C. Rounds 9 to 10.

Variable	Category	Round 9				Round 10			
		Positive	Total	Unweighted Prevalence	Weighted Prevalence*	Positive	Total	Unweighted Prevalence	Weighted Prevalence*
Age	05-12	81	13,605	0.60% (0.48% , 0.74%)	0.72% (0.54% , 0.95%)	33	10,750	0.31% (0.21% , 0.43%)	0.41% (0.27% , 0.62%)
	13-17	45	10,155	0.44% (0.33% , 0.59%)	0.57% (0.38% , 0.87%)	14	7,891	0.18% (0.10% , 0.30%)	0.19% (0.11% , 0.34%)
	18-24	40	6,788	0.59% (0.43% , 0.80%)	0.71% (0.45% , 1.12%)	13	5,799	0.22% (0.12% , 0.38%)	0.29% (0.16% , 0.53%)
	25-34	68	14,801	0.46% (0.36% , 0.58%)	0.46% (0.34% , 0.61%)	21	12,203	0.17% (0.11% , 0.26%)	0.18% (0.11% , 0.31%)
	35-44	94	21,088	0.45% (0.36% , 0.55%)	0.57% (0.42% , 0.77%)	35	17,339	0.20% (0.14% , 0.28%)	0.21% (0.14% , 0.31%)
	45-54	132	27,595	0.48% (0.40% , 0.57%)	0.50% (0.41% , 0.62%)	37	22,258	0.17% (0.12% , 0.23%)	0.20% (0.13% , 0.30%)
	55-64	135	31,727	0.43% (0.36% , 0.50%)	0.49% (0.40% , 0.60%)	39	25,811	0.15% (0.11% , 0.21%)	0.17% (0.12% , 0.25%)
	65+	94	39,697	0.24% (0.19% , 0.29%)	0.26% (0.21% , 0.33%)	35	38,793	0.09% (0.06% , 0.13%)	0.09% (0.06% , 0.13%)
Region	South East	119	36,642	0.32% (0.27% , 0.39%)	0.36% (0.29% , 0.44%)	18	31,128	0.06% (0.03% , 0.09%)	0.07% (0.04% , 0.12%)
	North East	41	5,836	0.70% (0.52% , 0.95%)	0.69% (0.49% , 0.98%)	16	4,935	0.32% (0.19% , 0.53%)	0.41% (0.24% , 0.70%)
	North West	112	19,402	0.58% (0.48% , 0.69%)	0.69% (0.54% , 0.88%)	41	16,106	0.25% (0.18% , 0.35%)	0.31% (0.21% , 0.45%)
	Yorkshire and The Humber	42	11,218	0.37% (0.28% , 0.51%)	0.60% (0.37% , 0.96%)	28	9,601	0.29% (0.19% , 0.42%)	0.36% (0.22% , 0.57%)
	East Midlands	103	21,193	0.49% (0.40% , 0.59%)	0.59% (0.45% , 0.77%)	34	17,934	0.19% (0.13% , 0.26%)	0.19% (0.13% , 0.28%)
	West Midlands	60	15,478	0.39% (0.30% , 0.50%)	0.36% (0.25% , 0.52%)	26	12,926	0.20% (0.13% , 0.29%)	0.20% (0.13% , 0.31%)
	East of England	86	23,672	0.36% (0.29% , 0.45%)	0.47% (0.36% , 0.60%)	27	20,103	0.13% (0.09% , 0.20%)	0.15% (0.10% , 0.24%)
	London	91	15,492	0.59% (0.48% , 0.72%)	0.60% (0.48% , 0.76%)	22	14,062	0.16% (0.10% , 0.24%)	0.16% (0.10% , 0.26%)
	South West	35	16,523	0.21% (0.15% , 0.29%)	0.23% (0.15% , 0.36%)	15	14,049	0.11% (0.06% , 0.18%)	0.13% (0.07% , 0.26%)

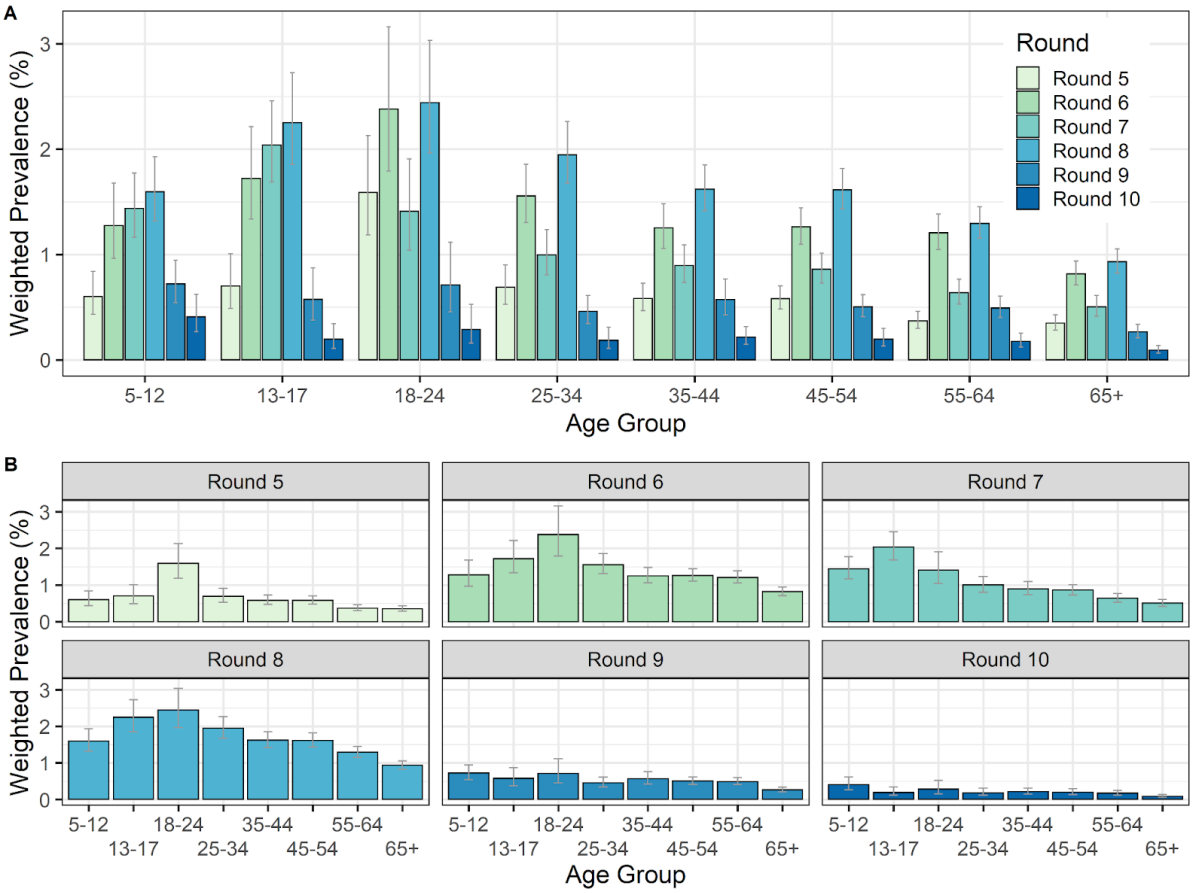


Figure 4.2 Weighted prevalence of swab positivity by age groups.

A. Weighted prevalence of swab-positivity by age group for rounds 5 to 10. **B**. Weighted prevalence of swab-positivity by age group and by round for rounds 5 to 10. Bars show 95% confidence intervals.

4.3.3 Heterogeneous Regional Pattern in Prevalence of SARS-CoV-2 Infection

We observed substantial sustained heterogeneity in the prevalence of infection across the nine regions of England. The weighted prevalence of SARS-CoV-2 was high at the start of the wave in five northern regions. In round 5 (18th September to 5th October 2020), the weighted prevalence was high in the: North West at 1.21% (0.99%, 1.43%), North East at 1.06% (0.76%, 1.37%), and Yorkshire and The Humber at 0.84% (0.56%, 1.12%), compared to lower prevalence in Midlands and Southern regions (Figure 4.3, Table 4.2). These differences persisted during round 6, with Yorkshire and The Humber increasing to 2.32% (1.90%, 2.81%) but were then less apparent in round 7.

However, the wave that would be later explained by the emergence of the Alpha variant (Volz et al., 2021) then caused southern regions to increase in prevalence. In round 8 between 6th and 22nd January 2021, the prevalence in London was 2.83% (2.53%, 3.16%) which was the highest among regions, with surrounding areas of South East and East of England also increasing high at 1.61% (1.46%, 1.77%) and 1.78% (1.57%, 2.02%) respectively.

There were substantial falls in regional weighted prevalence after the start of the third national lockdown in January (round 8). However, northern regions in England showed higher prevalence again: the highest prevalence in round 9 was found in North East at 0.41% (0.24%, 0.70%), followed by Yorkshire and The Humber at 0.36% (0.22%, 0.57%) and North West 0.31% (0.21%, 0.45%). The prevalence in London decreased to 0.16% (0.10%, 0.26%) from 0.60% (0.48%, 0.76%) in round 9, becoming the third lowest, while the lowest prevalence was in South East from 0.36% (0.29%, 0.44%) to 0.07% (0.04%, 0.12%). Similar levels of regional heterogeneity of infection risk have been observed in Spain, Italy, and Germany (Jalilian & Mateu, 2021). Also, the detailed characteristics of Alpha emerging

in the South East and then spreading via London to the rest of England have now been described using detailed genomic data (Kraemer et al., 2021).

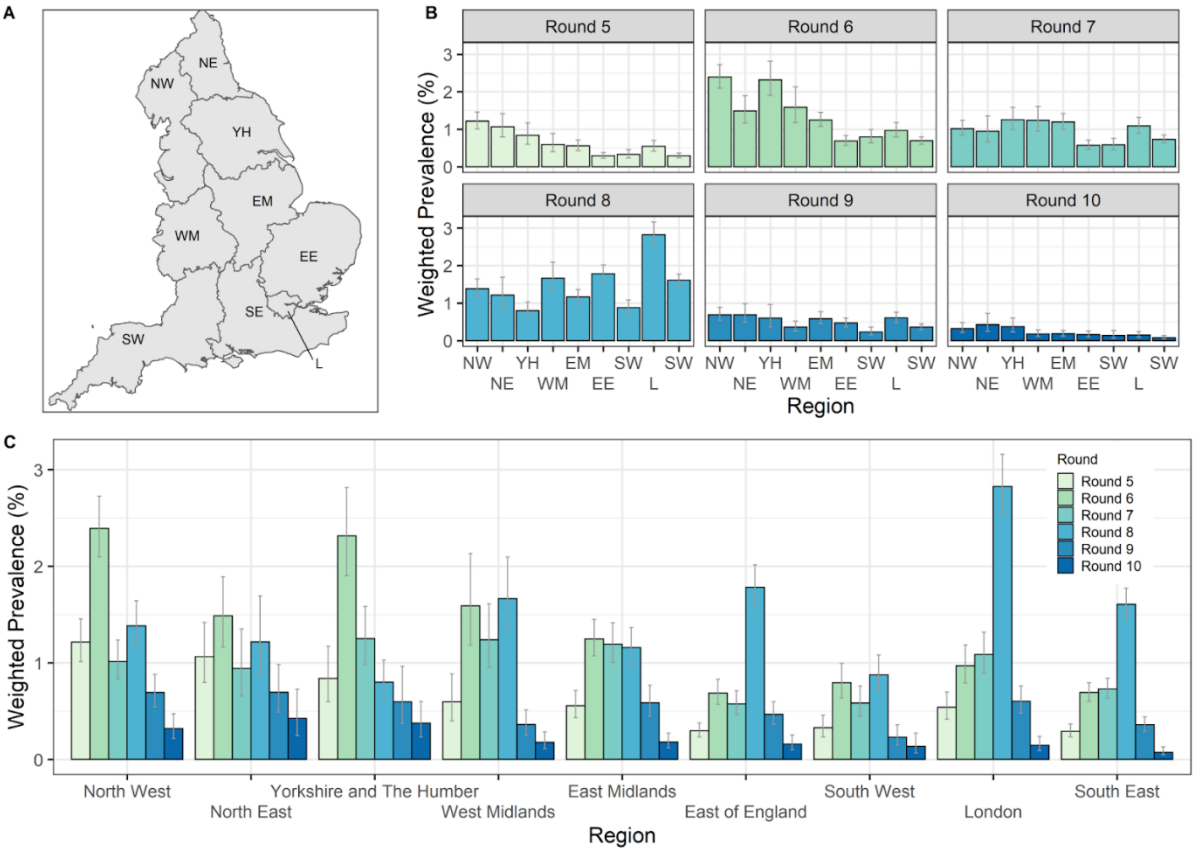


Figure 4.3 The weighted prevalence of swab positivity by geographical regions in England.

A. Map of England showing the location of the regions. **B**. Weighted prevalence of swab-positivity by region for rounds 5 to 10. C. Weighted prevalence of swab-positivity by region and by round for rounds 5 to 10. Bars show 95% confidence intervals. NW: North West, NE: North East, YH: Yorkshire and The Humble, WM: West Midlands, EM: East Midlands, EE: East of England, SW: South West, L: London and SE: South East.

4.3.4 Interaction of Weighted Prevalence between Age Group and Geographic Regions in England

Overall, the higher prevalence was always observed in younger adults aged below 35 years across rounds 5 to 10 in each region, whereas the highest prevalence was never observed among the older group (Figure 4.4, Table 4.3). This pattern was consistent with the overall age patterns showed in Figure 4.2, suggesting that the increased infections were very likely to start in younger age groups.

Variation in the patterns of prevalence of age groups by region was observed. In four northern regions North West, North East, Yorkshire and the Humble and West Midlands, the prevalence of participants aged below 35 years remained higher than older age across rounds 5 to 10, whereas participants aged between 13 and 24 years showed much higher prevalence in East Midlands, East of England, London and South East. In contrast, age groups in South West did not reflect obvious differences in weighted prevalence.

As observed in Figure 4.3, regional synchrony tends to be stronger than age synchrony. When the prevalence of one of the age groups in a region was high, all other age groups in the same region tended to have a higher prevalence. For instance, all age groups had the highest prevalence in London in round 8, whereas most of the age groups were the highest in Yorkshire and The Humber in round 6. South West and South West remained relatively low prevalence across six rounds, and the age groups in these two regions also consistently showed lower prevalence across the second wave.

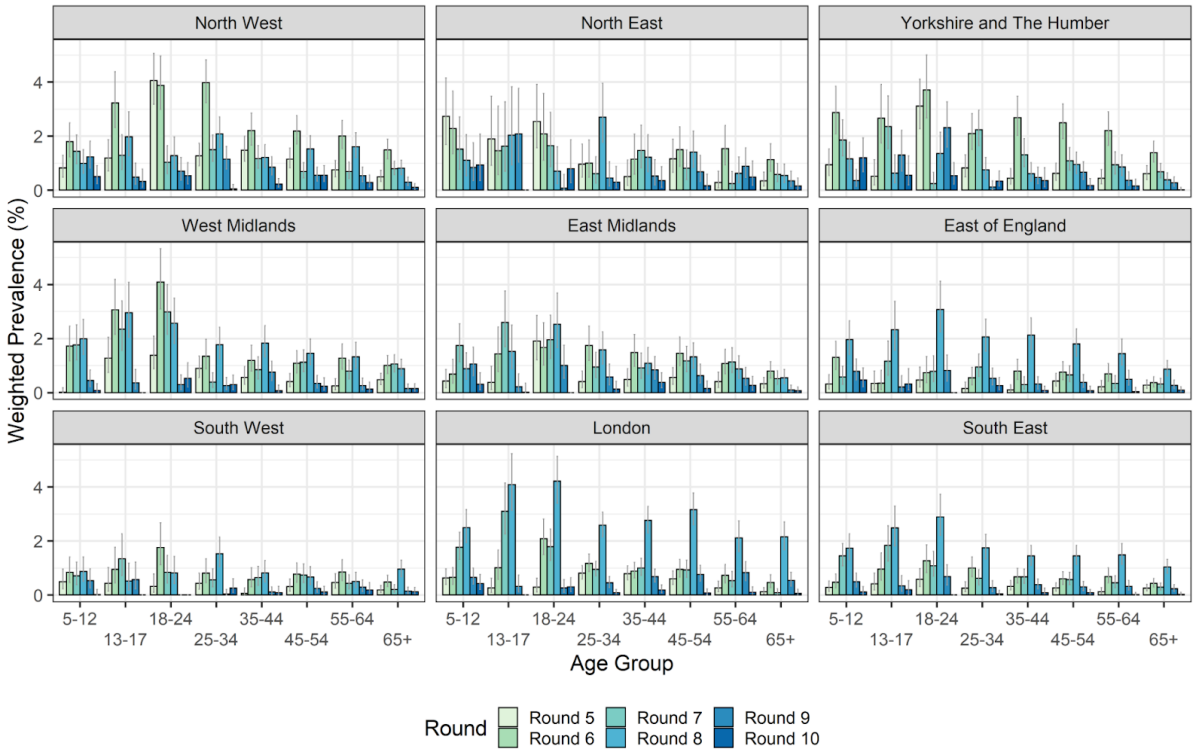


Figure 4.4 Nested weighted prevalence by age and region for rounds 5 to 10.

Bars show 95% credible intervals.

Table 4.3 Nested weighted prevalence by region and age groups.

A. Rounds 5 to 6.

Region	Age	Round 5			Round 6		
		Positive	Total	Weighted Prevalence	Positive	Total	Weighted Prevalence
North West	05-12	10	1,424	0.82% (0.48% , 1.29%)	21	1,095	1.79% (1.23% , 2.49%)
	13-17	11	1,149	1.19% (0.71% , 1.86%)	26	936	3.23% (2.32% , 4.38%)
	18-24	26	786	4.05% (3.18% , 5.06%)	26	712	3.87% (2.99% , 4.96%)
	25-34	23	1,678	1.27% (0.89% , 1.73%)	51	1,543	3.97% (3.24% , 4.82%)
	35-44	36	2,460	1.47% (1.07% , 1.99%)	38	2,119	2.20% (1.66% , 2.85%)
	45-54	32	3,189	1.15% (0.83% , 1.56%)	62	3,008	2.18% (1.70% , 2.76%)
	55-64	35	3,723	0.75% (0.49% , 1.10%)	67	3,535	2.00% (1.54% , 2.58%)
	65+	26	5,059	0.49% (0.32% , 0.73%)	81	4,918	1.49% (1.16% , 1.89%)
North East	05-12	8	439	2.73% (1.68% , 4.15%)	9	387	2.28% (1.35% , 3.67%)
	13-17	7	373	1.90% (0.91% , 3.47%)	5	289	1.45% (0.57% , 3.10%)
	18-24	9	312	2.54% (1.56% , 3.92%)	7	227	2.08% (1.11% , 3.57%)
	25-34	7	616	0.96% (0.48% , 1.70%)	7	538	1.00% (0.48% , 1.85%)
	35-44	5	820	0.50% (0.17% , 1.11%)	7	712	1.15% (0.59% , 2.07%)
	45-54	14	1,059	1.16% (0.66% , 1.88%)	15	943	1.49% (0.88% , 2.34%)
	55-64	5	1,312	0.28% (0.08% , 0.70%)	16	1,185	1.54% (0.93% , 2.39%)
	65+	6	1,771	0.34% (0.15% , 0.67%)	23	1,565	1.12% (0.70% , 1.71%)
Yorkshire and The	05-12	4	823	0.94% (0.54% , 1.53%)	17	677	2.87% (2.09% , 3.84%)
	13-17	5	687	0.51% (0.19% , 1.13%)	10	557	2.66% (1.75% , 3.91%)
	18-24	10	475	3.11% (2.28% , 4.11%)	8	375	3.70% (2.67% , 5.00%)
	25-34	8	973	0.82% (0.49% , 1.31%)	15	947	2.09% (1.51% , 2.83%)
	35-44	6	1,384	0.44% (0.22% , 0.79%)	28	1,283	2.68% (2.03% , 3.47%)
	45-54	13	1,868	0.62% (0.36% , 0.99%)	38	1,787	2.49% (1.90% , 3.19%)
	55-64	10	2,298	0.44% (0.23% , 0.76%)	37	2,159	2.20% (1.65% , 2.89%)
	65+	18	3,150	0.61% (0.40% , 0.91%)	36	2,912	1.39% (1.04% , 1.81%)
West Midlands	05-12	1	1,249	0.02% (0.00% , 0.18%)	10	975	1.72% (1.17% , 2.46%)
	13-17	7	1,028	1.27% (0.75% , 2.04%)	21	865	3.06% (2.15% , 4.18%)
	18-24	6	657	1.38% (0.89% , 2.09%)	12	586	4.08% (3.09% , 5.32%)
	25-34	11	1,424	0.89% (0.55% , 1.35%)	17	1,285	1.34% (0.88% , 1.97%)
	35-44	9	1,973	0.56% (0.31% , 0.96%)	18	1,827	1.19% (0.79% , 1.75%)
	45-54	14	2,805	0.41% (0.21% , 0.70%)	37	2,530	1.08% (0.72% , 1.55%)
	55-64	9	2,908	0.25% (0.10% , 0.52%)	43	2,781	1.27% (0.86% , 1.80%)
	65+	13	4,299	0.47% (0.30% , 0.72%)	38	4,011	1.00% (0.71% , 1.37%)
East Midlands	05-12	5	1,676	0.43% (0.18% , 0.86%)	8	1,455	0.68% (0.35% , 1.23%)
	13-17	5	1,282	0.38% (0.11% , 0.97%)	16	1,136	1.43% (0.78% , 2.42%)
	18-24	15	853	1.91% (1.21% , 2.85%)	15	818	1.67% (1.02% , 2.59%)
	25-34	6	1,817	0.40% (0.18% , 0.79%)	30	1,791	1.74% (1.19% , 2.46%)
	35-44	15	2,675	0.48% (0.24% , 0.88%)	31	2,522	1.48% (0.97% , 2.14%)
	45-54	20	3,747	0.55% (0.31% , 0.93%)	45	3,351	1.45% (0.99% , 2.06%)
	55-64	15	4,136	0.40% (0.19% , 0.75%)	44	3,893	1.08% (0.69% , 1.60%)
	65+	20	5,711	0.32% (0.17% , 0.55%)	44	5,492	0.80% (0.53% , 1.14%)
East of England	05-12	8	2,102	0.32% (0.13% , 0.66%)	18	1,702	1.31% (0.86% , 1.89%)
	13-17	8	1,563	0.33% (0.11% , 0.79%)	5	1,266	0.34% (0.11% , 0.82%)
	18-24	5	1,039	0.47% (0.21% , 0.94%)	9	941	0.73% (0.36% , 1.33%)
	25-34	4	2,147	0.15% (0.04% , 0.39%)	14	2,091	0.54% (0.29% , 0.93%)
	35-44	5	3,391	0.10% (0.02% , 0.28%)	19	3,043	0.79% (0.49% , 1.23%)
	45-54	17	4,336	0.43% (0.24% , 0.71%)	29	3,874	0.76% (0.47% , 1.14%)
	55-64	12	4,553	0.21% (0.09% , 0.45%)	32	4,361	0.69% (0.42% , 1.07%)
	65+	22	6,494	0.28% (0.15% , 0.47%)	24	5,953	0.36% (0.21% , 0.59%)
South West	05-12	6	1,178	0.49% (0.23% , 0.95%)	7	1,038	0.84% (0.46% , 1.40%)
	13-17	5	901	0.43% (0.14% , 1.02%)	8	781	0.95% (0.45% , 1.76%)
	18-24	2	545	0.32% (0.10% , 0.78%)	13	569	1.76% (1.12% , 2.67%)
	25-34	4	1,291	0.43% (0.21% , 0.81%)	11	1,281	0.81% (0.46% , 1.34%)
	35-44	3	1,873	0.06% (0.00% , 0.26%)	13	1,799	0.57% (0.30% , 1.01%)
	45-54	13	2,665	0.31% (0.14% , 0.59%)	18	2,529	0.76% (0.47% , 1.18%)
	55-64	13	3,073	0.46% (0.25% , 0.79%)	26	2,924	0.85% (0.53% , 1.30%)
	65+	9	4,734	0.18% (0.09% , 0.35%)	21	4,373	0.48% (0.31% , 0.73%)
London	05-12	8	1,629	0.63% (0.39% , 0.97%)	10	1,177	0.65% (0.39% , 1.01%)
	13-17	3	1,082	0.26% (0.09% , 0.62%)	9	812	1.01% (0.57% , 1.66%)
	18-24	2	716	0.29% (0.11% , 0.62%)	12	703	2.08% (1.50% , 2.81%)
	25-34	15	2,554	0.81% (0.59% , 1.09%)	29	2,622	1.17% (0.87% , 1.52%)
	35-44	24	3,167	0.79% (0.56% , 1.08%)	25	2,918	0.88% (0.63% , 1.20%)
	45-54	17	2,900	0.60% (0.39% , 0.88%)	28	2,653	0.94% (0.66% , 1.32%)
	55-64	6	2,616	0.26% (0.12% , 0.51%)	17	2,292	0.73% (0.45% , 1.13%)
	65+	4	2,835	0.12% (0.04% , 0.27%)	11	2,580	0.46% (0.27% , 0.75%)
South East	05-12	9	3,251	0.28% (0.13% , 0.52%)	13	2,739	0.47% (0.26% , 0.78%)
	13-17	8	2,415	0.41% (0.19% , 0.81%)	19	1,974	0.96% (0.55% , 1.56%)
	18-24	8	1,469	0.58% (0.31% , 0.97%)	18	1,377	1.27% (0.82% , 1.85%)
	25-34	8	3,183	0.26% (0.12% , 0.49%)	32	3,144	0.99% (0.68% , 1.41%)
	35-44	18	5,211	0.32% (0.18% , 0.54%)	35	4,780	0.67% (0.44% , 0.99%)
	45-54	17	6,811	0.26% (0.14% , 0.45%)	41	6,222	0.59% (0.39% , 0.86%)
	55-64	11	7,265	0.13% (0.05% , 0.27%)	44	6,622	0.68% (0.45% , 1.00%)
	65+	25	9,891	0.25% (0.15% , 0.39%)	42	9,306	0.43% (0.29% , 0.61%)

B. Rounds 7 to 8.

Region	Age	Round 7			Round 8		
		Positive	Total	Weighted Prevalence	Positive	Total	Weighted Prevalence
North West	05-12	14	1,112	1.43% (0.98% , 2.04%)	12	1,123	0.99% (0.63% , 1.50%)
	13-17	13	1,000	1.28% (0.76% , 2.05%)	15	863	1.97% (1.28% , 2.89%)
	18-24	7	668	1.03% (0.61% , 1.65%)	8	694	1.28% (0.79% , 1.96%)
	25-34	19	1,448	1.50% (1.07% , 2.04%)	24	1,493	2.09% (1.58% , 2.71%)
	35-44	22	2,042	1.17% (0.81% , 1.64%)	23	2,139	1.21% (0.84% , 1.68%)
	45-54	22	2,916	0.69% (0.45% , 1.02%)	43	2,947	1.52% (1.13% , 2.01%)
	55-64	26	3,680	0.69% (0.43% , 1.04%)	53	3,651	1.61% (1.20% , 2.12%)
	65+	34	5,471	0.78% (0.57% , 1.07%)	42	5,007	0.81% (0.57% , 1.10%)
North East	05-12	5	337	1.52% (0.77% , 2.71%)	6	358	1.11% (0.53% , 2.06%)
	13-17	6	295	1.62% (0.69% , 3.27%)	4	252	2.04% (0.93% , 3.82%)
	18-24	4	216	1.64% (0.86% , 2.88%)	3	233	0.70% (0.24% , 1.61%)
	25-34	2	526	0.61% (0.26% , 1.24%)	8	491	2.70% (1.78% , 3.95%)
	35-44	12	656	1.47% (0.83% , 2.40%)	9	733	1.21% (0.68% , 2.05%)
	45-54	8	850	0.81% (0.39% , 1.49%)	10	921	1.40% (0.85% , 2.18%)
	55-64	4	1,104	0.24% (0.06% , 0.68%)	9	1,173	0.62% (0.28% , 1.23%)
	65+	12	1,689	0.58% (0.31% , 0.99%)	13	1,570	0.54% (0.28% , 0.96%)
Yorkshire and The	05-12	11	750	1.85% (1.26% , 2.60%)	8	668	1.16% (0.72% , 1.78%)
	13-17	12	632	2.36% (1.54% , 3.48%)	4	550	0.63% (0.27% , 1.31%)
	18-24	3	408	0.24% (0.06% , 0.65%)	9	404	1.35% (0.80% , 2.14%)
	25-34	18	901	2.23% (1.64% , 2.96%)	9	852	0.75% (0.43% , 1.21%)
	35-44	16	1,241	1.31% (0.87% , 1.90%)	9	1,239	0.61% (0.33% , 1.02%)
	45-54	17	1,831	1.09% (0.74% , 1.57%)	16	1,709	0.94% (0.61% , 1.41%)
	55-64	15	2,239	0.94% (0.60% , 1.42%)	24	2,248	0.86% (0.54% , 1.31%)
	65+	15	3,286	0.68% (0.45% , 0.98%)	14	2,995	0.38% (0.22% , 0.64%)
West Midlands	05-12	18	1,050	1.76% (1.21% , 2.50%)	13	965	1.98% (1.41% , 2.71%)
	13-17	17	920	2.35% (1.56% , 3.39%)	17	802	2.95% (2.05% , 4.07%)
	18-24	11	623	2.98% (2.16% , 4.00%)	10	607	2.56% (1.82% , 3.49%)
	25-34	10	1,327	0.39% (0.19% , 0.73%)	20	1,263	1.77% (1.25% , 2.41%)
	35-44	14	1,819	0.84% (0.50% , 1.32%)	25	1,815	1.83% (1.32% , 2.47%)
	45-54	27	2,582	1.12% (0.77% , 1.58%)	32	2,538	1.45% (1.03% , 1.98%)
	55-64	26	3,000	0.79% (0.50% , 1.21%)	42	2,943	1.32% (0.90% , 1.86%)
	65+	33	4,436	1.05% (0.77% , 1.40%)	39	4,155	0.88% (0.62% , 1.24%)
East Midlands	05-12	21	1,468	1.74% (1.14% , 2.55%)	15	1,512	0.88% (0.50% , 1.47%)
	13-17	28	1,364	2.60% (1.70% , 3.76%)	15	1,185	1.52% (0.88% , 2.50%)
	18-24	19	844	1.95% (1.26% , 2.86%)	21	790	2.52% (1.67% , 3.68%)
	25-34	18	1,852	0.94% (0.56% , 1.49%)	27	1,756	1.58% (1.06% , 2.25%)
	35-44	19	2,484	0.91% (0.53% , 1.46%)	30	2,629	1.09% (0.68% , 1.67%)
	45-54	45	3,610	1.17% (0.78% , 1.71%)	48	3,665	1.32% (0.91% , 1.84%)
	55-64	41	4,211	1.13% (0.74% , 1.66%)	37	4,239	0.87% (0.53% , 1.36%)
	65+	30	6,168	0.51% (0.31% , 0.80%)	32	5,922	0.55% (0.33% , 0.85%)
East of England	05-12	12	1,840	0.57% (0.31% , 0.97%)	31	1,806	1.95% (1.41% , 2.65%)
	13-17	19	1,498	1.16% (0.64% , 1.91%)	32	1,387	2.33% (1.57% , 3.37%)
	18-24	9	984	0.78% (0.40% , 1.36%)	23	1,037	3.07% (2.25% , 4.12%)
	25-34	15	2,056	0.95% (0.60% , 1.43%)	44	2,274	2.05% (1.53% , 2.71%)
	35-44	11	3,087	0.29% (0.13% , 0.59%)	63	3,308	2.13% (1.60% , 2.76%)
	45-54	25	4,102	0.66% (0.41% , 1.00%)	72	4,166	1.80% (1.34% , 2.36%)
	55-64	13	4,403	0.33% (0.16% , 0.63%)	71	4,756	1.44% (1.02% , 1.97%)
	65+	21	6,657	0.32% (0.18% , 0.52%)	55	6,444	0.86% (0.62% , 1.18%)
South West	05-12	8	1,114	0.71% (0.38% , 1.22%)	9	1,085	0.87% (0.51% , 1.41%)
	13-17	14	940	1.33% (0.72% , 2.26%)	5	842	0.52% (0.18% , 1.18%)
	18-24	6	608	0.84% (0.44% , 1.46%)	8	663	0.82% (0.42% , 1.43%)
	25-34	9	1,293	0.56% (0.30% , 0.97%)	17	1,290	1.53% (1.04% , 2.14%)
	35-44	8	1,880	0.64% (0.36% , 1.06%)	14	1,888	0.81% (0.48% , 1.28%)
	45-54	18	2,533	0.74% (0.46% , 1.15%)	19	2,629	0.65% (0.39% , 1.05%)
	55-64	16	3,281	0.43% (0.22% , 0.74%)	18	3,341	0.50% (0.28% , 0.84%)
	65+	11	4,987	0.21% (0.10% , 0.38%)	35	4,830	0.95% (0.68% , 1.29%)
London	05-12	18	1,202	1.77% (1.31% , 2.33%)	27	1,231	2.49% (1.94% , 3.16%)
	13-17	23	897	3.10% (2.27% , 4.14%)	27	817	4.09% (3.14% , 5.23%)
	18-24	11	707	1.78% (1.27% , 2.44%)	32	779	4.21% (3.40% , 5.13%)
	25-34	22	2,384	0.94% (0.71% , 1.25%)	62	2,437	2.58% (2.15% , 3.07%)
	35-44	27	2,754	1.00% (0.73% , 1.35%)	74	2,855	2.76% (2.29% , 3.28%)
	45-54	20	2,578	0.93% (0.65% , 1.29%)	71	2,626	3.16% (2.60% , 3.77%)
	55-64	11	2,363	0.52% (0.30% , 0.87%)	48	2,380	2.11% (1.61% , 2.74%)
	65+	4	2,681	0.08% (0.02% , 0.23%)	49	2,516	2.15% (1.67% , 2.70%)
South East	05-12	40	3,000	1.45% (1.07% , 1.90%)	49	2,979	1.73% (1.31% , 2.26%)
	13-17	40	2,315	1.83% (1.26% , 2.56%)	53	2,249	2.48% (1.81% , 3.29%)
	18-24	16	1,417	1.08% (0.69% , 1.62%)	42	1,546	2.89% (2.21% , 3.73%)
	25-34	17	2,971	0.61% (0.37% , 0.94%)	52	3,175	1.75% (1.33% , 2.24%)
	35-44	31	4,883	0.67% (0.43% , 0.96%)	67	5,158	1.44% (1.10% , 1.84%)
	45-54	38	6,336	0.57% (0.38% , 0.84%)	101	6,781	1.44% (1.12% , 1.84%)
	55-64	36	7,115	0.45% (0.27% , 0.71%)	111	7,258	1.48% (1.14% , 1.91%)
	65+	36	10,259	0.29% (0.18% , 0.44%)	103	10,010	1.03% (0.81% , 1.32%)

C. Rounds 9 to 10.

Region	Age	Round 9			Round 10		
		Positive	Total	Weighted Prevalence	Positive	Total	Weighted Prevalence
North West	05-12	18	1,444	1.23% (0.82% , 1.81%)	4	1,127	0.30% (0.10% , 0.73%)
	13-17	8	1,196	0.48% (0.22% , 0.99%)	3	897	0.02% (0.00% , 0.35%)
	18-24	5	757	0.70% (0.37% , 1.21%)	3	652	0.00% (0.00% , 0.00%)
	25-34	18	1,602	1.15% (0.78% , 1.62%)	1	1,298	0.13% (0.02% , 0.42%)
	35-44	11	2,253	0.85% (0.56% , 1.24%)	6	1,786	0.38% (0.18% , 0.73%)
	45-54	17	3,229	0.55% (0.33% , 0.85%)	10	2,515	0.14% (0.04% , 0.42%)
	55-64	18	3,941	0.53% (0.31% , 0.85%)	10	3,154	0.27% (0.09% , 0.60%)
	65+	17	4,980	0.30% (0.17% , 0.48%)	4	4,677	0.07% (0.02% , 0.21%)
North East	05-12	5	429	0.84% (0.36% , 1.75%)	4	333	0.50% (0.24% , 0.91%)
	13-17	6	333	2.08% (1.03% , 3.76%)	0	238	0.32% (0.11% , 0.78%)
	18-24	1	234	0.07% (1.72E-05 , 0.58%)	2	212	0.53% (0.23% , 1.02%)
	25-34	2	522	0.45% (0.15% , 1.05%)	1	443	0.04% (0.00% , 0.21%)
	35-44	5	714	0.52% (0.20% , 1.13%)	2	574	0.22% (0.11% , 0.43%)
	45-54	8	941	0.68% (0.32% , 1.28%)	1	728	0.55% (0.32% , 0.91%)
	55-64	9	1,193	0.89% (0.47% , 1.56%)	3	977	0.28% (0.12% , 0.56%)
	65+	5	1,470	0.35% (0.14% , 0.69%)	3	1,430	0.10% (0.03% , 0.24%)
Yorkshire and The	05-12	3	847	0.35% (0.14% , 0.77%)	5	682	0.11% (0.03% , 0.32%)
	13-17	4	676	1.30% (0.71% , 2.21%)	4	533	0.19% (0.02% , 0.53%)
	18-24	3	451	2.31% (1.59% , 3.26%)	2	434	0.00% (0.00% , 0.00%)
	25-34	3	906	0.11% (0.03% , 0.34%)	3	751	0.04% (0.00% , 0.17%)
	35-44	3	1,296	0.47% (0.24% , 0.85%)	7	1,068	0.08% (0.04% , 0.25%)
	45-54	11	1,856	0.67% (0.40% , 1.05%)	3	1,518	0.08% (0.03% , 0.20%)
	55-64	6	2,270	0.36% (0.17% , 0.67%)	2	1,817	0% (0.00% , 0.09%)
	65+	9	2,916	0.27% (0.14% , 0.49%)	2	2,798	0.03% (0.00% , 0.10%)
West Midlands	05-12	6	1,221	0.44% (0.22% , 0.83%)	2	996	0.47% (0.22% , 0.92%)
	13-17	5	961	0.36% (0.12% , 0.85%)	0	762	0.32% (0.10% , 0.88%)
	18-24	4	640	0.29% (0.11% , 0.65%)	4	495	0.00% (0.00% , 0.00%)
	25-34	4	1,305	0.26% (0.11% , 0.55%)	4	1,019	0.26% (0.10% , 0.55%)
	35-44	11	1,906	0.76% (0.45% , 1.16%)	1	1,523	0.07% (0.01% , 0.23%)
	45-54	11	2,610	0.34% (0.17% , 0.63%)	6	2,018	0.06% (0.01% , 0.24%)
	55-64	11	3,051	0.26% (0.10% , 0.54%)	5	2,360	0.04% (0.00% , 0.18%)
	65+	8	3,784	0.15% (0.06% , 0.32%)	4	3,753	0.08% (0.02% , 0.22%)
East Midlands	05-12	14	1,768	1.06% (0.63% , 1.68%)	5	1,378	0.07% (0.00% , 0.32%)
	13-17	4	1,300	0.22% (0.04% , 0.69%)	1	1,025	0.00% (0.00% , 0.00%)
	18-24	8	824	1.00% (0.53% , 1.75%)	0	705	0.52% (0.21% , 1.10%)
	25-34	11	1,824	0.58% (0.28% , 1.05%)	3	1,425	0.29% (0.11% , 0.65%)
	35-44	17	2,453	0.84% (0.50% , 1.33%)	7	2,037	0.06% (0.01% , 0.27%)
	45-54	23	3,445	0.63% (0.36% , 1.05%)	5	2,817	0.24% (0.09% , 0.55%)
	55-64	20	4,249	0.53% (0.28% , 0.93%)	8	3,395	0.12% (0.03% , 0.38%)
	65+	6	5,330	0.10% (0.02% , 0.28%)	5	5,152	0.15% (0.06% , 0.33%)
East of England	05-12	12	2,046	0.78% (0.46% , 1.24%)	5	1,525	0.01% (0.00% , 0.20%)
	13-17	5	1,465	0.20% (0.05% , 0.62%)	2	1,082	0.00% (0.00% , 0.00%)
	18-24	8	1,005	0.81% (0.42% , 1.39%)	0	844	0.00% (0.00% , 4.30E-2)
	25-34	9	2,092	0.53% (0.29% , 0.91%)	4	1,598	0.25% (0.09% , 0.60%)
	35-44	8	3,066	0.31% (0.13% , 0.59%)	3	2,409	0.08% (0.02% , 0.32%)
	45-54	14	3,916	0.38% (0.20% , 0.66%)	4	3,152	0.11% (0.02% , 0.32%)
	55-64	20	4,453	0.49% (0.28% , 0.82%)	3	3,755	0.18% (0.06% , 0.46%)
	65+	10	5,629	0.27% (0.15% , 0.46%)	6	5,738	0.12% (0.05% , 0.28%)
South West	05-12	3	1,278	0.53% (0.26% , 0.96%)	1	987	1.19% (0.68% , 1.93%)
	13-17	3	936	0.57% (0.23% , 1.22%)	0	787	0.55% (0.20% , 1.27%)
	18-24	0	668	1.28E-26 (0.00% , 0.00%)	0	553	0.53% (0.22% , 1.02%)
	25-34	1	1,319	0.03% (1.47E-05 , 0.20%)	1	1,073	0.33% (0.15% , 0.71%)
	35-44	3	1,897	0.11% (0.03% , 0.31%)	1	1,565	0.35% (0.12% , 0.83%)
	45-54	9	2,676	0.24% (0.10% , 0.49%)	1	2,215	0.17% (0.05% , 0.43%)
	55-64	9	3,290	0.29% (0.13% , 0.56%)	5	2,647	0.15% (0.04% , 0.40%)
	65+	7	4,459	0.14% (0.06% , 0.29%)	6	4,222	0.01% (0.00% , 0.10%)
London	05-12	7	1,451	0.65% (0.40% , 0.99%)	5	1,213	0.93% (0.35% , 2.07%)
	13-17	3	937	0.33% (0.11% , 0.73%)	0	752	0.00% (0.00% , 0.00%)
	18-24	2	708	0.26% (0.10% , 0.57%)	2	657	0.79% (0.27% , 1.87%)
	25-34	12	2,317	0.45% (0.28% , 0.68%)	3	2,179	0.29% (0.08% , 0.88%)
	35-44	18	2,782	0.68% (0.48% , 0.96%)	5	2,525	0.35% (0.10% , 0.87%)
	45-54	18	2,614	0.76% (0.51% , 1.09%)	3	2,270	0.16% (0.02% , 0.59%)
	55-64	21	2,294	0.83% (0.53% , 1.24%)	2	2,015	0.48% (0.18% , 1.08%)
	65+	10	2,389	0.54% (0.33% , 0.83%)	2	2,451	0.15% (0.03% , 0.45%)
South East	05-12	13	3,121	0.49% (0.29% , 0.81%)	2	2,509	0.42% (0.20% , 0.76%)
	13-17	7	2,351	0.34% (0.14% , 0.72%)	4	1,815	0.00% (0.00% , 0.00%)
	18-24	9	1,501	0.67% (0.37% , 1.12%)	0	1,247	0.30% (0.12% , 0.64%)
	25-34	8	2,914	0.27% (0.13% , 0.53%)	1	2,417	0.09% (0.03% , 0.20%)
	35-44	18	4,721	0.38% (0.23% , 0.59%)	3	3,852	0.18% (0.08% , 0.37%)
	45-54	21	6,308	0.30% (0.16% , 0.50%)	4	5,025	0.07% (0.00% , 0.22%)
	55-64	21	6,986	0.32% (0.18% , 0.54%)	1	5,691	0.10% (0.03% , 0.30%)
	65+	22	8,740	0.23% (0.13% , 0.38%)	3	8,572	0.07% (0.02% , 0.21%)

4.4 Discussion

The findings of the REACT-1 study show substantial changes over periods in the prevalence of swab positivity during the second wave. At the national level, there was a resurgence in weighted prevalence as the second wave of the COVID-19 pandemic started in September (Riley et al., 2020a), followed by continuously increased prevalence between October 2020 and the beginning of January 2021, before the marked decline during February and the end of March 2021. Similar trends were also seen in national surveillance data (UK Health Security Agency, 2020a).

Prevalence was higher in young people aged below 35 years, indicating the high social mixings among young people. Lower swab positivity among older adults may suggest the effectiveness of the vaccination roll-out in England, which has started to provide vaccines to adults aged below 50 years (NHS website, 2023). At the regional scale, we observed heterogeneity in regional prevalence. The weighted prevalence was higher in the north of England than in the south of England and London during September and December 2020. Higher prevalence was seen in London and surrounding areas in the East of England and South East in January 2021. After that, there were substantial falls in all regions with a slightly higher prevalence in the north of England again.

We also showed the interaction effect between regional patterns and age groups. Within each region, the increased transmission of SARS-CoV-2 in the community tends to be started by school-age children and young adults below 35 years, which is consistent with the overall prevalence distribution among age groups in England as a whole. In addition, the weighted prevalence was synchronous within regions and asynchronous across regions.

Our results contrast with some other studies covering this period of the pandemic which suggested that school-aged children and adolescents had lower susceptibility to infection

than older adults (Goldstein, Lipsitch & Cevik, 2021; Zhu et al., 2021). However, it was also reported that the cumulative prevalence of asymptomatic and symptomatic SARS-CoV-2 infection among children aged 5 and 17 years was approximately equal to that among adults aged 18-49 years (CDC, 2021). However, results from the Office for National Statistics (ONS) Coronavirus Infection Survey from April to November 2020 (Pouwels et al., 2021) show a sharp increase in infections in younger adults (17 - 24 years) from late August to October 2020, consistent with our results in rounds 5 and 6. Overall, our results support the idea that children and young adults were an important source of transmission during the COVID-19 pandemic and that this may also be the case for future epidemics of viral respiratory pathogens.

Overall Rregional patterns of estimated prevalence during the second wave share key features with a prospective multicentre cohort study in England, even though the study was only among healthcare workers (Pople et al., 2022) which is a group that does not include children. Therefore, the more representative sample used in REACT recruitment is a key strength in allowing us to observe that younger age groups were driving increases in swab positivity in a subgroup of the population at a lower risk of hospitalisation and death (Van Vinh Chau et al., 2020).

A notable limitation of the analysis resides in the relatively low count of individuals with positive swab results when estimating nested weighted prevalence by region and age group. For example, the counts of being positive in round 10 were fewer than 10 for all age groups in all regions. Although we employed the Bayesian multilevel model to reduce the impact of small sample sizes, statistical power might still be limited and lead to an increase in uncertainty. However, individuals in our study were randomly sampled across England, our estimations should remain reflective and representative of the prevalence of SARS-CoV-2 swab positivity across England as a whole, and likely represent the least biased estimate of

age specific infection prevalence across different age groups and substantial geographical scale.

In conclusion, our findings show substantial differences in community prevalence of SARS-CoV-2 infections among age groups and geographic regions in England, indicating that the surveillance for the transmission of SARS-CoV-2 in the community can help identify suitable interventions to reduce these differences related to the risk of infections.

Chapter 5 Health Inequalities in SARS-CoV-2 Infection during the Second Wave in England: REACT-1 Study

This chapter has been submitted to BMJ Public Health and is under review. The preprint is available on Medrxiv (Wang et al., 2023).

5.1 Introduction

The COVID-19 pandemic has caused a substantial increase in hospitalisations and mortality globally since its emergence in China in late December 2019 and its subsequent spread (Huang et al., 2020). Populations continue to experience multiple waves of infection, with substantial reactive changes in social distancing still occurring in some countries either spontaneously or as a result of explicit government regulations (Mahase, 2020; Moorthy et al., 2020). Social distancing measures led to increased inequalities before the pandemic hit and interacted with pre-existing structural inequalities such as sex, age, region and ethnicity (Maestripieri, 2021; Blundell et al., 2020). Investigating how inequalities influence the risk of getting infected is crucial for policymakers in understanding the complex impacts of the pandemic and shaping their response to it.

In order to enable isolation and quarantine policies, routine testing programmes were implemented nationwide, which mainly targeted individuals who reported symptoms (Furukawa, Brooks & Sobel, 2020). While targeted testing provides valuable surveillance, sub-clinical cases that are likely to contribute substantially to transmission are not captured. Furthermore, the uptake of testing may be limited by test capability and geographic, demographic, and social factors, and thus not be fully representative even among those eligible for tests. Therefore, tracking the characteristics and underlying inequalities of individuals at risk of SARS-CoV-2 infections at the population level can provide additional

evidence for policymaking (Jordan & Adab, 2020). In the UK, large representative population-based studies were conducted in the UK to better understand the underlying trend of the epidemic and risk factors (Biobank, 2020; OpenSAFELY, n.d.; Office for National Statistics, 2020a).

There is considerable evidence across many geographies that minority ethnic groups were at increased per-person risk of severe disease and death from COVID-19. However, there remains considerable uncertainty as to what extent this was driven by differential exposure (Public Health England, n.d.; Windsor-Shellard & Nasir, 2021; Baqui et al., 2020), susceptibility (Khunti et al., 2020; Ogedegbe et al., 2020; Rodriguez-Diaz et al., 2020), infection (Poblador-Plou et al., 2020; Khunti et al., 2020; Bhala et al., 2020), access to testing (Kirby, 2020; Price-Haywood et al., 2020), or differential access to health care (Mackey et al., 2021; Chowkwanyun & Reed, 2020; Price-Haywood et al., 2020). Reducing these disparities during future similar emergencies will only be possible if we understand why they arose during this pandemic.

The REal time Assessment of Community Transmission-1 (REACT-1) study was designed to be representative of the population of England as a whole (Riley et al., 2020b). The REACT-1 study tracked the spread of the virus through a total of 19 cross-sectional surveys to collect self-administered throat and nose swabs which were tested for the presence of SARS-CoV-2 by reverse transcriptase polymerase chain reaction (rt- PCR) (Riley et al., 2020b). The first round of the study began on 1 May 2020, and each new round thereafter was conducted at approximately monthly intervals, with a two-week break between the end of the previous round and the beginning of the next round. In each round up to round 12, participants were sampled randomly with a similar achieved sample size from 315 lower-tier local authorities (LTLAs). Participants were asked to take part in the study regardless of whether they were symptomatic or not.

In this chapter, we present findings on risk factors and quantify the differences in risk of SARS-CoV-2 infection for rounds 5 to 10 between mid-September 2020 and the end of March 2021. Results on risk factors from the previous four rounds between May and the beginning of September 2020 have been published previously (Riley et al., 2021a).

5.2 Methods

5.2.1 Data Processing - Socio-demographic Groups

The detailed study protocol of the REACT-1 study has been reported elsewhere (Riley et al., 2020b) and described in Chapter 4. In brief, we randomly selected individuals aged 5 years and above from the National Health Service (NHS) patient list to send personalised invitations, balanced across 315 lower-tier local authority areas (LTLAs). Participants received test kits and instructions delivered by post and conducted self-administered throat and nose swabs (parent/guardian administered at ages 5 to 12 years). In addition to the information registered on the NHS list (name, age, sex, and address), participants were requested to complete an online or telephone questionnaire to provide other detailed information, e.g., employment type, ethnicity, household size, whether they were in contact with a confirmed or suspected COVID-19 case and symptom status.

Study questionnaires are available on the REACT 1 Study Materials webpage on the Imperial College London website (Imperial College London, 2020).

The outcome is swab positivity and the method we define positivity was described in Chapter 4. In order to investigate prevalence by subgroups, we converted continuous variables such as age into categories and regrouped some categorical variables into fewer categories by combining similar categories. We divided age into eight age groups as follows: 5 to 12, 13 to 17, 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64 years, and 65 years or older. Nineteen ethnic groups were re-grouped into five broad categories, white, Asian, black, mixed ethnic

group, and other ethnic groups for some analyses. Participants were asked to report the number of children or young people between the ages of 0 and 17 and adults aged over 18 years who currently live in or spend part of the week in the household of the respondent. We obtained the total household size by adding these numbers. We then created a new categorical variable for household size with 6 categories (1 person, 2 people, 3 people, 4 people, 5 people and 6 or more people), based on the number of people in each household. For the variable on COVID-19 case contact, participants answered whether they were in contact with a confirmed/tested or suspected COVID-19 case. We did not further re-categorise this variable.

We asked participants if they worked as health care workers or care home workers or other key workers. If they did not work as a key worker, they were asked if they were engaged in one or more public-facing activities or settings (11 different activities for rounds 5 to 8, 12 activities for rounds 9 and 10). In addition, participants were asked about their employment status (employed in full-time jobs, employed in part-time jobs, self-employed, unemployed, retired from work). We combined employment and key worker answers into a new single variable which we called key worker status that could take one of the following four values: 1) Health care or care home worker, 2) Other essential/key workers, 3) Other workers, and 4) Not full-time, part-time, or self-employed.

We also regrouped symptom status during the month before the test. Participants were classified as ‘Classic COVID-19 symptoms’ if they had experienced one or more of the following symptoms in the past month: loss or change of a sense of smell or taste, persistent cough, or fever. Other symptomatic people who did not show classic COVID-19 symptoms in the past month were classified as 'Other symptoms'. Participants who did not feel unwell and did not present any of the listed symptoms were classified as 'No symptoms'.

5.2.2 Statistical Analysis

The estimation of unweighted and weighted prevalence was described in Chapter 4.

We fitted multivariable logistic regression models to investigate whether particular subgroups were more likely to test positive for SARS-CoV-2. The binary outcome variable was swab-positivity (yes or no), while age, sex, region, broad categories of employment, broad categories of ethnicity, household size and deprivation were included as explanatory variables and were mutually adjusted for the logistic regression. The odds ratio (OR) was reported along with corresponding 95% confidence intervals. The associations of swab-positivity with detailed categories of occupation and ethnicity subgroups were also quantified using logistic regression models: for detailed occupation, the model was adjusted for age, sex, region, broad categories of ethnicity, and deprivation; and for ethnicity subgroups, the model was adjusted for age, sex, region, broad categories of employment, and deprivation.

To assess the overall risk for subgroups across rounds 5 to 10, we pooled log-odds ratios from each individual round using the inverse-variance method (Viechtbauer, 2010).

5.3 Results

5.3.1 Higher Prevalence of Test-positivity in Males

Higher rates of infection were associated with the male sex (Table 5.1). The weighted prevalence of swab positivity was slightly higher in males than in females in all rounds except round 6 [male: 1.29% (1.17%, 1.43%), female: 1.31% (1.19%, 1.44%)]. This difference was maintained after accounting for age, region, key worker status, ethnicity, and deprivation, with the female sex having lower odds of infection at 0.95 (0.90, 0.99) compared with the male sex across rounds 5 to 10 (Figure 5.1A, Table 5.2). The direction of effect is consistent with other published studies in the UK. Two UK studies, one between 1 March and 28 April 2020 (Martin et al., 2020) and the other between 28 January and 4 April 2020

(de Lusignan et al., 2020), also reported that male sex was associated with a higher risk of testing positive for SARS-CoV-2 than female. Other countries have shown similar sex differences with a higher chance of infection among males (Vahidy et al., 2021; Guan et al., 2020a; Huang et al., 2020; Jiang et al., 2020; Rodriguez-Morales et al., 2020).

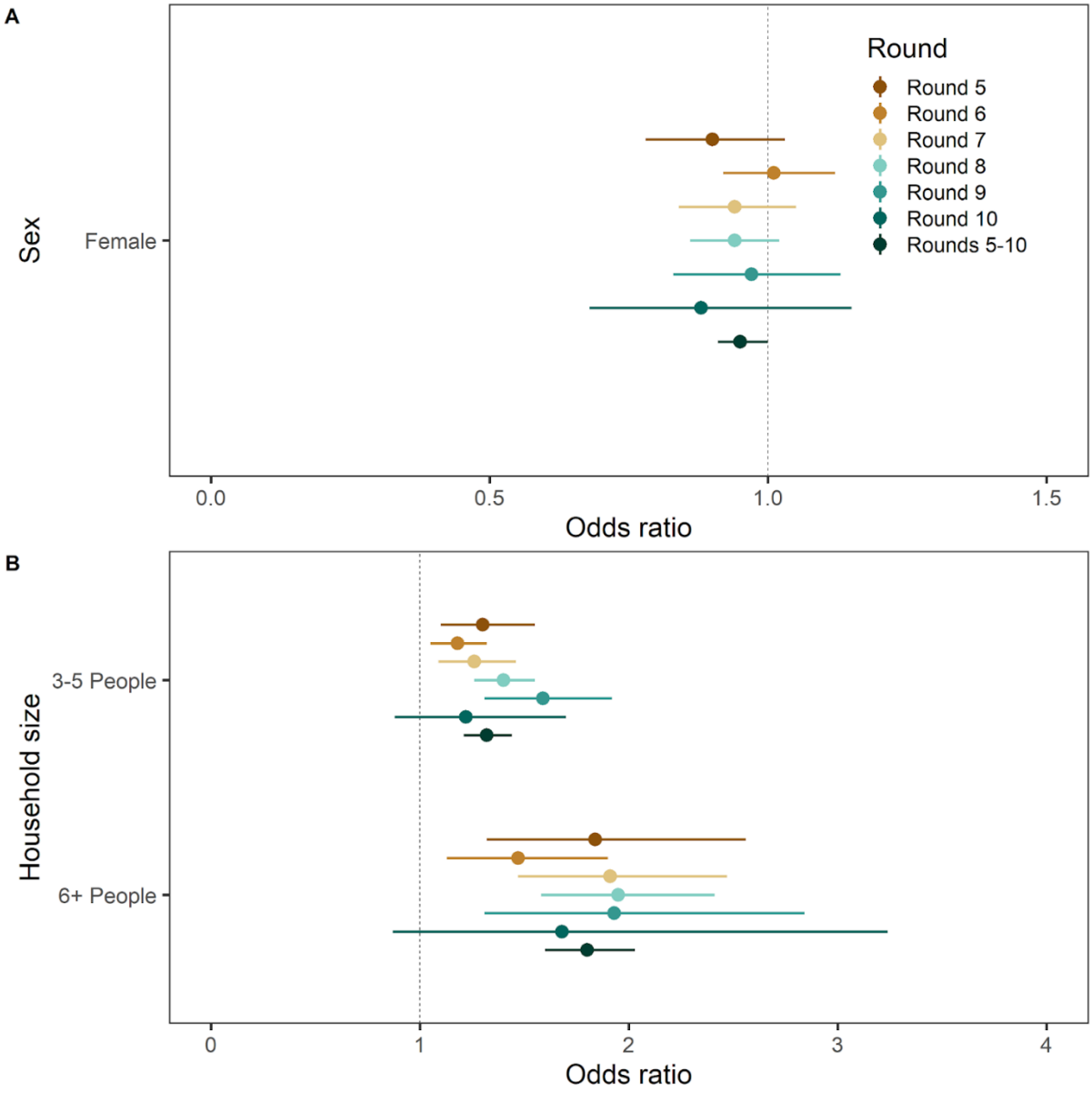


Figure 5.1 Forest plots of sex and household size.

A. Forest plot of sex for rounds 5 to 10 and an average across rounds weighted inverse to variance (rounds 5 to 10). The reference group is not shown in the figure and is male. **B**. Forest plot of household size for rounds 5 to 10 and an average weighted across rounds inverse to variance (rounds 5 to 10). The reference group is not shown in the figure and is households with 1-2 people. Individual round odds ratios are mutually adjusted for sex, key worker status, broad categories of ethnicity, household size and neighbourhood deprivation.

Table 5.1 The unweighted and weighted prevalence of swab-positivity by socio-demographic variables for rounds 5 to 10.

A. Rounds 5 to 6.

Variable	Category	Round 5				Round 6			
		Positive	Total	Unweighted Prevalence	Weighted Prevalence*	Positive	Total	Unweighted Prevalence	Weighted Prevalence*
Gender	Male	387	78,564	0.49% (0.44% , 0.54%)	0.65% (0.57% , 0.75%)	769	72,315	1.06% (0.99% , 1.14%)	1.29% (1.17% , 1.43%)
	Female	437	96,380	0.45% (0.41% , 0.50%)	0.54% (0.48% , 0.62%)	962	87,852	1.10% (1.03% , 1.17%)	1.31% (1.19% , 1.44%)
	NA	0	4	0.00% (0.00% , 60.24%)	NA (NA , NA)	0	6	0.00% (0.00% , 45.93%)	NA (NA , NA)
Age	05-12	59	13,771	0.43% (0.33% , 0.55%)	0.58% (0.47% , 0.73%)	113	11,245	1.00% (0.83% , 1.21%)	1.27% (0.96% , 1.68%)
	13-17	59	10,480	0.56% (0.43% , 0.73%)	0.60% (0.43% , 0.84%)	119	8,616	1.38% (1.15% , 1.65%)	1.72% (1.33% , 2.21%)
	18-24	83	6,852	1.21% (0.97% , 1.50%)	0.70% (0.49% , 1.00%)	120	6,308	1.90% (1.58% , 2.27%)	2.38% (1.79% , 3.16%)
	25-34	86	15,683	0.55% (0.44% , 0.68%)	1.59% (1.18% , 2.13%)	206	15,242	1.35% (1.17% , 1.55%)	1.56% (1.30% , 1.86%)
	35-44	121	22,954	0.53% (0.44% , 0.63%)	0.69% (0.53% , 0.90%)	214	21,003	1.02% (0.89% , 1.16%)	1.25% (1.05% , 1.48%)
	45-54	157	29,380	0.53% (0.45% , 0.62%)	0.58% (0.48% , 0.70%)	313	26,897	1.16% (1.04% , 1.30%)	1.26% (1.10% , 1.44%)
	55-64	116	31,884	0.36% (0.30% , 0.44%)	0.37% (0.30% , 0.46%)	326	29,752	1.10% (0.98% , 1.22%)	1.20% (1.05% , 1.38%)
	65+	143	43,944	0.33% (0.27% , 0.38%)	0.35% (0.28% , 0.43%)	320	41,110	0.78% (0.70% , 0.87%)	0.82% (0.71% , 0.94%)
Region	South East	104	39,496	0.26% (0.22% , 0.32%)	0.29% (0.23% , 0.37%)	244	36,164	0.67% (0.59% , 0.76%)	0.69% (0.60% , 0.80%)
	North East	61	6,702	0.91% (0.70% , 1.17%)	1.06% (0.80% , 1.42%)	89	5,846	1.52% (1.22% , 1.87%)	1.49% (1.17% , 1.89%)
	North West	199	19,468	1.02% (0.89% , 1.17%)	1.21% (1.01% , 1.46%)	372	17,866	2.08% (1.88% , 2.30%)	2.39% (2.10% , 2.73%)
	Yorkshire and The Humber	74	11,658	0.63% (0.50% , 0.80%)	0.84% (0.60% , 1.17%)	189	10,697	1.77% (1.53% , 2.03%)	2.32% (1.90% , 2.81%)
	East Midlands	101	21,897	0.46% (0.38% , 0.56%)	0.56% (0.43% , 0.72%)	233	20,458	1.14% (1.00% , 1.29%)	1.25% (1.07% , 1.45%)
	West Midlands	70	16,343	0.43% (0.33% , 0.54%)	0.60% (0.40% , 0.89%)	196	14,860	1.32% (1.14% , 1.52%)	1.59% (1.18% , 2.13%)
	East of England	81	25,625	0.32% (0.25% , 0.39%)	0.30% (0.23% , 0.38%)	150	23,231	0.65% (0.55% , 0.76%)	0.69% (0.57% , 0.83%)
	London	79	17,499	0.45% (0.36% , 0.56%)	0.54% (0.42% , 0.70%)	141	15,757	0.89% (0.75% , 1.05%)	0.97% (0.79% , 1.18%)
	South West	55	16,260	0.34% (0.25% , 0.44%)	0.33% (0.23% , 0.46%)	117	15,294	0.77% (0.63% , 0.92%)	0.80% (0.64% , 0.99%)
key worker status	Health care or care home worker	51	9,620	0.53% (0.40% , 0.70%)	0.56% (0.48% , 0.65%)	124	9,257	1.34% (1.12% , 1.60%)	1.78% (1.33% , 2.39%)
	Other essential/key worker	154	28,846	0.53% (0.45% , 0.62%)	0.66% (0.46% , 0.94%)	357	26,894	1.33% (1.19% , 1.47%)	1.64% (1.41% , 1.92%)
	Other workers	338	72,174	0.47% (0.42% , 0.52%)	0.70% (0.55% , 0.88%)	677	64,810	1.04% (0.97% , 1.13%)	1.22% (1.10% , 1.37%)
	Not full-time, part-time, or self-employed	256	59,762	0.43% (0.38% , 0.48%)	0.58% (0.49% , 0.69%)	519	55,444	0.94% (0.86% , 1.02%)	1.11% (0.98% , 1.25%)
	NA	25	4,546	0.55% (0.36% , 0.81%)	0.60% (0.37% , 0.95%)	54	3,768	1.43% 1.08% 1.87%	1.55% (1.04% , 2.31%)
Ethnic group	White	707	157,644	0.45% (0.42% , 0.48%)	0.54% (0.48% , 0.59%)	1549	145,334	1.07% (1.01% , 1.12%)	1.24% (1.16% , 1.33%)
	Asian	71	7,907	0.90% (0.70% , 1.13%)	1.11% (0.83% , 1.48%)	84	6,712	1.25% (1.00% , 1.55%)	2.05% (1.45% , 2.88%)
	Black	16	2,194	0.73% (0.42% , 1.18%)	1.17% (0.65% , 2.10%)	23	1,760	1.31% (0.83% , 1.95%)	1.23% (0.76% , 1.98%)
	Mixed	11	3,080	0.36% (0.18% , 0.64%)	0.44% (0.22% , 0.86%)	28	2,697	1.04% (0.69% , 1.50%)	1.01% (0.66% , 1.53%)
	Other	7	1,439	0.49% (0.20% , 1.00%)	0.47% (0.21% , 1.05%)	17	1,184	1.44% (0.84% , 2.29%)	1.81% (1.00% , 3.24%)
	NA	12	2,684	0.45% 0.23% 0.78%	0.68% (0.29% , 1.57%)	30	2,486	1.21% 0.82% 1.72%	1.38% (0.87% , 2.18%)
Household size	1	88	24,916	0.35% (0.28% , 0.44%)	0.44% (0.32% , 0.59%)	228	23,639	0.96% (0.84% , 1.10%)	1.08% (0.91% , 1.29%)
	2	227	61,717	0.37% (0.32% , 0.42%)	0.44% (0.37% , 0.53%)	553	59,221	0.93% (0.86% , 1.01%)	1.13% (0.99% , 1.28%)
	3	172	32,394	0.53% (0.45% , 0.62%)	0.61% (0.51% , 0.74%)	358	30,041	1.19% (1.07% , 1.32%)	1.47% (1.25% , 1.73%)
	4	211	37,231	0.57% (0.49% , 0.65%)	0.75% (0.61% , 0.92%)	349	31,911	1.09% (0.98% , 1.21%)	1.30% (1.11% , 1.52%)
	5	77	13,145	0.59% (0.46% , 0.73%)	0.76% (0.56% , 1.03%)	172	10,795	1.59% (1.37% , 1.85%)	1.73% (1.44% , 2.08%)
	6 and 6+	49	5,545	0.88% (0.65% , 1.17%)	1.05% (0.75% , 1.47%)	71	4,566	1.56% (1.22% , 1.96%)	2.16% (1.49% , 3.12%)
COVID case conta	No	556	139,853	0.40% (0.37% , 0.43%)	0.50% (0.44% , 0.56%)	1130	128,069	0.88% (0.83% , 0.94%)	1.02% (0.94% , 1.11%)
	Yes, contact with a confirmed/tested COVID-19 case	102	1,566	6.51% (5.34% , 7.85%)	8.20% (6.31% , 10.59%)	301	3,621	8.31% (7.43% , 9.26%)	8.92% (7.67% , 10.35%)
	Yes, contact with a suspected COVID-19 case	17	1,487	1.14% (0.67% , 1.82%)	1.09% (0.64% , 1.83%)	50	1,669	3.00% (2.23% , 3.93%)	3.55% (2.26% , 5.53%)
	NA	149	32,042	0.47% (0.39% , 0.55%)	0.54% (0.43% , 0.68%)	250	26,814	0.93% (0.82% , 1.05%)	1.20% (0.99% , 1.45%)
Symptom status	Classic COVID	187	8,376	2.23% (1.93% , 2.57%)	2.75% (2.25% , 3.36%)	422	6,283	6.72% (6.11% , 7.36%)	7.51% (6.63% , 8.49%)
	Other symptoms	146	25,832	0.57% (0.48% , 0.66%)	0.70% (0.56% , 0.87%)	266	20,739	1.28% (1.13% , 1.45%)	1.35% (1.17% , 1.57%)
	No symptoms	344	108,910	0.32% (0.28% , 0.35%)	0.39% (0.34% , 0.46%)	798	106,439	0.75% (0.70% , 0.80%)	0.90% (0.81% , 1.01%)
	NA	147	31,830	0.46% (0.39% , 0.54%)	0.54% (0.43% , 0.68%)	245	26,712	0.92% (0.81% , 1.04%)	1.19% (0.98% , 1.45%)
Deprivation	1 Most deprived	135	17,641	0.77% (0.64% , 0.91%)	1.00% (0.80% , 1.26%)	254	15,736	1.61% (1.42% , 1.82%)	2.09% (1.74% , 2.50%)
	2	129	27,483	0.47% (0.39% , 0.56%)	0.53% (0.43% , 0.65%)	289	24,957	1.16% (1.03% , 1.30%)	1.27% (1.11% , 1.46%)
	3	146	37,426	0.39% (0.33% , 0.46%)	0.48% (0.38% , 0.59%)	354	34,378	1.03% (0.93% , 1.14%)	1.09% (0.96% , 1.24%)
	4	190	43,109	0.44% (0.38% , 0.51%)	0.49% (0.41% , 0.58%)	407	39,792	1.02% (0.93% , 1.13%)	1.10% (0.98% , 1.25%)
	5 Least deprived	224	49,289	0.45% (0.40% , 0.52%)	0.52% (0.44% , 0.61%)	427	45,310	0.94% (0.86% , 1.04%)	1.02% (0.91% , 1.15%)

* We present weighted prevalence if the number of positives in a category is 10 or more.

B. Rounds 7 to 8.

Variable	Category	Round 7				Round 8			
		Positive	Total	Unweighted Prevalence	Weighted Prevalence*	Positive	Total	Unweighted Prevalence	Weighted Prevalence*
Gender	Male	599	75,656	0.79% (0.73% , 0.86%)	0.96% (0.86% , 1.07%)	1043	75,296	1.39% (1.30% , 1.47%)	1.63% (1.50% , 1.77%)
	Female	700	92,524	0.76% (0.70% , 0.81%)	0.92% (0.83% , 1.02%)	1239	92,340	1.34% (1.27% , 1.42%)	1.51% (1.41% , 1.63%)
	NA	0	1	0.00% (0.00% , 97.50%)	NA (NA , NA)	0	6	0.00% (0.00% , 31.08%)	NA (NA , NA)
Age	05-12	147	11,873	1.24% (1.05% , 1.45%)	1.44% (1.16% , 1.77%)	170	11,727	1.45% (1.25% , 1.68%)	1.59% (1.32% , 1.93%)
	13-17	172	9,861	1.74% (1.50% , 2.02%)	2.04% (1.69% , 2.46%)	172	8,947	1.92% (1.66% , 2.23%)	2.25% (1.85% , 2.73%)
	18-24	86	6,475	1.33% (1.06% , 1.64%)	1.41% (1.04% , 1.91%)	156	6,753	2.31% (1.98% , 2.70%)	2.44% (1.96% , 3.03%)
	25-34	130	14,758	0.88% (0.74% , 1.05%)	1.00% (0.80% , 1.23%)	263	15,031	1.75% (1.55% , 1.97%)	1.95% (1.68% , 2.26%)
	35-44	160	20,846	0.77% (0.65% , 0.90%)	0.89% (0.73% , 1.09%)	314	21,764	1.44% (1.29% , 1.61%)	1.62% (1.41% , 1.85%)
	45-54	220	27,338	0.80% (0.70% , 0.92%)	0.86% (0.73% , 1.01%)	412	27,982	1.47% (1.34% , 1.62%)	1.61% (1.43% , 1.81%)
	55-64	188	31,396	0.60% (0.52% , 0.69%)	0.64% (0.53% , 0.76%)	413	31,989	1.29% (1.17% , 1.42%)	1.29% (1.15% , 1.45%)
	65+	196	45,634	0.43% (0.37% , 0.49%)	0.50% (0.41% , 0.61%)	382	43,449	0.88% (0.80% , 0.97%)	0.93% (0.82% , 1.05%)
Region	South East	254	38,296	0.66% (0.58% , 0.75%)	0.73% (0.63% , 0.84%)	578	39,156	1.48% (1.36% , 1.60%)	1.61% (1.46% , 1.77%)
	North East	53	5,673	0.93% (0.70% , 1.22%)	0.94% (0.66% , 1.35%)	62	5,731	1.08% (0.84% , 1.38%)	1.22% (0.87% , 1.69%)
	North West	157	18,337	0.86% (0.73% , 1.00%)	1.02% (0.83% , 1.24%)	220	17,917	1.23% (1.08% , 1.40%)	1.38% (1.16% , 1.64%)
	Yorkshire and The Humber	107	11,288	0.95% (0.78% , 1.14%)	1.25% (0.99% , 1.58%)	93	10,665	0.87% (0.71% , 1.07%)	0.80% (0.62% , 1.03%)
	East Midlands	221	22,001	1.00% (0.88% , 1.15%)	1.19% (1.00% , 1.41%)	225	21,698	1.04% (0.91% , 1.18%)	1.16% (0.99% , 1.36%)
	West Midlands	156	15,757	0.99% (0.84% , 1.16%)	1.24% (0.95% , 1.61%)	198	15,088	1.31% (1.14% , 1.51%)	1.66% (1.32% , 2.10%)
	East of England	125	24,627	0.51% (0.42% , 0.60%)	0.58% (0.47% , 0.71%)	391	25,178	1.55% (1.41% , 1.71%)	1.78% (1.57% , 2.02%)
	London	136	15,566	0.87% (0.73% , 1.03%)	1.09% (0.89% , 1.32%)	390	15,641	2.49% (2.26% , 2.75%)	2.83% (2.53% , 3.16%)
	South West	90	16,636	0.54% (0.44% , 0.66%)	0.58% (0.45% , 0.76%)	125	16,568	0.75% (0.63% , 0.90%)	0.87% (0.71% , 1.08%)
key worker status	Health care or care home worker	122	9,460	1.29% (1.07% , 1.54%)	1.54% (1.23% , 1.94%)	161	8,259	1.95% (1.67% , 2.27%)	2.24% (1.84% , 2.72%)
	Other essential/key worker	267	29,910	0.89% (0.79% , 1.01%)	1.04% (0.88% , 1.22%)	531	30,239	1.76% (1.61% , 1.91%)	1.79% (1.61% , 2.00%)
	Other workers	457	64,423	0.71% (0.65% , 0.78%)	0.82% (0.72% , 0.93%)	867	64,772	1.34% (1.25% , 1.43%)	1.55% (1.42% , 1.70%)
	Not full-time, part-time, or self-employed	380	60,222	0.63% (0.57% , 0.70%)	0.85% (0.74% , 0.97%)	670	60,438	1.11% (1.03% , 1.20%)	1.36% (1.23% , 1.50%)
	NA	73	4,166	1.75% (1.38% , 2.20%)	1.80% (1.34% , 2.41%)	53	3,934	1.35% (1.03% , 1.76%)	1.58% (1.10% , 2.27%)
Ethnic group	White	1122	153,463	0.73% (0.69% , 0.78%)	0.84% (0.78% , 0.90%)	1938	152,111	1.27% (1.22% , 1.33%)	1.41% (1.33% , 1.49%)
	Asian	95	6,577	1.44% (1.17% , 1.76%)	2.07% (1.57% , 2.71%)	163	6,537	2.49% (2.14% , 2.90%)	2.80% (2.32% , 3.37%)
	Black	18	1,754	1.03% (0.61% , 1.62%)	1.07% (0.65% , 1.76%)	60	1,832	3.28% (2.55% , 4.19%)	3.07% (2.29% , 4.09%)
	Mixed	23	2,721	0.85% (0.54% , 1.27%)	0.84% (0.53% , 1.33%)	46	2,751	1.67% (1.26% , 2.22%)	1.78% (1.28% , 2.46%)
	Other	16	1,157	1.38% (0.79% , 2.24%)	1.97% (1.09% , 3.51%)	29	1,190	2.44% (1.70% , 3.48%)	2.42% (1.59% , 3.65%)
	NA	25	2,509	1.00% (0.65% , 1.47%)	1.15% (0.74% , 1.78%)	46	3,221	1.43% 1.07% 1.90%	1.76% (1.23% , 2.50%)
Household size	1	125	25,606	0.49% (0.41% , 0.58%)	0.56% (0.44% , 0.71%)	275	25,360	1.08% (0.96% , 1.22%)	1.24% (1.06% , 1.44%)
	2	351	62,446	0.56% (0.51% , 0.62%)	0.69% (0.60% , 0.79%)	622	61,379	1.01% (0.94% , 1.10%)	1.11% (1.01% , 1.23%)
	3	277	30,729	0.90% (0.80% , 1.01%)	0.96% (0.83% , 1.12%)	455	29,965	1.52% (1.39% , 1.66%)	1.67% (1.49% , 1.88%)
	4	315	33,498	0.94% (0.84% , 1.05%)	1.02% (0.89% , 1.17%)	568	34,847	1.63% (1.50% , 1.77%)	1.77% (1.59% , 1.96%)
	5	147	11,274	1.30% (1.10% , 1.53%)	1.59% (1.27% , 1.98%)	246	11,566	2.13% (1.88% , 2.41%)	2.48% (2.05% , 3.00%)
	6 and 6+	84	4,628	1.82% (1.45% , 2.24%)	2.40% (1.71% , 3.35%)	116	4,525	2.56% (2.14% , 3.07%)	3.01% (2.39% , 3.77%)
COVID case conta	No	705	133,663	0.53% (0.49% , 0.57%)	0.65% (0.59% , 0.71%)	1130	138,065	0.82% (0.77% , 0.87%)	0.94% (0.87% , 1.02%)
	Yes, contact with a confirmed/tested COVID-19 case	362	4,658	7.77% (7.02% , 8.58%)	7.77% (6.83% , 8.83%)	730	5,008	14.58% (13.63% , 15.58%)	15.61% (14.26% , 17.06%)
	Yes, contact with a suspected COVID-19 case	31	1,495	2.07% (1.41% , 2.93%)	1.96% (1.27% , 3.02%)	58	1,329	4.36% (3.39% , 5.60%)	4.94% (3.57% , 6.79%)
	NA	201	28,365	0.71% (0.61% , 0.81%)	0.92% (0.75% , 1.14%)	364	23,240	1.57% (1.41% , 1.73%)	1.64% (1.45% , 1.87%)
Symptom status	Classic COVID	404	5,280	7.65% (6.95% , 8.40%)	7.98% (7.07% , 9.01%)	822	6,475	12.70% (11.91% , 13.53%)	14.05% (12.89% , 15.30%)
	Other symptoms	220	18,185	1.21% (1.06% , 1.38%)	1.35% (1.14% , 1.60%)	350	19,609	1.78% (1.61% , 1.98%)	1.89% (1.66% , 2.15%)
	No symptoms	476	116,449	0.41% (0.37% , 0.45%)	0.51% (0.46% , 0.58%)	747	118,401	0.63% (0.59% , 0.68%)	0.76% (0.68% , 0.84%)
	NA	199	28,267	0.70% (0.61% , 0.81%)	0.92% (0.74% , 1.14%)	363	23,157	1.57% (1.42% , 1.74%)	1.65% (1.45% , 1.87%)
Deprivation	1 Most deprived	198	15,951	1.24% (1.08% , 1.43%)	1.41% (1.19% , 1.67%)	257	15,190	1.69% (1.50% , 1.91%)	1.79% (1.53% , 2.10%)
	2	238	26,259	0.91% (0.80% , 1.03%)	1.04% (0.86% , 1.25%)	449	26,012	1.73% (1.57% , 1.89%)	1.95% (1.74% , 2.19%)
	3	269	35,975	0.75% (0.66% , 0.84%)	0.80% (0.69% , 0.93%)	479	36,061	1.33% (1.22% , 1.45%)	1.52% (1.36% , 1.69%)
	4	293	42,350	0.69% (0.62% , 0.78%)	0.83% (0.71% , 0.95%)	540	42,002	1.29% (1.18% , 1.40%)	1.41% (1.27% , 1.56%)
	5 Least deprived	301	47,646	0.63% (0.56% , 0.71%)	0.68% (0.59% , 0.78%)	557	48,377	1.15% (1.06% , 1.25%)	1.22% (1.10% , 1.35%)

* We present weighted prevalence if the number of positives in a category is 10 or more.

C. Rounds 9 to 10.

Variable	Category	Round 9				Round 10			
		Positive	Total	Unweighted Prevalence	Weighted Prevalence*	Positive	Total	Unweighted Prevalence	Weighted Prevalence*
Gender	Male	315	75,027	0.42% (0.38% , 0.47%)	0.50% (0.43% , 0.59%)	108	63,077	0.17% (0.14% , 0.21%)	0.22% (0.17% , 0.28%)
	Female	374	90,425	0.41% (0.37% , 0.46%)	0.49% (0.43% , 0.56%)	119	77,765	0.15% (0.13% , 0.18%)	0.18% (0.14% , 0.22%)
	NA	0	4	0.00% (0.00% , 40.35%)	NA (NA , NA)	0	2	0 (0 , 84.19%)	NA (NA , NA)
Age	05-12	81	13,605	0.60% (0.48% , 0.74%)	0.72% (0.54% , 0.95%)	33	10,750	0.31% (0.21% , 0.43%)	0.41% (0.27% , 0.62%)
	13-17	45	10,155	0.44% (0.33% , 0.59%)	0.57% (0.38% , 0.87%)	14	7,891	0.18% (0.10% , 0.30%)	0.19% (0.11% , 0.34%)
	18-24	40	6,788	0.59% (0.43% , 0.80%)	0.71% (0.45% , 1.12%)	13	5,799	0.22% (0.12% , 0.38%)	0.29% (0.16% , 0.53%)
	25-34	68	14,801	0.46% (0.36% , 0.58%)	0.46% (0.34% , 0.61%)	21	12,203	0.17% (0.11% , 0.26%)	0.18% (0.11% , 0.31%)
	35-44	94	21,088	0.45% (0.36% , 0.55%)	0.57% (0.42% , 0.77%)	35	17,339	0.20% (0.14% , 0.28%)	0.21% (0.14% , 0.31%)
	45-54	132	27,595	0.48% (0.40% , 0.57%)	0.50% (0.41% , 0.62%)	37	22,258	0.17% (0.12% , 0.23%)	0.20% (0.13% , 0.30%)
	55-64	135	31,727	0.43% (0.36% , 0.50%)	0.49% (0.40% , 0.60%)	39	25,811	0.15% (0.11% , 0.21%)	0.17% (0.12% , 0.25%)
	65+	94	39,697	0.24% (0.19% , 0.29%)	0.26% (0.21% , 0.33%)	35	38,793	0.09% (0.06% , 0.13%)	0.09% (0.06% , 0.13%)
Region	South East	119	36,642	0.32% (0.27% , 0.39%)	0.36% (0.29% , 0.44%)	18	31,128	0.06% (0.03% , 0.09%)	0.07% (0.04% , 0.12%)
	North East	41	5,836	0.70% (0.52% , 0.95%)	0.69% (0.49% , 0.98%)	16	4,935	0.32% (0.19% , 0.53%)	0.41% (0.24% , 0.70%)
	North West	112	19,402	0.58% (0.48% , 0.69%)	0.69% (0.54% , 0.88%)	41	16,106	0.25% (0.18% , 0.35%)	0.31% (0.21% , 0.45%)
	Yorkshire and The Humber	42	11,218	0.37% (0.28% , 0.51%)	0.60% (0.37% , 0.96%)	28	9,601	0.29% (0.19% , 0.42%)	0.36% (0.22% , 0.57%)
	East Midlands	103	21,193	0.49% (0.40% , 0.59%)	0.59% (0.45% , 0.77%)	34	17,934	0.19% (0.13% , 0.26%)	0.19% (0.13% , 0.28%)
	West Midlands	60	15,478	0.39% (0.30% , 0.50%)	0.36% (0.25% , 0.52%)	26	12,926	0.20% (0.13% , 0.29%)	0.20% (0.13% , 0.31%)
	East of England	86	23,672	0.36% (0.29% , 0.45%)	0.47% (0.36% , 0.60%)	27	20,103	0.13% (0.09% , 0.20%)	0.15% (0.10% , 0.24%)
	London	91	15,492	0.59% (0.48% , 0.72%)	0.60% (0.48% , 0.76%)	22	14,062	0.16% (0.10% , 0.24%)	0.16% (0.10% , 0.26%)
	South West	35	16,523	0.21% (0.15% , 0.29%)	0.23% (0.15% , 0.36%)	15	14,049	0.11% (0.06% , 0.18%)	0.13% (0.07% , 0.26%)
key worker status	Health care or care home worker	55	9,042	0.61% (0.47% , 0.79%)	0.73% (0.49% , 1.08%)	19	8,614	0.22% (0.13% , 0.34%)	0.21% (0.12% , 0.36%)
	Other essential/key worker	165	31,393	0.53% (0.45% , 0.61%)	0.57% (0.46% , 0.70%)	44	24,094	0.18% (0.13% , 0.25%)	0.20% (0.14% , 0.30%)
	Other workers	244	59,907	0.41% (0.36% , 0.46%)	0.47% (0.40% , 0.56%)	90	52,495	0.17% (0.14% , 0.21%)	0.23% (0.18% , 0.30%)
	Not full-time, part-time, or self-employed	202	60,496	0.33% (0.29% , 0.38%)	0.43% (0.36% , 0.52%)	64	52,331	0.12% (0.09% , 0.16%)	0.14% (0.11% , 0.19%)
	NA	23	4,618	0.50% (0.33% , 0.75%)	0.55% (0.33% , 0.92%)	10	3,310	0.30% (0.15% , 0.55%)	0.33% (0.17% , 0.64%)
Ethnic group	White	596	149,859	0.40% (0.37% , 0.43%)	0.45% (0.41% , 0.50%)	194	126,966	0.15% (0.13% , 0.18%)	0.19% (0.16% , 0.23%)
	Asian	52	6,516	0.80% (0.61% , 1.04%)	0.95% (0.64% , 1.41%)	14	5,892	0.24% (0.13% , 0.40%)	0.27% (0.15% , 0.47%)
	Black	14	2,151	0.65% (0.39% , 1.09%)	0.79% (0.39% , 1.60%)	4	1,965	0.20% (0.06% , 0.52%)	NA (NA , NA)
	Mixed	8	2,926	0.27% (0.14% , 0.54%)	NA (NA , NA)	7	2,460	0.28% (0.11% , 0.59%)	NA (NA , NA)
	Other	5	1,185	0.42% (0.18% , 0.98%)	NA (NA , NA)	3	1,143	0.26% (0.05% , 0.77%)	NA (NA , NA)
	NA	14	2,819	0.50% (0.30% , 0.83%)	0.54% (0.30% , 0.94%)	5	2,418	0.21% (0.07% , 0.48%)	NA (NA , NA)
Household size	1	68	23,576	0.29% (0.23% , 0.37%)	0.31% (0.24% , 0.42%)	26	20,336	0.13% (0.08% , 0.19%)	0.15% (0.09% , 0.24%)
	2	177	59,046	0.30% (0.26% , 0.35%)	0.32% (0.27% , 0.38%)	64	53,128	0.12% (0.09% , 0.15%)	0.16% (0.11% , 0.22%)
	3	145	30,652	0.47% (0.40% , 0.56%)	0.52% (0.42% , 0.64%)	48	25,373	0.19% (0.14% , 0.25%)	0.22% (0.15% , 0.33%)
	4	182	35,280	0.52% (0.45% , 0.60%)	0.56% (0.46% , 0.67%)	58	28,568	0.20% (0.15% , 0.26%)	0.24% (0.18% , 0.33%)
	5	83	12,055	0.69% (0.56% , 0.85%)	0.84% (0.62% , 1.12%)	19	9,582	0.20% (0.12% , 0.31%)	0.21% (0.11% , 0.38%)
	6 and 6+	34	4,847	0.70% (0.50% , 0.98%)	1.41% (0.88% , 2.26%)	12	3,857	0.31% (0.16% , 0.54%)	0.35% (0.19% , 0.65%)
COVID case conta	No	350	139,231	0.25% (0.23% , 0.28%)	0.29% (0.25% , 0.33%)	136	119,963	0.11% (0.10% , 0.13%)	0.14% (0.11% , 0.17%)
	Yes, contact with a confirmed/tested COVID-19 case	215	2,587	8.31% (7.31% , 9.44%)	9.41% (7.85% , 11.23%)	46	977	4.71% (3.47% , 6.23%)	4.86% (3.37% , 6.96%)
	Yes, contact with a suspected COVID-19 case	15	640	2.34% (1.43% , 3.83%)	3.07% (1.02% , 8.90%)	5	397	1.26% (0.41% , 2.91%)	NA (NA , NA)
	NA	109	22,998	0.47% (0.39% , 0.57%)	0.52% (0.41% , 0.66%)	40	19,507	0.21% (0.15% , 0.28%)	0.27% (0.18% , 0.40%)
Symptom status	Classic COVID	223	4,968	4.49% (3.95% , 5.10%)	4.89% (4.02% , 5.92%)	45	3,508	1.28% (0.94% , 1.71%)	1.56% (1.09% , 2.24%)
	Other symptoms	105	17,099	0.61% (0.51% , 0.74%)	0.74% (0.56% , 0.96%)	30	14,801	0.20% (0.14% , 0.29%)	0.28% (0.18% , 0.42%)
	No symptoms	256	120,482	0.21% (0.19% , 0.24%)	0.26% (0.22% , 0.31%)	112	103,135	0.11% (0.09% , 0.13%)	0.12% (0.10% , 0.15%)
	NA	105	22,907	0.46% (0.38% , 0.55%)	0.50% (0.40% , 0.64%)	40	19,400	0.21% (0.15% , 0.28%)	0.27% (0.18% , 0.40%)
Deprivation	1 Most deprived	108	15,718	0.69% (0.57% , 0.83%)	0.85% (0.66% , 1.09%)	36	12,974	0.28% (0.19% , 0.38%)	0.33% (0.23% , 0.47%)
	2	142	25,369	0.56% (0.48% , 0.66%)	0.59% (0.49% , 0.72%)	46	21,750	0.21% (0.15% , 0.28%)	0.25% (0.18% , 0.35%)
	3	122	35,741	0.34% (0.29% , 0.41%)	0.34% (0.28% , 0.42%)	44	30,461	0.14% (0.11% , 0.19%)	0.17% (0.12% , 0.24%)
	4	157	41,455	0.38% (0.32% , 0.44%)	0.38% (0.32% , 0.46%)	48	35,151	0.14% (0.10% , 0.18%)	0.13% (0.09% , 0.17%)
	5 Least deprived	160	47,173	0.34% (0.29% , 0.40%)	0.36% (0.30% , 0.43%)	53	40,508	0.13% (0.10% , 0.17%)	0.13% (0.09% , 0.17%)

* We present weighted prevalence if the number of positives in a category is 10 or more.

The number of vaccinated female participants was higher than that of male participants in our study. Although we accounted for key worker status in our regression analysis, the number of female key workers (healthcare workers/care home workers) in our study was greater than the number of male key workers across rounds 5 to 10, so more female participants in this high-risk group were vaccinated (Table 5.2), which may contribute to our results that females had lower odds of testing positive compared with males.

Social behaviour may also contribute to the sex difference in infection risk (Keogh-Brown et al., 2020). Men showed a higher prevalence of smoking and drinking, which may have led to comorbidities associated with the risk of COVID-19 infection. In addition, other sex-related behaviours including lower rates of use of preventative health measures and seeking healthcare may put men at greater risk of infection (Gebhard et al., 2020; Sex, Gender and Covid-19 Project, n.d.).

Table 5.2 Odds ratios for rounds 5 to 10, and an average weighted inversely to variance for core variables.

A. Rounds 5 to 8.

Variable	Category	Round 5 Mutually adjusted	Round 6 Mutually adjusted	Round 7 Mutually adjusted	Round 8 Mutually adjusted
Sex	Male	Ref	Ref	Ref	Ref
	Female	0.90 [0.78,1.03]	1.01 [0.92,1.12]	0.94 [0.84,1.05]	0.94 [0.86,1.02]
Age Group	5-12	0.75 [0.54,1.02]	0.95 [0.75,1.19]	1.50 [1.20,1.89]	0.92 [0.76,1.11]
	13-17	0.97 [0.69,1.37]	1.38 [1.07,1.77]	2.30 [1.79,2.95]	1.55 [1.26,1.91]
	18-24	2.17 [1.63,2.88]	1.87 [1.49,2.35]	1.74 [1.34,2.27]	1.66 [1.36,2.02]
	25-34	1.10 [0.83,1.45]	1.37 [1.12,1.66]	1.17 [0.93,1.48]	1.26 [1.07,1.49]
	35-44	Ref	Ref	Ref	Ref
	45-54	1.07 [0.84,1.36]	1.16 [0.97,1.38]	1.10 [0.90,1.36]	1.11 [0.96,1.29]
	55-64	0.78 [0.60,1.02]	1.17 [0.97,1.40]	0.93 [0.74,1.16]	1.16 [0.99,1.35]
	65+	0.75 [0.56,1.00]	0.92 [0.75,1.13]	0.79 [0.62,1.01]	0.98 [0.82,1.17]
Region	North East	3.48 [2.52,4.81]	2.19 [1.71,2.80]	1.30 [0.96,1.76]	0.69 [0.53,0.90]
	North West	3.94 [3.09,5.02]	3.03 [2.56,3.57]	1.21 [0.99,1.49]	0.79 [0.67,0.92]
	Yorkshire and The Humber	2.52 [1.86,3.40]	2.62 [2.16,3.18]	1.40 [1.11,1.76]	0.58 [0.46,0.72]
	East Midlands	1.79 [1.36,2.36]	1.68 [1.40,2.02]	1.48 [1.24,1.78]	0.69 [0.59,0.80]
	West Midlands	1.62 [1.20,2.20]	1.92 [1.59,2.32]	1.43 [1.17,1.75]	0.86 [0.73,1.01]
	East of England	1.22 [0.91,1.63]	0.94 [0.77,1.16]	0.75 [0.60,0.93]	1.03 [0.90,1.17]
	London	1.47 [1.09,1.99]	1.20 [0.96,1.48]	1.09 [0.87,1.35]	1.45 [1.26,1.66]
	South East	Ref	Ref	Ref	Ref
	South West	1.40 [1.00,1.94]	1.14 [0.92,1.43]	0.82 [0.64,1.05]	0.50 [0.42,0.61]
Key Worker Status	HCW/CHW**	1.03 [0.76,1.38]	1.16 [0.95,1.41]	1.66 [1.35,2.03]	1.48 [1.25,1.76]
	Other essential/key workers	1.03 [0.85,1.25]	1.16 [1.01,1.32]	1.17 [1.00,1.36]	1.35 [1.21,1.51]
	Other worker	Ref	Ref	Ref	Ref
	Not FT, PT, SE**	1.03 [0.86,1.24]	0.93 [0.82,1.06]	0.91 [0.78,1.06]	0.93 [0.83,1.04]
	Unknown	0.97 [0.63,1.50]	1.10 [0.81,1.48]	1.38 [1.04,1.82]	0.76 [0.56,1.02]
Ethnicity	White	Ref	Ref	Ref	Ref
	Asian	1.89 [1.45,2.45]	1.19 [0.95,1.50]	1.53 [1.23,1.91]	1.36 [1.15,1.61]
	Black	1.58 [0.95,2.63]	1.24 [0.81,1.89]	1.02 [0.63,1.65]	1.61 [1.23,2.10]
	Mixed	0.79 [0.43,1.44]	0.98 [0.67,1.43]	0.84 [0.55,1.28]	0.99 [0.74,1.34]
	Other	1.08 [0.51,2.30]	1.45 [0.89,2.36]	1.66 [1.00,2.74]	1.35 [0.93,1.97]
	Unknown	1.11 [0.63,1.97]	1.24 [0.86,1.79]	1.44 [0.96,2.15]	1.10 [0.82,1.48]
Household Size	1-2 People	Ref	Ref	Ref	Ref
	3-5 People	1.30 [1.10,1.55]	1.18 [1.05,1.32]	1.26 [1.09,1.46]	1.40 [1.26,1.55]
	6+ People	1.84 [1.32,2.56]	1.47 [1.13,1.90]	1.91 [1.47,2.47]	1.95 [1.58,2.41]
Deprivation Index Quintile	1 - Most Deprived	1.11 [0.89,1.39]	1.25 [1.07,1.47]	1.69 [1.40,2.04]	1.44 [1.24,1.68]
	2	0.86 [0.69,1.08]	1.08 [0.93,1.26]	1.35 [1.13,1.60]	1.41 [1.24,1.61]
	3	0.81 [0.65,0.99]	1.04 [0.90,1.20]	1.17 [0.99,1.39]	1.15 [1.02,1.31]
	4	0.91 [0.75,1.11]	1.02 [0.89,1.17]	1.09 [0.93,1.28]	1.14 [1.01,1.29]
	5 - Least Deprived	Ref	Ref	Ref	Ref

* HCW/CHW: health care worker/care home worker; Not FT, PT SE: not self-employed, full-time, part-time, self-employed.

B. Rounds 9 to 10 and across rounds 5 to 10.

Variable	Category	Round 9 Mutually adjusted	Round 10 Mutually adjusted	Rounds 5 to 10 Mutually adjusted
Sex	Male	Ref	Ref	Ref
	Female	0.97 [0.83,1.13]	0.88 [0.68,1.15]	0.95 [0.91, 1.00]
Age Group	5-12	1.24 [0.91,1.68]	1.38 [0.85,2.25]	1.07 [0.87, 1.33]
	13-17	1.01 [0.68,1.49]	0.74 [0.37,1.48]	1.31 [0.96, 1.77]
	18-24	1.35 [0.93,1.96]	1.06 [0.56,2.02]	1.74 [1.55, 1.94]
	25-34	1.08 [0.79,1.48]	0.86 [0.50,1.48]	1.22 [1.10, 1.34]
	35-44	Ref	Ref	Ref
	45-54	1.16 [0.89,1.52]	0.85 [0.54,1.36]	1.11 [1.02, 1.21]
	55-64	1.24 [0.94,1.63]	0.84 [0.52,1.35]	1.03 [0.89, 1.2]
	65+	0.85 [0.61,1.19]	0.55 [0.31,0.96]	0.86 [0.77, 0.97]
Region	North East	1.92 [1.34,2.76]	5.32 [2.68,10.53]	1.97 [1.13, 3.44]
	North West	1.62 [1.24,2.11]	4.20 [2.39,7.37]	2.03 [1.17, 3.5]
	Yorkshire and The Humber	1.11 [0.78,1.58]	5.03 [2.77,9.14]	1.72 [0.95, 3.13]
	East Midlands	1.46 [1.12,1.90]	3.29 [1.85,5.84]	1.51 [1.03, 2.21]
	West Midlands	1.12 [0.82,1.53]	3.41 [1.86,6.24]	1.49 [1.06, 2.1]
	East of England	1.10 [0.83,1.45]	2.31 [1.27,4.19]	1.05 [0.86, 1.3]
	London	1.54 [1.16,2.05]	2.18 [1.15,4.13]	1.35 [1.18, 1.54]
	South East	Ref	Ref	Ref
	South West	0.65 [0.44,0.95]	1.90 [0.95,3.77]	0.93 [0.64, 1.34]
Key Worker Status	HCW/CHW**	1.36 [1.01,1.83]	1.09 [0.66,1.81]	1.32 [1.13, 1.54]
	Other essential/key workers	1.18 [0.97,1.44]	0.89 [0.62,1.29]	1.17 [1.07, 1.29]
	Other worker	Ref	Ref	Ref
	Not FT, PT, SE**	0.96 [0.78,1.18]	0.89 [0.62,1.27]	0.94 [0.88, 1.00]
	Unknown	1.03 [0.66,1.62]	1.61 [0.79,3.27]	1.07 [0.86, 1.32]
Ethnicity	White	Ref	Ref	Ref
	Asian	1.53 [1.14,2.07]	1.29 [0.73,2.27]	1.45 [1.27, 1.66]
	Black	1.10 [0.64,1.90]	1.03 [0.37,2.83]	1.34 [1.11, 1.63]
	Mixed	0.56 [0.28,1.13]	1.57 [0.73,3.38]	0.93 [0.77, 1.11]
	Other	0.83 [0.34,2.01]	1.53 [0.48,4.87]	1.36 [1.08, 1.72]
	Unknown	1.26 [0.74,2.15]	1.43 [0.59,3.48]	1.22 [1.03, 1.45]
Household Size	1-2 People	Ref	Ref	Ref
	3-5 People	1.59 [1.31,1.92]	1.22 [0.88,1.70]	1.32 [1.21, 1.44]
	6+ People	1.93 [1.31,2.84]	1.68 [0.87,3.24]	1.80 [1.60, 2.03]
Deprivation Index Quintile	1 - Most Deprived	1.72 [1.33,2.21]	1.44 [0.93,2.24]	1.42 [1.23, 1.63]
	2	1.52 [1.21,1.92]	1.37 [0.91,2.04]	1.24 [1.04, 1.47]
	3	1.00 [0.79,1.27]	1.01 [0.68,1.51]	1.04 [0.93, 1.17]
	4	1.11 [0.89,1.38]	0.97 [0.65,1.43]	1.06 [0.99, 1.14]
	5 - Least Deprived	Ref	Ref	Ref

* HCW/CHW: health care worker/care home worker; Not FT, PT SE: not self-employed, full-time, part-time, self-employed.

5.3.2. Ethnicity

Self-identified broad classes of ethnicity were a substantial risk factor for testing positive during this period of the epidemic in England and remained important even after adjustment for household size and level of local neighbourhood deprivation. Broad classes of ethnicity were defined by participants self-identifying as one of the white, Asian, black, mixed and other ethnic groups. The Asian ethnic group showed the highest weighted prevalence consistently during the second wave, while the white ethnic group was at the lowest level (Figure 5.2A, Table 5.1). Similar patterns of higher infection prevalence in black and Asian ethnic groups were seen in England in subsequent REACT-1 rounds.(Riley et al., 2021b, 2021c; Elliott et al., 2021; Chadeau-Hyam et al., 2022c, 2022a; Elliott et al., 2022a, 2022b; Chadeau-Hyam et al., 2022b; Elliott et al., 2022c) Across rounds 5 to 10, the odds of being swab-positive were 1.45 (1.27, 1.66) and 1.34 (1.11, 1.63) times greater for people of Asian and black ethnicity respectively, compared to white ethnicity, whereas the odds were 0.93 (0.77, 1.11) times lower for people of mixed ethnic group than white ethnic group (Figure 5.2A, Table 5.2).

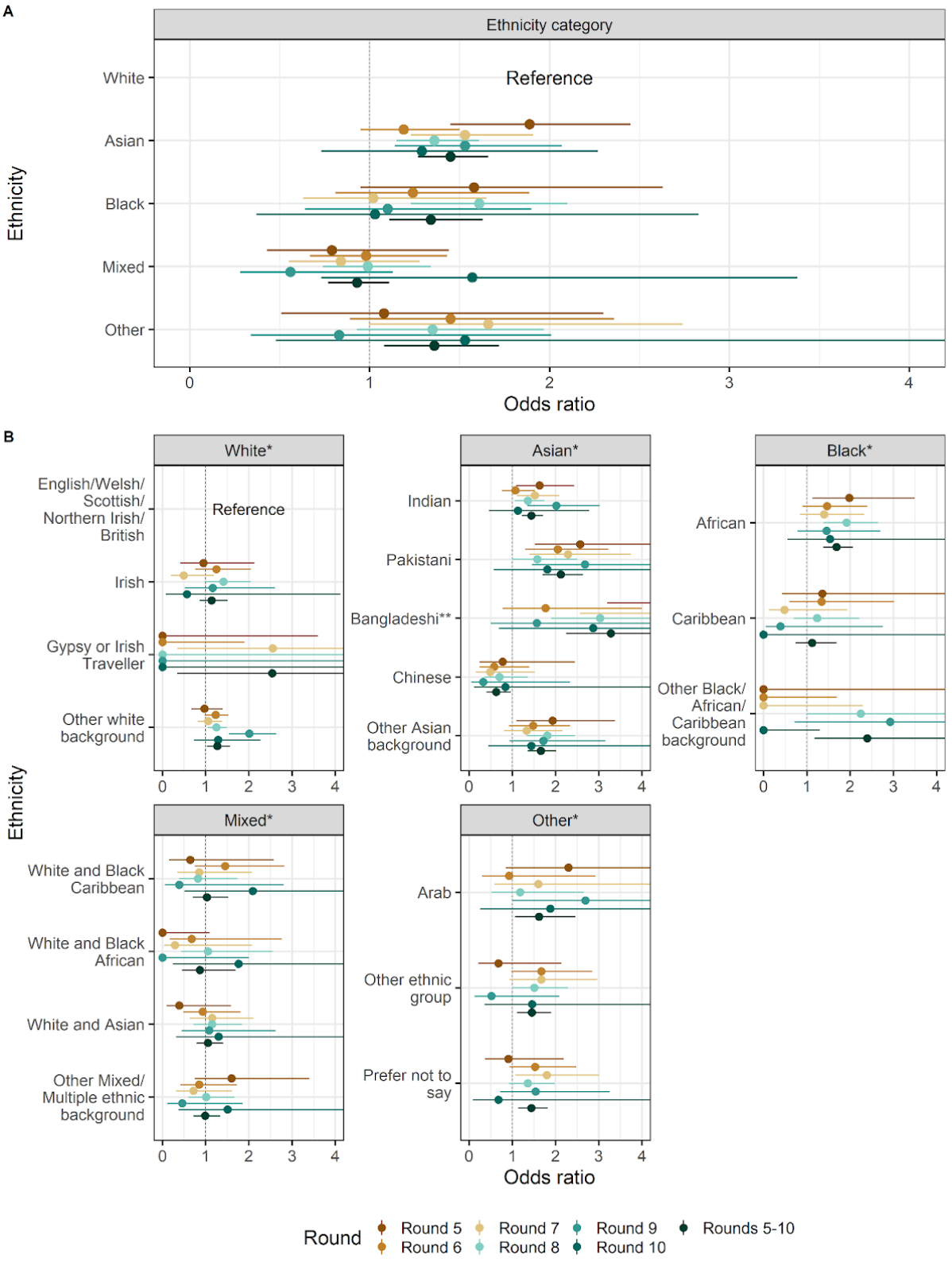


Figure 5.2 Forest plot of ethnicity categories and detailed subgroups of ethnicity.

A. Forest plot of ethnicity categories for rounds 5 to 10 and an average across rounds weighted inverse to variance (rounds 5 to 10), where the reference group is white ethnicity. **B**. Forest plot of detailed subgroups of ethnicity for rounds 5 to 10 and an average across rounds weighted inverse to variance (rounds 5 to 10), the reference group is those identifying as English, Welsh, Scottish, Northern Irish or British. Individual round odds ratios are mutually adjusted for sex, age group, region, key worker status, and deprivation index.
* For detailed subgroups of ethnicity, they are included in one model where the reference group is those identifying as English, Welsh, Scottish, Northern Irish or British.
** The odds ratio for Bangladeshi ethnicity in round 5 was 5.75 (3.29,10.07)

Participants were also asked to self-identify as a member of an ethnic subgroup. Some self-identified ethnic subgroups had a higher prevalence of SARS-CoV-2 infection in some rounds and lower in others (Table 5.3). These temporal differences were substantial and likely reflected underlying epidemic dynamics,(Padellini et al., 2022) rather than chance variation. For example, from rounds 5 to 8, the prevalence of swab-positivity in Bangladeshi participants was very high, reaching 6.12% (3.78%, 9.29%) in round 8. By comparison, the prevalence among white ethnic subgroups was less than 2%, with the highest value found in white Irish participants at 1.93% (1.30%, 2.74%). During the same period, black ethnic subgroups also showed a higher prevalence than white ethnic subgroups.

Excess risk of infection was also seen in these ethnic subgroups compared with the British white populations, even after accounting for the differences in age, sex, region, employment status, household size and level of local neighbourhood deprivation. Among subgroups, across rounds 5 to 10, the highest and the second-highest odds were found in Bangladeshi and Pakistani ethnic subgroups at 3.28 (2.24, 4.8) and 2.12 (1.7, 2.64), respectively, compared with British white participants, which contributed to the high odds in Asians (Figure 5.2B, Table 5.4).

Table 5.3 The unweighted and weighted prevalence of swab-positivity for detailed ethnicity subgroups for rounds 5 to 10.

A. Rounds 5 to 7.

Variable	Category	Round 5				Round 6				Round 7			
		Positive	Total	Unweighted Prevalence	Weighted Prevalence*	Positive	Total	Unweighted Prevalence	Weighted Prevalence*	Positive	Total	Unweighted Prevalence	Weighted Prevalence*
Ethnic group	White	702	157,370	0.45% (0.41% , 0.48%)	0.54% (0.48% , 0.59%)	1550	145,336	1.07% (1.02% , 1.12%)	1.24% (1.16% , 1.33%)	1122	153,463	0.73% (0.69% , 0.78%)	0.84% (0.78% , 0.90%)
	Asian	72	7,891	0.91% (0.73% , 1.15%)	1.19% (0.89% , 1.60%)	84	6,712	1.25% (1.01% , 1.55%)	2.05% (1.45% , 2.88%)	95	6,577	1.44% (1.18% , 1.76%)	2.07% (1.57% , 2.71%)
	Black	16	2,187	0.73% (0.45% , 1.19%)	1.17% (0.65% , 2.11%)	23	1,760	1.31% (0.87% , 1.95%)	1.23% (0.76% , 1.98%)	18	1,754	1.03% (0.65% , 1.62%)	1.07% (0.65% , 1.76%)
	Mixed	11	3,071	0.36% (0.20% , 0.64%)	0.44% (0.23% , 0.87%)	28	2,697	1.04% (0.72% , 1.50%)	1.01% (0.66% , 1.53%)	23	2,721	0.85% (0.56% , 1.27%)	0.84% (0.53% , 1.33%)
	Other	6	1,435	0.42% (0.19% , 0.91%)	0.41% (0.17% , 0.98%)	17	1,184	1.44% (0.90% , 2.29%)	1.81% (1.00% , 3.24%)	16	1,157	1.38% (0.85% , 2.23%)	1.97% (1.09% , 3.51%)
	NA	11	2,676	0.41% (0.23% , 0.73%)	0.62% (0.25% , 1.53%)	30	2,486	1.21% (0.85% , 1.72%)	1.38% (0.87% , 2.18%)	25	2,509	1.00% (0.68% , 1.47%)	1.15% (0.74% , 1.78%)
White ethnic group	English/Welsh/Scottish/ Northern Irish/British	663	147,483	0.45% (0.42% , 0.49%)	0.53% (0.48% , 0.59%)	1445	136,000	1.06% (1.01% , 1.12%)	1.24% (1.15% , 1.33%)	1052	144,148	0.73% (0.69% , 0.78%)	0.83% (0.77% , 0.90%)
	Irish	7	1,672	0.42% (0.17% , 0.86%)	NA (NA , NA)	16	1,458	1.10% (0.63% , 1.78%)	1.23% (0.71% , 2.13%)	5	1,600	0.31% (0.10% , 0.73%)	NA (NA , NA)
	Gypsy or Irish Traveller	0	48	0.00% (0.00% , 7.40%)	NA (NA , NA)	0	45	0.00% (0.00% , 7.87%)	NA (NA , NA)	1	49	2.04% (0.05% , 10.85%)	NA (NA , NA)
	Other white background	32	8,167	0.39% (0.27% , 0.55%)	0.59% (0.38% , 0.94%)	88	7,831	1.12% (0.90% , 1.38%)	1.37% (1.07% , 1.75%)	64	7,666	0.83% (0.64% , 1.06%)	0.96% (0.72% , 1.28%)
Asian ethnic group	Indian	26	3,849	0.68% (0.44% , 0.99%)	0.75% (0.49% , 1.13%)	34	3,304	1.03% (0.71% , 1.44%)	1.34% (0.88% , 2.03%)	42	3,250	1.29% (0.93% , 1.74%)	1.35% (0.95% , 1.91%)
	Pakistani	16	1,013	1.58% (0.91% , 2.55%)	1.94% (1.07% , 3.51%)	20	792	2.53% (1.55% , 3.87%)	3.42% (1.81% , 6.34%)	17	706	2.41% (1.41% , 3.83%)	2.52% (1.47% , 4.28%)
	Bangladeshi	14	499	2.81% (1.54% , 4.66%)	2.26% (1.24% , 4.08%)	6	330	1.82% (0.67% , 3.92%)	NA (NA , NA)	16	355	4.51% (2.60% , 7.22%)	4.56% (2.67% , 7.69%)
	Chinese	3	867	0.35% (0.07% , 1.01%)	NA (NA , NA)	5	884	0.57% (0.18% , 1.32%)	NA (NA , NA)	3	807	0.37% (0.08% , 1.08%)	NA (NA , NA)
	Other Asian background	13	1,663	0.78% (0.42% , 1.33%)	0.89% (0.49% , 1.61%)	19	1,402	1.36% (0.82% , 2.11%)	1.36% (0.76% , 2.44%)	17	1,459	1.17% (0.68% , 1.86%)	2.34% (1.06% , 5.07%)
Black ethnic group	African	13	1,417	0.92% (0.49% , 1.56%)	1.54% (0.80% , 2.94%)	17	1,145	1.48% (0.87% , 2.37%)	1.13% (0.65% , 1.96%)	16	1,082	1.48% (0.85% , 2.39%)	1.45% (0.85% , 2.46%)
	Caribbean	3	585	0.51% (0.11% , 1.49%)	NA (NA , NA)	6	465	1.29% (0.48% , 2.79%)	NA (NA , NA)	2	509	0.39% (0.05% , 1.41%)	NA (NA , NA)
	Other Black/African/Caribbean background	0	185	0.00% (0.00% , 1.97%)	NA (NA , NA)	0	150	0.00% (0.00% , 2.43%)	NA (NA , NA)	0	163	0.00% (0.00% , 2.24%)	NA (NA , NA)
Mixed ethnic group	White and Black Caribbean	2	639	0.31% (0.04% , 1.13%)	NA (NA , NA)	9	532	1.69% (0.78% , 3.19%)	NA (NA , NA)	5	541	0.92% (0.30% , 2.14%)	NA (NA , NA)
	White and Black African	0	316	0.00% (0.00% , 1.16%)	NA (NA , NA)	2	278	0.72% (0.09% , 2.57%)	NA (NA , NA)	1	303	0.33% (0.01% , 1.83%)	NA (NA , NA)
	White and Asian	3	1,095	0.27% (0.06% , 0.80%)	NA (NA , NA)	9	951	0.95% (0.43% , 1.79%)	NA (NA , NA)	11	973	1.13% (0.57% , 2.01%)	1.36% (0.71% , 2.59%)
	Other Mixed/Multiple ethnic background	6	1,021	0.59% (0.22% , 1.27%)	NA (NA , NA)	8	936	0.85% (0.37% , 1.68%)	NA (NA , NA)	6	904	0.66% (0.24% , 1.44%)	NA (NA , NA)
Other	Arab	4	349	1.15% (0.31% , 2.91%)	NA (NA , NA)	3	299	1.00% (0.21% , 2.90%)	NA (NA , NA)	4	278	1.44% (0.39% , 3.64%)	NA (NA , NA)
	Other ethnic group	2	1,086	0.18% (0.02% , 0.66%)	NA (NA , NA)	14	937	1.49% (0.82% , 2.49%)	1.84% (0.94% , 3.58%)	12	940	1.28% (0.66% , 2.22%)	1.87% (0.94% , 3.66%)
	Prefer not to say	4	1,314	0.30% (0.08% , 0.78%)	NA (NA , NA)	17	1,164	1.46% (0.85% , 2.33%)	1.66% (0.91% , 3.01%)	15	1,075	1.40% (0.78% , 2.29%)	1.36% (0.75% , 2.43%)
	NA	7	1,362	0.51% (0.21% , 1.06%)	NA (NA , NA)	13	1,270	1.02% (0.55% , 1.74%)	1.15% (0.55% , 2.36%)	10	1,373	0.73% (0.35% , 1.34%)	1.02% (0.52% , 1.99%)

* We present weighted prevalence if the number of positives in a category is 10 or more.

B. Rounds 8 to 10.

Variable	Category	Round 8				Round 9				Round 10			
		Positive	Total	Unweighted Prevalence	Weighted Prevalence*	Positive	Total	Unweighted Prevalence	Weighted Prevalence*	Positive	Total	Unweighted Prevalence	Weighted Prevalence*
Ethnic group	White	1938	152,111	1.27% (1.22%, 1.33%)	1.41% (1.33%, 1.49%)	596	149,859	0.40% (0.37%, 0.43%)	0.45% (0.40%, 0.50%)	194	126,966	0.15% (0.13%, 0.18%)	0.19% (0.16%, 0.23%)
	Asian	163	6,537	2.49% (2.14%, 2.90%)	2.80% (2.32%, 3.37%)	52	6,516	0.80% (0.61%, 1.04%)	0.91% (0.61%, 1.36%)	14	5,892	0.24% (0.14%, 0.40%)	0.27% (0.15%, 0.47%)
	Black	60	1,832	3.28% (2.55%, 4.19%)	3.07% (2.29%, 4.09%)	14	2,151	0.65% (0.39%, 1.09%)	0.83% (0.41%, 1.68%)	4	1,965	0.20% (0.08%, 0.52%)	NA (NA, NA)
	Mixed	46	2,751	1.67% (1.26%, 2.22%)	1.78% (1.28%, 2.46%)	8	2,926	0.27% (0.14%, 0.54%)	NA (NA, NA)	7	2,460	0.28% (0.14%, 0.59%)	NA (NA, NA)
	Other	29	1,190	2.44% (1.70%, 3.48%)	2.42% (1.59%, 3.65%)	5	1,185	0.42% (0.18%, 0.98%)	NA (NA, NA)	3	1,143	0.26% (0.09%, 0.77%)	NA (NA, NA)
	NA	46	3,221	1.43% (1.07%, 1.90%)	1.76% (1.23%, 2.50%)	14	2,819	0.50% (0.30%, 0.83%)	0.51% (0.28%, 0.92%)	5	2,418	0.21% (0.09%, 0.48%)	NA (NA, NA)
White ethnic group	English/Welsh/Scottish/ Northern Irish/British	1752	142,570	1.23% (1.17%, 1.29%)	1.35% (1.27%, 1.44%)	524	140,193	0.37% (0.34%, 0.41%)	0.41% (0.37%, 0.46%)	179	118,326	0.15% (0.13%, 0.18%)	0.19% (0.16%, 0.23%)
	Irish	30	1,558	1.93% (1.30%, 2.74%)	2.15% (1.40%, 3.29%)	6	1,428	0.42% (0.15%, 0.91%)	NA (NA, NA)	1	1,376	0.07% (0.00%, 0.40%)	NA (NA, NA)
	Gypsy or Irish Traveller	0	42	0.00% (0.00%, 8.41%)	NA (NA, NA)	0	37	0.00% (0.00%, 9.49%)	NA (NA, NA)	0	24	0.00% (0.00%, 14.25%)	NA (NA, NA)
	Other white background	156	7,941	1.96% (1.67%, 2.29%)	2.14% (1.78%, 2.56%)	66	8,201	0.80% (0.62%, 1.02%)	0.95% (0.71%, 1.27%)	14	7,240	0.19% (0.11%, 0.32%)	0.27% (0.14%, 0.51%)
Asian ethnic group	Indian	69	3,183	2.17% (1.69%, 2.74%)	2.42% (1.82%, 3.21%)	26	3,119	0.83% (0.55%, 1.22%)	0.72% (0.47%, 1.10%)	5	2,788	0.18% (0.06%, 0.42%)	NA (NA, NA)
	Pakistani	19	728	2.61% (1.58%, 4.05%)	2.29% (1.40%, 3.72%)	11	818	1.34% (0.67%, 2.39%)	2.08% (1.02%, 4.20%)	3	749	0.40% (0.08%, 1.17%)	NA (NA, NA)
	Bangladeshi	20	327	6.12% (3.78%, 9.29%)	6.03% (3.77%, 9.52%)	3	351	0.85% (0.18%, 2.48%)	NA (NA, NA)	2	337	0.59% (0.07%, 2.13%)	NA (NA, NA)
	Chinese	9	811	1.11% (0.51%, 2.10%)	NA (NA, NA)	1	748	0.13% (0.00%, 0.74%)	NA (NA, NA)	1	718	0.14% (0.00%, 0.77%)	NA (NA, NA)
	Other Asian background	46	1,488	3.09% (2.27%, 4.10%)	3.02% (2.16%, 4.20%)	11	1,480	0.74% (0.37%, 1.33%)	0.55% (0.29%, 1.04%)	3	1,300	0.23% (0.05%, 0.67%)	NA (NA, NA)
Black ethnic group	African	42	1,123	3.74% (2.71%, 5.02%)	3.43% (2.41%, 4.86%)	11	1,417	0.78% (0.39%, 1.38%)	1.02% (0.47%, 2.21%)	4	1,282	0.31% (0.09%, 0.80%)	NA (NA, NA)
	Caribbean	12	553	2.17% (1.13%, 3.76%)	1.90% (1.02%, 3.52%)	1	588	0.17% (0.00%, 0.94%)	NA (NA, NA)	0	547	0.00% (0.00%, 0.67%)	NA (NA, NA)
	Other Black/African/Caribbean background	6	156	3.85% (1.42%, 8.18%)	NA (NA, NA)	2	146	1.37% (0.17%, 4.86%)	NA (NA, NA)	0	136	0.00% (0.00%, 2.68%)	NA (NA, NA)
Mixed ethinic group	White and Black Caribbean	7	489	1.43% (0.58%, 2.93%)	NA (NA, NA)	1	540	0.19% (0.00%, 1.03%)	NA (NA, NA)	2	479	0.42% (0.05%, 1.50%)	NA (NA, NA)
	White and Black African	5	292	1.71% (0.56%, 3.95%)	NA (NA, NA)	0	323	0.00% (0.00%, 1.14%)	NA (NA, NA)	1	294	0.34% (0.01%, 1.88%)	NA (NA, NA)
	White and Asian	18	1,000	1.80% (1.07%, 2.83%)	1.90% (1.11%, 3.23%)	5	1,072	0.47% (0.15%, 1.09%)	NA (NA, NA)	2	906	0.22% (0.03%, 0.80%)	NA (NA, NA)
	Other Mixed/Multiple ethnic background	16	970	1.65% (0.95%, 2.66%)	1.98% (1.15%, 3.39%)	2	991	0.20% (0.02%, 0.73%)	NA (NA, NA)	2	781	0.26% (0.03%, 0.92%)	NA (NA, NA)
Other	Arab	6	296	2.03% (0.75%, 4.36%)	NA (NA, NA)	4	314	1.27% (0.35%, 3.23%)	NA (NA, NA)	1	285	0.35% (0.01%, 1.94%)	NA (NA, NA)
	Other ethnic group	23	930	2.47% (1.57%, 3.69%)	3.02% (2.16%, 4.20%)	2	916	0.22% (0.03%, 0.79%)	NA (NA, NA)	2	899	0.22% (0.03%, 0.80%)	NA (NA, NA)
	Prefer not to say	28	1,430	1.96% (1.31%, 2.82%)	2.55% (1.64%, 3.96%)	7	1,125	0.62% (0.25%, 1.28%)	NA (NA, NA)	1	899	0.11% (0.00%, 0.62%)	NA (NA, NA)
	NA	18	1,755	1.03% (0.61%, 1.62%)	0.99% (0.57%, 1.70%)	6	1,649	0.36% (0.13%, 0.79%)	NA (NA, NA)	4	1,478	0.27% (0.07%, 0.69%)	NA (NA, NA)

* We present weighted prevalence if the number of positives in a category is 10 or more.

Table 5.4 Odds ratios for rounds 5 to 10, and an average weighted inversely to variance for detailed subgroups of ethnicity.

A. Rounds 5 to 8.

Variable	Category	Round 5 Mutually adjusted	Round 6 Mutually adjusted	Round 7 Mutually adjusted	Round 8 Mutually adjusted
Gender	Male	Ref	Ref	Ref	Ref
	Female	0.90 [0.78,1.03]	1.01 [0.91,1.11]	0.93 [0.83,1.04]	0.93 [0.86,1.01]
Age Group	5-12	0.81 [0.59,1.10]	1.00 [0.79,1.25]	1.62 [1.29,2.03]	1.01 [0.84,1.22]
	13-17	1.04 [0.74,1.48]	1.44 [1.13,1.85]	2.48 [1.93,3.18]	1.71 [1.39,2.11]
	18-24	2.25 [1.70,2.99]	1.91 [1.52,2.40]	1.81 [1.39,2.36]	1.73 [1.43,2.11]
	25-34	1.05 [0.79,1.38]	1.33 [1.09,1.61]	1.13 [0.89,1.42]	1.19 [1.01,1.41]
	35-44	Ref	Ref	Ref	Ref
	45-54	1.03 [0.81,1.31]	1.13 [0.95,1.35]	1.07 [0.87,1.31]	1.08 [0.93,1.25]
	55-64	0.69 [0.54,0.90]	1.09 [0.91,1.29]	0.83 [0.67,1.03]	1.02 [0.88,1.18]
	65+	0.63 [0.48,0.82]	0.83 [0.68,1.00]	0.67 [0.54,0.85]	0.80 [0.67,0.94]
Region	North East	3.44 [2.49,4.75]	2.17 [1.69,2.78]	1.28 [0.95,1.73]	0.68 [0.52,0.89]
	North West	3.92 [3.07,5.00]	3.01 [2.55,3.55]	1.20 [0.98,1.47]	0.78 [0.66,0.91]
	Yorkshire	2.49 [1.85,3.37]	2.60 [2.14,3.15]	1.38 [1.10,1.74]	0.57 [0.45,0.71]
	East Midlands	1.78 [1.35,2.35]	1.67 [1.40,2.00]	1.48 [1.23,1.77]	0.68 [0.58,0.79]
	West Midlands	1.62 [1.19,2.20]	1.92 [1.59,2.32]	1.42 [1.16,1.74]	0.85 [0.73,1.01]
	East of England	1.22 [0.91,1.63]	0.94 [0.77,1.16]	0.75 [0.60,0.92]	1.03 [0.90,1.17]
	London	1.46 [1.08,1.98]	1.19 [0.96,1.47]	1.07 [0.86,1.34]	1.43 [1.24,1.64]
	South East	Ref	Ref	Ref	Ref
	South West	1.39 [1.00,1.93]	1.14 [0.91,1.42]	0.82 [0.64,1.04]	0.50 [0.41,0.61]
Key Worker Status	HCW/CHW*	1.02 [0.76,1.38]	1.16 [0.96,1.42]	1.66 [1.35,2.04]	1.48 [1.25,1.77]
	Other essential/key workers	1.04 [0.85,1.26]	1.16 [1.02,1.33]	1.18 [1.01,1.37]	1.36 [1.22,1.52]
	Other worker	Ref	Ref	Ref	Ref
	Not FT, PT, SE*	1.04 [0.86,1.25]	0.93 [0.82,1.06]	0.91 [0.78,1.07]	0.93 [0.82,1.04]
	Unknown	0.98 [0.63,1.51]	1.10 [0.82,1.48]	1.39 [1.05,1.84]	0.76 [0.57,1.02]
Deprivation Index Quintile	1 - Most Deprived	1.10 [0.88,1.38]	1.25 [1.06,1.47]	1.70 [1.41,2.05]	1.41 [1.21,1.65]
	2	0.86 [0.69,1.07]	1.08 [0.92,1.25]	1.34 [1.13,1.60]	1.39 [1.22,1.58]
	3	0.80 [0.65,0.99]	1.04 [0.90,1.20]	1.17 [0.99,1.38]	1.14 [1.00,1.29]
	4	0.91 [0.75,1.10]	1.02 [0.89,1.17]	1.09 [0.93,1.28]	1.13 [1.00,1.28]
	5 - Least Deprived	Ref	Ref	Ref	Ref
Ethnicity	English/Welsh/Scottish/Northern Irish/British	Ref	Ref	Ref	Ref
	Irish	0.95 [0.42,2.13]	1.25 [0.76,2.05]	0.49 [0.20,1.19]	1.41 [0.98,2.04]
	Gypsy or Irish Traveller	NA	NA	2.55 [0.35,18.71]	NA
	Other white background	0.97 [0.67,1.39]	1.23 [0.98,1.53]	1.06 [0.82,1.38]	1.25 [1.05,1.49]
	White and Black Caribbean	0.64 [0.16,2.58]	1.45 [0.75,2.82]	0.85 [0.35,2.07]	0.82 [0.39,1.74]
	White and Black African	NA	0.68 [0.17,2.76]	0.29 [0.04,2.07]	1.05 [0.43,2.55]
	White and Asian	0.39 [0.10,1.59]	0.93 [0.48,1.81]	1.15 [0.63,2.11]	1.15 [0.72,1.84]
	Other Mixed/Multiple ethnic background	1.60 [0.75,3.40]	0.85 [0.42,1.72]	0.72 [0.32,1.61]	1.01 [0.61,1.67]
	Indian	1.63 [1.09,2.44]	1.07 [0.76,1.52]	1.52 [1.11,2.09]	1.36 [1.06,1.74]
	Pakistani	2.57 [1.52,4.35]	2.05 [1.30,3.22]	2.29 [1.40,3.74]	1.58 [0.99,2.50]
	Bangladeshi	5.59 [3.20,9.78]	1.77 [0.78,4.00]	4.32 [2.57,7.27]	3.03 [1.90,4.82]
	Chinese	0.78 [0.25,2.45]	0.58 [0.24,1.40]	0.49 [0.16,1.52]	0.70 [0.36,1.36]
	Other Asian background	1.93 [1.10,3.38]	1.48 [0.93,2.34]	1.33 [0.81,2.16]	1.81 [1.34,2.45]
	African	1.98 [1.13,3.49]	1.47 [0.90,2.40]	1.40 [0.84,2.33]	1.92 [1.39,2.65]
	Caribbean	1.36 [0.43,4.28]	1.34 [0.60,3.02]	0.48 [0.12,1.94]	1.24 [0.70,2.22]
	Other Black/African/Caribbean background	NA	NA	NA	2.25 [0.99,5.12]
	Arab	2.30 [0.85,6.22]	0.93 [0.30,2.92]	1.60 [0.59,4.33]	1.18 [0.52,2.66]
	Other ethnic group	0.68 [0.22,2.14]	1.67 [0.98,2.85]	1.67 [0.94,2.97]	1.51 [0.99,2.29]
	Prefer not to say	0.91 [0.37,2.19]	1.53 [0.94,2.48]	1.80 [1.07,3.01]	1.36 [0.93,1.99]
	Unknown	1.35 [0.64,2.87]	1.07 [0.61,1.85]	1.18 [0.63,2.20]	0.93 [0.58,1.49]

* HCW/CHW: health care worker/care home worker; Not FT, PT SE: not self-employed, full-time, part-time, self-employed.

B. Round 9 to 10 and across rounds 5 to 10.

Variable	Category	Round 9 Mutually adjusted	Round 10 Mutually adjusted	Rounds 5 to 10 Mutually adjusted
Gender	Male	Ref	Ref	Ref
	Female	0.96 [0.82,1.12]	0.88 [0.67,1.15]	0.95 [0.90, 0.99]
Age Group	5-12	1.38 [1.02,1.86]	1.48 [0.91,2.39]	1.16 [0.94, 1.45]
	13-17	1.12 [0.76,1.65]	0.79 [0.40,1.57]	1.41 [1.04, 1.92]
	18-24	1.40 [0.96,2.04]	1.09 [0.57,2.08]	1.8 [1.61, 2.01]
	25-34	1.01 [0.74,1.38]	0.83 [0.48,1.42]	1.16 [1.06, 1.28]
	35-44	Ref	Ref	Ref
	45-54	1.11 [0.85,1.45]	0.83 [0.52,1.33]	1.08 [0.99, 1.17]
	55-64	1.04 [0.79,1.36]	0.77 [0.48,1.22]	0.92 [0.8, 1.06]
	65+	0.64 [0.46,0.87]	0.48 [0.29,0.81]	0.71 [0.63, 0.81]
Region	North East	1.89 [1.32,2.72]	5.28 [2.67,10.46]	1.95 [1.11, 3.41]
	North West	1.60 [1.23,2.09]	4.18 [2.38,7.34]	2.01 [1.16, 3.48]
	Yorkshire	1.09 [0.76,1.55]	5.00 [2.75,9.08]	1.7 [0.94, 3.1]
	East Midlands	1.44 [1.10,1.88]	3.27 [1.85,5.81]	1.5 [1.02, 2.19]
	West Midlands	1.11 [0.81,1.52]	3.41 [1.86,6.23]	1.49 [1.06, 2.09]
	East of England	1.09 [0.83,1.45]	2.31 [1.27,4.19]	1.05 [0.86, 1.3]
	London	1.52 [1.14,2.02]	2.16 [1.14,4.10]	1.34 [1.17, 1.52]
	South East	Ref	Ref	Ref
	South West	0.65 [0.44,0.94]	1.89 [0.95,3.77]	0.92 [0.64, 1.34]
Key Worker Status	HCW/CHW*	1.37 [1.02,1.84]	1.09 [0.66,1.81]	1.32 [1.13, 1.54]
	Other essential/key workers	1.20 [0.99,1.47]	0.90 [0.62,1.29]	1.18 [1.07, 1.3]
	Other worker	Ref	Ref	Ref
	Not FT, PT, SE*	0.95 [0.77,1.18]	0.89 [0.62,1.28]	0.94 [0.88, 1]
	Unknown	1.05 [0.67,1.64]	1.61 [0.79,3.28]	1.08 [0.87, 1.33]
Deprivation Index Quintile	1 - Most Deprived	1.67 [1.29,2.15]	1.44 [0.93,2.23]	1.41 [1.22, 1.62]
	2	1.48 [1.17,1.86]	1.35 [0.91,2.02]	1.22 [1.04, 1.44]
	3	0.98 [0.77,1.24]	1.01 [0.67,1.50]	1.03 [0.92, 1.16]
	4	1.09 [0.88,1.37]	0.96 [0.65,1.43]	1.06 [0.98, 1.13]
	5 - Least Deprived	Ref	Ref	Ref
Ethnicity	English/Welsh/Scottish/Northern Irish/British	Ref	Ref	Ref
	Irish	1.16 [0.52,2.61]	0.57 [0.08,4.12]	1.14 [0.86, 1.51]
	Gypsy or Irish Traveller	NA	0.00 [0.00,Inf]	2.54 [0.35, 18.57]
	Other white background	2.01 [1.54,2.64]	1.29 [0.73,2.27]	1.27 [1.03, 1.57]
	White and Black Caribbean	0.39 [0.06,2.81]	2.09 [0.51,8.53]	1.03 [0.70, 1.52]
	White and Black African	NA	1.76 [0.24,12.75]	0.87 [0.45, 1.69]
	White and Asian	1.08 [0.44,2.62]	1.30 [0.32,5.29]	1.05 [0.79, 1.41]
	Other Mixed/Multiple ethnic background	0.46 [0.11,1.86]	1.51 [0.37,6.16]	0.99 [0.72, 1.34]
	Indian	2.02 [1.35,3.02]	1.13 [0.46,2.78]	1.44 [1.22, 1.71]
	Pakistani	2.68 [1.46,4.93]	1.81 [0.57,5.77]	2.12 [1.70, 2.64]
	Bangladeshi	1.57 [0.50,4.96]	2.87 [0.69,11.89]	3.28 [2.24, 4.80]
	Chinese	0.33 [0.05,2.34]	0.84 [0.12,6.03]	0.63 [0.41, 0.96]
	Other Asian background	1.72 [0.94,3.16]	1.44 [0.45,4.58]	1.66 [1.36, 2.01]
	African	1.46 [0.79,2.70]	1.54 [0.56,4.28]	1.69 [1.38, 2.07]
	Caribbean	0.39 [0.05,2.76]	NA	1.12 [0.75, 1.69]
	Other Black/African/Caribbean background	2.93 [0.72,11.94]	NA	2.40 [1.18, 4.89]
	Arab	2.69 [0.99,7.28]	1.88 [0.26,13.66]	1.62 [1.06, 2.47]
	Other ethnic group	0.52 [0.13,2.09]	1.46 [0.36,5.94]	1.45 [1.11, 1.9]
	Prefer not to say	1.54 [0.72,3.25]	0.68 [0.09,4.87]	1.44 [1.14, 1.82]
	Unknown	1.08 [0.48,2.42]	2.10 [0.77,5.68]	1.12 [0.87, 1.46]

* HCW/CHW: health care worker/care home worker; Not FT, PT SE: not self-employed, full-time, part-time, self-employed.

5.3.3 Household Size and Neighbourhood Deprivation

Participants in larger households consistently had higher rates of infection than in smaller households. Prevalence increased monotonically with household size, and households with six or more people consistently had the highest prevalence in rounds 5 through 10 (Table 5.1). These results were maintained when estimating odds ratios adjusted for age and sex, where participants in households of 3 to 5 people and 6 or more people had overall increased odds of 1.30 (1.18, 1.43) and 1.32 (1.21, 1.44), respectively, compared with participants in households of size 1 or 2. This trend was accentuated by a model adjusted for region, key worker status, ethnicity and deprivation in addition to age and sex, with odds ratios of 1.92 (1.64, 2.24) and 1.80 (1.60, 2.03) respectively. (Figure 5.1B, Table 5.2).

In rounds 5 to 10, participants in the most deprived neighbourhoods consistently had a higher risk of infection than those in the least deprived areas, with an overall increased odds of 1.69 (1.51, 1.89) adjusted for age and sex. The deprivation difference remained after adjusting for age, sex, region, broad categories of employment type, broad categories of ethnicity, and household size, with people in the most deprived quintile having an increased odds of 1.42 (1.23, 1.63) compared to the least deprived (Figure 5.3, Table 5.2).

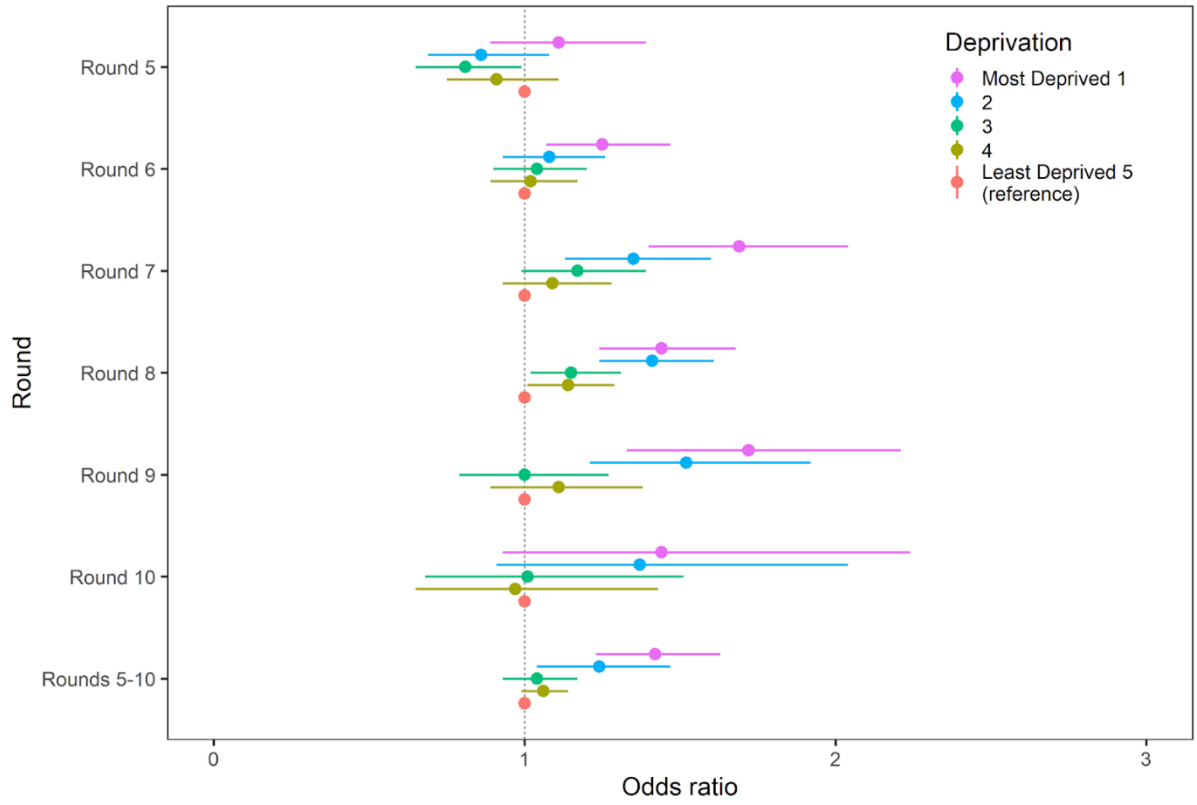


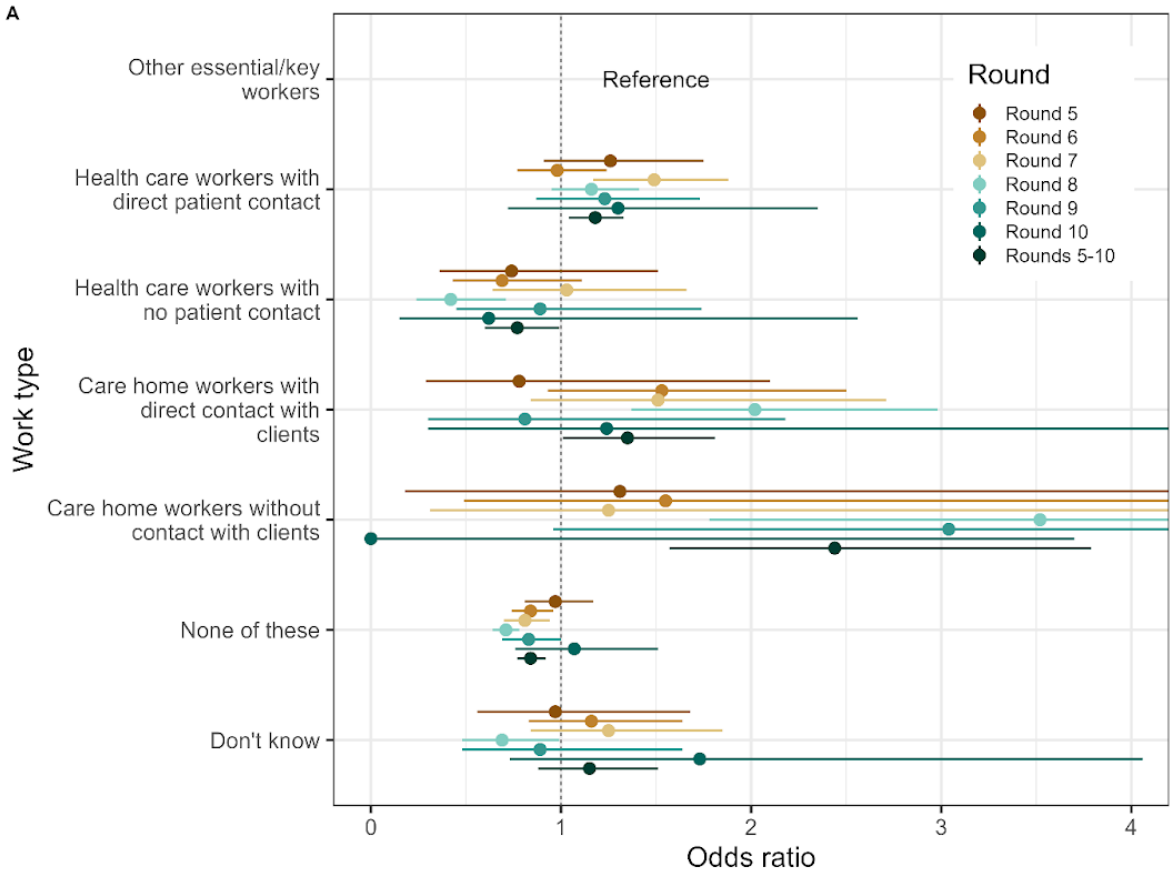
Figure 5.3 Forest plot of deprivation.

Odds ratios were calculated for rounds 5 to 10 and an average across rounds weighted inverse to variance (rounds 5 to 10). The reference group is those living in the least deprived areas (Least Deprived 5). Individual round odds ratios are mutually adjusted for sex, key worker status, broad categories of ethnicity, household size and deprivation index.

5.3.4 Occupation

Occupation was also a risk factor for swab-positivity. Across rounds 5 to 10, compared to non-key workers, the odds of infections in health care and care home workers were 1.32 (1.13, 1.54) (Table 5.2). Among key workers, health care workers with direct patient contact had increased odds of 1.18 (1.04, 1.33), while the odds for health care workers with no patient contact was 0.77 (0.6, 0.99), compared with other essential/key workers (Table 5.5). The odds of swab-positivity were increased among care home workers, both those with direct contact with clients at 1.35 (1.01, 1.81) and those without contact with clients at 2.44 (1.57, 3.79) compared to other essential/key workers (Figure 5.4A, Table 5.5).

We also asked non-key workers about specific activities that they may or may not have had to undertake related to their work. For example, compared to those not occupationally exposed to the activity, participants whose work activity included: food retail or other shop work had increased odds of 1.24 (1.11, 1.38); hospitality had odds of 1.27 (1.09, 1.47); public transport had odds of 1.53 (1.08, 2.17); and participants working in education, school, nursery or childcare had odds of 1.21 (1.11, 1.31) (Figure 5.4B, Table 5.5). The odds of infections for participants who were not required to work outside their homes were lower at 0.73 (0.68, 0.79) compared with those who were required to work outside their homes.


A
Other essential/key workers
Health care workers with direct patient contact
Health care workers with no patient contact
Care home workers with direct contact with clients
Care home workers without contact with clients
None of these
Don't know
Work type
Reference
Round
Round 5
Round 6
Round 7
Round 8
Round 9
Round 10
Rounds 5-10
0
1
2
3
4
Odds ratio

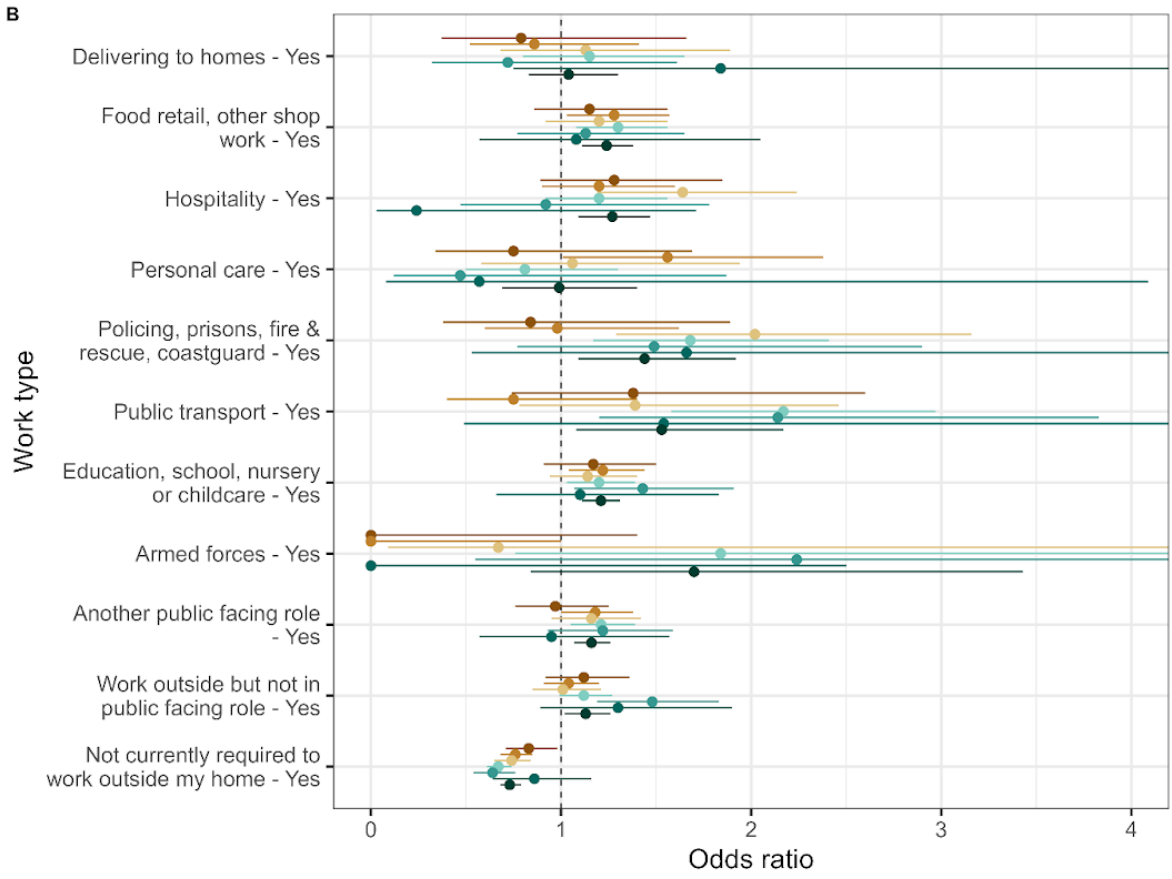

B
Delivering to homes - Yes
Food retail, other shop work - Yes
Hospitality - Yes
Personal care - Yes
Policing, prisons, fire & rescue, coastguard - Yes
Public transport - Yes
Education, school, nursery or childcare - Yes
Armed forces - Yes
Another public facing role - Yes
Work outside but not in public facing role - Yes
Not currently required to work outside my home - Yes
Work type
0
1
2
3
4
Odds ratio

Figure 5.4 Forest plot of occupations for rounds 5 to 10 and an average across rounds weighted inverse to variance (rounds 5 to 10).

A. Reference group is other essential/key workers. Note that the bottom four categories are combined in Figure 4B as HCW/CHW, and combined with full-time and part-time status in the variable named 'key worker status'. **B**. Only people who did not work in health or social care (top four boxes, Figure 5.4A) were asked whether their work involved these activities or settings. Each work activity or setting has a different reference group - those who answered that they did not work with that activity or in that setting. For example, the reference group for those that answered that they delivered to homes was the group who answered that they did not deliver to homes. Individual round odds ratios are mutually adjusted for sex, age group, region, broad category of ethnicity, deprivation index and the specific work activity or setting represented.

Table 5.5 Odds ratios for rounds 5 to 10, and an average weighted inversely to variance for detailed subgroups of occupations.

A. Rounds 5 to 8.

Variable	Category	Round 5 Mutually adjusted	Round 6 Mutually adjusted	Round 7 Mutually adjusted	Round 8 Mutually adjusted
Gender	Male	Ref	Ref	Ref	Ref
	Female	0.90 [0.78,1.03]	1.01 [0.91,1.11]	0.93 [0.83,1.04]	0.93 [0.86,1.01]
Age Group	5-12	0.81 [0.59,1.10]	1.00 [0.79,1.25]	1.62 [1.29,2.03]	1.01 [0.84,1.22]
	13-17	1.04 [0.74,1.48]	1.44 [1.13,1.85]	2.48 [1.93,3.18]	1.71 [1.39,2.11]
	18-24	2.25 [1.70,2.99]	1.91 [1.52,2.40]	1.81 [1.39,2.36]	1.73 [1.43,2.11]
	25-34	1.05 [0.79,1.38]	1.33 [1.09,1.61]	1.13 [0.89,1.42]	1.19 [1.01,1.41]
	35-44	Ref	Ref	Ref	Ref
	45-54	1.03 [0.81,1.31]	1.13 [0.95,1.35]	1.07 [0.87,1.31]	1.08 [0.93,1.25]
	55-64	0.69 [0.54,0.90]	1.09 [0.91,1.29]	0.83 [0.67,1.03]	1.02 [0.88,1.18]
	65+	0.63 [0.48,0.82]	0.83 [0.68,1.00]	0.67 [0.54,0.85]	0.80 [0.67,0.94]
Region	North East	3.44 [2.49,4.75]	2.17 [1.69,2.78]	1.28 [0.95,1.73]	0.68 [0.52,0.89]
	North West	3.92 [3.07,5.00]	3.01 [2.55,3.55]	1.20 [0.98,1.47]	0.78 [0.66,0.91]
	Yorkshire	2.49 [1.85,3.37]	2.60 [2.14,3.15]	1.38 [1.10,1.74]	0.57 [0.45,0.71]
	East Midlands	1.76 [1.35,2.35]	1.67 [1.40,2.00]	1.48 [1.23,1.77]	0.68 [0.58,0.79]
	West Midlands	1.62 [1.19,2.20]	1.92 [1.59,2.32]	1.42 [1.16,1.74]	0.85 [0.73,1.01]
	East of England	1.22 [0.91,1.63]	0.94 [0.77,1.16]	0.75 [0.60,0.92]	1.03 [0.90,1.17]
	London	1.46 [1.08,1.98]	1.19 [0.96,1.47]	1.07 [0.86,1.34]	1.43 [1.24,1.64]
	South East	Ref	Ref	Ref	Ref
	South West	1.39 [1.00,1.93]	1.14 [0.91,1.42]	0.82 [0.64,1.04]	0.50 [0.41,0.61]
Ethnic group	White	Ref	Ref	Ref	Ref
	Asian	2.02 [1.57,2.62]	1.24 [0.98,1.55]	1.63 [1.31,2.03]	1.45 [1.23,1.72]
	Black	1.65 [0.99,2.75]	1.27 [0.83,1.94]	1.05 [0.65,1.68]	1.67 [1.28,2.19]
	Mixed	0.78 [0.43,1.43]	0.97 [0.66,1.42]	0.84 [0.55,1.27]	0.99 [0.73,1.33]
	Other	1.13 [0.53,2.40]	1.48 [0.91,2.40]	1.71 [1.04,2.83]	1.40 [0.96,2.04]
	Unknown	1.12 [0.63,1.99]	1.25 [0.87,1.80]	1.45 [0.97,2.16]	1.11 [0.82,1.49]
Deprivation Index Quintile	1 - Most Deprived	1.10 [0.88,1.38]	1.25 [1.06,1.47]	1.70 [1.41,2.05]	1.41 [1.21,1.65]
	2	0.86 [0.69,1.07]	1.08 [0.92,1.25]	1.34 [1.13,1.60]	1.39 [1.22,1.58]
	3	0.80 [0.65,0.99]	1.04 [0.90,1.20]	1.17 [0.99,1.38]	1.14 [1.00,1.29]
	4	0.91 [0.75,1.10]	1.02 [0.89,1.17]	1.09 [0.93,1.28]	1.13 [1.00,1.28]
	5 - Least Deprived	Ref	Ref	Ref	Ref
Work type 1	1. Health care workers with direct patient contact	1.26 [0.91,1.75]	0.98 [0.77,1.24]	1.49 [1.17,1.88]	1.16 [0.95,1.41]
	2. Health care workers with no patient contact	0.74 [0.36,1.51]	0.69 [0.43,1.11]	1.03 [0.64,1.66]	0.42 [0.24,0.71]
	3. Care home workers with direct contact with clients	0.78 [0.29,2.10]	1.53 [0.93,2.50]	1.51 [0.84,2.71]	2.02 [1.37,2.98]
	4. Care home workers without contact with clients	1.31 [0.18,9.47]	1.55 [0.49,4.90]	1.25 [0.31,5.06]	3.52 [1.78,6.95]
	5. Other essential/key workers	Ref	Ref	Ref	Ref
	6. None of these	0.97 [0.81,1.17]	0.84 [0.74,0.96]	0.81 [0.70,0.94]	0.71 [0.64,0.78]
	7. Don't know	0.97 [0.56,1.68]	1.16 [0.83,1.64]	1.25 [0.84,1.85]	0.69 [0.48,0.99]
Work type 2	1. Delivering to homes - No	Ref	Ref	Ref	Ref
	Delivering to homes - Yes	0.79 [0.37,1.66]	0.86 [0.52,1.41]	1.13 [0.68,1.89]	1.15 [0.80,1.65]
	2. Food retail, other shop work - No	Ref	Ref	Ref	Ref
	Food retail, other shop work - Yes	1.15 [0.86,1.56]	1.28 [1.03,1.57]	1.20 [0.92,1.56]	1.30 [1.08,1.56]
	10. Hospitality - No	Ref	Ref	Ref	Ref
	Hospitality - Yes	1.28 [0.89,1.85]	1.20 [0.90,1.60]	1.64 [1.19,2.24]	1.20 [0.92,1.56]
	11. Personal care - No	Ref	Ref	Ref	Ref
	Personal care - Yes	0.75 [0.34,1.69]	1.56 [1.01,2.38]	1.06 [0.58,1.94]	0.81 [0.50,1.30]
	3. Policing, prisons, fire & rescue, coastguard - No	Ref	Ref	Ref	Ref
	Policing, prisons, fire & rescue, coastguard - Yes	0.84 [0.38,1.89]	0.98 [0.60,1.62]	2.02 [1.29,3.16]	1.68 [1.17,2.41]
	4. Public transport - No	Ref	Ref	Ref	Ref
	Public transport - Yes	1.38 [0.74,2.60]	0.75 [0.40,1.40]	1.39 [0.78,2.46]	2.17 [1.58,2.97]
	5. Education, school, nursery or childcare - No	Ref	Ref	Ref	Ref
	Education, school, nursery or childcare - Yes	1.17 [0.91,1.50]	1.22 [1.04,1.44]	1.14 [0.94,1.40]	1.20 [1.03,1.39]
	12. Childcare - No	Ref	Ref	Ref	Ref
	Childcare - Yes	NA	NA	NA	NA
	6. Armed forces - No	Ref	Ref	Ref	Ref
	Armed forces - Yes	NA	NA	0.67 [0.09,4.80]	1.84 [0.76,4.48]
	7. Another public facing role - No	Ref	Ref	Ref	Ref
	Another public facing role - Yes	0.97 [0.76,1.25]	1.18 [1.00,1.38]	1.16 [0.95,1.42]	1.21 [1.05,1.39]
	8. Work outside but not in public facing role - No	Ref	Ref	Ref	Ref
	Work outside but not in public facing role - Yes	1.12 [0.92,1.36]	1.04 [0.91,1.20]	1.01 [0.85,1.21]	1.12 [0.99,1.27]
	9. Not currently required to work outside my home	Ref	Ref	Ref	Ref
	Not currently required to work outside my home- Y	0.83 [0.71,0.98]	0.76 [0.68,0.85]	0.74 [0.65,0.84]	0.67 [0.61,0.74]

B. Rounds 9 to 10 and across rounds 5 to 10.

Variable	Category	Round 9 Mutually adjusted	Round 10 Mutually adjusted	Rounds 5 to 10 Mutually adjusted
Gender	Male	Ref	Ref	Ref
	Female	0.96 [0.82,1.12]	0.88 [0.67,1.15]	0.95 [0.90, 0.99]
Age Group	5-12	1.38 [1.02,1.86]	1.48 [0.91,2.39]	1.16 [0.94, 1.45]
	13-17	1.12 [0.76,1.65]	0.79 [0.40,1.57]	1.41 [1.04, 1.92]
	18-24	1.40 [0.96,2.04]	1.09 [0.57,2.08]	1.80 [1.61, 2.01]
	25-34	1.01 [0.74,1.38]	0.83 [0.48,1.42]	1.16 [1.06, 1.28]
	35-44	Ref	Ref	Ref
	45-54	1.11 [0.85,1.45]	0.83 [0.52,1.33]	1.08 [0.99, 1.17]
	55-64	1.04 [0.79,1.36]	0.77 [0.48,1.22]	0.92 [0.80, 1.06]
	65+	0.64 [0.46,0.87]	0.48 [0.29,0.81]	0.71 [0.63, 0.81]
Region	North East	1.89 [1.32,2.72]	5.28 [2.67,10.46]	1.95 [1.11, 3.41]
	North West	1.60 [1.23,2.09]	4.18 [2.38,7.34]	2.01 [1.16, 3.48]
	Yorkshire	1.09 [0.76,1.55]	5.00 [2.75,9.08]	1.70 [0.94, 3.10]
	East Midlands	1.44 [1.10,1.88]	3.27 [1.85,5.81]	1.50 [1.02, 2.19]
	West Midlands	1.11 [0.81,1.52]	3.41 [1.86,6.23]	1.49 [1.06, 2.09]
	East of England	1.09 [0.83,1.45]	2.31 [1.27,4.19]	1.05 [0.86, 1.30]
	London	1.52 [1.14,2.02]	2.16 [1.14,4.10]	1.34 [1.17, 1.52]
	South East	Ref	Ref	Ref
	South West	0.65 [0.44,0.94]	1.89 [0.95,3.77]	0.92 [0.64, 1.34]
Ethnic group	White	Ref	Ref	Ref
	Asian	1.65 [1.22,2.21]	1.36 [0.77,2.37]	1.55 [1.34, 1.78]
	Black	1.15 [0.67,1.98]	1.06 [0.39,2.93]	1.39 [1.14, 1.69]
	Mixed	0.55 [0.27,1.12]	1.56 [0.72,3.36]	0.92 [0.77, 1.11]
	Other	0.86 [0.36,2.10]	1.58 [0.50,5.02]	1.41 [1.12, 1.77]
	Unknown	1.27 [0.75,2.17]	1.43 [0.59,3.49]	1.23 [1.03, 1.46]
Deprivation Index Quintile	1 - Most Deprived	1.67 [1.29,2.15]	1.44 [0.93,2.23]	1.41 [1.22, 1.62]
	2	1.48 [1.17,1.86]	1.35 [0.91,2.02]	1.22 [1.04, 1.44]
	3	0.98 [0.77,1.24]	1.01 [0.67,1.50]	1.03 [0.92, 1.16]
	4	1.09 [0.88,1.37]	0.96 [0.65,1.43]	1.06 [0.96, 1.13]
	5 - Least Deprived	Ref	Ref	Ref
Work type 1	1. Health care workers with direct patient contact	1.23 [0.87,1.73]	1.30 [0.72,2.35]	1.18 [1.04, 1.33]
	2. Health care workers with no patient contact	0.89 [0.45,1.74]	0.62 [0.15,2.56]	0.77 [0.60, 0.99]
	3. Care home workers with direct contact with clients	0.81 [0.30,2.18]	1.24 [0.30,5.15]	1.35 [1.01, 1.81]
	4. Care home workers without contact with clients	3.04 [0.96,9.64]	NA	2.44 [1.57, 3.79]
	5. Other essential/key workers	Ref	Ref	Ref
	6. None of these	0.83 [0.69,1.00]	1.07 [0.76,1.51]	0.84 [0.77, 0.92]
	7. Don't know	0.89 [0.48,1.64]	1.73 [0.73,4.06]	1.15 [0.88, 1.51]
Work type 2	1. Delivering to homes - No	Ref	Ref	Ref
	Delivering to homes - Yes	0.72 [0.32,1.61]	1.84 [0.75,4.50]	1.04 [0.83, 1.30]
	2. Food retail, other shop work - No	Ref	Ref	Ref
	Food retail, other shop work - Yes	1.13 [0.77,1.65]	1.08 [0.57,2.05]	1.24 [1.11, 1.38]
	10. Hospitality - No	Ref	Ref	Ref
	Hospitality - Yes	0.92 [0.47,1.78]	0.24 [0.03,1.71]	1.27 [1.09, 1.47]
	11. Personal care - No	Ref	Ref	Ref
	Personal care - Yes	0.47 [0.12,1.87]	0.57 [0.08,4.09]	0.99 [0.69, 1.40]
	3. Policing, prisons, fire & rescue, coastguard - No	Ref	Ref	Ref
	Policing, prisons, fire & rescue, coastguard - Yes	1.49 [0.77,2.90]	1.66 [0.53,5.22]	1.44 [1.09, 1.92]
	4. Public transport - No	Ref	Ref	Ref
	Public transport - Yes	2.14 [1.20,3.83]	1.54 [0.49,4.86]	1.53 [1.08, 2.17]
	5. Education, school, nursery or childcare - No	Ref	Ref	Ref
	Education, school, nursery or childcare - Yes	1.43 [1.07,1.91]	1.10 [0.66,1.83]	1.21 [1.11, 1.31]
	12. Childcare - No	Ref	Ref	Ref
	Childcare - Yes	0.57 [0.14,2.29]	1.69 [0.42,6.83]	0.98 [0.34, 2.83]
	6. Armed forces - No	Ref	Ref	Ref
	Armed forces - Yes	2.24 [0.55,9.07]	NA	1.70 [0.84, 3.43]
	7. Another public facing role - No	Ref	Ref	Ref
	Another public facing role - Yes	1.22 [0.93,1.59]	0.95 [0.57,1.57]	1.16 [1.07, 1.26]
	8. Work outside but not in public facing role - No	Ref	Ref	Ref
	Work outside but not in public facing role - Yes	1.48 [1.19,1.83]	1.30 [0.89,1.90]	1.13 [1.02, 1.26]
	9. Not currently required to work outside my home - No	Ref	Ref	Ref
	Not currently required to work outside my home- Yes	0.64 [0.54,0.76]	0.86 [0.64,1.16]	0.73 [0.68, 0.79]

5.3.5 Sensitivity Analysis for Vaccination Status

As the rollout of the vaccination programme in England began on 8 December 2020 (England, 2020), we included vaccination-related questions in the questionnaire from round 8 (January 2021) and asked participants for consent to longer-term follow-up through linkage to their NHS records. Due to the low number of people vaccinated in round 8, we only linked to NHS vaccination records during rounds 9 and 10 with a match to 114, 747 (69%) participants in round 9 and 98,983 (70%) participants in round 10.

We compared the results from two regression models fitted on linked participants: 1) adjusted for sex, age, region, key worker status, ethnicity, household size and neighbourhood deprivation, 2) adjusted for sex, age, region, key worker status, ethnicity, household size, neighbourhood deprivation, and vaccination status. There were no marked differences between these two sets of models (Table 5.6). For example, the odds of SARS-CoV-2 infection remained higher among health care and care home workers at 1.87 (1.03, 3.38) after adjustment for vaccination status. As expected, we also found participants who received one or two doses of vaccination had lower odds of infection than unvaccinated participants, with odds ratios (ORs) of 0.63 (0.45, 0.86) and 0.61 (0.23, 1.59) respectively (Elliott et al., 2021).

Table 5.6 Odds ratios for rounds 9 and 10, and an average weighted inversely to variance for linked vaccination data.

Variable	Category	Round 9		Round 10		Rounds 9 to 10	
		Mutually adjusted (linked data)**	Mutually adjusted (including vaccination status)	Mutually adjusted (linked data)**	Mutually adjusted (including vaccination status)	Mutually adjusted (linked data)	Mutually adjusted (including vaccination status)
Sex	Male	Ref	Ref	Ref	Ref	Ref	Ref
	Female	1.09 [0.91,1.32]	1.12 [0.93,1.35]	0.86 [0.62,1.19]	0.86 [0.62,1.19]	1.01 [0.8, 1.26]	1.02 [0.79, 1.31]
Age Group	5-12	1.19 [0.82,1.71]	0.94 [0.64,1.37]	1.08 [0.60,1.96]	1.07 [0.58,1.96]	1.16 [0.85, 1.58]	0.97 [0.71, 1.34]
	13-17	1.04 [0.64,1.69]	0.87 [0.53,1.41]	0.78 [0.34,1.79]	0.76 [0.33,1.77]	0.97 [0.64, 1.47]	0.84 [0.55, 1.28]
	18-24	1.17 [0.73,1.88]	1.08 [0.67,1.73]	0.87 [0.39,1.93]	0.86 [0.38,1.91]	1.08 [0.72, 1.63]	1.02 [0.68, 1.53]
	25-34	1.07 [0.74,1.55]	1.03 [0.71,1.49]	0.67 [0.34,1.34]	0.67 [0.34,1.33]	0.93 [0.61, 1.42]	0.92 [0.63, 1.33]
	35-44	Ref	Ref	Ref	Ref	Ref	Ref
	45-54	1.10 [0.80,1.50]	1.24 [0.91,1.70]	0.73 [0.42,1.25]	0.74 [0.43,1.28]	0.95 [0.65, 1.4]	1.01 [0.62, 1.66]
	55-64	1.14 [0.82,1.57]	1.50 [1.06,2.12]	0.74 [0.43,1.29]	0.79 [0.43,1.46]	0.98 [0.65, 1.46]	1.15 [0.62, 2.13]
	65+	0.69 [0.46,1.06]	1.27 [0.79,2.05]	0.39 [0.19,0.79]	0.38 [0.17,0.86]	0.56 [0.33, 0.96]	0.73 [0.23, 2.37]
Region	North East	1.91 [1.22,2.98]	1.88 [1.20,2.94]	4.49 [1.90,10.60]	4.48 [1.89,10.58]	2.69 [1.18, 6.13]	2.67 [1.16, 6.18]
	North West	1.80 [1.31,2.48]	1.82 [1.32,2.51]	4.45 [2.28,8.66]	4.43 [2.27,8.62]	2.69 [1.12, 6.5]	2.7 [1.14, 6.41]
	Yorkshire and The Humber	1.17 [0.76,1.81]	1.17 [0.76,1.80]	5.27 [2.61,10.66]	5.26 [2.60,10.64]	2.42 [0.56, 10.54]	2.41 [0.55, 10.53]
	East Midlands	1.68 [1.23,2.31]	1.69 [1.23,2.32]	3.07 [1.54,6.11]	3.08 [1.55,6.12]	2.1 [1.19, 3.71]	2.1 [1.19, 3.71]
	West Midlands	1.20 [0.82,1.75]	1.22 [0.84,1.79]	2.88 [1.37,6.09]	2.89 [1.37,6.11]	1.75 [0.75, 4.1]	1.77 [0.77, 4.07]
	East of England	1.19 [0.85,1.67]	1.20 [0.86,1.68]	2.20 [1.07,4.49]	2.19 [1.07,4.48]	1.49 [0.84, 2.64]	1.49 [0.85, 2.62]
	London	1.49 [1.04,2.13]	1.52 [1.06,2.18]	2.77 [1.31,5.88]	2.76 [1.30,5.85]	1.86 [1.04, 3.33]	1.86 [1.07, 3.25]
	South East	Ref	Ref	Ref	Ref	Ref	Ref
	South West	0.82 [0.54,1.26]	0.82 [0.54,1.26]	2.10 [0.95,4.62]	2.10 [0.96,4.62]	1.24 [0.5, 3.07]	1.24 [0.5, 3.08]
Key Worker Status	HCW/CHW*	1.52 [1.08,2.13]	2.39 [1.62,3.53]	1.33 [0.73,2.45]	1.29 [0.66,2.51]	1.47 [1.09, 1.98]	1.87 [1.03, 3.38]
	Other essential/key workers	1.18 [0.92,1.50]	1.23 [0.96,1.57]	1.08 [0.69,1.70]	1.08 [0.69,1.69]	1.15 [0.93, 1.43]	1.19 [0.96, 1.48]
	Other worker	Ref	Ref	Ref	Ref	Ref	Ref
	Not FT, PT, SE*	0.94 [0.72,1.22]	1.01 [0.77,1.32]	1.10 [0.69,1.75]	1.11 [0.69,1.76]	0.98 [0.78, 1.23]	1.03 [0.82, 1.3]
	Unknown	1.30 [0.64,2.67]	1.33 [0.65,2.73]	3.34 [1.42,7.88]	3.34 [1.42,7.88]	2.02 [0.81, 5.08]	2.05 [0.83, 5.02]
Ethnicity	White	Ref	Ref	Ref	Ref	Ref	Ref
	Asian	1.89 [1.31,2.72]	1.91 [1.32,2.76]	0.80 [0.32,2.00]	0.80 [0.32,2.00]	1.37 [0.61, 3.08]	1.38 [0.6, 3.13]
	Black	0.62 [0.23,1.68]	0.60 [0.22,1.62]	1.61 [0.50,5.20]	1.62 [0.50,5.24]	0.95 [0.38, 2.39]	0.93 [0.35, 2.48]
	Mixed	0.42 [0.16,1.13]	0.42 [0.15,1.12]	1.33 [0.48,3.67]	1.33 [0.49,3.67]	0.75 [0.24, 2.31]	0.74 [0.24, 2.32]
	Other	0.91 [0.29,2.85]	0.90 [0.29,2.83]	0.87 [0.12,6.32]	0.87 [0.12,6.31]	0.9 [0.33, 2.42]	0.89 [0.33, 2.41]
	Unknown	1.71 [0.91,3.22]	1.71 [0.91,3.21]	1.57 [0.50,4.96]	1.58 [0.50,4.97]	1.68 [0.96, 2.92]	1.68 [0.96, 2.92]
Household Size	1-2 People	Ref	Ref	Ref	Ref	Ref	Ref
	3-5 People	1.55 [1.24,1.94]	1.52 [1.22,1.91]	1.31 [0.88,1.96]	1.31 [0.88,1.96]	1.49 [1.22, 1.81]	1.47 [1.21, 1.79]
	6+ People	1.78 [1.09,2.91]	1.76 [1.07,2.88]	2.00 [0.90,4.44]	2.01 [0.90,4.45]	1.84 [1.21, 2.8]	1.82 [1.2, 2.77]
Deprivation Index Quintile	1 - Most Deprived	1.95 [1.45,2.63]	1.95 [1.44,2.63]	0.98 [0.56,1.74]	0.98 [0.56,1.75]	1.45 [0.75, 2.82]	1.45 [0.75, 2.81]
	2	1.55 [1.17,2.04]	1.54 [1.17,2.04]	1.00 [0.61,1.65]	1.00 [0.61,1.65]	1.31 [0.87, 1.98]	1.31 [0.87, 1.98]
	3	0.85 [0.63,1.15]	0.85 [0.63,1.14]	0.93 [0.59,1.48]	0.93 [0.59,1.48]	0.87 [0.68, 1.12]	0.87 [0.68, 1.12]
	4	1.07 [0.82,1.39]	1.07 [0.82,1.40]	0.89 [0.57,1.40]	0.89 [0.57,1.40]	1.02 [0.81, 1.28]	1.02 [0.81, 1.28]
	5 - Least Deprived	Ref	Ref	Ref	Ref	Ref	Ref
Vaccination status	Unvaccinated	Ref	Ref	Ref	Ref	Ref	Ref
	Vaccinated - 1 dose	NA	0.57 [0.44,0.73]	NA	0.82 [0.47,1.42]	NA	0.63 [0.45, 0.86]
	Vaccinated - 2 doses	NA	0.38 [0.26,0.55]	NA	1.01 [0.59,1.74]	NA	0.61 [0.23, 1.59]

* HCW/CHW: health care worker/care home worker; Not FT, PT SE: not self-employed, full-time, part-time, self-employed.
**The sample size for linked vaccination data is smaller than REACT data used in previous regression analysis. The sample size for round 9 is 114,747, and for round 10 is 98,983.

5.4 Discussion

In this large community-based study of SARS-CoV-2 infection in England, we found during the period September 2020 to March 2021 that minority ethnic subgroups, larger household sizes, increasing neighbourhood deprivation, and key worker status were independently associated with increased odds of testing positive for SARS-CoV-2. During that time, few people were vaccinated, and the perceived risk of infection was high. Our results suggest that substantial inequalities arose because of differential risk of infection rather than access to testing and healthcare.

Our findings build on prior descriptions of antigenic exposure to SARS-CoV-2 in a similar study in England during the period immediately prior to September 2020 (Ward et al., 2021). which suggested that higher hospitalisation and deaths in ethnic minority subgroups were because of their higher infection risk, not because they were intrinsically more vulnerable. Therefore, the results presented here imply that health inequality driven by a higher risk of infection persisted in England from September 2020 through to the following winter, during which there was high excess mortality (Meslé et al., 2021).

Factors driving increased infection in ethnic minorities are not immediately obvious: for example, one study of social mixing patterns during 2020 (Bridgen, Jewell & Read, 2022) found reduced rates of mixing in some ethnic groups, rather than increased rates which would provide a mechanism for increased rates of infection. However, this may have been driven by response bias during a period when many people were at home with more time to enrol in surveys compared with those going out to work (Koltai et al., 2022). In contrast, a US study analysed mobility data as a proxy for mixing at the level of a county, using variation in the ethnic diversity in each county to infer relative levels of mixing for different ethnic groups (Hu et al., 2022). They found that African Americans had the lowest reduction in

mobility and the highest rates of severe disease compared with white and Asian American populations.

These results represent substantial evidence that variation in infection drives variation in disease among ethnic groups. While, other studies have speculated that other factors explain differences in patterns of disease, such as genetic susceptibility (Ishak et al., 2022; D'Antonio et al., 2021; Beckman, Mougeot & Mougeot, 2021), treatment access, and comorbidities (Wang et al., 2020a), such as obesity (Sattar, McInnes & McMurray, 2020), diabetes (Apicella et al., 2020), and hypertension (Zhang et al., 2020), few studies are representative or measure infection, making it difficult to identify the root causes of increased disease accurately. Further studies are warranted to better characterise the contribution of these different factors to health inequality during the current post-pandemic phase of SARS-CoV-2 circulation.

Our study has limitations. First, we aimed to obtain representative samples for the population of England. Although we were able to correct for variations in response and estimate weighted prevalence for the population of England and specific subgroups, the number of people taking part in some subgroups was relatively low; this resulted in higher uncertainty in the estimates (with wider 95% confidence intervals). For example, we chose to present weighted prevalence only if the number of positives in a category was 10 or more; among people of ethnic minorities during some periods, the number of positives was too small to obtain robust estimates for weighted prevalence using classic survey-weighting methods. As a result, some important differences between subgroups were not fully captured. Although other analytical approaches, such as Bayesian hierarchical regression, may be better able to infer patterns where the sample size is small (Gelman & Carpenter, 2020), they would also require additional assumptions about the relationship between hierarchies, which we felt were not justified.

Also, the effects of risk factors might vary over time, for example, due to changes in contact or behavioural patterns among subgroups. Thus, students returning to universities from home after the holiday will increase contacts while travelling and on arrival at university. In addition, as the vaccine programme in England prioritised certain subgroups of the population, they might have acquired sufficient immunity to reduce the risk of infections (Sewell et al., 2020) in ways not captured by our sensitivity analyses. Although we investigated odds ratios for risk factors by rounds, we did not investigate within-round temporal trends in the relationship between risk factors and swab positivity.

Furthermore, response rates were lower for minority ethnic groups. In theory, this could have biased these results. However, we would expect such bias to have decreased rates of infection in these groups, rather than increased them, because people who take part in surveys tend to have lower rates of exposure (National Institute for Health and Care Excellence (NICE), 2020). Further, being present at home during lockdown periods made participation in our study easier, and less likely to have been in contact with other people (Geldsetzer, 2020; Kwok et al., 2020).

In conclusion, our findings show substantial differences in the risk of SARS-CoV-2 infections among minority ethnic groups and other subgroups of the population in England during a period when there was a substantial risk of disease and death from COVID-19. Planning for future severe waves of respiratory pathogens should include policies to reduce inequalities driven by ethnicity, household size, and occupational activity (Government Equalities Office, Race Disparity Unit, 2021).

Chapter 6 Forecasting Regional-level COVID-19 Hospitalisation in England as an Ordinal Variable Using the Machine Learning Method

6.1 Introduction

The COVID-19 pandemic placed a considerable strain on hospitals at varying times and in different locations between January 2020 and late 2021. By 31st December 2022, the World Health Organization (WHO) has reported more than 600 million confirmed cases, including 6.6 million deaths (World Health Organization (WHO), 2020). Given the significant variation in disease incidence across both space and time, even within the same country, local public health authorities often faced challenges in obtaining adequate insights to effectively prioritise health services. This deficiency led to even greater disruption in healthcare delivery than would have been otherwise anticipated. Having advanced and accurate knowledge of higher disease incidence allowed for the postponement of elective care, whereas insight into lower disease incidence facilitated a more expedited refocusing on the backlog of postponed elective procedures. Heightened levels of uncertainty were especially pronounced during the emergence of new COVID-19 variants, which were associated with increased transmission rates and, at times, greater severity.

Specifically, in England and Wales, there were repeated rapid surges in SARS-CoV-2 infections, resulting in high demands for hospitalisations and medical resources for COVID-19 patients. While vaccination efforts have significantly reduced the need for additional interventions to prevent overwhelming hospitals, even as of December 2022, the persistently high rates of infections and mortality have placed substantial pressure on the resources of the National Health Service (NHS). Consequently, the need for frequently updated short-term forecasting at the local level is evident, as it has the potential to markedly enhance the efficiency with which limited healthcare capacity is employed (Sarker et al., 2021).

Non-epidemiological time-varying factors may have collectively influenced the transmission dynamics of SARS-CoV-2 (Ilin et al., 2021; Ganslmeier, Furceri & Ostry, 2021). Numerous studies have detailed the association between COVID-19 cases and climate or meteorological conditions (McClymont & Hu, 2021; Wu et al., 2020c). Multiple investigations conducted in France have confirmed that the integration of weather factors into models can enhance the model's capability to reproduce observed patterns, encompassing the progression of hospital admissions (d'Albis et al., 2021; Paireau et al., 2022).

Changes in human social behaviours swiftly impact the trajectory of an epidemic and the influence of severe disease (Moran et al., 2016; Ainslie et al., 2020). Comprehensive mobility data can be harnessed to illustrate shifts in customary commuting behaviours, patterns of social interaction, and inter-regional transits. These changes in metrics have shown correlations with infection rates and mortality (Klein et al., 2022). In this chapter, we endeavour to expand upon these findings by incorporating weather conditions and mobility patterns into a model designed to predict COVID-19 admissions at the regional level.

Our previous work (Chapter 3) involved the development of a machine learning short-term forecasting technique which demonstrated success through retrospective evaluations in predicting the incidence of influenza-like illness (ILI) as an ordinal variable (Wang, Kwok & Riley, 2023). The method exhibited satisfactory performance in these retrospective assessments. In this study, we expand upon our previous method by integrating local weather conditions and mobility patterns to forecast short-term, weekly COVID-19 admissions for seven NHS regions of England. As such, we conduct a two-step evaluation of this novel method: 1) We examine whether the inclusion of weather and mobility data enhances the accuracy of 1-to-4-week forecast individually and collectively drawing from previous admission patterns; 2) We explore whether the machine learning models can consistently achieve high accuracy across various levels of application.

6.2 Methods

6.2.1 Data

The dataset amalgamates information encompassing epidemiological aspects of COVID-19 (hospital admissions, cases, and deaths), human mobility patterns and weather data. All data were sourced from publicly available outlets.

Epidemiological data were publicly available on the GOV.UK Coronavirus (COVID-19) in the dashboard (UK Health Security Agency, 2020a). This source provided daily hospitalisation data for NHS regions in England, capturing instances of positive COVID-19 tests within 14 days prior to hospitalisation, along with post-admission positive cases. However, for NHS regions, daily cases and death statistics were unavailable. Instead, we extracted case and death data for Lower Tier Local Authorities (LTLAs) and subsequently matched the data of 315 LTLAs in England with their respective NHS regions. This facilitated the aggregation of NHS region-specific- case and death data. Daily death data recorded individuals who died within 28 days of being confirmed COVID-19 cases, with the date reflecting the date of death instead of the date of reporting. Hospital admission data became available on 19th March 2020 which postdates the availability of case and death data. To ensure temporal consistency, the analysis encompassed epidemiological data spanning from 19th March 2020 to 31st December 2022.

Daily hospitalisation, cases and deaths were aggregated by ISO week. Subsequently, the weekly numbers were adjusted relative to the regional population, scaled per 100000 people, thereby generating per capita hospitalisation, cases and death statistics. The regional population were deduced from population estimates at the LTLA levels by the Office for National Statistics (ONS) (Office for National Statistics, 2022). Per capita cases and deaths remained as continuous variables while per capita hospitalisation was converted into an ordinal variable using distinct methodologies described below.

Per capita hospitalisation was converted into an ordinal variable with numerical levels through two techniques: 1) n-tile, and 2) n-uniform interval, where n = 3, 5, and 10. The n-tile method constitutes an unsupervised discretisation method that segments the value range into a specified number of bins (i.e., 3, 5 and 10 in this context), aiming to maintain nearly uniform instances within each bin This effectively translates numeric target values into ordinal quantities by ensuring equivalent-frequency binning. On the other hand, the n-uniform interval strategy divides the range of values into equidistant bins, each containing varying observation quantities. For both methodologies, the higher the value of the level, the higher the number of hospital admissions, with level 1 representing minimal hospitalisation and level n indicating the highest level of hospitalisation.

Human mobility data was obtained from Google, which provides aggregated anonymised information collected from its users of online platforms (Google, 2020). Google mobility data reports the percentage of mobility changes in 7 different categories of location compared to the baseline day to indicate the movement trends in response to the pandemic and lockdown restrictions. The baseline day is defined as the median value from 3rd January to 6th February 2020, and each category of location has its own baseline day. Google stopped reporting new data on 15th October 2022.

Google provides mobility data in England at the sub-regional levels, which is not straightforward for NHS regions. To analyse the data at NHS regions, we re-mapped the sub-regions in Google mobility data to LTLA with the help of the lookup table posted on the GitHub repository *datasciencecampus / google-mobility-reports-data (Data Science Campus, 2020)*. By calculating the mean value of daily mobility data at the LTLA level belonging to different NHS regions, we can derive the daily mobility data for NHS regions. Finally, we obtained weekly mobility data at the NHS-regional level by taking the average of daily mobility by ISO weeks.

Weekly weather data including air temperature at 2 meters (m) and total precipitation were obtained from the fifth-generation reanalysis (ERA5) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hersbach et al., 2020). Time-series data were extracted by identifying the nearest grid point to the centre of each of the 315 LTLAs. To obtain the weekly temperature and weekly total precipitation by NHS regions, the average value across LTLAs in the corresponding NHS regions were calculated respectively. The overall descriptions of data type and predictors were summarised in Table 6.1.

Table 6.1 Detailed descriptions of predictors.

Data type	Predictor	Description
Epidemiological data	Hospitalisation	COVID-19 Patients admitted to hospital
	Cases	New cases by specimen date
	Deaths	New deaths with COVID-19 on the death certificate by date registered
Weather data	Temperature	Temperature at 2 meters (°C)
	Precipitation	Total precipitation (mm)
Mobility data	Grocery and pharmacy	Mobility trends for places such as grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies, etc.
	Parks	Mobility trends for places such as local parks, national parks, public beaches, marinas, dog parks, plazas, and public gardens,etc.
	Transit stations	Mobility trends for places such as public transport hubs such as subway, bus, and train stations, etc.
	Retail and recreation	Mobility trends for places such as restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters, etc.
	Residential	Mobility trends for residential places.
	Workplaces	Mobility trends for places of work.

6.2.2 Primary Forecast Model

XGBoost is a standard classification algorithm for nominal classes, the ordinal information in the class attributes may be discarded partly when applied to ordinal prediction problems, while information can potentially enhance the predictability of the classifier. Therefore, we used a simple method developed by (Frank & Hall, 2001) to enable the underlying learning algorithm, standard XGBoost, in this work, to take advantage of the ordering information contained in the hospital admission levels.

To utilise the ordered class values, the essential step of this method is to transform a k-class ordinal problem into a *k-1* binary classification problem. This is achieved by converting an ordinal attribute, denoted as A^*, with ordinal values V_1, V_2, V_3, and so on up to V_k, into *k-1* binary attributes. Specifically, for each of the first *k-1* values of the original attribute, a binary attribute is created. Each of these binary attributes represents the test $A^* > V_i$, where *i* refers to the corresponding value of the original attribute. By adopting this method, we can effectively convert an ordinal regression problem into a binary classification problem, thereby enabling the application of various binary classification algorithms to the original ordinal problem.

The training process commences by generating novel datasets from the primary dataset, where a distinct dataset is generated for each of the *k-1* newly-formed binary class attributes. We take $k = 5$ as an example here for illustration, i.e., hospital admissions are divided into five levels, where the hospitalisation increases as the value of levels increases. We then can convert it into 4 binary classification problems from the original dataset such that

the binary target is 1 if levels > 1, so the classifier will predict *Pr (Level > 1)*

the binary target is 2 if levels > 2, so the classifier will predict *Pr (Level > 2)*

the binary target is 3 if levels > 3, so the classifier will predict *Pr (Level > 3)*

the binary target is 4 if levels > 4, so the classifier will predict *Pr (Level > 4)*

Subsequently, the XGBoost algorithm is applied to generate a model for each of the newly-created binary datasets separately. In order to predict the class value of an unseen instance, it is necessary to estimate the probabilities of the k original ordinal classes, utilising the (k - 1) models developed in the previous step. The estimation of probabilities for the first and last ordinal class values is determined by a single classifier. The probability of the first ordinal value (*Level = 1*) is computed as 1 - *Pr (Level > 1)*. In the same manner, the probability of the last ordinal value (*Level = 5*) is determined by calculating the probability of *Pr (Level > 4)*. For class values that fall within the range between 1 and 5, the probability is given by a pair

of classifiers. For example, the probability of Pr (*Level = 2*) is given by *Pr (Level > 1) - Pr (Level > 2)*.

Generally, for any ordinal hospital admission level values Vi the probability can be estimated as:

$$Pr\,(V_1) = 1 - Pr\,(Level > V_1)$$

$$Pr\,(V_i) = Pr\,(Level > V_{i-1}) - Pr\,(Level > V_i),\ 1 < i < k$$

$$Pr\,(V_k) = 1 - Pr\,(Level > V_{k-1})$$

During the prediction phase, the *k-1* classifiers are involved to calculate probability of each of the k ordinal class values for an unknown instance by employing the aforementioned approach. The class value with the highest probability is assigned to the instance.

We applied the method described above to the standard XGBoost algorithm to construct a new model that uses ordering information, referred to as the XGBoost ordered model in the later section. The XGBoost ordered model is our primary forecast model and its performance is evaluated in comparison with other baseline models. In addition, the standard XGBoost, which treats each class attribute as a set of unordered values, was used and compared its performance to the ordered XGBoost model. For clarity, in subsequent sections, we refer to the standard XGBoost model as the XGBoost category model.

Comparison null model

Ordered logistic regression (OLR) was employed as a baseline model to serve as a benchmark for evaluating the main model we explained above. The OLR model is an extension of logistic regression, which assumes proportional odds where the effect of the predictors is constant across all levels of the outcome variable (Parry, 2016). It provides a simple but useful starting point for exploring the relationship between the predictors and the

ordinal outcome. The predictors used in this model were limited to epidemiological predictors only (prior one- and two-week hospitalisation levels and prior one-week cases and deaths).

In addition to the OLR model, the null model in which the prediction of the target week is the same as the most recent available observation week, was used as a baseline model as well.

6.2.3 Combinations of Predictors

Hospital admission level is the ordinal outcome for each model. There are three types of data incorporated as predictors in the models: 1) epidemiological data, including ordinal hospitalisation levels, COVID-19 cases and deaths; 2) weather data, including temperature and total precipitation; 3) community mobility trends data in different types of locations. We only applied epidemiological data to the baseline models to provide baseline accuracy for model evaluation. For the XGBoost ordered and category models, we adapted four different combinations of predictors: 1) epidemiological data only; 2) epidemiological and weather data; 3) epidemiological and mobility data and 4) epidemiological, weather and mobility data. Detailed descriptions of data and predictors used for each model are summarised in Table 6.2.

Table 6.2 Summary of predictors incorporated by each model.

Model	Predictor
Baseline model 1 - Null model	Epidemiological data
Baseline model 2 - OLR	Epidemiological data
XGBoost category model	Epidemiological data
	Epidemiological data + Weather conditon
	Epidemiological data + Mobility data
	Epidemiological data + Weather conditon + Mobility data
XGBoost ordered model	Epidemiological data
	Epidemiological data + Weather conditon
	Epidemiological data + Mobility data
	Epidemiological data + Weather conditon + Mobility data

6.2.4 Forecasting

The COVID-19 hospitalisation data can be well-described as a non-stationary time series, with strong autocorrelation. Therefore, instead of randomly splitting data into training and test set. We selected data from the period 9th March 2020 to 31st December 2021 as the training set, and data from 1st January to 31st December 2022 as the test set. We also adopted the extending window approach which has been described in the previous paper (Wang, Kwok & Riley, 2023) to develop multiple-step-ahead forecasts. In brief, we added one new observation to the fit period (training set) in each update of the forecasting. This process of defining training and test sets was applied to all the models used in this work.

In terms of the hyperparameters for the standard XGBoost model and XGboost ordered model, we applied the same values that have been tuned in Chapter 3 (Table 6.3) (Wang, Kwok & Riley, 2023). The analysis was conducted in R version 4.1.3.

Table 6.3 Hyperparameter values used by the XGBoost models.

Hyperparameter	Value
maximum depth of a tree (max_depth)	4
minimum sum of sample weight (hessian) needed in a leaf node (min_child_weight)	1
subsample ratio of the training set (subsample)	1
subsample ratio of conlums used to construct each tree (colsample_bytree)	0.8
minimum loss reduction (gamma)	0
learning rate (eta)	0.02

6.2.5 Evaluation

Models were evaluated according to macro-averaged mean absolute error (mMAE) and accuracy. Macro-averaged mean absolute error (mMAE) refers to the average of the mean absolute error (MAE) of each class. More detail is given in Chapters 2 and 3. Accuracy is defined as the proportion of the number of weeks with correctly predicted hospitalisation levels to the total number of weeks over the test period.

6.3 Results

Key features of the COVID-19 pandemic in England such as the timing and height of peaks in healthcare demand, can be seen in the raw hospitalisation data and in the n-tile ordinal version of these data (Figure 6.1). During the period in which the Alpha variant was dominant, hospitalisation reached a peak level across all the NHS regions in January 2021. However, following the start of the third national lockdown and the rollout of vaccination programs and public health measures, hospitalisation levels began to decrease in the spring and summer. With the emergence of the Delta variants in the middle of 2021, hospitalisations again began to rise during the autumn and early winter of 2021. Omicron drove a further increase in late 2021 into the start of 2022 after which the hospitalisation level exhibited a fluctuating downward trend.

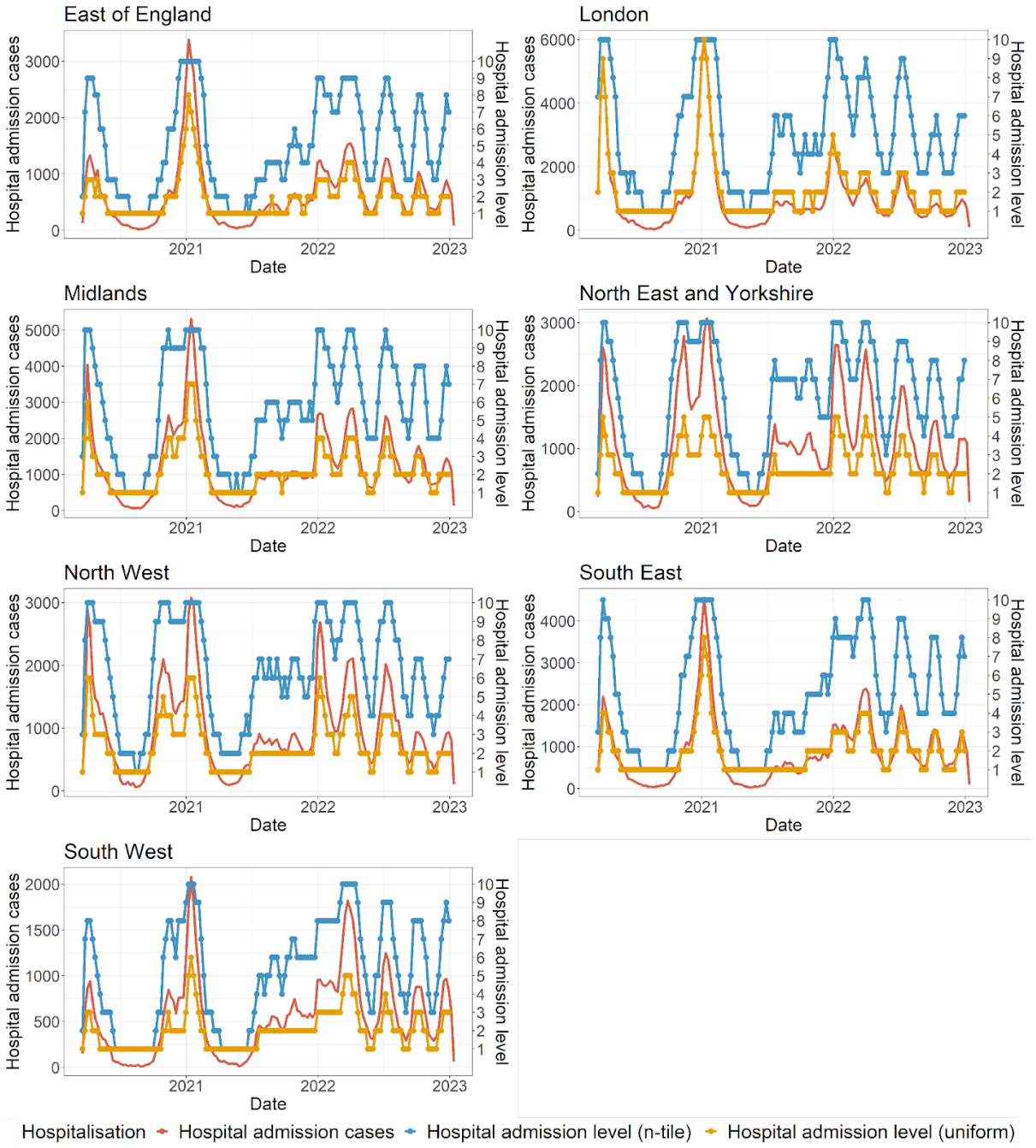


Figure 6.1 Epidemic curve of weekly hospitalisation by NHS regions in England from 19 March 2020 to 31 December 2022.

The left y-axis is the numerical number of weekly new hospital admissions (red line), while the right y-axis is the ordinal weekly hospitalisation level defined by the n-tile method (blue line) and the uniform method (yellow line), where the total number of level values is ten.

Regional COVID-19 hospitalisation data from England provides additional statistical power with which to illustrate how the n-tile ordinal version of these data differs from the uniform ordinal version. The frequency distribution did not discriminate well for the ordinal levels defined by the uniform method (Figure 6.2). For example, with three levels, 90.02% of levels fall into level 1; with 5 levels, 94.81% of them fall into levels 1 and 2; with 10 levels, 88% of the observations between 1 and 3 inclusive (Figure 6.2). Thus, the trend displayed by the uniform ordinal levels appears to be smoother and exhibits less fluctuation compared to the actual numerical trend (Figure 6.1, Figure 6.3). Therefore, we chose to use n-tile levels as the primary outcome and concentrate on evaluating the predictive performance of models using the ordinal outcome defined by the n-tile method. We return to the potential impact of this choice in a later sensitivity analysis.

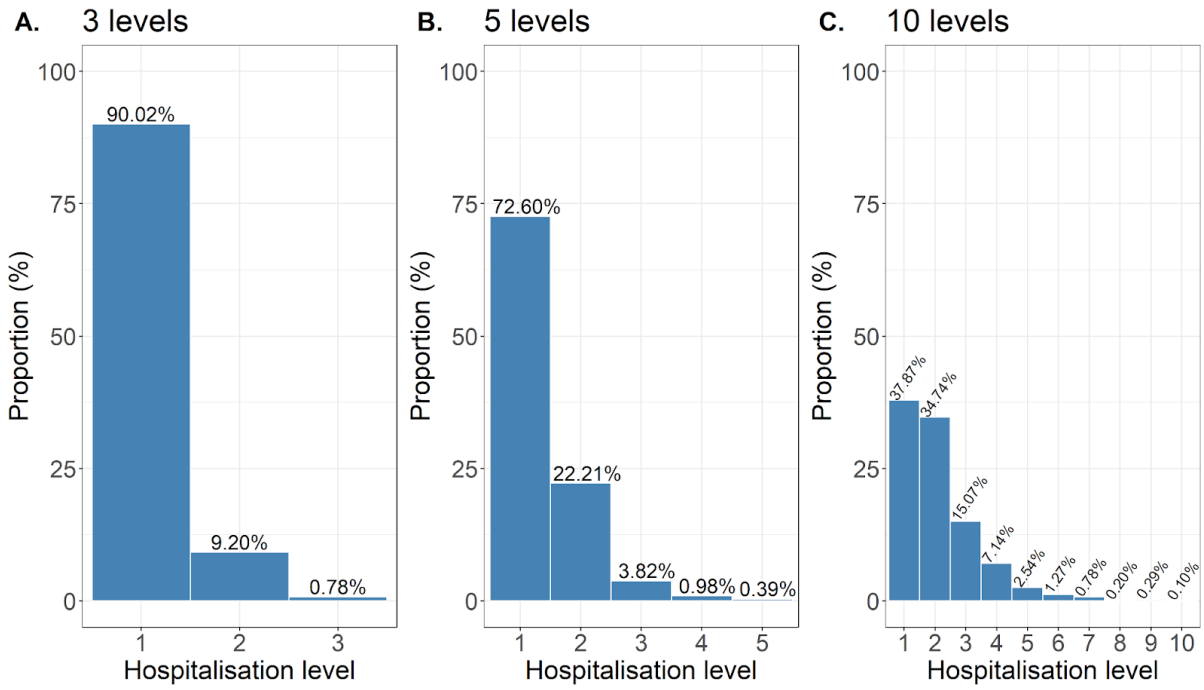
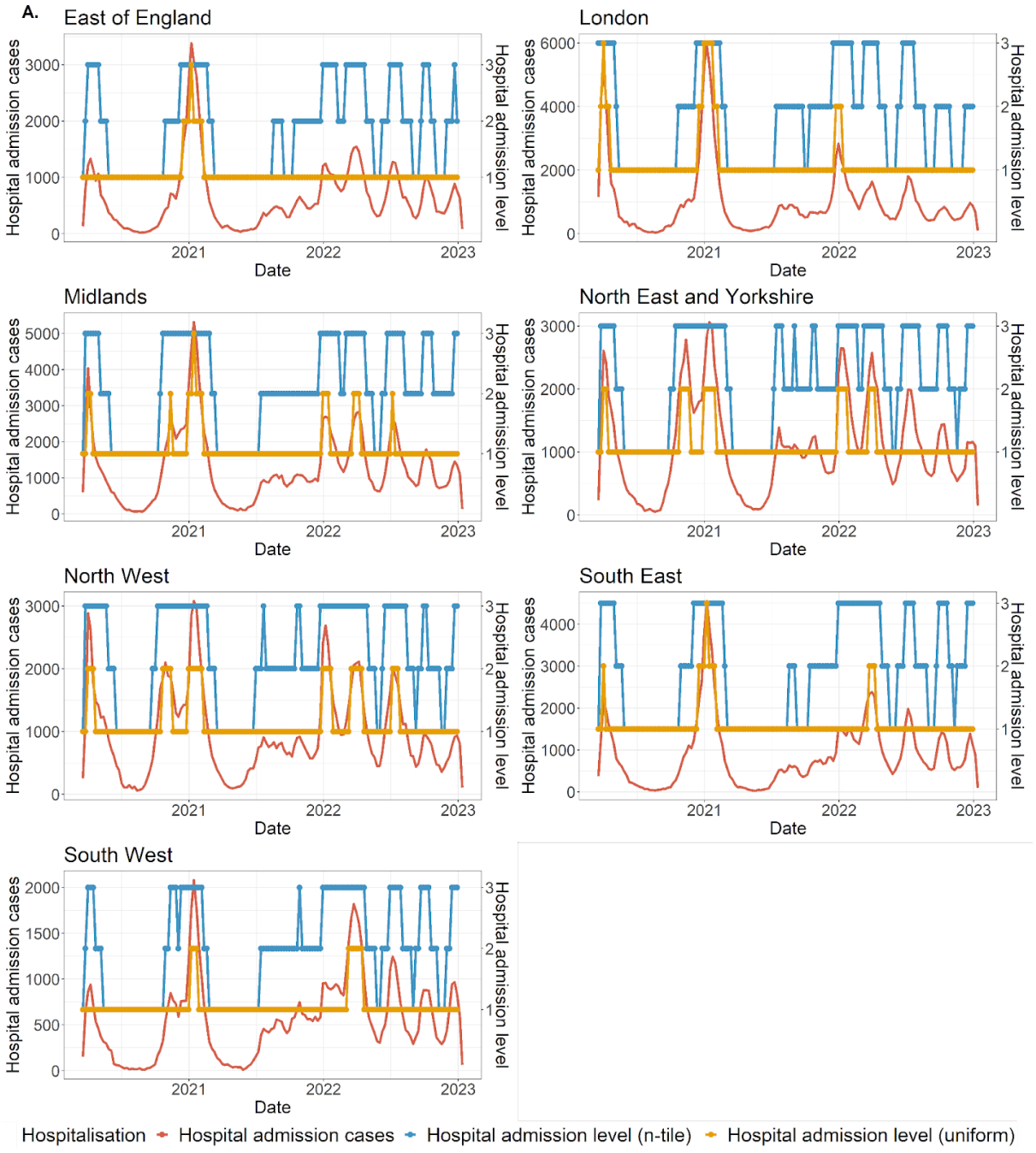


Figure 6.2 Distribution of ordinal levels for weekly hospitalisations for NHS regions discretised by the uniform method.

A. The distribution of hospital admission that discretised into three levels. **B**. The distribution of hospital admission that discretised into five levels. **C**. The distribution of hospital admission that discretised into ten levels.


A.
East of England
London
Midlands
North East and Yorkshire
North West
South East
South West
Hospital admission cases
Hospital admission level
Date
2021
2022
2023
Hospitalisation
Hospital admission cases
Hospital admission level (n-tile)
Hospital admission level (uniform)

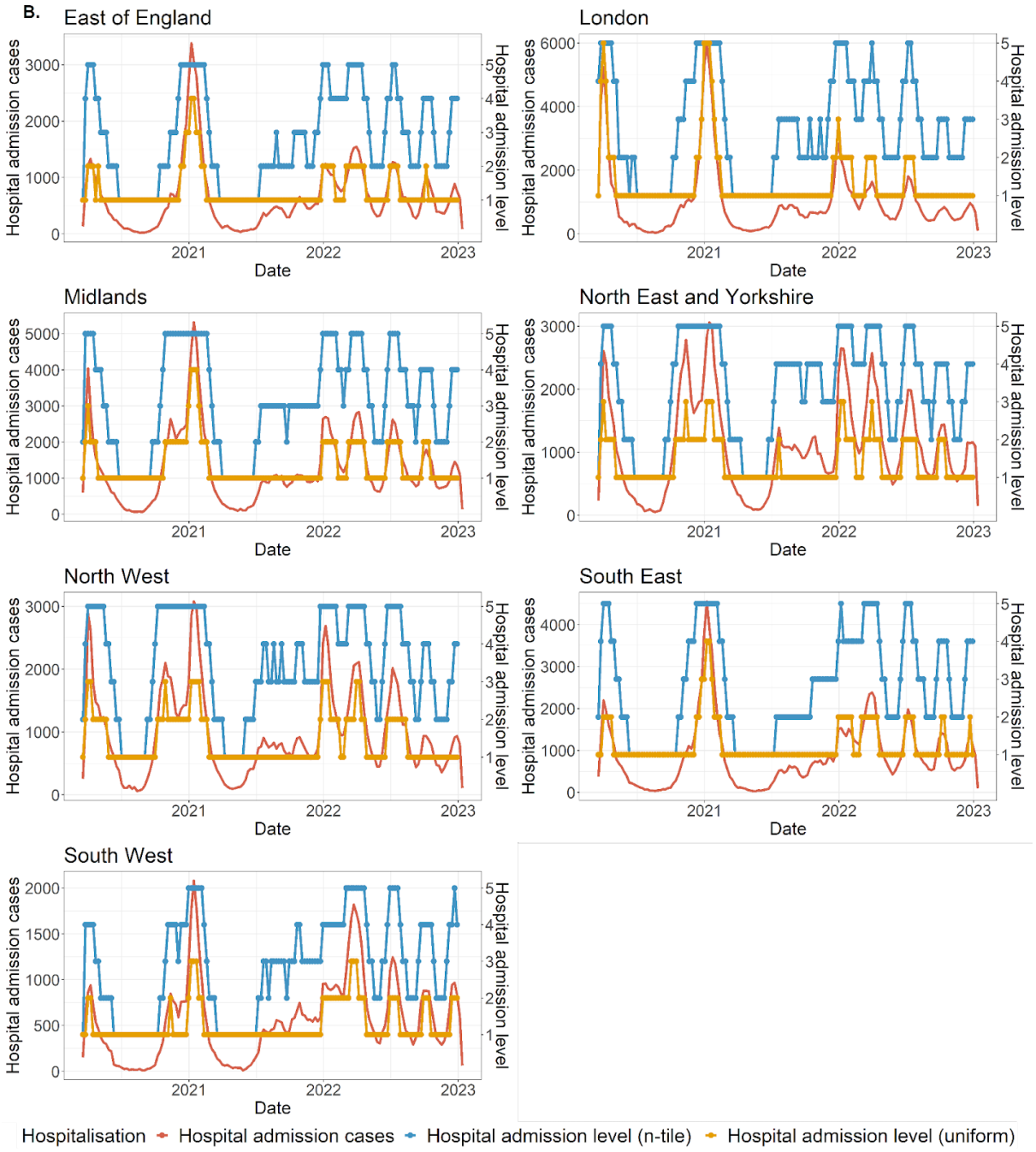


Figure 6.3 Epidemic curve of weekly hospitalisation by NHS regions in England from 19 March 2020 to 31 December 2022.

The left y-axis is the numerical number of weekly new hospital admissions (red line), while the right y-axis is the ordinal weekly hospitalisation level defined by the n-tile method (blue line) and the uniform method (yellow line). **A**. The number of levels equals three. **B**. The number of levels equals five.

We initially assessed the performance of the model by only including epidemiological data, which may be the only option for many populations for which similar data are available during future similar periods. We can consistently generate more accurate forecasts by the XGboost ordered model than baseline models for 1- to 4-week ahead predictions even without incorporating additional potential features such as weather and mobility in the model (Figure 6.4).

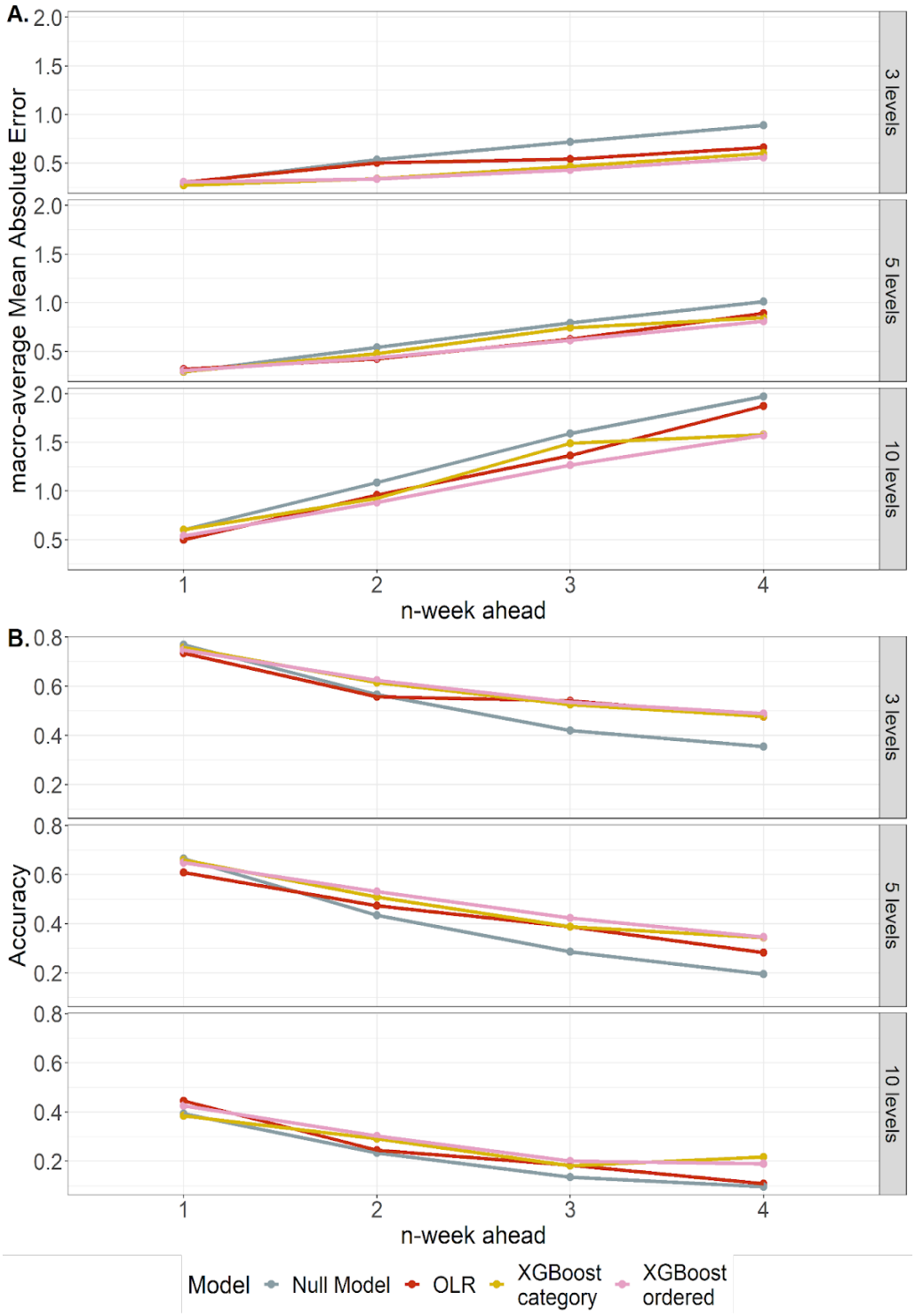


Figure 6.4 Model performance of predicting n-tile hospitalisation levels with only epidemiological predictors.

A. macro-averaged Mean Absolute Errors (mMAE) of the null model (grey line), ordered logistic regression model (red line), the XGBoost category model (yellow line) and the XGBoost model (pink line). **B**. Accuracy of the null model (grey line), ordered logistic regression model (red line), the XGBoost category model (yellow line) and the XGBoost model (pink line).

We then added weather and mobility data into both the XGBoost ordered and category models as predictors respectively, and evaluated model performance by macro-averaged Mean Absolute Error (mMAE). The overall mMAE is the average value across seven NHS regions. The use of XGBoost models with weather data and epidemiological data improved the performance of forecasting models when compared to the baseline models (Null model and OLR) (Figure 6.5). However, when XGBoost models with both weather and epidemiological models were compared with XGBoost models with only epidemiological data, there was no improvement in accuracy (Figure 6.6).

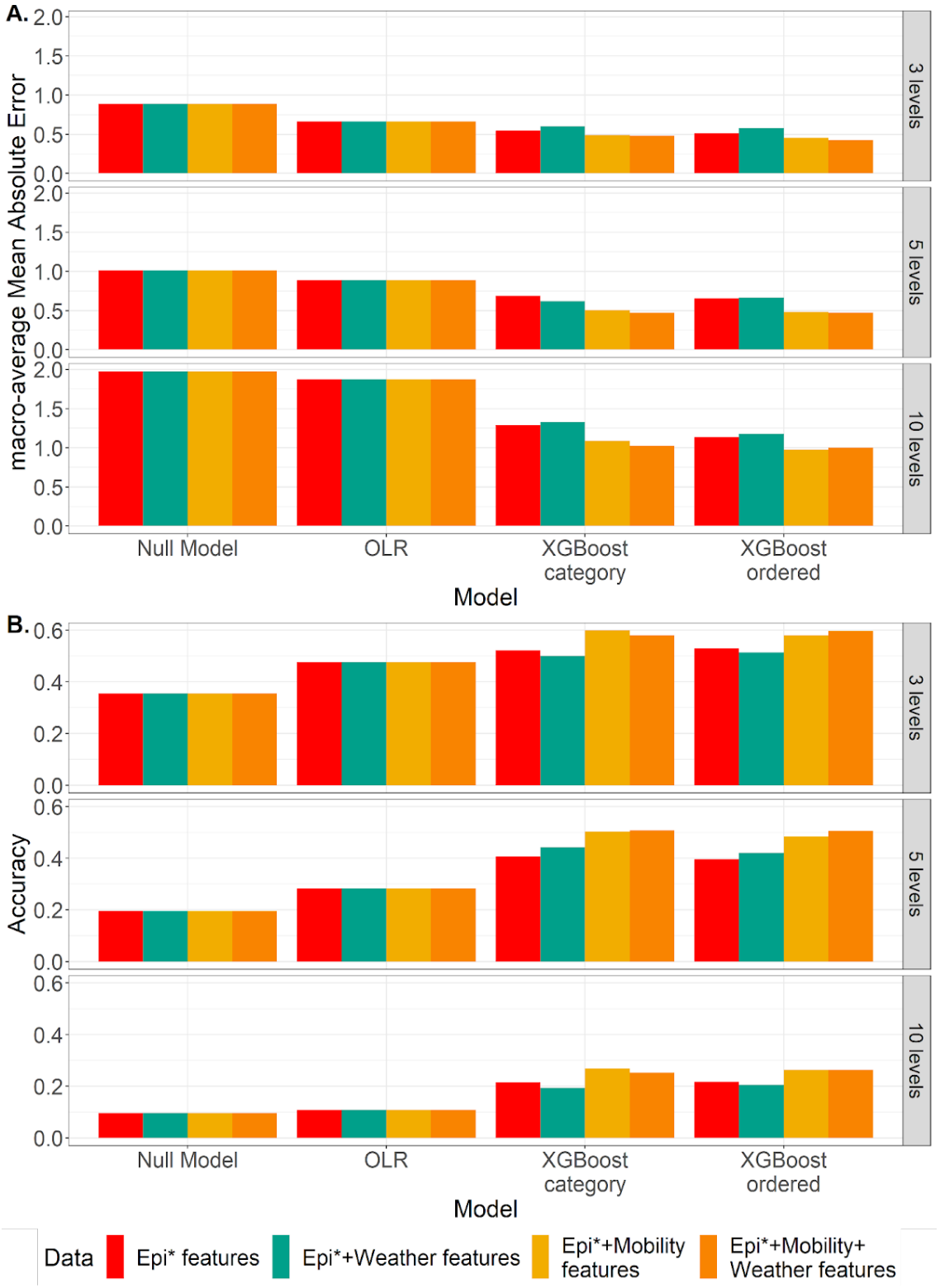


Figure 6.5 Evaluation metrics comparison of 4-week ahead forecasts between models trained with different combinations of predictors.

A. Average macro-averaged Mean Absolute Error (mMAE) over all seven NHS regions. **B**. Average accuracy over all seven NHS regions.
*Epidemiological

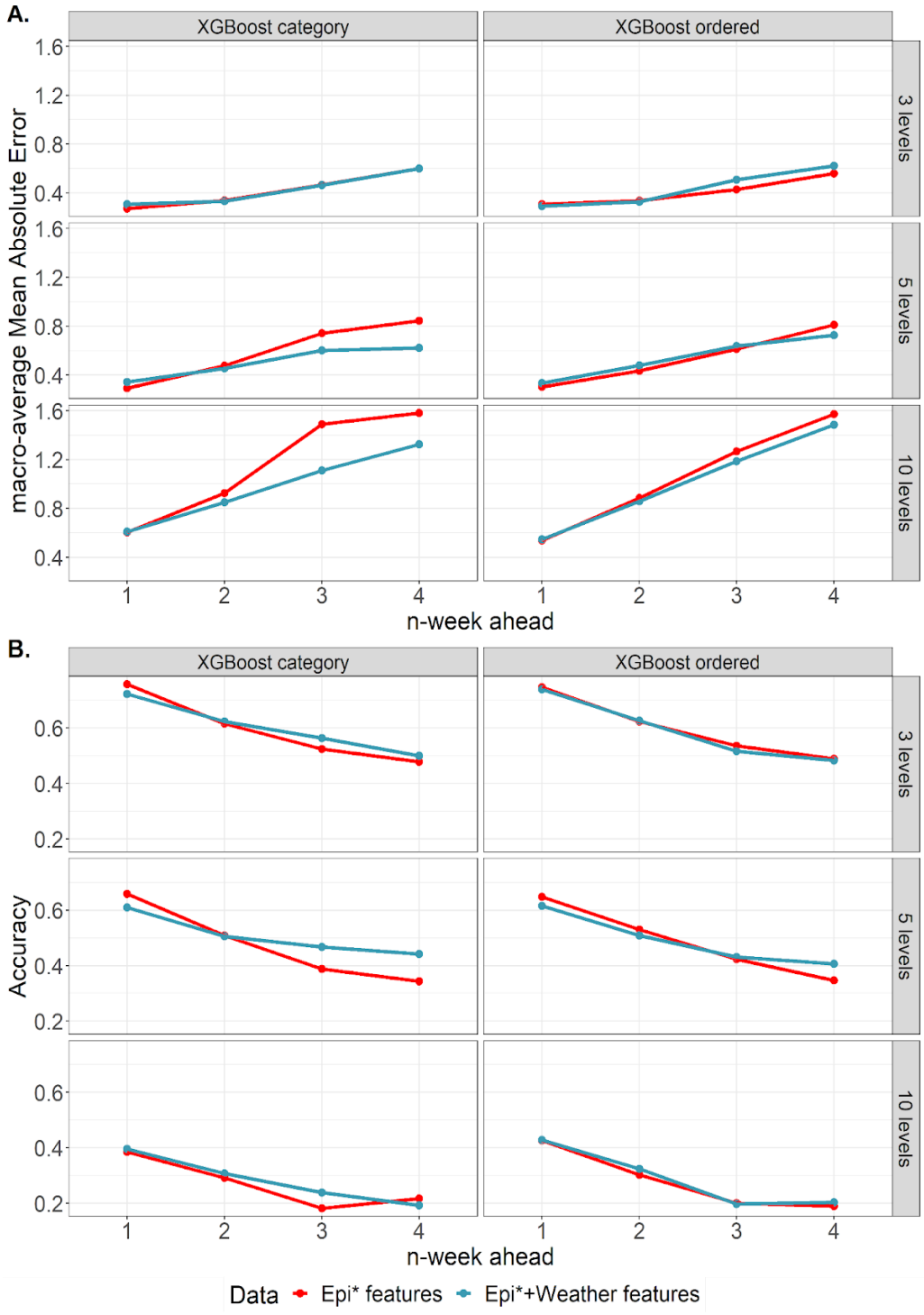


Figure 6.6 Performance of the XGBoost models incorporated with epidemiological and weather data, measured by macro-averaged Mean Absolute Error (mMAE) and accuracy.

The overall mMAE and accuracy were averaged over seven NHS regions for each prediction horizon (1- to 4-week ahead). **A**. mMAEs of the XGBoost models trained with epidemiological and weather features (blue line) and epidemiological features (red line). **B**. Accuracy of the XGBoost models trained with epidemiological and weather features (blue line) and epidemiological features (red line). *Epidemiological

In contrast, our findings suggest that the mobility data does improve forecast accuracy over that which can be achieved with epidemiological data alone (Figure 6.5). Notably, the XGBoost models that solely incorporates epidemiological data performed significantly worse than those incorporating mobility data (Figure 6.7), indicating mobility data should be a key consideration in forecasting models for COVID-19 hospitalisation levels.

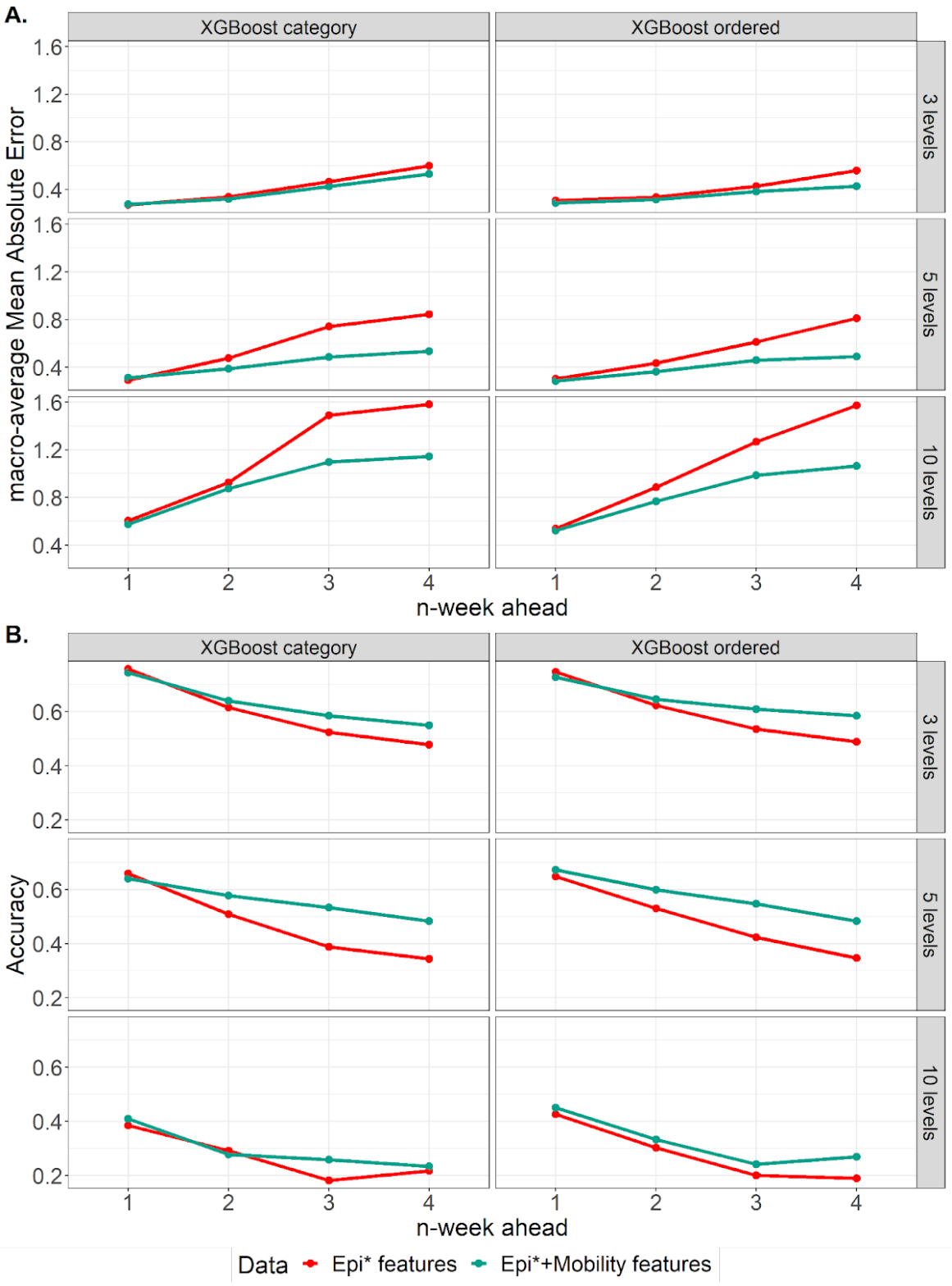


Figure 6.7 Performance of the XGBoost models incorporated with epidemiological and mobility data, measured by macro-averaged Mean Absolute Error (mMAE) and accuracy.

The overall mMAE and accuracy were averaged over seven NHS regions for each prediction horizon (1- to 4-week ahead). **A**. mMAEs of the XGBoost models trained with epidemiological and mobility features (green line) and epidemiological features (red line). **B**. Accuracy of the XGBoost models trained with epidemiological and mobility features (green line) and epidemiological features (red line). *Epidemiological

We finally incorporated both mobility and weather data together in the models. Figure 6.8 illustrates the average mMAE for each of the four combinations of predictors of the XGBoost models being compared over various prediction horizons (1- to 4-week ahead). For each combination, mMAE increases as the prediction horizon increases, highlighting the greater challenge of making predictions further ahead in time. Nevertheless, within each prediction horizon, the models that incorporate mobility data demonstrate better performance. We observe even lower mMAE for 3- and 4-week ahead prediction when we include both weather and mobility as predictors.

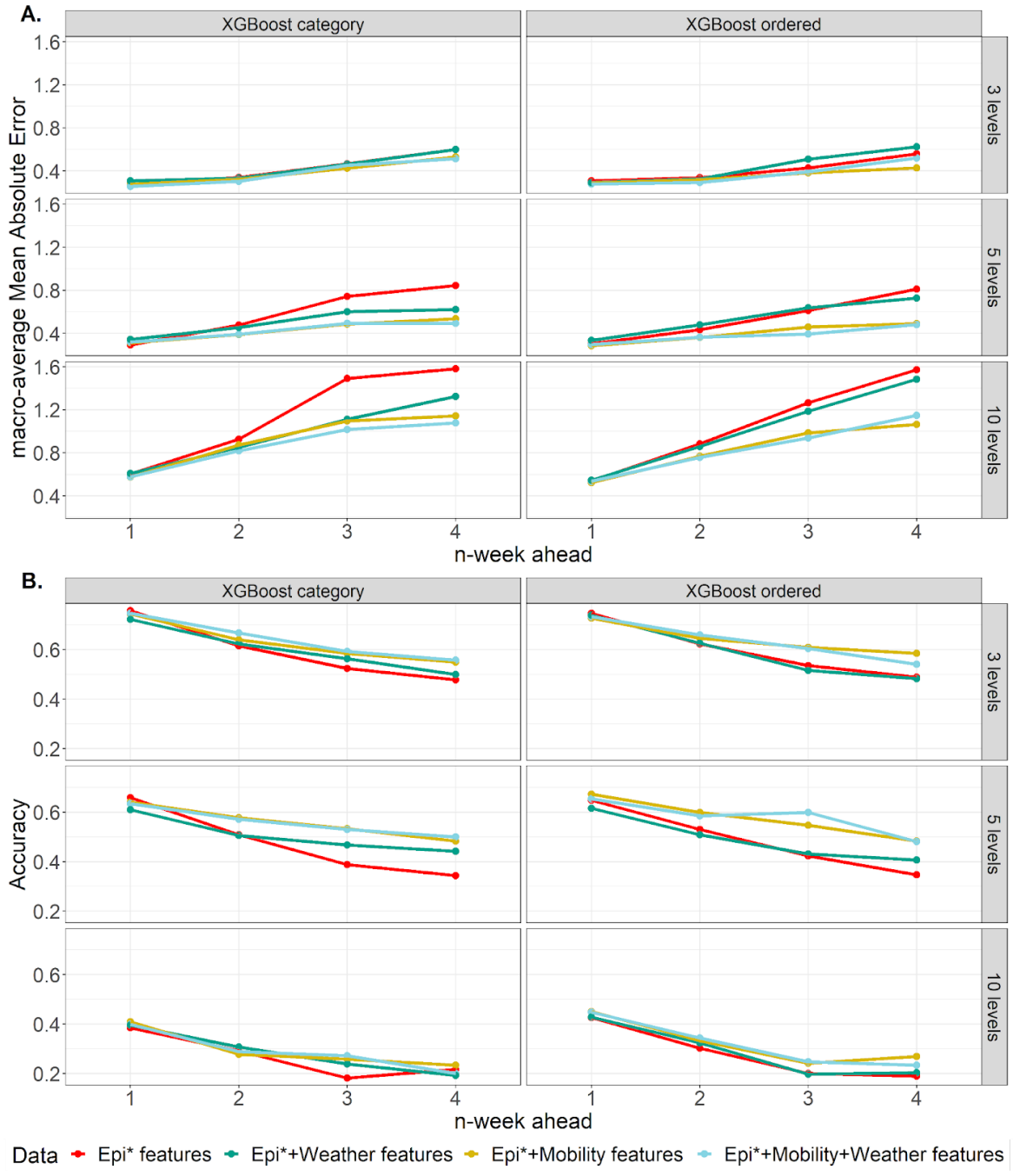


Figure 6.8 Performance of XGBoost models with different sets of features, measured by macro-averaged Mean Absolute Error (mMAE) and accuracy.

The overall mMAE was averaged over 7 NHS regions for each prediction horizon (1- to 4-week ahead). **A**. mMAE of the XGBoost category model and the XGBoost ordered model trained with 1) epidemiological data (red line); 2) epidemiological and weather data (green line); 3) epidemiological and mobility data (yellow line); and 4) epidemiological, mobility and weather features (blue line). **B**. Accuracy of the XGBoost category model and the XGBoost ordered model trained with 1) epidemiological data (red line); 2) epidemiological and weather data (green line); 3) epidemiological and mobility data (yellow line); and 4) epidemiological, mobility and weather features (blue line). *Epidemiological

We also investigated the relative merits of different numbers of ordinal levels and Table 6.3 summarises the results for the 3-, 5-, and 10-bin levels for n-tile bins. It is evident that the XGBoost ordered model outperforms the baseline models in terms of accuracy for both 3 and 5 bins based on the n-tile method, although it is interesting to see that the XGBoost ordered model does not always perform remarkably better. For 10 bins, we find greater differences in accuracy (Table 6.4, Figure 6.8), in particular when compared to the XGBoost category model under the n-tile method. The difference in performance between the XGBoost ordered model and the baseline models becomes more significant, and the XGBoost ordered model outperforms the XGBoost category model on all four combinations of predictors over all prediction horizons. The performance of the XGBoost ordered model across each NHS region also shows the same patterns of accuracy (Figure 6.9). This marked improvement indicates that the contribution of ordering information becomes more pronounced as the number of classes increases.

Table 6.4 Predictive performance of models when the ordinal level of hospital admissions is defined by the n-tile method.

The predictive performance was evaluated for models incorporated with epidemiological, weather conditions, and mobility predictors.

n-week ahead		Bins = 3		Bins = 5		Bins = 10	
		mMAE	Accuracy (%)	mMAE	Accuracy (%)	mMAE	Accuracy (%)
1-week ahead	Null model	0.294	76.92%	0.290	66.48%	0.604	39.29%
	OLR	0.302	73.47%	0.317	60.88%	0.498	44.56%
	XGBoost category	0.298	72.80%	0.305	64.56%	0.605	37.91%
	XGBoost ordered	0.293	71.43%	0.279	67.03%	0.494	45.60%
2-week ahead	Null model	0.536	56.59%	0.540	43.41%	1.088	23.35%
	OLR	0.503	55.78%	0.424	47.28%	0.960	24.49%
	XGBoost category	0.315	65.38%	0.376	58.79%	0.811	31.87%
	XGBoost ordered	0.300	65.93%	0.340	61.26%	0.668	35.44%
3-week ahead	Null model	0.717	42.03%	0.793	28.57%	1.593	13.46%
	OLR	0.540	54.08%	0.624	38.78%	1.366	18.37%
	XGBoost category	0.349	61.81%	0.473	55.77%	0.971	29.40%
	XGBoost ordered	0.346	64.56%	0.388	60.44%	0.832	28.85%
4-week ahead	Null model	0.888	35.44%	1.011	19.51%	1.974	9.62%
	OLR	0.663	47.62%	0.891	28.23%	1.874	10.88%
	XGBoost category	0.483	57.97%	0.470	50.82%	1.024	25.27%
	XGBoost ordered	0.427	59.62%	0.475	50.55%	1.002	26.37%

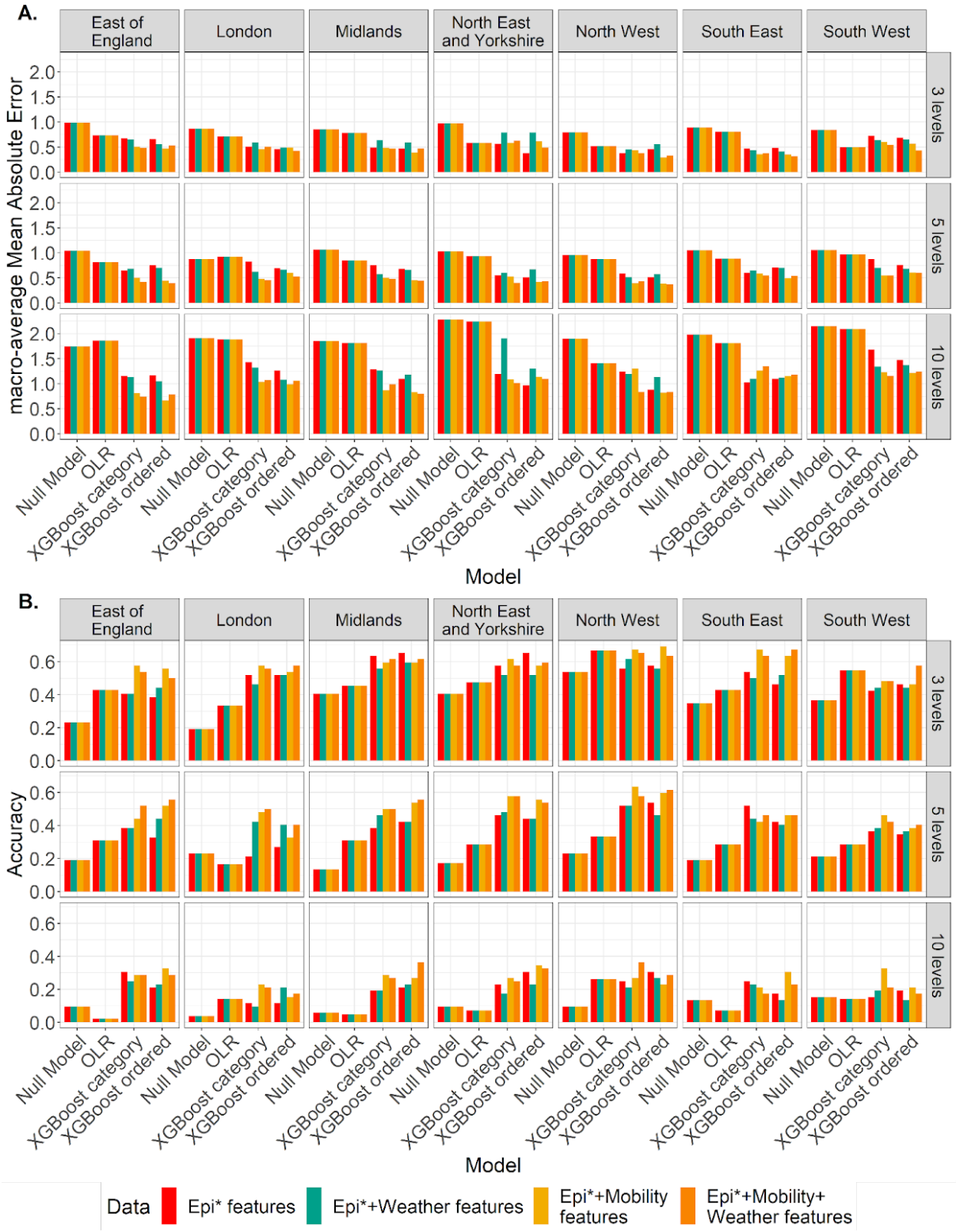


Figure 6.9 Comparison of model performance by NHS regions for XGBoost models and baseline models trained with different combinations of predictors when the prediction horizon is 4 week-ahead for the test period 2022.

The levels are defined by the n-tile method. **A**. macro-averaged Mean Absolute Error (mMAE) of forecasts for hospitalisation levels. **B**. Accuracy of forecasts for hospitalisation levels.
*Epidemiological

To assess the contribution of predictors, “the Gain”, or the relative contribution of the corresponding feature to the model, is used to assess the importance of predictors included in the XGBoost ordered model. The five most important predictors are hospitalisation levels from previous weeks, the number of COVID-19 cases, temperature, mobility changes in retail and recreation places, and the number of deaths with COVID-19 (Figure 6.10). These key predictors confirm the importance of epidemiological data and further improvement of model accuracy with the help of weather and mobility data. Lower importance is observed for total precipitation and other types of mobility trends.

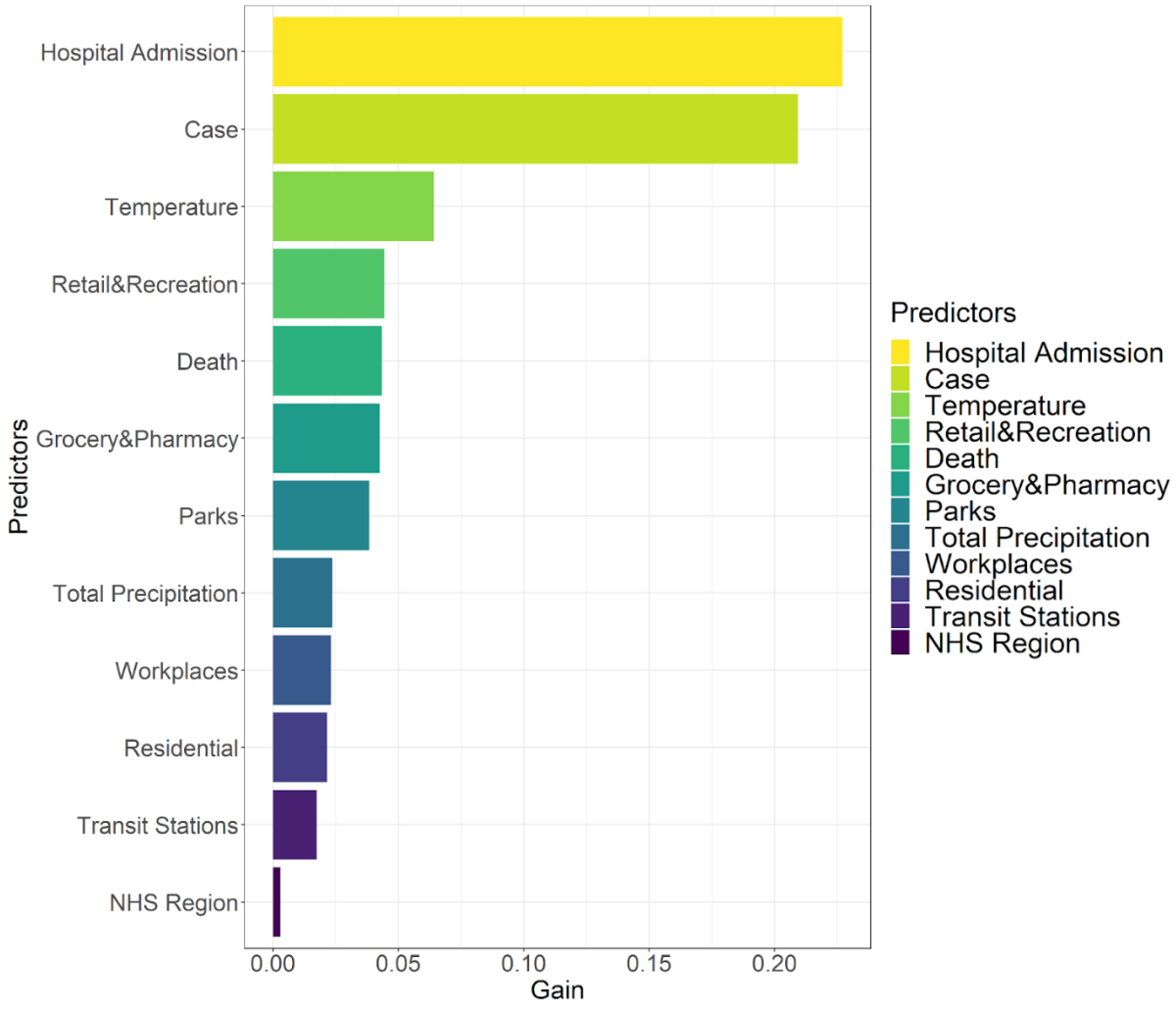


Figure 6.10 The relative importance (Gain) of predictors used by the XGBoost ordered model.

Our findings are confirmed when we compare results for individual regions (Figure 6.11). The hospitalisation levels are defined by the n-quantile method, and they are predicted 1- to 4-week ahead for the entire test period. The prediction errors (mMAEs) for models that add mobility, and mobility and weather are less compared to those with epidemiological or weather alone for all regions. As the class value increases, the magnitude of the difference in prediction errors becomes more pronounced. For example, in the East of England, when the class number is 10, the mMAE of the 2-week ahead forecast only uses epidemiological data is about two times that of the 4-week ahead forecast that includes both weather and mobility data; when the prediction horizon increases to 4-week ahead, the difference of mMAE increase to three times.

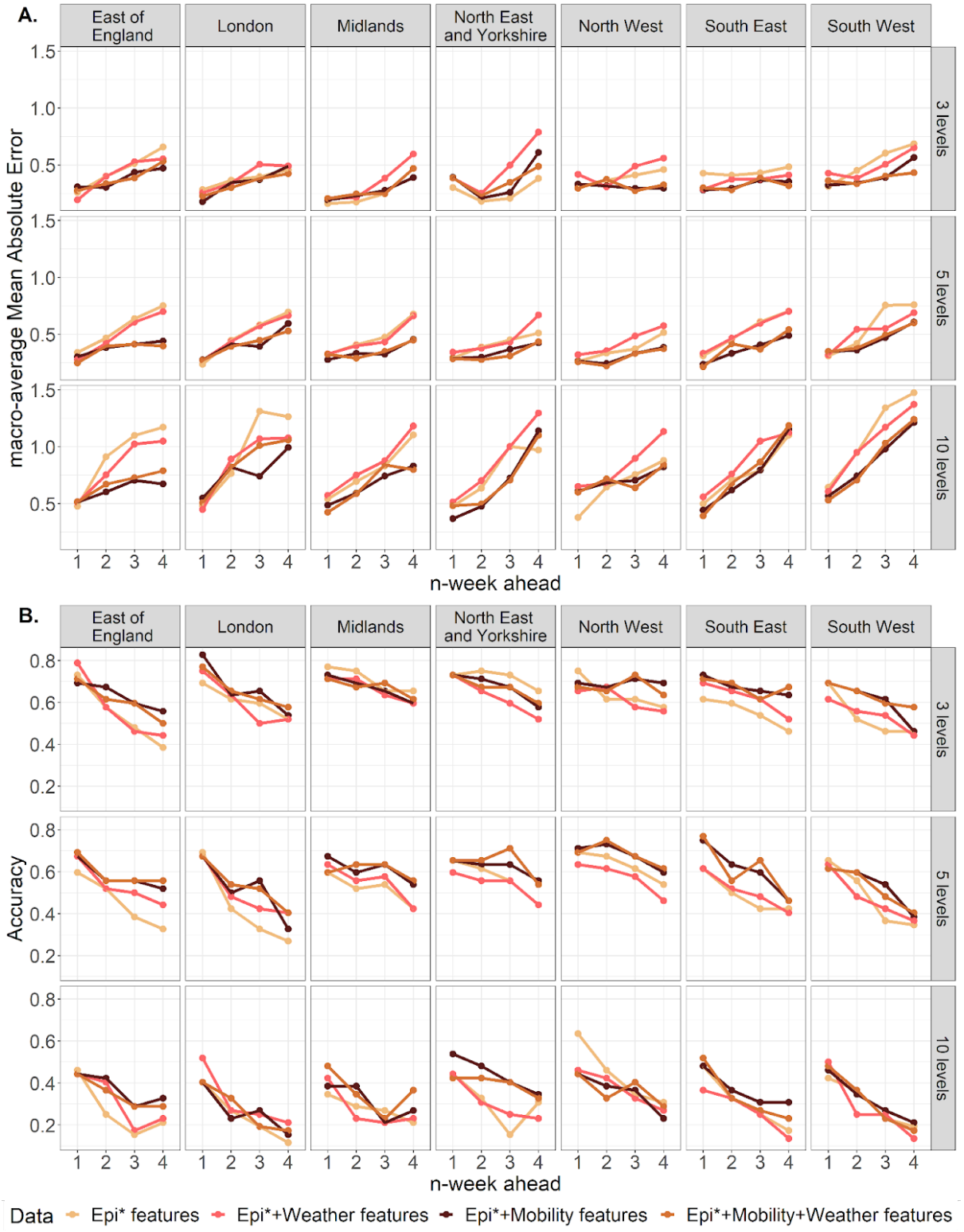
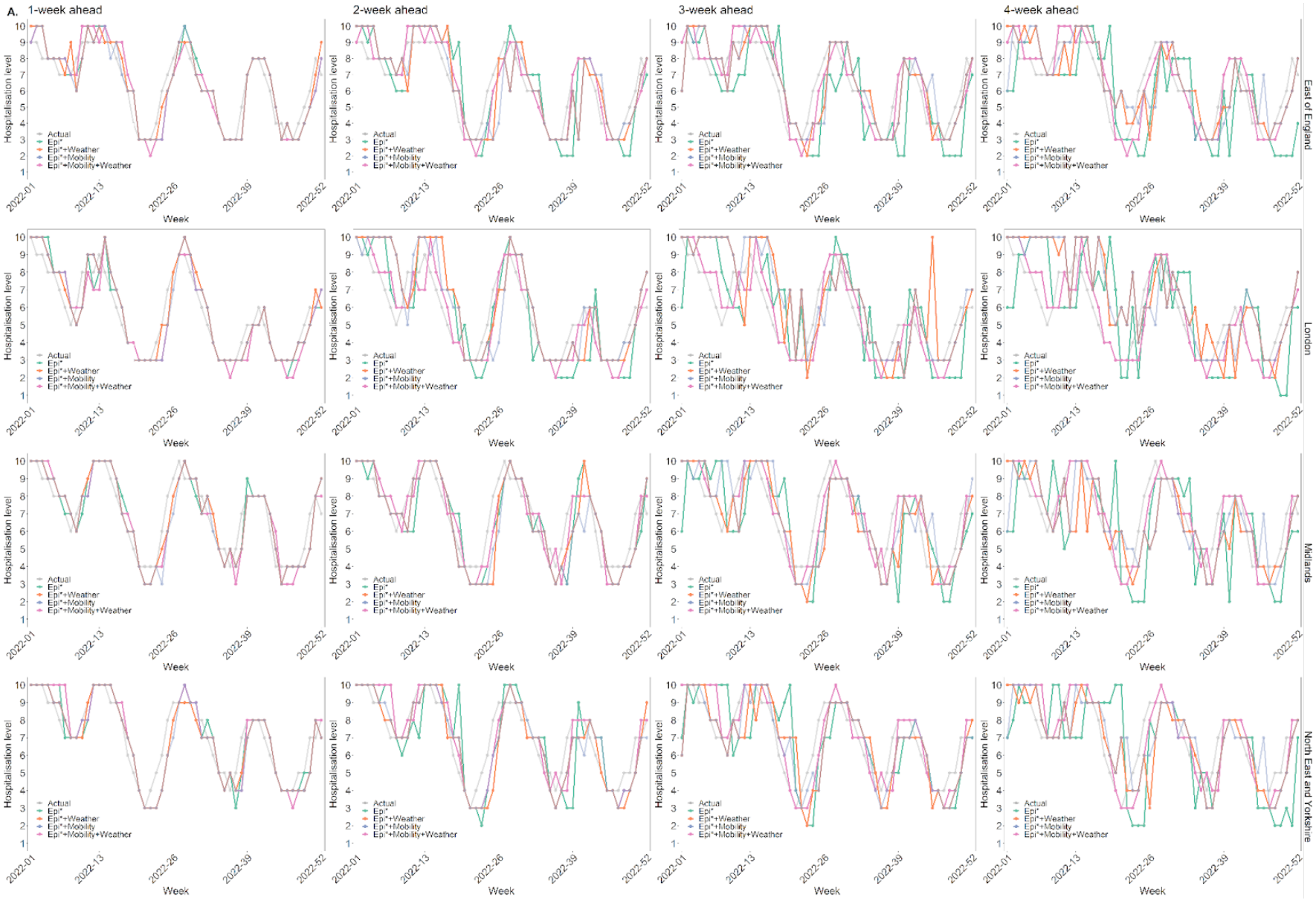


Figure 6.11 Comparison of NHS-region-specific performance for the XGBoost ordered model.

The XGBoost ordered model was trained with 1) epidemiological features (beige line); 2) epidemiological and weather features (pink line); 3) epidemiological and mobility features (brown line); and 4) epidemiological, weather and mobility features (yellow line). The prediction horizon is from 1 to 4 week-ahead for the test period 2022. The number of total levels orders the rows from 3 (top) to 10 (bottom) levels, and levels are defined by the n-tile method. Each column represents one NHS region. **A**. macro-averaged Mean Absolute Error (mMAE) of forecasts for hospitalisation levels. **B**. Accuracy of forecasts for hospitalisation levels.

*Epidemiological

In Figure 6.12, we compared the predicted values from the XGBoost ordered models with the time series trace of the actual hospitalisation values. For all NHS regions, all four combinations of predictors are successful in depicting the general trend of hospitalisation and identifying the peak admission levels for 1- and 2-week forecasts, with the models that include mobility data only and include both mobility and weather data being the most accurate; however, the model with only epidemiological data shows obvious deviations from the actual levels for 3- and 4-week ahead forecast, mainly reflected in the lag in the predicted peaks. The models with mobility data and with both mobility and weather data do not fully capture the fluctuations that happened at the beginning of 2020 for 3- and 4-week ahead forecasts, but their predictions up to 4-week ahead are still the closest to the actual observed peak.


A.
1-week ahead
2-week ahead
3-week ahead
4-week ahead
East of England
London
Midlands
North East and Yorkshire
Hospitalisation level
Week
Actual
Epi*
Epi*+Weather
Epi*+Mobility
Epi*+Mobility+Weather
2022-01
2022-13
2022-26
2022-39
2022-52

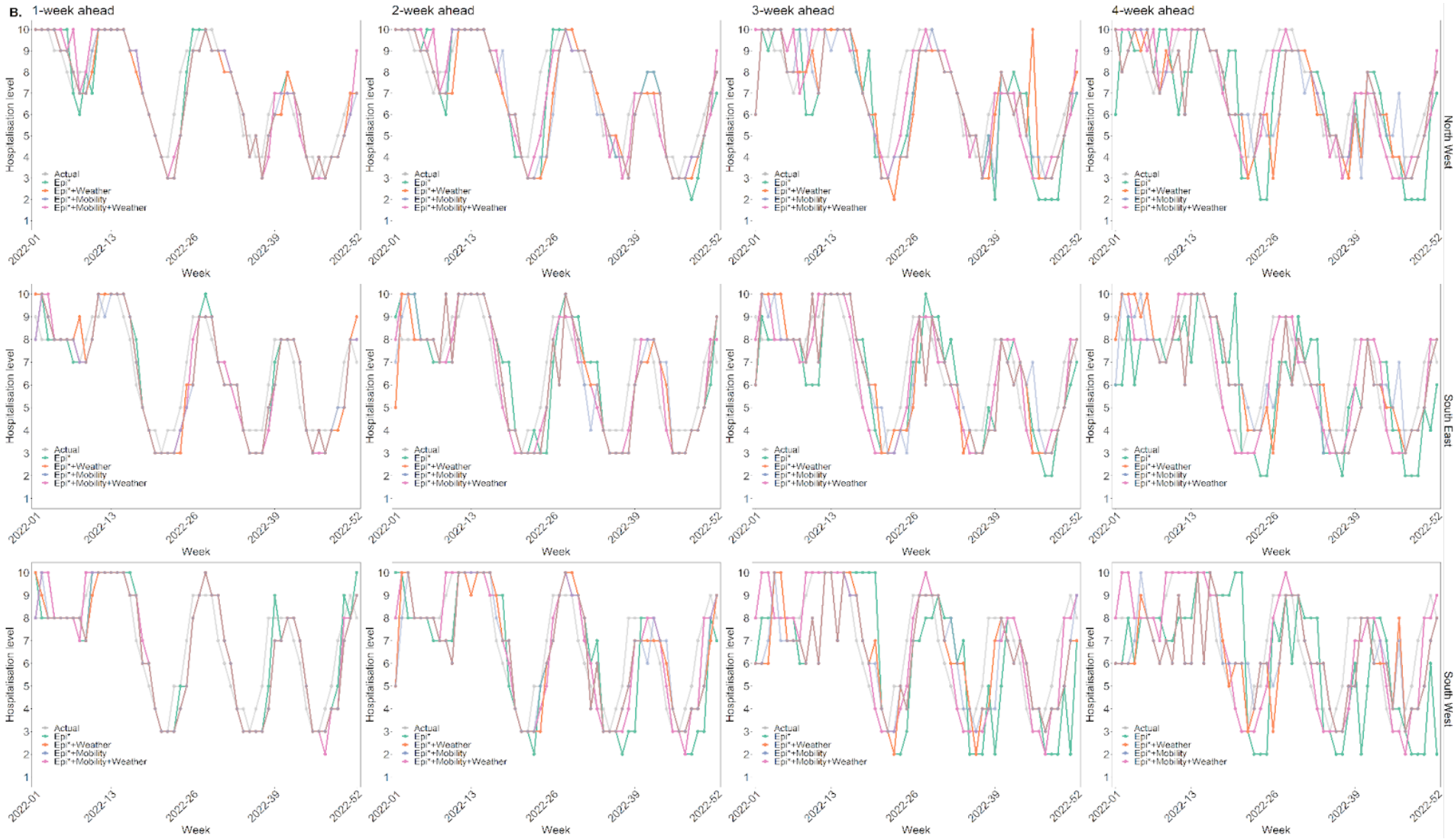

B.
1-week ahead
2-week ahead
3-week ahead
4-week ahead
North West
South East
South West
Hospitalisation level
Week
Actual
Epi*
Epi*+Weather
Epi*+Mobility
Epi*+Mobility+Weather
2022-01
2022-13
2022-26
2022-39
2022-52

Figure 6.12 Comparison of actual and predicted hospitalisation levels by NHS regions.

The predicted hospitalisation levels are predicted by the XGBoost ordered model with 10 levels defined by the n-tile method, and the time ranges from week 1 to week 52 in 2022. Each row represents one NHS region. Columns are prediction horizons ordered from 1- (left) to 4-week (right) ahead. **A**. East of England, London, Midlands and North East and Yorkshire; **B**. North West, South East and South West.

Finally, as a sensitivity analysis, we return to the possibility of using the uniform levels rather than n-tile levels in these analyses. Because of the skewed frequency distribution with uniform levels (Figure 6.2), absolute accuracy is not a useful way to compare the scales: we would expect the mMAE to be lower or the accuracy to be higher for uniform than than the n-tile method because far more observations are in the same level. Therefore, in order to gain some insights into the merits of using the n-tile versus uniform method, we compared how the accuracy changed with different model choices and different combinations of predictors under the two scales: n-tile and uniform method. We observed that with the same number of bins, compared to the baseline models, mMAE and accuracy of XGBoost models changed little, even including additional predictors (Table 6.5, Figure 6.13). On the contrary, the n-tile methods demonstrated improvement in both mMAE and accuracy when using the XGBoost ordered model and incorporating additional weather conditions and mobility predictors (Table 6.4, Figure 6.13).

Table 6.5 Predictive performance of models when the ordinal level of hospital admissions is defined by the uniform method.

The predictive performance was evaluated for models incorporated with epidemiological, weather conditions, and mobility predictors.

n-week ahead		Bins = 3		Bins = 5		Bins = 10	
		mMAE	Accuracy (%)	mMAE	Accuracy (%)	mMAE	Accuracy (%)
1-week ahead	Null model	0.068	95.05%	0.100	84.62%	0.187	65.93%
	OLR	0.169	89.12%	0.167	75.85%	0.309	56.12%
	XGBoost category	0.077	94.23%	0.146	77.75%	0.224	63.74%
	XGBoost ordered	0.078	93.96%	0.133	79.40%	0.206	65.38%
2-week ahead	Null model	0.147	89.84%	0.203	70.05%	0.373	38.74%
	OLR	0.220	81.63%	0.261	65.65%	0.504	35.71%
	XGBoost category	0.170	89.84%	0.205	75.55%	0.270	51.92%
	XGBoost ordered	0.239	88.19%	0.182	73.63%	0.283	53.85%
3-week ahead	Null model	0.274	84.34%	0.289	57.69%	0.523	23.63%
	OLR	0.308	74.83%	0.348	58.16%	0.736	32.65%
	XGBoost category	0.172	87.36%	0.209	74.18%	0.460	50.55%
	XGBoost ordered	0.186	85.71%	0.216	73.08%	0.367	52.47%
4-week ahead	Null model	0.303	74.15%	0.344	49.73%	0.630	17.86%
	OLR	0.278	82.14%	0.405	54.08%	0.823	27.55%
	XGBoost category	0.211	88.19%	0.251	71.15%	0.452	47.53%
	XGBoost ordered	0.211	88.19%	0.239	70.33%	0.412	47.80%

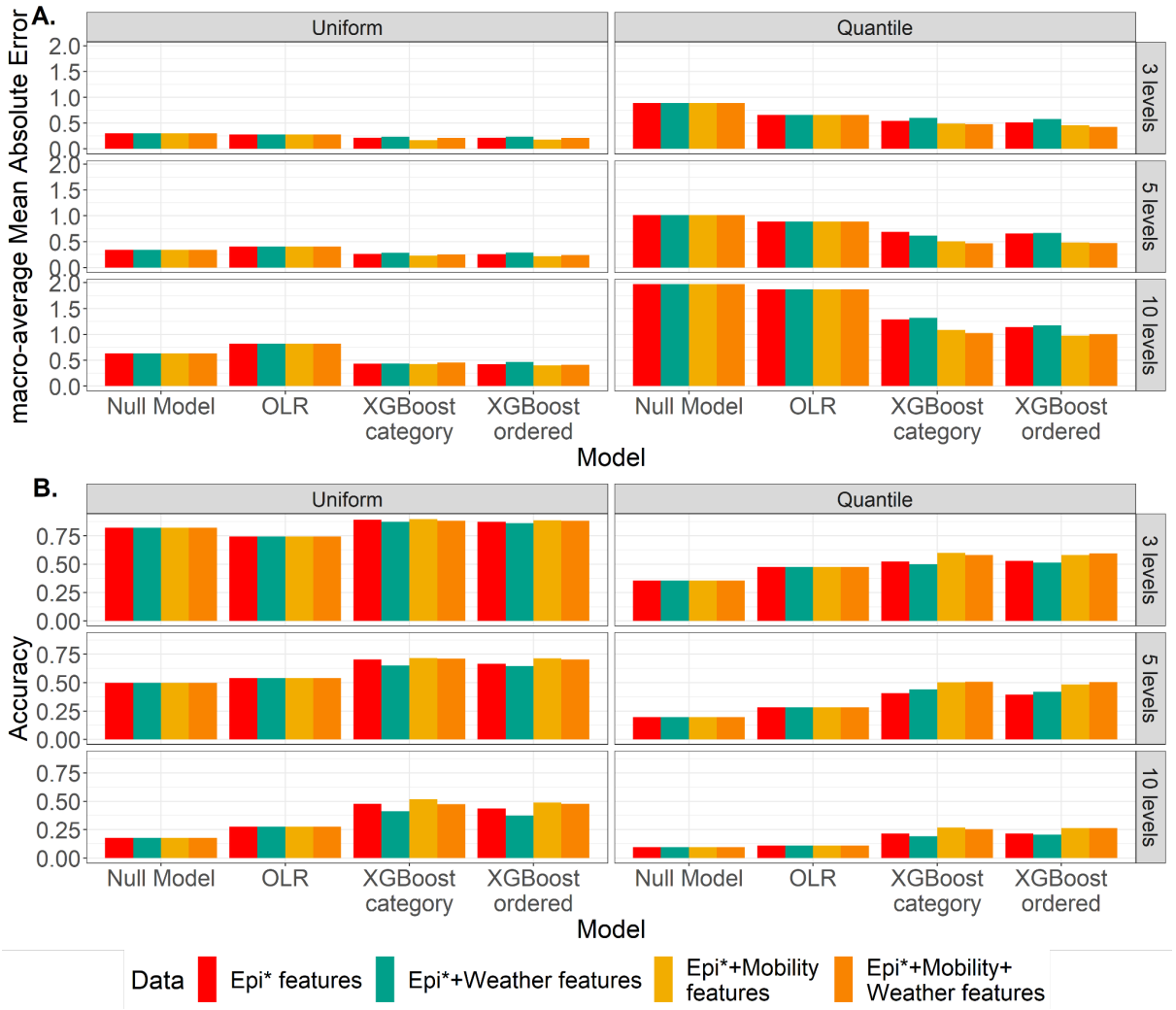


Figure 6.13 Comparison of changes in evaluation metrics when using the n-tile method versus the uniform method.

The left panel is the uniform method while the right panel is the n-tile method. Each row represents the number of bins from 3 to 10. **A**. The comparison of changes in macro-averaged Mean Absolute Error (mMAE). **B**. The comparison of changes in accuracy.

6.4 Discussion

In this chapter, we aim to incorporate different data sources to improve the forecasting accuracy of the regional hospitalisation level in England as an ordinal outcome. The ability of different bin partitioning methods to exploit ordering information is assessed and the n-tile method shows better insights than the uniform method. Across seven NHS regions, we evaluate the performance of short-term forecasts based on: 1) predictions that solely relied on epidemiological data - previous hospitalisations, cases, and deaths; 2) predictions that separately incorporated mobility data and weather as additional predictors; and 3) predictions that involved both mobility and weather as additional predictors. The XGboost ordered model based on epidemiological data achieves higher accuracy than baseline models, which can benefit countries that only have constant access to epidemiological data rather than other data sources. Additionally, we demonstrate that the inclusion of aggregated mobility data in the predictions leads to a further enhancement in the overall prediction accuracy even up to 4-weeks ahead compared to epidemiological data only, while weather conditions show little improvements in accuracy.

A key insight from our study is that mobility data emerged as a stronger predictor of COVID-19 transmission compared to weather conditions. Our results align with prior research that has highlighted the critical role of mobility as a key predictor of COVID-19 transmission dynamics (Badr et al., 2020; Kraemer et al., 2020; Klein et al., 2023). Several other studies have also explored the association between mobility and COVID-19 transmission, but with variable conclusions. For instance, a study conducted by (Xiong et al., 2020) found mobility to be a significant predictor of COVID-19 cases in certain regions, while it had a limited impact on others. (Salje et al., 2020) demonstrated a strong correlation between mobility and COVID-19 transmission across various geographical areas. The disparity in findings among studies could be attributed to variations in study design, geographical contexts, and specific modelling approaches employed. Differences in data collection methods and time periods

considered in each study may also contribute to the observed discrepancies. In our analysis, we believe that the extensive and diverse dataset used, encompassing various mobility patterns and weather conditions, allowed us to discern mobility as the primary predictor. Additionally, the consideration of local factors and spatiotemporal variations in mobility patterns might have contributed to our more accurate predictions.

Change in mobility data likely reflects more general changes in contact behaviour occurring within a population in response to COVID-19, when weather factors may have an interactive effect on these changes that may contribute to the spread of COVID-19 and lead to the increase in hospitalisation. For example, during mild weather conditions, people are more inclined to spend time outdoors in parks and engage in socialising activities that involve close contact with others without adhering to appropriate safety measures such as wearing masks and maintaining physical distancing.

Our study is subject to certain limitations. First, with respect to its generalisability, the robustness of our model forecasts largely relies on the accessibility and representativeness of mobility data. However, it is important to note that Google stopped updating their mobility data from 15 October 2022 onwards, therefore our results may be less generalisable going forwards because fewer mobility datasets are available. Moreover, in resource-constrained settings with limited smartphone usage, the aggregated mobility data may not fully characterise the movement patterns of the entire population. Our results provide additional motivation for the collection and curation of mobility data, which likely has numerous other health applications. For these reasons, we also described here the performance of the XGBoost model of ordinal data using only intrinsic variables in the time series.

We were not able to obtain sufficiently accurate datasets with which to consider the application of non-pharmaceutical interventions (NPIs) and vaccination at the regional level

as additional features in our predictive model. Specifically, the lack of suitable weekly vaccination data by NHS regions hindered our assessment of the potential impact of vaccination coverage rate on the accuracy of our predictions. Furthermore, we did not incorporate the effectiveness of NPIs and behavioural patterns, such as intentions of wearing face masks and following social distancing, as predictors in our model. However, it is well-established that vaccination and non-pharmaceutical interventions have played crucial roles in mitigating the spread of COVID-19 and reducing infection and hospitalisation rates in various regions (Davies et al., 2020; Walker et al., 2020). The rollout of vaccination campaigns and the implementation of NPIs, including lockdowns, travel restrictions, and mask mandates, have demonstrated their efficacy in curbing transmission and minimizing the burden on healthcare systems (Wilder-Smith & Freedman, 2020; Lurie et al., 2020). We suggest that the framework we have proposed here could be used with such data to assess the likely value of those data for predicting key patterns of hospitalisation.

In conclusion, our findings demonstrate that incorporating weather and mobility data to explore the ordering information of ordinal hospitalisation levels can improve the precision of hospital admission level predictions with a 4-week ahead time horizon. This provides policymakers with extra precious time to plan for hospital resource arrangements.

Chapter 7 Discussion

The main motivation of my PhD was to investigate the transmission of influenza epidemics and the COVID-19 pandemic and develop a forecasting framework based on a machine learning algorithm that can be applied to epidemiological data and incorporated with external information flexibly to predict epidemics accurately.

7.1 Summary of Findings

An overview of the general epidemiology of influenza and SARS-CoV-2 was given from multiple perspectives in Chapter 1. Biological background including gene structures, viral replication and immune responses were described and compared for the influenza virus and the SARS-CoV-2 virus, and surveillance systems and diagnostics methods commonly used for them were also discussed. Additionally, epidemiological characteristics such as mortality and morbidity, and disease burdens were described. Chapter 2 mainly focused on infectious disease forecasting. Key concepts involved in infectious disease forecasting and evaluation metrics were introduced. This chapter also reviewed different categories of mathematical, machine learning and deep learning models and their applications in predicting influenza and COVID-19 related outcomes.

In Chapter 3, a machine learning forecasting framework was initiated based on the XGBoost algorithm, to predict the forecast target ordinal influenza-like illness (ILI) incidence level. This framework was applied to surveillance data of 32 countries between 2010 and 2018, and the predictive performance was compared to baseline models, measured by mean-zero error (MZE) and macro-averaged mean absolute error (mMAE). The proposed forecasting framework maintained higher accuracy than baseline models for all prediction horizons from 1 to 4-week ahead. Compared to the baseline model, the proposed XGBoost forecasting

model reduced mMAE by 78% for 1-week ahead predictions and by 62% for 4-week ahead predictions.

In Chapter 4, the investigation focused on the prevalence and transmission trends of SARS-CoV-2 between September 2020 and March 2021, using data on swab-positivity from self-administered throat and nose swabs tested by reverse transcriptase polymerase chain reaction (rt- PCR) in the REal-time Assessment of Community Transmission-1 (REACT-1) study. Prevalence by age group and geographical regions in England and the interaction effects on prevalence patterns between regions and age groups were estimated. Swab-positivity increased after the fall of 2020, reaching a peak in January at 1.57% (1.49%, 1.66%). School-age children (5 - 17 years) and young adults (18 - 24 years) were over-represented among those testing positive, and regional prevalence showed substantial heterogeneity. Interactions between regions and age groups were observed, where the variations in regional prevalence were reflected in the prevalence distributions among age groups within each region.

Chapter 5 extended the scope to assess risk factors influencing SARS-CoV-2 positivity through multivariable logistic regression models. Health inequalities of risk of being positive for SARS-CoV-2 were observed among subgroups of the population in England during the second wave when the risk of hospitalisations and death from infection was substantial. Clear health inequalities with higher risks of infection among Asian and Black ethnic groups compared to white ethnicity were observed, underlining the disproportionate impact of the pandemic on these communities. Furthermore, Bangladeshi and Pakistan participants showed the highest and the second highest risk of swab positivity respectively when compared to British white individuals. The analysis also identified occupational risks, with healthcare and care home workers facing significantly higher odds of infection. In addition, the elevated risk for individuals engaged in public-facing activities or settings highlights the importance of mitigating strategies in public settings.

In Chapter 6, the initial forecasting framework developed in Chapter 3 was further extended to predict regional COVID-19 hospitalisation levels in England for the period from 1 January to 31 December 2022, treating the number of hospital admissions in each region as an ordinal variable. A simple approach (Frank & Hall, 2001) was applied to enable the framework to exploit the ordering information contained in level attributes. In addition to the epidemiological data including the weekly number of COVID-19 cases, hospital admissions and COVID-19 related deaths, mobility data for multiple categories of locations and weather conditions data were incorporated into the forecasting framework. As expected, the results showed that the forecasting accuracy was improved when ordered information was used compared to baseline models without that information. The inclusion of weather data consistently increases the accuracy of our forecasts compared with models that rely only on intrinsic epidemiological data, but only by a small amount. Mobility data brought about a more substantial increase in the accuracy of our forecasts.

7.2 Limitations of the Research and Future Directions

The research presented in this thesis has limitations. First of all, data availability and quality are major concerns, which will limit the predictivity and accuracy of the forecasting model. For example, population mobility data used in Chapter 6 stopped updating on 15th October 2022, leading to partly missingness in training data which may reduce the predictivity of the model. The fluID platform (World Health Organization, n.d.) allows countries or regions to submit surveillance data on influenza by themselves, therefore the quality and coverage of data vary across countries. Countries or regions located in tropical or sub-tropical regions with less advanced economic status are less likely to provide sufficient data that is eligible to be used by the forecasting framework, which impedes our understanding of transmission in these countries, leading to poor predictions in the end. In Chapter 3, the data was reported at the national level. Although the aggregated data offers the benefit of presenting a

comprehensive overview of a country or a part, it may inadvertently overlook heterogeneities across subregions and demographic subgroups of the population.

In the analyses presented in Chapters 4 and 5, the utilisation of a round-based approach, while methodologically sound within the design constraints of the REACT-1 study, may have constrained the robustness of evidence generated in response to specific scientific questions. A notable example is assessing the extent of children's role in COVID-19 transmission. Currently, the study's methodology involves estimating all parameters separately for each round, which, while providing round-specific insights, might limit the capacity to draw more generalized conclusions across different stages of the pandemic. An alternative approach could have involved the construction of a model that estimated a single, cohesive parameter for key effects, utilising data from multiple rounds. This could be achieved, for instance, through the implementation of a Bayesian hierarchical model employing both fixed and random effects. Such a model would potentially allow for a more integrated analysis, capturing the nuances of transmission dynamics over time, and offering a more comprehensive understanding of the spread of the virus, particularly the role played by different age groups, including children.

This approach was considered during the research process. However, after careful deliberation, the decision was made to adhere to the round-based analysis. This choice was influenced by the desire to maintain consistency in the methodological framework and to judge the evidence based on its coherence and alignment with the overall design of the REACT study, essentially following the design rationale for the method used to gather the data.

However, due to the study design of the REACT-1 study, in which were collected rounds with 2-week break between each round, the discontinuity prevented us from using those data for forecasting in Chapter 6. Conversely, data from the Coronavirus (COVID-19)

Infection Survey (Office for National Statistics, 2020b) collected by the Office for National Statistics (ONS) were continuous in time. The Coronavirus (COVID-19) Infection Survey and the REACT-1 studies represent two significant efforts in understanding the spread of infectious diseases, particularly COVID-19, within the UK. Therefore, a possible extension to the work in this would be to devise a study design that would allow us to test the utility of the proposed forecasting framework on REACT-1 data (overcoming its epochal nature), or we could use data from the Coronavirus (COVID-19) Infection Survey to if it could be used to increase predictive accuracy.

Both studies offer valuable insights into estimating the prevalence of COVID-19 in the community, and contribute significantly to public health policy, informing government decisions on lockdowns, social distancing measures, and vaccination strategies. High-quality information sources provided by the Coronavirus (COVID-19) Infection Survey and the REACT-1 study offer the potential to analyse these two studies together for a comprehensive, more nuanced understanding of the virus's transmission and impact, but we need to be careful about their differences in design, methodology, and conclusions. The Coronavirus (COVID-19) Infection Survey is a household survey designed initially to describe transmission dynamics in household and later extended to measure the prevalence of the virus among the community population. It involves repeated cross-sectional and longitudinal sampling, where participants provide swab tests and answer questionnaires (Pouwels et al., 2021). This method allows for a continuous assessment of the infection's prevalence over time and provides insights into the socio-demographic factors affecting transmission (Office for National Statistics, 2023). The REACT-1 study, on the other hand, primarily focuses on the rapid assessment of COVID-19 community transmission. It utilises a large-scale, random sampling of the population, with participants providing self-administered swab tests. The design of the REACT-1 study is more focused on immediate, real-time data collection and analysis, which aids in quick decision-making and public health response. Another main difference between the two studies lies in their sampling methods. The

Coronavirus (COVID-19) Infection Survey uses a relatively stable cohort over time, allowing for longitudinal data analysis and understanding of infection dynamics over a longer period. However, the REACT-1 study employs a cross-sectional approach, focusing on capturing a snapshot of the infection at a particular point in time, which is crucial for understanding immediate transmission rates. In terms of scale, the REACT-1 study typically involved a larger sample size, which enables a broader geographical coverage and thus, a more generalised understanding of the virus's spread across different regions. The Coronavirus (COVID-19) Infection Survey, while often slightly smaller in scale, provides more in-depth data per participant, including socio-economic variables, which are crucial for understanding the impact of the virus on different demographic groups.

Another limitation that needs to be considered is the potential bias caused by imperfect responses, particularly the underrepresentation of minority ethnic groups. Although in the REACT-1 study, the swab samples were sent to individuals randomly across England, the number of responses from minor ethnic backgrounds was much lower than white ethnicity, indicating that minority ethnic groups may be under-represented, which might have led to increased uncertainty of estimated prevalence with wider confidence intervals. Multiple factors contribute to the lower participation of minorities. First, sociocultural backgrounds including language barriers, cultural differences, and the extent of understanding of the purpose of the study can lead to different levels of participation. Historical mistrust in medical research within certain communities and lack of awareness about the study further exacerbate this issue (Scharff et al., 2010; Miller & Miller, 2021).

Second, socioeconomic status may also be an influencing factor, including economic conditions, accessibility of healthcare resources, etc. Moreover, study design and recruitment strategies, such as whether the questionnaire was designed to be sufficiently inclusive of culturally and linguistically diverse groups, can also have an impact. This imbalance in participation across groups may have led to incomplete and skewed data,

which in turn may affect our understanding of overall population characteristics and epidemiological transmission patterns. To address these challenges, future studies should employ targeted outreach strategies, such as using multilingual materials and involving community leaders in the recruitment process.

Furthermore, ensuring ethical and cultural sensitivity in the study's design and implementation is also important. This involves respecting cultural norms and addressing concerns about privacy and data usage, while clearly communicating the study's benefits. Building long-term relationships with minority communities can foster trust and improve participation in future research. Collaborating with multidisciplinary teams, including sociologists and community outreach experts, can help design and conduct studies that are more inclusive and representative of the diverse population. By adopting these strategies, we can aim to achieve higher response rates and fully reflect the characteristics of different races and communities when conducting similar studies, so as to obtain more accurate and reliable findings.

Knowledge of changes in control measures during the COVID-19 pandemic was not captured by the forecasting framework in this thesis, presenting a key area for improvement. During the COVID-19 pandemic, interventions such as phased lockdowns and social distancing rules, as well as vaccination rollout for different groups were implemented in England, which significantly influenced case trends, hospitalisation rates and the pattern of spread of the epidemic. Dynamically incorporating changes of interventions including specific lockdown phases and vaccination coverage in the forecasting framework may have markedly improved its accuracy. In order to more accurately predict and explain epidemic trends, future research could focus on quantitatively assessing the impact of temporal change in interventions and incorporating these factors into the prediction framework. Utilising real-time data to reflect immediate policy changes and analysing the complex interplay of simultaneous interventions could enhance the model's responsiveness and

predictive capability. Creating scenarios based on different policy interventions can also provide valuable foresight for policymakers. While integrating real-time policy changes presents challenges, such as data immediacy and varying compliance levels, its potential benefits in accurately forecasting epidemic trends are substantial, offering critical insights for informed public health strategies.

Another interesting direction worth exploring in future work would be to further develop the forecasting framework of this thesis into an ensemble prediction model based on multiple machine learning (ML) and deep learning (DL) algorithms (Opitz & Maclin, 1999; Latha & Jeeva, 2019). To our knowledge, ensemble modelling has not been applied to predict influenza or COVID-19 related outcomes as an ordinal variable, but it shows significant improvement in prediction accuracy when non-ordinal outcomes are used (Dean et al., 2020; Mahajan et al., 2023; Panja et al., 2023). Merging multiple algorithms into a unified framework that can leverage the strengths of different algorithms while mitigating their weaknesses, can lead to more robust predictions (Opitz & Maclin, 1999) . Furthermore, the ability of DL algorithms to process and interpret large volumes of unstructured data has raised researchers' attention. They have been employed in epidemiological forecasting and have shown some promising results in predicting disease outbreaks and transmission dynamics (Wu, 2018; Yang et al., 2023; Majeed et al., 2023). We would therefore potentially expect improvements ensemble modelling with ML and DL to be applied to these forecast targets.

It is useful to compare ML approaches and Bayesian methods in epidemiological forecasting. The thesis primarily focused on applying ML algorithms to infectious disease forecasting, which, while effective in certain aspects, present a different set of strengths and weaknesses compared to Bayesian methods. ML algorithms are renowned for their proficiency in processing extensive datasets efficiently and uncovering intricate, non-linear patterns in data (LeCun, Bengio & Hinton, 2015). These algorithms are highly adaptable and

can improve their performance iteratively as more data becomes available. Nevertheless, they often require large amounts of training data and are prone to overfitting, which can ultimately diminish their predictive accuracy. They might also lack transparency in their assumptions and decision-making processes, making it challenging to interpret the results (Lo Piano, 2020). On the other hand, Bayesian methods offer a probabilistic approach to forecasting, formally integrating prior knowledge with observed data. This framework is particularly advantageous in situations where data is sparse or uncertain, as it allows for the incorporation of expert knowledge and previous research findings into the model (Gelman, 2021). Bayesian methods are inherently designed to handle uncertainty and provide estimates of confidence, which is vital in epidemiological forecasting. However, these methods can be computationally intensive and may require careful calibration of prior distributions to avoid introducing biases (Gelman, 2021).

In this thesis, the absence of a detailed comparison between these two methodologies signifies a missed opportunity to explore how Bayesian methods might complement or enhance the ML approaches used. For instance, Bayesian methods could potentially be integrated into ensemble models to improve the handling of uncertainty and provide a more comprehensive understanding of epidemiological trends. Future research should aim to incorporate and compare both ML and Bayesian methodologies in epidemiological forecasting. Such an integrative approach could leverage the strengths of both methods, utilising the predictive power and adaptability of ML models while incorporating the probabilistic and uncertainty-handling capabilities of Bayesian methods. This combination could provide a more robust and nuanced forecasting framework, especially valuable in the rapidly evolving field of epidemiological research.

In addition, techniques in data pre-processing, feature selection and hyperparameter tuning need to be considered to ensure that the ensemble model maintains good performance in different scenarios. By introducing diversity and complexity, future predictive models may be

better able to cope with the variability and uncertainty of outbreak data and provide more reliable support and guidance for public health decision-making.

In the current phase of circulation of SARS-CoV-2 virus, far-referred resources are available for community surveillance. Hence the REACT and the Coronavirus (COVID-19) Infection Survey studies are no longer active. However, the precision of the epidemiological results presented in this thesis suggests there could be high public health value in governments considering similar but less expensive options. In particular, to tease out the degree to which health inequality is driven by rates of infection rather than poor treatment of disease, data such as those presented here are essential.

7.3 Implications of the Research

The analysis and results presented in this thesis focus on community surveillance of SARS-CoV-2 transmission in England and the development of a high-precision forecasting framework applicable to different infectious diseases, which provides a new direction and paradigm for infectious disease surveillance, response, and forecasting.

The epidemiology of influenza and SARS-CoV-2, while both concerning respiratory viruses, presents distinct challenges in terms of surveillance methods and data sources. Influenza surveillance has traditionally relied on established syndromic surveillance systems and laboratory confirmations, which focus on tracking seasonal patterns and identifying strains to inform vaccine development (Torner et al., 2019; Dong et al., 2017). In contrast, the emergence of the SARS-CoV-2 required the rapid development of new surveillance strategies. The widespread use of PCR testing for the SARS-CoV-2 virus, more extensive than is seen in influenza monitoring, provided an unprecedented level of detail about the virus transmission, including the identification of asymptomatic cases (Aitken et al., 2020; Barbian et al., 2024; Love et al., 2021).

This thesis highlights the need for adaptable and responsive surveillance systems in certain circumstances. The REACT-1 study provides a compelling example of community surveillance during the COVID-19 pandemic. It demonstrates how real-time, widespread data collection can significantly improve our understanding and management of a novel virus, providing a strong reference for pandemic that will occur in the future. One of the key lessons from the COVID-19 pandemic is the flexibility and scalability of surveillance systems. The rapid global spread of SARS-CoV-2 underscored the need for surveillance strategies that can be quickly adapted and scaled up in response to emerging threats. Digital tools for contact tracing and data sharing, as well as innovative approaches such as wastewater surveillance, have shown great potential in this regard (Budd et al., 2020; Burdinski, Brockmann & Maier, 2022). Additionally, the rapid spread of the virus across continents indicates global collaboration in data sharing and analysis is essential for enhancing future surveillance of respiratory viruses.

Apart from the technical aspects of data collection and analysis, effective communication of research findings to the public needs to be considered by those designing pandemic response plans. Transparent and comprehensible dissemination of information significantly enhances the public's understanding and interpretation of the ongoing situation (Varghese et al., 2021b). This clarity is essential not only for encouraging adherence to control measures but also for fostering public trust and cooperation. An informed public, cognizant of the rationale behind specific health directives and the critical role of their compliance, is more likely to participate actively in research surveys (Berg et al., 2021). Such participation is crucial in garnering comprehensive and representative epidemiological data, enabling us to gain a more comprehensive understanding of the dynamics of epidemics at the community level and more accurately predict and assess the spread of infectious diseases, which in turn provided strong support for decision-making and resource allocation.

Overall, the takeaway from this thesis in terms of the REACT-1 study is that significant inequalities were found among infection in subgroups of the population such as ethnicity and occupation providing valuable insights into the impact of socio-economic factors, population density, and healthcare resources on the spread of epidemics, indicating that the continued surveillance for the transmission of SARS-CoV-2 in population subgroups can help identify suitable interventions to reduce these differences related to the risk of infections. With the experience and knowledge gained in the REACT-1 study, we can more effectively address public health challenges when epidemics happen in the future.

The forecasting framework developed in this thesis, specifically in Chapters 3 and 6, has demonstrated its potential to be applied to other infectious disease outbreaks and epidemics. The outcome was predicted as an ordinal variable, providing a more straightforward method to capture epidemiological trends and to assess the insensitivity of epidemics. Furthermore, this thesis highlights the importance of infectious disease surveillance and prediction in public health strategies. Our findings suggest that active surveillance and accurate prediction of infectious disease transmission trends can play a key role in early risk identification, resource deployment, and the implementation of targeted interventions. We therefore call for the integration of infectious disease surveillance and forecasting into broader public health strategies to respond more efficiently to public health emergencies and to safeguard the health and safety of society. This study provides a useful reference for further research in the field of public health and forecasting to promote better epidemic prevention and health management practices.

Reference

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H. & Liu, B. (2011) Predicting Flu Trends using Twitter data. In: *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. April 2011 pp. 702–707. doi:10.1109/INFCOMW.2011.5928903.

Adam, D.C., Wu, P., Wong, J.Y., Lau, E.H.Y., Tsang, T.K., Cauchemez, S., Leung, G.M. & Cowling, B.J. (2020) Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nature medicine*. 26 (11), 1714–1719. doi:10.1038/s41591-020-1092-0.

Afriyie, D.K., Asare, G.A., Amponsah, S.K. & Godman, B. (2020) COVID-19 pandemic in resource-poor countries: challenges, experiences and opportunities in Ghana. *Journal of infection in developing countries*. 14 (8), 838–843. doi:10.3855/jidc.12909.

Ahmad, K., Erqou, S., Shah, N., Nazir, U., Morrison, A.R., Choudhary, G. & Wu, W.-C. (2020) Association of poor housing conditions with COVID-19 incidence and mortality across US counties. *PloS one*. 15 (11), e0241327. doi:10.1371/journal.pone.0241327.

Aiello, A.E., Renson, A. & Zivich, P.N. (2020) Social Media- and Internet-Based Disease Surveillance for Public Health. *Annual review of public health*. 41, 101–118. doi:10.1146/annurev-publhealth-040119-094402.

Aiken, E.L., Nguyen, A.T., Viboud, C. & Santillana, M. (2021) Toward the use of neural networks for influenza prediction at multiple spatial resolutions. *Science advances*. 7 (25). doi:10.1126/sciadv.abb1237.

Ainslie, K.E.C., Walters, C.E., Fu, H., Bhatia, S., Wang, H., et al. (2020) Evidence of initial success for China exiting COVID-19 social distancing policy after achieving containment. *Wellcome open research*. 5, 81. doi:10.12688/wellcomeopenres.15843.2.

Aitken, J., Ambrose, K., Barrell, S., Beale, R., Bineva-Todd, G., et al. (2020) Scalable and robust SARS-CoV-2 testing in an academic center. *Nature biotechnology*. 38 (8), 927–931. doi:10.1038/s41587-020-0588-y.

Alabool, H., Alarabiat, D., Abualigah, L., Habib, M., Khasawneh, A.M., Alshinwan, M. & Shehab, M. (2020) *Artificial intelligence techniques for Containment COVID-19 Pandemic: A Systematic Review*. doi:10.21203/rs.3.rs-30432/v1.

Albahri, O.S., Zaidan, A.A., Albahri, A.S., Zaidan, B.B., Abdulkareem, K.H., Al-Qaysi, Z.T., Alamoodi, A.H., Aleesa, A.M., Chyad, M.A., Alesa, R.M., Kem, L.C., Lakulu, M.M., Ibrahim, A.B. & Rashid, N.A. (2020) Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects. *Journal of infection and public health*. 13 (10), 1381–1396. doi:10.1016/j.jiph.2020.06.028.

Alballa, N. & Al-Turaiki, I. (2021) Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in medicine unlocked*. 24, 100564. doi:10.1016/j.imu.2021.100564.

d'Albis, H., Coulibaly, D., Roumagnac, A., de Carvalho Filho, E. & Bertrand, R. (2021) Quantification of the effects of climatic conditions on French hospital admissions and deaths induced by SARS-CoV-2. *Scientific reports*. 11 (1), 21812. doi:10.1038/s41598-021-01392-2.

Aleta, A., Martín-Corral, D., Pastore Y Piontti, A., Ajelli, M., Litvinova, M., Chinazzi, M., Dean, N.E., Halloran, M.E., Longini, I.M., Jr, Merler, S., Pentland, A., Vespignani, A., Moro, E. & Moreno, Y. (2020) Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nature human behaviour*. 4 (9), 964–971. doi:10.1038/s41562-020-0931-9.

Ali, S.T. & Cowling, B.J. (2021) Influenza Virus: Tracking, Predicting, and Forecasting. *Annual review of public health*. 42, 43–57. doi:10.1146/annurev-publhealth-010720-021049.

Al Khatib, H.A., Al Thani, A.A., Gallouzi, I. & Yassine, H.M. (2019) Epidemiological and genetic characterization of pH1N1 and H3N2 influenza viruses circulated in MENA region during 2009-2017. *BMC infectious diseases*. 19 (1), 314. doi:10.1186/s12879-019-3930-6.

Al-Omari, A., Rabaan, A.A., Salih, S., Al-Tawfiq, J.A. & Memish, Z.A. (2019) MERS coronavirus outbreak: Implications for emerging viral infections. *Diagnostic microbiology and infectious disease*. 93 (3), 265–285. doi:10.1016/j.diagmicrobio.2018.10.011.

Al-Qaness, M.A.A., Ewees, A.A., Fan, H. & Abd El Aziz, M. (2020a) Optimization Method for Forecasting Confirmed Cases of COVID-19 in China. *Journal of clinical medicine research*. 9 (3). doi:10.3390/jcm9030674.

Al-Qaness, M.A.A., Ewees, A.A., Fan, H., Abualigah, L. & Abd Elaziz, M. (2020b) Marine Predators Algorithm for Forecasting Confirmed Cases of COVID-19 in Italy, USA, Iran and Korea. *International journal of environmental research and public health*. 17 (10). doi:10.3390/ijerph17103520.

Anderson, R.M., Heesterbeek, H., Klinkenberg, D. & Déirdre Hollingsworth, T. (2020) How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet*. 395 (10228), 931–934. doi:10.1016/S0140-6736(20)30567-5.

Angers, J.-F., Biswas, A. & Maiti, R. (2017) Bayesian forecasting for time series of categorical data. *Journal of forecasting*. 36 (3), 217–229. doi:10.1002/for.2426.

Apicella, M., Campopiano, M.C., Mantuano, M., Mazoni, L., Coppelli, A. & Del Prato, S. (2020) COVID-19 in people with diabetes: understanding the reasons for worse outcomes. *The lancet. Diabetes & endocrinology*. 8 (9), 782–792. doi:10.1016/S2213-8587(20)30238-2.

Ardabili, S.F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A.R., Reuter, U., Rabczuk, T. & Atkinson, P.M. (2020) COVID-19 Outbreak Prediction with Machine Learning. *Algorithms*. 13 (10), 249. doi:10.3390/a13100249.

Arduin, H., Domenech de Cellès, M., Guillemot, D., Watier, L. & Opatowski, L. (2017) An agent-based model simulation of influenza interactions at the host level: insight into the influenza-related burden of pneumococcal infections. *BMC infectious diseases*. 17 (1), 382. doi:10.1186/s12879-017-2464-z.

Armstrong, B.G. & Sloan, M. (1989) Ordinal regression models for epidemiologic data. *American journal of epidemiology*. 129 (1), 191–204. doi:10.1093/oxfordjournals.aje.a115109.

Arya, R., Kumari, S., Pandey, B., Mistry, H., Bihani, S.C., Das, A., Prashar, V., Gupta, G.D., Panicker, L. & Kumar, M. (2021) Structural insights into SARS-CoV-2 proteins. *Journal of molecular biology*. 433 (2), 166725. doi:10.1016/j.jmb.2020.11.024.

Athanasiou, M., Fragkozidis, G., Zarkogianni, K. & Nikita, K.S. (2023) Long Short-term Memory-Based Prediction of the Spread of Influenza-Like Illness Leveraging Surveillance,

Weather, and Twitter Data: Model Development and Validation. *Journal of medical Internet research*. 25, e42519. doi:10.2196/42519.

Axelsen, J.B., Yaari, R., Grenfell, B.T. & Stone, L. (2014) Multiannual forecasting of seasonal influenza dynamics reveals climatic and evolutionary drivers. *Proceedings of the National Academy of Sciences of the United States of America*. 111 (26), 9538–9542. doi:10.1073/pnas.1321656111.

Ayyoubzadeh, S.M., Ayyoubzadeh, S.M., Zahedi, H., Ahmadi, M. & R Niakan Kalhori, S. (2020) Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR public health and surveillance*. 6 (2), e18828. doi:10.2196/18828.

Azziz Baumgartner, E., Dao, C.N., Nasreen, S., Bhuiyan, M.U., Mah-E-Muneer, S., Al Mamun, A., Sharker, M.A.Y., Zaman, R.U., Cheng, P.-Y., Klimov, A.I., Widdowson, M.-A., Uyeki, T.M., Luby, S.P., Mounts, A. & Bresee, J. (2012) Seasonality, timing, and climate drivers of influenza activity worldwide. *The Journal of infectious diseases*. 206 (6), 838–846. doi:10.1093/infdis/jis467.

Babii, A., Ghysels, E. & Striaukas, J. (2022) Machine Learning Time Series Regressions With an Application to Nowcasting. *Journal of business & economic statistics: a publication of the American Statistical Association*. 40 (3), 1094–1106. doi:10.1080/07350015.2021.1899933.

Baccianella, S., Esuli, A. & Sebastiani, F. (2009) Evaluation Measures for Ordinal Regression. In: *2009 Ninth International Conference on Intelligent Systems Design and Applications*. November 2009 pp. 283–287. doi:10.1109/ISDA.2009.230.

Badr, H.S., Du, H., Marshall, M., Dong, E., Squire, M.M. & Gardner, L.M. (2020) Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet infectious diseases*. 20 (11), 1247–1254. doi:10.1016/S1473-3099(20)30553-3.

Balaman, Ş.Y. (2019) Chapter 5 - Uncertainty Issues in Biomass-Based Production Chains. In: Ş.Y. Balaman (ed.). *Decision-Making for Biomass-Based Production Chains*. Academic Press. pp. 113–142. doi:10.1016/B978-0-12-814278-3.00005-4.

Balogh, A., Harman, A. & Kreuter, F. (2022) Real-Time Analysis of Predictors of COVID-19 Infection Spread in Countries in the European Union Through a New Tool. *International journal of public health*. 67, 1604974. doi:10.3389/ijph.2022.1604974.

Baqui, P., Bica, I., Marra, V., Ercole, A. & van der Schaar, M. (2020) Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study. *The Lancet Global Health*. 8 (8), e1018–e1026. doi:10.1016/S2214-109X(20)30285-0.

Barbian, H.J., Kittner, A., Teran, R., Bobrovska, S., Qiu, X., English, K., Green, S.J., Ghinai, I., Pacilli, M. & Hayden, M.K. (2024) A response playbook for early detection and population surveillance of new SARS-CoV-2 variants in a regional public health laboratory. *BMC public health*. 24 (1), 59. doi:10.1186/s12889-023-17536-0.

Bartolomei, S.M. & Sweet, A.L. (1989) A note on a comparison of exponential smoothing methods for forecasting seasonal series. *International journal of forecasting*. 5 (1), 111–116. https://ideas.repec.org/a/eee/intfor/v5y1989i1p111-116.html.

Beckman, M.F., Mougeot, F.B. & Mougeot, J.-L.C. (2021) Comorbidities and Susceptibility to COVID-19: A Generalized Gene Set Data Mining Approach. *Journal of clinical medicine*

research. 10 (8). doi:10.3390/jcm10081666.

Bedford, T., Riley, S., Barr, I.G., Broor, S., Chadha, M., et al. (2015) Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*. 523 (7559), 217–220. doi:10.1038/nature14460.

Berg, S.H., O'Hara, J.K., Shortt, M.T., Thune, H., Brønnick, K.K., Lungu, D.A., Røislien, J. & Wiig, S. (2021) Health authorities' health risk communication with the public during pandemics: a rapid scoping review. *BMC public health*. 21 (1), 1401. doi:10.1186/s12889-021-11468-3.

Bhala, N., Curry, G., Martineau, A.R., Agyemang, C. & Bhopal, R. (2020) Sharpening the global focus on ethnicity and race in the time of COVID-19. *The Lancet*. 395 (10238), 1673–1676. doi:10.1016/S0140-6736(20)31102-8.

Bhatraju, P.K., Ghassemieh, B.J., Nichols, M., Kim, R., Jerome, K.R., Nalla, A.K., Greninger, A.L., Pipavath, S., Wurfel, M.M., Evans, L., Kritek, P.A., West, T.E., Luks, A., Gerbino, A., Dale, C.R., Goldman, J.D., O'Mahony, S. & Mikacenic, C. (2020) Covid-19 in Critically Ill Patients in the Seattle Region - Case Series. *The New England journal of medicine*. 382 (21), 2012–2022. doi:10.1056/NEJMoa2004500.

Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I.C.-H., Hickmann, K.S., Lewis, B., Rosenfeld, R., Shaman, J., Tsou, M.-H., Velardi, P., Vespignani, A., Finelli, L. & Influenza Forecasting Contest Working Group (2016) Results from the centers for disease control and prevention's predict the 2013-2014 Influenza Season Challenge. *BMC infectious diseases*. 16, 357. doi:10.1186/s12879-016-1669-x.

Biggerstaff, M., Johansson, M., Alper, D., Brooks, L.C., Chakraborty, P., Farrow, D.C., Hyun, S., Kandula, S., McGowan, C., Ramakrishnan, N., Rosenfeld, R., Shaman, J., Tibshirani, R., Tibshirani, R.J., Vespignani, A., Yang, W., Zhang, Q. & Reed, C. (2018) Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics*. 24, 26–33. doi:10.1016/j.epidem.2018.02.003.

Biobank, U.K. (2020) *COVID-19 hub*. 2020. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/covid-19-hub [Accessed: 16 February 2022].

Birrell, P.J., Ketsetzis, G., Gay, N.J., Cooper, B.S., Presanis, A.M., Harris, R.J., Charlett, A., Zhang, X.-S., White, P.J., Pebody, R.G. & Angelis, D.D. (2011) Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *Proceedings of the National Academy of Sciences*. 108 (45), 18238–18243. doi:10.1073/pnas.1103002108.

Blazquez, D. & Domenech, J. (2018) Big Data sources and methods for social and economic analyses. *Technological forecasting and social change*. 130, 99–113. doi:10.1016/j.techfore.2017.07.027.

Bloom-Feshbach, K., Alonso, W.J., Charu, V., Tamerius, J., Simonsen, L., Miller, M.A. & Viboud, C. (2013) Latitudinal variations in seasonal activity of influenza and respiratory syncytial virus (RSV): a global comparative review. *PloS one*. 8 (2), e54445. doi:10.1371/journal.pone.0054445.

Blundell, R., Costa Dias, M., Joyce, R. & Xu, X. (2020) COVID-19 and Inequalities. *Fiscal studies*. 41 (2), 291–319. doi:10.1111/1475-5890.12232.

Boddington, N.L., Pearson, I., Whitaker, H., Mangtani, P. & Pebody, R.G. (2021) Effectiveness of Influenza Vaccination in Preventing Hospitalization Due to Influenza in Children: A Systematic Review and Meta-analysis. *Clinical infectious diseases: an official*

publication of the Infectious Diseases Society of America. 73 (9), 1722–1732. doi:10.1093/cid/ciab270.

Bourouiba, L. (2020) Turbulent Gas Clouds and Respiratory Pathogen Emissions: Potential Implications for Reducing Transmission of COVID-19. *JAMA: the journal of the American Medical Association*. 323 (18), 1837–1838. doi:10.1001/jama.2020.4756.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C. & Ljung, G.M. (2015) *Time Series Analysis: Forecasting and Control, 5th Edition | Wiley*. https://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021.

Bracher, J., Ray, E.L., Gneiting, T. & Reich, N.G. (2021) Evaluating epidemic forecasts in an interval format. *PLoS computational biology*. 17 (2), e1008618. doi:10.1371/journal.pcbi.1008618.

Brauer, F., Castillo-Chavez, C. & Feng, Z. (n.d.) *Mathematical Models in Epidemiology*. Springer New York. doi:10.1007/978-1-4939-9828-9.

Bridgen, J.R., Jewell, C. & Read, J.M. (2022) Social mixing patterns in the UK following the relaxation of COVID-19 pandemic restrictions, July-August 2020: a cross-sectional online survey. *BMJ open*. 12 (12), e059231. doi:10.1136/bmjopen-2021-059231.

Brooks, L.C., Farrow, D.C., Hyun, S., Tibshirani, R.J. & Rosenfeld, R. (2018) Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS computational biology*. 14 (6), e1006134. doi:10.1371/journal.pcbi.1006134.

Budd, J., Miller, B.S., Manning, E.M., Lampos, V., Zhuang, M., Edelstein, M., Rees, G., Emery, V.C., Stevens, M.M., Keegan, N., Short, M.J., Pillay, D., Manley, E., Cox, I.J., Heymann, D., Johnson, A.M. & McKendry, R.A. (2020) Digital technologies in the public-health response to COVID-19. *Nature medicine*. 26 (8), 1183–1192. doi:10.1038/s41591-020-1011-4.

Buehler, J.W., Sonricker, A., Paladini, M., Soper, P. & Mostashari, F. (2007) *Syndromic surveillance practice in the United States: Findings from a survey of state, territorial, and selected local health departments*. 6 (3). http://dx.doi.org/.

Burdinski, A., Brockmann, D. & Maier, B.F. (2022) Understanding the impact of digital contact tracing during the COVID-19 pandemic. *PLOS digital health*. 1 (12), e0000149. doi:10.1371/journal.pdig.0000149.

Buri, M., Curt, A., Steeves, J. & Hothorn, T. (2020) Baseline-adjusted proportional odds models for the quantification of treatment effects in trials with ordinal sum score outcomes. *BMC medical research methodology*. 20 (1), 104. doi:10.1186/s12874-020-00984-2.

Bzdok, D. (2017) Classical Statistics and Statistical Learning in Imaging Neuroscience. *Frontiers in neuroscience*. 11, 543. doi:10.3389/fnins.2017.00543.

Bzdok, D., Altman, N. & Krzywinski, M. (2018) Statistics versus machine learning. *Nature methods*. 15 (4), 233–234. doi:10.1038/nmeth.4642.

Bzdok, D., Krzywinski, M. & Altman, N. (2017) Points of Significance: Machine learning: a primer. *Nature methods*. 14 (12), 1119–1120. doi:10.1038/nmeth.4526.

Caini, S., Huang, Q.S., Ciblak, M.A., Kusznierz, G., Owen, R., et al. (2015) Epidemiological and virological characteristics of influenza B: results of the Global Influenza B Study.

Influenza and other respiratory viruses. 9 Suppl 1 (Suppl 1), 3–12. doi:10.1111/irv.12319.

Callen, J.L., Kwan, C.C.Y., Yip, P.C.Y. & Yuan, Y. (1996) Neural network forecasting of quarterly accounting earnings. *International journal of forecasting*. 12 (4), 475–482. doi:10.1016/S0169-2070(96)00706-6.

Castro, B.M., Reis, M. de M. & Salles, R.M. (2022) Multi-agent simulation model updating and forecasting for the evaluation of COVID-19 transmission. *Scientific reports*. 12 (1), 1–18. doi:10.1038/s41598-022-22945-z.

Cauchemez, S., Van Kerkhove, M.D., Archer, B.N., Cetron, M., Cowling, B.J., et al. (2014) School closures during the 2009 influenza pandemic: national and local experiences. *BMC infectious diseases*. 14, 207. doi:10.1186/1471-2334-14-207.

CDC (2021) *Coronavirus Disease 2019 (COVID-19)*. 2 September 2021. https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/transmission_k_12_schools.html [Accessed: 11 September 2021].

CDC (2023) *FluSight: Flu Forecasting*. 6 January 2023. Centers for Disease Control and Prevention. https://www.cdc.gov/flu/weekly/flusight/index.html [Accessed: 16 February 2023].

CDC Epidemic Prediction Initiative (2021) *CDC FluSight Hospitalization Forecasts*. Github. https://github.com/cdcepi/Flusight-forecast-data.

CDC Epidemic Prediction Initiative (2016) *FluSight-forecasts: FluSight Forecast Submissions*. Github. https://github.com/cdcepi/FluSight-forecasts.

Centers for Disease Control and Prevention (2020a) *CDC 2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel*. https://www.fda.gov/media/134922/download.

Centers for Disease Control and Prevention (2020b) *COVID Data Tracker*. 2020. https://covid.cdc.gov/covid-data-tracker/#datatracker-home [Accessed: 23 July 2023].

Centers for Disease Control and Prevention (2023) *Risk Assessment Summary for SARS CoV-2 Sublineage BA.2.86*. 24 August 2023. https://www.cdc.gov/respiratory-viruses/whats-new/covid-19-variant.html [Accessed: 1 September 2023].

Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD) (2022) *U.S. Influenza Surveillance: Purpose and Methods*. 14 October 2022. https://www.cdc.gov/flu/weekly/overview.htm [Accessed: 27 February 2023].

Chadeau-Hyam, M., Eales, O., Bodinier, B., Wang, H., Haw, D., et al. (2022a) Breakthrough SARS-CoV-2 infections in double and triple vaccinated adults and single dose vaccine effectiveness among children in Autumn 2021 in England: REACT-1 study. *eClinicalMedicine*. 48. doi:10.1016/j.eclinm.2022.101419.

Chadeau-Hyam, M., Tang, D., Eales, O., Bodinier, B., Wang, H., et al. (2022b) The Omicron SARS-CoV-2 epidemic in England during February 2022. *bioRxiv*. doi:10.1101/2022.03.10.22272177.

Chadeau-Hyam, M., Wang, H., Eales, O., Haw, D., Bodinier, B., et al. (2022c) SARS-CoV-2 infection and vaccine effectiveness in England (REACT-1): a series of cross-sectional random community surveys. *The Lancet. Respiratory medicine*. 10 (4), 355–366. doi:10.1016/S2213-2600(21)00542-7.

Chang, H.-J., Chuang, J.-H., Chern, T.-C., Stein, M., Coker, R., Wang, D.-W. & Hsu, T.-S. (2014) A comparison between a deterministic, compartmental model and an individual based-stochastic model for simulating the transmission dynamics of pandemic influenza. In: *Proceedings of the 4th International Conference on Simulation and Modeling Methodologies, Technologies and Applications*. 2014 SCITEPRESS - Science and Technology Publications. doi:10.5220/0005040905860594.

Chang, S.L., Harding, N., Zachreson, C., Cliff, O.M. & Prokopenko, M. (2020) Modelling transmission and control of the COVID-19 pandemic in Australia. *Nature communications*. 11 (1), 1–13. doi:10.1038/s41467-020-19393-6.

Chatfield, C. (2013) *The Analysis of Time Series : an Introduction, Sixth Edition*. Hoboken, Taylor and Francis. https://www.worldcat.org/title/analysis-of-time-series-an-introduction-sixth-edition/oclc/880737910.

Chen, C.W.S. & Chiu, L.M. (2021) Ordinal Time Series Forecasting of the Air Quality Index. *Entropy* . 23 (9). doi:10.3390/e23091167.

Cheng, C., Jiang, W.-M., Fan, B., Cheng, Y.-C., Hsu, Y.-T., Wu, H.-Y., Chang, H.-H. & Tsou, H.-H. (2023) Real-time forecasting of COVID-19 spread according to protective behavior and vaccination: autoregressive integrated moving average models. *BMC public health*. 23 (1), 1500. doi:10.1186/s12889-023-16419-8.

Cheng, H.-Y., Wu, Y.-C., Lin, M.-H., Liu, Y.-L., Tsai, Y.-Y., Wu, J.-H., Pan, K.-H., Ke, C.-J., Chen, C.-M., Liu, D.-P., Lin, I.-F. & Chuang, J.-H. (2020) Applying Machine Learning Models with An Ensemble Approach for Accurate Real-Time Influenza Forecasting in Taiwan: Development and Validation Study. *Journal of medical Internet research*. 22 (8), e15394. doi:10.2196/15394.

Chen, S., Xu, J., Wu, Y., Wang, X., Fang, S., Cheng, J., Ma, H., Zhang, R., Liu, Y., Zhang, L., Zhang, X., Chen, L. & Liu, X. (2019) Predicting temporal propagation of seasonal influenza using improved gaussian process model. *Journal of biomedical informatics*. 93, 103144. doi:10.1016/j.jbi.2019.103144.

Chen, T. & Guestrin, C. (2016a) XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. 13 August 2016 New York, NY, USA, Association for Computing Machinery. pp. 785–794. doi:10.1145/2939672.2939785.

Chen, T. & Guestrin, C. (2016b) XGBoost: A Scalable Tree Boosting System. *arXiv [cs.LG]*. http://arxiv.org/abs/1603.02754.

Chen, X. & Ishwaran, H. (2012) Random forests for genomic data analysis. *Genomics*. 99 (6), 323–329. doi:10.1016/j.ygeno.2012.04.003.

Chen, Y., Chu, C.W., Chen, M.I.C. & Cook, A.R. (2018) The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison. *Journal of biomedical informatics*. 81, 16–30. doi:10.1016/j.jbi.2018.02.014.

Choi, S.B., Kim, J. & Ahn, I. (2019) Forecasting type-specific seasonal influenza after 26 weeks in the United States using influenza activities in other countries. *PloS one*. 14 (11), e0220423. doi:10.1371/journal.pone.0220423.

Chowell, G., Dahal, S., Tariq, A., Roosa, K., Hyman, J.M. & Luo, R. (2022) An ensemble n-sub-epidemic modeling framework for short-term forecasting epidemic trajectories: Application to the COVID-19 pandemic in the USA. *PLoS computational biology*. 18 (10),

e1010602. doi:10.1371/journal.pcbi.1010602.

Chowkwanyun, M. & Reed, A.L., Jr (2020) Racial Health Disparities and Covid-19 - Caution and Context. *The New England journal of medicine*. 383 (3), 201–203. doi:10.1056/NEJMp2012910.

Chretien, J.-P., George, D., Shaman, J., Chitale, R.A. & McKenzie, F.E. (2014) Influenza forecasting in human populations: a scoping review. *PloS one*. 9 (4), e94130. doi:10.1371/journal.pone.0094130.

Chretien, J.-P., Riley, S. & George, D.B. (2015) Mathematical modeling of the West Africa Ebola epidemic. *eLife*. 4. doi:10.7554/eLife.09186.

Christofoletti, M., Benedetti, T.R.B., Mendes, F.G. & Carvalho, H.M. (2021) Using Multilevel Regression and Poststratification to Estimate Physical Activity Levels from Health Surveys. *International journal of environmental research and public health*. 18 (14). doi:10.3390/ijerph18147477.

Chutinimitkul, S., Herfst, S., Steel, J., Lowen, A.C., Ye, J., et al. (2010) Virulence-associated substitution D222G in the hemagglutinin of 2009 pandemic influenza A(H1N1) virus affects receptor binding. *Journal of virology*. 84 (22), 11802–11813. doi:10.1128/JVI.01136-10.

Cocucci, T.J., Pulido, M., Aparicio, J.P., Ruíz, J., Simoy, M.I. & Rosa, S. (2022) Inference in epidemiological agent-based models using ensemble-based data assimilation. *PloS one*. 17 (3), e0264892. doi:10.1371/journal.pone.0264892.

Conly, J., Seto, W.H., Pittet, D., Holmes, A., Chu, M., et al. (2020) Use of medical face masks versus particulate respirators as a component of personal protective equipment for health care workers in the context of the COVID-19 pandemic. *Antimicrobial resistance and infection control*. 9 (1), 126. doi:10.1186/s13756-020-00779-6.

Cooley, P., Brown, S., Cajka, J., Chasteen, B., Ganapathi, L., Grefenstette, J., Hollingsworth, C.R., Lee, B.Y., Levine, B., Wheaton, W.D. & Wagener, D.K. (2011) The role of subway travel in an influenza epidemic: a New York City simulation. *Journal of urban health: bulletin of the New York Academy of Medicine*. 88 (5), 982–995. doi:10.1007/s11524-011-9603-4.

Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., et al. (2020) Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*. 25 (3). doi:10.2807/1560-7917.ES.2020.25.3.2000045.

de Courville, C., Cadarette, S.M., Wissinger, E. & Alvarez, F.P. (2022) The economic burden of influenza among adults aged 18 to 64: A systematic literature review. *Influenza and other respiratory viruses*. 16 (3), 376–385. doi:10.1111/irv.12963.

Cummings, M.J., Baldwin, M.R., Abrams, D., Jacobson, S.D., Meyer, B.J., Balough, E.M., Aaron, J.G., Claassen, J., Rabbani, L.E., Hastie, J., Hochman, B.R., Salazar-Schicchi, J., Yip, N.H., Brodie, D. & O'Donnell, M.R. (2020) Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. *The Lancet*. 395 (10239), 1763–1770. doi:10.1016/S0140-6736(20)31189-2.

Dai, S. & Han, L. (2023) Influenza surveillance with Baidu index and attention-based long short-term memory model. *PloS one*. 18 (1), e0280834. doi:10.1371/journal.pone.0280834.

Daley, D.J. & Gani, J. (1999) *Epidemic Modelling: An Introduction*. Cambridge University

Press. doi:10.1017/CBO9780511608834.

D'Antonio, M., Nguyen, J.P., Arthur, T.D., Matsui, H., COVID-19 Host Genetics Initiative, D'Antonio-Chronowska, A. & Frazer, K.A. (2021) SARS-CoV-2 susceptibility and COVID-19 disease severity are associated with genetic variants affecting gene expression in a variety of tissues. *Cell reports*. 37 (7), 110020. doi:10.1016/j.celrep.2021.110020.

Data Science Campus (2020) *google-mobility-reports-data: Archive of data extracted from the google community mobility reports*. Github. https://github.com/datasciencecampus/google-mobility-reports-data.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., et al. (2021) Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*. 372 (6538). doi:10.1126/science.abg3055.

Davies, N.G., Klepac, P., Liu, Y., Prem, K., Jit, M. & Eggo, R.M. (2020) Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nature medicine*. 26 (8), 1205–1211. doi:10.1038/s41591-020-0962-9.

Dean, N.E., Pastore Y Piontti, A., Madewell, Z.J., Cummings, D.A.T., Hitchings, M.D.T., Joshi, K., Kahn, R., Vespignani, A., Halloran, M.E. & Longini, I.M., Jr (2020) Ensemble forecast modeling for the design of COVID-19 vaccine efficacy trials. *Vaccine*. 38 (46), 7213–7216. doi:10.1016/j.vaccine.2020.09.031.

Delpino, F.M., Costa, Â.K., Farias, S.R., Chiavegatto Filho, A.D.P., Arcêncio, R.A. & Nunes, B.P. (2022) Machine learning for predicting chronic diseases: a systematic review. *Public health*. 205, 14–25. doi:10.1016/j.puhe.2022.01.007.

Deschepper, M., Eeckloo, K., Malfait, S., Benoit, D., Callens, S. & Vansteelandt, S. (2021) Prediction of hospital bed capacity during the COVID- 19 pandemic. *BMC health services research*. 21 (1), 468. doi:10.1186/s12913-021-06492-3.

Distante, C., Pereira, I.G., Gonçalves, L.M.G., Piscitelli, P. & Miani, A. (2020) Forecasting Covid-19 Outbreak Progression in Italian Regions: A model based on neural network training from Chinese data. *bioRxiv*. doi:10.1101/2020.04.09.20059055.

Distributed (Deep) Machine Learning Community (n.d.) *XGBoost Documentation — xgboost 2.1.0-dev documentation*. https://xgboost.readthedocs.io/en/latest/index.html [Accessed: 28 August 2023].

Docherty, A.B., Harrison, E.M., Green, C.A., Hardwick, H.E., Pius, R., et al. (2020) Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* . 369, m1985. doi:10.1136/bmj.m1985.

Dong, X., Boulton, M.L., Carlson, B., Montgomery, J.P. & Wells, E.V. (2017) Syndromic surveillance for influenza in Tianjin, China: 2013-14. *Journal of public health* . 39 (2), 274–281. doi:10.1093/pubmed/fdw022.

van Doremalen, N., Bushmaker, T., Morris, D.H., Holbrook, M.G., Gamble, A., Williamson, B.N., Tamin, A., Harcourt, J.L., Thornburg, N.J., Gerber, S.I., Lloyd-Smith, J.O., de Wit, E. & Munster, V.J. (2020) Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *The New England journal of medicine*. 382 (16), 1564–1567. doi:10.1056/NEJMc2004973.

Dorjee, S., Poljak, Z., Revie, C.W., Bridgland, J., McNab, B., Leger, E. & Sanchez, J. (2013)

A review of simulation modelling approaches used for the spread of zoonotic influenza viruses in animal and human populations. *Zoonoses and public health*. 60 (6), 383–411. doi:10.1111/zph.12010.

Dou, D., Revol, R., Östbye, H., Wang, H. & Daniels, R. (2018) Influenza A Virus Cell Entry, Replication, Virion Assembly and Movement. *Frontiers in immunology*. 9, 1581. doi:10.3389/fimmu.2018.01581.

Downes, M. & Carlin, J.B. (2020) Multilevel regression and poststratification as a modeling approach for estimating population quantities in large population health studies: A simulation study. *Biometrical journal. Biometrische Zeitschrift*. 62 (2), 479–491. doi:10.1002/bimj.201900023.

Drosten, C., Günther, S., Preiser, W., van der Werf, S., Brodt, H.-R., et al. (2003) Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *The New England journal of medicine*. 348 (20), 1967–1976. doi:10.1056/NEJMoa030747.

Dugas, A.F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T. & Rothman, R.E. (2013) Influenza forecasting with Google Flu Trends. *PloS one*. 8 (2), e56176. doi:10.1371/journal.pone.0056176.

Du, Z., Shao, Z., Zhang, X., Chen, R., Chen, T., Bai, Y., Wang, L., Lau, E.H.Y. & Cowling, B.J. (2023) Nowcasting and Forecasting Seasonal Influenza Epidemics - China, 2022-2023. *China CDC weekly*. 5 (49), 1100–1106. doi:10.46234/ccdcw2023.206.

Elliott, P., Bodinier, B., Eales, O., Wang, H., Haw, D., et al. (2022a) Rapid increase in Omicron infections in England during December 2021: REACT-1 study. *Science*. 375 (6587), 1406–1411. doi:10.1126/science.abn8347.

Elliott, P., Eales, O., Bodinier, B., Tang, D., Wang, H., et al. (2022b) Post-peak dynamics of a national Omicron SARS-CoV-2 epidemic during January 2022. *bioRxiv*. doi:10.1101/2022.02.03.22270365.

Elliott, P., Eales, O., Steyn, N., Tang, D., Bodinier, B., et al. (2022c) *Twin peaks: the Omicron SARS-CoV-2 BA.1 and BA.2 epidemics in England*. http://spiral.imperial.ac.uk/handle/10044/1/96170.

Elliott, P., Haw, D., Wang, H., Eales, O., Walters, C.E., et al. (2021) Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the Delta variant. *Science*. 374 (6574), eabl9551. doi:10.1126/science.abl9551.

England, N.H.S. (2020) *Landmark moment as first NHS patient receives COVID-19 vaccination*. 8 December 2020. https://www.england.nhs.uk/2020/12/landmark-moment-as-first-nhs-patient-receives-covid-19-vaccination/ [Accessed: 16 August 2023].

Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., et al. (2020) *Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand*. doi:10.25561/77482.

Fischer, L.S., Santibanez, S., Hatchett, R.J., Jernigan, D.B., Meyers, L.A., Thorpe, P.G. & Meltzer, M.I. (2016) CDC Grand Rounds: Modeling and Public Health Decision-Making. *MMWR. Morbidity and mortality weekly report*. 65 (48), 1374–1377. doi:10.15585/mmwr.mm6548a4.

Fleming, D.M. & Elliot, A.J. (2008) Lessons from 40 years' surveillance of influenza in England and Wales. *Epidemiology and infection*. 136 (7), 866–875.

doi:10.1017/S0950268807009910.

FluView, C.D.C. (2012) *ILINet State Activity Indicator Map*. 2012. https://gis.cdc.gov/grasp/fluview/main.html [Accessed: 26 May 2023].

Frank, E. & Hall, M. (2001) A Simple Approach to Ordinal Classification. In: *Machine Learning: ECML 2001*. Lecture notes in computer science. Berlin, Heidelberg, Springer Berlin Heidelberg. pp. 145–156. doi:10.1007/3-540-44795-4_13.

Freeman, E.E., McMahon, D.E., Lipoff, J.B., Rosenbach, M., Kovarik, C., Desai, S.R., Harp, J., Takeshita, J., French, L.E., Lim, H.W., Thiers, B.H., Hruza, G.J. & Fox, L.P. (2020) The spectrum of COVID-19-associated dermatologic manifestations: An international registry of 716 patients from 31 countries. *Journal of the American Academy of Dermatology*. 83 (4), 1118–1129. doi:10.1016/j.jaad.2020.06.1016.

Fung, I.C.-H., Tse, Z.T.H. & Fu, K.-W. (2015) The use of social media in public health surveillance. *Western Pacific surveillance and response journal : WPSAR*. 6 (2), 3–6. doi:10.5365/WPSAR.2015.6.1.019.

Furukawa, N.W., Brooks, J.T. & Sobel, J. (2020) Evidence Supporting Transmission of Severe Acute Respiratory Syndrome Coronavirus 2 While Presymptomatic or Asymptomatic. *Emerging infectious diseases*. 26 (7). doi:10.3201/eid2607.201595.

Galasso, J., Cao, D.M. & Hochberg, R. (2022) A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data. *Chaos, solitons, and fractals*. 156, 111779. doi:10.1016/j.chaos.2021.111779.

Galhardas, H. & Oliveira, P. (2005) A taxonomy of data quality problems. *International Workshop on Databases and Expert Systems Applications: proceedings*. https://www.academia.edu/15363781/A_Taxonomy_of_Data_Quality_Problems.

Ganslmeier, M., Furceri, D. & Ostry, J.D. (2021) The impact of weather on COVID-19 pandemic. *Scientific reports*. 11 (1), 22027. doi:10.1038/s41598-021-01189-3.

Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H.K., Morgan, R. & Klein, S.L. (2020) Impact of sex and gender on COVID-19 outcomes in Europe. *Biology of sex differences*. 11 (1), 29. doi:10.1186/s13293-020-00304-9.

Geldsetzer, P. (2020) Use of Rapid Online Surveys to Assess People's Perceptions During Infectious Disease Outbreaks: A Cross-sectional Survey on COVID-19. *Journal of medical Internet research*. 22 (4), e18790. doi:10.2196/18790.

Gelman, A. (2021) *Bayesian Data Analysis Third edition*. 15 February 2021. http://www.stat.columbia.edu/~gelman/book/BDA3.pdf [Accessed: 7 January 2024].

Gelman, A. & Carpenter, B. (2020) Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society. Series C, Applied statistics*. 69 (5), 1269–1283. doi:10.1111/rssc.12435.

Gelman, A., Phillips, J., Gabry, J. & Trangucci, R. (2016) *Using multilevel regression and poststratification to estimate dynamic public opinion*. 2016. http://www.stat.columbia.edu/~gelman/research/unpublished/MRT(1).pdf [Accessed: 14 August 2023].

Generous, N., Fairchild, G., Deshpande, A., Del Valle, S.Y. & Priedhorsky, R. (2014) Global disease monitoring and forecasting with Wikipedia. *PLoS computational biology*. 10 (11),

e1003892. doi:10.1371/journal.pcbi.1003892.

van Genugten, M.L.L., Heijnen, M.-L.A. & Jager, J.C. (2003) Pandemic influenza and healthcare demand in the Netherlands: scenario analysis. *Emerging infectious diseases*. 9 (5), 531–538. doi:10.3201/eid0905.020321.

Glasser, J.W., Hupert, N., McCauley, M.M. & Hatchett, R. (2011) Modeling and public health emergency responses: lessons from SARS. *Epidemics*. 3 (1), 32–37. doi:10.1016/j.epidem.2011.01.001.

Global Influenza Surveillance and Response System (GISRS) (n.d.) *Global Influenza Surveillance and Response System (GISRS)*. Global Influenza Surveillance and Response System (GISRS). https://www.who.int/initiatives/global-influenza-surveillance-and-response-system [Accessed: 10 July 2023].

Global Outbreak Alert and Response Network (n.d.) *GOARN*. https://goarn.who.int/ [Accessed: 23 July 2023].

Gneiting, T. & Raftery, A.E. (2007) Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*. 102 (477), 359–378. doi:10.1198/016214506000001437.

Goldstein, E., Lipsitch, M. & Cevik, M. (2021) On the Effect of Age on the Transmission of SARS-CoV-2 in Households, Schools, and the Community. *The Journal of infectious diseases*. 223 (3), 362–369. doi:10.1093/infdis/jiaa691.

Google (2020) *COVID-19 Community Mobility Reports*. https://www.google.com/covid19/mobility/.

Government Equalities Office, Race Disparity Unit (2021) *Second quarterly report on progress to address COVID-19 health inequalities*. 2021. GOV.UK. https://www.gov.uk/government/publications/second-quarterly-report-on-progress-to-address-covid-19-health-inequalities/second-quarterly-report-on-progress-to-address-covid-19-health-inequalities [Accessed: 1 February 2023].

Gov, G. (2019) *View Opportunity*. 25 January 2019. VIEW GRANT OPPORTUNITY. https://www.grants.gov/web/grants/view-opportunity.html?oppId=311615 [Accessed: 20 February 2023].

Grasselli, G., Zangrillo, A., Zanella, A., Antonelli, M., Cabrini, L., et al. (2020) Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. *JAMA: the journal of the American Medical Association*. 323 (16), 1574–1581. doi:10.1001/jama.2020.5394.

Grohskopf, L.A., Blanton, L.H., Ferdinands, J.M., Chung, J.R., Broder, K.R., Talbot, H.K., Morgan, R.L. & Fry, A.M. (2022) Prevention and Control of Seasonal Influenza with Vaccines: Recommendations of the Advisory Committee on Immunization Practices - United States, 2022-23 Influenza Season. *MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and reports / Centers for Disease Control*. 71 (1), 1–28. doi:10.15585/mmwr.rr7101a1.

Groves, H.E., Papenburg, J., Mehta, K., Bettinger, J.A., Sadarangani, M., Halperin, S.A., Morris, S.K., Bancej, C., Burton, C., Embree, J. & Others (2022) The effect of the COVID-19 pandemic on influenza-related hospitalization, intensive care admission and mortality in children in Canada: A population-based study. *The Lancet Regional Health--Americas*. 7. https://www.thelancet.com/journals/lanam/article/PIIS2667-193X(21)00128-9/fulltext.

Guan, W.-J., Liang, W.-H., Zhao, Y., Liang, H.-R., Chen, Z.-S., et al. (2020a) Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. *The European respiratory journal: official journal of the European Society for Clinical Respiratory Physiology*. 55 (5). doi:10.1183/13993003.00547-2020.

Guan, W.-J., Ni, Z.-Y., Hu, Y., Liang, W.-H., Ou, C.-Q., et al. (2020b) Clinical Characteristics of Coronavirus Disease 2019 in China. *The New England journal of medicine*. 382 (18), 1708–1720. doi:10.1056/NEJMoa2002032.

Guefrechi, S., Jabra, M.B., Ammar, A., Koubaa, A. & Hamam, H. (2021) Deep learning based detection of COVID-19 from chest X-ray images. *Multimedia tools and applications*. 80 (21-23), 31803–31820. doi:10.1007/s11042-021-11192-5.

Guerra, J., Acharya, P. & Barnadas, C. (2019) Community-based surveillance: A scoping review. *PloS one*. 14 (4), e0215278. doi:10.1371/journal.pone.0215278.

Guo, Z.-D., Wang, Z.-Y., Zhang, S.-F., Li, X., Li, L., Li, C., Cui, Y., Fu, R.-B., Dong, Y.-Z., Chi, X.-Y., Zhang, M.-Y., Liu, K., Cao, C., Liu, B., Zhang, K., Gao, Y.-W., Lu, B. & Chen, W. (2020) Aerosol and Surface Distribution of Severe Acute Respiratory Syndrome Coronavirus 2 in Hospital Wards, Wuhan, China, 2020. *Emerging infectious diseases*. 26 (7), 1583–1591. doi:10.3201/eid2607.200885.

Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., et al. (2020) Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science*. 369 (6504), 718–724. doi:10.1126/science.abc6027.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S. & Tatlow, H. (2021) A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature human behaviour*. 5 (4), 529–538. doi:10.1038/s41562-021-01079-8.

Halloran, M.E., Ferguson, N.M., Eubank, S., Longini, I.M., Cummings, D.A.T., Lewis, B., Xu, S., Fraser, C., Vullikanti, A., Germann, T.C., Wagener, D., Beckman, R., Kadau, K., Barrett, C., Macken, C.A., Burke, D.S. & Cooley, P. (2008) Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences*. 105 (12), 4639–4644. doi:10.1073/pnas.0706849105.

Hasan, N. (2020) A Methodological Approach for Predicting COVID-19 Epidemic Using EEMD-ANN Hybrid Model. *Internet of Things*. 11, 100228. doi:10.1016/j.iot.2020.100228.

van der Heijden, K. (2005) *Scenarios: The Art of Strategic Conversation*. John Wiley & Sons. https://play.google.com/store/books/details?id=kEPKCQAAQBAJ.

Hernandez, J.B.R. & Kim, P.Y. (2022) *Epidemiology Morbidity And Mortality*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK547668/.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., et al. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*. 146 (730), 1999–2049. doi:10.1002/qj.3803.

Hossain, M.A., Rashid, M.U.B., Khan, M.A.S., Sayeed, S., Kader, M.A. & Hawlader, M.D.H. (2021) Healthcare Workers' Knowledge, Attitude, and Practice Regarding Personal Protective Equipment for the Prevention of COVID-19. *Journal of multidisciplinary healthcare*. 14, 229–238. doi:10.2147/JMDH.S293717.

Howerton, E., Contamin, L., Mullany, L.C., Qin, M., Reich, N.G., et al. (2023) Evaluation of

the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty. *Nature communications*. 14 (1), 7260. doi:10.1038/s41467-023-42680-x.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., et al. (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*. 395 (10223), 497–506. doi:10.1016/S0140-6736(20)30183-5.

Huang, N., Pérez, P., Kato, T., Mikami, Y., Okuda, K., et al. (2021) SARS-CoV-2 infection of the oral cavity and saliva. *Nature medicine*. 27 (5), 892–903. doi:10.1038/s41591-021-01296-8.

Hu, H., Wang, H., Wang, F., Langley, D., Avram, A. & Liu, M. (2018) Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network. *Scientific reports*. 8 (1), 4895. doi:10.1038/s41598-018-23075-1.

Hu, S., Xiong, C., Younes, H., Yang, M., Darzi, A. & Jin, Z.C. (2022) Examining spatiotemporal evolution of racial/ethnic disparities in human mobility and COVID-19 health outcomes: Evidence from the contiguous United States. *Sustainable cities and society*. 76, 103506. doi:10.1016/j.scs.2021.103506.

Hyndman, R.J. & Athanasopoulos, G. (2021) *Forecasting: Principles and Practice (3rd ed)*. May 2021. https://otexts.com/fpp3/ [Accessed: 12 August 2023].

Hyndman, R.J. & Koehler, A.B. (2006) Another look at measures of forecast accuracy. *International journal of forecasting*. 22 (4), 679–688. doi:10.1016/j.ijforecast.2006.03.001.

Ilin, C., Annan-Phan, S., Tai, X.H., Mehra, S., Hsiang, S. & Blumenstock, J.E. (2021) Public mobility data enables COVID-19 forecasting and management at local and global scales. *Scientific reports*. 11 (1), 13531. doi:10.1038/s41598-021-92892-8.

Imperial College London (2020) *REACT 1 Study Materials*. 2020. https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/for-researchers/react-1-study-materials/ [Accessed: 28 August 2023].

Institute for Health Metrics and Evaluation (2019) *GBD Results*. 2019. Institute for Health Metrics and Evaluation. https://vizhub.healthdata.org/gbd-results/ [Accessed: 20 October 2022].

Ipsos (n.d.) *Ipsos*. https://www.ipsos.com/en-uk [Accessed: 2 January 2024].

Ishak, A., Mehendale, M., AlRawashdeh, M.M., Sestacovschi, C., Sharath, M., Pandav, K. & Marzban, S. (2022) The association of COVID-19 severity and susceptibility and genetic risk factors: A systematic review of the literature. *Gene*. 836, 146674. doi:10.1016/j.gene.2022.146674.

Islam, M.M., Karray, F., Alhajj, R. & Zeng, J. (2021) A Review on Deep Learning Techniques for the Diagnosis of Novel Coronavirus (COVID-19). *IEEE access : practical innovations, open solutions*. 9, 30551–30572. doi:10.1109/ACCESS.2021.3058537.

Iuliano, A.D., Roguski, K.M., Chang, H.H., Muscatello, D.J., Palekar, R., et al. (2018) Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *The Lancet*. 391 (10127), 1285–1300. doi:10.1016/S0140-6736(17)33293-2.

Iwasaki, A. & Pillai, P.S. (2014) Innate immunity to influenza virus infection. *Nature reviews. Immunology*. 14 (5), 315–328. doi:10.1038/nri3665.

Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S.,

Sharma Mittal, R. & Munigala, V. (2020) Overview and Importance of Data Quality for Machine Learning Tasks. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '20. 23 August 2020 New York, NY, USA, Association for Computing Machinery. pp. 3561–3562. doi:10.1145/3394486.3406477.

Jain, S., Kamimoto, L., Bramley, A.M., Schmitz, A.M., Benoit, S.R., et al. (2009) Hospitalized patients with 2009 H1N1 influenza in the United States, April-June 2009. *The New England journal of medicine*. 361 (20), 1935–1944. doi:10.1056/NEJMoa0906695.

Jajosky, R.A. & Groseclose, S.L. (2004) Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. *BMC public health*. 4, 29. doi:10.1186/1471-2458-4-29.

Jalilian, A. & Mateu, J. (2021) A hierarchical spatio-temporal model to analyze relative risk variations of COVID-19: a focus on Spain, Italy and Germany. *Stochastic environmental research and risk assessment: research journal*. 1–16. doi:10.1007/s00477-021-02003-2.

Janssens, Y., Joye, J., Waerlop, G., Clement, F., Leroux-Roels, G. & Leroux-Roels, I. (2022) The role of cell-mediated immunity against influenza and its implications for vaccine evaluation. *Frontiers in immunology*. 13, 959379. doi:10.3389/fimmu.2022.959379.

Jiang, F., Deng, L., Zhang, L., Cai, Y., Cheung, C.W. & Xia, Z. (2020) Review of the Clinical Characteristics of Coronavirus Disease 2019 (COVID-19). *Journal of general internal medicine*.35 (5) pp.1545–1549. doi:10.1007/s11606-020-05762-w.

Johns Hopkins University & Medicine (2020) *COVID-19 Map*. 2020. Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html [Accessed: 18 July 2023].

Johnson, M.R., Naik, H., Chan, W.S., Greiner, J., Michaleski, M., Liu, D., Silvestre, B. & McCarthy, I.P. (2023) Forecasting ward-level bed requirements to aid pandemic resource planning: Lessons learned and future directions. *Health care management science*. 1–24. doi:10.1007/s10729-023-09639-2.

Jordan, R.E. & Adab, P. (2020) Who is most likely to be infected with SARS-CoV-2? *The Lancet infectious diseases*.20 (9) pp.995–996. doi:10.1016/S1473-3099(20)30395-9.

Joslyn, S. & Savelli, S. (2021) Visualizing Uncertainty for Non-Expert End Users: The Challenge of the Deterministic Construal Error. *Frontiers in Computer Science*. 2. doi:10.3389/fcomp.2020.590232.

Kenah, E., Chao, D.L., Matrajt, L., Halloran, M.E. & Longini, I.M., Jr (2011) The global transmission and control of influenza. *PloS one*. 6 (5), e19515. doi:10.1371/journal.pone.0019515.

Keogh-Brown, M.R., Jensen, H.T., Edmunds, W.J. & Smith, R.D. (2020) The impact of Covid-19, associated behaviours and policies on the UK economy: A computable general equilibrium model. *SSM - population health*. 12, 100651. doi:10.1016/j.ssmph.2020.100651.

Kermack, W.O., McKendrick, A.G. & Walker, G.T. (1927) A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*. 115 (772), 700–721. doi:10.1098/rspa.1927.0118.

Kesper, A., Wenz, V. & Taentzer, G. (2020) Detecting quality problems in research data: a

model-driven approach. In: *Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems*. MODELS '20. 16 October 2020 New York, NY, USA, Association for Computing Machinery. pp. 354–364. doi:10.1145/3365438.3410987.

Khanna, M., Kumar, P., Choudhary, K., Kumar, B. & Vijayan, V.K. (2008) Emerging influenza virus: a global threat. *Journal of biosciences*. 33 (4), 475–482. doi:10.1007/s12038-008-0066-z.

Khunti, K., Singh, A.K., Pareek, M. & Hanif, W. (2020) Is ethnicity linked to incidence or outcomes of covid-19? *BMJ* . 369. doi:10.1136/bmj.m1548.

Kim, J. & Ahn, I. (2019) Weekly ILI patient ratio change prediction using news articles with support vector machine. *BMC bioinformatics*. 20 (1), 259. doi:10.1186/s12859-019-2894-2.

Kirby, J.E., Riedel, S., Dutta, S., Arnaout, R., Cheng, A., Ditelberg, S., Hamel, D.J., Chang, C.A. & Kanki, P.J. (2023) Sars-Cov-2 antigen tests predict infectivity based on viral culture: comparison of antigen, PCR viral load, and viral culture testing on a large sample cohort. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*. 29 (1), 94–100. doi:10.1016/j.cmi.2022.07.010.

Kirby, T. (2020) Evidence mounts on the disproportionate effect of COVID-19 on ethnic minorities. *The Lancet. Respiratory medicine*. 8 (6), 547–548. doi:10.1016/S2213-2600(20)30228-9.

Klein, B., LaRock, T., McCabe, S., Torres, L., Friedland, L., Kos, M., Privitera, F., Lake, B., Kraemer, M.U.G., Brownstein, J.S., Gonzalez, R., Lazer, D., Eliassi-Rad, T., Scarpino, S.V., Vespignani, A. & Chinazzi, M. (2022) Characterizing collective physical distancing in the U.S. during the first nine months of the COVID-19 pandemic. *arXiv [physics.soc-ph]*. http://arxiv.org/abs/2212.08873.

Klein, B., Zenteno, A.C., Joseph, D., Zahedi, M., Hu, M., Copenhaver, M.S., Kraemer, M.U.G., Chinazzi, M., Klompas, M., Vespignani, A., Scarpino, S.V. & Salmasian, H. (2023) Forecasting hospital-level COVID-19 admissions using real-time mobility data. *Communication & medicine*. 3 (1), 25. doi:10.1038/s43856-023-00253-5.

Kobres, P.-Y., Chretien, J.-P., Johansson, M.A., Morgan, J.J., Whung, P.-Y., Mukundan, H., Del Valle, S.Y., Forshey, B.M., Quandelacy, T.M., Biggerstaff, M., Viboud, C. & Pollett, S. (2019) A systematic review and evaluation of Zika virus forecasting and prediction research during a public health emergency of international concern. *PLoS neglected tropical diseases*. 13 (10), e0007451. doi:10.1371/journal.pntd.0007451.

Koel, B.F., Burke, D.F., Bestebroer, T.M., van der Vliet, S., Zondag, G.C.M., Vervaet, G., Skepner, E., Lewis, N.S., Spronken, M.I.J., Russell, C.A., Eropkin, M.Y., Hurt, A.C., Barr, I.G., de Jong, J.C., Rimmelzwaan, G.F., Osterhaus, A.D.M.E., Fouchier, R.A.M. & Smith, D.J. (2013) Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution. *Science*. 342 (6161), 976–979. doi:10.1126/science.1244730.

Koltai, J., Vásárhelyi, O., Röst, G. & Karsai, M. (2022) Reconstructing social mixing patterns via weighted contact matrices from online and representative surveys. *Scientific reports*. 12 (1), 4690. doi:10.1038/s41598-022-07488-7.

Koutsandreas, D., Spiliotis, E., Petropoulos, F. & Assimakopoulos, V. (2022) On the selection of forecasting accuracy measures. *The Journal of the Operational Research Society*. 73 (5), 937–954. doi:10.1080/01605682.2021.1892464.

Kovács, G., Kaló, Z., Jahnz-Rozyk, K., Kyncl, J., Csohan, A., Pistol, A., Leleka, M., Kipshakbaev, R., Durand, L. & Macabeo, B. (2014) Medical and economic burden of influenza in the elderly population in central and eastern European countries. *Human vaccines & immunotherapeutics*. 10 (2), 428–440. doi:10.4161/hv.26886.

Kraemer, M.U.G., Hill, V., Ruis, C., Dellicour, S., Bajaj, S., et al. (2021) Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science*. 373 (6557), 889–895. doi:10.1126/science.abj0113.

Kraemer, M.U.G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D.M., Group†, O.C.-19 D.W., du Plessis, L., Faria, N.R., Li, R., Hanage, W.P., Brownstein, J.S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O.G. & Scarpino, S.V. (2020) The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*. 368 (6490), 493–497. doi:10.1126/science.abb4218.

Kramer, S.C. & Shaman, J. (2019) Development and validation of influenza forecasting for 64 temperate and tropical countries. *PLoS computational biology*. 15 (2), e1006742. doi:10.1371/journal.pcbi.1006742.

Krammer, F. (2019) The human antibody response to influenza A virus infection and vaccination. *Nature reviews. Immunology*. 19 (6), 383–397. doi:10.1038/s41577-019-0143-6.

Krammer, F. & Palese, P. (2015) Advances in the development of influenza virus vaccines. *Nature reviews. Drug discovery*. 14 (3), 167–182. doi:10.1038/nrd4529.

Kretzschmar, M.E., Rozhnova, G., Bootsma, M.C.J., van Boven, M., van de Wijgert, J.H.H.M. & Bonten, M.J.M. (2020) Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *The Lancet. Public health*. 5 (8), e452–e459. doi:10.1016/S2468-2667(20)30157-2.

Krug, R.M. & Fodor, E. (2013) *The virus genome and its replication*. 57–66. doi:10.1002/9781118636817.ch4.

Krylova, O., Kazmi, O., Wang, H., Lam, K., Logar-Henderson, C. & Gapanenko, K. (2020) Estimating surge in COVID-19 cases, hospital resources and PPE demand with the interactive and locally-informed COVID-19 Health System Capacity Planning Tool. *International journal of population data science*. 5 (4), 1710. doi:10.23889/ijpds.v5i4.1710.

Kucharski, A.J., Klepac, P., Conlan, A.J.K., Kissler, S.M., Tang, M.L., Fry, H., Gog, J.R., Edmunds, W.J. & CMMID COVID-19 working group (2020) Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study. *The Lancet infectious diseases*. 20 (10), 1151–1160. doi:10.1016/S1473-3099(20)30457-6.

Kucirka, L.M., Lauer, S.A., Laeyendecker, O., Boon, D. & Lessler, J. (2020) Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction-Based SARS-CoV-2 Tests by Time Since Exposure. *Annals of internal medicine*. 173 (4), 262–267. doi:10.7326/M20-1495.

Kumaresan, V., Balachandar, N., Poole, S.F., Myers, L.J., Varghese, P., Washington, V., Jia, Y. & Lee, V.S. (2023) Fitting and validation of an agent-based model for COVID-19 case forecasting in workplaces and universities. *PloS one*. 18 (3), e0283517. doi:10.1371/journal.pone.0283517.

Kwok, K.O., Cowling, B., Wei, V., Riley, S. & Read, J.M. (2018) Temporal variation of human encounters and the number of locations in which they occur: a longitudinal study of Hong

Kong residents. *Journal of the Royal Society, Interface / the Royal Society*. 15 (138). doi:10.1098/rsif.2017.0838.

Kwok, K.O., Li, K.K., Chan, H.H.H., Yi, Y.Y., Tang, A., Wei, W.I. & Wong, S.Y.S. (2020) Community Responses during Early Phase of COVID-19 Epidemic, Hong Kong. *Emerging infectious diseases*. 26 (7), 1575–1579. doi:10.3201/eid2607.200500.

Lamb, R.A. (2008) Influenza. In: B.W.J. Mahy & M.H.V. Van Regenmortel (eds.). *Encyclopedia of Virology (Third Edition)*. Oxford, Academic Press. pp. 95–104. doi:10.1016/B978-012374410-4.00654-3.

Laskowski, M., Demianyk, B.C.P., Witt, J., Mukhi, S.N., Friesen, M.R. & McLeod, R.D. (2011) Agent-based modeling of the spread of influenza-like illness in an emergency department: a simulation study. *IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society*. 15 (6), 877–889. doi:10.1109/TITB.2011.2163414.

Latha, C.B.C. & Jeeva, S.C. (2019) Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*. 16, 100203. doi:10.1016/j.imu.2019.100203.

Lauer, S.A., Brown, A.C. & Reich, N.G. (2020) *Infectious Disease Forecasting for Public Health*. doi:10.1093/oso/9780198853244.003.0004.

Lau, M.S.Y., Grenfell, B., Thomas, M., Bryan, M., Nelson, K. & Lopman, B. (2020) Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. *Proceedings of the National Academy of Sciences of the United States of America*. 117 (36), 22430–22435. doi:10.1073/pnas.2011802117.

LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*. 521 (7553), 436–444. doi:10.1038/nature14539.

Levin, A.T., Hanage, W.P., Owusu-Boaitey, N., Cochran, K.B., Walsh, S.P. & Meyerowitz-Katz, G. (2020) Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. *European journal of epidemiology*. 35 (12), 1123–1138. doi:10.1007/s10654-020-00698-1.

Liang, F., Guan, P., Wu, W. & Huang, D. (2018) Forecasting influenza epidemics by integrating internet search queries and traditional surveillance data with the support vector machine regression model in Liaoning, from 2011 to 2015. *PeerJ*. 6, e5134. doi:10.7717/peerj.5134.

Liang, J.-B., Yuan, H.-Y., Li, K.-K., Wei, W.-I., Wong, S.Y.S., Tang, A., Riley, S. & Kwok, K.O. (2022) Path to normality: Assessing the level of social-distancing measures relaxation against antibody-resistant SARS-CoV-2 variants in a partially-vaccinated population. *Computational and structural biotechnology journal*. 20, 4052–4059. doi:10.1016/j.csbj.2022.07.048.

Li, L., Wong, J.Y., Wu, P., Bond, H.S., Lau, E.H.Y., Sullivan, S.G. & Cowling, B.J. (2018) Heterogeneity in Estimates of the Impact of Influenza on Population Mortality: A Systematic Review. *American journal of epidemiology*. 187 (2), 378–388. doi:10.1093/aje/kwx270.

Lin, C.Y. (2008) Modeling infectious diseases in humans and animals by KEELING, M. j. and ROHANI, P. *Biometrics*. 64 (3), 993–993. doi:10.1111/j.1541-0420.2008.01082_7.x.

Liu, R., Sheng, Z., Huang, C., Wang, D. & Li, F. (2020) Influenza D virus. *Current opinion in*

virology. 44, 154–161. doi:10.1016/j.coviro.2020.08.004.

Li, Z., Luo, X., Wang, B., Bertozzi, A.L. & Xin, J. (2020) A Study on Graph-Structured Recurrent Neural Networks and Sparsification with Application to Epidemic Forecasting. In: *Optimization of Complex Systems: Theory, Models, Algorithms and Applications*. 2020 Springer International Publishing. pp. 730–739. doi:10.1007/978-3-030-21803-4_73.

Lofgren Eric, Fefferman N. H., Naumov Y. N., Gorski J. & Naumova E. N. (2007) Influenza Seasonality: Underlying Causes and Modeling Theories. *Journal of virology*. 81 (11), 5429–5436. doi:10.1128/jvi.01680-06.

Lombardo, J., Burkom, H., Elbert, E., Magruder, S., Lewis, S.H., Loschen, W., Sari, J., Sniegoski, C., Wojcik, R. & Pavlin, J. (2003) A systems overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II). *Journal of urban health: bulletin of the New York Academy of Medicine*. 80 (2 Suppl 1), i32–i42. doi:10.1007/pl00022313.

Lo Piano, S. (2020) Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications*. 7 (1), 1–7. doi:10.1057/s41599-020-0501-9.

Love, J., Wimmer, M.T., Toth, D.J.A., Chandran, A., Makhija, D., Cooper, C.K., Samore, M.H. & Keegan, L.T. (2021) Comparison of antigen- and RT-PCR-based testing strategies for detection of SARS-CoV-2 in two high-exposure settings. *PloS one*. 16 (9), e0253407. doi:10.1371/journal.pone.0253407.

Lowen, A.C. & Steel, J. (2014) Roles of humidity and temperature in shaping influenza seasonality. *Journal of virology*. 88 (14), 7692–7695. doi:10.1128/JVI.03544-13.

Lumley, T. (2004) Analysis of Complex Survey Samples. *Journal of statistical software*. 9, 1–19. doi:10.18637/jss.v009.i08.

Lumley, T. (2023) *pacsurvey: Analysis of Complex Survey Samples*. 3 May 2023. pacsurvey: Analysis of Complex Survey Samples. https://www.jstatsoft.org/issue/view/v009 [Accessed: 28 August 2023].

Lurie, N., Saville, M., Hatchett, R. & Halton, J. (2020) Developing Covid-19 Vaccines at Pandemic Speed. *The New England journal of medicine*. 382 (21), 1969–1973. doi:10.1056/NEJMp2005630.

de Lusignan, S., Dorward, J., Correa, A., Jones, N., Akinyemi, O., et al. (2020) Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study. *The Lancet infectious diseases*. 20 (9), 1034–1042. doi:10.1016/S1473-3099(20)30371-6.

Lutz, C.S., Huynh, M.P., Schroeder, M., Anyatonwu, S., Dahlgren, F.S., Danyluk, G., Fernandez, D., Greene, S.K., Kipshidze, N., Liu, L., Mgbere, O., McHugh, L.A., Myers, J.F., Siniscalchi, A., Sullivan, A.D., West, N., Johansson, M.A. & Biggerstaff, M. (2019) Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC public health*. 19 (1), 1659. doi:10.1186/s12889-019-7966-8.

Lynch, P. (2008) The origins of computer weather prediction and climate modeling. *Journal of computational physics*. 227 (7), 3431–3444. doi:10.1016/j.jcp.2007.02.034.

Mackey, K., Ayers, C.K., Kondo, K.K., Saha, S., Advani, S.M., Young, S., Spencer, H., Rusek, M., Anderson, J., Veazie, S., Smith, M. & Kansagara, D. (2021) Racial and Ethnic

Disparities in COVID-19-Related Infections, Hospitalizations, and Deaths : A Systematic Review. *Annals of internal medicine*. 174 (3), 362–373. doi:10.7326/M20-6306.

Maestripieri, L. (2021) The Covid-19 Pandemics: why Intersectionality Matters. *Frontiers in sociology*. 6, 642662. doi:10.3389/fsoc.2021.642662.

Mahajan, P., Uddin, S., Hajati, F. & Moni, M.A. (2023) Ensemble Learning for Disease Prediction: A Review. *Healthcare (Basel, Switzerland)*. 11 (12). doi:10.3390/healthcare11121808.

Mahase, E. (2020) Covid-19: UK starts social distancing after new model points to 260 000 potential deaths. *BMJ* . 368, m1089. doi:10.1136/bmj.m1089.

Mahmood, S.U., Crimbly, F., Khan, S., Choudry, E. & Mehwish, S. (2020) Strategies for Rational Use of Personal Protective Equipment (PPE) Among Healthcare Providers During the COVID-19 Crisis. *Cureus*. 12 (5), e8248. doi:10.7759/cureus.8248.

Mahy, B.W.J. (2010) The Evolution and Emergence of RNA Viruses. *Emerging infectious diseases*. 16 (5), 899. doi:10.3201/eid1605.100164.

Majeed, M.A., Shafri, H.Z.M., Zulkafli, Z. & Wayayok, A. (2023) A Deep Learning Approach for Dengue Fever Prediction in Malaysia Using LSTM with Spatial Attention. *International journal of environmental research and public health*. 20 (5). doi:10.3390/ijerph20054130.

Makridakis, S., Spiliotis, E. & Assimakopoulos, V. (2018) Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one*. 13 (3), e0194889. doi:10.1371/journal.pone.0194889.

Mandl, K.D., Overhage, J.M., Wagner, M.M., Lober, W.B., Sebastiani, P., Mostashari, F., Pavlin, J.A., Gesteland, P.H., Treadwell, T., Koski, E., Hutwagner, L., Buckeridge, D.L., Aller, R.D. & Grannis, S. (2004) Implementing syndromic surveillance: a practical guide informed by the early experience. *Journal of the American Medical Informatics Association: JAMIA*. 11 (2), 141–150. doi:10.1197/jamia.M1356.

Martin, C.A., Jenkins, D.R., Minhas, J.S., Gray, L.J., Tang, J., et al. (2020) Socio-demographic heterogeneity in the prevalence of COVID-19 during lockdown is associated with ethnicity and household size: Results from an observational cohort study. *EClinicalMedicine*. 25, 100466. doi:10.1016/j.eclinm.2020.100466.

Martin-Delgado, J., Viteri, E., Mula, A., Serpa, P., Pacheco, G., Prada, D., Campos de Andrade Lourenção, D., Campos Pavan Baptista, P., Ramirez, G. & Mira, J.J. (2020) Availability of personal protective equipment and diagnostic and treatment facilities for healthcare workers involved in COVID-19 care: A cross-sectional study in Brazil, Colombia, and Ecuador. *PloS one*. 15 (11), e0242185. doi:10.1371/journal.pone.0242185.

Martin-Moreno, J.M., Alegre-Martinez, A., Martin-Gorgojo, V., Alfonso-Sanchez, J.L., Torres, F. & Pallares-Carratala, V. (2022) Predictive Models for Forecasting Public Health Scenarios: Practical Experiences Applied during the First Wave of the COVID-19 Pandemic. *International journal of environmental research and public health*. 19 (9). doi:10.3390/ijerph19095546.

Mass, C. (2012) Nowcasting: The Promise of New Technologies of Communication, Modeling, and Observation. *Bulletin of the American Meteorological Society*. 93 (6), 797–809. doi:10.1175/BAMS-D-11-00153.1.

McClymont, H. & Hu, W. (2021) Weather Variability and COVID-19 Transmission: A Review

of Recent Research. *International journal of environmental research and public health*. 18 (2). doi:10.3390/ijerph18020396.

McCullagh, P. (1980) Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B, Statistical methodology*. 42 (2), 109–142. http://www.jstor.org/stable/2984952.

McGough, S.F., Johansson, M.A., Lipsitch, M. & Menzies, N.A. (2020) Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLoS computational biology*. 16 (4), e1007735. doi:10.1371/journal.pcbi.1007735.

Meakin, S., Abbott, S., Bosse, N., Munday, J., Gruson, H., Hellewell, J., Sherratt, K., CMMID COVID-19 Working Group & Funk, S. (2022) Comparative assessment of methods for short-term forecasts of COVID-19 hospital admissions in England at the local level. *BMC medicine*. 20 (1), 86. doi:10.1186/s12916-022-02271-x.

Mellor, J., Christie, R., Paton, R.S., Leslie, R., Tang, M., Fyles, M., Deeny, S., Ward, T. & Overton, C.E. (2023) Real-time COVID-19 hospital admissions forecasting with leading indicators and ensemble methods in England. *arXiv [stat.AP]*. http://arxiv.org/abs/2306.05762.

Meslé, M.M., Brown, J., Mook, P., Hagan, J., Pastore, R., et al. (2021) Estimated number of deaths directly averted in people 60 years and older as a result of COVID-19 vaccination in the WHO European Region, December 2020 to November 2021. *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*. 26 (47). doi:10.2807/1560-7917.ES.2021.26.47.2101021.

Meyer, B., Reimerink, J., Torriani, G., Brouwer, F., Godeke, G.-J., Yerly, S., Hoogerwerf, M., Vuilleumier, N., Kaiser, L., Eckerle, I. & Reusken, C. (2020) Validation and clinical evaluation of a SARS-CoV-2 surrogate virus neutralisation test (sVNT). *Emerging microbes & infections*. 9 (1), 2394–2403. doi:10.1080/22221751.2020.1835448.

Midway, S.R. (2020) Principles of Effective Data Visualization. *Patterns (New York, N.Y.)*. 1 (9), 100141. doi:10.1016/j.patter.2020.100141.

Miller, F. & Miller, P. (2021) Transgenerational Trauma and Trust Restoration. *AMA journal of ethics*. 23 (6), E480–E486. doi:10.1001/amajethics.2021.480.

Ming, R.-X., Liu, J.-M., W Cheung, W.K. & Wan, X. (2016) Stochastic modelling of infectious diseases for heterogeneous populations. *Infectious diseases of poverty*. 5 (1), 107. doi:10.1186/s40249-016-0199-5.

Molinari, N.-A.M., Ortega-Sanchez, I.R., Messonnier, M.L., Thompson, W.W., Wortley, P.M., Weintraub, E. & Bridges, C.B. (2007) The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine*. 25 (27), 5086–5096. doi:10.1016/j.vaccine.2007.03.046.

Moller, A.R. (2001) Severe Local Storms Forecasting. In: C.A. Doswell (ed.). *Severe Convective Storms*. Boston, MA, American Meteorological Society. pp. 433–480. doi:10.1007/978-1-935704-06-5_11.

Moore, J.B. & June, C.H. (2020) Cytokine release syndrome in severe COVID-19. *Science*. 368 (6490), 473–474. doi:10.1126/science.abb8925.

Moorthy, V., Henao Restrepo, A.M., Preziosi, M.-P. & Swaminathan, S. (2020) Data sharing for novel coronavirus (COVID-19). *Bulletin of the World Health Organization*. 98 (3), 150.

doi:10.2471/BLT.20.251561.

de Moraes Batista, A.F., Miraglia, J.L., Rizzi Donato, T.H. & Porto Chiavegatto Filho, A.D. (2020) COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. *bioRxiv*. doi:10.1101/2020.04.04.20052092.

Moran, K.R., Fairchild, G., Generous, N., Hickmann, K., Osthus, D., Priedhorsky, R., Hyman, J. & Del Valle, S.Y. (2016) Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast. *The Journal of infectious diseases*. 214 (suppl_4), S404–S408. doi:10.1093/infdis/jiw375.

Morawska, L. & Milton, D.K. (2020) It Is Time to Address Airborne Transmission of Coronavirus Disease 2019 (COVID-19). *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*. 71 (9), 2311–2313. doi:10.1093/cid/ciaa939.

Morens, D.M., Folkers, G.K. & Fauci, A.S. (2004) The challenge of emerging and re-emerging infectious diseases. *Nature*. 430 (6996), 242–249. doi:10.1038/nature02759.

Morss, R.E., Demuth, J.L. & Lazo, J.K. (2008) Communicating Uncertainty in Weather Forecasts: A Survey of the U.S. Public. *Weather and Forecasting*. 23 (5), 974–991. doi:10.1175/2008WAF2007088.1.

Murray, J.L.K., Marques, D.F.P., Cameron, R.L., Potts, A., Bishop, J., von Wissmann, B., William, N., Reynolds, A.J., Robertson, C. & McMenamin, J. (2018) Moving epidemic method (MEM) applied to virology data as a novel real time tool to predict peak in seasonal influenza healthcare utilisation. The Scottish experience of the 2017/18 season to date. *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*. 23 (11). doi:10.2807/1560-7917.ES.2018.23.11.18-00079.

Mustafa, S., Zhang, Y., Zibwowa, Z., Seifeldin, R., Ako-Egbe, L., McDarby, G., Kelley, E. & Saikat, S. (2021) COVID-19 Preparedness and Response Plans from 106 countries: a review from a health systems resilience perspective. *Health policy and planning*. 37 (2), 255–268. doi:10.1093/heapol/czab089.

Myers, M.F., Rogers, D.J., Cox, J., Flahault, A. & Hay, S.I. (2000) Forecasting disease risk for increased epidemic preparedness in public health. *Advances in parasitology*. 47, 309–330. doi:10.1016/s0065-308x(00)47013-2.

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A. & Bürkner, P.-C. (2019) An Introduction to Bayesian Multilevel Models Using brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian. *Journal of speech, language, and hearing research: JSLHR*. 62 (5), 1225–1242. doi:10.1044/2018_JSLHR-S-18-0006.

Napoli, C., Fabiani, M., Rizzo, C., Barral, M., Oxford, J., Cohen, J., Niddam, L., Goryński, P., Pistol, A., Lionis, C., Briand, S., Nicoll, A., Penttinen, P., Gauci, C., Bounekkar, A., Bonnevay, S. & Beresniak, A. (2015) Assessment of human influenza pandemic scenarios in Europe. *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*. 20 (7), 29–38. doi:10.2807/1560-7917.es2015.20.7.21038.

National Institute for Health and Care Excellence (NICE) (2020) *COVID-19 rapid guideline: managing the long-term effects of COVID-19 (NG188)*. National Institute for Health and Care Excellence (NICE). https://www.ncbi.nlm.nih.gov/books/NBK567264/.

National Institutes of Health (2023) *Testing for SARS-CoV-2 Infection*. 2023. COVID-19 Treatment Guidelines. https://www.covid19treatmentguidelines.nih.gov/overview/sars-cov-2-testing/ [Accessed: 29 August 2023].

Newcomb, L.L., Kuo, R.-L., Ye, Q., Jiang, Y., Tao, Y.J. & Krug, R.M. (2009) Interaction of the influenza a virus nucleocapsid protein with the viral RNA polymerase potentiates unprimed viral RNA replication. *Journal of virology*. 83 (1), 29–36. doi:10.1128/JVI.02293-07.

NHS website (2023) *Coronavirus (COVID-19) vaccine*. 2023. https://www.nhs.uk/conditions/coronavirus-covid-19/coronavirus-vaccination/coronavirus-vaccine/?gclid=CjwKCAjwy42FBhB2EiwAJY0yQiifS6ixZmtt4xJjrRvyi7d3KFkzWecfbnIKSb1VOdg8xC01rrBZsxoC2PYQAvD_BwE [Accessed: 18 May 2021].

Nishiura, H., Oshitani, H., Kobayashi, T., Saito, T., Sunagawa, T., Matsui, T., Wakita, T., Suzuki, M. & MHLW COVID-19 Response Team (2020) Closed environments facilitate secondary transmission of coronavirus disease 2019 (COVID-19). *bioRxiv*. doi:10.1101/2020.02.28.20029272.

Noordzij, M., Dekker, F.W., Zoccali, C. & Jager, K.J. (2010) Measures of disease frequency: prevalence and incidence. *Nephron. Clinical practice*. 115 (1), c17–c20. doi:10.1159/000286345.

Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team, Dawood, F.S., Jain, S., Finelli, L., Shaw, M.W., Lindstrom, S., Garten, R.J., Gubareva, L.V., Xu, X., Bridges, C.B. & Uyeki, T.M. (2009) Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *The New England journal of medicine*. 360 (25), 2605–2615. doi:10.1056/NEJMoa0903810.

Nsoesie, E.O., Brownstein, J.S., Ramakrishnan, N. & Marathe, M.V. (2014) A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses*. 8 (3), 309–316. doi:10.1111/irv.12226.

Nsoesie, E.O., Oladeji, O., Abah, A.S.A. & Ndeffo-Mbah, M.L. (2021) Forecasting influenza-like illness trends in Cameroon using Google Search Data. *Scientific reports*. 11 (1), 1–11. doi:10.1038/s41598-021-85987-9.

Office for National Statistics (n.d.) *Annual Population Survey/Labour Force Survey*. https://www.nomisweb.co.uk/sources/aps [Accessed: 28 August 2023].

Office for National Statistics (2020a) *Coronavirus (COVID-19)*. 2020. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases [Accessed: 16 February 2022].

Office for National Statistics (2020b) *Coronavirus (COVID-19) Infection Survey*. 2020. https://www.ons.gov.uk/surveys/informationforhouseholdsandindividuals/householdandindividualsurveys/covid19infectionsurvey [Accessed: 27 August 2023].

Office for National Statistics (2023) Coronavirus (COVID-19) Infection Survey: methods and further information. 20 April 2023. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/methodologies/covid19infectionsurveypilotmethodsandfurtherinformation [Accessed: 4 January 2024].

Office for National Statistics (2022) *Estimates of the population for the UK, England, Wales, Scotland and Northern Ireland*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland.

Ogedegbe, G., Ravenell, J., Adhikari, S., Butler, M., Cook, T., Francois, F., Iturrate, E., Jean-Louis, G., Jones, S.A., Onakomaiya, D., Petrilli, C.M., Pulgarin, C., Regan, S., Reynolds, H., Seixas, A., Volpicelli, F.M. & Horwitz, L.I. (2020) Assessment of Racial/Ethnic Disparities in

Hospitalization and Mortality in Patients With COVID-19 in New York City. *JAMA network open*. 3 (12), e2026881. doi:10.1001/jamanetworkopen.2020.26881.

Ogunleye, A. & Wang, Q.-G. (2020) XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*. 17 (6), 2131–2140. doi:10.1109/TCBB.2019.2911071.

O'Halloran, A.C., Holstein, R., Cummings, C., Daily Kirley, P., Alden, N.B., et al. (2021) Rates of Influenza-Associated Hospitalization, Intensive Care Unit Admission, and In-Hospital Death by Race and Ethnicity in the United States From 2009 to 2019. *JAMA network open*. 4 (8), e2121880. doi:10.1001/jamanetworkopen.2021.21880.

Olson, D.R., Heffernan, R.T., Paladini, M., Konty, K., Weiss, D. & Mostashari, F. (2007) Monitoring the impact of influenza by age: emergency department fever and respiratory complaint surveillance in New York City. *PLoS medicine*. 4 (8), e247. doi:10.1371/journal.pmed.0040247.

OpenSAFELY (n.d.) *OpenSAFELY: Home*. https://www.opensafely.org/ [Accessed: 16 February 2022].

Opitz, D. & Maclin, R. (1999) Popular ensemble methods: an empirical study. *The journal of artificial intelligence research*. 11 (1), 169–198. https://dl.acm.org/doi/10.5555/3013545.3013549.

Ortiz, J.R., Neuzil, K.M., Shay, D.K., Rue, T.C., Neradilek, M.B., Zhou, H., Seymour, C.W., Hooper, L.G., Cheng, P.-Y., Goss, C.H. & Cooke, C.R. (2014) The burden of influenza-associated critical illness hospitalizations. *Critical care medicine*. 42 (11), 2325–2332. doi:10.1097/CCM.0000000000000545.

Osthus, D., Gattiker, J., Priedhorsky, R. & Del Valle, S.Y. (2019) Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy (with Discussion). *Bayesian Analysis*. 14 (1), 261–312. doi:10.1214/18-BA1117.

Padellini, T., Jersakova, R., Diggle, P.J., Holmes, C., King, R.E., Lehmann, B.C.L., Mallon, A.-M., Nicholson, G., Richardson, S. & Blangiardo, M. (2022) Time varying association between deprivation, ethnicity and SARS-CoV-2 infections in England: A population-based ecological study. *The Lancet regional health. Europe*. 15 (100322), 100322. doi:10.1016/j.lanepe.2022.100322.

Padmanabhan, R., Abed, H.S., Meskin, N., Khattab, T., Shraim, M. & Al-Hitmi, M.A. (2021) A review of mathematical model-based scenario analysis and interventions for COVID-19. *Computer methods and programs in biomedicine*. 209, 106301. doi:10.1016/j.cmpb.2021.106301.

Paireau, J., Andronico, A., Hozé, N., Layan, M., Crépey, P., Roumagnac, A., Lavielle, M., Boëlle, P.-Y. & Cauchemez, S. (2022) An ensemble model based on early predictors to forecast COVID-19 health care demand in France. *Proceedings of the National Academy of Sciences of the United States of America*. 119 (18), e2103302119. doi:10.1073/pnas.2103302119.

Palese, P. (2004) Influenza: old and new threats. *Nature medicine*. 10 (12 Suppl), S82–S87. doi:10.1038/nm1141.

Pang, I.K. & Iwasaki, A. (2012) Control of antiviral immunity by pattern recognition and the microbiome. *Immunological reviews*. 245 (1), 209–226. doi:10.1111/j.1600-065X.2011.01073.x.

Panja, M., Chakraborty, T., Nadim, S.S., Ghosh, I., Kumar, U. & Liu, N. (2023) An ensemble neural network approach to forecast Dengue outbreak based on climatic condition. *Chaos, Solitons & Fractals*. 167, 113124. doi:10.1016/j.chaos.2023.113124.

Panovska-Griffiths, J., Kerr, C.C., Waites, W. & Stuart, R.M. (2021) Chapter 10 - Mathematical modeling as a tool for policy decision making: Applications to the COVID-19 pandemic. In: A.S.R. Srinivasa Rao & C.R. Rao (eds.). *Handbook of Statistics*. Elsevier. pp. 291–326. doi:10.1016/bs.host.2020.12.001.

Pant, P.N. & Starbuck, W.H. (1990) Innocents in the Forest: Forecasting and Research Methods. *https://papers.ssrn.com › sol3 › papershttps://papers.ssrn.com › sol3 › papers*. https://papers.ssrn.com/abstract=2707885.

Park, A. & Iwasaki, A. (2020) Type I and Type III Interferons - Induction, Signaling, Evasion, and Application to Combat COVID-19. *Cell host & microbe*. 27 (6), 870–878. doi:10.1016/j.chom.2020.05.008.

Park, N. (2020) *Population estimates for the UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics*. 23 June 2020. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2019estimates [Accessed: 1 June 2021].

Park, S., Bekemeier, B., Flaxman, A. & Schultz, M. (2022) Impact of data visualization on decision-making and its implications for public health practice: a systematic literature review. *Informatics for health & social care*. 47 (2), 175–193. doi:10.1080/17538157.2021.1982949.

Parry, S. (2016) *91_ordlogistic.pdf*. https://cscu.cornell.edu/wp-content/uploads/91_ordlogistic.pdf.

Perkel, J.M. (2018) Data visualization tools drive interactivity and reproducibility in online publishing. *Nature*. 554 (7690), 133–134. doi:10.1038/d41586-018-01322-9.

Perra, N. & Gonçalves, B. (2015) Modeling and Predicting Human Infectious Diseases. *Social Phenomena*. 59. doi:10.1007/978-3-319-14011-7_4.

Perrotta, D., Tizzoni, M. & Paolotti, D. (2017) Using Participatory Web-based Surveillance Data to Improve Seasonal Influenza Forecasting in Italy. In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. 3 April 2017 Republic and Canton of Geneva, CHE, International World Wide Web Conferences Steering Committee. pp. 303–310. doi:10.1145/3038912.3052670.

Pflug, A., Lukarska, M., Resa-Infante, P., Reich, S. & Cusack, S. (2017) Structural insights into RNA synthesis by the influenza virus transcription-replication machine. *Virus research*. 234, 103–117. doi:10.1016/j.virusres.2017.01.013.

Pinter, G., Felde, I., Mosavi, A., Ghamisi, P. & Gloaguen, R. (2020) COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach. *Science in China, Series A: Mathematics*. 8 (6), 890. doi:10.3390/math8060890.

Poblador-Plou, B., Carmona-Pírez, J., Ioakeim-Skoufa, I., Poncel-Falcó, A., Bliek-Bueno, K., Cano-Del Pozo, M., Gimeno-Feliú, L.A., González-Rubio, F., Aza-Pascual-Salcedo, M., Bandrés-Liso, A.C., Díez-Manglano, J., Marta-Moreno, J., Mucherino, S., Gimeno-Miguel, A., Prados-Torres, A. & EpiChron Group (2020) Baseline Chronic Comorbidity and Mortality in Laboratory-Confirmed COVID-19 Cases: Results from the PRECOVID Study in Spain. *International journal of environmental research and public health*. 17 (14).

doi:10.3390/ijerph17145171.

Pople, D., Monk, E.J.M., Evans, S., Foulkes, S., Islam, J., Wellington, E., Atti, A., Hope, R., Robotham, J., Hopkins, S., Brown, C.S., Hall, V.J. & SIREN Study Group (2022) Burden of SARS-CoV-2 infection in healthcare workers during second wave in England and impact of vaccines: prospective multicentre cohort study (SIREN) and mathematical model. *BMJ* . 378, e070379. doi:10.1136/bmj-2022-070379.

Porte, L., Legarraga, P., Vollrath, V., Aguilera, X., Munita, J.M., Araos, R., Pizarro, G., Vial, P., Iruretagoyena, M., Dittrich, S. & Weitzel, T. (2020) Evaluation of a novel antigen-based rapid detection test for the diagnosis of SARS-CoV-2 in respiratory samples. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases*. 99, 328–333. doi:10.1016/j.ijid.2020.05.098.

Pouwels, K.B., House, T., Pritchard, E., Robotham, J.V., Birrell, P.J., et al. (2021) Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. *The Lancet. Public health*. 6 (1), e30–e38. doi:10.1016/S2468-2667(20)30282-6.

Price-Haywood, E.G., Burton, J., Fort, D. & Seoane, L. (2020) Hospitalization and Mortality among Black Patients and White Patients with Covid-19. *The New England journal of medicine*. 382 (26), 2534–2543. doi:10.1056/NEJMsa2011686.

Prime Minister's Office (2020a) *PM announces easing of lockdown restrictions: 23 June 2020*. 23 June 2020. GOV.UK. https://www.gov.uk/government/news/pm-announces-easing-of-lockdown-restrictions-23-june-2020 [Accessed: 7 August 2023].

Prime Minister's Office (2020b) *Prime Minister's statement on coronavirus (COVID-19): 19 December 2020*. 19 December 2020. GOV.UK. https://www.gov.uk/government/speeches/prime-ministers-statement-on-coronavirus-covid-19-19-december-2020 [Accessed: 3 August 2023].

Prime Minister's Office & Street, 10 Downing (2020) *Prime Minister's statement on coronavirus (COVID-19): 31 October 2020*. 31 October 2020. GOV.UK. https://www.gov.uk/government/speeches/prime-ministers-statement-on-coronavirus-covid-19-31-october-2020 [Accessed: 31 May 2021].

Public Health England (n.d.) *Disparities_in_the_risk_and_outcomes_of_COVID_August_2020_update.pdf*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/908434/Disparities_in_the_risk_and_outcomes_of_COVID_August_2020_update.pdf.

Puhach, O., Meyer, B. & Eckerle, I. (2023) SARS-CoV-2 viral load and shedding kinetics. *Nature reviews. Microbiology*. 21 (3), 147–161. doi:10.1038/s41579-022-00822-w.

Punn, N.S., Sonbhadra, S.K. & Agarwal, S. (2020) COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms. *medRxiv*.p.2020.04.08.20057679. doi:10.1101/2020.04.08.20057679.

Qin, R. (2022) The Construction of Corporate Financial Management Risk Model Based on XGBoost Algorithm. *Journal of Mathematics*. 2022. doi:10.1155/2022/2043369.

Rajaraman, S., Antani, S.K., Poostchi, M., Silamut, K., Hossain, M.A., Maude, R.J., Jaeger, S. & Thoma, G.R. (2018) Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*. 6, e4568. doi:10.7717/peerj.4568.

Rao, D.M., Chernyakhovsky, A. & Rao, V. (2009) Modeling and analysis of global epidemiology of avian influenza. *Environmental Modelling & Software*. 24 (1), 124–134. doi:10.1016/j.envsoft.2008.06.011.

REal-time Assessment of Community Transmission (REACT) Study (n.d.) *REACT- – Antigen (PCR) tests*. https://www.imperial.ac.uk/media/imperial-college/institute-of-global-health-innovation/REACT1-R19-REACT2-R6---Fieldwork-info-table.pdf [Accessed: 19 August 2023].

Reed, C., Chaves, S.S., Daily Kirley, P., Emerson, R., Aragon, D., Hancock, E.B., Butler, L., Baumbach, J., Hollick, G., Bennett, N.M., Laidler, M.R., Thomas, A., Meltzer, M.I. & Finelli, L. (2015) Estimating influenza disease burden from population-based surveillance data in the United States. *PloS one*. 10 (3), e0118369. doi:10.1371/journal.pone.0118369.

Reich, N.G., Brooks, L.C., Fox, S.J., Kandula, S., McGowan, C.J., Moore, E., Osthus, D., Ray, E.L., Tushar, A., Yamana, T.K., Biggerstaff, M., Johansson, M.A., Rosenfeld, R. & Shaman, J. (2019a) A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences of the United States of America*. 116 (8), 3146–3154. doi:10.1073/pnas.1812594116.

Reich, N.G., Lauer, S.A., Sakrejda, K., Iamsirithaworn, S., Hinjoy, S., Suangtho, P., Suthachana, S., Clapham, H.E., Salje, H., Cummings, D.A.T. & Lessler, J. (2016a) Challenges in Real-Time Prediction of Infectious Disease: A Case Study of Dengue in Thailand. *PLoS neglected tropical diseases*. 10 (6), e0004761. doi:10.1371/journal.pntd.0004761.

Reich, N.G., Lessler, J., Sakrejda, K., Lauer, S.A., Iamsirithaworn, S. & Cummings, D.A.T. (2016b) Case study in evaluating time series prediction models using the relative mean absolute error. *The American statistician*. 70 (3), 285–292. doi:10.1080/00031305.2016.1148631.

Reich, N.G., McGowan, C.J., Yamana, T.K., Tushar, A., Ray, E.L., Osthus, D., Kandula, S., Brooks, L.C., Crawford-Crudell, W., Gibson, G.C., Moore, E., Silva, R., Biggerstaff, M., Johansson, M.A., Rosenfeld, R. & Shaman, J. (2019b) Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLoS computational biology*. 15 (11), e1007486. doi:10.1371/journal.pcbi.1007486.

Resa-Infante, P., Jorba, N., Coloma, R. & Ortin, J. (2011) The influenza virus RNA synthesis machine: advances in its structure and function. *RNA biology*. 8 (2), 207–215. doi:10.4161/rna.8.2.14513.

Reveil, M. & Chen, Y.-H. (2022) Predicting and preventing COVID-19 outbreaks in indoor environments: an agent-based modeling study. *Scientific reports*. 12 (1), 16076. doi:10.1038/s41598-022-18284-8.

Richardson, S., Hirsch, J.S., Narasimhan, M., Crawford, J.M., McGinn, T., et al. (2020) Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA: the journal of the American Medical Association*. 323 (20), 2052–2059. doi:10.1001/jama.2020.6775.

Riley, S., Ainslie, K.E.C., Eales, O., Walters, C.E., Wang, H., Atchison, C., Fronterre, C., Diggle, P.J., Ashby, D., Donnelly, C.A., Cooke, G., Barclay, W., Ward, H., Darzi, A. & Elliott, P. (2020a) High and increasing prevalence of SARS-CoV-2 swab positivity in England during end September beginning October 2020: REACT-1 round 5 updated report. *bioRxiv*. doi:10.1101/2020.10.12.20211227.

Riley, S., Ainslie, K.E.C., Eales, O., Walters, C.E., Wang, H., Atchison, C., Fronterre, C., Diggle, P.J., Ashby, D., Donnelly, C.A., Cooke, G., Barclay, W., Ward, H., Darzi, A. & Elliott, P. (2021a) Resurgence of SARS-CoV-2: detection by community viral surveillance. *Science*. doi:10.1126/science.abf0874.

Riley, S., Atchison, C., Ashby, D., Donnelly, C.A., Barclay, W., Cooke, G., Ward, H., Darzi, A., Elliott, P. & REACT study group (2020b) REal-time Assessment of Community Transmission (REACT) of SARS-CoV-2 virus: Study protocol. *Wellcome Open Research*.5 p.200. doi:10.12688/wellcomeopenres.16228.1.

Riley, S., Fraser, C., Donnelly, C.A., Ghani, A.C., Abu-Raddad, L.J., et al. (2003) Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science*. 300 (5627), 1961–1966. doi:10.1126/science.1086478.

Riley, S., Haw, D., Walters, C.E., Wang, H., Eales, O., et al. (2021b) REACT-1 round 11 report: low prevalence of SARS-CoV-2 infection in the community prior to the third step of the English roadmap out of lockdown. *bioRxiv*. doi:10.1101/2021.05.13.21257144.

Riley, S., Wang, H., Eales, O., Haw, D., Walters, C.E., et al. (2021c) REACT-1 round 12 report: resurgence of SARS-CoV-2 infections in England associated with increased frequency of the Delta variant. *bioRxiv*. doi:10.1101/2021.06.17.21259103.

Rodriguez-Diaz, C.E., Guilamo-Ramos, V., Mena, L., Hall, E., Honermann, B., Crowley, J.S., Baral, S., Prado, G.J., Marzan-Rodriguez, M., Beyrer, C., Sullivan, P.S. & Millett, G.A. (2020) Risk for COVID-19 infection and death among Latinos in the United States: examining heterogeneity in transmission dynamics. *Annals of epidemiology*. 52, 46–53.e2. doi:10.1016/j.annepidem.2020.07.007.

Rodriguez-Morales, A.J., Cardona-Ospina, J.A., Gutiérrez-Ocampo, E., Villamizar-Peña, R., Holguin-Rivera, Y., et al. (2020) Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis. *Travel medicine and infectious disease*. 34, 101623. doi:10.1016/j.tmaid.2020.101623.

Safarishahrbijari, A., Teyhouee, A., Waldner, C., Liu, J. & Osgood, N.D. (2017) Predictive accuracy of particle filtering in dynamic models supporting outbreak projections. *BMC infectious diseases*. 17 (1), 648. doi:10.1186/s12879-017-2726-9.

Sahai, S.Y., Gurukar, S., KhudaBukhsh, W.R., Parthasarathy, S. & Rempała, G.A. (2022) A machine learning model for nowcasting epidemic incidence. *Mathematical biosciences*. 343, 108677. doi:10.1016/j.mbs.2021.108677.

Saif, S., Das, P. & Biswas, S. (2021) A Hybrid Model based on mBA-ANFIS for COVID-19 Confirmed Cases Prediction and Forecast. *Journal of The Institution of Engineers (India): Series B*. 102 (6), 1123–1136. doi:10.1007/s40031-021-00538-0.

Salje, H., Tran Kiem, C., Lefrancq, N., Courtejoie, N., Bosetti, P., Paireau, J., Andronico, A., Hozé, N., Richet, J., Dubost, C.-L., Le Strat, Y., Lessler, J., Levy-Bruhl, D., Fontanet, A., Opatowski, L., Boelle, P.-Y. & Cauchemez, S. (2020) Estimating the burden of SARS-CoV-2 in France. *Science*. 369 (6500), 208–211. doi:10.1126/science.abc3517.

Santosh, K.C. & Gaur, L. (2021) *Artificial Intelligence and Machine Learning in Public Healthcare*. Springer Nature Singapore. doi:10.1007/978-981-16-6768-8.

Sarkar, J. & Chakrabarti, P. (2020) A machine learning model reveals older age and delayed hospitalization as predictors of mortality in patients with COVID-19. *bioRxiv*. doi:10.1101/2020.03.25.20043331.

Sarker, S., Jamal, L., Ahmed, S.F. & Irtisam, N. (2021) Robotics and artificial intelligence in healthcare during COVID-19 pandemic: A systematic review. *Robotics and autonomous systems*. 146, 103902. doi:10.1016/j.robot.2021.103902.

Sattar, N., McInnes, I.B. & McMurray, J.J.V. (2020) Obesity Is a Risk Factor for Severe COVID-19 Infection: Multiple Potential Mechanisms. *Circulation*. 142 (1), 4–6. doi:10.1161/CIRCULATIONAHA.120.047659.

Scharff, D.P., Mathews, K.J., Jackson, P., Hoffsuemmer, J., Martin, E. & Edwards, D. (2010) More than Tuskegee: understanding mistrust about research participation. *Journal of health care for the poor and underserved*. 21 (3), 879–897. doi:10.1353/hpu.0.0323.

Schnaars, S.P. (1986) Long-range forecasting: From crystal ball to computer: J. Scott Armstrong, 2nd ed. (Wiley, New York, 1985) [UK pound]22.95 (paper), pp. 689. *International journal of forecasting*. 2 (3), 387–390. https://EconPapers.repec.org/RePEc:eee:intfor:v:2:y:1986:i:3:p:387-390.

Schoeman, D. & Fielding, B.C. (2019) Coronavirus envelope protein: current knowledge. *Virology journal*. 16 (1), 69. doi:10.1186/s12985-019-1182-0.

Scott Armstrong, J. (2011) *Principles of Forecasting*. Springer US. doi:10.1007/978-0-306-47630-3.

Sederdahl, B.K. & Williams, J.V. (2020) Epidemiology and Clinical Characteristics of Influenza C Virus. *Viruses*. 12 (1). doi:10.3390/v12010089.

Serfling, R.E. (1963) Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports*. 78 (6), 494–506. https://www.ncbi.nlm.nih.gov/pubmed/19316455.

Sewell, H.F., Agius, R.M., Kendrick, D. & Stewart, M. (2020) Covid-19 vaccines: delivering protective immunity. *BMJ* . 371, m4838. doi:10.1136/bmj.m4838.

Sex, Gender and Covid-19 Project (n.d.) *Men, sex, gender and COVID-19*. https://globalhealth5050.org/the-sex-gender-and-covid-19-project/men-sex-gender-and-covid-19/ [Accessed: 21 February 2022].

Shaghaghi, N., Calle, A., Kouretas, G., Mirchandani, J. & Castillo, M. (2021) eVision: Epidemic Forecasting on COVID-19. *Current Directions in Biomedical Engineering*. 7 (2), 839–842. doi:10.1515/cdbme-2021-2214.

Shaman, J. & Karspeck, A. (2012) Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences of the United States of America*. 109 (50), 20425–20430. doi:10.1073/pnas.1208772109.

Shaman, J., Karspeck, A., Yang, W., Tamerius, J. & Lipsitch, M. (2013) Real-time influenza forecasts during the 2012–2013 season. *Nature communications*. 4 (1), 1–10. doi:10.1038/ncomms3837.

Shaman, J., Pitzer, V.E., Viboud, C., Grenfell, B.T. & Lipsitch, M. (2010) Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS biology*. 8 (2), e1000316. doi:10.1371/journal.pbio.1000316.

Sharot, T. (1986) *Weighting survey results*. 1986. http://redresearch.com/wp/wp-content/uploads/2016/01/Weighting-Survey-Results.pdf [Accessed: 1 June 2021].

Shereen, M.A., Khan, S., Kazmi, A., Bashir, N. & Siddique, R. (2020) COVID-19 infection:

Origin, transmission, and characteristics of human coronaviruses. *Journal of advertising research*. 24, 91–98. doi:10.1016/j.jare.2020.03.005.

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y. & Shen, D. (2021) Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. *IEEE reviews in biomedical engineering*. 14, 4–15. doi:10.1109/RBME.2020.2987975.

Shor, B., Bafumi, J., Keele, L. & Park, D. (2007) A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data. *Political analysis: an annual publication of the Methodology Section of the American Political Science Association*. 15 (2), 165–181. doi:10.1093/pan/mpm006.

Simonsen, L., Clarke, M.J., Williamson, G.D., Stroup, D.F., Arden, N.H. & Schonberger, L.B. (1997) The impact of influenza epidemics on mortality: introducing a severity index. *American journal of public health*. 87 (12), 1944–1950. doi:10.2105/ajph.87.12.1944.

Simonsen, L., Reichert, T.A., Viboud, C., Blackwelder, W.C., Taylor, R.J. & Miller, M.A. (2005) Impact of influenza vaccination on seasonal mortality in the US elderly population. *Archives of internal medicine*. 165 (3), 265–272. doi:10.1001/archinte.165.3.265.

Singh, S. & Kaur, H. (2021) Influenza prediction from social media texts using machine learning. *Journal of physics. Conference series*. 1950 (1), 012018. doi:10.1088/1742-6596/1950/1/012018.

Si, Y., Trangucci, R., Gabry, J.S. & Gelman, A. (2017) Bayesian hierarchical weighting adjustment and survey inference. *arXiv [stat.ME]*. http://arxiv.org/abs/1707.08220.

Skehel, J.J. & Wiley, D.C. (2000) Receptor Binding and Membrane Fusion in Virus Entry: The Influenza Hemagglutinin. *Annual review of biochemistry*. 69 (1), 531–569. doi:10.1146/annurev.biochem.69.1.531.

Soebiyanto, R.P., Adimi, F. & Kiang, R.K. (2010) Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PloS one*. 5 (3), e9450. doi:10.1371/journal.pone.0009450.

Soebiyanto, R.P., Clara, W., Jara, J., Castillo, L., Sorto, O.R., Marinero, S., de Antinori, M.E.B., McCracken, J.P., Widdowson, M.-A., Azziz-Baumgartner, E. & Kiang, R.K. (2014) The Role of Temperature and Humidity on Seasonal Influenza in Tropical Areas: Guatemala, El Salvador and Panama, 2008–2013. *PloS one*. 9 (6), e100659. doi:10.1371/journal.pone.0100659.

Souza, G.N. de, Jr, Mendes, A.G.B., Costa, J.D.S., Oliveira, M.D.S., Lima, P.V.C., de Moraes, V.N., Silva, D.C.C., Rocha, J.E.C. da, Botelho, M. do N., Araujo, F.A., Fernandes, R. da S., Souza, D.L. & Braga, M. de B. (2023) Deep learning framework for epidemiological forecasting: A study on COVID-19 cases and deaths in the Amazon state of Pará, Brazil. *PloS one*. 18 (11), e0291138. doi:10.1371/journal.pone.0291138.

Soyiri, I.N. & Reidpath, D.D. (2013) An overview of health forecasting. *Environmental health and preventive medicine*. 18 (1), 1–9. doi:10.1007/s12199-012-0294-6.

Spaeder, M.C., Stroud, J.R. & Song, X. (2012) Time-series model to predict impact of H1N1 influenza on a children's hospital. *Epidemiology and infection*. 140 (5), 798–802. doi:10.1017/S0950268811001749.

Ssentongo, P., Ssentongo, A.E., Heilbrunn, E.S., Ba, D.M. & Chinchilli, V.M. (2020)

Association of cardiovascular disease and 10 other pre-existing comorbidities with COVID-19 mortality: A systematic review and meta-analysis. *PloS one*. 15 (8), e0238215. doi:10.1371/journal.pone.0238215.

Stadnytskyi, V., Bax, C.E., Bax, A. & Anfinrud, P. (2020) The airborne lifetime of small speech droplets and their potential importance in SARS-CoV-2 transmission. *Proceedings of the National Academy of Sciences*. 117 (22), 11875–11877. doi:10.1073/pnas.2006874117.

Starr, T.N., Greaney, A.J., Addetia, A., Hannon, W.W., Choudhary, M.C., Dingens, A.S., Li, J.Z. & Bloom, J.D. (2021) Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science*. 371 (6531), 850–854. doi:10.1126/science.abf9302.

Steel, J. & Lowen, A.C. (2014) Influenza A virus reassortment. *Current topics in microbiology and immunology*. 385, 377–401. doi:10.1007/82_2014_395.

Sturniolo, S., Waites, W., Colbourn, T., Manheim, D. & Panovska-Griffiths, J. (2021) Testing, tracing and isolation in compartmental models. *PLoS computational biology*. 17 (3), e1008633. doi:10.1371/journal.pcbi.1008633.

Sutherland, E., Headicar, J. & Delong, P. (2021) *Coronavirus (COVID-19) Infection Survey technical article - Office for National Statistics*. 28 June 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsurveytechnicalarticle/wavesandlagsofcovid19inenglandjune2021 [Accessed: 18 July 2023].

Swain, S.L., McKinstry, K.K. & Strutt, T.M. (2012) Expanding roles for CD4+ T cells in immunity to viruses. *Nature reviews. Immunology*. 12 (2), 136–148. doi:10.1038/nri3152.

Swanson, N.R. & White, H. (1997) Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International journal of forecasting*. 13 (4), 439–461. doi:10.1016/S0169-2070(97)00030-7.

Tabataba, F.S., Chakraborty, P., Ramakrishnan, N., Venkatramanan, S., Chen, J., Lewis, B. & Marathe, M. (2017) A framework for evaluating epidemic forecasts. *BMC infectious diseases*. 17 (1), 345. doi:10.1186/s12879-017-2365-1.

Tamerius, J.D., Shaman, J., Alonso, W.J., Bloom-Feshbach, K., Uejio, C.K., Comrie, A. & Viboud, C. (2013) Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS pathogens*. 9 (3), e1003194. doi:10.1371/journal.ppat.1003194.

Tamerius, J., Nelson, M.I., Zhou, S.Z., Viboud, C., Miller, M.A. & Alonso, W.J. (2011) Global influenza seasonality: reconciling patterns across temperate and tropical regions. *Environmental health perspectives*. 119 (4), 439–445. doi:10.1289/ehp.1002383.

Tang, L., Zhou, Y., Wang, L., Purkayastha, S., Zhang, L., He, J., Wang, F. & Song, P.X.-K. (2020) A Review of Multi-Compartment Infectious Disease Models. *International statistical review = Revue internationale de statistique*. 88 (2), 462–513. doi:10.1111/insr.12402.

Tashman, L.J. (2000) Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*. 16 (4), 437–450. https://EconPapers.repec.org/RePEc:eee:intfor:v:16:y:2000:i:4:p:437-450.

Taubenberger, J.K. & Morens, D.M. (2008) The pathology of influenza virus infections. *Annual review of pathology*. 3, 499–522. doi:10.1146/annurev.pathmechdis.3.121806.154316.

Tay, M.Z., Poh, C.M., Rénia, L., MacAry, P.A. & Ng, L.F.P. (2020) The trinity of COVID-19: immunity, inflammation and intervention. *Nature reviews. Immunology*. 20 (6), 363–374. doi:10.1038/s41577-020-0311-8.

Ten Have, T.R. (1996) A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics*. 52 (2), 473–491. https://www.ncbi.nlm.nih.gov/pubmed/8672699.

Thacker, S.B. & Stroup, D.F. (1994) Future directions for comprehensive public health surveillance and health information systems in the United States. *American journal of epidemiology*. 140 (5), 383–397. doi:10.1093/oxfordjournals.aje.a117261.

Thompson, W.W., Comanor, L. & Shay, D.K. (2006) Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease. *The Journal of infectious diseases*. 194 Suppl 2, S82–S91. doi:10.1086/507558.

Thompson, W.W., Shay, D.K., Weintraub, E., Brammer, L., Cox, N., Anderson, L.J. & Fukuda, K. (2003) Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA: the journal of the American Medical Association*. 289 (2), 179–186. doi:10.1001/jama.289.2.179.

Torner, N., Basile, L., Martínez, A., Rius, C., Godoy, P., Jané, M., Domínguez, Á. & Working Group on PIDIRAC Sentinel Surveillance of Catalonia (2019) Assessment of two complementary influenza surveillance systems: sentinel primary care influenza-like illness versus severe hospitalized laboratory-confirmed influenza using the moving epidemic method. *BMC public health*. 19 (1), 1089. doi:10.1186/s12889-019-7414-9.

Trionfetti, F. (2023) *Mechanisms of response of Mesothelial Cells to viral infections: role of epigenetic HDAC1/2 regulation*. iris.uniroma1.it. https://iris.uniroma1.it/handle/11573/1679488.

UK Health Security Agency (2020a) *England Summary*. 2020. https://coronavirus.data.gov.uk/ [Accessed: 20 May 2021].

UK Health Security Agency (2020b) *GOV.UK Coronavirus (COVID-19) in the UK*. 2020. https://coronavirus.data.gov.uk/details/cases [Accessed: Augist 08 2023].

UK Health Security Agency (2020c) *GOV.UK Coronavirus (COVID-19) in the UK*. 2020. https://coronavirus.data.gov.uk/about [Accessed: 8 August 2023].

UKHSA (n.d.) *UKHSA data dashboard*. UKHSA data dashboard. https://ukhsa-dashboard.data.gov.uk/ [Accessed: 14 January 2024].

Ullah, R., Al Mdallal, Q., Khan, T., Ullah, R., Al Alwan, B., Faiz, F. & Zhu, Q. (2023) The dynamics of novel corona virus disease via stochastic epidemiological model with vaccination. *Scientific reports*. 13 (1), 3805. doi:10.1038/s41598-023-30647-3.

Uyeki, T.M., Bernstein, H.H., Bradley, J.S., Englund, J.A., File, T.M., Fry, A.M., Gravenstein, S., Hayden, F.G., Harper, S.A., Hirshon, J.M., Ison, M.G., Johnston, B.L., Knight, S.L., McGeer, A., Riley, L.E., Wolfe, C.R., Alexander, P.E. & Pavia, A.T. (2019) Clinical Practice Guidelines by the Infectious Diseases Society of America: 2018 Update on Diagnosis, Treatment, Chemoprophylaxis, and Institutional Outbreak Management of Seasonal Influenzaa. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*. 68 (6), e1–e47. doi:10.1093/cid/ciy866.

Vahidy, F.S., Pan, A.P., Ahnstedt, H., Munshi, Y., Choi, H.A., Tiruneh, Y., Nasir, K., Kash,

B.A., Andrieni, J.D. & McCullough, L.D. (2021) Sex differences in susceptibility, severity, and outcomes of coronavirus disease 2019: Cross-sectional analysis from a diverse US metropolitan area. *PloS one*. 16 (1), e0245556. doi:10.1371/journal.pone.0245556.

Van Vinh Chau, N., Lam, V.T., Dung, N.T., Yen, L.M., Minh, N.N.Q., et al. (2020) The Natural History and Transmission Potential of Asymptomatic Severe Acute Respiratory Syndrome Coronavirus 2 Infection. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*. 71 (10), 2679–2687. doi:10.1093/cid/ciaa711.

Varghese, A., Kolamban, S., Sherimon, V., Lacap, E.M., Jr, Ahmed, S.S., Sreedhar, J.P., Al Harthi, H. & Al Shuaily, H.S. (2021a) SEAMHCRD deterministic compartmental model based on clinical stages of infection for COVID-19 pandemic in Sultanate of Oman. *Scientific reports*. 11 (1), 11984. doi:10.1038/s41598-021-91114-5.

Varghese, N.E., Sabat, I., Neumann-Böhme, S., Schreyögg, J., Stargardt, T., Torbica, A., van Exel, J., Barros, P.P. & Brouwer, W. (2021b) Risk communication during COVID-19: A descriptive study on familiarity with, adherence to and trust in the WHO preventive measures. *PloS one*. 16 (4), e0250872. doi:10.1371/journal.pone.0250872.

Vega, T., Lozano, J.E., Meerhoff, T., Snacken, R., Mott, J., Ortiz de Lejarazu, R. & Nunes, B. (2013) Influenza surveillance in Europe: establishing epidemic thresholds by the moving epidemic method. *Influenza and other respiratory viruses*. 7 (4), 546–558. doi:10.1111/j.1750-2659.2012.00422.x.

Venna, S.R., Tavanaei, A., Gottumukkala, R.N., Raghavan, V.V., Maida, A.S. & Nichols, S. (2019) A Novel Data-Driven Model for Real-Time Influenza Forecasting. *IEEE Access*. 7, 7691–7701. doi:10.1109/ACCESS.2018.2888585.

Verikios, G., McCaw, J.M., McVernon, J. & Harris, A.H. (2012) H1N1 influenza and the Australian macroeconomy. *Journal of the Asia Pacific Economy*. 17 (1), 22–51. doi:10.1080/13547860.2012.639999.

Verity, R., Okell, L.C., Dorigatti, I., Winskill, P., Whittaker, C., et al. (2020) Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*. 20 (6), 669–677. doi:10.1016/S1473-3099(20)30243-7.

Viboud, C., Bjørnstad, O.N., Smith, D.L., Simonsen, L., Miller, M.A. & Grenfell, B.T. (2006) Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*. 312 (5772), 447–451. doi:10.1126/science.1125237.

Viboud, C. & Vespignani, A. (2019) The future of influenza forecasts. *Proceedings of the National Academy of Sciences of the United States of America*.116 (8) pp.2802–2804. doi:10.1073/pnas.1822167116.

Viechtbauer, W. (2010) Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software, Articles*. 36 (3), 1–48. doi:10.18637/jss.v036.i03.

Vijaykrishna, D., Bahl, J., Riley, S., Duan, L., Zhang, J.X., Chen, H., Peiris, J.S.M., Smith, G.J.D. & Guan, Y. (2008) Evolutionary dynamics and emergence of panzootic H5N1 influenza viruses. *PLoS pathogens*. 4 (9), e1000161. doi:10.1371/journal.ppat.1000161.

Virtual meeting (2022) *WHO consultation to adapt influenza sentinel surveillance systems to include COVID-19 virological surveillance*. 2022. https://apps.who.int/iris/bitstream/handle/10665/356309/9789240050426-eng.pdf?sequence=1 [Accessed: 17 August 2023].

Viziflu (2018) *Viziflu: An open-source visualization tool to help make multi-model seasonal influenza forecasts more actionable for decision-makers*. Github. https://github.com/IQTLabs/Viziflu.

V'kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. (2021) Coronavirus biology and replication: implications for SARS-CoV-2. *Nature reviews. Microbiology*. 19 (3), 155–170. doi:10.1038/s41579-020-00468-6.

Volz, E., Mishra, S., Chand, M., Barrett, J.C., Johnson, R., et al. (2021) Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*. 593 (7858), 266–269. doi:10.1038/s41586-021-03470-x.

Walker, P.G.T., Whittaker, C., Watson, O.J., Baguelin, M., Winskill, P., et al. (2020) The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. *Science*. 369 (6502), 413–422. doi:10.1126/science.abc0035.

Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T. & Veesler, D. (2020) Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*. 181 (2), 281–292.e6. doi:10.1016/j.cell.2020.02.058.

Wang, B., Li, R., Lu, Z. & Huang, Y. (2020a) Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis. *Aging*. 12 (7), 6049–6057. doi:10.18632/aging.103000.

Wang, H., Ainslie, K.E.C., Walters, C.E., Eales, O., Haw, D., Atchison, C., Fronterre, C., Diggle, P.J., Ashby, D., Cooke, G., Barclay, W., Ward, H., Darzi, A., Donnelly, C.A., Riley, S. & Elliott, P. (2023) Health inequalities in SARS-CoV-2 infection during the second wave in England: REACT-1 study. *bioRxiv*. doi:10.1101/2023.08.01.23293491.

Wang, H., Kwok, K.O. & Riley, S. (2023) Forecasting influenza incidence as an ordinal variable using machine learning. *medRxiv*. doi:10.1101/2023.02.09.23285705.

Wang, H., Kwok, K.O. & Riley, S. (2024) Forecasting influenza incidence as an ordinal variable using machine learning[version 1; peer review: awaiting peer review]. *Wellcome Open Res*. 9 (11). doi:10.12688/wellcomeopenres.19599.1.

Wang, J., Cheng, Q. & Dong, Y. (2022) An XGBoost-based multivariate deep learning framework for stock index futures price forecasting. *Kybernetes. The International Journal of Cybernetics, Systems and Management Sciences*. ahead-of-print (ahead-of-print). doi:10.1108/K-12-2021-1289.

Wang, X., Li, Y., O'Brien, K.L., Madhi, S.A., Widdowson, M.-A., et al. (2020b) Global burden of respiratory infections associated with seasonal influenza in children under 5 years in 2018: a systematic review and modelling study. *The Lancet. Global health*. 8 (4), e497–e510. doi:10.1016/S2214-109X(19)30545-5.

Ward, H., Cooke, G.S., Atchison, C., Whitaker, M., Elliott, J., Moshe, M., Brown, J.C., Flower, B., Daunt, A., Ainslie, K., Ashby, D., Donnelly, C.A., Riley, S., Darzi, A., Barclay, W. & Elliott, P. (2021) Prevalence of antibody positivity to SARS-CoV-2 following the first peak of infection in England: Serial cross-sectional studies of 365,000 adults. *The Lancet regional health. Europe*. 4, 100098. doi:10.1016/j.lanepe.2021.100098.

Warshaw, C. & Rodden, J. (2012) How Should We Measure District-Level Public Opinion on Individual Issues? *The journal of politics*. 74 (1), 203–219. doi:10.1017/s0022381611001204.

Weisblum, Y., Schmidt, F., Zhang, F., DaSilva, J., Poston, D., et al. (2020) Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife*. 9. doi:10.7554/eLife.61312.

Weissgerber, T.L., Garovic, V.D., Savic, M., Winham, S.J. & Milic, N.M. (2016) From Static to Interactive: Transforming Data Visualization to Improve Transparency. *PLoS biology*. 14 (6), e1002484. doi:10.1371/journal.pbio.1002484.

Wernstedt, K., Roberts, P.S., Arvai, J. & Redmond, K. (2019) How emergency managers (mis?)interpret forecasts. *Disasters*. 43 (1), 88–109. doi:10.1111/disa.12293.

Wieczorek, M., Siłka, J., Połap, D., Woźniak, M. & Damaševičius, R. (2020) Real-time neural network based predictor for cov19 virus spread. *PloS one*. 15 (12), e0243189. doi:10.1371/journal.pone.0243189.

Wilder-Smith, A. & Freedman, D.O. (2020) Isolation, quarantine, social distancing and community containment: pivotal role for old-style public health measures in the novel coronavirus (2019-nCoV) outbreak. *Journal of travel medicine*. 27 (2). doi:10.1093/jtm/taaa020.

Wilke, C.O. (2019) *Fundamentals of Data Visualization*. 14 May 2019. https://clauswilke.com/dataviz/ [Accessed: 15 August 2023].

Williamson, E.J., Walker, A.J., Bhaskaran, K., Bacon, S., Bates, C., et al. (2020) Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 584 (7821), 430–436. doi:10.1038/s41586-020-2521-4.

Windsor-Shellard, B. & Nasir, R. (2021) *Coronavirus (COVID-19) related deaths by occupation, England and Wales - Office for National Statistics*. 25 January 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/bulletins/coronaviruscovid19relateddeathsbyoccupationenglandandwales/deathsregisteredbetween9marchand28december2020 [Accessed: 5 April 2023].

Wong, K.C.-Y., Xiang, Y., Yin, L. & So, H.-C. (2021) Uncovering Clinical Risk Factors and Predicting Severe COVID-19 Cases Using UK Biobank Data: Machine Learning Approach. *JMIR public health and surveillance*. 7 (9), e29544. doi:10.2196/29544.

World Health Organization (2021) *Ethics and governance of artificial intelligence for health*. 28 June 2021. https://www.who.int/publications/i/item/9789240029200 [Accessed: 3 May 2023].

World Health Organization (n.d.) *FluID*. https://www.who.int/teams/global-influenza-programme/surveillance-and-monitoring/fluid [Accessed: 28 July 2022a].

World Health Organization (n.d.) *FluID*. https://www.who.int/teams/global-influenza-programme/surveillance-and-monitoring/fluid [Accessed: 24 August 2023b].

World Health Organization (2019) *Global influenza strategy 2019-2030*. World Health Organization (WHO). https://apps.who.int/iris/handle/10665/311184.

World Health Organization (2020a) *Global surveillance for COVID-19 caused by human infection with COVID-19 virus: interim guidance, 20 March 2020*. https://apps.who.int/iris/handle/10665/331506.

World Health Organization (n.d.) *Influenza & COVID19*. https://www.who.int/teams/global-influenza-programme/influenza-covid19 [Accessed: 21 July 2023c].

World Health Organization (2023a) *Influenza (Seasonal)*. 12 January 2023. https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal) [Accessed: 14 July 2023].

World Health Organization (2023b) *Influenza (Seasonal)*. 12 January 2023. https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal) [Accessed: 9 March 2022].

World Health Organization (2020b) *Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)*. 30 January 2020. https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) [Accessed: 18 July 2023].

World Health Organization (2020c) *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. 11 March 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 [Accessed: 3 August 2023].

World Health Organization. Regional Office for South-East Asia (2002) *Data management for evidence based decision making*. https://apps.who.int/iris/handle/10665/204907.

World Health Organization (WHO) (2020) *WHO Coronavirus (COVID-19) Dashboard*. 2020. https://covid19.who.int/ [Accessed: 11 January 2023].

Wright, G. & Goodwin, P. (2009) Decision making and planning under low levels of predictability: Enhancing the scenario method. *International journal of forecasting*. 25 (4), 813–825. doi:10.1016/j.ijforecast.2009.05.019.

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Tan, W., Cheng, G. & Jiang, T. (2020a) Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell host & microbe*. 27 (3), 325–328. doi:10.1016/j.chom.2020.02.001.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., et al. (2020b) A new coronavirus associated with human respiratory disease in China. *Nature*. 579 (7798), 265–269. doi:10.1038/s41586-020-2008-3.

Wu, Y. (2018) *DL4Epi: Deep Learning for Epidemiological Predictions*. Github. doi:10.1145/3209978.3210077.

Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., Du, M. & Liu, M. (2020c) Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. *The Science of the total environment*. 729, 139051. doi:10.1016/j.scitotenv.2020.139051.

Wu, Z. & McGoogan, J.M. (2020) Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA: the journal of the American Medical Association*. 323 (13), 1239–1242. doi:10.1001/jama.2020.2648.

Wynants, L., Van Calster, B., Collins, G.S., Riley, R.D., Heinze, G., et al. (2020) Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* . 369, m1328. doi:10.1136/bmj.m1328.

Xia, S., Wang, L., Zhu, Y., Lu, L. & Jiang, S. (2022) Origin, virological features, immune evasion and intervention of SARS-CoV-2 Omicron sublineages. *Signal transduction and targeted therapy*. 7 (1), 241. doi:10.1038/s41392-022-01105-9.

Xiong, C., Hu, S., Yang, M., Luo, W. & Zhang, L. (2020) Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences of the United States of America*. 117 (44), 27087–27089. doi:10.1073/pnas.2010836117.

Xu, Q., Gel, Y.R., Leticia Ramirez Ramirez, L., Nezafati, K., Zhang, Q. & Tsui, K.-L. (2017) Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. *PloS one*. 12 (5), e0176690. doi:10.1371/journal.pone.0176690.

Yadav, P.D., Sapkal, G.N., Abraham, P., Ella, R., Deshpande, G., Patil, D.Y., Nyayanit, D.A., Gupta, N., Sahay, R.R., Shete, A.M., Panda, S., Bhargava, B. & Mohan, V.K. (2022) Neutralization of Variant Under Investigation B.1.617.1 With Sera of BBV152 Vaccinees. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*. 74 (2), 366–368. doi:10.1093/cid/ciab411.

Yamey, G., Schäferhoff, M., Hatchett, R., Pate, M., Zhao, F. & McDade, K.K. (2020) Ensuring global access to COVID-19 vaccines. *The Lancet*. 395 (10234), 1405–1406. doi:10.1016/S0140-6736(20)30763-7.

Yang, L., Li, G., Yang, J., Zhang, T., Du, J., Liu, T., Zhang, X., Han, X., Li, W., Ma, L., Feng, L. & Yang, W. (2023) Deep-Learning Model for Influenza Prediction From Multisource Heterogeneous Data in a Megacity: Model Development and Evaluation. *Journal of medical Internet research*. 25, e44238. doi:10.2196/44238.

Yang, W., Cowling, B.J., Lau, E.H.Y. & Shaman, J. (2015) Forecasting Influenza Epidemics in Hong Kong. *PLoS computational biology*. 11 (7), e1004383. doi:10.1371/journal.pcbi.1004383.

Yang, W., Lau, E.H.Y. & Cowling, B.J. (2020) Dynamic interactions of influenza viruses in Hong Kong during 1998-2018. *PLoS computational biology*. 16 (6), e1007989. doi:10.1371/journal.pcbi.1007989.

Yang, W., Lipsitch, M. & Shaman, J. (2015) Inference of seasonal and pandemic influenza transmission dynamics. *Proceedings of the National Academy of Sciences*. 112 (9), 2723–2728. doi:10.1073/pnas.1415012112.

Yang, Y., Sugimoto, J.D., Halloran, M.E., Basta, N.E., Chao, D.L., Matrajt, L., Potter, G., Kenah, E. & Longini, I.M., Jr (2009) The transmissibility and control of pandemic influenza A (H1N1) virus. *Science*. 326 (5953), 729–733. doi:10.1126/science.1177373.

York, A., Hengrung, N., Vreede, F.T., Huiskonen, J.T. & Fodor, E. (2013) Isolation and characterization of the positive-sense replicative intermediate of a negative-strand RNA virus. *Proceedings of the National Academy of Sciences of the United States of America*. 110 (45), E4238–E4245. doi:10.1073/pnas.1315068110.

Young-Xu, Y., van Aalst, R., Russo, E., Lee, J.K.H. & Chit, A. (2017) The Annual Burden of Seasonal Influenza in the US Veterans Affairs Population. *PloS one*. 12 (1), e0169344. doi:10.1371/journal.pone.0169344.

Yuan, H., Kramer, S.C., Lau, E.H.Y., Cowling, B.J. & Yang, W. (2021) Modeling influenza seasonality in the tropics and subtropics. *PLoS computational biology*. 17 (6), e1009050. doi:10.1371/journal.pcbi.1009050.

Yu, M., Yang, C. & Li, Y. (2018) Big Data in Natural Disaster Management: A Review. *Geosciences Journal*. 8 (5), 165. doi:10.3390/geosciences8050165.

Zaki, A.M., van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D.M.E. & Fouchier, R.A.M. (2012) Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *The New England journal of medicine*. 367 (19), 1814–1820. doi:10.1056/NEJMoa1211721.

Zhang, G. & Liu, X. (2021) Prediction and control of COVID-19 spreading based on a hybrid intelligent model. *PloS one*. 16 (2), e0246360. doi:10.1371/journal.pone.0246360.

Zhang, J.-J., Dong, X., Cao, Y.-Y., Yuan, Y.-D., Yang, Y.-B., Yan, Y.-Q., Akdis, C.A. & Gao, Y.-D. (2020) Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. *Allergy*. 75 (7), 1730–1741. doi:10.1111/all.14238.

Zhang, Q., Perra, N., Perrotta, D., Tizzoni, M., Paolotti, D. & Vespignani, A. (2017) Forecasting Seasonal Influenza Fusing Digital Indicators and a Mechanistic Disease Model. In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. 3 April 2017 Republic and Canton of Geneva, CHE, International World Wide Web Conferences Steering Committee. pp. 311–319. doi:10.1145/3038912.3052678.

Zhang, X., Liu, H., Tang, H., Zhang, M., Yuan, X. & Shen, X. (2021) The effect of population size for pathogen transmission on prediction of COVID-19 spread. *Scientific reports*. 11 (1), 18024. doi:10.1038/s41598-021-97578-9.

Zhang, X.-X., Deng, T. & Jia, G.-Z. (2020) Nuclear spin-spin coupling constants prediction based on XGBoost and LightGBM algorithms. *Molecular physics*. 118 (14), e1696478. doi:10.1080/00268976.2019.1696478.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., et al. (2020a) Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*. 395 (10229), 1054–1062. doi:10.1016/S0140-6736(20)30566-3.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., et al. (2020b) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 579 (7798), 270–273. doi:10.1038/s41586-020-2012-7.

Zhou, Y., Wu, Y., Ding, L., Huang, X. & Xiong, Y. (2021) Point-of-care COVID-19 diagnostics powered by lateral flow assay. *Trends in analytical chemistry: TRAC*. 145, 116452. doi:10.1016/j.trac.2021.116452.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., et al. (2020) A Novel Coronavirus from Patients with Pneumonia in China, 2019. *The New England journal of medicine*. 382 (8), 727–733. doi:10.1056/NEJMoa2001017.

Zhu, Y., Bloxham, C.J., Hulme, K.D., Sinclair, J.E., Tong, Z.W.M., Steele, L.E., Noye, E.C., Lu, J., Xia, Y., Chew, K.Y., Pickering, J., Gilks, C., Bowen, A.C. & Short, K.R. (2021) A Meta-analysis on the Role of Children in Severe Acute Respiratory Syndrome Coronavirus 2 in Household Transmission Clusters. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*. 72 (12), e1146–e1153. doi:10.1093/cid/ciaa1825.

Centers for Disease Control and Prevention (2023) Risk Assessment Summary for SARS CoV-2 Sublineage BA.2.86. 24 August 2023. https://www.cdc.gov/respiratory-viruses/whats-new/covid-19-variant.html [Accessed: 1 September 2023].

Mellor, J., Christie, R., Paton, R.S., Leslie, R., Tang, M., Fyles, M., Deeny, S., Ward, T. & Overton, C.E. (2023) Real-time COVID-19 hospital admissions forecasting with leading

indicators and ensemble methods in England. arXiv [stat.AP]. http://arxiv.org/abs/2306.05762.