# Identifying populations at high risk of infections and antimicrobial resistance using large-scale electronic health record data


Emma Pritchard

Wolfson College

University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2024

# Table of Contents

# List of Figures

# List of Tables

# Abstract

Identifying populations at increased risk of infections can help inform public health strategies to reduce the incidence of disease. Electronic health records (EHRs) and large cohort data offer an opportunity to consider and identify many risk factors for various infections; however, it is unclear how to achieve this, especially in near real-time scenarios. This was particularly important during the rapidly evolving COVID-19 pandemic but is also important for monitoring common bloodstream infections (BSIs) such as *Escherichia coli* (*E. coli*) BSIs. This thesis therefore aimed to identify populations at a higher risk of infections using large datasets.

I first developed a real-time screening process to monitor associations between SARS-CoV-2 infection and demographic and behavioural risk factors. I considered potential confounders, multiple testing, collinearity, and reverse causality during the development of the process and demonstrated its use between July 2020-2021. I then explored methods to identify changes in growth rates of SARS-CoV-2 prevalence, comparing Iterative Sequential Regression and second derivates of generalised additive models. I found that both methods could find change-points around 3-5 weeks after they occurred in the data and that change-points could be detected earlier within specific subgroups. I next explored whether I could extend my learning and the methods I developed for use during the COVID-19 pandemic to investigate different diseases in varying data sources. I investigated the challenges in using large-scale EHRs when conducting case-control studies to identify risk factors for *E. coli* BSIs, specifically how to define control groups and risk factors for analyses of routinely collected data. I found missing data to be a key component when choosing a control group and that reverse causality could impact associations between calculated risk factors and *E. coli* BSIs. Finally, I extended and implemented the screening process developed originally for the COVID-19 pandemic on EHR data to identify associations between risk factors and *E. coli* BSIs. I discussed potential interventions based on these findings.

Overall, this thesis demonstrated effective methods for identifying populations at increased risk of infectious diseases using large datasets. With the continuing growth of EHRs, leveraging these resources to monitor at-risk populations could enhance the targeting of future interventions, ultimately aiming to reduce the burden of disease.

# Acknowledgements

This thesis would not have been possible without the brilliant support and mentorship of my supervisors. I am very grateful to have learned so much from Sarah since first joining the group in 2018. Sarah's consistent support and encouragement throughout this work has been invaluable. As well as being incredibly inspiring at work, Sarah has shown me great compassion when life outside work got harder - I can't overstate how important this support and flexibility has been. I would also like to thank David and Susan for their guidance throughout this project and their fascinating insights into the clinical world. I first chose to do my Master's in medical statistics because I really enjoy the interdisciplinary nature of the field so it has been fantastic getting to work with David and Susan.

I would like to thank everyone in the Modernising Medical Microbiology group who has supported me throughout my DPhil. I would particularly like to thank Karina for all her insight on this work and for providing me with boundless optimism each week, keeping me positive throughout this project. I would also like to thank Carla Wright for her support - our weekly Teams catch-ups were the highlight of my week during the pandemic!

I owe an enormous thank you to my family for all the continuous support they have given me, not just over the last four years. To my Mum for always pushing me to challenge myself and to my Dad for showing me how you can always find positivity and optimism even in the hardest times. And to my sisters, Megan and Becca, for brightening up every day.

Loads of thanks go to my partner (and now fiancé!) Jamie for his incredible patience over the last four years. From supporting me when things got stressful to celebrating with me when things went well, he has been on this whole rollercoaster ride with me from beginning to end. I am so incredibly grateful for his tireless support throughout.

Finally, I would like to thank all participants of the COVID-19 Infection Survey and acknowledge all the patients whose data is used throughout this thesis. This research in this thesis would not have been possible without their contributions.

# Funding

This work was supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with the UK Health Security Agency (UKHSA), and the NIHR Oxford Biomedical Research Centre (NIHR200915).

## Ethics Statement

Data from the COVID-19 Infection Survey used in Chapters 2 and 3 received ethical approval from the South Central Berkshire B Research Ethics Committee (20/SC/0195).

Data from the Infections in Oxfordshire Research Database (IORD; https://oxfordbrc.nihr.ac.uk/research-themes/modernising-medical-microbiology-and-big-infection-diagnostics/iord-about/) used in Chapters 4 and 5 has generic Research Ethics Committee, Health Research Authority and Confidentiality Advisory Group approvals (19/SC/0403, 19/CAG/0144) as a de-identified Electronic Research Database of routinely collected NHS electronic healthcare record data from the Oxford University Hospitals NHS Foundation Trust Clinical Systems Data Warehouse and research data (e.g. sequencing data) from the Antimicrobial Resistance and Modernising Microbiology Theme of the Oxford NIHR Biomedical Research Centre, Oxford. IORD records are identified by a specific, random, number ensuring that no patient-identifiable information is shared with researchers using this resource.

## Statement of contribution to work

I, Emma Pritchard, designed and conducted all of the analyses presented in this thesis with appropriate support from my supervisors and colleagues. I hereby declare and gratefully acknowledge the assistance I have received from others in relation to this work below.

**Chapter 2**

I am grateful to Joel Jones from the Office for National Statistics for his statistical insight during the development of the methods presented in Chapter 2.

**Chapter 3**

Karina-Doris Vihta provided example code for Iterative Sequential Regression. I am also grateful to Koen Pouwels for reviewing R code for calculating second derivatives.

**Chapters 4 and 5**

Phuong Quan and Jack Cregan provided the data extract from IORD for these Chapters. I am grateful to Koen Pouwels, Phuong Quan, Sam Lipworth, Amelia Andrews, Karina-Doris Vihta, Qingze Gu, Jack Cregan, Susan Hopkins, Dimple Chudasama, Russell Hope, Elisabeth Dietz for meeting to discuss risk factors defined for these Chapters, offering their expertise and guidance on which risk factors I should be using and how these should be defined. Chris Middlemass, an accredited clinical coder at the Oxford University Hospital NHS Foundation Trust, offered insight into how ICD-10 codes are recorded in IORD.

# List of abbreviations

| | |
|---|---|
| A&E | Accident and Emergency |
| AMR | Antimicrobial Resistance |
| BH | Benjamini-Hochberg |
| BIC | Bayesian Information Criterion |
| BLC | Blood culture |
| BMI | Body Mass Index |
| BSI | Bloodstream Infection |
| *C. diff* | *Clostridioides difficile* |
| CI | Confidence Interval |
| CIS | COVID-19 Infection Survey |
| CMV | Cytomegalovirus |
| CNS | Central Nervous System |
| COVID-19 | Coronavirus Disease 2019 |
| CPRD | Clinical Practice Research Datalink |
| CRP | C-reactive protein |
| CT | Computed Tomography |
| CT | Cycle Threshold |
| CVD | Cardiovascular disease |
| DBP | Diastolic Blood Pressure |
| *E. coli* | *Escherichia coli* |
| EBV | Epstein–Barr virus |
| EDF | Effective Degrees of Freedom |
| EHR | Electronic Health Record |
| EMR | Electronic Medical Record |
| ESPAUR | English Surveillance Programme for Antimicrobial Utilisation and Resistance |
| EUCAST | European Committee on Antimicrobial Susceptibility Testing |
| FY | Financial Year |
| GAM | Generalised Additive Model |
| GP | General Practice |
| GWAS | Genome-Wide Association Studies |

| | |
|---|---|
| HES | Hospital Episode Statistics |
| HH | Household |
| ICD-10 | International Classification of Diseases, 10th Revision |
| ICU | Intensive Care Unit |
| IMD | Index of Multiple Deprivation |
| IORD | Infections in Oxfordshire Research Database |
| IQR | Interquartile Range |
| IRR | Incidence Rate Ratio |
| ISR | Iterative Sequential Regression |
| IT | Information Technology |
| LIMS | Laboratory Information Management Systems |
| LOS | Length of Stay |
| MICE | Multiple Imputation by Chained Equations |
| MRSA | Methicillin-resistant Staphylococcus aureus |
| NAP | National Action Plan |
| NIMS | National Immunisation Management System. |
| ONS | Office for National Statistics |
| OPCS | Office of Population Censuses and Surveys |
| PAF | Population Attributable Fraction |
| PCR | Polymerase Chain Reaction |
| PSA | Prostate-Specific Antigen |
| RC | Reference Category |
| RSV | Respiratory Syncytial Virus |
| *S. aureus* | *Staphylococcus aureus* |
| SARS-CoV-2 | Severe Acute Respiratory Syndrome Coronavirus 2 |
| SBP | Systolic Blood Pressure |
| SDE | Secure Data Environment |
| SGSS | Second Generation Surveillance System |
| SGTF | S-Gene Target Failure |
| SGTP | S-Gene Target Positive |
| SY | School Year |
| UKHSA | United Kingdom Health Security Agency |
| VOC | Variant Of Concern |
| WHO | World Health Organisation |

## Papers arising directly from this work

Pritchard, E., Jones, J., Vihta, K. D., Stoesser, N., Matthews, P. C., Eyre, D. W., House, T., Bell, J. I., Newton, J. N., Farrar, J., Crook, D., Hopkins, S., Cook, D., Rourke, E., Studley, R., Diamond, I., Peto, T., Pouwels, K. B., Walker, A. S., & COVID-19 Infection Survey Team (2022). Monitoring populations at increased risk for SARS-CoV-2 infection in the community using population-level demographic and behavioural surveillance. *The Lancet regional health. Europe*, *13*, 100282. https://doi.org/10.1016/j.lanepe.2021.100282

Pritchard, E., Vihta, K. D., Eyre, D. W., Hopkins, S., Peto, T. E. A., Matthews, P. C., Stoesser, N., Studley, R., Rourke, E., Diamond, I., Pouwels, K. B., Walker, A. S., & COVID-19 Infection Survey Team (2024). Detecting changes in population trends in infection surveillance using community SARS-CoV-2 prevalence as an exemplar. *American Journal of Epidemiology*, kwae091. Advance online publication. https://doi.org/10.1093/aje/kwae091

## Other papers arising during the course of this work

**Peer-reviewed**

Pouwels, K. B., House, T., Pritchard, E., Robotham, J. V., Birrell, P. J., Gelman, A., Vihta, K. D., Bowers, N., Boreham, I., Thomas, H., Lewis, J., Bell, I., Bell, J. I., Newton, J. N., Farrar, J., Diamond, I., Benton, P., Walker, A. S., & COVID-19 Infection Survey Team (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. https://doi.org/10.1016/S2468-2667(20)30282-6

Pritchard, E., Fawcett, N., Quan, T. P., Crook, D., Peto, T. E., & Walker, A. S. (2021). Combining Charlson and Elixhauser scores with varying lookback predicated mortality better than using individual scores. *Journal of clinical epidemiology*, *130*, 32–41. https://doi.org/10.1016/j.jclinepi.2020.09.020

Pritchard, E., Matthews, P. C., Stoesser, N., Eyre, D. W., Gethings, O., Vihta, K. D., Jones, J., House, T., VanSteenHouse, H., Bell, I., Bell, J. I., Newton, J. N., Farrar, J., Diamond, I., Rourke, E., Studley, R., Crook, D., Peto, T. E. A., Walker, A. S., & Pouwels, K. B. (2021). Impact of vaccination on new SARS-CoV-2 infections in the United Kingdom. *Nature medicine*, *27*(8), 1370–1378. https://doi.org/10.1038/s41591-021-01410-w

Walker, A. S., Pritchard, E., House, T., Robotham, J. V., Birrell, P. J., Bell, I., Bell, J. I., Newton, J. N., Farrar, J., Diamond, I., Studley, R., Hay, J., Vihta, K. D., Peto, T. E., Stoesser, N., Matthews, P. C., Eyre,

D. W., Pouwels, K. B., & COVID-19 Infection Survey team (2021). Ct threshold values, a proxy for viral load in community SARS-CoV-2 cases, demonstrate wide variation across populations and over time. *eLife*, *10*, e64683. https://doi.org/10.7554/eLife.64683

Pouwels, K. B., Pritchard, E., Matthews, P. C., Stoesser, N., Eyre, D. W., Vihta, K. D., House, T., Hay, J., Bell, J. I., Newton, J. N., Farrar, J., Crook, D., Cook, D., Rourke, E., Studley, R., Peto, T. E. A., Diamond, I., & Walker, A. S. (2021). Effect of Delta variant on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK. *Nature medicine*, *27*(12), 2127–2135. https://doi.org/10.1038/s41591-021-01548-7

Vihta, K. D., Pouwels, K. B., Peto, T. E. A., Pritchard, E., Eyre, D. W., House, T., Gethings, O., Studley, R., Rourke, E., Cook, D., Diamond, I., Crook, D., Matthews, P. C., Stoesser, N., Walker, A. S., & COVID-19 Infection Survey (2022). Symptoms and Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Positivity in the General Population in the United Kingdom. *Clinical Infectious Diseases*, *75*(1), e329–e337. https://doi.org/10.1093/cid/ciab945

Walker, A. S., Vihta, K. D., Gethings, O., Pritchard, E., Jones, J., House, T., Bell, I., Bell, J. I., Newton, J. N., Farrar, J., Diamond, I., Studley, R., Rourke, E., Hay, J., Hopkins, S., Crook, D., Peto, T., Matthews, P. C., Eyre, D. W., Stoesser, N., Pouwels, K. B., & COVID-19 Infection Survey Team (2021). Tracking the Emergence of SARS-CoV-2 Alpha Variant in the United Kingdom. *The New England Journal of Medicine*, *385*(27), 2582–2585. https://doi.org/10.1056/NEJMc2103227

Dietz, E., Pritchard, E., Pouwels, K., Ehsaan, M., Blake, J., Gaughan, C., Haduli, E., Boothe, H., Vihta, K. D., Peto, T., Stoesser, N., Matthews, P., Taylor, N., Diamond, I., Studley, R., Rourke, E., Birrell, P., De Angelis, D., Fowler, T., Watson, C., Eyre, D., House, T., Walker, A. S. (2024). SARS-CoV-2, influenza A/B and respiratory syncytial virus positivity and association with influenza-like illness and self-reported symptoms, over the 2022/23 winter season in the UK: a longitudinal surveillance cohort. *BMC Medicine*, *22*(1), 143. https://doi.org/10.1186/s12916-024-03351-w

**Preprints**

Vihta, K. D., Pritchard, E., Pouwels, K. B., Hopkins, S., Guy, R. L., Henderson, K., Chudasama, D., Hope, R., Muller-Pebody, B., Walker, A. S., & Eyre, D. W. (2023). Predicting future hospital antimicrobial resistance prevalence using machine learning. *medRxiv*, 2023-11.

# Chapter 1 Introduction

## 1.1 Identifying at-risk populations using data

The idea of using data to find people most at risk of disease is not new. Many have previously recognised that identifying the risk factors for a disease can lead to interventions that reduce spread and protect people from harm. One of the earliest, and probably most famous, examples of this in modern epidemiology is the work of John Snow in 1854. Through mapping the location of water pumps in London, he traced cases of cholera to a single water pump contaminated with sewage[1] leading to the removal of the pump handle to stop continued cholera transmission.[2] Around a similar time, Florence Nightingale was working as a nurse during the Crimean War. Through plotting the causes of mortality on innovative graphs, she linked the high mortality rate to poor sanitation and overcrowding, leading to the intervention of improved hand hygiene.[3]

Epidemiologists have also used cohort studies to identify at-risk populations for different diseases, with the Framingham Heart Study being one of the first long-term studies of its kind.[4] Originally established in 1948, the study recruited individuals without cardiovascular disease (CVD) and followed them over time, leading to the identification of many risk factors including elevated blood pressure, blood triglyceride, and cholesterol levels.[4] This study has contributed to numerous healthcare interventions,[5] including CVD risk calculation tools for clinicians.[6,7] Also pioneers in the field of epidemiology, Austin Bradford Hill and Richard Doll established the British Doctors Study in 1951: a prospective cohort study investigating the association between smoking and lung cancer.[8] By 1954 they published preliminary results with evidence supporting this association,[9] and twenty years into the study, alongside Richard Peto, published strong evidence for this association, along with associations between smoking and other diseases.[10] The rate of smoking in England has since massively declined.[11]

While, as demonstrated through the examples above, using data for disease surveillance is not new, large-scale data collected through Electronic Health Records (EHRs) and increasing computing power is. In this Introduction, I will first introduce the history and growth of EHRs particularly for research purposes. In the two following sections, I will introduce surveillance systems for the two disease interests in this thesis: bloodstream infections (BSIs) and COVID-19. I will then present the aims and outline of this thesis.

## 1.2 Electronic Health Records

### 1.2.1 A Brief History of Electronic Health Records

Electronic Health Records (EHRs), also known as Electronic Medical Records (EMRs), are a collection of electronic (or digital) medical information about individuals including, for example, diagnoses, medications, and procedures.[12] The primary purpose of EHRs is often for hospital reimbursement[13,14] but they also have been shown to improve patient care via better access to previous healthcare records for clinicians managing patients and better communication between different hospital providers.[15] Naturally, the use of EHRs grew in parallel with improvements to computers. The transition from paper to electronic records began in the 1960s, with US academic medical centres developing their own systems.[16-18] By the 1990s, it was recognised that computer-based systems could improve patient care (for example through point-of-care reminders) and computers were more affordable and powerful.[19] The late 1990s and 2000s saw increased initiatives to incentivise adoption of EHRs, for example in 1998[20] and 2002[21] in the UK, although implementation throughout the 2000s was slow.[22] The 2010s saw substantial growth in EHR implementation. One study estimated that, in the US, EHR adoption rate increased from 6.6% in 2009 to 81.2% in 2019.[23] EHRs are now well-established in England with 90% of NHS Trusts using EHRs in 2023.[24]

While the focus above is on EHR implementation in high-income countries, the picture in low and low-middle-income countries is different and worth contextualising. A systematic review in 2023 summarising the adoption of EHRs in low-income countries concluded that EHRs are currently at an early stage of implementation.[25] The main barriers the studies mentioned included lack of training, poor infrastructure, and lack of management commitment. However, they gave examples where EHRs positively impacted patient care, for example, in Uganda[26] and Rwanda,[27] stating that the adoption of EHRs is becoming feasible and more possible in the future. Analyses of EHRs in middle-income countries show a similar picture, with interoperability and integration of standalone electronic systems being particularly challenging and barriers to progress including a lack of national leadership and conflict.[28] The disparity in provision of EHRs in high vs low- and low-middle-income countries was recognised by the World Health Organization (WHO) in the Global Strategy on Digital Health for 2020-2025 with one of four guiding principles being to "recognize the urgent need to address the major impediments faced by least-developed countries implementing digital health technologies".[29] It is important to recognise that, due to the barriers explored above, the discussions around EHR use in this thesis are most applicable to high-income countries such as the UK.

### 1.2.2 Electronic Health Records for research and surveillance

As the number of EHRs and the data they include has grown, EHRs have been increasingly used for research. The large amount of data on a large variety of factors, continuously updated over time, offers a wealth of information with which to explore medical questions ranging from assessing the impact of healthcare interventions[30] such as the cost-effectiveness of different hip replacements[31] to predicting healthcare outcomes for the future, particularly with advances in machine learning and artificial intelligence.[32] EHRs have been used previously to identify risk factors for diseases, for example finding risk factors for cardiovascular diseases in those who previously had cancer[33] and investigating whether body mass index (BMI) is a risk factor for dementia.[34]

Near real-time surveillance using EHRs has challenges. Real-time surveillance relies on up-to-date data streams that may lag for varying reasons, for example, due to the time between patient discharge and electronic recording and the time from data input to researcher access. Once the researcher has access to the data, processes such as data cleaning may take time, further impacting surveillance timeliness. In addition, data quality may impede the use of EHRs for surveillance due to incomplete case ascertainment, for example where tests are done outside of hospital settings (e.g. for COVID-19) or due to factors such as cause of death being difficult to ascertain with confidence.[35] EHRs can also be messy with large amounts of missing data and unexpected change-points due to changes in hospital infrastructure[36] such as a new analyser measuring blood test results reporting with different units. "Garbage in, garbage out" holds true for all statistical analyses and hence these inconsistencies are vital to consider when using EHRs for research.[37] These challenges notwithstanding, studies have demonstrated the value use of EHRs for surveillance, for example providing a real-time early warning system for COVID-19 mortality based on laboratory results and clinical measurements,[38] and providing near real-time estimates of COVID-19 case numbers.[39]

A common challenge encountered when using EHR data for research in England is the lack of access to linked datasets. Data linkage involves combining data from different sources for the same person to create a new and enhanced dataset.[40] This could be, for example, linking inpatient admissions with laboratory test results, or linking inpatient admissions with general practice (GP) (primary care) records. Being able to link data from different sources has numerous benefits for research.[41] It allows analysis of associations that may otherwise be difficult to determine through one data source alone, for example, linking inpatient admissions with GP records would allow greater ascertainment of comorbidities for individuals getting admitted to hospital. While the benefits of data linkage are evident, in reality there are many barriers and practical challenges that may slow down or completely inhibit the linkage of data from different sources. A systematic review in 2022 identified the main barriers to data linkage in the UK, with the key issue reported being limits to technical

capabilities and data quality issues.[42] This included the absence of secure data transfer methods and old systems unable to share data across organisations. Differences in data quality and coding algorithms between organisations were also noted as being particularly challenging. The underpinning legal and ethical framework was cited as the second most common barrier, with laws set up to improve data protection often limiting the sharing of data which is vital for data linkage.

There is currently no provision for national linked healthcare data across the population of England. This is due to some of the barriers mentioned above including a lack of secure mechanism for researchers to access data and conduct population-wide research and no national linkable primary care data until very recently.[43] This means that research on datasets in England is usually carried out on data in one of the three following ways:

1. Linked and detailed local level data e.g. Infections in Oxfordshire Research Database,[44]
2. National-level unlinked datasets such as Hospital Episode Statistics (HES)[45] or the Clinical Practice Research Datalink (CPRD) for primary care data,[46]
3. National-level datasets with linkage specifically for a project e.g. linking primary care data from CPRD with secondary care data from Hospital Episode Statistics (HES).[47,48]

COVID-19 has accelerated the development of national-level linked data, with studies successfully linking multiple data sources to obtain data on 96% of the English population, linking primary care data with HES, the death registry, laboratory testing for COVID-19, and community dispensing.[43] The OpenSAFELY platform was also established,[49] covering >99% of primary care records in England and linked to HES.[50] The COVID-19 pandemic demonstrated that with increased capacity and need, there could be faster and more efficient access to data and increased data linkage while still carrying out analysis safely and securely under data protection principles,[51] however specific legislation had to be passed to allow this.[52]

The way researchers access EHR data to conduct research must be balanced to ensure data privacy while allowing timely access to data. There are currently two broad ways researchers access data in England:

1. Researchers gain access to data through the raw data being shared directly with them. This will be under strict specific restrictions. Some examples of these restrictions, as outlined by the Data Access Request Service for accessing NHS digital data, include the requirement to store data on an encrypted laptop and to delete the data after the research is completed.[53]
2. Researchers gain access to data from working within a Secure Data Environment (SDE). Raw data never leaves the SDE and all research outputs (e.g. graphs and tables) taken out of the

SDE have to be approved before being shared. Examples of SDEs include the Office for National Statistics (ONS) Secure Research Service[54] and the OpenSAFELY platform.[49]

SDEs are becoming an attractive approach to data sharing and access, especially since the publication of the "Goldacre Review" in April 2022 that recommended using SDEs (called Trusted Research Environments in the review) as the norm for all analyses of NHS patient records.[55] By never sharing the raw data outside of the SDE, patient data is safer as it is never stored on an individual's laptop which could, for example, be lost. All output taken from within the SDE for publication or sharing has to be approved further ensuring patients cannot be identified from output. The OpenSAFELY platform also requires that all code which is written inside the SDE be shared for potential re-use by subsequent users, a further benefit saving researcher time and deduplicating efforts. However, SDEs have limitations, as they are resource-intensive and require large teams of employees to review files before adding them to or removing them from the SDE. They can also be slow to use with many individuals working within a shared environment. While they sound good in principle, in practice they can therefore be challenging to use.

## 1.3 Surveillance of bloodstream infections in England

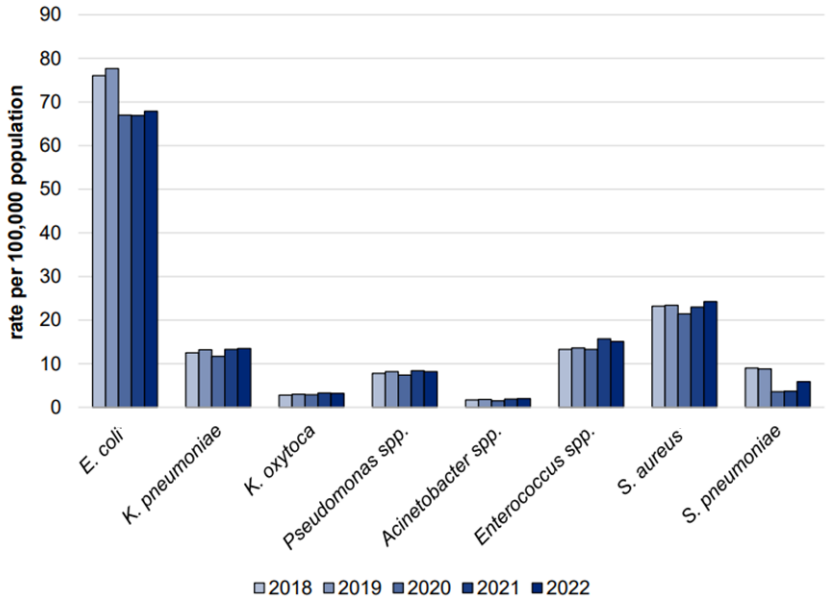
Before starting analyses investigating how to improve identification and surveillance of risk factors for BSIs using EHRs, it is important to understand the history of BSI surveillance and the current surveillance situation.

Mandatory surveillance of BSIs in England began in 2001 in response to observed rising rates of *Staphylococcus aureus (S. aureus)* bacteraemia.[56] Initially, this surveillance compromised quarterly reports from Trusts reporting the total number of positive blood cultures, the number of blood cultures positive for *S. aureus,* and of those, how many were Methicillin-resistant.[57] In 2005, additional characteristics such as date of birth, sex, and patient's location were added to enhance surveillance.[56] Mandatory reporting of *Clostridioides difficile* was added in 2004 for patients aged ≥65 years, expanding to all cases in patients aged ≥2 years in 2007. Reporting of *E. coli* bacteraemias was initially voluntary but was made mandatory in June 2011 following observed year-on-year increases. *Klebsiella* species and *Pseudomonas aeruginosa* bacteraemia were added to the mandatory surveillance in 2017 following government targets to reduce Gram-negative infections by 50% by 2021. For surveillance of all these pathogens, information on infections was collected separately in each acute hospital trust and then reported to UKHSA via a web-based surveillance system; the Healthcare Associated Infection Data Capture System (HCAI-DCS). One output of mandatory surveillance was the establishment of the English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR). This program aimed to bring together mandatory

surveillance (including resistance) with antimicrobial use to monitor the number of infections and antimicrobial stewardship.[58] The first report from the programme was published in 2014[59] and reports have been published annually ever since.[60]

The current ESPAUR report published in November 2023 reported the trends in the incidence of priority pathogens causing BSIs, with the incidence of most pathogens being consistent or marginally increased from 2018-2022, apart from *E. coli* and *Streptococcus pneumoniae* which decreased over this timeframe, coincident with the COVID-19 pandemic (**Figure 1.1**).[61] *E. coli* was the most common key BSI pathogen identified. Regional estimates of BSIs were also presented with rates generally being highest in the North East. Key characteristics of age, ethnicity, and deprivation were also summarised for BSIs, particularly with a focus on the number of AMR BSIs in these subgroups, finding that those in the highest age group (>74 years), in the most deprived quintile for deprivation, and those of white ethnicity had the highest rates of BSIs.

**Figure 1.1: Annual incidence rate of key pathogen BSI, per 100,000 population, England**




*Source: UKHSA, ESPAUR Report 2022/23.*[61]

The use of the current surveillance data to monitor populations at risk of BSIs is limited for several reasons. First, only identified infections are collected and stored within the HCAI-DCS. Positive microbiological test results are stored in a separate system, the Second Generation Surveillance System (SGSS). However this does not have negative test results. Without negative test results, it can be challenging to identify those at higher risk as there is no negative control group to compare to – the only comparisons possible are with national data, for example on age, sex and deprivation as above. Further, data completed by the Infection Control teams at the local level into the HCAI-DCS

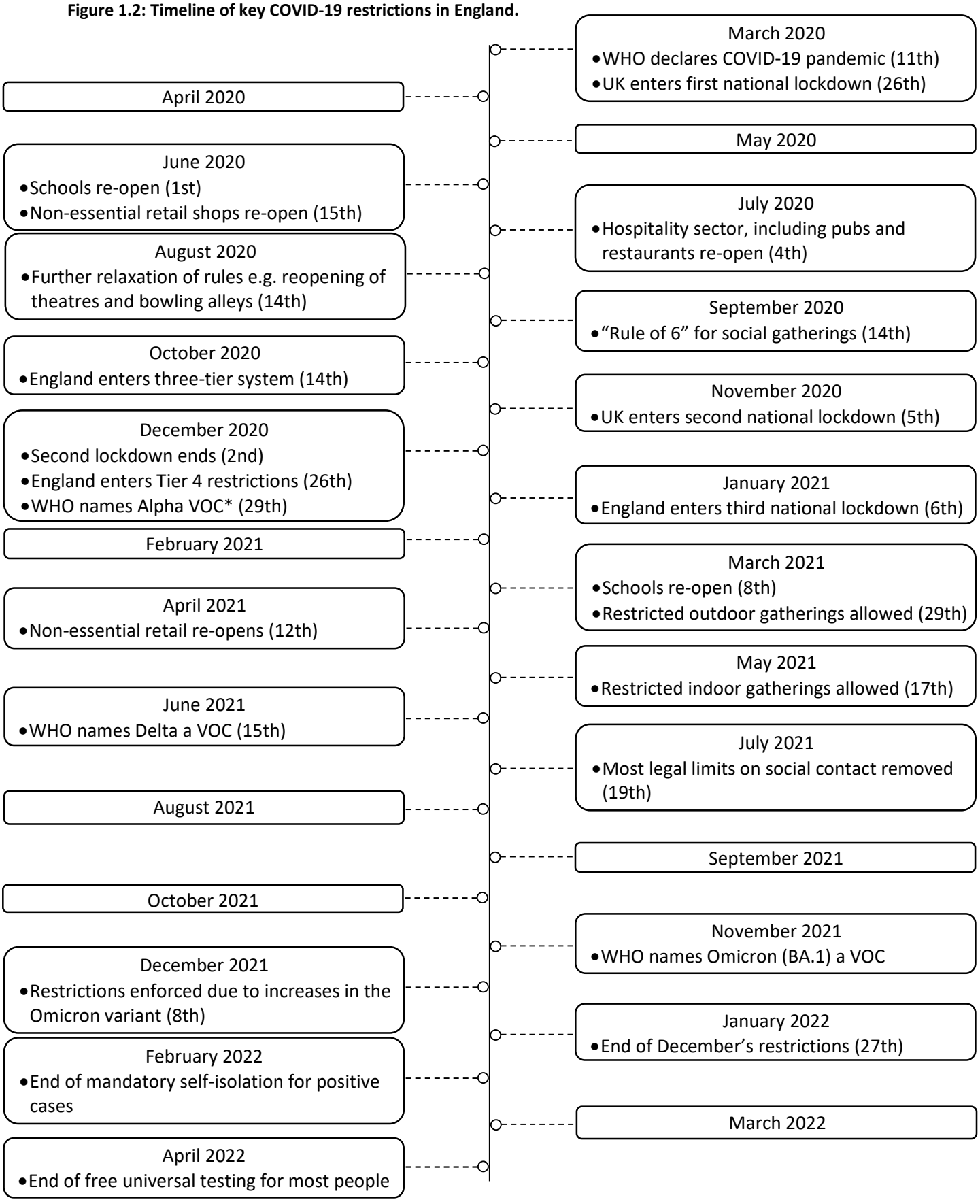
has previously been shown to have high levels of missingness for some characteristics.[62] Completeness of characteristics varied by both pathogen isolated and the data field. For example, fields were on average 81% complete for *S. aureus* BSIs but only 48% complete for *C. difficile*. This highlights the importance of trying to use linked EHRs, as a potentially more complete data source, to better understand what is impacting the numbers of BSIs in England and hence what interventions might be most useful to reduce the number of BSIs.

## 1.4 Surveillance of COVID-19

In December 2019, doctors noticed a cluster of pneumonia cases in Wuhan, China.[63] This cluster of cases was identified as a novel coronavirus, subsequently termed "SARS-CoV-2" and most commonly known as COVID-19. On 30th January 2020, the World Health Organisation (WHO) declared a Public Health Emergency of International Concern due to the rapid spread of the virus, and by 11th March 2020, a pandemic was declared.[64] Over the following years, COVID-19 would spread around the globe, causing an estimated 776 million infections[65] and 7.1 million deaths globally[66] as of 8th July 2024. The impact that COVID-19 had on public health and society was enormous. Restrictions to control the spread of the virus were varied and ever-changing, as demonstrated in England over the first two years of the pandemic in **Figure 1.2**. Many people are still bearing the cost of having had the disease through long COVID, and many economies are still recovering.

The unprecedented nature of the COVID-19 pandemic was matched by the response from the scientific community. Within a year of COVID-19 first emerging, a vaccine was developed and administered, with the first vaccine being given on 8th December 2020 in the UK.[67] This was much faster than usual vaccine development which typically takes upwards of ten years.[68] Large-scale testing facilities were also established within the first months of the pandemic being declared. In England, NHS Test and Tract was launched on 28 May 2020, allowing quick and easy access to PCR (Polymerase Chain Reaction) tests for those with pre-defined COVID-19 symptoms.[69]

**Figure 1.2: Timeline of key COVID-19 restrictions in England.**




**VOC = Variant of concern as defined by WHO.[70] Adapted from Institute for Government.[71]*

Near real-time surveillance of COVID-19 in the UK was performed using several studies, with diversity in data collection methods and the type of data collected. Large, community-based surveillance studies aimed to assess who had COVID-19 in the general population in near real-time allowing effective and informative policy decisions. The first survey launched in the UK was the Office for National Statistics (ONS) COVID-19 Infection Survey (CIS) which started collecting data on 26 April, 2020.[72] This survey was set up to monitor who was testing positive for SARS-CoV-2 in the community and recruited randomly selected households across the UK. Participants were swabbed regularly regardless of symptoms, meaning both symptomatic and asymptomatic cases were identified. Another study which also launched in April 2020 and focused on community testing was REACT: Real-time Assessment of Community Transmission.[73] REACT was a point-prevalence survey that swabbed 100,000 different randomly selected people each month, again regardless of symptoms, testing them for SARS-CoV-2 and recording other characteristics.

The examples provided above were both community surveys with randomly selected participants, but other projects took different approaches. The COVID Symptom Study, now known more broadly as the ZOE Health Study, was a partnership between British company Zoe and Kings College London.[74] This study was run through a free phone app where participants could log symptoms they were experiencing daily and were advised to seek testing dependent on symptoms, allowing estimation of the number of symptomatic COVID-19 cases. Another study taking a different approach was the SARS-CoV-2 immunity and reinfection evaluation study, known as SIREN.[75] This study was restricted to those who worked in healthcare settings, with participants being recruited through one of 135 hospital sites they worked at. This study focused on trying to understand the immune response to COVID-19 and evaluate the protection offered by vaccines and the nature of reinfections in healthcare workers, who had high exposure. The four studies mentioned above are not an exhaustive list of studies monitoring COVID-19 in the UK, but they portray the scale and diversity of studies done.

Other studies took advantage of pre-collected data by using EHRs. While data linkage in EHRs can be challenging (as discussed above), the accelerated need to understand the COVID-19 pandemic led to insightful studies using linked EHR data. One study linked all SARS-CoV-2 PCR tests in England to hospital attendance and mortality data where available, estimating the number of hospital-onset cases to better understand the onset of COVID-19 in hospitals and the community.[76] Another study linked these data to find the risks of hospitalisation and death associated with the Omicron and Delta variants, finding that hospitalisation and death were reduced for Omicron compared with Delta while also showing that pre-existing immunity through either vaccination or previous infection reduced the risk of hospitalisation.[77] Many studies also used data to assess vaccine response, for

example linking test results from PCR and lateral flow tests with vaccination data from the National Immunization Management System (NIMS) to estimate vaccine effectiveness against the Omicron Variant.[78] With a new disease and multiple new vaccines, studies such as these were able to use pre-collected data to importantly add to the knowledge base at the time.

Using data from the national PCR and lateral flow test results had limitations. Primarily, these tests were only from individuals who sought testing. This was likely to miss many infections as an estimated 41% of SARS-CoV-2 cases were asymptomatic[79] and the national testing programs restricted the availability of tests based on symptoms for substantial periods of time. Additionally, the uptake of testing was influenced by various factors, such as the availability of testing facilities and the cost of self-isolation, which was more likely to negatively impact those experiencing high deprivation levels who could not afford to miss work.[80] As CIS and REACT tested everyone in their respective cohorts regardless of symptoms they were likely to be less affected by these issues. However, looking at all the evidence from all the studies together gave increased confidence of true effects when they were observed in multiple studies, and allowed further investigation and understanding when studies from different sources provided different results.

This section has highlighted the vast amount of work done during the pandemic. Moving forward, there are opportunities to use the large amounts of data collected to better understand the pandemic and how to respond in the future and also use the methods developed during the pandemic to better understand other infectious diseases.

## 1.5 Thesis Outline

The aim of this thesis is to identify populations at a high risk of infections using large datasets.

In Chapter 2, I developed a process to monitor populations at an increased risk of COVID-19 in the community using population-level data from CIS. I focused on creating a process that could be used in near real-time to allow key risk factors to be monitored throughout the pandemic to give a better understanding of what was driving current increases in positivity rates. As I considered many risk factors, I developed a process that could handle large amounts of missing data and I assessed the impact of multiple testing.

In Chapter 3, I looked at the overall population trends of COVID-19 which was an important part of monitoring the risk of COVID-19. There was no clear guidance on what method would be best to identify when population trends were increasing or decreasing so I compared two methods that looked for changes in population trends: Iterative Sequential Regression and second derivatives of generalised additive models. I compared the location of the change-points these methods found

retrospectively over two years of the pandemic as well as assessing how quickly these methods detected change-points in near real-time.

I went on to apply the methods I developed in Chapter 2 to a new context, specifically seeing if I could apply the methodology to EHRs to monitor *E. coli* BSIs. I first had to define a control group and risk factors using EHRs. In Chapter 4, I explored the challenges encountered when doing this and suggested various ways to define control groups and risk factors while accounting for different types of bias using EHRs.

In Chapter 5, I combined the learning from Chapters 2 and 4 to run the screening process on EHR data to find risk factors for *E. coli* BSIs. I investigated risk factors annually over multiple years from 2018 to 2022. I also considered whether risk factors varied across different types of *E. coli* BSIs, for example, nosocomial BSIs or those resistant to certain antibiotics. I discussed the risk factors found within the context of the literature and explored potential reasons for new risk factors which I identified.

I conclude this thesis with a discussion on the potential future direction of this work, focusing on expanding the methodology developed to national-level data.

# Chapter 2 Monitoring populations at increased risk for SARS-CoV-2 infection in the community

The work presented in this Chapter was published in *The Lancet Regional Health – Europe* in 2022.[81] I authored the text and figures below (which are reproduced here largely unaltered from the published work) with input from supervisors commensurate with the amount of input/advice that would be considered appropriate for a DPhil thesis. All authors have agreed to the inclusion of this published work in my thesis.

## 2.1 Introduction

As of 8th July 2024, there have been over 776 million SARS-CoV-2 cases worldwide.[65] Disparities in COVID-19 risk and outcomes based on demographics and behaviours have been described in the UK[82,83] and globally,[84,85] but emerging variants[86] coupled with varying control policies, including differential vaccine roll-out programmes, reinforce the need to monitor characteristics of individuals "at increased risk" for SARS-CoV-2 infection continuously.

In particular, as of 26th October 2021, when this analysis was conducted, the World Health Organisation (WHO) had identified four Variants of Concern (VOC) based on evidence of greater transmissibility, virulence, and/or decreased effectiveness of public health and social measures.[86] Identifying groups in whom newly identified variants of concern were spreading in the community was considered vital in preventing widespread transmission. In England, since 26th March 2020, there had been three national lockdowns, a tiered system[87] with varying restrictions in smaller geographical areas, and various other restrictions between these,[71] all affecting behaviour and risk of acquiring and spreading SARS-CoV-2. Finding societal factors or specific behaviours where these restrictions were less effective may assist policy development. With restrictions being relaxed in many countries from 2021 onwards, rapidly identifying groups where positivity was rising in real-time could help monitor the spread of SARS-CoV-2 and target advice.

High-quality surveillance is challenging, particularly given the large proportion of asymptomatic SARS-CoV-2-infected individuals,[88] with a balance between missing important but potentially imprecisely estimated signals (false-negatives) and noise (false-positives). With large datasets containing many potential risk factors, multiple testing is inevitably problematic,[89] but standard approaches to building regression models restricting to smaller numbers of hypothesised associated factors risk missing true signals with a rapidly evolving pathogen and societal responses. The cumulative effect of missing data across many risk factors can mean substantial proportions of the original sample are excluded from penalised regression or backwards elimination, losing power,[90]

and risking bias if missingness depends on the outcome.[91] Further, the most informative way to collapse small sub-categories amongst explanatory variables is often unknown. A method allowing numerous variable parametrisations of many individual variables would therefore be useful, provided collinearity and confounding can be taken into account given their potential to impact effects sizes and/or direction and incorrectly influence interpretation.[92]

Using the Office for National Statistics (ONS) COVID-19 Infection Survey, a large community-based surveillance study, I therefore developed a process with the potential to monitor groups with the highest SARS-CoV-2 positivity week by week.

## 2.2 Methods

### 2.2.1 Study design

The ONS COVID-19 Infection Survey was a large household survey with longitudinal follow-up (ISRCTN21086382; https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/protocol-and-information-sheets). Private households were randomly selected on a continuous basis from address lists and previous surveys to provide a representative sample across the UK. Following verbal consent, a study worker visited each household to take written informed consent for individuals aged >2 years (from parents/carers for those 2–15 years; those 10–15 years also provided written assent). The study received ethical approval from the South Central Berkshire B Research Ethics Committee (20/SC/0195).

Participants were asked about demographics, behaviours, work, and vaccination uptake, as can be seen on the Case Report Forms available here: https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/case-record-forms. New iterations of the forms added new characteristics as they became relevant (e.g. COVID-19 vaccination) and occasionally changed the wording of questions for clarity. At the first visit, participants were asked for consent for optional follow-up visits every week for the next month, then monthly thereafter. At each visit, participants provided a nose and throat self-swab.

### 2.2.2 Inclusion/exclusion criteria

This analysis included visits from 19th July 2020 to 17th July 2021 with a positive or negative swab result, including one visit per participant within each discrete fortnight in this period, namely the first test-positive visit, otherwise the last (negative) visit. This mimics repeated point-prevalence surveys, similar to the English Real-time Assessment of Community Transmission (REACT) study.[93]

### 2.2.3 Outcome and exposures

The outcome of the analysis was any SARS-CoV-2 PCR-positive swab in each fortnight. For exposures, I identified eight non-missing key potential confounders ("core" variables): age (years), sex, ethnicity (white vs non-white as relatively small numbers in the latter), geographical region (12 levels; 9 English regions and 3 devolved administrations: Wales, Scotland, Northern Ireland), rural/urban classification (major urban area, urban town/city, rural town, and rural village), deprivation percentile (see below), household size, and whether the household was multigenerational (defined as households including individuals aged school year 11 or younger AND school year 12 to age 49 AND aged 50+). Deprivation was assessed using the index of multiple deprivation (IMD) in England, a score based on lower-layer super output areas with an average population of 1,500 people and incorporating seven domains to produce an overall relative measure of deprivation (income,

employment, education, skills and training, health and disability, crime, barriers to housing services and living environment).[94] Equivalent scores were used in the other three countries comprising the UK.[95-98]

I next defined 60 non-core "screening" variables that could dynamically identify those at increased risk of testing positive from questions detailing participant's current work/school status, including the ability to social distance and patient-facing healthcare/social-care roles, current health status including COVID-19 vaccination and smoking, household and living environment, and contacts including with care homes, hospitals, and confirmed COVID-19 cases. IMD sub-components were assessed in the variable screening, however these were only available for England.

Although participants were tested predominantly monthly, most behavioural questions related to the last 7 days. As public health laws at the time required self-isolation if one tested positive for COVID-19, those with zero physical/social contacts were potentially more likely to have already had a positive swab result compared with those with one or more contacts. As some participants already knew/thought they had COVID-19 (from symptoms or testing outside the study) this could affect behaviours reported immediately before study tests, leading to reverse causality. The screening variables were therefore grouped into those most plausibly preceding any current infection (47 variables), or potentially modified through knowledge of recent prior infection (13 variables). The latter included the number of social/physical contacts, frequency of shopping and/or socialising, and time spent in other's homes/other people spent in participants' homes. Rather than using the value self-reported at the included visit, I considered the maximum reported value across all visits in the preceding 35 days, excluding the included visit (with the positive/negative swab included in the analysis), and only including participants with at least one test-negative visit in the preceding 10-35 days.

I incorporated work sector into the screening process as 16 separate binary variables (each work sector vs all other work sectors) rather than a 16-level categorical variable because the expectation was that only one or two sectors might have higher or lower positivity, with little difference between most sectors. Inclusion as one 16-level variable could risk missing effects of important individual work sectors versus an arbitrary reference category given the 16 degree of freedom global test, and, dependent on the reference category, individual effects may not be significant enough to capture at the individual p-value level. Instead, the chosen parameterisation compared participants in each work sector to those currently working in all other work sectors. I treated all other categorical variables (with a maximum of 5 levels) as mutually exclusive categories with a fixed reference category.

### 2.2.4 Statistical analysis

Within each fortnight, associations between SARS-CoV-2 positivity and the eight "core" characteristics were estimated using logistic regression (numbers included per fortnight in **Table 2.1**). These characteristics were included in all subsequent models regardless of statistical significance. All analyses used complete cases (all "core" variables were non-missing); models with household-level random effects would not converge due to low positivity rates. For the geographic region, South West England was the reference as this had the lowest SARS-CoV-2 positivity across the study, facilitating the identification of where infections were increasing.

Given the large number of effect estimates over the 52-week study period (e.g. shown for rural/ urban classification in **Figure 2.1**), I summarised the importance of each characteristic over time using two properties simultaneously: 1) global (Wald) p-value and 2) overall effect size, to create a standard error-weighted mean effect estimate:[99]

$$Overall\ effect\ size = \exp\left(\frac{\sum \frac{1}{se(\beta_i)}\beta_i}{\sum \frac{1}{se(\beta_i)}}\right)$$, where $\beta_i$ is the log odds ratio for each level.

For categorical variables with multiple levels, I set the reference category to the level with the lowest positivity in each fortnight to avoid overall effect sizes netting to zero if some category levels had an odds ratio >1, while others had an odds ratio <1, compared with the reference category. To incorporate non-linear effects, a restricted natural cubic spline was used with 4 internal knots at 20, 40, 60, and 80th percentiles of unique ages, and boundary knots at 5th and 95th percentiles; the overall effect size combined estimates at ages 10, 25, 40, 55 vs 70 years (reference category) as above.

I tested interactions between the eight core variables individually in fortnights where positivity was >0.5% (arbitrary threshold to avoid small numbers), conducting backwards elimination on all interactions with individual global heterogeneity p-value<0.001 (Bonferroni adjustment, 0.05/26 (number of interaction tests)), creating the "core model". An overall effect size was calculated for interactions as above, but taking the absolute coefficient values. Interactions between household size and multigenerational households, and region and rural/urban classification were not considered as, by definition, all those living in multigenerational households had a household size of 3 or more, and not all regions included major urban conurbations.

Given missing data (**Table 2.3; Table 2.4)**, I used forward selection to retain as many participants as possible when screening each non-core characteristic based on complete-cases, first adding each of the 47 "screening" variables individually to the "core model", thus estimating the total effects not explained by core characteristics. For all work-related variables, work status was included regardless

of significance so that effects reflected additional effects of the characteristic for those currently employed and working. To monitor multiple testing, I plotted observed p-values (global per variable and individual level vs reference) against expected p-values assuming no difference (randomly distributed between 0 and 1 given the number of tests), creating a Q-Q plot, including 0.05, Bonferroni and Benjamini-Hochberg adjusted p-values (0.05/tests) as references. As the goal was to identify signals of "at-risk" populations, I included all characteristics with either global $p<0.05$ or any level with $p<0.001$ vs reference, and then used backward elimination (exit $p=0.05$) to identify a final "main model". I used a similar methodology on the behavioural variables, also adjusting for variables identified from the main screen, regardless of significance. I categorised screening variables after backwards elimination into five broad groups dependent on the persistence of effects as follows:

- **Never:** The effect is never significant at a $p<0.05$ threshold in any fortnight
- **Inconsistent**: The variable is significant at a $p<0.05$ threshold in at least one fortnight, but never with an odds ratio in a consistent direction in any consecutive fortnights
- **Isolated**: The variable is significant at a $p<0.05$ threshold in two consecutive fortnights at most once, and "never consecutive" at all other times
- **Comes/goes**: The variable is significant at a $p<0.05$ threshold in three or more consecutive fortnights, or two consecutive fortnights at least twice, and is not significant with a gap of at least three fortnights, or two gaps of two fortnights, if the effect appears again.
- **Persistent**: The variable is significant at a $p<0.05$ threshold for the entire period after the first significant fortnight, with no more than one gap of two fortnights separating the consistency of the effect.

### 2.2.5 Sensitivity Analyses

To assess the impact of small numbers of positive tests in some fortnights on power, I repeated the process using 28-day periods. I also evaluated how many "true" effects (defined as seen in at least two consecutive fortnights), I would have detected earlier, later, or at the same time given the time period used.

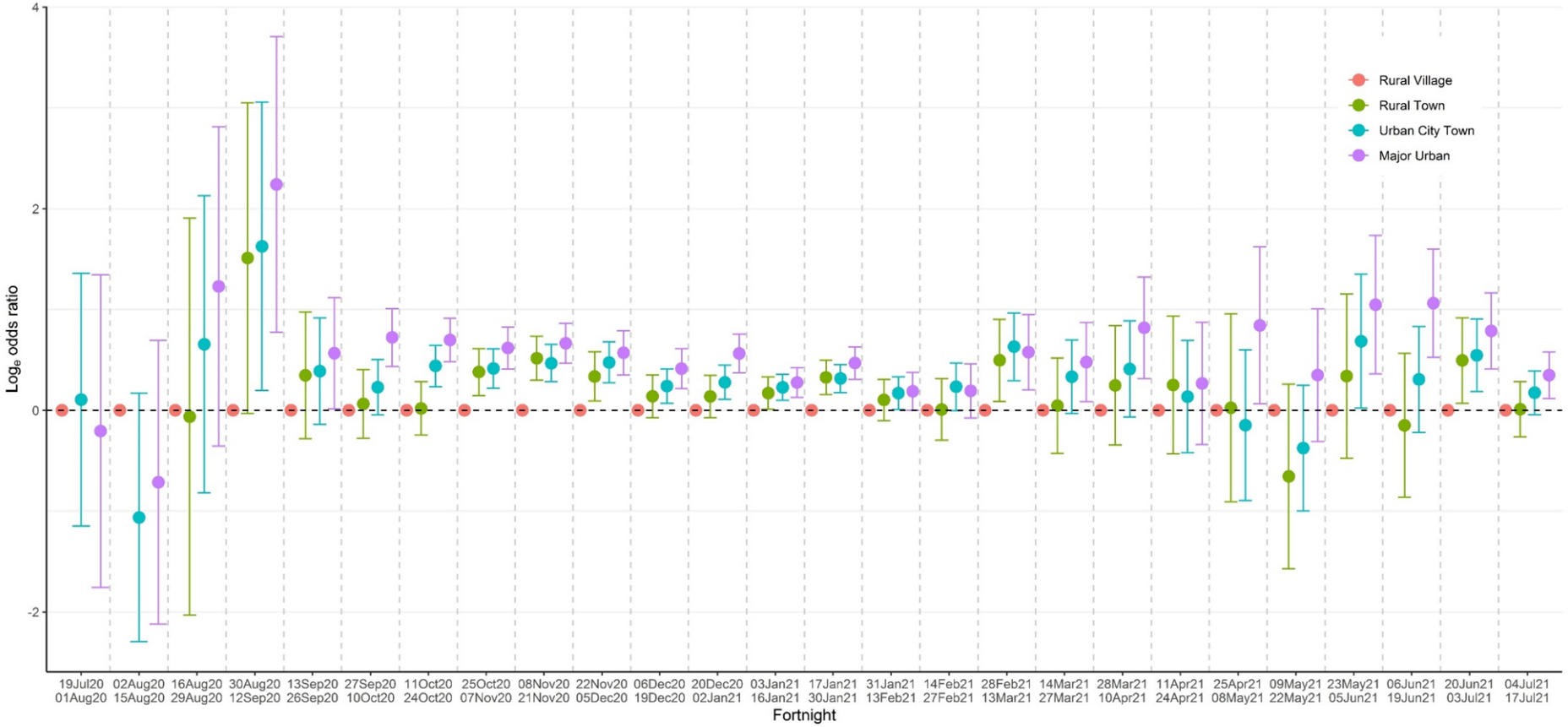
Given logistic regression can have higher bias and variability with low rates, and hence lose accuracy and precision,[100] I also compared the core variable estimates with those from ridge regression. This penalised regression technique can help overcome these issues by adding a penalty term with shrinks estimated coefficients towards zero. To compare, I quantified the number of ridge regression coefficients falling with the 95% confidence intervals of the logistic regression estimates.

**Table 2.1: Count of visits included in each fortnightly model, including the number not included in the core model.**

| Fortnight | Positive visits, n (%) | Negative visits, n (%) | Total, n (%) | Negative visits excluded from core models*, n (% of negatives) |
|---|---|---|---|---|
| 19Jul20 - 01Aug20 | 27 (0.1) | 32,157 (99.9) | 32,184 (100) | 4,074 (12.7) |
| 02Aug20 - 15Aug20 | 22 (0.1) | 43,073 (99.9) | 43,095 (100) | 8,672 (20.1) |
| 16Aug20 - 29Aug20 | 41 (0.1) | 57,895 (99.9) | 57,936 (100) | 0 (0.0) |
| 30Aug20 - 12Sep20 | 111 (0.1) | 76,276 (99.9) | 76,387 (100) | 0 (0.0) |
| 13Sep20 - 26Sep20 | 320 (0.3) | 116,467 (99.7) | 116,787 (100) | 0 (0.0) |
| 27Sep20 - 10Oct20 | 1,090 (0.6) | 171,298 (99.4) | 172,388 (100) | 0 (0.0) |
| 11Oct20 - 24Oct20 | 1,995 (1.0) | 194,123 (99.0) | 196,118 (100) | 0 (0.0) |
| 25Oct20 - 07Nov20 | 2,109 (1.2) | 169,735 (98.8) | 171,844 (100) | 0 (0.0) |
| 08Nov20 - 21Nov20 | 2,316 (1.2) | 192,715 (98.8) | 195,031 (100) | 0 (0.0) |
| 22Nov20 - 05Dec20 | 1,874 (1.0) | 192,534 (99.0) | 194,408 (100) | 0 (0.0) |
| 06Dec20 - 19Dec20 | 2,286 (1.2) | 190,313 (98.8) | 192,599 (100) | 0 (0.0) |
| 20Dec20 - 02Jan21 | 2,710 (1.9) | 136,703 (98.1) | 139,413 (100) | 0 (0.0) |
| 03Jan21 - 16Jan21 | 3,891 (1.9) | 198,116 (98.1) | 202,007 (100) | 0 (0.0) |
| 17Jan21 - 30Jan21 | 3,275 (1.7) | 194,157 (98.3) | 197,432 (100) | 0 (0.0) |
| 31Jan21 - 13Feb21 | 2,171 (1.0) | 205,148 (99.0) | 207,319 (100) | 0 (0.0) |
| 14Feb21 - 27Feb21 | 1,058 (0.5) | 196,410 (99.5) | 197,468 (100) | 0 (0.0) |
| 28Feb21 - 13Mar21 | 621 (0.3) | 193,549 (99.7) | 194,170 (100) | 0 (0.0) |
| 14Mar21 - 27Mar21 | 475 (0.3) | 173,734 (99.7) | 174,209 (100) | 0 (0.0) |
| 28Mar21 - 10Apr21 | 364 (0.2) | 169,692 (99.8) | 170,056 (100) | 0 (0.0) |
| 11Apr21 - 24Apr21 | 189 (0.1) | 164,958 (99.9) | 165,147 (100) | 0 (0.0) |
| 25Apr21 - 08May21 | 123 (0.1) | 172,931 (99.9) | 173,054 (100) | 0 (0.0) |
| 09May21 - 22May21 | 137 (0.1) | 164,249 (99.9) | 164,386 (100) | 0 (0.0) |
| 23May21 - 05Jun21 | 240 (0.1) | 160,888 (99.9) | 161,128 (100) | 0 (0.0) |
| 06Jun21 - 19Jun21 | 309 (0.2) | 167,862 (99.8) | 168,171 (100) | 0 (0.0) |
| 20Jun21 - 03Jul21 | 675 (0.4) | 159,246 (99.6) | 159,921 (100) | 0 (0.0) |
| 04Jul21 - 17Jul21 | 1,474 (0.9) | 167,405 (99.1) | 168,879 (100) | 0 (0.0) |

* Negative visits were excluded in the two earliest fortnights due to perfect prediction

**Figure 2.1: Log odds ratios with 95% confidence intervals for the effect of rural/urban classification (reference category rural village) across the 52-week study period**



## 2.3 Results

Analyses included 4,091,537 RT-PCR results from nose and throat swabs from 482,677 individuals (median (IQR) swabs per participant=9 (6-11)) in 240,490 households (median (IQR) swabs per household per fortnight=2 (1-2)) from 19th July 2020-17th July 2021. 29,903 (0·7%) swabs were positive. Characteristics of included visits are shown in **Table 2.2** (core variables) and **Table 2.3** and **Table 2.4** (screening variables). Overall, the median (IQR) age was 52 years (33-66), 7% visits occurred in those reporting non-white ethnicity, 53% in females, 36% in those residing in major urban areas and 43% in urban cities/towns, most (42%) in two-person households, and with a median deprivation percentile of 60 (34-81) (1=most deprived, 100=least deprived) (**Table 2.2**; screened variables **Table 2.3**, **Table 2.4**). The highest fortnightly positivity was 1.9% (95% CI 1.9-2.0%) from 20th December to 2nd January 2020, and the lowest was 0.05% (0.03-0.08%) from 2nd to 15th August 2020 (**Figure 2.2**). Numbers within each fortnight increased as the study expanded from August to October 2020,[101] with 32,184 participants from 19th July to 1st August 2020 to a median of 173,054 (IQR 168,171-195,031) from 27th September 2020 onwards (**Figure 2.3**).

**Table 2.2 Characteristics of the core variables for visits included in analysis.**

| Characteristic | Positive, n (%) or median (IQR) | Negative, n (%) or median (IQR) | Total, n (%) or median (IQR) |
|---|---|---|---|
| **Age (years)** | 43 (23, 58) | 52 (33, 66) | 52 (33, 66) |
| **Sex** | | | |
| Male | 14,405 (48) | 1,911,299 (47) | 1,925,704 (47) |
| Female | 15,498 (52) | 2,150,335 (53) | 2,165,833 (53) |
| **Ethnicity** | | | |
| White | 26,702 (89) | 3,764,627 (93) | 3,791,329 (93) |
| Non-White | 3,201 (11) | 297,007 (7) | 300,208 (7) |
| **Deprivation percentile** (1=most deprived, 100=least deprived) | 54 (29, 78) | 60 (34, 81) | 60 (34, 81) |
| **Household (HH) size** | | | |
| One | 3,842 (13) | 675,623 (17) | 679,465 (17) |
| Two | 10,124 (34) | 1,725,494 (42) | 1,735,618 (42) |
| Three | 5,797 (19) | 657,828 (16) | 663,625 (16) |
| Four | 6,639 (22) | 686,036 (17) | 692,675 (17) |
| Five or more | 3,501 (12) | 316,653 (8) | 320,154 (8) |
| **Multigenerational HH** | | | |
| No | 27,311 (91) | 3,796,655 (93) | 3,823,966 (93) |
| Yes | 2,592 (9) | 264,979 (7) | 267,571 (7) |
| **Rural/urban classification** | | | |
| Major urban area | 14,044 (47) | 1,449,580 (36) | 1,463,624 (36) |
| Urban city/town | 11,425 (38) | 1,735,105 (43) | 1,746,530 (43) |
| Rural town | 2,445 (8) | 435,296 (11) | 437,741 (11) |
| Rural village | 1,989 (7) | 441,653 (11) | 443,642 (11) |
| **Region** | | | |
| London | 6,498 (22) | 698,608 (17) | 705,106 (17) |
| North West England | 5,077 (17) | 477,380 (12) | 482,457 (12) |
| North East England | 1,390 (5) | 156,119 (4) | 157,509 (4) |

| Characteristic | Positive, n (%) or median (IQR) | Negative, n (%) or median (IQR) | Total, n (%) or median (IQR) |
|---|---|---|---|
| Yorkshire | 2,861 (10) | 343,353 (8) | 346,214 (8) |
| West Midlands | 2,266 (8) | 311,661 (8) | 313,927 (8) |
| East Midlands | 1,893 (6) | 264,293 (7) | 266,186 (7) |
| South East England | 2,986 (10) | 531,594 (13) | 534,580 (13) |
| South West England | 1,332 (4) | 320,869 (8) | 322,201 (8) |
| East England | 2,425 (8) | 405,304 (10) | 407,729 (10) |
| Northern Ireland | 665 (2) | 106,660 (3) | 107,325 (3) |
| Wales | 969 (3) | 179,900 (4) | 180,869 (4) |
| Scotland | 1,541 (5) | 265,893 (7) | 267,434 (7) |

**Table 2.3: Characteristics of screening variables for visits included in the main screen.**

| Characteristic | Positive, n (%) or median (IQR) N=29,903 | Negative, n (%) or median (IQR) N=4,061,634 | Total, n (%) or median (IQR) N=4,091,537 |
|---|---|---|---|
| **Contact with other people** | | | |
| **Contact with known Covid-19 (last 28 days)** | | | |
| No | 13,999 (47) | 3,640,835 (90) | 3,654,834 (89) |
| Yes | 15,904 (53) | 420,799 (10) | 436,703 (11) |
| Missing | 0 (0) | 0 (0) | 0 (0) |
| **Contact hospital (last 28 days)** | | | |
| No | 22,699 (76) | 3,124,538 (77) | 3,147,237 (77) |
| Yes, I have | 3,677 (12) | 500,711 (12) | 504,388 (12) |
| No, but someone in my household has | 2,967 (10) | 359,387 (9) | 362,354 (9) |
| Missing | 560 (2) | 76,998 (2) | 77,558 (2) |
| **Contact care home (last 28 days)** | | | |
| No | 28,007 (94) | 3,825,176 (94) | 3,853,183 (94) |
| Yes, I have | 623 (2) | 77,503 (2) | 78,126 (2) |
| No, but someone in my household has | 592 (2) | 67,317 (2) | 67,909 (2) |
| Missing | 681 (2) | 91,638 (2) | 92,319 (2) |
| **Travel abroad in the last 28 days** | | | |
| No | 29,662 (99) | 4,034,194 (99) | 4,063,856 (99) |
| Yes | 241 (1) | 27,440 (1) | 27,681 (1) |
| Missing | 0 (0) | 0 (0) | 0 (0) |
| **Face covering** | | | |
| Yes, other situations only | 15,479 (52) | 2,394,819 (59) | 2,410,298 (59) |
| Yes, work and other situations | 10,254 (34) | 1,224,461 (30) | 1,234,715 (30) |
| Yes, work only | 471 (2) | 40,593 (1) | 41,064 (1) |
| Yes, face already covered | 632 (2) | 52,980 (1) | 53,612 (1) |
| No | 1,746 (6) | 188,210 (5) | 189,956 (5) |
| Missing | 1,321 (4) | 160,571 (4) | 161,892 (4) |
| **Face covering (binary)** | | | |
| Yes (any) | 26,836 (90) | 3,712,853 (91) | 3,739,689 (91) |
| No | 1,746 (6) | 188,210 (5) | 189,956 (5) |
| Missing | 1,321 (4) | 160,571 (4) | 161,892 (4) |
| **Visit frequency‡** | | | |
| Last visit >14 days ago | 19,043 (64) | 2,863,978 (71) | 2,883,021 (70) |
| Last visit ≤ 14 days ago | 7,852 (26) | 916,167 (23) | 924,019 (23) |
| Enrolment | 3,008 (10) | 281,489 (7) | 284,497 (7) |
| Missing | 0 (0) | 0 (0) | 0 (0) |
| **Household and living environment** | | | |
| **IMD indoors*** | 50 (27, 73) | 51 (27, 75) | 51 (27, 75) |
| Missing | 3,175 (11) | 552,453 (14) | 555,628 (14) |
| **IMD outdoors*** | 44 (20, 71) | 51 (26, 76) | 51 (26, 76) |

| Characteristic | Positive, n (%) or median (IQR) N=29,903 | Negative, n (%) or median (IQR) N=4,061,634 | Total, n (%) or median (IQR) N=4,091,537 |
|---|---|---|---|
| Missing | 3,175 (11) | 552,453 (14) | 555,628 (14) |
| **IMD education*** | 59 (34, 82) | 64 (39, 84) | 64 (39, 84) |
| Missing | 3,175 (11) | 552,453 (14) | 555,628 (14) |
| **IMD health*** | 55 (29, 78) | 62 (37, 82) | 62 (37, 82) |
| Missing | 3,175 (11) | 552,453 (14) | 555,628 (14) |
| **IMD crime*** | 49 (26, 72) | 57 (32, 79) | 57 (32, 79) |
| Missing | 3,175 (11) | 552,453 (14) | 555,628 (14) |
| **IMD housing*** | 49 (26, 75) | 49 (25, 75) | 49 (25, 75) |
| Missing | 3,175 (11) | 552,453 (14) | 555,628 (14) |
| **Number of people per room*** | 1 (0, 1) | 1 (0, 1) | 1 (0, 1) |
| Missing | 3,633 (12) | 562,589 (14) | 566,222 (14) |
| **Number of people per bedroom*** | 1 (1, 1) | 1 (1, 1) | 1 (1, 1) |
| Missing | 3,640 (12) | 562,795 (14) | 566,435 (14) |
| **Number of people per 100m2*** | 3 (2, 4) | 2 (2, 3) | 2 (2, 3) |
| Missing | 3,669 (12) | 566,273 (14) | 569,942 (14) |
| **Energy efficiency decile*** | 6 (3, 10) | 6 (3, 10) | 6 (3, 10) |
| Missing | 3,526 (12) | 551,740 (14) | 555,266 (14) |
| **Age of house (decades) *** | 3 (3, 4) | 3 (3, 4) | 3 (3, 4) |
| Missing | 14,910 (50) | 2,174,664 (54) | 2,189,574 (54) |
| **Work, school, and nursery[§]** | | | |
| **Work status** | | | |
| Employed, working | 14,713 (49) | 1,832,299 (45) | 1,847,012 (45) |
| Employed, not working | 1,858 (6) | 134,876 (3) | 136,734 (3) |
| Not working | 1,631 (5) | 213,550 (5) | 215,181 (5) |
| Retired | 5,455 (18) | 1,281,213 (32) | 1,286,668 (31) |
| Child/student | 6,239 (21) | 599,352 (15) | 605,591 (15) |
| Missing | 7 (0) | 344 (0) | 351 (0) |
| **Work location** | | | |
| Working from home | 7,868 (26) | 1,005,480 (25) | 1,013,348 (25) |
| Elsewhere | 12,528 (42) | 1,433,415 (35) | 1,445,943 (35) |
| NA | 8,511 (28) | 1,537,192 (38) | 1,545,703 (38) |
| Missing | 996 (3) | 85,547 (2) | 86,543 (2) |
| **Work social distancing** | | | |
| Working from home | 7,868 (26) | 1,005,480 (25) | 1,013,348 (25) |
| Elsewhere, easy to maintain 2m | 3,239 (11) | 437,667 (11) | 440,906 (11) |
| Elsewhere, relatively easy to maintain 2m | 1,826 (6) | 214,528 (5) | 216,354 (5) |
| Elsewhere, difficult to maintain 2m | 2,004 (7) | 214,690 (5) | 216,694 (5) |
| Elsewhere, very difficult to maintain 1m | 4,247 (14) | 449,980 (11) | 454,227 (11) |
| NA | 8,511 (28) | 1,537,192 (38) | 1,545,703 (38) |
| Missing | 2,208 (7) | 202,097 (5) | 204,305 (5) |
| **Work travel[†]** | | | |
| Working from home | 7,868 (26) | 1,005,480 (25) | 1,013,348 (25) |
| On foot/bike or other | 2,616 (9) | 295,024 (7) | 297,640 (7) |
| Car/taxi | 7,986 (27) | 937,529 (23) | 945,515 (23) |
| Train/bus | 1,413 (5) | 137,124 (3) | 138,537 (3) |
| NA | 8,511 (28) | 1,537,192 (38) | 1,545,703 (38) |
| Missing | 1,509 (5) | 149,285 (4) | 150,794 (4) |
| **Work in direct contact with patients, service users, clients, customers** | | | |
| No | 25,962 (87) | 3,630,423 (89) | 3,656,385 (89) |
| Yes | 3,685 (12) | 404,714 (10) | 408,399 (10) |
| Missing | 256 (1) | 26,497 (1) | 26,753 (1) |
| **Ever reported working in person facing social care** | | | |

| Characteristic | Positive, n (%) or median (IQR) N=29,903 | Negative, n (%) or median (IQR) N=4,061,634 | Total, n (%) or median (IQR) N=4,091,537 |
|---|---|---|---|
| No | 29,464 (99) | 4,020,303 (99) | 4,049,767 (99) |
| Yes | 439 (1) | 41,331 (1) | 41,770 (1) |
| Missing | 0 (0) | 0 (0) | 0 (0) |
| **Ever reported working in a care home** | | | |
| No | 29,426 (98) | 4,019,274 (99) | 4,048,700 (99) |
| Yes | 477 (2) | 42,360 (1) | 42,837 (1) |
| Missing | 0 (0) | 0 (0) | 0 (0) |
| **Ever reported working in patient-facing healthcare** | | | |
| No | 29,031 (97) | 3,970,666 (98) | 3,999,697 (98) |
| Yes | 872 (3) | 90,968 (2) | 91,840 (2) |
| Missing | 0 (0) | 0 (0) | 0 (0) |
| **Work sector** | | | |
| Teaching and education | 2,832 (9) | 295,102 (7) | 297,934 (7) |
| Health care | 2,034 (7) | 225,167 (6) | 227,201 (6) |
| Social care | 534 (2) | 60,746 (1) | 61,280 (1) |
| Transport (incl. storage, logistics) | 752 (3) | 77,628 (2) | 78,380 (2) |
| Retail sector (incl. wholesale) | 1,384 (5) | 150,473 (4) | 151,857 (4) |
| Hospitality (e.g. hotel, restaurant) | 705 (2) | 67,521 (2) | 68,226 (2) |
| Food production, agriculture, farming | 268 (1) | 35,235 (1) | 35,503 (1) |
| Personal services (e.g. hairdressers) | 235 (1) | 27,437 (1) | 27,672 (1) |
| Information technology and communication | 1,014 (3) | 148,805 (4) | 149,819 (4) |
| Financial services incl. insurance | 1,303 (4) | 168,590 (4) | 169,893 (4) |
| Manufacturing or construction | 1,737 (6) | 195,676 (5) | 197,413 (5) |
| Civil service or Local Government | 1,087 (4) | 143,774 (4) | 144,861 (4) |
| Armed forces | 50 (0) | 6,847 (0) | 6,897 (0) |
| Arts, Entertainment or Recreation | 399 (1) | 55,956 (1) | 56,355 (1) |
| Other occupation sector | 2,341 (8) | 324,118 (8) | 326,459 (8) |
| NA (not currently working) | 9,863 (33) | 1,534,348 (38) | 1,544,211 (38) |
| Missing | 3,365 (11) | 544,211 (13) | 547,576 (13) |
| **Additional paid employment** | | | |
| No | 10,342 (35) | 2,241,224 (55) | 2,251,566 (55) |
| Yes | 127 (0) | 21,981 (1) | 22,108 (1) |
| Missing | 19,434 (65) | 1,798,429 (44) | 1,817,863 (44) |
| **Current health status** | | | |
| **Think have had COVID-19 (last 90 days)** | | | |
| No | 10,288 (34) | 3,970,284 (98) | 3,980,572 (97) |
| Yes | 19,615 (66) | 91,350 (2) | 110,965 (3) |
| Missing | 0 (0) | 0 (0) | 0 (0) |
| **Self-isolating** | | | |
| No | 20,121 (67) | 3,804,735 (94) | 3,824,856 (93) |
| Yes, I or some in my HH is | 8,003 (27) | 24,497 (1) | 32,500 (1) |
| Yes, other reasons | 845 (3) | 74,019 (2) | 74,864 (2) |
| Missing | 934 (3) | 158,383 (4) | 159,317 (4) |
| **Smoke now** | | | |
| Non-smoker | 27,520 (92) | 3,695,283 (91) | 3,722,803 (91) |
| Tobacco smoker | 1,583 (5) | 268,245 (7) | 269,828 (7) |
| Only vape | 693 (2) | 82,037 (2) | 82,730 (2) |
| Missing | 107 (0) | 16,069 (0) | 16,176 (0) |
| **Smoke ever regularly** | | | |
| No | 22,120 (74) | 2,843,859 (70) | 2,865,979 (70) |
| Yes | 7,283 (24) | 1,139,616 (28) | 1,146,899 (28) |
| Missing | 500 (2) | 78,159 (2) | 78,659 (2) |

| Characteristic | Positive, n (%) or median (IQR) N=29,903 | Negative, n (%) or median (IQR) N=4,061,634 | Total, n (%) or median (IQR) N=4,091,537 |
|---|---|---|---|
| **Any disability** | | | |
| No | 26,607 (89) | 3,513,264 (86) | 3,539,871 (87) |
| Yes | 3,296 (11) | 548,370 (14) | 551,666 (13) |
| Missing | 0 (0) | 0 (0) | 0 (0) |
| **Long-term health conditions** | | | |
| No | 24,755 (83) | 3,243,863 (80) | 3,268,618 (80) |
| Yes | 4,765 (16) | 751,236 (18) | 756,001 (18) |
| Missing | 383 (1) | 66,535 (2) | 66,918 (2) |
| **Impact of health conditions** | | | |
| No health conditions | 24,755 (83) | 3,243,863 (80) | 3,268,618 (80) |
| No impact at all | 2,164 (7) | 332,664 (8) | 334,828 (8) |
| A little impact | 1,526 (5) | 239,834 (6) | 241,360 (6) |
| A lot of impact | 1,017 (3) | 172,191 (4) | 173,208 (4) |
| Missing | 441 (1) | 73,082 (2) | 73,523 (2) |
| **Covid vaccination status** | | | |
| Not vaccinated, no prior positive, >21 days before vaccination | 25,254 (84) | 2,431,522 (60) | 2,456,776 (60) |
| 1-21 days before vaccination or 0-7 days post-vaccination | 1,422 (5) | 313,585 (8) | 315,007 (8) |
| Vaccinated 8-20 days ago | 665 (2) | 141,629 (3) | 142,294 (3) |
| Vaccinated ≥ 21 days ago, no second dose | 1,162 (4) | 495,471 (12) | 496,633 (12) |
| Post second dose or not vaccinated before positive | 1,400 (5) | 679,427 (17) | 680,827 (17) |
| Missing | 0 (0) | 0 (0) | 0 (0) |
| **Regular LFT testing** | | | |
| No | 719 (2) | 59,169 (1) | 59,888 (1) |
| Yes | 1,055 (4) | 116,773 (3) | 117,828 (3) |
| Missing | 28,129 (94) | 3,885,692 (96) | 3,913,821 (96) |

*Characteristic available for England only

** Question introduced or expanded part way through the study so missing data also reflects time periods when the question was not included.

† 6,744/945,515 visits in the car/taxi group were taxi; numbers were too few to assess whether another grouping might be preferable.

‡ Visit frequency was calculated based on completed survey visits

§ Questions on work, such as work location and work social distancing were asked phrased as "if currently working".

Note: For more details on the questions from which these characteristics were derived, the current questionnaire used in the survey, as well as all previous versions can be found at: https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/case-record-forms.

**Table 2.4: Characteristics of screening variables for visits included in the behaviour screen.**

| Characteristic† | Positive, n (%) or median (IQR) N=29,903 | Negative, n (%) or median (IQR) N=4,061,634 | Total, n (%) or median (IQR) N=4,091,537 |
|---|---|---|---|
| **Number of physical contacts aged <18y** | | | |
| 0 | 11,898 (40) | 2,160,467 (53) | 2,172,365 (53) |
| 1-5 | 4,146 (14) | 608,127 (15) | 612,273 (15) |
| 6-10 | 675 (2) | 71,849 (2) | 725,24 (2) |
| 11-20 | 2,294 (8) | 206,076 (5) | 208,370 (5) |
| 21 or more | 10,890 (36) | 1,015,115 (25) | 1,026,005 (25) |
| **Number of physical contacts aged 18-69y** | | | |
| 0 | 10,031 (34) | 1,848,906 (46) | 1,858,937 (45) |

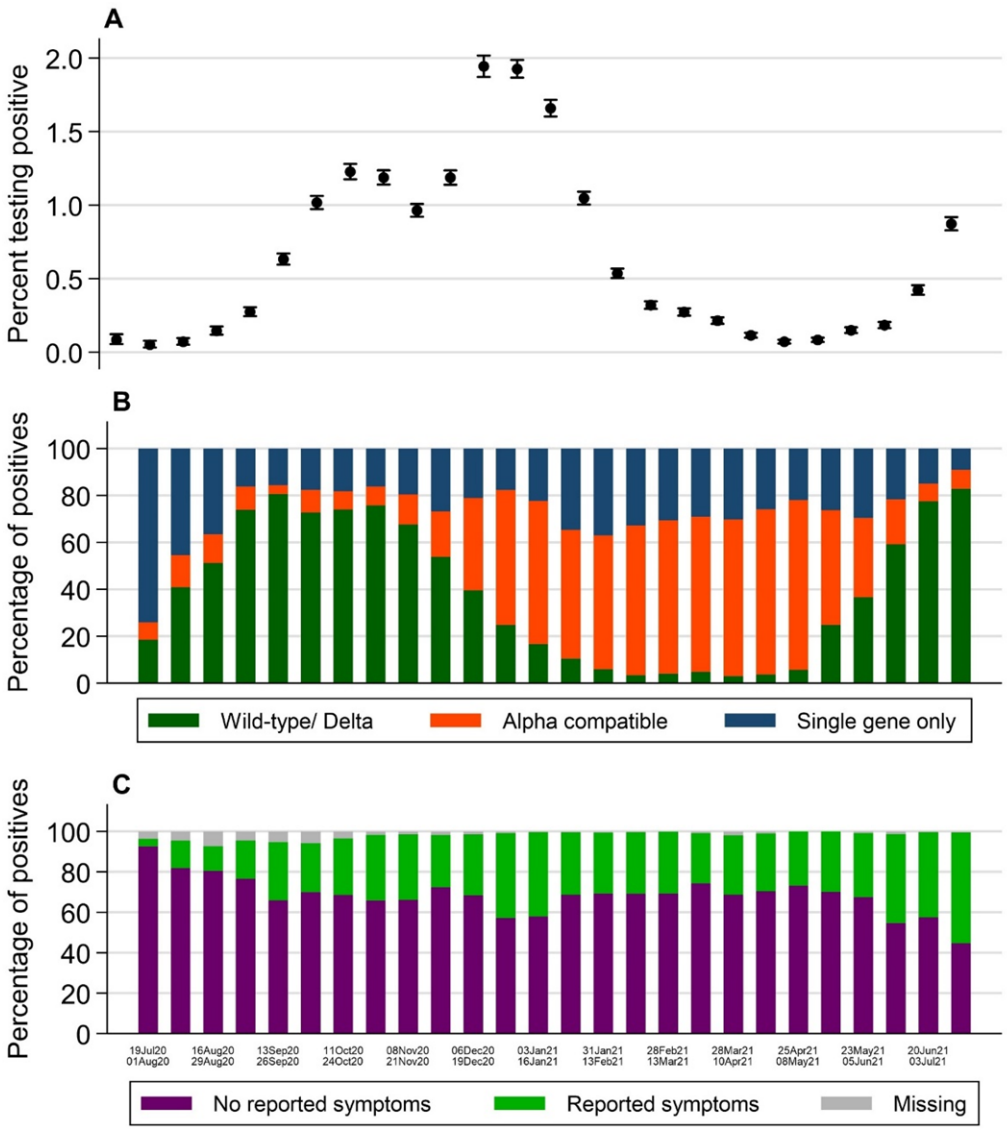
| Characteristic† | Positive, n (%) or median (IQR) N=29,903 | Negative, n (%) or median (IQR) N=4,061,634 | Total, n (%) or median (IQR) N=4,091,537 |
|---|---|---|---|
| 1-5 | 6,487 (22) | 950,800 (23) | 957,287 (23) |
| 6-10 | 1,233 (4) | 128,817 (3) | 130,050 (3) |
| 11-20 | 1,269 (4) | 119,866 (3) | 121,135 (3) |
| 21 or more | 10,883 (36) | 1,013,245 (25) | 1,024,128 (25) |
| **Number of physical contacts aged ≥70y** | | | |
| 0 | 15,293 (51) | 2,530,655 (62) | 2,545,948 (62) |
| 1-5 | 3,034 (10) | 449,008 (11) | 452,042 (11) |
| 6-10 | 205 (1) | 22,434 (1) | 22,639 (1) |
| 11-20 | 423 (1) | 41,165 (1) | 41,588 (1) |
| 21 or more | 10,948 (37) | 1,018,372 (25) | 1,029,320 (25) |
| **Number of social contacts aged <18y** | | | |
| 0 | 12,138 (41) | 1,935,681 (48) | 1,947,819 (48) |
| 1-5 | 4,797 (16) | 835,491 (21) | 840,288 (21) |
| 6-10 | 696 (2) | 106,921 (3) | 107,617 (3) |
| 11-20 | 1,294 (4) | 163,396 (4) | 164,690 (4) |
| 21 or more | 10,978 (37) | 1,020,145 (25) | 1,031,123 (25) |
| **Number of social contacts aged 18-69y** | | | |
| 0 | 4,243 (14) | 803,071 (20) | 807,314 (20) |
| 1-5 | 6,351 (21) | 1,191,642 (29) | 1,197,993 (29) |
| 6-10 | 3,033 (10) | 425,739 (10) | 428,772 (10) |
| 11-20 | 5,385 (18) | 628,365 (15) | 633,750 (15) |
| 21 or more | 10,891 (36) | 1,012,817 (25) | 1,023,708 (25) |
| **Number of social contacts aged ≥70y** | | | |
| 0 | 12,138 (41) | 1,935,681 (48) | 1,947,819 (48) |
| 1-5 | 4,797 (16) | 835,491 (21) | 840,288 (21) |
| 6-10 | 696 (2) | 106,921 (3) | 107,617 (3) |
| 11-20 | 1,294 (4) | 163,396 (4) | 164,690 (4) |
| 21 or more | 10,978 (37) | 1,020,145 (25) | 1,031,123 (25) |
| **Outside socialising times** | | | |
| None | 409 (1) | 80,290 (2) | 80,699 (2) |
| Once | 345 (1) | 52,733 (1) | 53,078 (1) |
| Twice | 208 (1) | 28,400 (1) | 28,608 (1) |
| Three times | 128 (0) | 14,056 (0) | 14,184 (0) |
| Four times | 54 (0) | 6,834 (0) | 6,888 (0) |
| Five times | 52 (0) | 4,194 (0) | 4,246 (0) |
| Six times | 19 (0) | 1,468 (0) | 1,487 (0) |
| Seven times or more | 43 (0) | 4,718 (0) | 4,761 (0) |
| Missing | 28,645 (96) | 3,868,941 (95) | 3,897,586 (95) |
| **Outside shopping only times** | | | |
| None | 260 (1) | 32,514 (1) | 32,774 (1) |
| Once | 297 (1) | 47,098 (1) | 47,395 (1) |
| Twice | 291 (1) | 48,764 (1) | 49,055 (1) |
| Three times | 180 (1) | 30,207 (1) | 30,387 (1) |
| Four times | 84 (0) | 13,948 (0) | 14,032 (0) |
| Five times | 56 (0) | 7,835 (0) | 7,891 (0) |
| Six times | 14 (0) | 2,663 (0) | 2,677 (0) |
| Seven times or more | 76 (0) | 9,669 (0) | 9,745 (0) |
| Missing | 28,645 (96) | 3,868,936 (95) | 3,897,581 (95) |
| **Time spent shopping or socializing outside** | | | |
| None | 3,180 (11) | 513,784 (13) | 516,964 (13) |
| Once | 3,687 (12) | 634,651 (16) | 638,338 (16) |
| Twice | 3,719 (12) | 602,006 (15) | 605,725 (15) |
| Three times | 2,236 (7) | 356,644 (9) | 358,880 (9) |

| Characteristic† | Positive, n (%) or median (IQR) N=29,903 | Negative, n (%) or median (IQR) N=4,061,634 | Total, n (%) or median (IQR) N=4,091,537 |
|---|---|---|---|
| Four times | 1,133 (4) | 180,386 (4) | 181,519 (4) |
| Five times | 737 (2) | 111,370 (3) | 112,107 (3) |
| Six times | 293 (1) | 44,966 (1) | 45,259 (1) |
| Seven times or more | 1,196 (4) | 177,612 (4) | 178,808 (4) |
| Missing | 13,722 (46) | 1,440,215 (35) | 1,453,937 (36) |
| **Hours spent in other's homes** | | | |
| None | 11,597 (39) | 1,954,027 (48) | 1,965,624 (48) |
| Once | 2,619 (9) | 397,279 (10) | 399,898 (10) |
| Twice | 830 (3) | 125,436 (3) | 126,266 (3) |
| Three | 356 (1) | 49,845 (1) | 50,201 (1) |
| Four | 175 (1) | 23,312 (1) | 23,487 (1) |
| Five | 129 (0) | 19,331 (0) | 19,460 (0) |
| Six | 49 (0) | 6,566 (0) | 6,615 (0) |
| Seven or more | 261 (1) | 34,453 (1) | 34,714 (1) |
| Missing | 13,887 (46) | 1,451,385 (36) | 1,465,272 (36) |
| **Hours others spent in own home** | | | |
| None | 10,753 (36) | 1,780,000 (44) | 1,790,753 (44) |
| Once | 2,906 (10) | 490,148 (12) | 493,054 (12) |
| Twice | 1,139 (4) | 171,857 (4) | 172,996 (4) |
| Three times | 502 (2) | 69,515 (2) | 70,017 (2) |
| Four times | 201 (1) | 32,278 (1) | 32,479 (1) |
| Five times | 180 (1) | 23,641 (1) | 23,821 (1) |
| Six times | 60 (0) | 7,818 (0) | 7,878 (0) |
| Seven times or more | 254 (1) | 32,741 (1) | 32,995 (1) |
| Missing | 13,908 (47) | 1,453,636 (36) | 1,467,544 (36) |

† All characteristics except hours spent with someone else in one's own home per day related to the past 7 days.

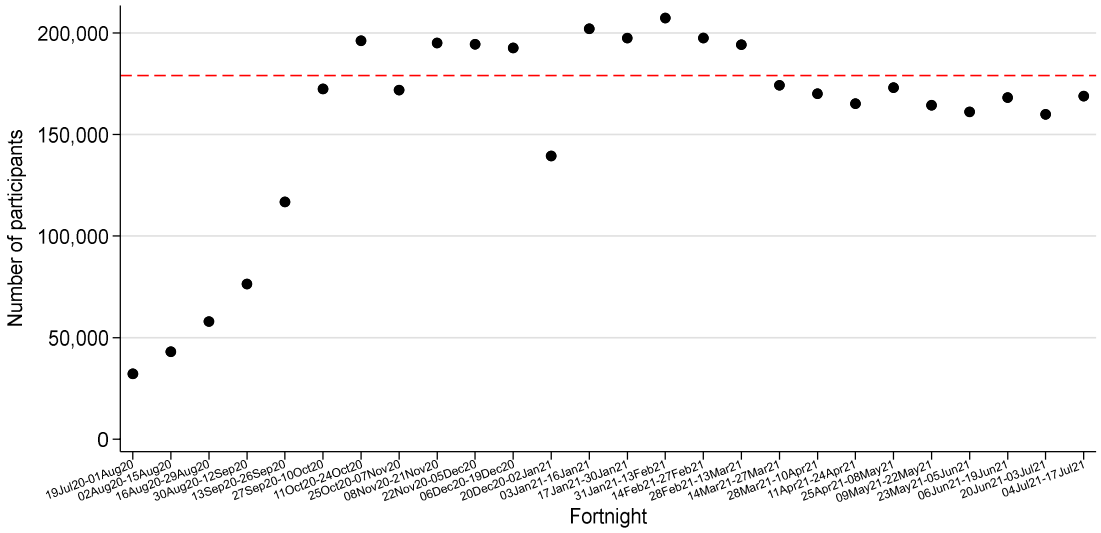
Note: To distinguish between physical and social contacts, participants were asked either "how many adults not living in your home have you had physical contact with (e.g. handshake, personal care), including with PPE if you wear it?" or "how many adults not living in your home have you had direct, but not physical contact with in person, e.g. with social distancing only?"

**Figure 2.2: Unadjusted percentage (95% CI) of positive swabs per fortnight (A), positive swabs split by gene positivity pattern (B), and symptoms (C).**




Note: Wild-type/Delta compatible=positive on all three genes (N, S, ORF1ab) or S plus one other gene. Alpha-compatible=positive on N+ORF1ab. Single gene=positive on N or ORF1ab only (S only not considered positive). A separate manuscript investigated associations with symptoms in positives and negatives in detail.[102]

**Figure 2.3: Total number of participants per fortnight.**




Note: The red dashed line shows the recruitment target of 179,000 swabs from unique participants across the UK from 1st October onwards

### 2.3.1 Core model

From 19th July to 1st August 2020, there was no evidence that any core variable was associated with positivity, potentially related to power given both low positivity (0.08% [95% CI 0.06%-0.12%]) and sample size (32,184 swabs, 27 positive). The first characteristic associated with positivity was ethnicity, the only characteristic associated with positivity in the fortnights between 2nd to 29th August 2020 (**Figure 2.4A**), with 3.3 (1.1-10.0; p-value=0.034) and 3.5 (1.5-7.9; p-value=0.003) higher odds of positivity in those of non-white ethnicity, respectively.

As positivity began to increase in early September 2020, geographical region, rural/urban classification, and household size became independently associated with positivity, with odds of positivity highest in Wales, Northern Ireland, and northern English regions, in more urban areas, and those living in larger households (**Figure 2.4B**). For most subsequent fortnights, evidence of higher positivity persisted in participants living in more urban areas, and larger households.

As positivity rates rose further through October 2020, age and deprivation became associated with positivity, with rates highest in those aged 16-30y, and living in more deprived areas. Positivity was also heavily concentrated in northern and then midland English regions until 21st November 2020. From 22nd November, positivity increased overall, particularly in southern England, with higher odds of positivity in London, East, and South East England, reflecting the rise of the Alpha variant.[103] Age remained strongly associated with positivity, but with less excess risk at younger ages, and instead decreased odds of positivity in those over 60y (**Figure 2.4B, Figure 2.5**). This lower risk in older

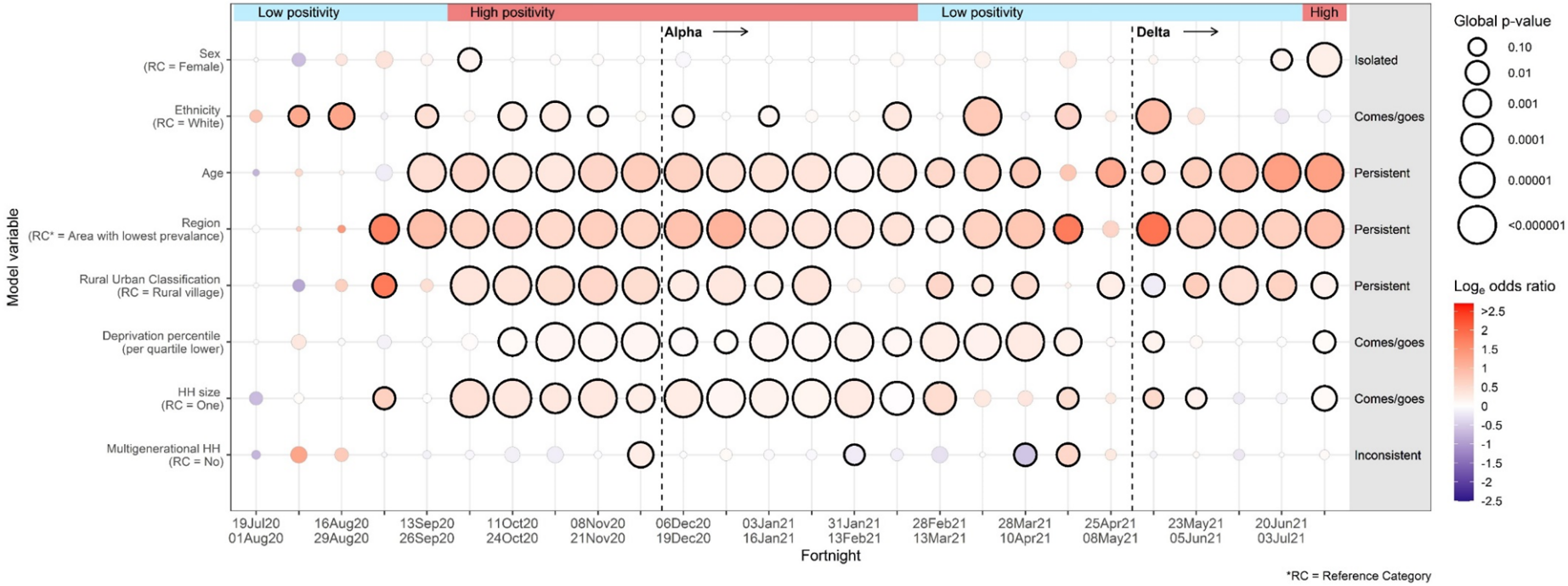
individuals persisted for most subsequent fortnights. During February-May 2021, as positivity decreased, associations between positivity and age, region, and deprivation persisted, but their strength attenuated. As positivity rose during 17th May-17th July 2021, reflecting the rise of the Delta variant[104] and major sporting events, sex was associated with positivity in two consecutive fortnights for the first time in the study, with higher odds in males compared with females. Age again became strongly associated, with a large peak in those aged 16-30y (**Figure 2.5**).

Overall, the effects of age, region, and rural/urban classification were persistent throughout the time period, while effects of deprivation and household size remained significant during periods of high positivity but fewer associations were found when positivity was low. Effects of ethnicity also came/went throughout the period of analysis; those reporting non-white ethnicity always had higher positivity than those reporting white ethnicity but this was not clearly linked with periods of high positivity. The effect of sex was isolated only to the last two fortnights in the study period. Multigenerational households were never consistently associated with positivity.

Few interactions between core variables were significant at the p=0.001 threshold, with no evidence of the same significant interactions in any consecutive fortnight (**Figure 2.6**). For model comparability, none were therefore included in any fortnight for screening other variables.

**Figure 2.4: Effects of the 8 core variables across the 52-week study period**

*(A) Overall effects*




Note: RC=reference category. HH=household size. The size of the circles is proportional to -$\log_{10}$ of the global p-value for each variable in each fortnight. Circles with black outlines indicate p<0·05. The colour of the circles represents the size of the odds ratio (vs the reference category shown). See Methods for details on combined effects for categorical variables with >2 levels, and details of classification as isolated, persistent etc.

*(B) Individual level effects*

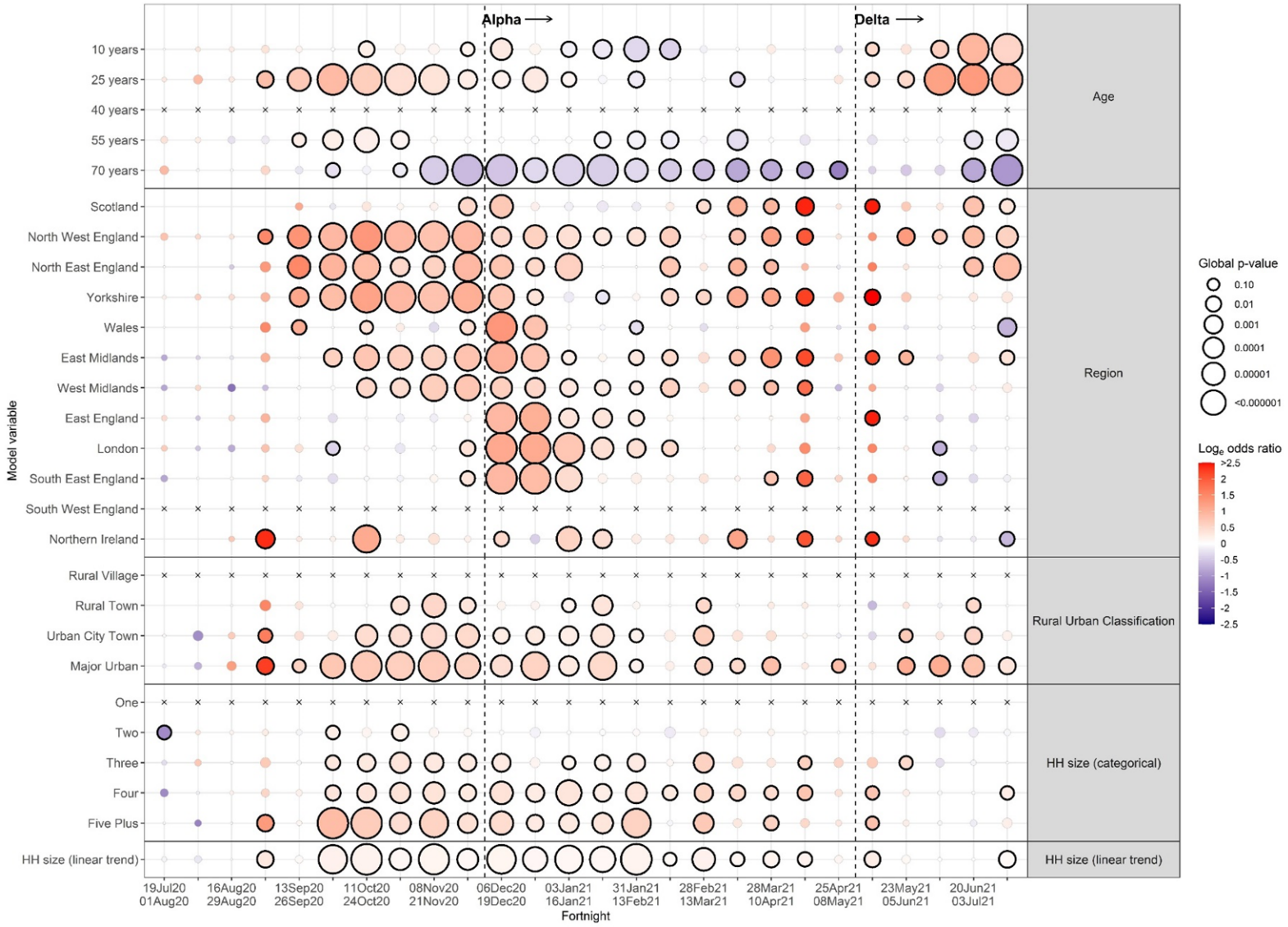
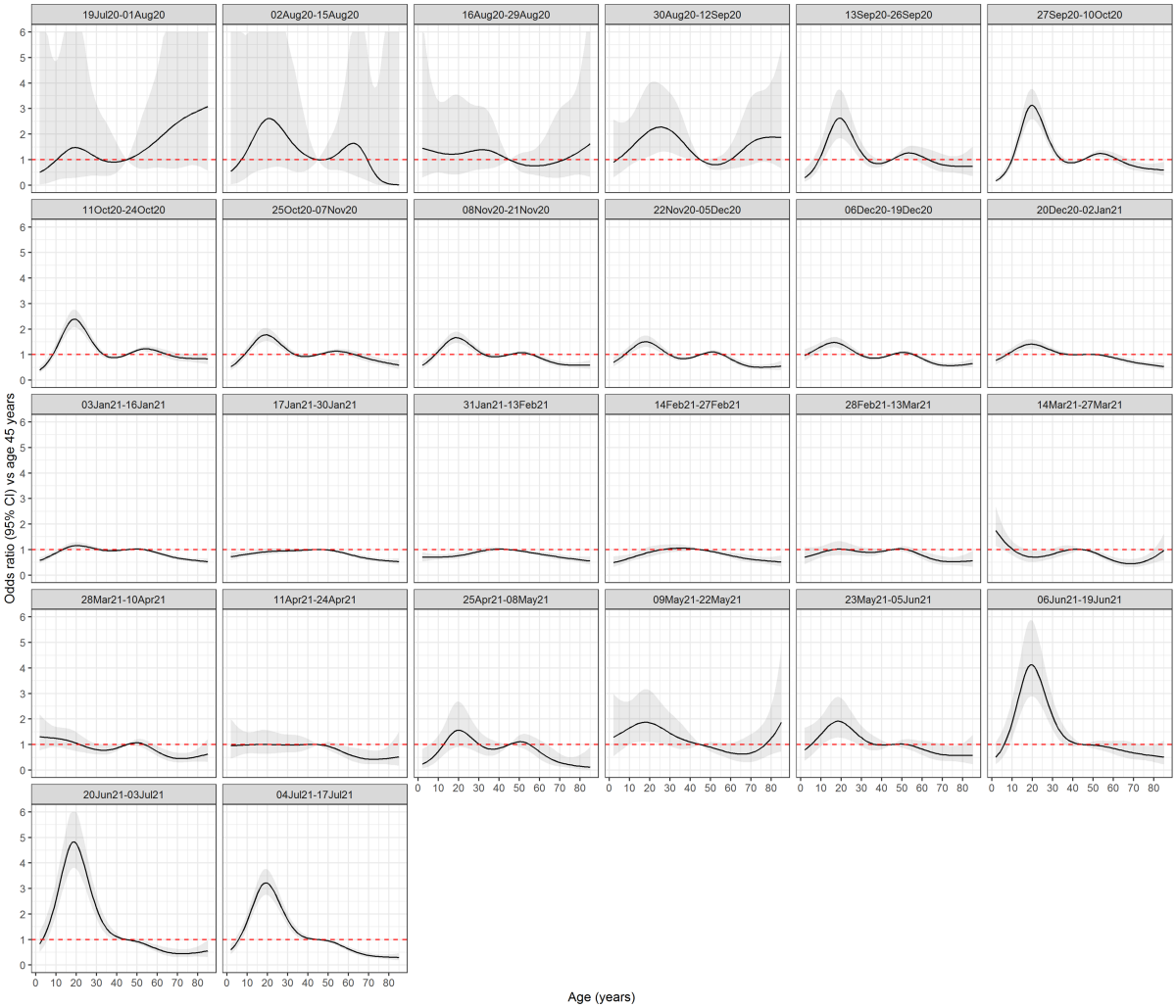


**Figure 2.5: Adjusted effect of age (years) on positivity over the 52-week study period.**




Note: Odds ratios are predicted for each age versus a reference age of 45 years.

**Figure 2.6: Summary of odds ratio and p-values for interactions between all of the core variables using fortnights.**


Note: The size of the circles is proportional to -log$_{10}$ of the global heterogeneity p-value for each interaction in each fortnight. The colour of the circles represents the average size of the interaction terms, converted to the odds ratio scale.

### 2.3.2 Screening

As positivity increased, the screening process identified more variables and at a greater significance than expected by chance (**Figure 2.7**, **Figure 2.8A**). Contact with anyone who had recently had COVID-19, currently self-isolating and thinking one had had COVID-19 recently, strongly and consistently predicted higher positivity. As these characteristics are potential mediators of the effects of other factors, they were not considered further.

*Work and employment characteristics*

Work and employment were significantly associated with positivity throughout the study. Initially from 2nd August-12th September 2020, there was independently higher positivity for those working in care/nursing homes or patient-facing healthcare roles **(Figure 2.8A**). This effect returned from 25th October onwards, along with increased odds in those reporting working in healthcare sectors and specifically in person-facing social-care roles. From 25th October 2020 to 27th March 2021, there was consistently higher positivity in those working outside compared with those from home, with risk increasing as social distancing in the workplace became more difficult. An increased risk was also associated with all modes of travel to work (foot/bike, car/taxi, train/bus), compared with those not travelling to work (**Figure 2.8B**), with the highest odds for car/taxi, then train/bus then foot/bike. Higher positivity was observed in the teaching work sector during October/November 2020, while those working in information technology (IT) had consistently lower odds (**Figure 2.8A**). As the Delta variant became prominent during June/July 2021, lower (rather than higher) positivity was observed in those reporting working outside the home. In contrast, from 20th June-17th July 2021 those in additional paid employment were at a higher risk of testing positive compared with those not undertaking additional work.

*Contact characteristics*

From 16th August to 7th November 2020, positivity was consistently higher in those who had travelled abroad in the last 28 days. This effect returned during 28th March-12th April 2021 and 9th-22nd May 2021. Contact with hospitals and care homes was associated with increased odds of positivity, particularly from 3rd January to 27th February 2021, when positivity rates were very high due to Alpha. From 27th September 2020 to 27th February 2021 (when positivity was consistently >0.3%), participants were more likely to test positive on enrolment visits (**Figure 2.8B**), most likely reflecting the identification of longer-term PCR-positives at these visits.

From 25th October 2020-13th February 2021, the odds of testing positive were higher in those who did not wear a face covering compared with those wearing a face covering in other situations only. There was also a persistent association with higher positivity between 8th November 2020-27th

March in those wearing a face covering in work and other situations, compared with those wearing a face covering in other situations only (**Figure 2.8B**). This effect did not appear again when positivity rose during the Delta period.

*Health-related characteristics*

Health-related variables varied in importance. Notably, there was no evidence of an association between long-term health conditions and positivity. From 13th September 2020-13th March 2021, lower positivity in those who smoked tobacco products was observed, compared with non-smokers. From 20th December 2020, there was a very strong effect of COVID-19 vaccination, with lower positivity in those vaccinated, compared with unvaccinated (**Figure 2.8B**).

The question on regular use of lateral flow tests was only introduced into the questionnaires later in the survey, with enough data for analysis from the fortnight commencing 23rd May 2021 onwards. After this, the odds of testing positive were persistently higher in those who regularly took lateral flow tests, compared with those who did not.

*Deprivation and living environment characteristics*

IMD components (available only for England) had little impact on positivity after adjusting for the overall deprivation index and household size from the core model (**Figure 2.10**). This is likely due to high correlations between individual components with overall deprivation. The indoors deprivation index was the only component which was additionally associated with positivity, from 6th December 2020-16th January 2021. Higher values of the indoor index are associated with improved quality of housing, measured though the proportion of housing without central heating and the proportion of housing failing to meet the Decent Homes standard.[105] As the index increased, the odds of testing positivity decreased. Characteristics detailing household composition (also available for England only), such as persons per household area, room, and bedroom, were rarely associated with positivity, most likely due to these variables being highly correlated with household size.

*Confounding and backwards elimination*

Many of the effects described above were independent of each other on top of the core model, however many of the characteristics collected explained similar associations with positivity, thus backwards elimination was an important step in highlighting variables of most importance. This was most evident for variables relating to work which included a mixture of binary variables describing professions associated with higher contact roles, for example, social care, care home work, manufacturing, and teaching, and also an overarching variable documenting one's ability to social distance in the workplace. After backwards elimination each fortnight, a varying combination of these variables was selected (particularly from 11th October 2020-13th March 2021); however, the

consistent message of increased risk in higher contact and risk professions was evident. The characteristics of COVID-19 vaccination and lateral flow tests were persistently significant after backwards elimination in all models they were included in and hence were included in all models on top of the core model.

*Behavioural characteristics*

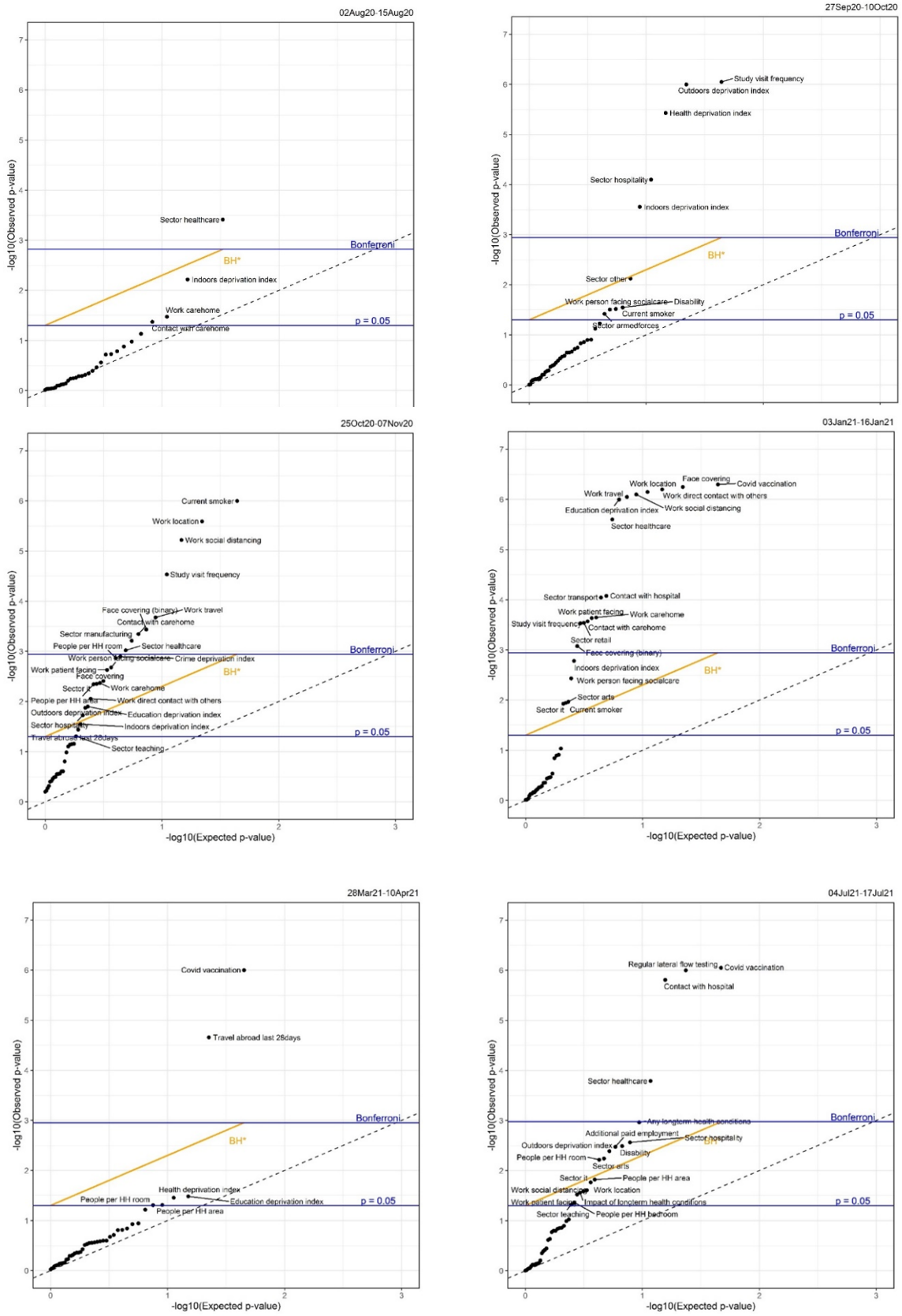
Independently of the core model, higher odds of positivity with higher numbers of social and physical contacts were seen during periods when positivity rates were high (**Figure 2.11**). After also adjusting for variables identified from the main screen and after backwards elimination, there were higher odds of positivity with higher numbers of physical contacts with 18-69-year-olds between 20th December 2020-13th February 2021, and with higher numbers of physical contacts with those <18y between 14th February 2021-27th March 2021 (after schools re-opened). As lockdown restrictions eased and the Delta variant became prominent from 20th June 2021-17th July 2021, odds of positivity were higher in those spending more time socialising outside the home.

The majority of contact variables were included in the screening process both as categorical factor variables and as continuous variables with higher numbers denoting increasing numbers of contacts. While often both were significant, the continuous parametrisations were often more significant, likely due to increased power, and hence selected for the final model.

*Assessing characteristic persistency*

After backwards elimination, of the 71 variables screened (47 in the main screen, 13 in the behavioural screen with 24 parameterisations across the latter), two (3%) effects were persistent, 13 (18%) had effects which came and went, nine (13%) had effects isolated to only two consecutive fortnights, 30 (42%) were associated inconsistently in fortnights, and 17 (24%) were never associated. Covid-19 vaccination and regular lateral flow testing were the only effects which were persistently associated with positivity since their introduction to the study questionnaires. Many of the 13 effects which came and went were associated with positivity during October 2020 through the rise of the Alpha variant to early-March 2021 when positivity decreased. The isolated effects of work location, work sector hospitality, work sector IT, additional paid employment, and time socialising outside home had a period of consecutive association during the last two fortnights in the study period (20th June-17th July 2021), and therefore may have been more persistently associated with positivity going forward. Of the 17 characteristics never associated with positivity, six were behavioural characteristics which were not collected from the beginning of the study period, with the remaining 11 being tested each fortnight for the full year period.

**Figure 2.7: Global heterogeneity p-values per factor from the main screen for a selection of fortnights across the study period.**

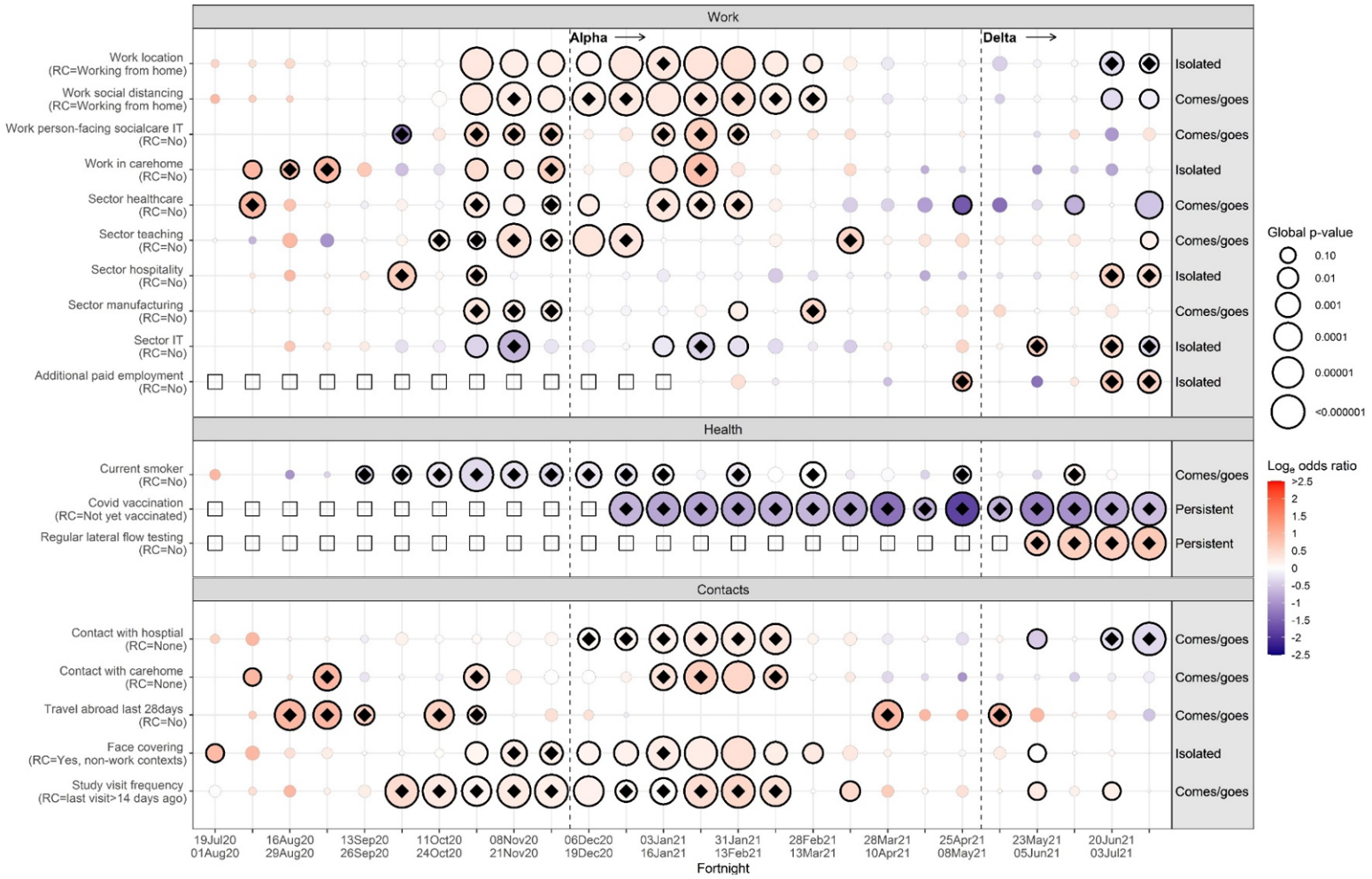



Note: *Benjamini-Hochberg threshold; calculated by ordering p-values from smallest to largest (k = 1,…n), and using the formula: *B-H threshold = k(0·05/N)*, where N is the total number of tests. The black dashed line

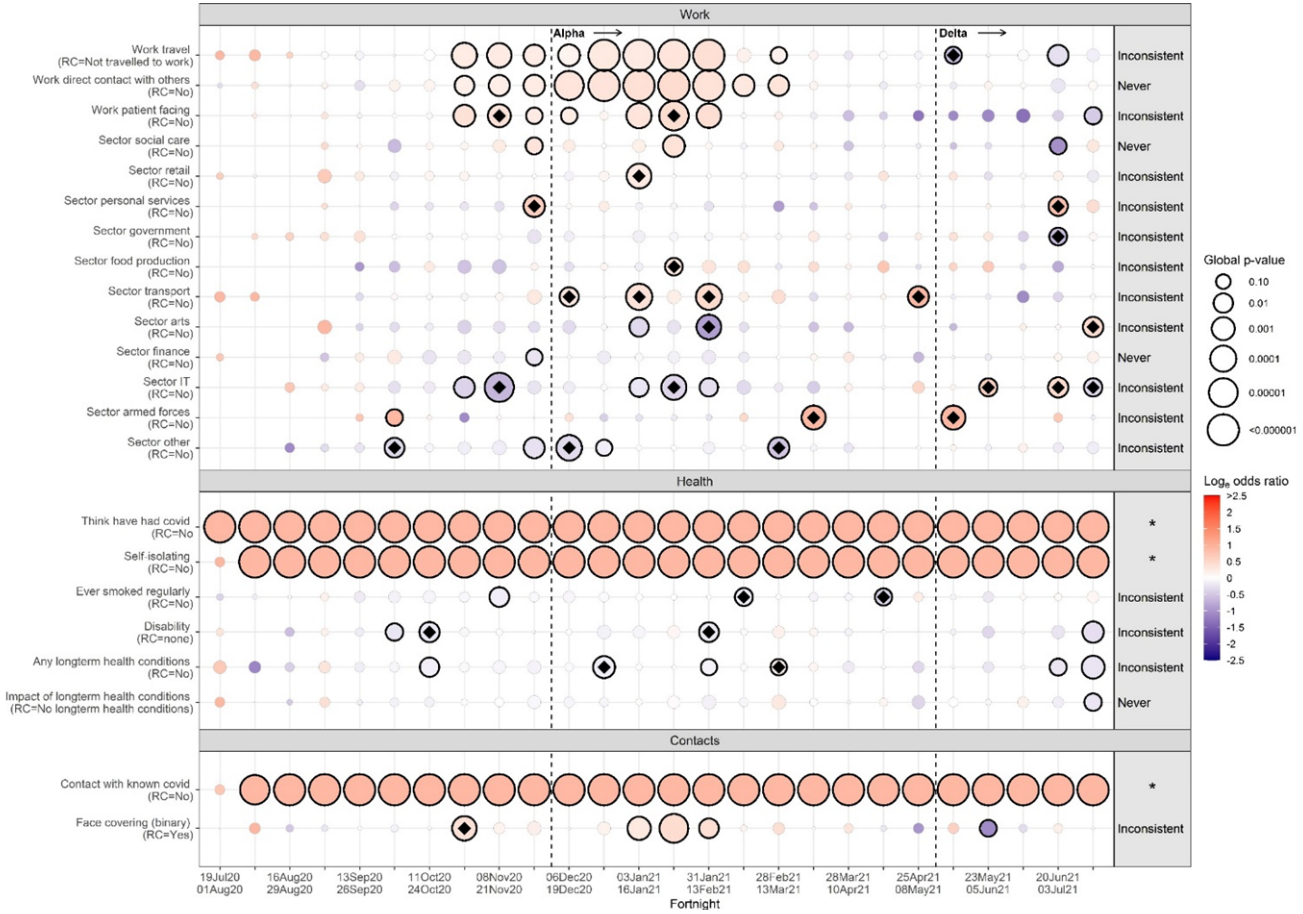
shows y = x, corresponding to the expected distribution of p-values under the global null hypothesis of no effect of any variable.

**Figure 2.8 Overall effects of additional factors from the main screening, adjusted for the core variables, over the 52-week study period.**

(A) *Effects which were persistent, come/go, or were isolated*



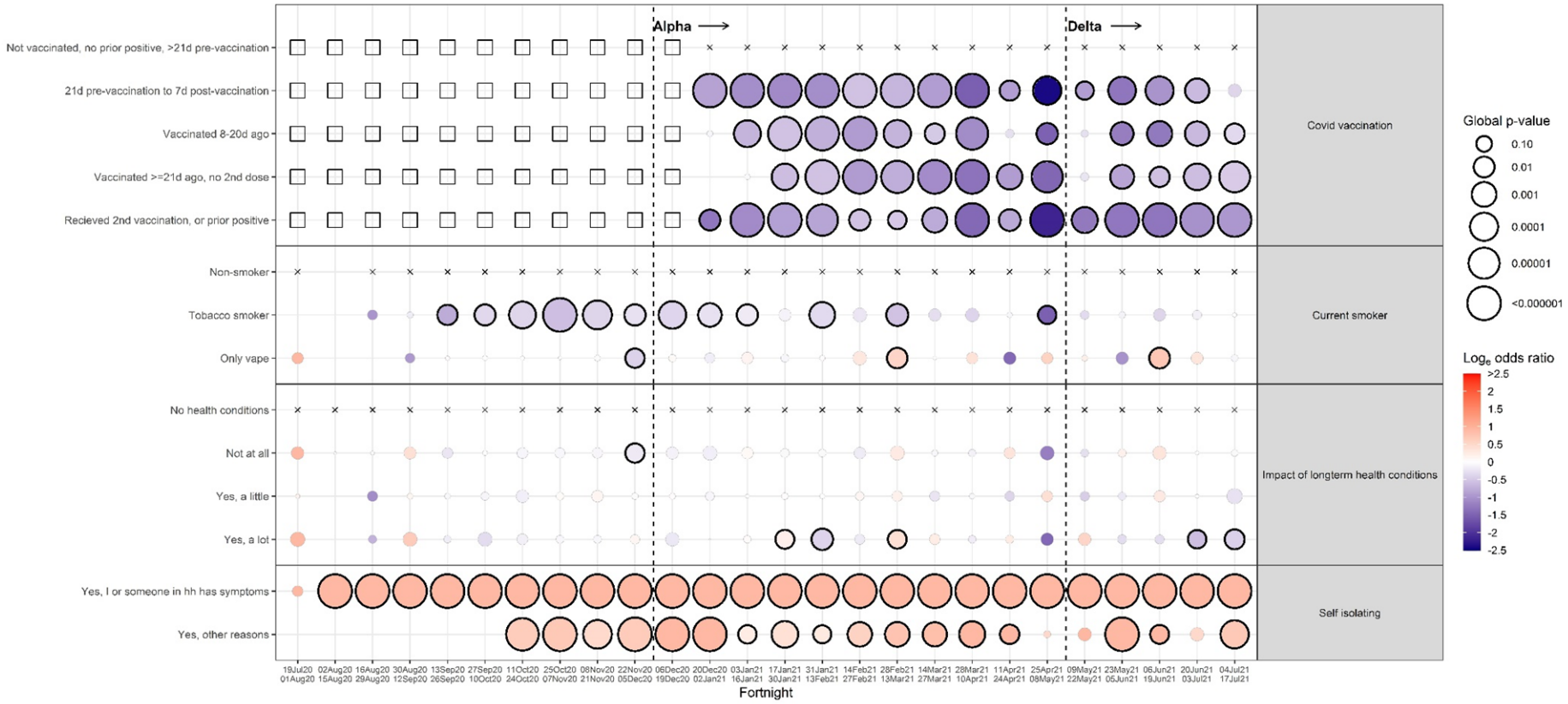
*(B) Effects which were either inconsistent or never associated with positivity*



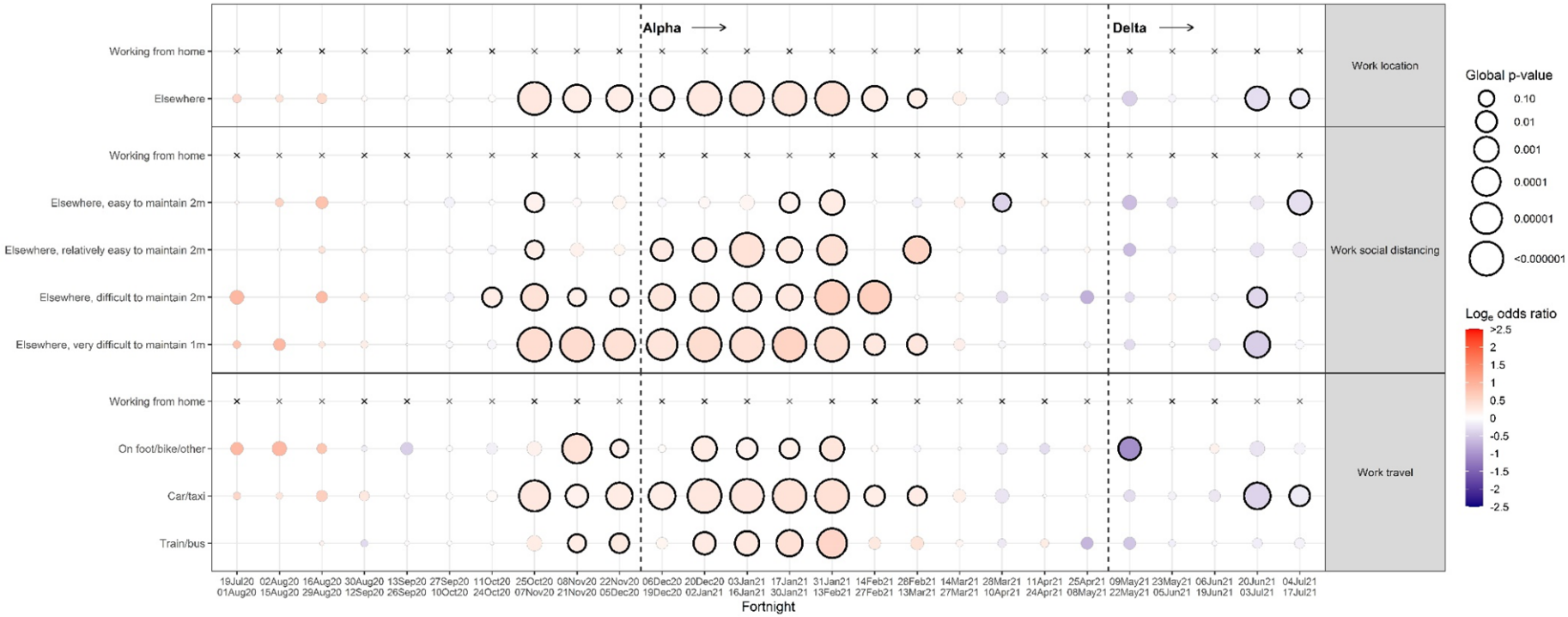

*potential mediators of effects of other factors so not considered in main effects model further. Note: each factor included in addition to the core variables in each fortnight. Black diamonds indicate factors which remain after backwards elimination of all factors with p<0·05 in each fortnight. White squares indicate fortnights where characteristic was not collected by the survey.

**Figure 2.9 Individual effects of categorical variables with >2 categories from the main screening, adjusted for the core variables, over the 52-week study period.**

*Health status*



*Work and employment*



*Contacts*

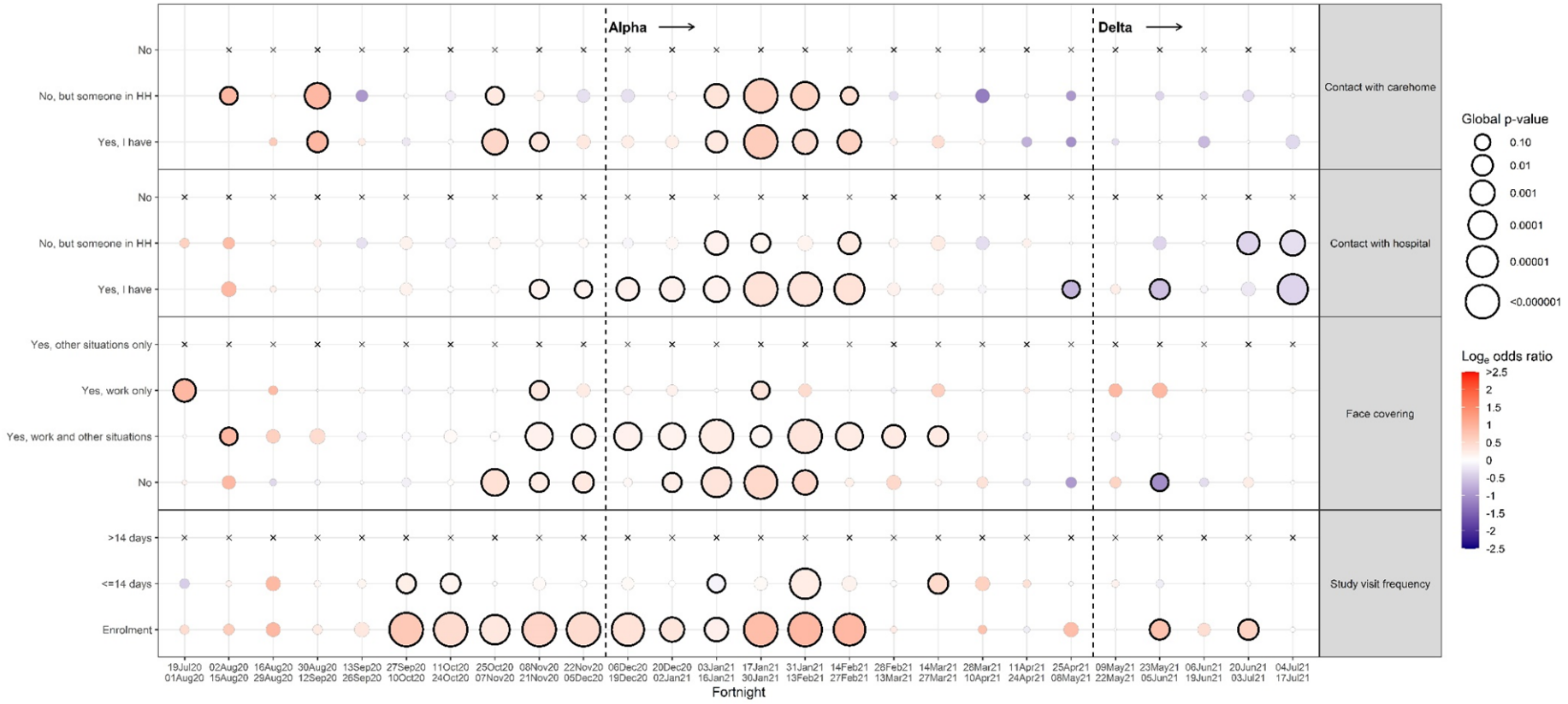
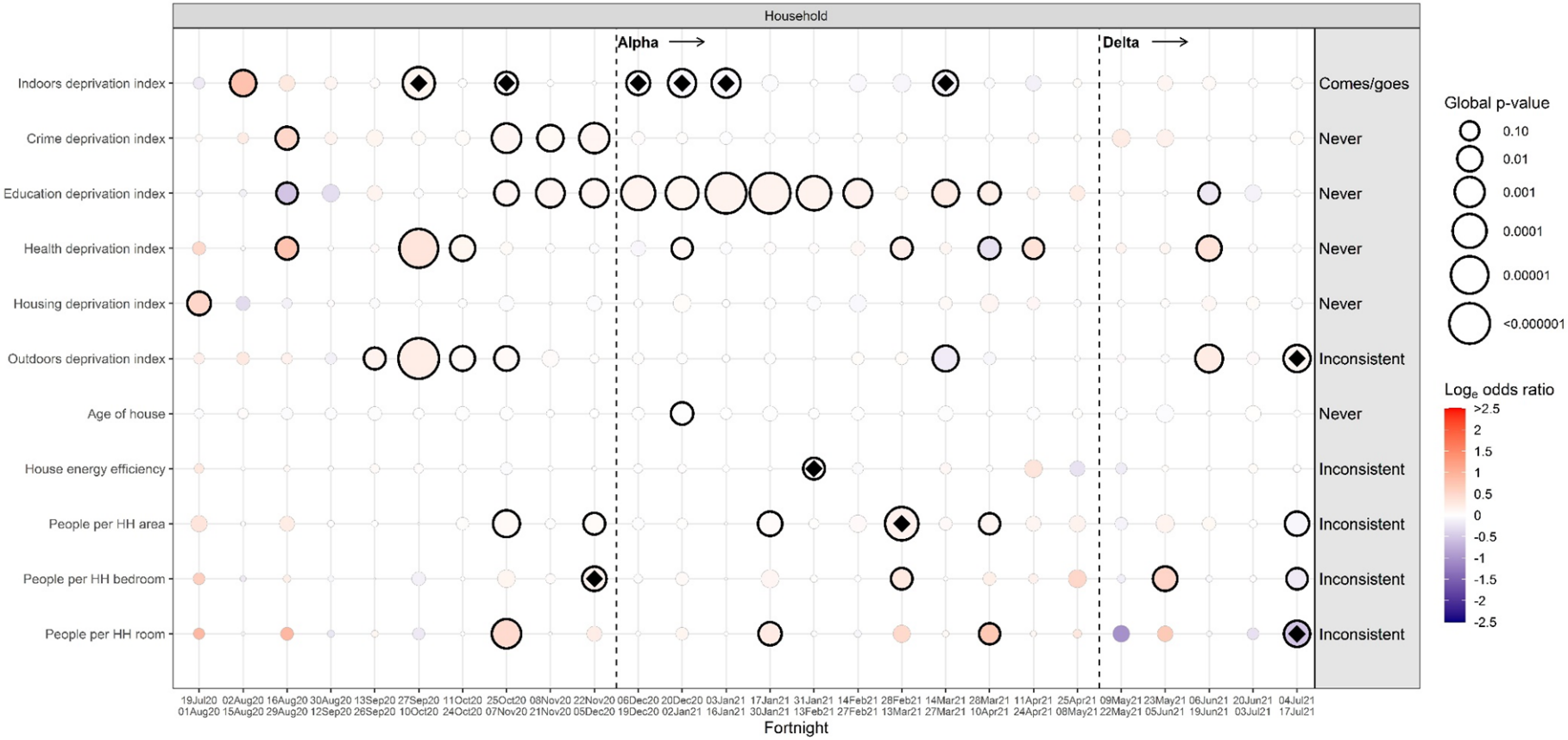

Alpha
Delta
No
No, but someone in HH
Yes, I have
Contact with carehome
No
No, but someone in HH
Yes, I have
Contact with hospital
Yes, other situations only
Yes, work only
Yes, work and other situations
No
Face covering
>14 days
<=14 days
Enrolment
Study visit frequency
19Jul20 01Aug20
02Aug20 15Aug20
16Aug20 29Aug20
30Aug20 12Sep20
13Sep20 26Sep20
27Sep20 10Oct20
11Oct20 24Oct20
25Oct20 07Nov20
08Nov20 21Nov20
22Nov20 05Dec20
06Dec20 19Dec20
20Dec20 02Jan21
03Jan21 16Jan21
17Jan21 30Jan21
31Jan21 13Feb21
14Feb21 27Feb21
28Feb21 13Mar21
14Mar21 27Mar21
28Mar21 10Apr21
11Apr21 24Apr21
25Apr21 08May21
09May21 22May21
23May21 05Jun21
06Jun21 19Jun21
20Jun21 03Jul21
04Jul21 17Jul21
Fortnight
Global p-value
0.10
0.01
0.001
0.0001
0.00001
<0.000001
Log$_e$ odds ratio
>2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5

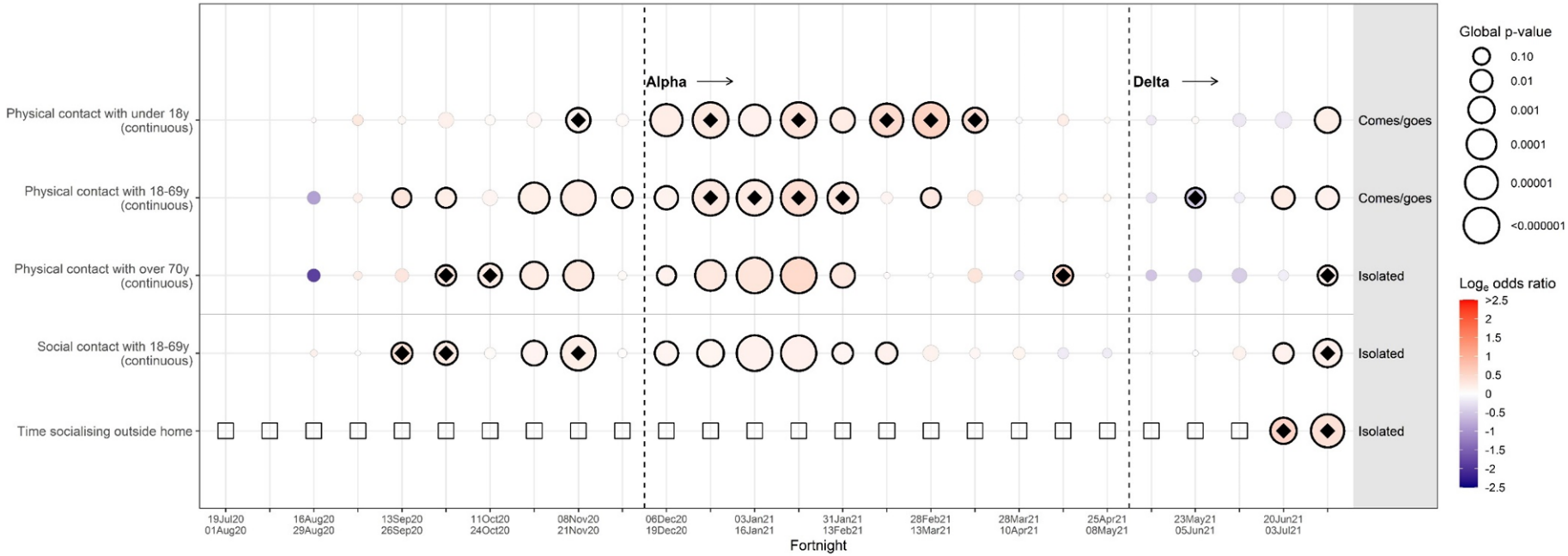
**Figure 2.10: Global heterogeneity p-values per factor from the screening for household and living environment characteristics.**



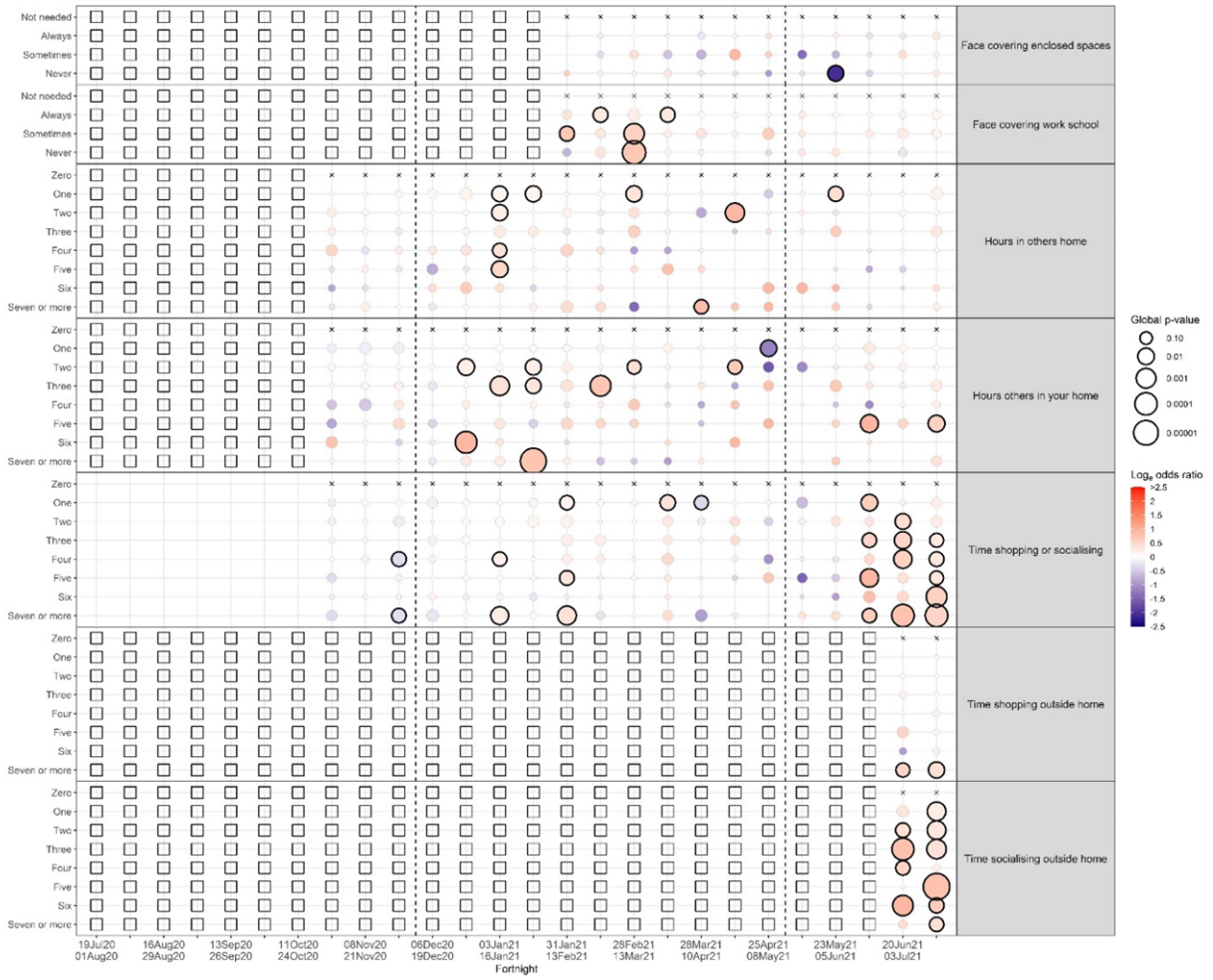

Note: Each factor is included in addition to the core variables in each fortnight. Black diamonds indicate factors which remain after backwards elimination of all factors with p<0.05 in each fortnight.

**Figure 2.11: Adjusted effects of behavioural variables from the main screen.**

(A) *Effects which were persistent, come/go, or were isolated*



*(B) Effects which were either inconsistent or never associated with positivity*



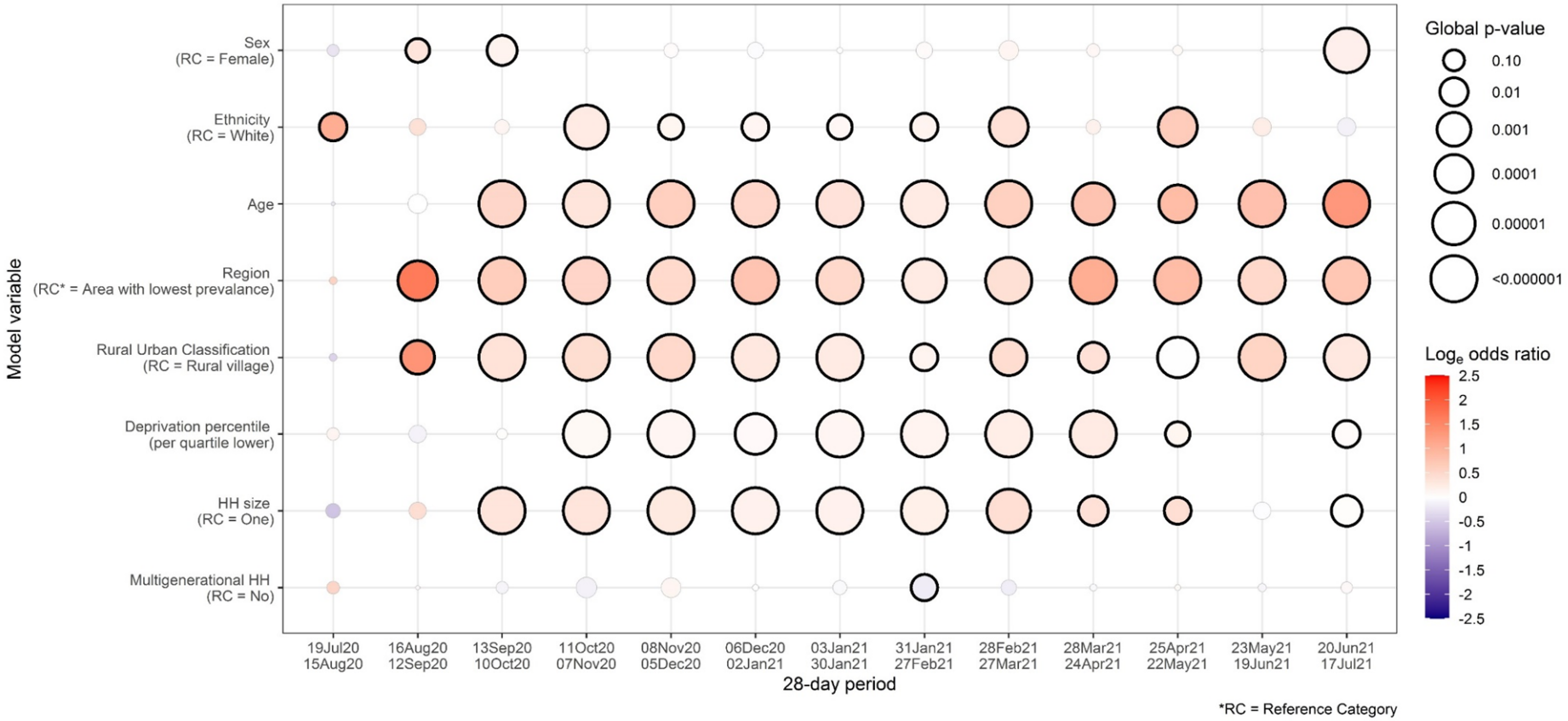

Note: Each factor is included in addition to the core variables in each fortnight. Black diamonds indicate factors which remain after adjustment for all variables identified in the main screen and backwards elimination of all factors with p<0·05 in each fortnight. White squares indicate fortnights where characteristic was not collected.

### 2.3.3 Sensitivity analysis

*28-day periods versus fortnights*

Similar key predictors of positivity were obtained using 28-day periods in the core model (**Figure 2.12**; **Figure 2.13**; **Figure 2.14**). Notably, there was a more consistent signal of higher positivity in non-white ethnicities from 11th October 2020-27th March 2021, while this signal was more intermittent using fortnights (**Figure 2.4A**). Again, no significant interactions persisted across any two consecutive 28-day periods. Similar key associations were also identified from the main screen. Of the 45 consecutive occurrences of effects with p<0·05 in fortnights, 25 (56%) would have been detected later in 28-day periods, 14 (31%) at the same time, five (11%) earlier, and one (2%) never detected. For example, the variable describing current smoking status would have been identified on 26th September 2020 when using fortnights, but 10th October 2020 if using 28-day periods (**Figure 2.15A**). The effect of work travel would have first been identified on the same day (7th November 2020; **Figure 2.15B**). In contrast, the association of visit frequency on positivity would have been identified later in fortnights, with it being detected from the beginning of the study period in 28-day periods (1st August 2020), but not until 10th October when using fortnights (**Figure 2.15C**).

**Figure 2.12: Summary of odds ratios and p-values for the 8 core variables over 28-day periods.**




Note: RC=reference category. HH=household size. The size of the circles is proportional to -$\log_{10}$ of the global heterogeneity p-value for each variable in each 28-day period. Circles with black outlines represent p<0·05. The colour of the circles represents the size of the odds ratio (vs the reference category shown). For categorical variables with >2 levels (region, rural/urban classification, and household size), the reference category was set as the level with the lowest prevalence in each fortnight, and the overall "odds ratio" was calculated as $\exp\left(\frac{\sum \frac{1}{se(\beta_i)}\beta_i}{\sum \frac{1}{se(\beta_i)}}\right)$. As age was included in the model as a restricted natural cubic spline, odds ratios were predicted at ages 10, 25, 40, and 55 vs 70 (reference) years and then combined in the same way.

**Figure 2.13: Summary of odds ratio and p-values for interactions between all of the core variables for 28-day periods.**

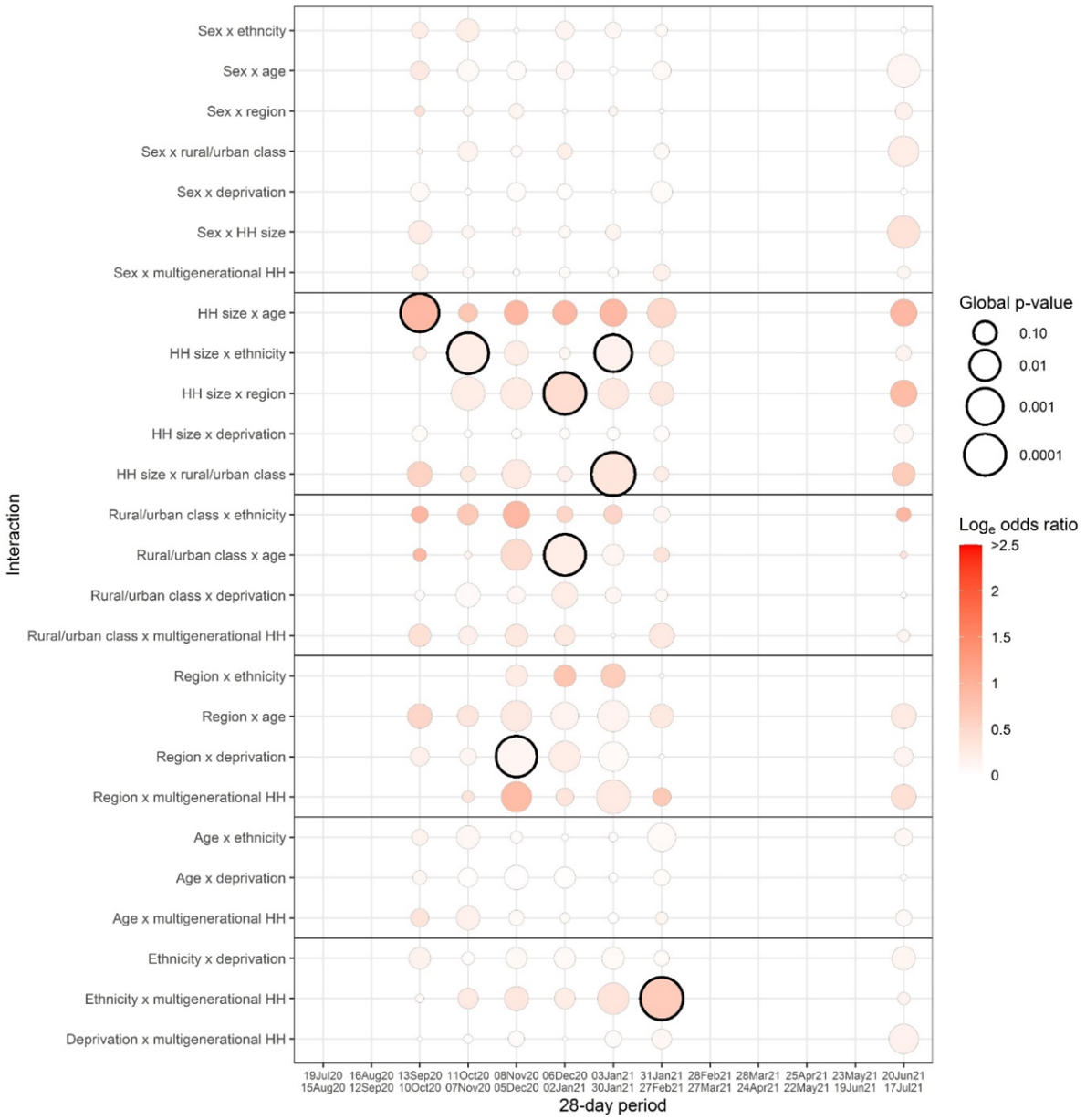


**Figure 2.14: Global heterogeneity p-values per factor from the main screen for 28-day periods for characterises based on work, health status and contacts.**

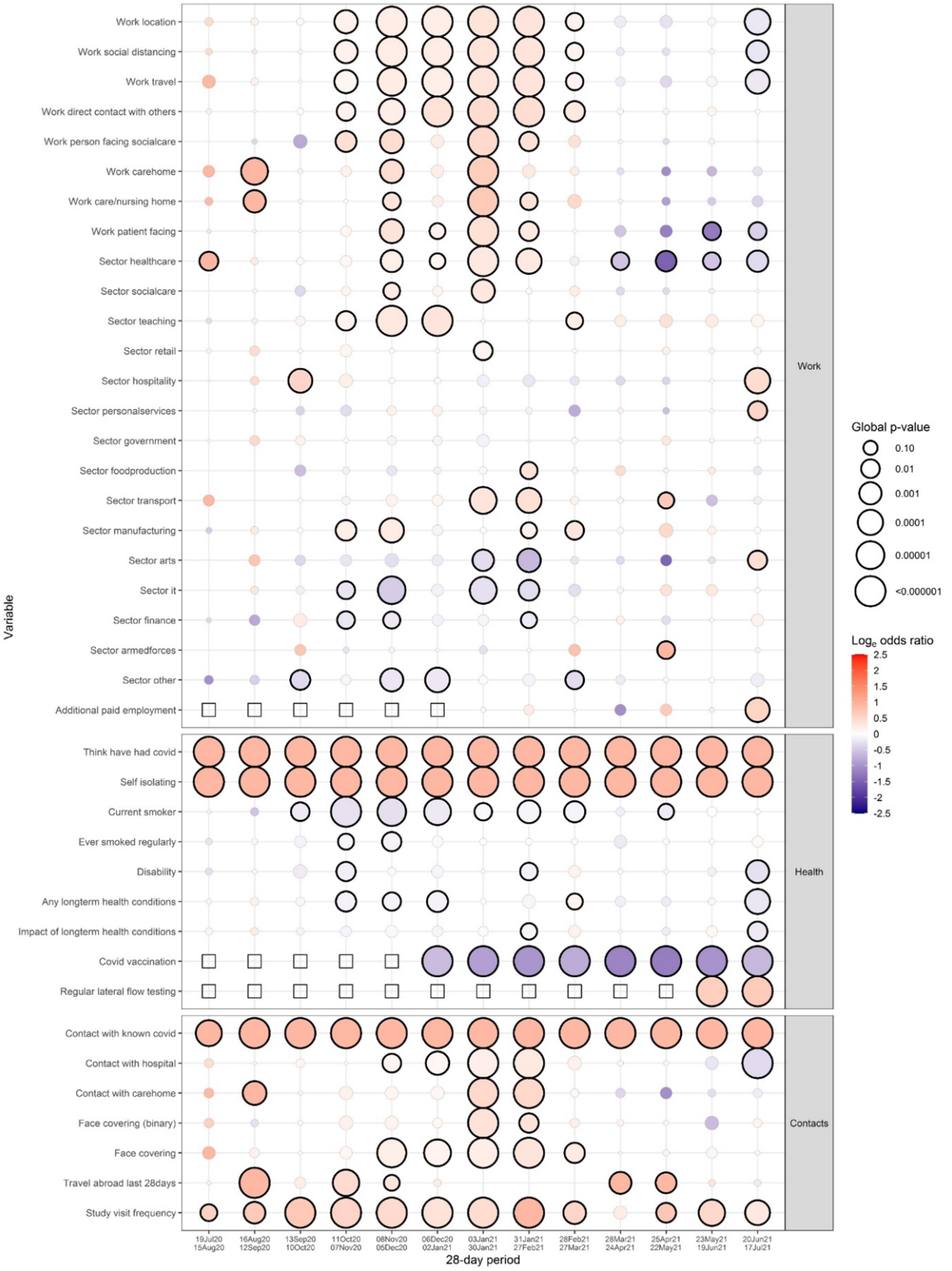
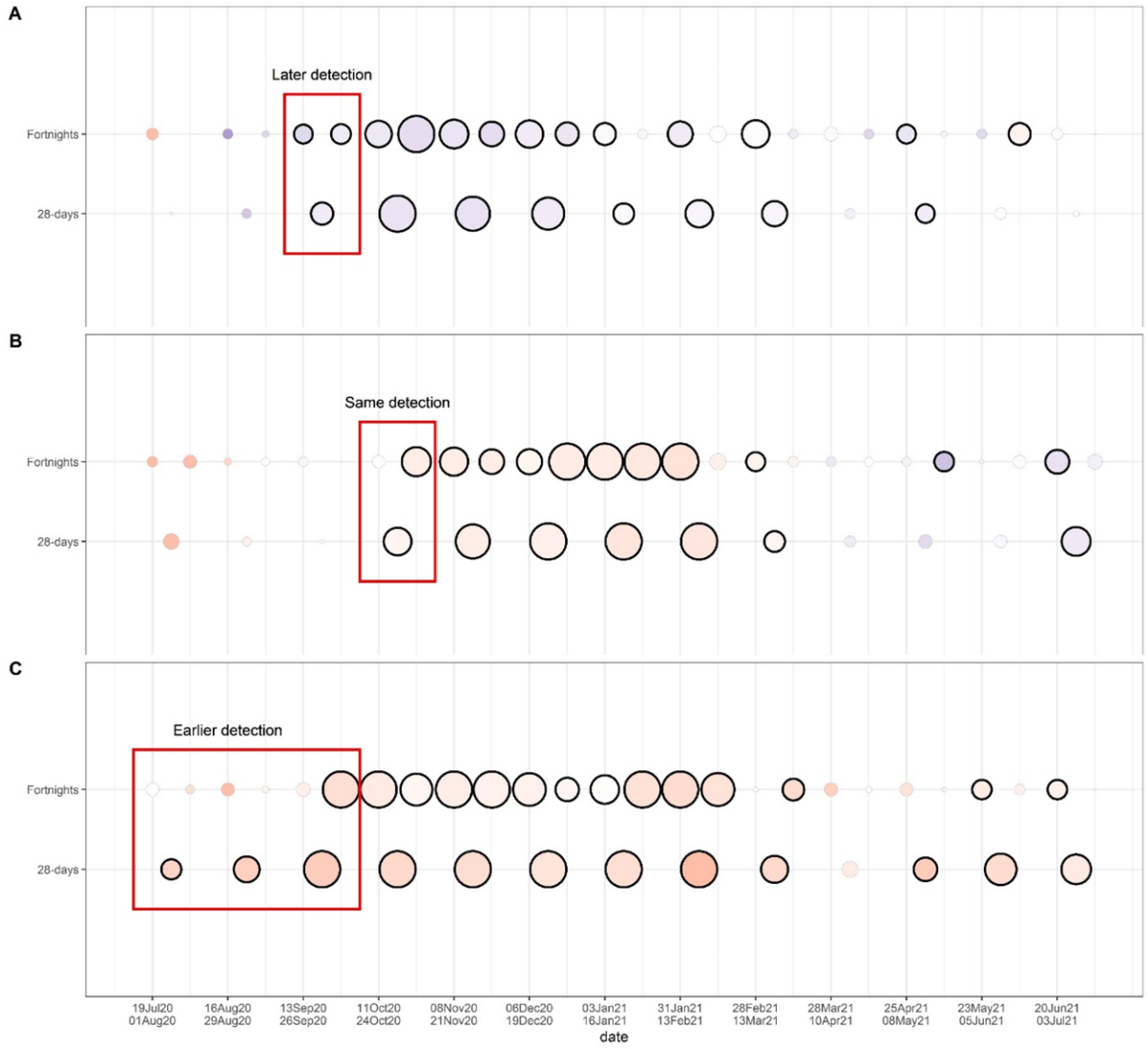


**Figure 2.15: Examples of later, same, and earlier detection of effects for current smoking status (A) work travel (B) and study visit frequency (C).**

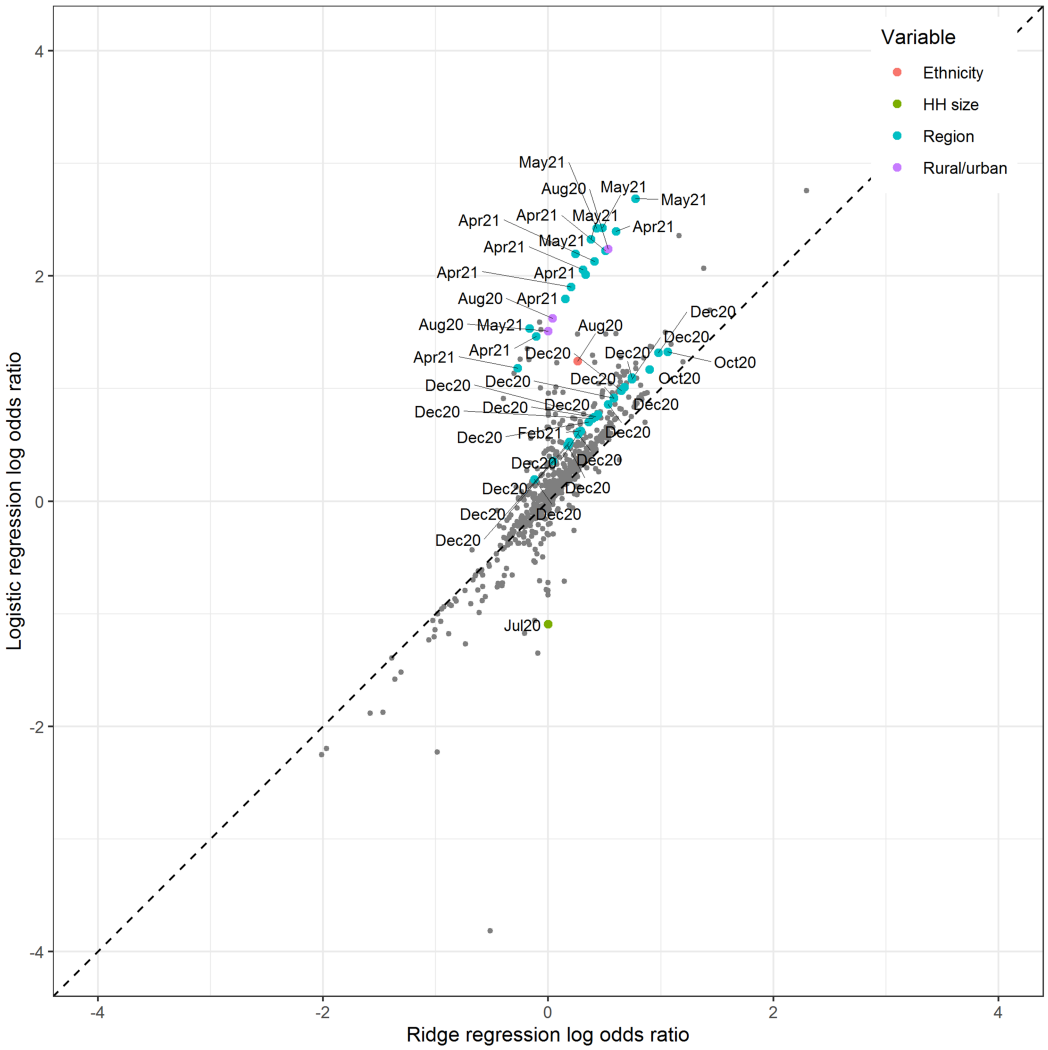


*2.3.4 Ridge regression*

Of the 692 coefficients from the core models over the 52-week study period, 43 (6%) coefficients produced from ridge regression did not fall within the 95% confidence intervals from the original logistic regression models presented above (**Figure 2.16**). Of these, the majority (38 coefficients; 88%) were effects of geographical region. These were mostly in the first fortnights of the study period when event rates and sample size were smallest, and during December 2020, when there were strong regional effects due to the rise of the Alpha variant in Southern regions of England. Many of the inconsistencies within geographical regions occurred within the same fortnight i.e. either none or all of the effect estimates for regions were within the 95% confidence intervals.

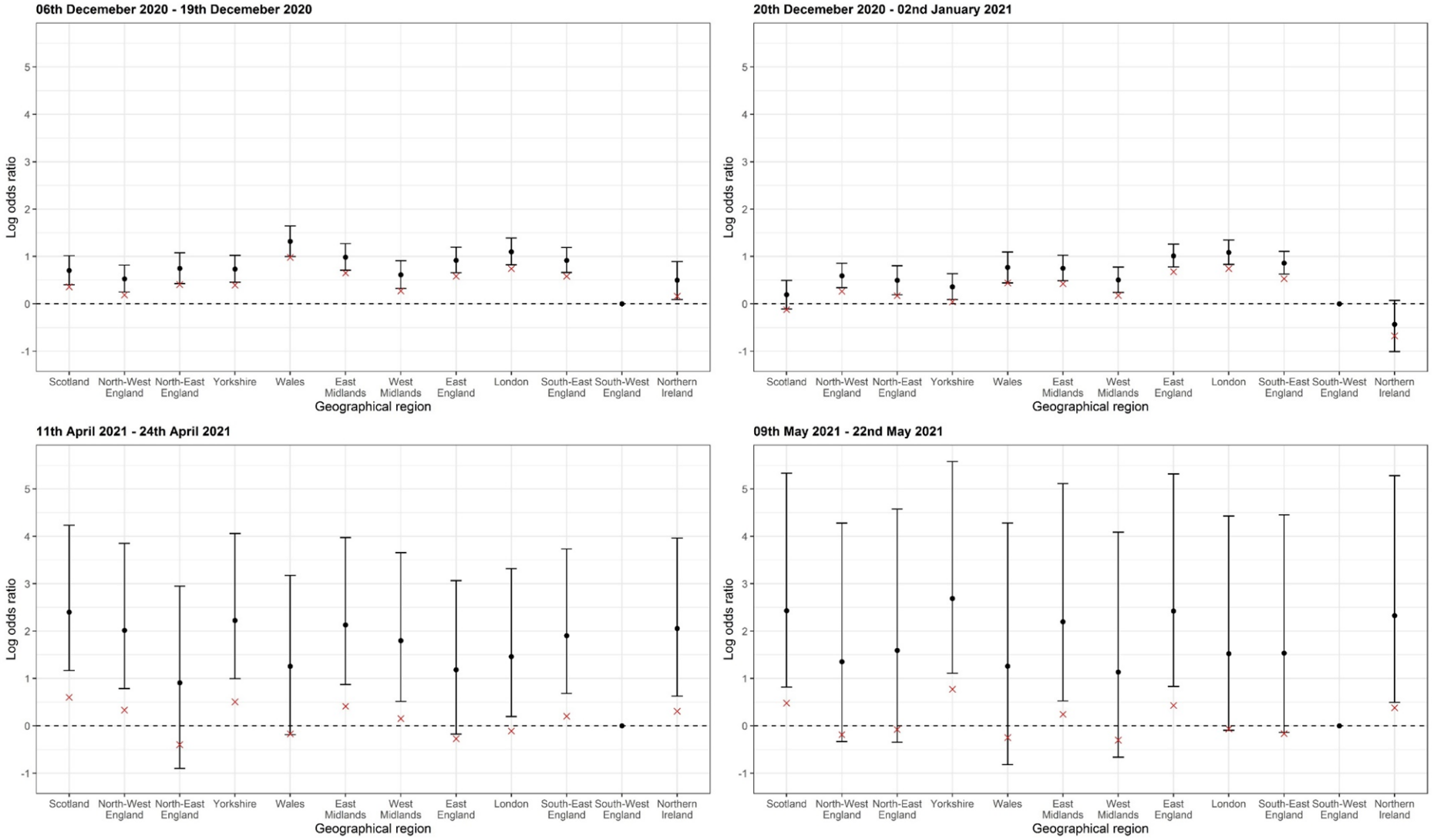
The differences observed between coefficients in December 2020 while the Alpha variant was rising suggest that the ridge regression penalised early signal for the regional effect, while logistic regression models identified this. The odds ratios for geographical regions from ridge regression tended towards zero during the four fortnights where the estimates were outside of the 95% logistic regression confidence intervals (**Figure 2.17**). While often challenging to distinguish between signal and noise, through triangulation with other data sources, the regional effects observed in the logistic regression model were accurate and representative of rises in the Alpha variant in London and the South East, while ridge regression missed this effect, hence justifying my choice of method.

**Figure 2.16: Results from ridge regression and logistic regression.**




| | Ridge coefficients outside of logistic regression 95% confidence interval, n (%) |
|---|---|
| **Total** | 43 (6% of all 692 coefficients) |
| **By Variable** | |
| Region | 38 (88) |
| Rural/Urban Classification | 3 (7) |
| Household size | 1 (2) |
| Ethnicity | 1 (2) |
| **By fortnight** | |
| 19Jul20-01Aug20 | 1 (2) |
| 16Aug20-29Aug20 | 1 (2) |
| 30Aug20-12Sep20 | 3 (7) |
| 11Oct20-24Oct20 | 2 (5) |
| 06Dec20-19Dec20 | 10 (23) |
| 20Dec20-02Jan21 | 10 (23) |
| 14Feb21-27Feb21 | 1 (2) |
| 11Apr21-24Apr21 | 9 (21) |
| 09May21-22May21 | 6 (14) |

**Figure 2.17: Odds ratios from logistic regression with 95% confidence intervals (black) and ridge regression (red crosses) for geographical region for four fortnights.**



## 2.4 Discussion

Over one year from 19th July 2020-17th July 2021, I estimated and summarised the key predictors of SARS-CoV-2 positivity in the UK, using a method designed to be run weekly in real-time to provide up-to-date information on changes in populations at increased risk. In the first fortnight from 19th July-1st August 2020, there was no evidence that any characteristic impacted positivity. As positivity rose through September-November 2020, rates were independently higher in those of younger ages, living in Northern areas of England, in major urban conurbations, in more deprived areas, and in larger households. Additionally, rates were higher in those who had recently travelled abroad, worked in healthcare roles, or worked outside of the home. As positivity peaked in December 2020-January 2021, while strong effects of living in urban areas and large households remained, there was a major shift in high positivity to more southern geographical regions (reflecting the emergence of the Alpha variant), with risk no longer concentrated in younger ages. Those working outside of home and in healthcare roles still had a higher risk. As the national vaccine programme rolled out in December 2020, there were large reductions in positivity in vaccinated individuals. From February-May 2021 as rates decreased, the impact of work on positivity decreased, while the effect of vaccination remained. As the Delta variant became prominent and positivity rates rose from mid-May through July 2021, there were higher odds of positivity in younger ages, in men, and in those not yet vaccinated.

Whilst my observed associations were consistent with other community infection surveys in the UK, particularly the English REACT study,[93] no other studies have assessed as many characteristics in a community population over time as this study was able to, with many focusing on outcomes of mortality and hospital admissions. Variation in positivity by region was well documented[106,107] as was increased positivity in non-white ethnic groups during September-November 2020,[106] and those working in hospitals and care homes during November 2020-January 2021,[108] but not as the Delta variant rose[109]. As well as demonstrating the increased risk of infection in those not vaccinated as Delta came to dominate,[110,111] the screening process I developed facilitated continuous monitoring of waning vaccine-associated protection going forward. However, I was also able to monitor characteristics including behaviours and work, many of which affected positivity inconsistently over time. For example, between October 2020-March 2021, working from home was associated with lower positivity, whereas during June/July 2021 working from home was associated with higher positivity. As working from home was recommended during the former period, working from home likely preceded infection. In contrast, as returning to the workplace became encouraged from May 2021, working from home may have sometimes been a consequence of exposure and hence self-

isolation, leading to a degree of reverse causality, and a higher risk of positivity in those working from home. Interpreting associations contextually with current restrictions is therefore critical.

Work sectors with evidence of higher positivity over time generally involved roles with more contact with others, namely teaching, hospitality, and manufacturing. This is consistent with other studies showing increases in SARS-CoV-2 cases with increasing numbers of close contacts.[112] Physical and social contact was also associated with a higher positivity. Specifically, there was higher positivity in those with more physical contact with those under 18s, particularly when schools re-opened in March 2021,[113] more physical contact with those aged 18-69y and over 70s, more social contact with those aged 18-69y, and those spending more time socialising outside their home. Increased positivity associated with those reporting additional paid employment may also reflect an increased propensity for such work to involve close contact. Increased positivity in those having had recent contact with hospitals and care-homes during December 2020-February 2021 is likely due to the high number of SARS-CoV-2 cases in these environments as the Alpha variant emerged and came to dominate.[114] The persistent increased risk of positivity in those self-reporting taking regular lateral flow tests is likely due to their use, as recommended, for those whose activities reflect increased risk such as working outside the home or in teaching, healthcare or social care.

The impact of close contact may also explain the increased risk observed in those not self-reporting wearing face-coverings, with evidence suggesting face-coverings are effective at reducing transmission.[115] The higher risk observed in those wearing face-coverings in work and other situations, compared with those only wearing face-coverings in non-work situations may reflect the increased risk of working away from home which was observed in other work-related characteristics. Higher positivity in those who had recently travelled abroad in August-November 2020 may also be due to the increased number of close contacts involved with travel (e.g. being at an airport), or higher risks of infection in the destination country at the time.

There was a lower risk of positivity in those who smoked tobacco products consistently from September 2020 to January 2021 and intermittently from then onwards. There was no protective effect in those only reporting vaping. While some have outlined biologically plausible mechanisms which may explain this reduced risk, this also could be attributed to residual confounding based on the demographic of people who smoke offering a protective effect, not otherwise adjusted for.[116] Vaccination has been shown to reduce community infections elsewhere;[117] prior infection is well-recognised to give at least as good protection, and the fact that these known associations were easily identified within my model provides confidence that important confounders are adjusted for in estimating other associations.

### 2.4.1 Limitations of the screening process

The screening process demonstrated here has several limitations. First, low event numbers and smaller sample sizes reduce statistical power, reducing the chance of detecting true associations (false-negatives) and increasing the likelihood that the magnitude of "true" effects are inflated (false-positives).[118] Increased statistical power using 28-day periods rather than fortnights more consistently detected associations with ethnicity in the core model and found more evidence of interactions. The screening process, however, detected the same characteristics using both time-periods, with earlier detection in most cases using fortnights. As there were no major differences and I aimed to identify associations most relevant to current positivity, the benefit of more regular estimates may outweigh the power gained from evaluating longer time-frames, although this will depend on event numbers. When event numbers are low, logistic regression can be biased and/or imprecise.[119,120] Sensitivity analyses using penalised regression techniques showed most coefficients were within the logistic regression 95% confidence intervals, suggesting that, while there was some attenuation of estimates, for example for geographical regions in a few fortnights, the logistic regression models were not substantially overfitting. Further, differences observed between coefficients in December 2020 while the Alpha variant was rising suggest that ridge regression penalised early signal for the regional effect, while logistic regression models identified this, hence justifying my method choice.

Multiple testing is an unavoidable limitation of my screening methodology. Doing many multiple independent tests increases the risk of false positives;[121] however, a priori the questionnaire was based on potential risk factors so the "correct" degree of adjustment is unclear. I therefore used Q-Q plots with Bonferroni and Benjamini-Hochberg adjustments to monitor the potential for false-positives, rather than as strict thresholds.[122,123] Even using stricter Bonferroni criteria, many screening variables were associated with positivity. Considering sex as a "negative control" (no effect expected), an association with positivity was only observed in one of 24 fortnights before 20th June 2021. The association between sex and positivity from 20th June-17th July 2021 coincided with the European Football Championship, thus plausibly reflecting changes in social behaviour by sex, as observed elsewhere.[124] My results suggest more emphasis should be placed on effects that appear at least twice, interpreting effects that are inconsistent or appear sporadically with caution.

To show whether the screening process is viable and assess the impact of the above limitations, one could conduct a simulation study. Specifically, it would be of interest to assess the impact of multiple testing and evaluate the rate of false-positive results, while also estimating the rate of missed true associations with an outcome. Designing such a simulation study which would be both effective and useful for evaluating the methods would have its challenges. Extensive work would need to go into

the data-generating mechanism to produce data on which to assess methods which is as complex as that used in the screening methods, particularly with high levels of confounding between explanatory variables. To both adequately and usefully assess the process in a simulation study, it would be recommended to follow the robust guidelines of the "ADEMP" (Aims, Data-generating mechanisms, Methods, Estimands, Performance measure) procedure as outlined in the tutorial by Morris et al.[125]

### 2.4.2 Strengths and Limitations of the COVID-19 Infection Survey

The underpinning design, namely a large community-based survey including randomly selected private households, is a major study strength. Participants being regularly asked about behaviours, work, and health status provided a rich opportunity to identify associations between positivity and many important demographic and behavioural characteristics. As participants were tested regardless of symptoms, characteristics could be assessed in an unbiased population, thus avoiding selection bias through only observing those choosing to take a COVID-19 test, for example, in the English national testing programme[126] or through presenting to hospital with severe disease.

The study design also had limitations, particularly with individuals tested initially at weekly and then monthly visits. As fragments of the virus can be detectable in the respiratory tract long after the onset of infection, positives included in my outcome include both new infections and lingering PCR-positivity. Associations from the screening process may therefore not necessarily be related to new infections. Whilst I could have grouped positive tests into "episodes", for example, considering only the first positive in 90-day periods,[127] I chose to mirror other point-prevalence studies, such as REACT,[93] also expecting that many characteristics would be reasonably stable over time and therefore even associations with ongoing PCR-positivity could still be relevant to the original infection. This may however dilute effects if participants with long carriage have different characteristics to those testing positive with new infections. Ongoing PCR-positivity may also reduce sensitivity to detect specific "at-risk" populations as new variants emerge. My analysis was also conducted over periods when both Alpha and Delta were prominent in the UK. Whilst theoretically this might have allowed me to estimate changes associated with each variant, the differing control policies over these periods, coupled with a large proportion of the population being vaccinated once Delta became prominent, make it challenging to disentangle whether different associations were due to control strategy versus variant.

As the CIS took swabs from participants regardless of symptoms, I included both symptomatic and asymptomatic infection as my outcome. While analyses could have considered these groups separately, many participants on monthly visits risk either catching infections late and thus missing

symptoms at the beginning of infection or conversely finding infection positivity early before the onset of symptoms. For future work, it could be appropriate to use multinomial logistic regression models to assess differences in explanatory variables between those testing positive with reported symptoms, testing positive without reported symptoms, and those testing negative. It would be important to consider the impact of these additional tests on multiple testing, and also the reduced power as the positive outcome is split. There is published evidence of differences in associations between age, sex, and ethnicity with SARS-CoV-2 positivity with/without symptoms.[102]

### 2.4.3 Conclusion

In conclusion, the screening process presented could be a valuable tool in understanding the characteristics driving current SARS-CoV-2 positivity, allowing enhanced up-to-date understanding of the pandemic across the UK. Looking forward, this could be used to target public health messages to detected groups to increase uptake of symptomatic and asymptomatic testing. The process went on to be used repeatedly by the Office for National Statistics throughout the pandemic, particularly estimating the impact of waning protection from vaccination, with regular Statistical Bulletins entitled "Characteristics of people testing positive for COVID-19" (full list on https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/results/results-longer-articles-and-adhoc-publications-from-ons).

# Chapter 3 Detecting changes in population trends in infection surveillance using community SARS-CoV-2 prevalence as an exemplar

The work presented in the Chapter was published in the *American Journal of Epidemiology* in 2024.[128] I authored the text and figures below (which are reproduced here largely unaltered from the published work) with input from supervisors commensurate with the amount of input/advice that would be considered appropriate for a DPhil thesis. All authors have agreed to the inclusion of this published work in my thesis.

## 3.1 Introduction

Infectious disease surveillance has two broad goals; identifying outbreaks which lead to sudden changes in incidence/prevalence, and detecting the emergence of more virulent/resistant strains. Reasons for changes in infectious disease trends vary, for example: changing population susceptibility causing increasing group A streptococcal infections;[129] emerging antimicrobial resistant strains e.g. ribotype-027 *Clostridium difficile* or Gram-negative pathogens carrying extended-spectrum beta-lactamases;[130] and mutations affecting transmissibility in COVID-19.[131] While laboratory sequencing methods can accurately identify variants,[132] limited sampling and resources mean retrospective statistical models are often more practical for monitoring infectious diseases, particularly with increasing availability of linked electronic health records.

While change-point detection methods are numerous, many are sub-optimal for use on infectious disease time-series in near real-time. Statistical methods involve locating points in a time-series where some property of the data (e.g. distribution, scale) changes.[133] Given the epidemiological drivers above, identifying points where the rate of change in the trend is increasing/decreasing are most useful in infectious disease surveillance. Many statistical methods however identify step changes, i.e. changes in mean levels in a time-series, rather than the more gradual trend changes[134] characteristic of changing infectious disease epidemiology. Other methods require pre-specifying the number of change-points, and many are computationally expensive and therefore not practical for near real-time use. During the COVID-19 pandemic, while studies used change-point analysis to retrospectively assess the impact of interventions e.g. lockdowns and gatherings,[135-137] change-point detection methods for near real-time use have been less commonly assessed.[138] Two methods considering more gradual changes and finding change-points in trends are iterative sequential regression[139,140] (ISR) and second derivatives of generalised additive models (GAMs).[141] While both have been evaluated separately,[139,142] to my knowledge, they have never been directly compared. ISR provides a clear statistical assessment of when rates change and estimates constant growth rates between change-points, potentially maximizing power when this is close to true underlying trends.

However, it considers data sequentially, fixing change-points as it iterates, thus not necessarily optimising overall fit. Second derivatives of GAMs have been used to identify periods of change,[142,143] and quantify change-points.[141] Their flexibility allows estimates to closely reflect reality, but to what extent smoothing through penalized splines reduces the ability to detect change-points in near real-time is unclear.

I aimed to compare the performance of GAMs and ISR for change-point detection for infectious disease surveillance, both retrospectively and in near real-time, using COVID-19 as an exemplar for surveillance more generally, e.g. using linked electronic health records. Rapid changes in SARS-CoV-2 prevalence coupled with multiple emerging variants over the COVID-19 pandemic provides an ideal opportunity to test these methods in real-world data which is more complex than simulations. I compared the consistency and timeliness of detection between ISR and second derivatives of GAMs for identifying changes in growth rates of SARS-CoV-2 positivity over time using the UK's Office for National Statistics (ONS) COVID-19 Infection Survey. I assessed whether earlier detection was possible considering positivity separately by age group or by available proxies for viral variants.

## 3.2 Methods

### 3.2.1 Study design

Similarly to Chapter 2, data from the ONS COVID-19 Infection Survey (CIS) was used in this study. To briefly recap the structure of the CIS, the ONS COVID-19 Infection Survey was a large household survey with longitudinal follow-up (ISRCTN21086382). Private households were selected randomly from address lists and previous ONS surveys on a continuous basis to provide a representative sample across the UK. Following verbal consent, a study worker visited each household to take written informed consent for individuals aged ≥2 years (from parents/carers for those 2–15 years; those 10–15 years also provided written assent). The study received ethical approval from the South Central Berkshire B Research Ethics Committee (20/SC/0195). At the first visit, participants were asked for consent for optional follow-up visits every week for the next month, then monthly thereafter (>98.5% providing such consent). At each visit, participants provided a nose and throat self-swab and completed questionnaires (https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/case-record-forms).

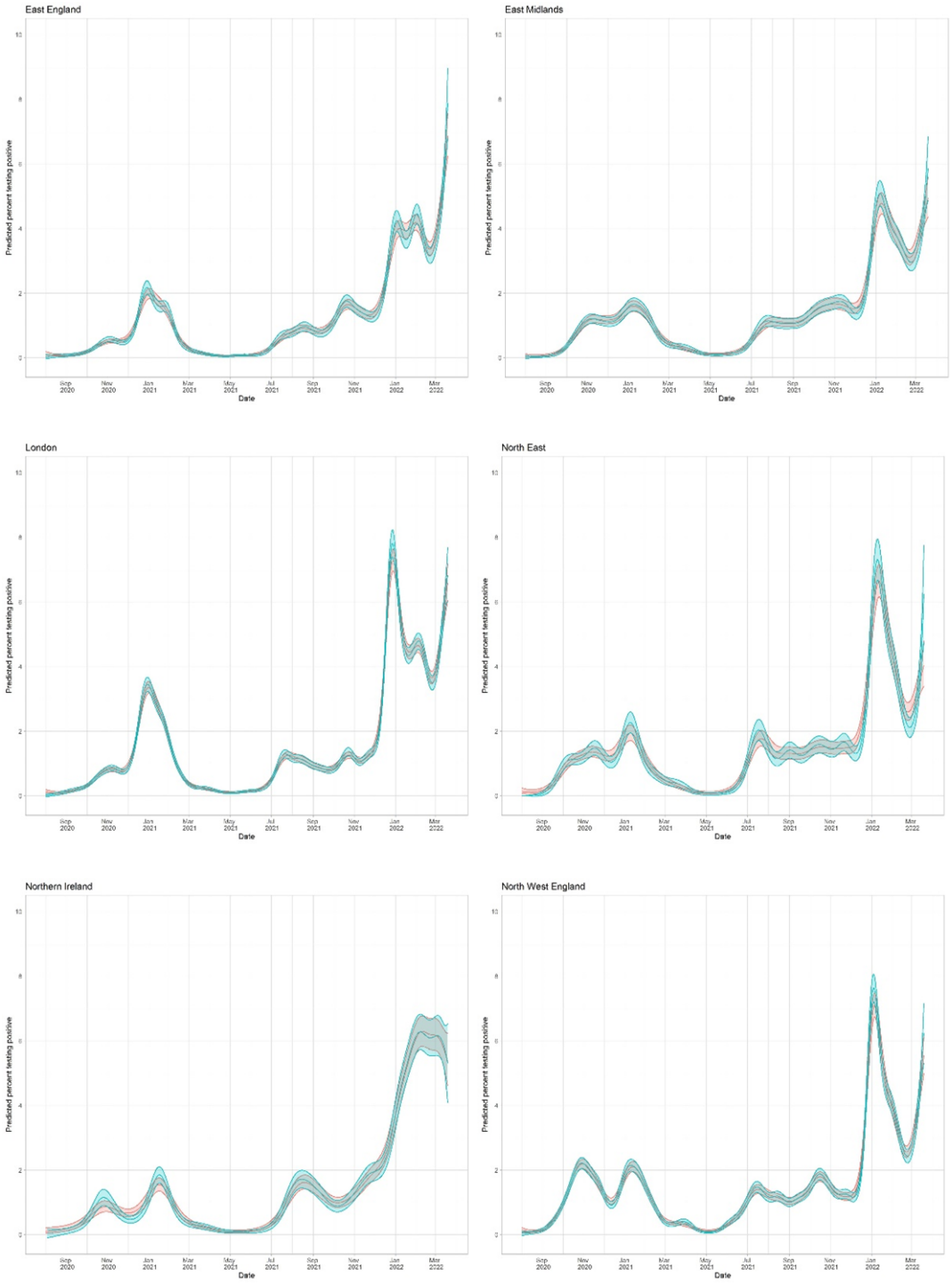
### 3.2.2 Study Population

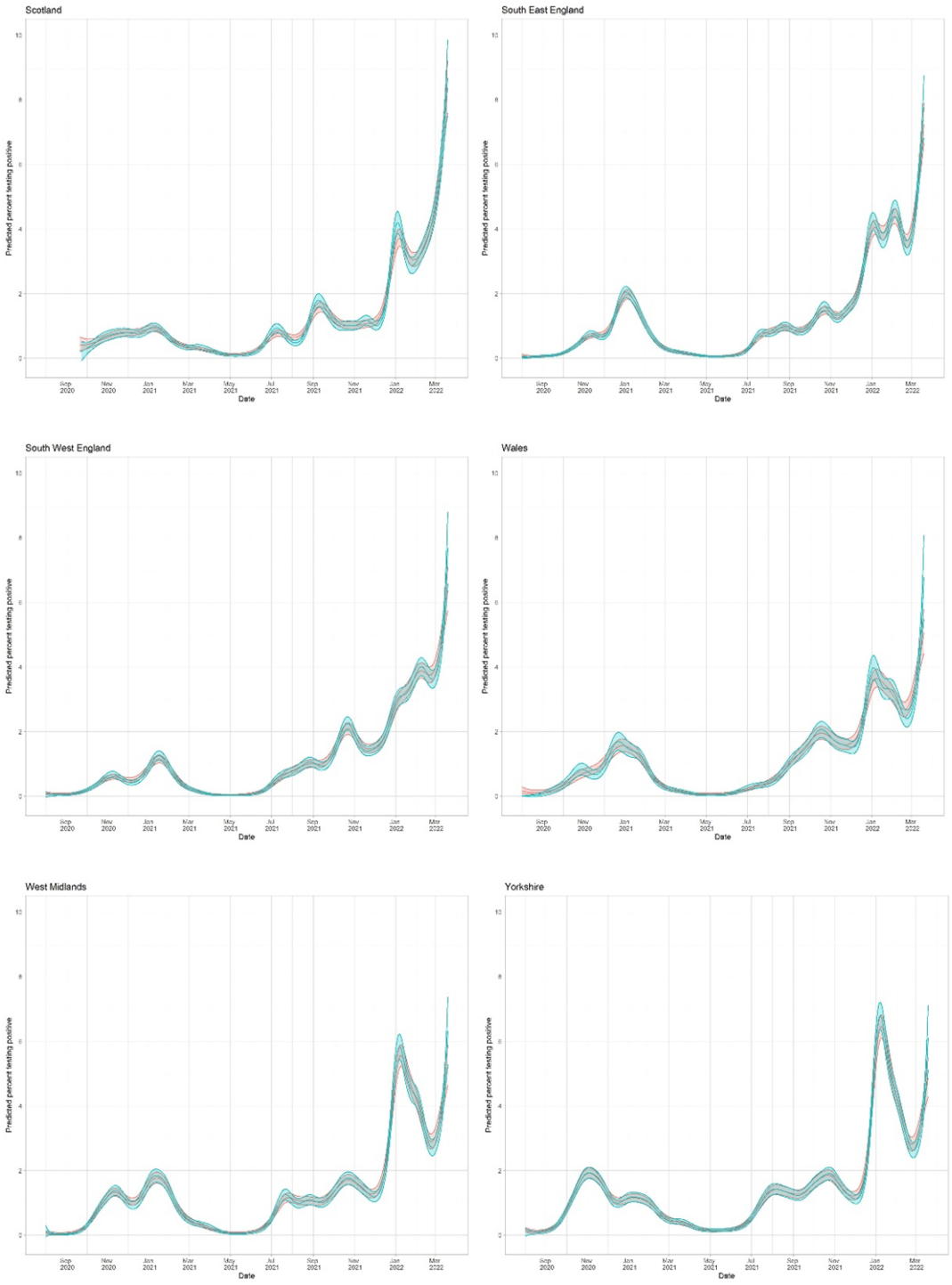
The analysis included all visits with positive or negative swabs from 1st August 2020-30th June 2022 (n=225,348 (2%) visits with void/missing results excluded).

### 3.2.3 Statistical Analyses

The outcome measure was the proportion of visits with PCR-positive SARS-CoV-2 tests. I compared two methods for detecting changes in trend over time: ISR[139] and second derivatives of GAMs.[141] All models were run separately for 12 geographical regions (9 English regions and 3 devolved administrations: Wales, Scotland, and Northern Ireland) due to positivity trend differences and ISR estimating change-points in a single time-series. Running GAMs separately by region made only very small differences to predictions versus including region-time interactions, and reduced computational time (**Figure 3.1**).

**Figure 3.1: Comparison of GAMs with region included at an interaction with time (red) and as separate models for each region (blue).**





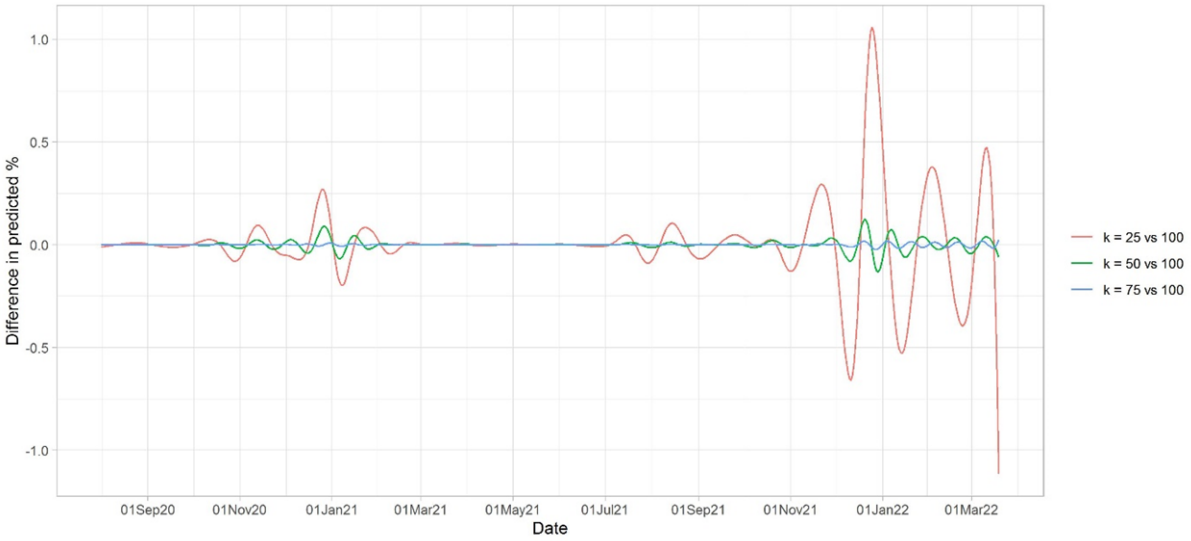

Note: The run-time for the model including the interaction between region-time was 84.6 hours, versus approximately 4 hours when running all regions separately, before derivatives were calculated.

ISR, using a negative binomial distribution with a log link allowing for overdispersion, initially fitted a log-linear trend within the first month of data to 1st September 2020 allowing no change-points to be found in this period. The model then considered the subsequent n days of the time-series, and fit two models: one extending the current trend (one-trend), the second allowing a change of trend

(two-trend). If the two-trend reduced the AIC by at least 6.635 (critical value at p=0.01, to reduce the impact of false positives), the change-point was permanently fixed in the model, otherwise, the one-trend model was chosen. A minimum length of time between change-points (interval length, 7 days) had to be pre-specified, as well as a minimum length of time between a change-point and its detection time (minimum distance, 3 days). If the one-trend model was selected, the endpoint was moved forward one day. If the two-trend model was selected, the endpoint was moved forward three days. The algorithm repeated this process until the end of the time-series. Change-points and dates change-points were permanently fixed into the model ("detection date") and were extracted from fitted models.

GAMs, using a negative binomial distribution with log link, included a single explanatory variable of time, measured in days since 1st August 2020 and modelled using thin plate splines.[144] The number of basis functions, *k*, determining smoothness, was selected from 25, 50, 75, or 100 as the lowest value with predicted positivity within ±0.25% (absolute scale) compared with *k*=100, optimising computational time, without large increases in the effective degrees of freedom[145] (**Figure 3.2**). Splines were penalised based on the third derivative as the second derivative was the measure of interest.

**Figure 3.2: Difference in predicted percentage testing positive from GAMs with varying numbers of basis functions (k) of 25, 50, 75, and 100, for London only.**




Note: The median (IQR) [range] of differences between GAMs with k = 25, 50, 75 vs k = 100 were -0.0005 (-0.08, 0.01) [-1.09, 1.79], 0.0002 (-0.007, 0.010) [-0.1287, 0.2087], and 0.000007 (-0.0015, 0.0017) [-0.036, 0.032], respectively. The effective degrees of freedom (EDF) were 23.4, 39.5, 44.6, and 45.6 for k = 25, 50, 75, and 100, respectively.

Derivatives were estimated for the smooth function using posterior simulation on the absolute scale. If positivity was relatively common throughout the entire period, coefficients from the GAM would approximately follow a multivariate normal distribution with mean vector and covariance matrix specified by the model estimates of the coefficients and their covariances, respectively.[142] Posterior simulation involves taking random draws from this distribution, whereby each draw represents a new trend that is consistent with the fitted model while also incorporating uncertainty in the estimated trend. However, this Gaussian approximation will be poor in periods where data consists of mostly zeros due to low positivity, as observed for some periods in this exemplar. To overcome this problem, I used a simple Metropolis-Hastings sampler to generate samples from the posterior distribution of the fitted model (as implemented in the *gam.mh* function from the *mgcv* R package).[146,147] This approach alternates fixed proposals – based on the typical Gaussian approximation to the posterior – with random walk proposals, based on a shrunken version of the approximate posterior covariance matrix. The random walk component ensures that the chain does not get stuck in regions for which the Gaussian proposal density is much lower than the posterior density.[146]

Code was adapted from the *derivatives* function in the R *gratia* package, which currently can only obtain derivatives on the linear predictor scale.[148] First, 2000 curves were simulated from the fitted GAMs using a Metropolis-Hastings algorithm. The first and second derivatives on the absolute scale were then calculated for each simulation using backwards finite differences, allowing estimation on the final day of data (not possible with forward of central differences). The median, 2.5th, and 97.5th percentiles were calculated across the simulations, obtaining an average with credible intervals. Change-points were defined on the first day zero was no longer within the 95% credible interval of the second derivative, corresponding to a 97.5% probability of a change, from 1st September 2020 onwards.

Positivity trends over the full time-series were compared between ISR and GAMs. Change-points were classified as found by both methods if within ±7 days, an arbitrary but pragmatic window based on the distribution of time between change-points identified by both methods and the timeliness of public health responses. To make the decision based on the distribution of time between change points, for all change-points found by the GAMs for each region (n=199), I identified the closest ISR change-point and calculated the number of days between the GAM change-point and the ISR change-point. I repeated this for the ISR change-points i.e. for all change-points found by ISR for each region (n=230), I identified the closest GAM change-point and calculated days between the ISR change-point and the GAM change-point. I plotted the number of days between change-points for

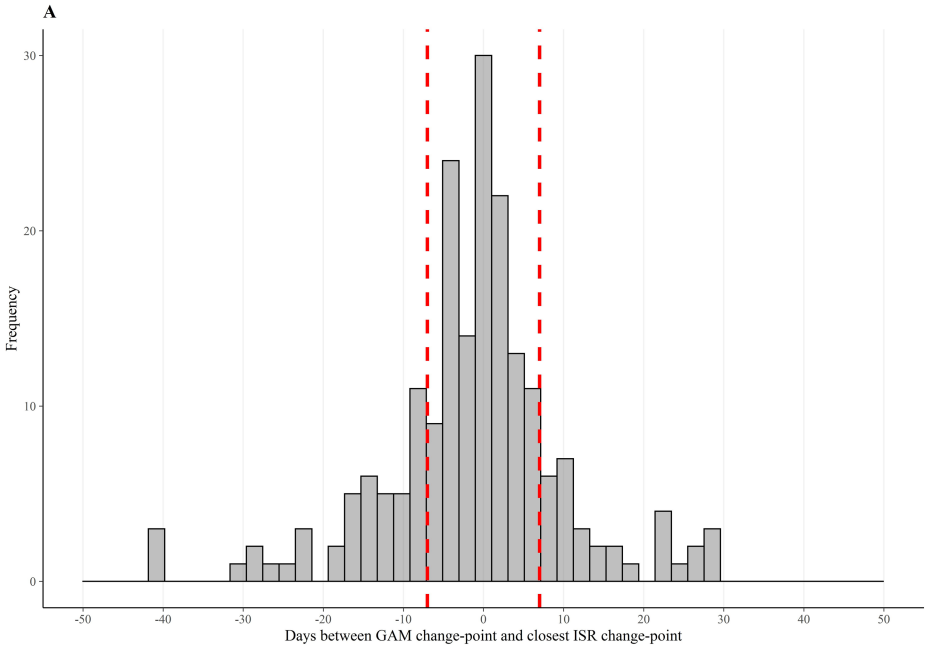
each of these analyses to identify a sensible number of days in which to classify ISR and GAM changes-points as the same.

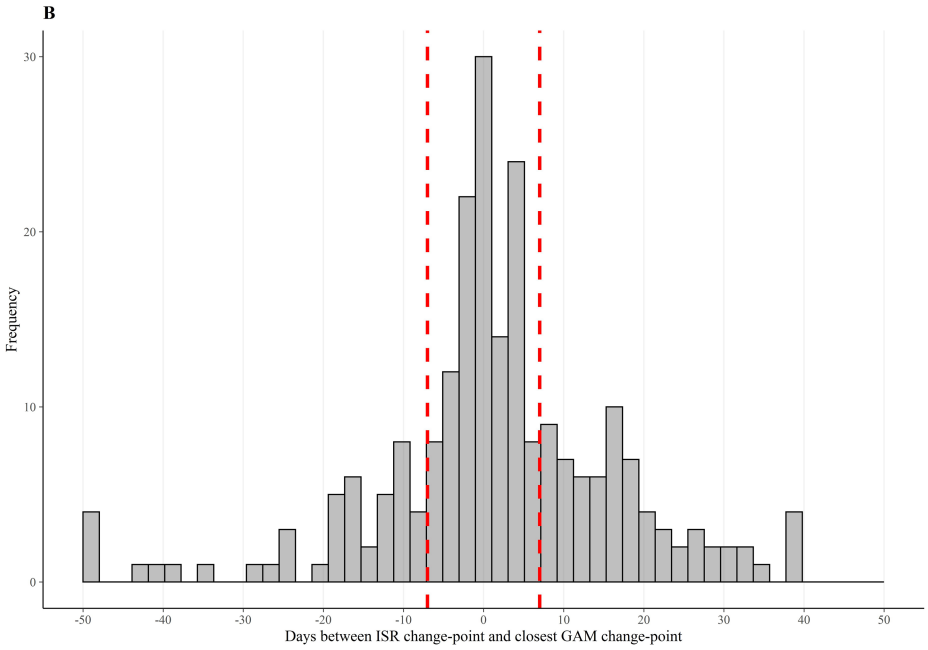
For the GAM change-points, the closest ISR change-points were a median 1 day earlier (IQR 6 days earlier, 4 days later) (**Figure 3.3**). For ISR change-points, the closest GAM change-points were a median 1 day later (IQR 5 days earlier, 10 days later). Given this distribution and the potential timeframe for public health responses, change-points were therefore assumed to reflect the same change in underlying trend if they were within ±7 days of each other, allowing the majority of closest change-points to be classified as reflecting the same underlying change in trend, whilst also being sensitive to change-points further away in time being more unlikely to be capturing the same underlying change in the data.

Change-points corresponding to the emergence of Alpha, Delta, and BA.1 variants were also compared between methods.

To summarise the importance of change-points, the relative percentage change in positivity after change-points was estimated. For each change-point identified by GAMs and ISR run on the full time-series, for each region separately, the predicted percentage testing positive on the day of each change-point was compared with the predicted percentage testing positive 4 weeks later, as estimated by each respective model. The relative percentage change was calculated between these two predicted percentages.

**Figure 3.3: Distribution of the number of days between change-points identified by GAMs for all regions and the closest ISR change-point (A), and the number of days between change-points identified by ISR and the closest GAM change-point (B).**






### 3.2.4 Detection of change-points in 'near real-time'

As ISR fixes change-points once identified and adds data progressively, it does not need to be run on segments of data sequentially to assess real-time detection. When applied in real-time, one could run ISR from the latest detected change-point onwards to decrease fitting time, albeit change-points may differ slightly versus models incorporating the full time-series as previous data can impact AIC. To assess consistency in near real-time detection between the methods, GAMs were run sequentially adopting a sliding-window approach. Sliding window length was determined by running GAMs on shorter time periods (16, 24, and 32 weeks) and assessing whether similar change-points were found in the final 8 weeks of each model, as most recent changes are of most interest in near real-time. Starting from 1st October 2020 (including data from 1st August 2020), seven-day increments of data were added until the sliding-window length was reached, from which seven days of data were removed from the start of the time-series each time seven days were added on. I selected *k* as before for sliding-window length, scaling *k* down proportionally for the shorter time-series. I checked whether all change-points identified in the last 8 weeks of each model were detected within ±7 days in five subsequent models and/or by ISR. Due to long runtimes (~12-36 hours per region including derivatives calculation), I compared GAM detection dates for the largest (London) and smallest (Northern Ireland) regions. A "detection date" for change-points identified in the GAM including data from the full time-series was defined as the last date included in the earliest successive GAM which also confirmed the change-point within ±7 days.

### 3.2.5 Sensitivity Analysis

Using the second derivative of GAMs risks potentially missing change-points if positivity decreases and increases at the same rate over a short period of time. While the second derivative will be significantly different from zero, a new change-point will not be found when positivity changes direction as the second derivative may not cross zero. I summarised the number and position of additional change-points added if placed where, over a period of the second derivative being significantly different from zero, the first derivative changed from significantly positive to negative, or visa-versa.

To assess whether earlier detection of change-points was possible by focusing on high-risk population subgroups, I estimated change-points from separate ISR and GAMs in those aged 2 years (y)-school year (sy) 11 (~aged 16y), 12sy–49y, and 50y+, and then compared these with combined all-age estimates. I also considered separate analysis by PCR gene positivity as a proxy for SARS-CoV-2 variant; Delta and BA.2 being spike (S) gene target positive (SGTP), whereas Alpha and BA.1 had S-gene target failure (SGTF). Models were run separately with SGTP and SGTF positivity as outcomes, with all other positives (including those positive on only the N gene or ORF1ab) included in the negative comparator group, comparing change-points to the "all positives" model.

All analysis was conducted in R version 4.0.2. Key analysis code is available at https://github.com/EmmaPritchard.

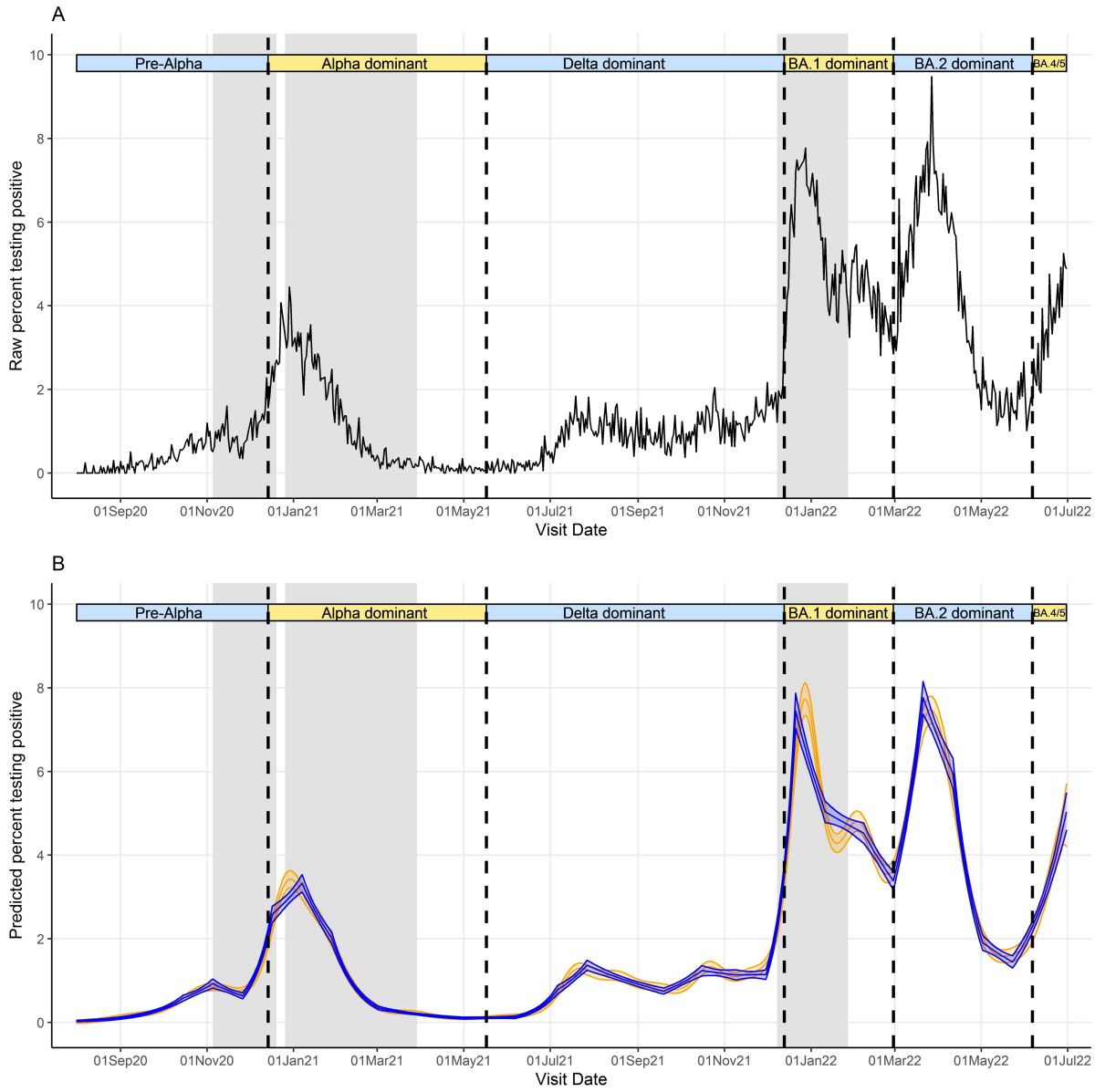
## 3.3 Results

From 1st August 2020-30th June 2022, 8,799,079 visits from 533,157 CIS participants in 266,400 households returned 147,278 (1.7%) SARS-CoV-2 positive swabs (visit characteristics in **Table 3.1**). From August-November 2020 (pre-Alpha), positivity rose to ~1%, before increasing to ~2% in January 2021, reflecting the emergence and rise of the Alpha variant (**Figure 3.4; Figure 3.5**). Positivity then decreased until June 2021 before increasing to ~1-2% in July-December 2021 (emergence of Delta). Positivity rose sharply to ~6% from December 2021 (BA.1), decreasing to ~3.5% by February 2022, before increasing to ~7.5% by mid-March (BA.2). Rises began again June 2022 (BA.4/BA.5). During the pre-Alpha period, 10% of strong positives (Ct<30) had SGTF, versus 79%, 1%, 84%, and 9% in Alpha, Delta, BA.1, and BA.2-dominant period (**Table 3.2**). Positivity varied by region, particularly between Northern/Southern English regions e.g. higher positivity pre-Alpha in Yorkshire, versus London (**Figure 3.5**).

**Table 3.1: Characteristics of all visits included in the analysis, split by swab result**

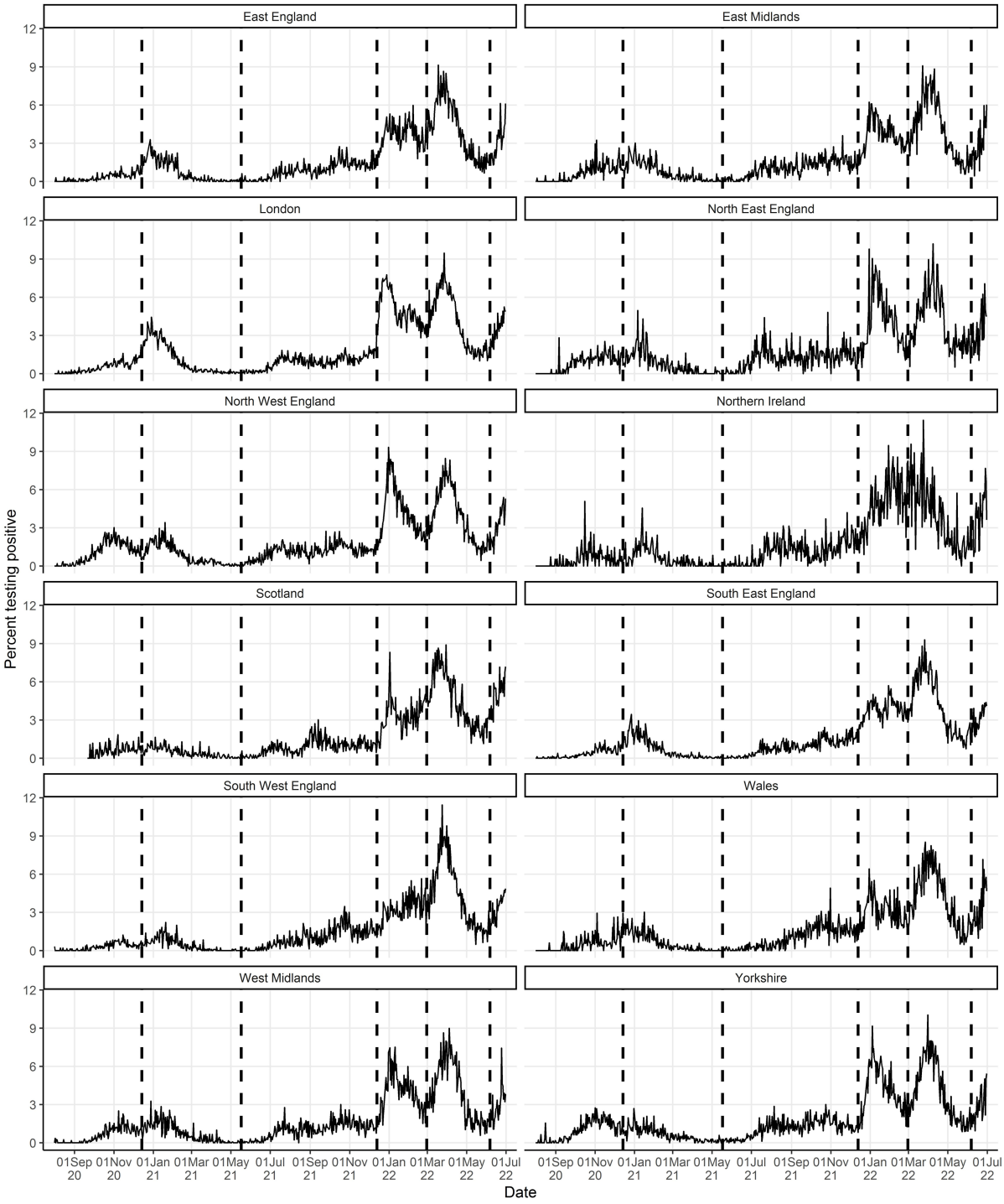
| Characteristic | Positive, n (%) or median (IQR) | Negative, n (%) or median (IQR) | Total, n (%) or median (IQR) |
|---|---|---|---|
| **Age (years)** | 47 (27, 62) | 53 (34, 67) | 53 (34, 67) |
| **Age group** | | | |
| 2y-11sy (school year) | 25,168 (17) | 977,775 (11) | 1,002,943 (11) |
| 12sy-49y | 54,950 (37) | 2,868,355 (33) | 2,923,305 (33) |
| 50+ | 67,160 (45) | 4,805,671 (55) | 4,872,831 (55) |
| **Sex** | | | |
| Male | 70,733 (48) | 4,035,518 (46) | 4,106,251 (46) |
| Female | 76,545 (51) | 4,616,283 (53) | 4,692,828 (53) |
| **Geographical region** | | | |
| Scotland | 10,679 (7) | 648,775 (7) | 659,454 (7) |
| North West England | 18,801 (12) | 994,441 (11) | 1,013,242 (11) |
| North East England | 5,831 (3) | 321,657 (3) | 327,488 (3) |
| Yorkshire | 12,889 (8) | 719,130 (8) | 732,019 (8) |
| East Midlands | 9,089 (6) | 547,238 (6) | 556,327 (6) |
| West Midlands | 11,042 (7) | 658,057 (7) | 669,099 (7) |
| East England | 12,834 (8) | 829,889 (9) | 842,723 (9) |
| Wales | 6,725 (4) | 428,267 (4) | 434,992 (4) |
| London | 26,286 (17) | 1,443,346 (16) | 1,469,632 (16) |
| South East England | 17,758 (12) | 1,128,635 (13) | 1,146,393 (13) |
| South West England | 10,823 (7) | 686,459 (7) | 697,282 (7) |
| Northern Ireland | 4,521 (3) | 245,907 (2) | 250,428 (2) |

**Figure 3.4: Raw percentage testing positive (A) and predicted percentage of visits testing positive (B) for SARS-CoV-2 from ISR (blue) and GAMs (orange) for London only**




Note: Vertical dashed lines indicate periods when new variants became dominant, defined as >50% of positive swabs with cycle threshold (Ct)<30 being S-gene target positive (ORF1ab+N+S, ORF1ab+S, N+S gene positivity) in the Covid-19 Infection Survey for the pre-Alpha period (01 August 2020 - 13 December 2020), the Delta variant (17 May 2021 – 12 December 2021), and the Omicron BA.2 variant (28 February 2022 – 5 June 2022), and >50% Ct<30 S-gene target negative (ORF1ab+N gene positivity) for the Alpha variant (14 December 2020 – 16 May 2021), Omicron BA.1 variant (13 December 2021 – 27 February 2022), and Omicron BA4/BA.5 (6 June 2022 onwards). Grey shaded areas indicate periods where stay/work from home laws were enforced, although specific restrictions varied across the time series.

**Figure 3.5: Raw daily percentage of visits with a SARS-CoV-2 positive test over the study period split by region.**




Note: Vertical dashed lines indicate periods when new variants became dominant across the UK, as defined in **Figure 3.4**.

**Table 3.2: Characteristics of SARS-CoV-2 positive swabs, split by period in which different variants dominated.**

| Characteristic | Pre-alpha | Alpha | Delta | Omicron BA.1 | Omicron BA.2 | Omicron BA.4/BA.5 |
|---|---|---|---|---|---|---|
| **Date range** | 01 August 2020 to 13 December 2020 | 14 December 2020 to 16 May 2021 | 17 May 2021 to 12 December 2021 | 13 December 2021 to 27 February 2022 | 28 February 2022 to 05 June 2022 | 06 June 2022, 30 June 2022 |
| **Number of positives, n** | 12,263 | 16,667 | 26,805 | 39,620 | 45,318 | 6,582 |
| **Ct value, median (IQR)** | 28 (21, 32) | 30 (22, 33) | 25 (19, 31) | 24 (19, 30) | 24 (20, 30) | 23 (19, 28) |
| **Ct ≥ 30 n (% pos in epoch)** | 5,106 (41) | 8,405 (50) | 7,719 (29) | 9,233 (23) | 11,127 (25) | 1,183 (18) |
| **Ct < 30, n (% pos in epoch)** | 7,203 (59) | 8,307 (50) | 19,107 (71) | 30,432 (77) | 34,200 (75) | 5,402 (82) |
| **SGTF (% pos in epoch) [% Ct <30]** | 705 (6) [10] | 6,593 (40) [79] | 242 (1) [1] | 25,462 (64) [84] | 3,206 (7) [9] | 4,054 (62) [75] |
| **S-gene detected (% pos in epoch) [% Ct <30]** | 6,452 (53) [90] | 1,669 (10) [20] | 18,844 (70) [99] | 4,925 (12) [16] | 30,985 (68) [91] | 1,345 (20) [25] |

Note: Excluding 23 positive results without Ct values or genes detected available. Epochs were defined as in **Figure 3.4**.

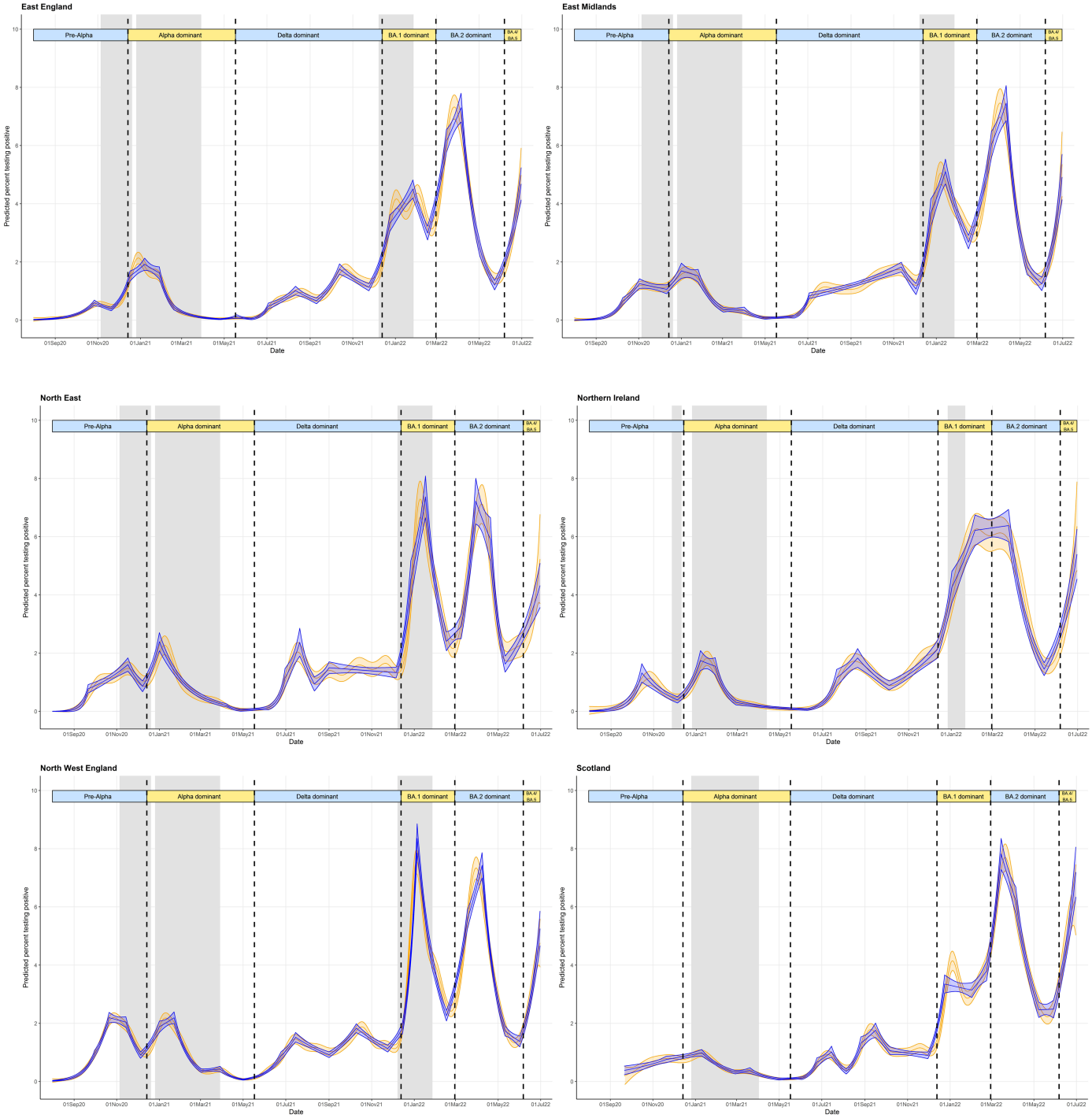
### 3.3.1 Detection of changes in growth rates using ISR and GAMs

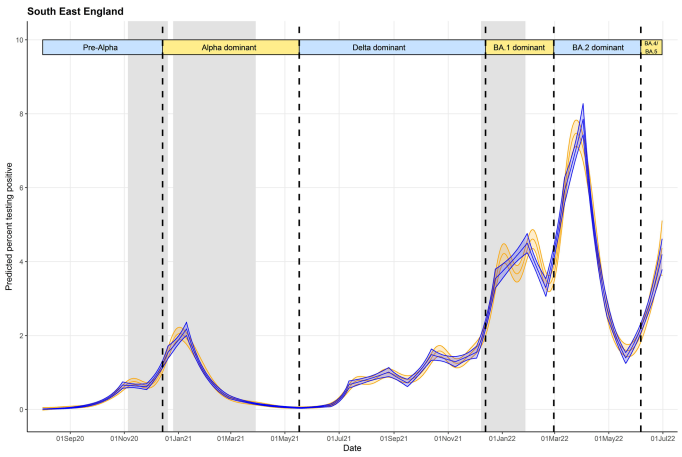
I compared change-points detected across the entire study period by the two methods, first considering the emergence of dominant SARS-CoV-2 variants. ISR and GAMs made broadly similar predictions of changing positivity trends across geographical regions over the study period (**Figure 3.4, Figure 3.6**). Focussing on London in particular, change-points corresponding to the emergence of Alpha, Delta, BA.1, and BA.2 occurred on 26 November 2020, 6 June 2021, 30 November 2021, and 28 February 2022 using ISR (**Figure 3.7**, **Table 3.3**), and 6 days earlier, 3 days later, 6 and 13 days earlier, respectively, using GAM derivatives. Across all regions, change-points for the three variants were estimated to occur a median 4 days earlier (IQR 0-8) [range 22 days later-26 days earlier] in GAMs versus ISR. 33/48 (69%) of change-points occurred earlier using GAMs. No change-point was detected for Alpha in the East Midlands or Scotland using GAMs but were detected using ISR. Change-points were also found where the key variants started to decline; for example, after the rise of BA.1, change-points indicating decreasing positivity trends were found on 20 and 21 December 2021 for GAMs and ISR, respectively. A decline in BA.2 was established on 16 and 21 March 2022 followed by increases in BA.4/BA.5 on 26 and 23 May 2022 using GAMs vs ISR, respectively.

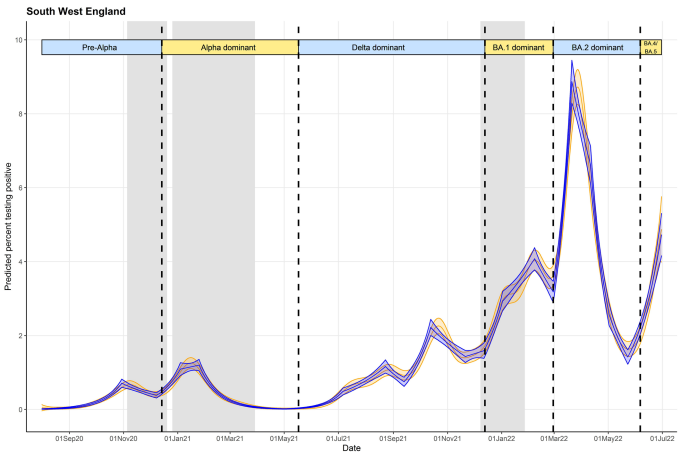
Both methods also identified other change-points aside from trend increases resulting from the emergence of these variants. For example, in London, a change-point indicating faster growth pre-Alpha occurred on 26 September and 15 October 2020 using GAMs and ISR, respectively (**Table 3.4**). ISR and GAMs identified slowing of Alpha growth on 17 and 19 December 2020, respectively. ISR identified a change-point of positive to negative growth on 7 January 2021. Both methods identified a faster decline of Alpha from late-January 2021. GAMs identified a subsequent slowing in declines on 5 February 2021, with the equivalent slowing on 2 March 2021 using ISR. A slowing in the growth of Delta was observed on 12 and 6 July 2021 using GAMs and ISR, respectively.

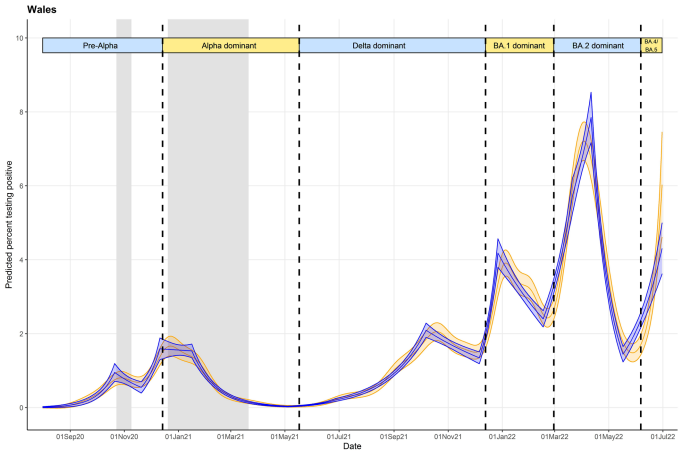
63% (12/19) of all change-points in London identified in GAMs were identified using ISR within ±7 days (**Table 3.4**). 57% (12/21) of all change-points in London identified by ISR were identified by GAMs within ±7 days. Inconsistent change-points between methods generally reflected small fluctuations when positivity was low (explored for all regions in the next section).

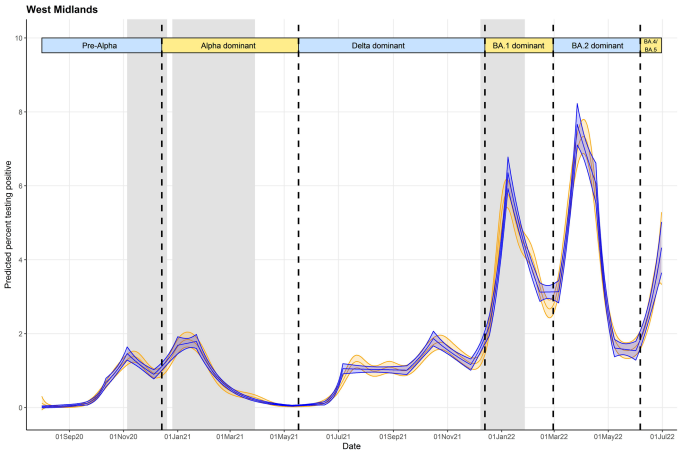
**Figure 3.6: Predicted percentage of visits testing positive for SARS-CoV-2 from ISR (blue) and GAMs (orange) for all regions**

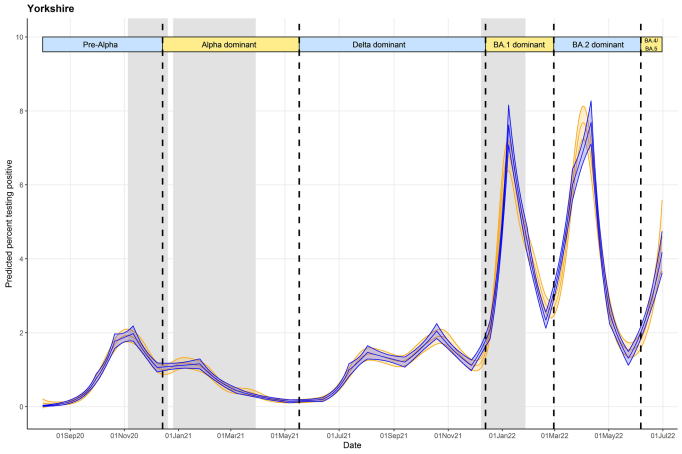




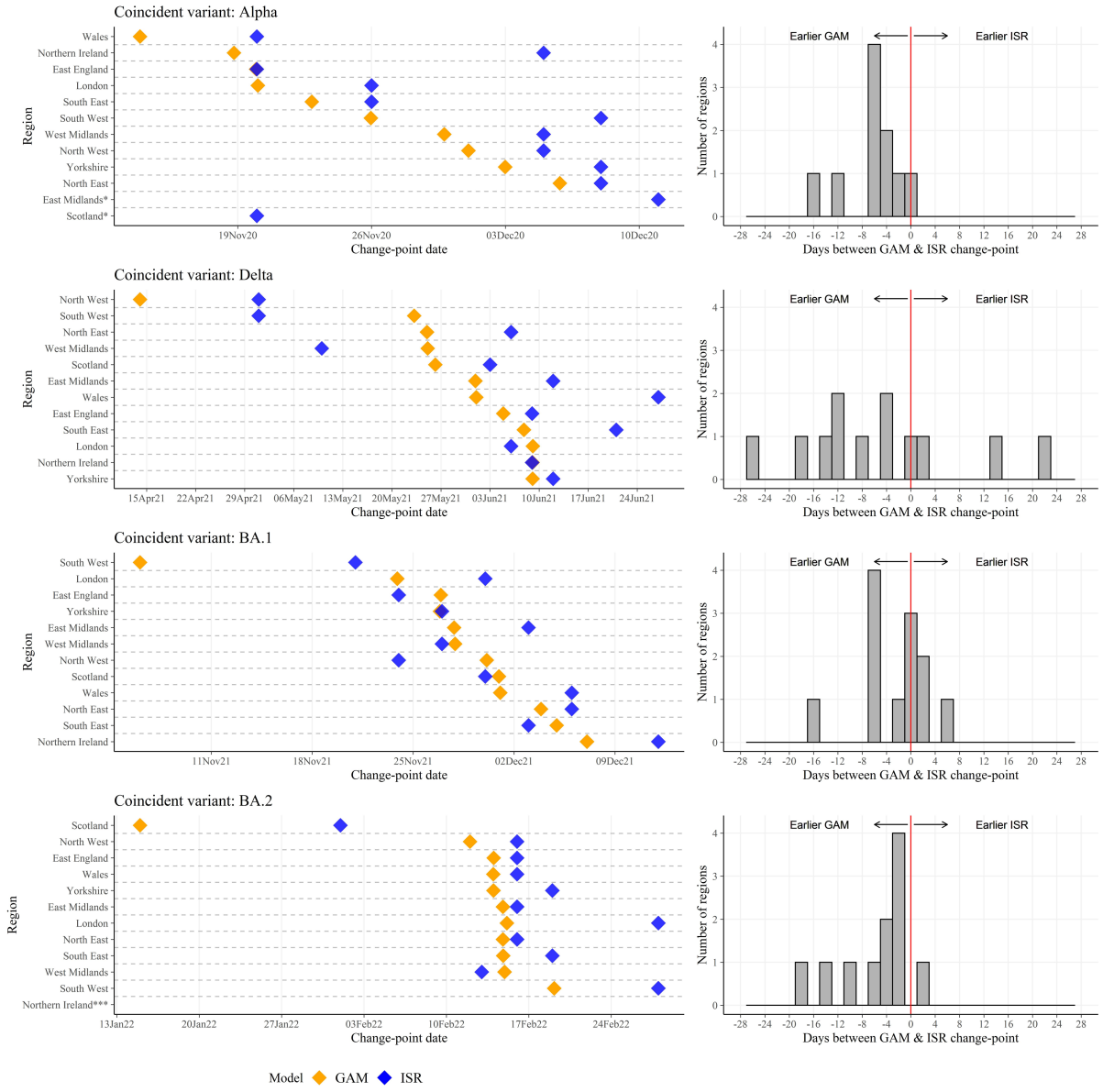




Note: Vertical dashed lines and grey shaded areas are defined in **Figure 3.4.**

**Figure 3.7: Change-points corresponding to the emergence of three key SARS-CoV-2 variants found by iterative sequential regression (ISR) and second derivatives of generalised additive models (GAM) for each geographical region, run on the full time-series.**




*No change-point coincident with variant using GAMs (East Midlands and Scotland while Alpha was the coincident variant, top left panel);***No change-point coincident with variant for GAMs and ISR (Northern Ireland while BA.2 was the coincident variant). Exact dates of change-points are in **Table 3.3**.

**Table 3.3: Change-points corresponding to periods corresponding to the emergence of four key SARS-CoV-2 variants found by ISR and second derivatives of GAMs for each geographical region, run on the full time-series.**

| Region | Coincident Variant | GAM breakpoint (DD.MM.YYYY) | ISR breakpoint (DD.MM.YYYY) | ISR detection date | Days between ISR and GAM change-point* | Days between ISR change-point and detection |
|---|---|---|---|---|---|---|
| **East England** | Alpha | 20.11.2020 | 20.11.2020 | 14.12.2020 | 0 | 24 |
| | Delta | 05.06.2021 | 09.06.2021 | 06.07.2021 | -4 | 27 |
| | BA.1 | 27.11.2021 | 24.11.2021 | 18.12.2021 | 3 | 24 |
| | BA.2 | 14.02.2022 | 16.02.2022 | 12.03.2022 | -2 | 24 |
| **East Midlands** | Alpha | **Not found** | 11.12.2020 | 04.01.2021 | n/a | 24 |
| | Delta | 01.06.2021 | 12.06.2021 | 06.07.2021 | -11 | 24 |
| | BA.1 | 28.11.2021 | 03.12.2021 | 27.12.2021 | -5 | 24 |
| | BA.2 | 15.02.2022 | 16.02.2022 | 12.03.2022 | -1 | 24 |
| **London** | Alpha | 20.11.2020 | 26.11.2020 | 20.12.2020 | -6 | 24 |
| | Delta | 09.06.2021 | 06.06.2021 | 30.06.2021 | 3 | 24 |
| | BA.1 | 24.11.2021 | 30.11.2021 | 24.12.2021 | -6 | 24 |
| | BA.2 | 15.02.2022 | 28.02.2022 | 24.03.2022 | -13 | 24 |
| **North East** | Alpha | 06.12.2020 | 08.12.2020 | 01.01.2021 | -2 | 24 |
| | Delta | 25.05.2021 | 06.06.2021 | 09.07.2021 | -12 | 33 |
| | BA.1 | 04.12.2021 | 06.12.2021 | 30.12.2021 | -2 | 24 |
| | BA.2 | 15.02.2022 | 16.02.2022 | 12.03.2022 | -1 | 24 |
| **Northern Ireland** | Alpha | 19.11.2020 | 05.12.2020 | 04.01.2021 | -16 | 30 |
| | Delta | 09.06.2021 | 09.06.2021 | 03.07.2021 | 0 | 24 |
| | BA.1 | 07.12.2021 | 12.12.2021 | 05.01.2022 | -5 | 24 |
| | BA.2 | n/a | n/a | n/a | n/a | n/a |
| **North West** | Alpha | 01.12.2020 | 05.12.2020 | 29.12.2020 | -4 | 24 |
| | Delta | 14.04.2021 | 01.05.2021 | 25.05.2021 | -17 | 24 |
| | BA.1 | 30.11.2021 | 24.11.2021 | 18.12.2021 | 6 | 24 |
| | BA.2 | 12.02.2022 | 16.02.2022 | 12.03.2022 | -4 | 24 |
| **Scotland** | Alpha | **Not found** | 20.11.2020 | 23.12.2020 | n/a | 33 |
| | Delta | 26.05.2021 | 03.06.2021 | 27.06.2021 | -8 | 24 |
| | BA.1 | 01.12.2021 | 30.11.2021 | 24.12.2021 | 1 | 24 |
| | BA.2 | 15.01.2022 | 01.02.2022 | 25.02.2022 | -17 | 24 |
| **South East** | Alpha | 23.11.2020 | 26.11.2020 | 20.12.2020 | -3 | 24 |
| | Delta | 08.06.2021 | 21.06.2021 | 15.07.2021 | -13 | 24 |
| | BA.1 | 05.12.2021 | 03.12.2021 | 27.12.2021 | 2 | 24 |
| | BA.2 | 15.02.2022 | 19.02.2022 | 15.03.2022 | -4 | 24 |
| **South West** | Alpha | 26.11.2020 | 08.12.2020 | 01.01.2021 | -12 | 24 |
| | Delta | 23.05.2021 | 01.05.2021 | 25.05.2021 | 22 | 24 |
| | BA.1 | 06.11.2021 | 21.11.2021 | 18.12.2021 | -15 | 27 |
| | BA.2 | 19.02.2022 | 28.02.2022 | 24.03.2022 | -9 | 24 |
| **Wales** | Alpha | 14.11.2020 | 20.11.2020 | 14.12.2020 | -6 | 24 |
| | Delta | 01.06.2021 | 27.06.2021 | 01.10.2021 | -26 | 96 |
| | BA.1 | 01.12.2021 | 06.12.2021 | 30.12.2021 | -5 | 24 |
| | BA.2 | 14.02.2022 | 16.02.2022 | 12.03.2022 | -2 | 24 |
| **West Midlands** | Alpha | 30.11.2020 | 05.12.2020 | 29.12.2020 | -5 | 24 |
| | Delta | 25.05.2021 | 10.05.2021 | 06.06.2021 | 15 | 27 |

| Region | Coincident Variant | GAM breakpoint (DD.MM.YYYY) | ISR breakpoint (DD.MM.YYYY) | ISR detection date | Days between ISR and GAM change-point* | Days between ISR change-point and detection |
|---|---|---|---|---|---|---|
| | BA.1 | 28.11.2021 | 27.11.2021 | 21.12.2021 | 1 | 24 |
| | BA.2 | 15.02.2022 | 13.02.2022 | 09.03.2022 | 2 | 24 |
| **Yorkshire** | Alpha | 03.12.2020 | 08.12.2020 | 01.01.2021 | -5 | 24 |
| | Delta | 09.06.2021 | 12.06.2021 | 06.07.2021 | -3 | 24 |
| | BA.1 | 27.11.2021 | 27.11.2021 | 21.12.2021 | 0 | 24 |
| | BA.2 | 14.02.2022 | 19.02.2022 | 15.03.2022 | -5 | 24 |

**Negative values indicate earlier occurrence of change-points using GAMs, compared with ISR.*

**Table 3.4: All changepoints found by iterative sequential regression (ISR) and second derivatives of generalised additive models (GAM) for London.**

| GAM change-point date | ISR change-point date | Description of trend | Days between ISR and GAM change-point |
|---|---|---|---|
| 26/09/2020 | - | Faster growth | - |
| - | 15/10/2020 | Faster growth | - |
| 02/11/2020 | 05/11/2020 | Increase to decrease | -3 |
| 20/11/2020 | 26/11/2020 | Decrease to increase (rise in Alpha) | -6 |
| 19/12/2020 | 17/12/2020 | Slower growth (slowing down of growth in Alpha) | 2 |
| - | 07/01/2021 | Increase to decrease | - |
| 23/01/2021 | 28/01/2021 | Faster decline | -5 |
| 05/02/2021 | - | Slower decline | - |
| - | 02/03/2021 | Slower decline | - |
| - | 01/05/2021 | Decrease to increase | - |
| 09/06/2021 | 06/06/2021 | Faster growth (rise of Delta) | 3 |
| 12/07/2021 | 06/07/2021 | Slower growth | 6 |
| - | 27/07/2021 | Increase to decrease | - |
| 25/09/2021 | 19/09/2021 | Decrease to increase | 6 |
| 15/10/2021 | 16/10/2021 | Increase to decrease | -1 |
| 01/11/2021 | - | Decrease to increase | - |
| - | 09/11/2021 | Decrease to increase | - |
| 24/11/2021 | 30/11/2021 | Faster growth (rise of BA.1) | -6 |
| 20/12/2021 | 21/12/2021 | Increase to decrease (decline of BA.1) | -1 |
| 06/01/2022 | 11/01/2022 | Slower decline | -5 |
| 29/01/2022 | - | Slower decline | - |
| - | 07/02/2022 | Decrease to increase | - |
| 15/02/2022 | - | Faster growth (rise of BA.2) | - |
| - | 28/02/2022 | Decrease to increase (rise of BA.2) | - |
| 16/03/2022 | 21/03/2022 | Increase to decrease (decline of BA.2) | -5 |

| GAM change-point date | ISR change-point date | Description of trend | Days between ISR and GAM change-point |
|---|---|---|---|
| - | 11/04/2022 | Faster decline | - |
| 19/04/2022 | - | Faster decline | - |
| - | 02/05/2022 | Slower decline | - |
| 26/05/2022 | 23/05/2022 | Decrease to increase (rise of BA.4/BA.5) | 3 |

### 3.3.2 Relative percentage change in positivity after change-points

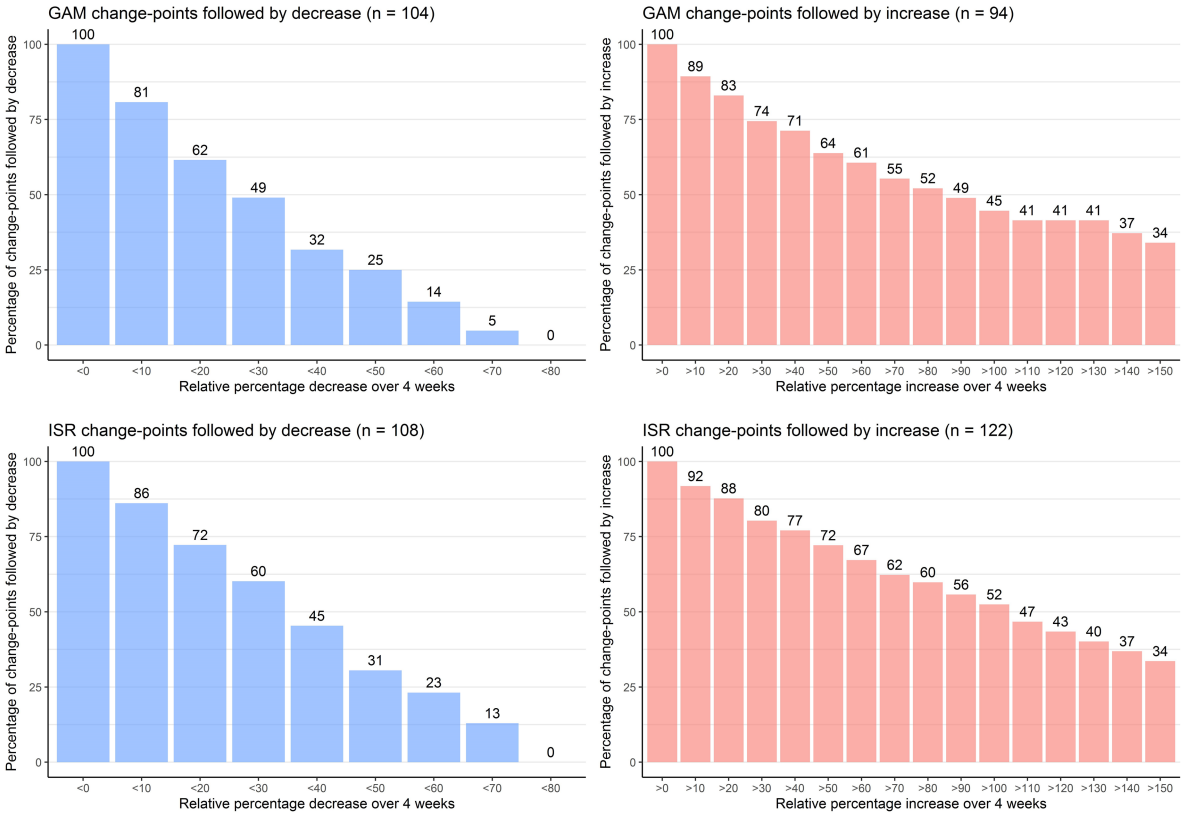
One way to evaluate the importance of change-points across the full time series was to quantify the size of the change in the percentage testing positive after a change-point occurred.

Across all regions, 199 change-points were identified using GAMs run on the full time-series. Of these change-points, 104 (52%) were followed by a decrease in positivity over 4 weeks, 94 (47%) were followed by an increase, and one (1%) change-point was less than 4 weeks before the end of the time series (**Figure 3.8**). Of those change-points followed by decreasing positivity, 25% had a relative decrease of more than 50%; 62% had a relative decrease of more than 20%, and only 19% of decreasing GAM change-points had a relative decrease of 0-10% (remainder 10-20%). For change-points followed by increasing positivity, 34% had a relative increase of more than 150%; 45% had a more than 100% relative increase; 83% had a more than 20% relative increase, and only 11% of increasing GAM change-points had a relative increase of between 0-10% (remainder 10-20%).

For ISR, 230 changes-points were identified across all regions of which 108 (47%) were followed by decreasing positivity over 4 weeks and 122 (53%) were followed by increasing positivity (**Figure 3.8**). Of those change-points followed by decreasing positivity, 31% had a relative decrease of more than 50%; 72% had a relative decrease of more than 20%, and only 14% of decreasing ISR change-points had a relative decrease of 0-10% (remainder 10-20%). For change-points followed by increasing positivity, 34% had a relative increase of more than 150%; 52% had a more than 100% relative increase; 88% had a more than 20% relative increase, and only 8% of increasing ISR change-points had a relative increase of between 0-10% (remainder 10-20%).

Overall, GAMs and ISR had a similarly low proportion of change-points with small relative changes in the percentage testing positive. Increases in trends were generally larger than decreases but the vast majority of change-points were followed by a relative increase or decrease >20%.

**Figure 3.8: The relative percentage decrease (blue) or increase (red) in positivity on the date of the detected change-point compared with positivity 4 weeks later.**




Note: Results are presented for second derivatives of generalised additive models (GAMs; top) and iterative sequential regression (ISR; bottom). Models run on all 12 regions across the full time-series are included.

### 3.3.3 Detection of change-points in 'near real-time'

In real-time, often most interest is in change-points at the end of a time-series, for example, the final 8-weeks. Rather than running GAMs from the start of the time series (1st August 2020) each time I wanted to find new change-points at the end of the time-series, to improve computational efficiency I assessed whether the same change-points were estimated if GAMs were only run on double (16-weeks), triple (24-weeks), or quadruple (32-weeks) the period of interest. The shorter time-frames of 16-weeks and 24-weeks missed over half the change-points in the full time-series in both cases so were not considered further (**Table 3.5**).

In contrast, the 32-week model found the majority of change-points in the full time-series (8/10). In the 32-week model, one change-point on 21 February 2021 was not identified by the full model in the model ending 18 March 2021. This change in the second derivative was only significant for two days, thus may not be a meaningful change. The second derivative became significant again on 24 February 2021 for nine days, matching the change-point in the full model on 23 February 2021. With models ending on 23 December 2021 and 17 February 2022, the 32-week model missed two change-points identified in the full time-series (4 Nov 2021, 2 Feb 2022). While these were not change-

points indicating substantial growth/decay of variants, they were both identified by ISR (9 Nov 2021, 7 Feb 2022; **Table 3.4**). Thus, while a 32-week GAM appeared to identify the majority of change-points, if the capacity is available to run models on the full time-series, statistical power will likely be increased.

Running GAMs sequentially adding new data for London every week from 1st October 2020-30th June, 96 change-points were found in the final 8-weeks across all GAMs (**Figure 3.9**). The majority (64/96: 67%) of change-points were identified by five successive GAMs. Eight (8%) change-points were not identified in any of the five subsequent GAM models, but four of these were identified by ISR. Overall, 77% (74/96) of change-points in the last 8-weeks of successive GAMs were identified by ISR, and 23% (22/96) were never identified by ISR.

Results were broadly similar for Northern Ireland (the smallest region in the dataset), where 52 change-points were found in the final 8-weeks across all GAMs (**Figure 3.9**). The majority (31/52: 60%) of these change-points were identified by five successive GAMs. One change-point was not identified in any of the five subsequent GAM models but was identified by ISR. Overall, 52% (27/52) of change-points in the last 8-weeks of successive GAMs were identified by ISR, and 48% (25/52) were never identified by ISR.

Using the final date of the first successive GAM to estimate when change-points in the full time-series GAM would have been detected, for London, change-points were detected a median 21 (IQR 17-26; range 10-128) days after the change in growth was estimated to occur (**Figure 3.10**). ISR generally fixed change-points into the model (based on lower AIC vs linear trend) around 24 days after the change. When identified by both GAMs and ISR, GAMs detected change-points a median 4 days earlier (IQR 10 days earlier, 1 day later; range 17 days earlier-35 days later). Four change-points identified in the final GAM for London were not identified in any successive GAMs, hence a detection date could not be determined.

When considering change-points for Northern Ireland (around a fifth of the number of visits in London), while ISR still consistently detected change-points ~24 days after the change occurred, GAMs detected changes median 30 days after (IQR: 24-54; range: 8-108) (**Figure 3.10**). When identified by both ISR and GAMs, in contrast to London, ISR detected change-points a median 10 days earlier (IQR: 0, 32).

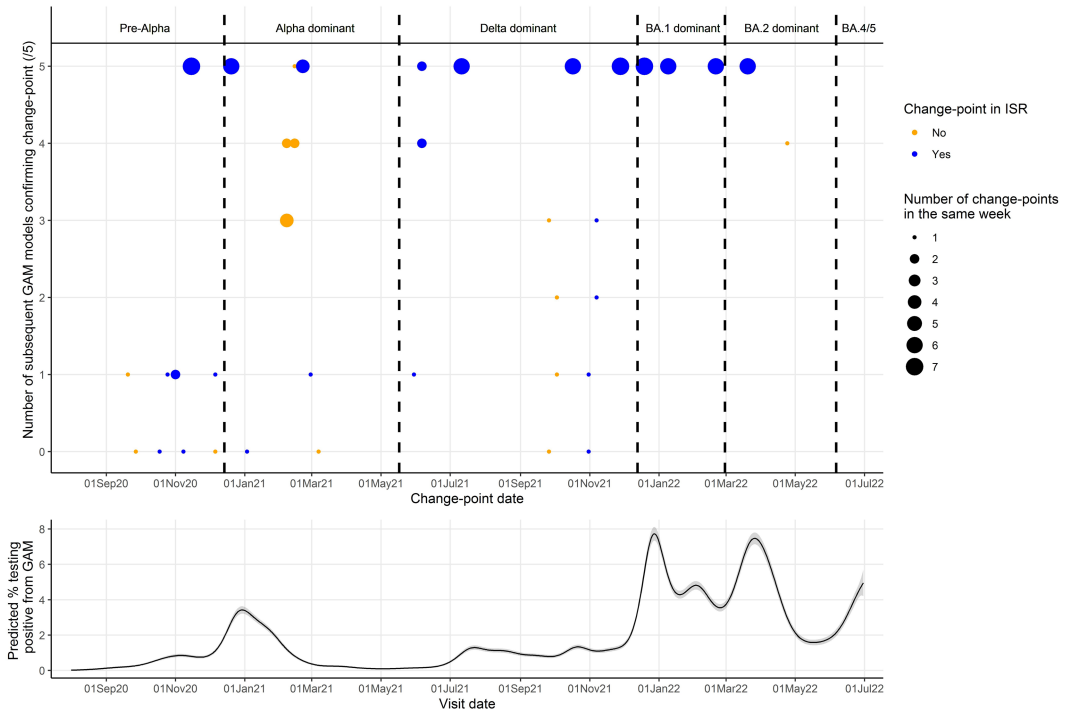
**Table 3.5: Comparison of change-points detected by generalised additive models run on the full time series from 1st August 2020, 16-week, 24-week, and 32-week periods for London**

| Model end date (days from 1st August 2020) | Change-point dates from model run from 1st August 2020 [duration*, days] | Change-point dates (days between current change-point and change-point in full model) [duration*, days] | | |
|---|---|---|---|---|
| | | 16-week model (112 days) | 24-week model (168 days) | 32-week model (224 days) |
| 26-11-2020 (118 days) | 06-11-2020 [14] | 04-11-2020 (-2) [17] | n/a | n/a |
| 21-01-2021 (174 days) | 09-12-2020 [5] | - | 11-12-2020 (2) [3] | n/a |
| | 23-12-2020 [9] | 25-12-2020 (2) [6] | 24-12-2020 (1) [7] | n/a |
| | - | 05-01-2021 (n/a) [3] | - | n/a |
| 18-03-2021 (230 days) | 10-02-2021 [1] | - | - | 10-02-2021 (0) [3] |
| | 12-02-2021 [2] | - | - | 14-02-2021 (-2) [1] |
| | - | - | - | 21-02-2021 (n/a) [2] |
| | 23-02-2021 [9] | 26-02-2021 (3) [7] | - | 24-02-2021 (-1) [9] |
| 13-05-2021 (286 days) | No change-points | No change-points | No change-points | No change-points |
| 08-07-2021 (342 days) | 13-06-2021 [17] | - | - | 14-06-2021 (-1) [19] |
| 02-09-2021 (398 days) | 13-07-2021 [12] | 14-07-2021 (1) [9] | 14-07-2021 (1) [11] | 14-07-2021 (1) [12] |
| 28-10-2021 (454 days) | - | 10-09-2021 (n/a) [45] | - | - |
| | 04-10-2021 [6] | - | - | 04-10-2021 (0) [5] |
| 23-12-2021 (510 days) | 04-11-2021 [10] | - | - | - |
| | 27-11-2021 [27] | 04-12-2021 (7) [12] | 03-12-2021 (6) [14] | 02-12-2021 (5) [16] |
| 17-02-2022 (566 days) | 06-01-2022 [18] | 12-01-2022 (6) [6] | 12-01-2022 (6) [5] | 11-01-2022 (5) [8] |
| | 02-02-2022 [5] | - | - | - |

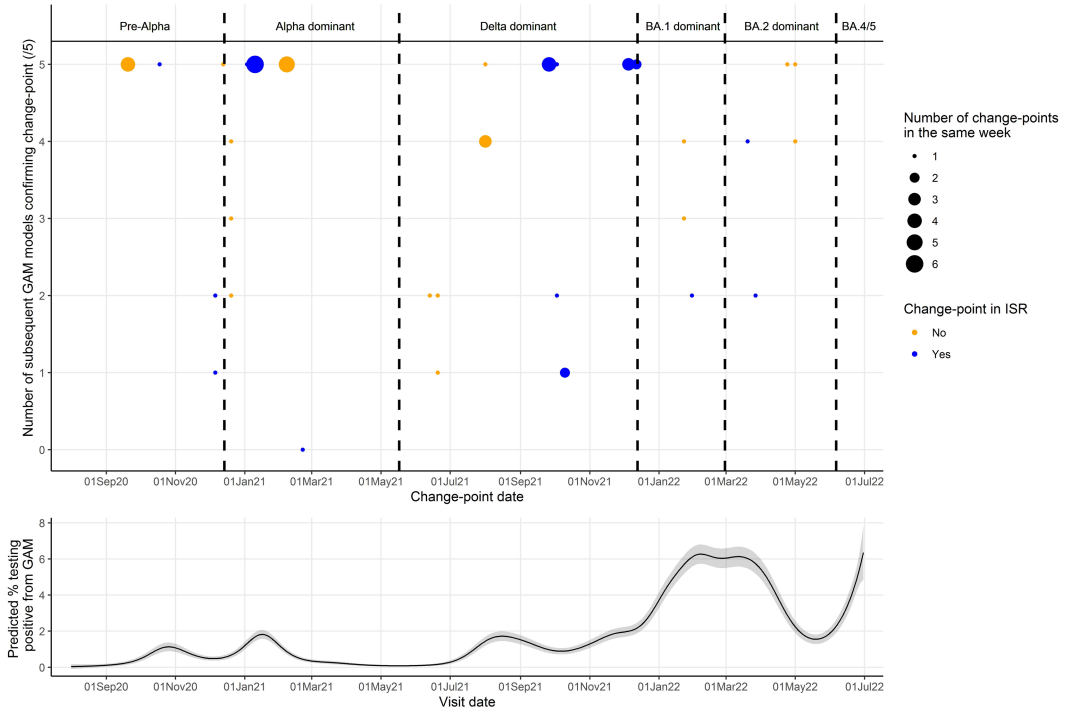
*Duration = number of days that the credible interval of the second derivative did not contain zero Note: Change-points recorded as “n/a” are not applicable as the 24-week and/or 32-week model is identical to the model run from 1st August 2020.

**Figure 3.9: Number of successive GAMs (zero to five), and ISR, finding the same change-points (top panel). Predicted positivity from final GAM for reference (bottom panel). Results are for London (A) and Northern Ireland (B).**

**A: London**



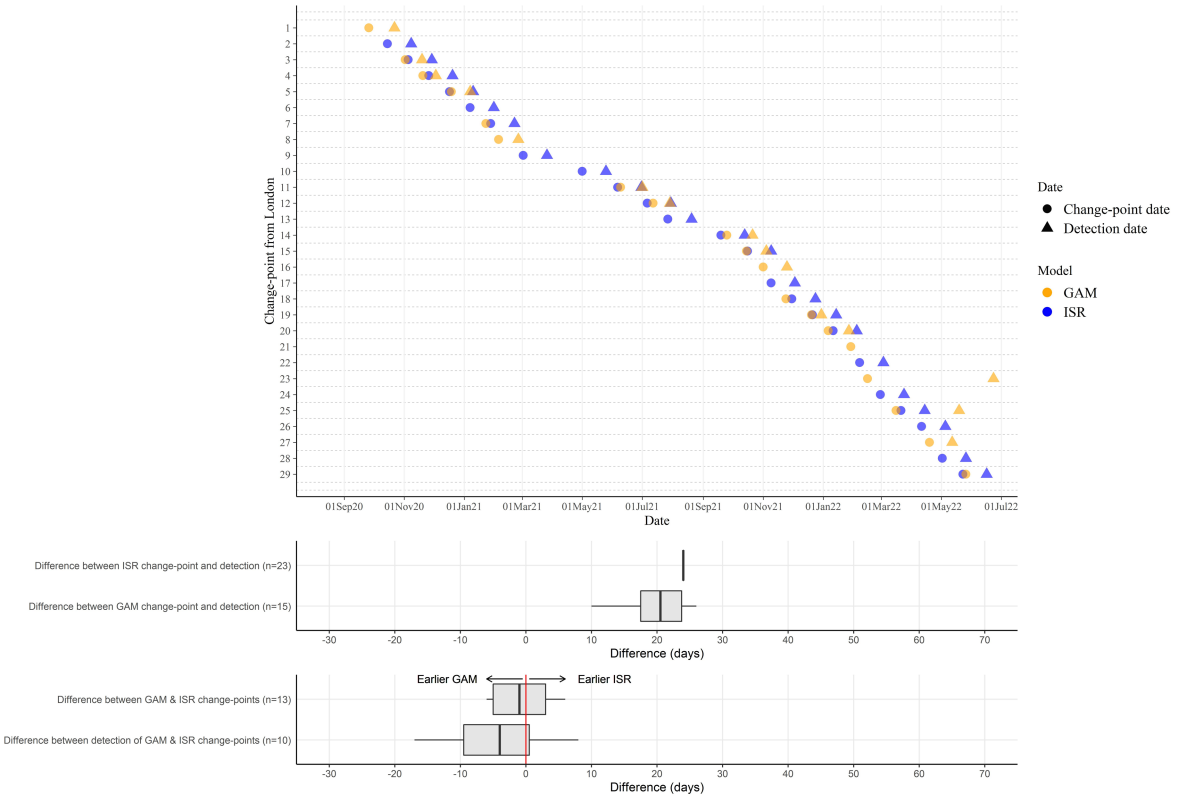

**B: Northern Ireland**



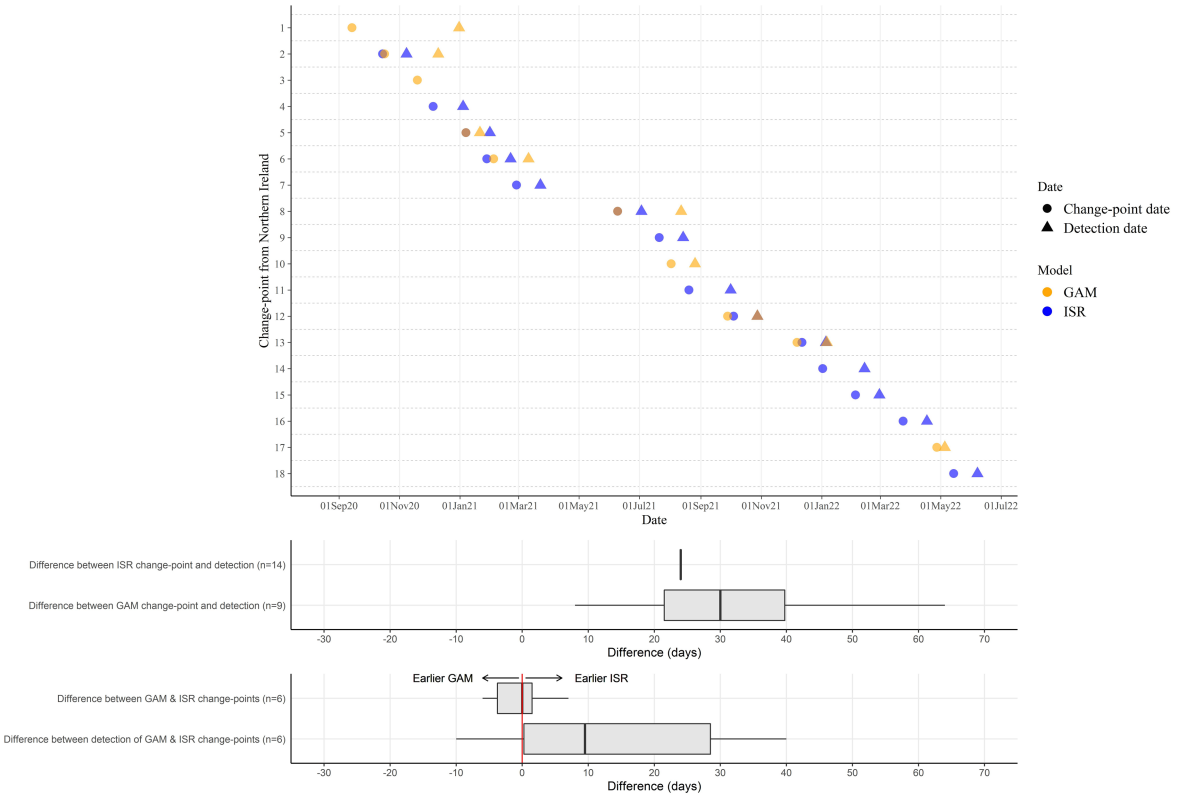

Note: Change-points in the same week (starting Monday) found in the same number of subsequent models were grouped together (indicated by size of circle). Points are blue if at least one change-point in that week was also found by ISR, and orange is no change-points in that week were found by ISR.

**Figure 3.10: Difference in detection dates between GAMs (orange) and ISR (blue) for London (A) and Northern Ireland (B).**

## A: London




## B: Northern Ireland



### 3.3.4 Incorporating change-points based on the first derivative

Change-points corresponding to the emergence of BA.4/BA.5 were found for all regions using ISR but were not found using GAMs for 9/12 regions (**Table 3.6**). The growth rate of BA.4/BA.5 was similar to that of BA.2 declining so, while the second derivative was significantly different from zero, a new change-point was not established. Adding in additional change-points where the first derivative switched signs, all regions found change-points for BA.4/BA.5 using GAMs (**Figure 3.11**).

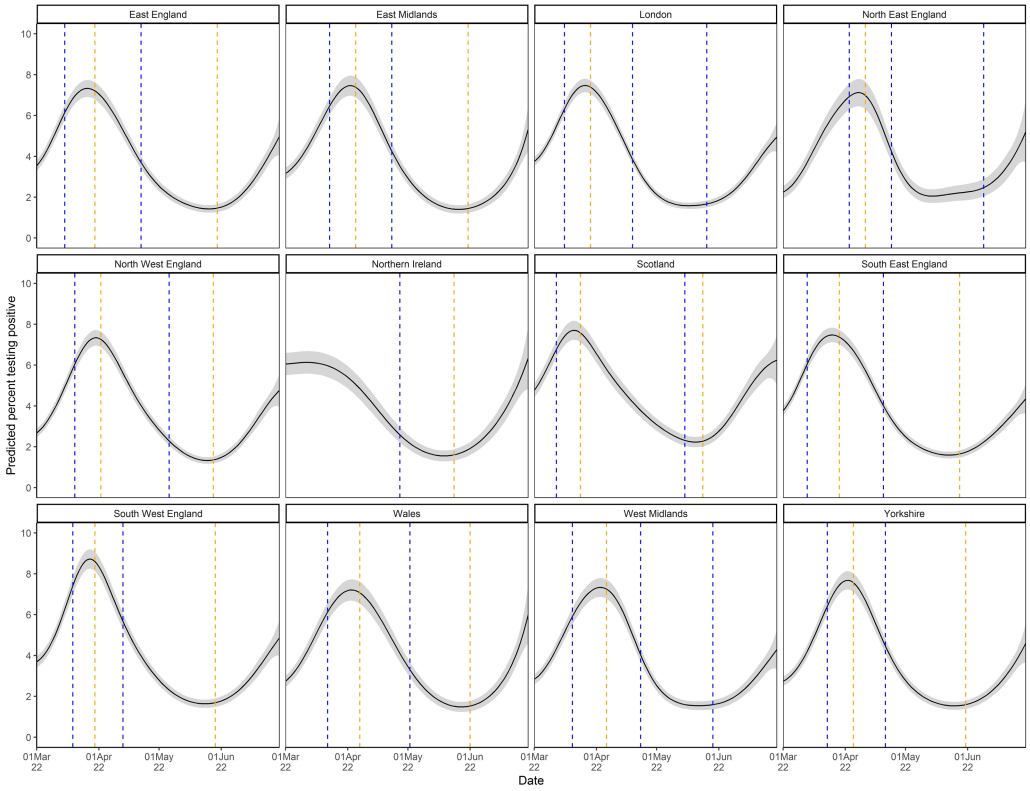
Overall, adding in additional change-points where the first derivative switched signs during a period of significance in the second derivative, an additional 78 change-points were established across the full time-series for all regions using GAM models (199 change-points based on the second derivative only). The largest number of additional change-points occurred in South East England, with 10 additional change-points above the 20 original change-points established using the second derivative only (**Table 3.6**). The majority of the additional change-points occurred in January and February 2022, concurrent with the rise and fall of BA.1.

**Table 3.6: Change-points from GAMs and ISR fitted on the full time-series for BA.4/ BA.5. Change-points from GAMs are presented for both change-points defined by the second derivative alone, as well as additional change-points based on the first derivative.**

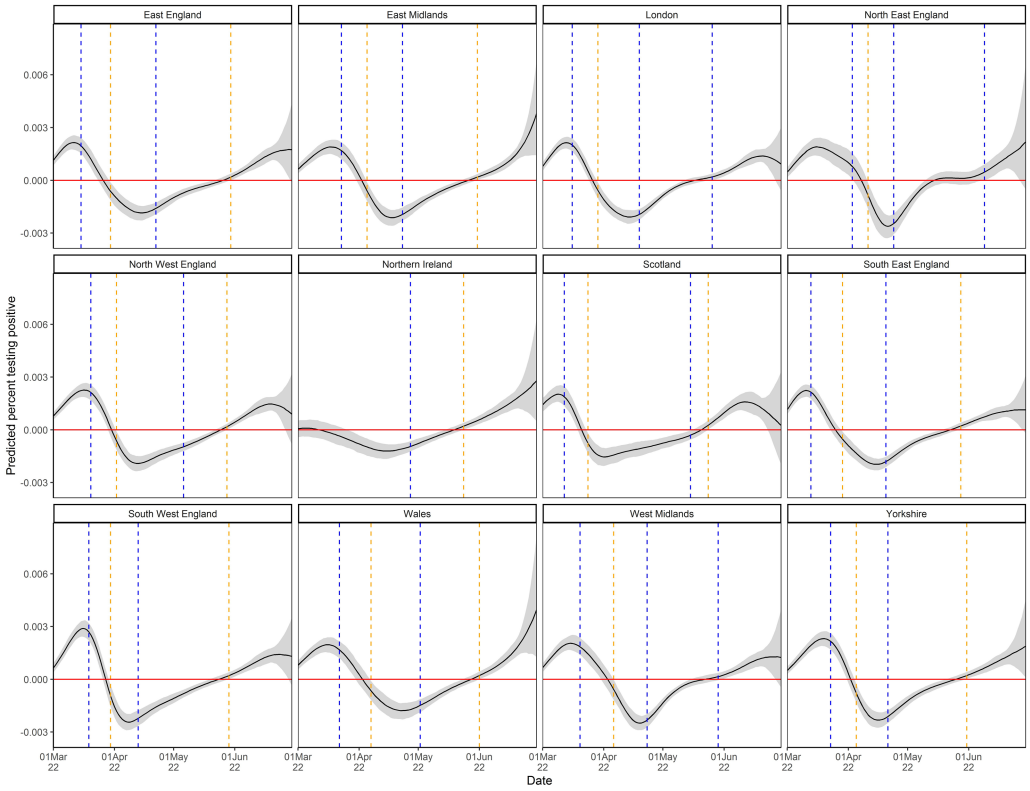
| Geographical region | GAM change-point date calculated using second derivative only (DD.MM.YYYY) | GAM change-point date incorporating additional change-points from first derivative (DD.MM.YYYY) | ISR change-point date (DD.MM.YYYY) |
|---|---|---|---|
| **East England** | Not found | 30.05.2022 | 23.05.2022 |
| **East Midlands** | Not found | 31.05.2022 | 01.06.2022 |
| **London** | 26.05.2022 | No change | 23.05.2022 |
| **North East** | 09.06.2022 | No change | 11.05.2022 |
| **Northern Ireland** | Not found | 24.05.2022 | 14.05.2022 |
| **North West** | Not found | 28.05.2022 | 01.06.2022 |
| **Scotland** | Not found | 24.05.2022 | 29.05.2022 |
| **South East** | Not found | 28.05.2022 | 20.05.2022 |
| **South West** | Not found | 29.05.2022 | 23.05.2022 |
| **Wales** | Not found | 01.06.2022 | 17.05.2022 |
| **West Midlands** | 29.05.2022 | No change | 01.06.2022 |
| **Yorkshire** | Not found | 31.05.2022 | 23.05.2022 |

**Figure 3.11: Predicted positivity (A), first derivatives (B), and second derivatives (C) calculated from GAMs fitted on the entire time-series for each geographical region, but only presented from 1st March to 30th June 2020. Original change-points based on the second derivative are shown in vertical blue lines, and additional change-points based on the first derivative are shown in vertical orange lines.**

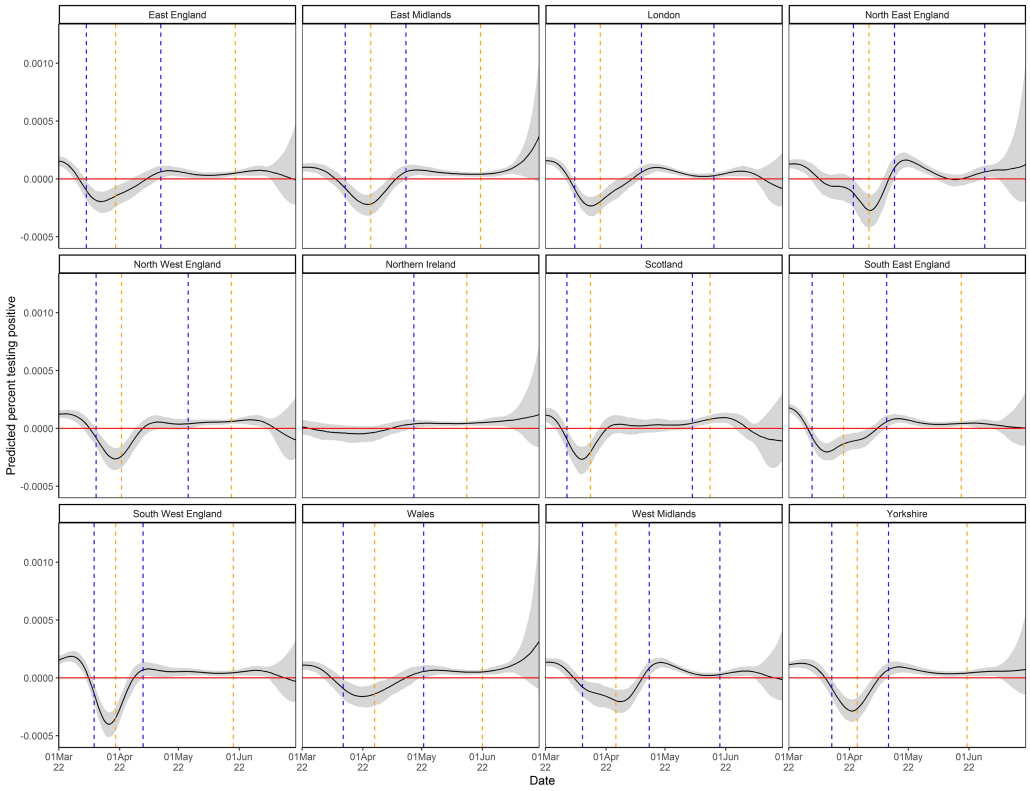
**A: Predicted positivity**



## B: First derivative




## C: Second derivative



### 3.3.5 Estimating change-points in target subgroups

Analogous to "sentinel surveillance", I assessed whether change-points could be established earlier by modelling population subgroups, here focusing on age as schools remained open for much of the pandemic after the initial lockdowns. In the dataset used for this analysis, as others,[149] large rises in positivity associated with the emergence of Alpha occurred earlier in those aged 2y-11sy, with steeper increases in positivity in late-August 2021 (Delta) and late-January 2022 (BA.1) compared with older age-groups (**Figure 3.12**).

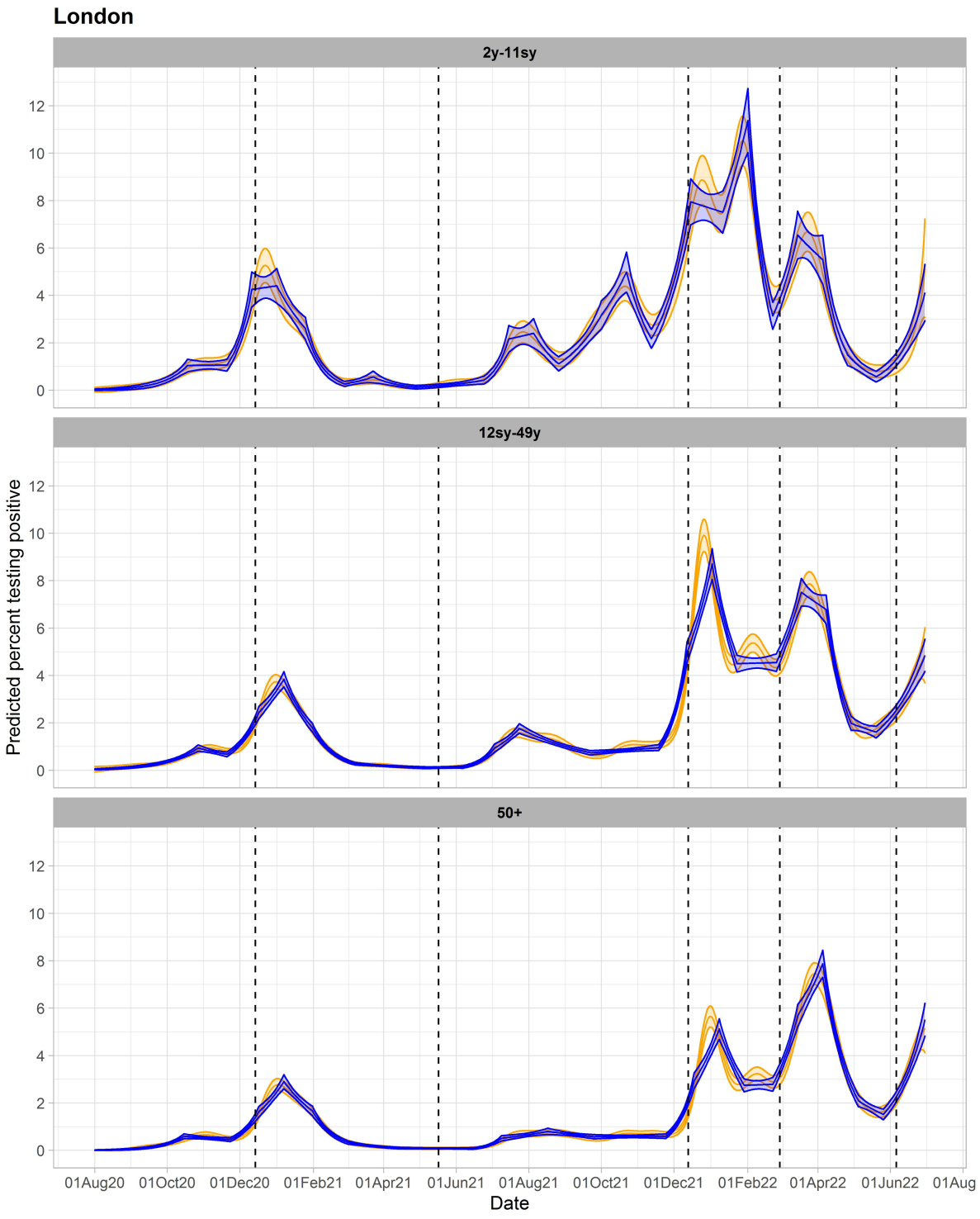
Little difference was seen across age groups for GAM change-points associated with Alpha (**Figure 3.13**). For Delta, change-points occurred earliest in the overall model and those aged 12sy-49y, and latest in those aged 2y-11sy. Rises in BA.1 occurred 18 days earlier in the youngest age group versus all ages using ISR, and 19 days earlier using GAMs. Rises in BA.2 were found earliest in the 2y-11sy age group using GAMs (9 February 2022).

### 3.3.6 Estimating change-points by type of outcome

Analogous to surveillance of different types of infection, e.g. resistant vs susceptible *Staphylococcus aureus*, I considered whether change-points could be established earlier by modelling PCR S-gene positivity as a proxy for SARS-CoV-2 variant. There were distinct differences in trend between SGTF and SGTP positivity over time (**Figure 3.14**) and GAM and ISR predictions for London closely followed these trends (**Figure 3.15**).

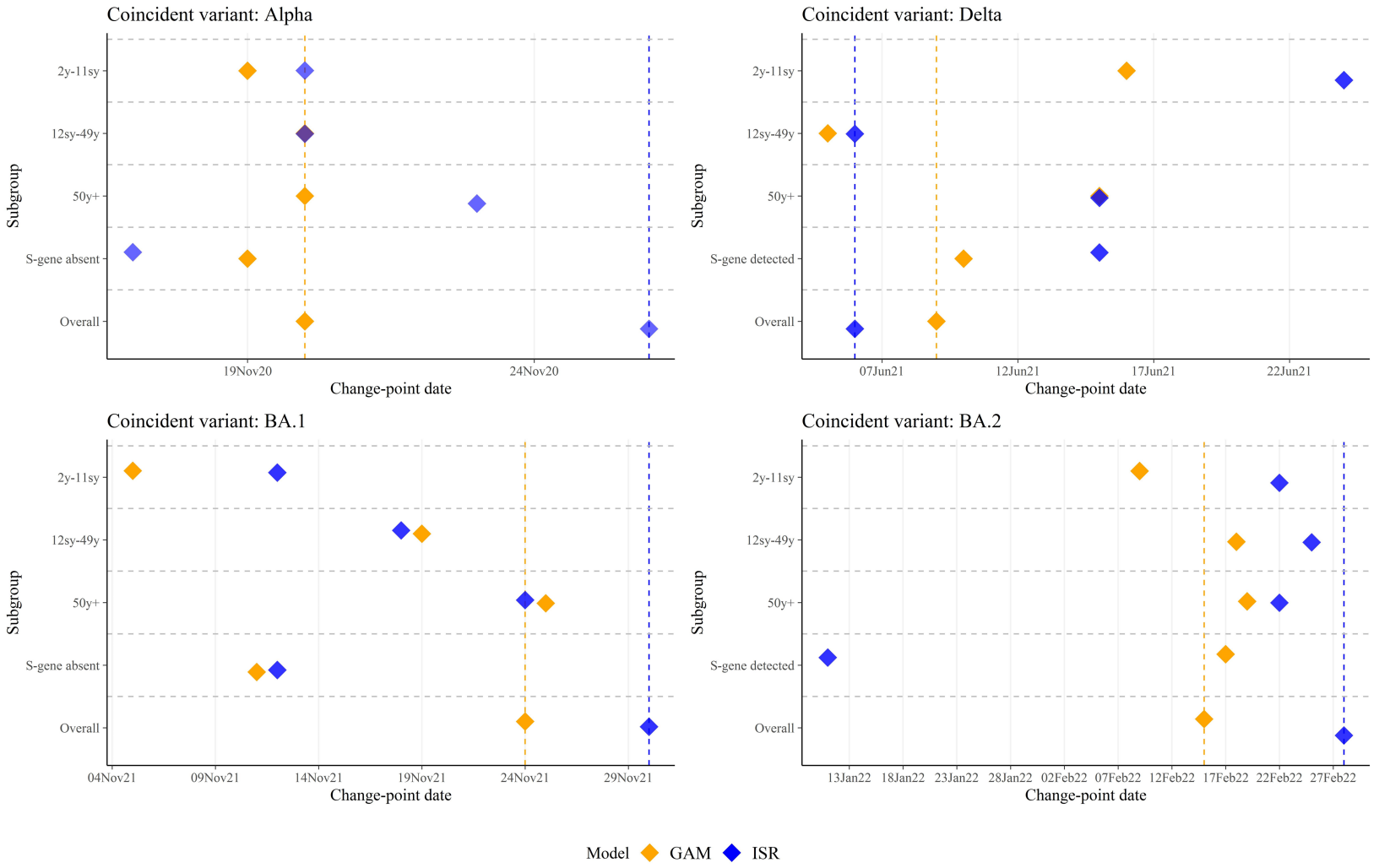
For London, change-points associated with the emergence of Alpha occurred one and nine days earlier using SGTF versus all positives for GAMs and ISR, respectively (**Figure 3.13**). Change-points for Delta occurred and were detected nine days later using SGTP versus all positives using ISR, and occurred one day later using GAMs. Change-points for BA.1 occurred on 11 and 12 November 2021 using GAMs and ISR for SGTF, with all-positive change-points on 24 and 30 November 2021, respectively, a 15-day earlier detection for the ISR change-point. For SGTP, ISR estimated a change-point for BA.2 on 11 January 2022 (detected 4 February 2022) and did not find a change-point for all-positives until 48 days later.

**Figure 3.12: Predicted percentage testing positive for SARS-CoV-2 from ISR (blue) and GAMs (orange) for models run separately by age group for London.**



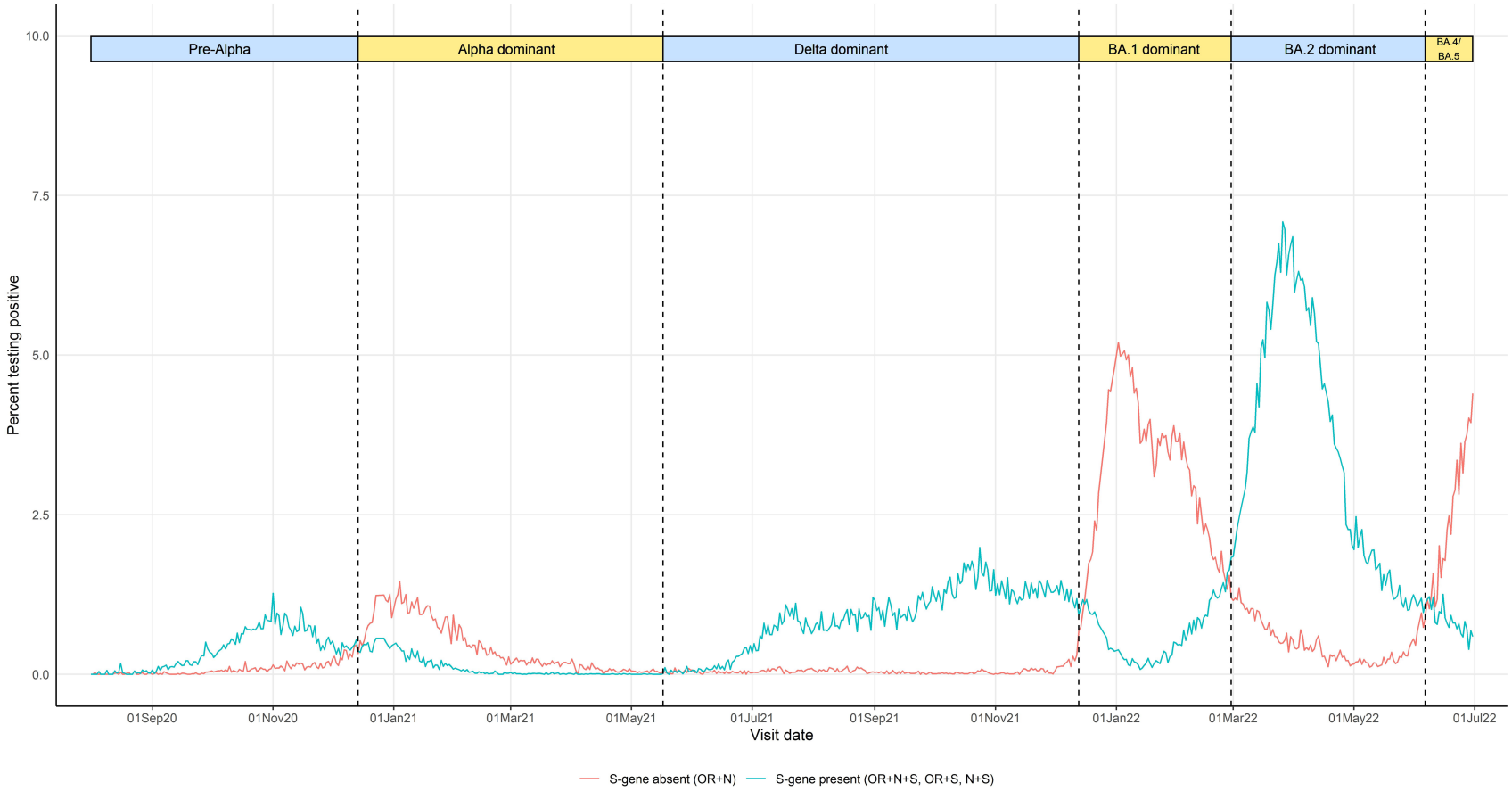

Note: Vertical dashed lines and grey shaded areas are defined in **Figure 3.4**.

**Figure 3.13: Change-points from GAMs (run on the full time-series; orange) and ISR (blue) run separately by age, separately by S gene detection, and overall in London.**



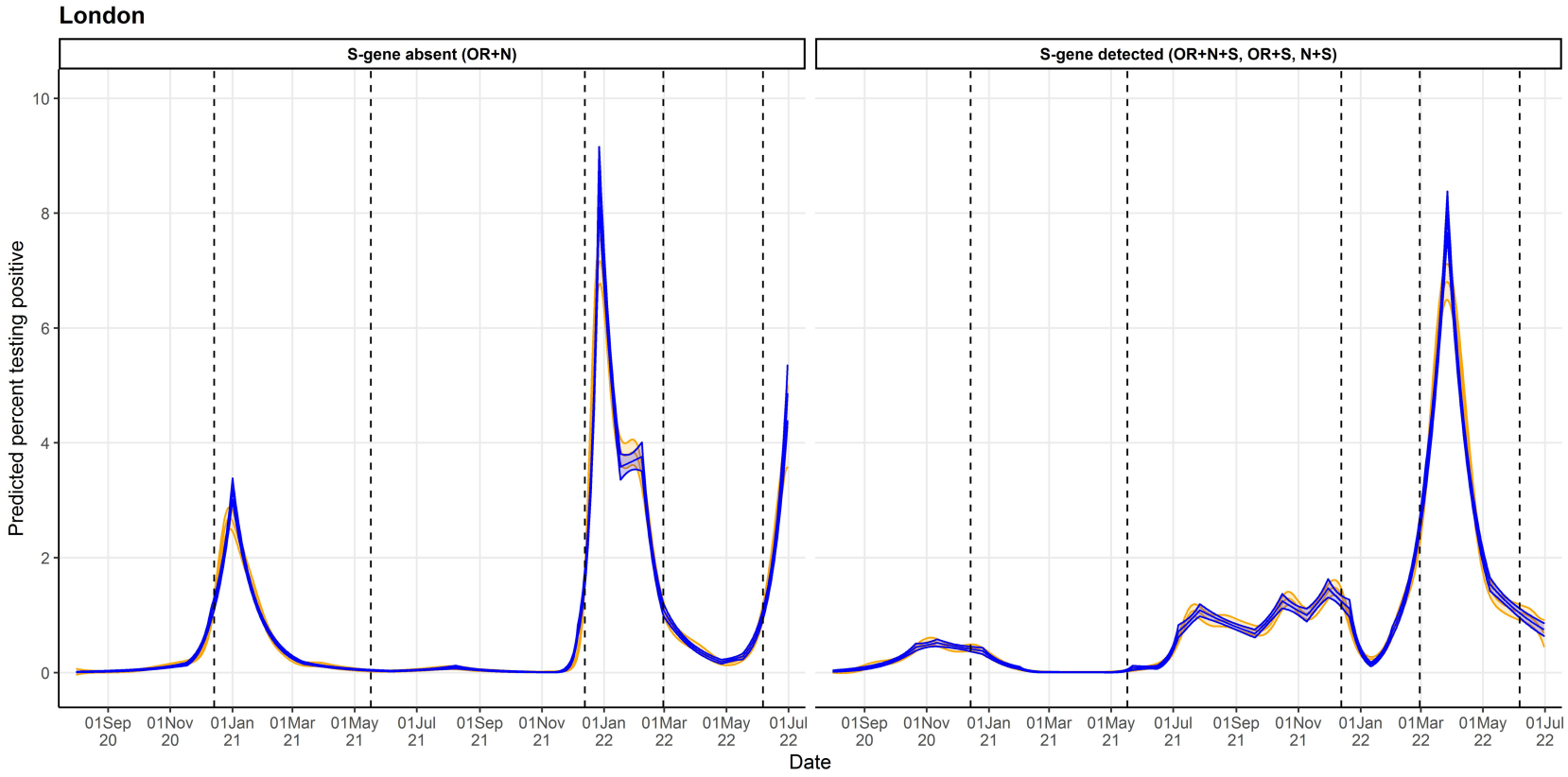

Note: Vertical dashed lines show the position of change-points in overall GAM and ISR.

**Figure 3.14: Raw daily percentage testing positive split by S-gene target positive and S-gene target failure, for all regions.**




Note: Vertical dashed lines are defined in **Figure 3.4**.

**Figure 3.15: Predicted daily percentage of visits testing positive for SARS-CoV-2 from ISR (blue) and GAMs (orange) for London only, split by SGTP and SGTF.**




Note: Vertical dashed lines are defined in **Figure 3.4**.

## 3.4 Discussion

Here, I compared two methods for detecting changes in growth rates in surveillance data, using SARS-CoV-2 as an exemplar. Both methods detected trend increases and decreases associated with the emergence of Alpha, Delta, BA.1, and BA.2 variants, and other smaller growth rate fluctuations, at similar dates. Considering near real-time analysis, the most recent change-points detected using GAMs were found in successive GAMs including five subsequent weeks' data and using ISR, demonstrating consistency between GAM model runs and between methods. However, GAMs needed at least 4 times the duration of data over which there was interest in identifying change-points to provide stable estimates. Change-points were, on average, detected slightly earlier using GAMs versus ISR considering larger geographical regions, but this was not consistent across different sized regions or subgroups. Considering positivity trends separately in different age subgroups allowed earlier BA.1 detection using data from children alone versus all positives, and similarly considering positivity trends separately for SGTP allowed earlier detection of BA.2 emergence.

Supporting using these methods for sentinel surveillance, I found change-points could be detected earlier through modelling age groups separately, often, but not always, in those 2y-11sy. This is likely due to faster SARS-CoV-2 transmission at younger ages, specifically during Omicron BA.1 emergence, driven by higher contact levels through school attendance while other guidance (e.g. working from home[150]) slowed transmission between adults until infections were later transmitted onwards from younger to older individuals. While studies have found no evidence of increased transmission on school premises,[151-154] increased person-to-person contact associated with attending school (e.g. public transport, gatherings at school pick-up/drop-off) measurably impact the reproduction number.[155] Rising SARS-CoV-2 positivity in younger individuals may therefore be a useful early warning signal for rises in older individuals, where hospitalisation risk and mortality are higher,[156] although trend changes were not consistently identified earlier in younger age groups. Implementing surveillance systems separately by subgroups may be an efficient way to detect changes earlier more generally, although its value may depend on specifics of transmission dynamics.

Similarly, I found change-points were generally estimated to occur slightly earlier when considering positivity split by S-gene detection. This was particularly useful when BA.2 emerged, as BA.1 declines concealed fast BA.2 growth when combining all positives. However, again this was not universal, likely reflecting an accuracy-precision trade-off, since the number of positives of specific types was always lower, even if they were a more sensitive measure of the new variant. More broadly, surveillance of pathogens with different susceptibilities could allow similar shifts in underlying variants to be elucidated.[157]

The methods I considered in this chapter have much wider applicability to infection surveillance, but SARS-CoV-2 provided an ideal opportunity to test them due to rapid changes in positivity and emerging variants with different epidemiology. Change-points estimated for Alpha using ISR and GAMs were generally consistent with changing UK public health policy. The first Alpha sequence came from a sample on 20 September 2020, but it was not widely recognised until December after its rapid growth throughout November,[158,159] with regional lockdowns implemented on 23rd December 2020.[160] By then, change-points had been detected by ISR in four geographical regions, including London and South East England where Alpha rose earliest and fastest. In contrast, Delta was named a variant of concern on 6 May 2021,[161] approximately a month earlier than change-points occurred in most regions in this analysis, reflecting its earlier identification through rapid increases in infections in India. The appropriate length of time between changes occurring and change-point detection thus depends on the surrounding context. The average 3-week lag observed here may not be generalisable under large-scale specific testing but may be relevant if surveillance is reliant on passive data. The infection/disease being monitored could also influence the relevance of this lag. Fundamentally ISR and GAMs identify when infection epidemiology changes, which may be independent of or coincident with recognition of new variants with different transmission potential, virulence, or resistance through genetic sequencing or changes in epidemiology in other countries. Whilst the methods presented here could be applied to the proportion of genetic sequences which are a specific variant, to date this has generally shown log-linear growth for SARS-CoV-2,[162] without change-points before a new variant becomes the majority sequence.

Real-time surveillance is mostly concerned with recent data, where uncertainty is greatest. Using ISR, most change-points were detected slightly later versus using GAMs – at least in part a limitation of ISR requiring a minimum number of days between the current and last identified change-point. While GAMs detected some change-points earlier, they also found a small number of change-points during the last 7-days of successive model runs which were not confirmed when adding a further 7-days' data. The increased flexibility afforded through GAMs may therefore cause false-positives at data boundaries. Most change-points found in the last 4-weeks of successive GAMs were confirmed by the full time-series GAMs, albeit sometimes later due to reduced power over shorter timeframes. Requiring at least 7-days of data after change-points, or confirmation in two successive models, would increase certainty. Change-points were found slightly later in GAMs versus ISR on smaller datasets, likely due to ISR fixing one parameter at a time whereas GAMs optimise over the entire time-series, hence being more influenced by sample size. Overall, this illustrates the inherent trade-offs between the two methods; ISR forcing log-linear trends between change-points will identify

change-points more efficiently when this is close to the truth but will be inefficient if trends are volatile.

While most change-points found in this study were related to increases and decreases in major variants, I also identified other fluctuations. Most of these other change-points were associated with large relative percentage changes in positivity over 4 weeks, with most smaller relative changes indicating growth/decline slowing down or flattening off, so these were still epidemiologically important. Some change-points identified by GAMs were significant only for short durations e.g. one day, and of small magnitude in the second derivative. Whilst statistically significant, these changes may not be meaningful, with policy decisions more likely made on larger changes in growth/decay. For both GAMs and ISR, I would recommend interpreting change-points in the context of current underlying prevalence. Further, using second derivatives of GAMs, change-points for BA.4/BA.5 were mostly not found by the 30th June 2022 but could be established when considering additional change-points based on the first derivative swapping sign. Considering changes in the first derivative may be important to avoid missing change-points moving forward. As regards methods, estimating derivatives using a Metropolis-Hastings or similar sampler is recommended during low prevalence periods.

In this exemplar, demonstrating that relevant change-points can be detected in a randomly sampled community population is useful for future SARS-CoV-2 surveillance, as this could trigger targeted testing in different regions and/or age groups to help control spread and identify new variants,[163,164] ultimately aiming to reduce cases/hospitalisations. The large sample size allowed power to detect change-points, despite relatively low positivity rates, enabling me to compare the two methods. Whilst SARS-CoV-2 is a respiratory virus, there is no reason that the methods would not apply more broadly to different infection surveillance data streams.

Study limitations include comparing the two methods on a single dataset, albeit including multiple change-points of different magnitudes. While ISR has been evaluated independently on other datasets,[139] further comparisons in other settings may be useful. Comparing methods on complex real-world data is practically useful, but future simulation studies could systematically evaluate statistical properties of these methods against a known “gold standard”, albeit likely based on much simpler underlying trends. I arbitrarily decided to define change-points within ±7-days between methods as matched (i.e. identifying the same true underlying epidemiological trend) which may have led to a small amount of misclassification. The amount of data required to detect change-points will depend on the specific outcome, and speed of underlying changes, which will differ between respiratory pathogens (e.g. SARS-CoV-2) and antimicrobial resistance determinants, for

example. While ISR and second derivatives of GAMs are two options, other change-point detection methods may also be suitable, although I was not able to identify other methods that aimed to identify trend changes on an initial literature search (also confirmed through 2024 conversation between TP Quan and author; unreferenced).

In summary, ISR and second derivatives of GAMs could potentially detect changes in trend in multiple different types of infections in near real-time surveillance, including SARS-CoV-2, but more widely including hospital-acquired infections and antimicrobial-resistant pathogens. While both methods gave a generally consistent pattern, some known changes in the epidemiology of SARS-CoV-2 caused by different variants emerging were identified earlier by GAMs than by ISR and vice-versa, therefore running both methods in parallel would be ideal.

# Chapter 4 Choosing control groups and defining exposures in studies using Electronic Health Records

The unprecedented nature of the COVID-19 pandemic resulted in scientific research focusing on how to better understand, manage, and treat COVID-19. Chapters 2 and 3 of this thesis focused on some key questions about the COVID-19 response both relating to how we can identify those at risk of disease in "near real-time" scenarios. I wanted to explore whether I could extend the methods I used during the COVID-19 pandemic to investigate those at risk of different diseases in varying data sources. This Chapter explores the challenges encountered when applying these methods in different contexts.

## 4.1 Introduction

Electronic Health Records (EHRs) offer a wealth of information to explore at-risk populations. While research is not their primary purpose, EHRs have been increasingly used for research purposes. The real-world nature of the data can reduce selection bias compared with other study types such as clinical trials, allow large sample sizes which is particularly useful for less common diseases, and reduce costs due to data already being collected for routine purposes.[165] The use of EHRs has increased massively over the last 30 years, as outlined in the Introduction to this thesis, resulting in 90% of NHS Trusts in England using EHRs in 2023.[24] There has also been increased linkage between microbiology samples and hospital episode statistics, for example in national level data held by UKHSA.[76] This allows more accurate data on laboratory-confirmed infections and gives more opportunity to assess the impact of a more diverse and higher number of risk factors for these infections. In particular, using linked data may allow better monitoring of risk factor associations with the most serious type of infection, bloodstream infections (BSIs), over time.

To investigate risk factors using EHRs, a control group had to be established so risk factor differences could be compared between cases and a population that did not get the disease of interest (controls). The COVID-19 Infection Survey (CIS) used in Chapters 2 and 3 was a prospective cohort study. These studies work by choosing a group of people before the event of interest is observed and following them up over time, measuring exposures and observing whether they have the event of interest. Those who do have the event become "cases" while those who do not become "controls" and therefore the group is not conditioned on the outcome. This contrasts with case-control studies where individuals are selected specifically for the study based on the outcome of interest and defined as either cases or controls.[166] Potential risk factors are then observed retrospectively from the pre-collected data. When using EHR data, a case-control design was

suitable as the data were already collected; however, this did require defining both cases and controls at the beginning of the study.

Who should be included in a control group when using EHRs is unclear. There are general recommendations for all study types in the literature about conditions individuals should meet to be considered in a control group. For example, controls should have been included as cases if they had developed the disease (i.e. the control group should come to the same hospital as the included cases if cases are defined by hospitalisation at a specific location), and they should have the same source population as cases.[167] The choice of control group can impact results and yield incorrect results if not carefully selected. Collider bias can be particularly problematic in studies using EHRs due to the probability of being selected into the study (i.e. healthcare attendance) often being influenced by the exposure and the outcome, possibly inducing incorrect associations (i.e. estimated associations do not have a causal interpretation).[168] This has been well documented in hospital-based studies for COVID-19 where, for example, both frailty and COVID-19 infection predict hospitalisation. Considering frailty as a risk factor could therefore induce collider bias.[168] Further, missing data for risk factors can influence results when choosing a control group. Many risk factors are calculable in EHRs from International Classification of Disease 10th edition (ICD-10) codes which are predominately recorded in inpatient admissions. An absent ICD-10 code may not reflect absent disease but instead the absence of an inpatient admission where the disease could be recorded. The amount of missing data could be impacted by selection of a control population. Finally, while one benefit of EHRs is the large number of people included in them, information governance laws may be in place to restrict access to all individuals with relevant records and therefore this has to be considered in control group choice.

There is no gold standard for defining risk factors using EHRs and, as described above, bias can be introduced. The proximity of risk factors to the outcome of interest is important and can cause reverse causality if measurements taken very close to the outcome are included. For example, increases in C-reactive protein (CRP) are a consequence of BSIs so CRP measurements taken just before a positive blood culture collection could be caused by the BSI, rather than affect the risk of getting a subsequent BSI through e.g. increased background inflammation since CRP could also be raised when diseases are in a sub-clinical state.[169] The location where samples were collected may also impact the test results and therefore the suitability of values for inclusion in risk-based analyses. For example, a heart rate measurement taken at a planned outpatient appointment may be closer to what is physiologically normal for a person versus a heart rate measurement taken at the emergency department from the same person. Further, the amount of missingness in risk factors may be dependent on the location of the patient. The type of laboratory blood tests requested can also

differ in inpatient admissions compared with requested tests in ambulatory, outpatient and emergency department settings, for example,[170] perhaps resulting in different missingness patterns across different healthcare settings.

Finally, EHRs are imperfect for use in research. EHRs are primarily collected for clinical care and reimbursement purposes and can have imperfect data input, impacting the quality of risk factors calculated. For example, ICD-10 codes reflect the ultimate most serious (in terms of cost) final diagnosis rather than the initial diagnosis the patient was treated for when the disease was first diagnosed, or even other diagnoses made during admission.[171] Abnormal test results will only be present if the test was conducted and considered relevant to the main condition the patient was admitted for.[172] There is unlikely to be one correct definition for exposures, and instead multiple ways to represent the same risk factors could be considered and give useful information.

*Escherichia coli* (*E. coli*) BSIs were selected as the disease of interest for this study. They are the most common Gram-negative bloodstream infection in the UK and can lead to serious illness and death.[61] Understanding which populations are at an increased risk of *E. coli* BSIs could help target interventions to reduce the number of *E. coli* BSIs. Further motivation for choosing *E. coli* BSIs as the disease of interest will be outlined in the following Chapter.

Using data from the Infections in Oxfordshire Research Database (IORD), I first compared the impact of different control groups, varied by the time since previous hospital contact and type of hospital contact, on missing data and results from models including demographics only. Second, I assessed the impact of differing definitions of risk factors on model results, specifically the proximity of risk factors to *E. coli* BSI collection as well as where in the hospital measurements were taken.

## 4.2 Methods

I used data from the Infections in Oxfordshire Research Database (IORD): a large dataset including inpatient admissions, outpatient appointments, and accident and emergency (A&E) visits, as well as microbiology and biochemistry/haematology test results from samples sent within hospital and from general practice. These were linked with diagnostic codes, procedure codes, and vital signs taken during hospital attendance. IORD includes four large teaching hospitals covering a catchment area of around 660,000 individuals. The dataset goes back to 1997, with electronic recording of most variables, for example, diagnostic codes, from 2007 onwards. Electronic recording of vital signs began in 2016. I included all admissions from 1st April 2018 to 31st March 2022 to allow for a lookback period for vital signs. I analysed data separately over financial years (FY, i.e. April to March) to keep winter months together in the same year.

### 4.2.1 Case definition

I defined cases using results from microbiological isolations. All *E. coli* cultured from blood samples were considered as cases. Positive isolations were de-duplicated within 90 days from an index positive to avoid multiple inclusions reflecting the same infection episode or incompletely treated ongoing BSIs. Of note, this is different from the surveillance definition which deduplicates using a 28-day window as I only wanted to consider *de novo*, rather than recurrent, infections in this study. A small number of cultures from individuals <16 years old and those missing sex were excluded. Those <16 years old were excluded as risk factors may differ in children versus adults and infections in babies may be influenced by characteristics of the mother that, due to deidentified data, were not definable using EHRs. For each financial year, only one positive blood culture was considered per person, selecting the first *E. coli* BSI if multiple de-duplicated positive blood cultures were present. Recurrent *E. coli* BSIs were included but, as per the definition above, only if BSIs were >90 days apart and in separate FYs.

### 4.2.2 Control definitions

I defined controls using all individuals with any of an inpatient admission, outpatient appointment, A&E attendance, microbiology sample taken, or blood test recorded in the Laboratory Information Management System (LIMS). Importantly, the microbiology results contained both positive and negative results from samples and results from both patients who had samples taken in the hospital and people in the community who had a sample sent for testing from the GP to the hospital-based laboratories. Records from individuals <16 years old and those missing sex were dropped from the analysis. All records from individuals who had had an *E. coli* BSI since 1st April 2013 were removed to avoid contamination of the control group. I included each person once per FY if individuals had multiple healthcare visits, selecting the last available contact within the FY.

### 4.2.3 Defining the analysis cohorts

Ideally, the control group would include everyone in Oxfordshire; however, IORD only includes a subset of individuals who have had healthcare contact (as defined above) with Oxford University Hospitals. Further, many risk factors of interest can only be calculated from data recorded in inpatient admissions, for example, diagnosis or procedure codes and vital signs. To account for the fact that individuals with no previous or current inpatient episode would have missing data for many characteristics, two analysis cohorts were subsequently defined based on previous and current healthcare contact recorded in IORD. (Note: Inpatient admissions, or an inpatient spell, are a continuous hospital stay in the same hospital from admission to discharge. A spell is split into one, or more, contiguous consultant episodes, with a new episode generated when the patient is transferred to a different consultant or specialty.[173] Diagnosis and procedure codes are assigned to each episode.) Within each of these cohorts, five subsets were initially created based on a varying amount of "lookback" included to calculate risk factors, from 1FY to 5FYs.

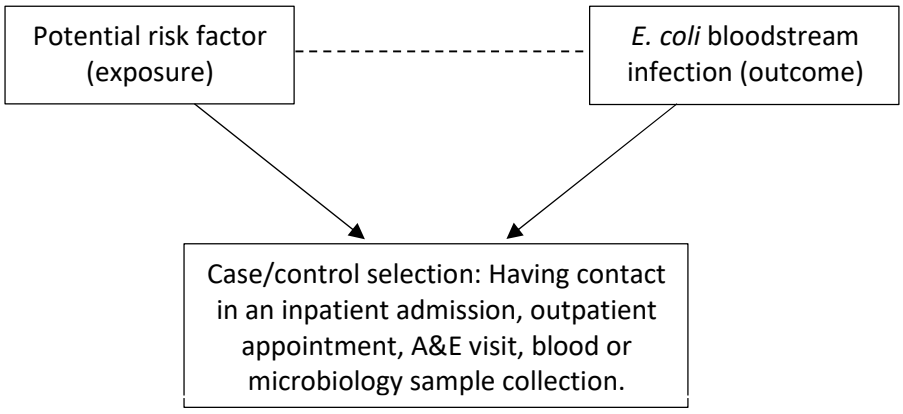
One cohort was defined based on previous inpatient contact, subsequently referred to as the "**inpatient only**" cohort, to minimise the amount of missing data for risk factors.

- The inclusion of *E. coli* BSIs (cases) into this cohort was conditional on them having a previous inpatient episode. A previous inpatient episode was defined as any inpatient episode in the previous 1-5 FYs which ended (based on episode end date and time) at least 72 hours before the collection date/time of the blood sample defining the *E. coli* BSI (their most recent contact date).
    - The exclusion of inpatient exposure in the 72 hours before blood culture collection was to avoid issues of reverse causality as many *E. coli* BSIs result in an inpatient admission; factors recorded in these episodes could therefore be a consequence, rather than a cause, of the infection. Inpatient episodes were used rather than inpatient spells to allow for nosocomial BSIs to have risk factors based on current hospital exposure, provided the episodes ended >72 hours before blood culture collection.
- The inclusion of potential controls into this cohort was conditional on having either a current inpatient episode or a previous inpatient episode within the previous 1-5 FYs which ended on or before the "most recent contact" date (the last date in each FY where individuals did not have an *E. coli* BSI). The 72-hour exclusion was therefore not applied to the control group because there were no concerns about reverse causality; however collider bias was a concern (see below).

The second cohort was defined conditional on having any previous healthcare contact, subsequently referred to as the "**any healthcare**" cohort, to allow the retention of as many individuals as possible for analysis.

- The inclusion of *E. coli* BSIs (cases) into this cohort was conditional on them having a previous inpatient episode, as defined above, or an outpatient appointment, A&E visit, blood test, or microbiology sample which occurred >72 hours before the collection date/time of the *E. coli* BSI and within the previous 1-5 FYs. Again, this was to avoid issues of reverse causality.
- The inclusion of potential controls into the "any healthcare" cohort was conditional on having a current or previous inpatient episode in the prior 1-5 FYs as defined above, or a previous outpatient appointment, A&E visit, blood test, or microbiology sample strictly before the "most recent contact" date. Non-inpatient contact was specified to be strictly before the "most recent contact" date (rather than including current contact) to reduce the impact of collider bias in this cohort by reducing the number of individuals being selected into the cohort based on their current health status. This is demonstrated below where both potential risk factors (e.g. frailty) and *E. coli* BSIs impact case/control selection, creating spurious relationships between exposures and the outcome (**Figure 4.1**).

**Figure 4.1: A simplified Directed Acyclic Graph (DAG) illustrating potential collider bias.**




Note: Directed arrows indicate causal effects and dotted lines indicate induced associations. Other variables and confounders are excluded for illustrative purposes only.

### 4.2.4 Defining exposures

Risk factors were defined from all available data sources, specifically inpatient admissions and outpatient appointments including ICD-10 codes and procedure codes (Office of Population Censuses and Surveys [OPCS] Classification of Interventions and Procedures version 4), A&E

attendances (which do not have these codes but a different coding system for recording reasons for attendance which is not aligned with ICD-10 and therefore I did not use), blood test results, microbiology results, and vital signs. A full list of variables with their definitions is provided in **Appendix A**. Details on how variables were selected are outlined below, with variables broadly being selected based on risk factors defined in previously published research (e.g.[62,174]), clinical advice and knowledge, and the availability of data.

For ICD-10 and procedure codes, I exploited the hierarchical nature of the coding systems to aggregate information to create variables.[175,176] At the highest level, both types of codes are split into Chapters broadly encompassing the diagnosis or type of procedure a patient has received. For example, ICD-10 Chapter I includes codes related to "certain infectious and parasitic diseases", while Chapter A for procedure codes encompasses procedures related to the nervous system. I therefore created 22 variables denoting all ICD-10 code chapters and 24 variables for all procedure code chapters.

I also aggregated ICD-10 codes based on pre-defined scores. I calculated the Charlson co-morbidity index[177] using predefined ICD-10 codes[178] with updated weightings from Quan et al.[179]. I considered the composite score as a continuous variable, but also the presence of the individual components separately as binary variables. I defined frailty using a pre-defined frailty score for use with EHRs by Gilbert et al.[180]

I used published coding lists to define additional variables from procedure codes for chemotherapy,[181] dialysis,[182] and transplant.[183] Eighteen specific surgery groupings (e.g. including large bowel surgery and hip replacement) were defined from previously published UKHSA documentation.[184] This document also flagged whether each surgery within each grouping was mostly clean, clean-contaminated, or contaminated, and hence three additional groupings of surgeries under these headings were considered as potential risk factors. Some variables were defined based on key phrases in procedure code descriptions. This was done for suspected risk factors, such as urine catheterisation, or common procedures, such as endoscopies, where there was not an available pre-defined coding list.

I created variables from microbiological samples considering both whether a test was requested and the test result. Variables were created for pathogens commonly isolated from urine samples: urine positive for *E. coli*, *Enterococcus*, *Klebsiella*, and Enterobacterales (including *E. coli* and *Klebsiella*). Similarly, for blood samples, any blood culture positive for *S. aureus* was considered as a binary variable with all other isolated pathogens occurring very infrequently in the *E. coli* cases and therefore not included. Any urine or blood sample collected for culture were also created as

variables alongside the number of both these tests requested irrespective of results as, based on previous research looking at blood tests, requesting a test can be as predictive of ill-health as the positive test/test result itself.[172] I also considered any of the following tests requested as variables: faeces culture (and separately those positive for *Clostridioides difficile*), COVID-19 swabs (and those positive) and influenza/Respiratory Syncytial Virus (RSV) PCRs (polymerase chain reaction tests), respiratory samples (and those positive for *Pseudomonas aeruginosa*), surface swab culture taken, screen for MRSA (Methicillin-resistant Staphylococcus aureus), blood PCR test for CMV (Cytomegalovirus) or EBV (Epstein–Barr virus), or CNS (central nervous system) culture taken.

I selected blood test results from twenty tests based on advice from clinicians to capture a variety of clinical information. I selected routine tests such as those done as part of a full blood count, including white cell count, platelets, and haemoglobin levels, to assess general health indicators.[185] I also included tests which were markers for risk factors of interest, including CRP as it is a marker for inflammation, alanine aminotransferase, albumin, and bilirubin for liver function, and creatinine and urea to indicate kidney function. Other tests were selected as they are used widely to screen for specific illnesses, including HbA1c measurements for diabetes/pre-diabetes and the Prostate-Specific Antigen test (PSA) for prostate cancer. I created variables using test results from six key vital signs: heart rate, respiratory rate, systolic blood pressure (SBP), diastolic blood pressure (DBP), oxygen saturation, and temperature. These vital sign tests were measured routinely in inpatient admissions.

For all vital signs and blood test results, extreme values incompatible with life (based on clinician recommendation) were dropped from the analysis. To avoid undue influence of marked outliers, I truncated all continuous variables at the 5th and 95th percentiles for variables that did not include a natural zero (BMI, height, systolic and diastolic blood pressure, heart rate, respiratory rate, temperature, Charlson score, albumin, creatinine, haemoglobin), and at the 0th and 95th percentiles otherwise (eosinophils, neutrophils). I truncated oxygen saturation at the 5th and 100th percentiles since an oxygen saturation of 100% is within the normal physiological range.

I matched all blood test results and vital signs to their closed inpatient admission, outpatient appointment, or A&E attendance. If measurements did not fall strictly within any of these, measurements were matched to admissions/attendances within ±72 hours of their collection. A small proportion of vital signs (<1%) were not strictly within, or within ±72 hours, of an inpatient admission, outpatient appointment, or A&E visit. These measurements were subsequently matched to their closest healthcare attendance. A larger proportion of blood test results did not fall strictly within, or within ±72 hours of, a hospital attendance but these were expected as blood tests outside

of the hospital setting (i.e. those from samples taken at GPs) were tested within the hospital and hence recorded in the EHR. These measurements were therefore left unmatched to healthcare attendances.

### 4.2.5 Statistical analyses

Summary statistics were presented as medians with interquartile ranges for continuous variables and proportions for categorical variables.

The outcome of all models was the presence of *E. coli* BSI (case) versus absence (control) as a binary outcome. Poisson models with cluster robust standard errors were used for all models with incidence rate ratios reported – rates were very low so this is equivalent to a logistic regression but allows a more natural presentation of risk rather than odds. Explanatory variables chosen for inclusion in the models were "core variables" selected as key confounders for risk of *E. coli* BSIs and likely to be stable over time. These were:

- Age (calculated at last contact in exact values compared with the month and year of birth and included as a restricted cubic spline with one internal knot and two boundary knots).
- Sex (binary, male and female)
- Ethnicity (binary, white and non-white due to small numbers in the latter group)
- IMD percentile (continuous, tested for non-linearity)
- Rural/urban classification (included as a three-level factor: urban city/town, town/fringe, rural village)
- Catchment percentage (percentage of individuals in local area visiting an Oxfordshire hospital within Oxford University Hospitals NHS Foundation Trust versus another Trust as defined by the Office of Health Disparities; 0 = none, 100 = all).[186]

If the core variables were missing at the "most recent contact" date, the closest recorded measurement was used from either past or future values. Age was truncated at the 95th percentile of the unique value distribution. Restricted cubic splines with between one to five internal knots were considered for all continuous core variables, with a minimum of one internal knot included for age due to expected variation and linear effects allowed for deprivation and catchment percentage. Internal knots were placed at even percentiles throughout the range of values for each variable (after truncation at the 5th and 95th percentiles) with boundary knots included at the 10th and 90th percentiles of the range of values for each variable. Incidence-rate ratios with 95% confidence intervals were compared between all the cohorts across all lookback periods for all study years to assess the impact of the cohort on model estimates.

After selecting a cohort ("inpatient only" or "any healthcare"), I investigated the impact of both the timing and location of measurements on the values of all the other variables defined above, denoted "screening variables" following Chapter 2. For all the screening variables, I explored the timing of measurements relative to the "most recent contact date" for cases and controls. To investigate issues of reverse causality, I investigated removing observations 72 hours before the "most recent contact" date on the values included for both cases and controls. For blood test results and vital signs, I also investigated the timing of measurements within inpatient admissions on the blood test results compared with normal ranges. I also investigated the impact of the location of where the tests were taken on the values selected for cases and controls relative to the "most recent contact" date. The blood test results available in the dataset included tests from inpatient admissions, outpatient appointments, A&E visits, and those sent from GPs in Oxfordshire. I summarised the differences in values across these four locations using medians (IQRs) and histograms. For vital signs, I summarised the differences in values across inpatient admissions, outpatient appointments, and A&E visits only, since values taken at GPs are not available in IORD.

### 4.2.6 Sensitivity analyses

Large amounts of missing data can influence the results of a complete case analysis. To assess the impact of missing ethnicity, results from multiple imputation were compared with complete-case estimates. Multiple imputing using chained equations (MICE) was used to impute ethnicity, using Poisson regression and including age (one-knot restricted cubic spline), sex, rural/urban classification, IMD percentile, catchment percentage, presence of *E. coli* BSI and 30-day mortality including the Nelson-Aalen estimator as recommended. The number of imputation datasets was set at 25 using the recommended rule of thumb that the number of imputations should be at least equal to the percentage of incomplete cases, allowing enough datasets to reduce the Monte Carlo error while still being computationally efficient.[187]

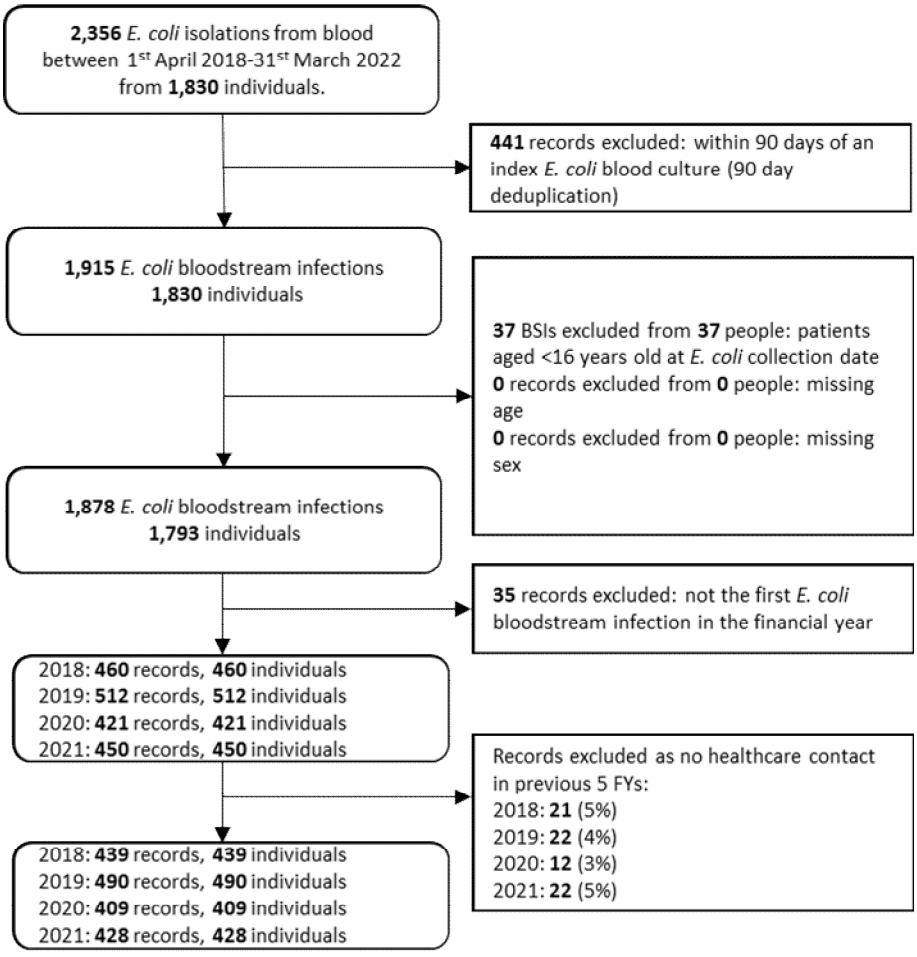
As one of the aims when defining the control group was to get as close to the population of Oxfordshire as possible, I compared some distributions of the core variables in IORD to available census data. I compared the age and sex distribution and the proportion in different ethnic groups to the 2021 census estimates,[188,189] and the distribution of IMD percentile to the most recent population estimates from 2019,[190] using the closest relevant year of IORD data (either 2021 or 2019 respectively).

## 4.3 Results

### 4.3.1 Summary of data

Between 1st April 2018 and 31st March 2022, there were 2,356 *E. coli* isolations from blood from 1,830 individuals (**Figure 4.2**). 441 isolations were excluded as they were within 90 days of an index *E. coli* blood culture, leaving 1,915 de-duplicated BSIs. 37 (2%) BSIs from 37 (2%) individuals were aged <16y at the time of index blood culture collection and were hence dropped from subsequent analysis. Between 3-5% of remaining *E. coli* BSIs had no healthcare contact recorded in IORD in the previous 5FYs and were dropped from all following analyses (investigated in more detail in Chapter 5). The highest number of *E. coli* BSIs was in the financial year (FY) 2019 (n=490) and the lowest in FY2020 (n=409).

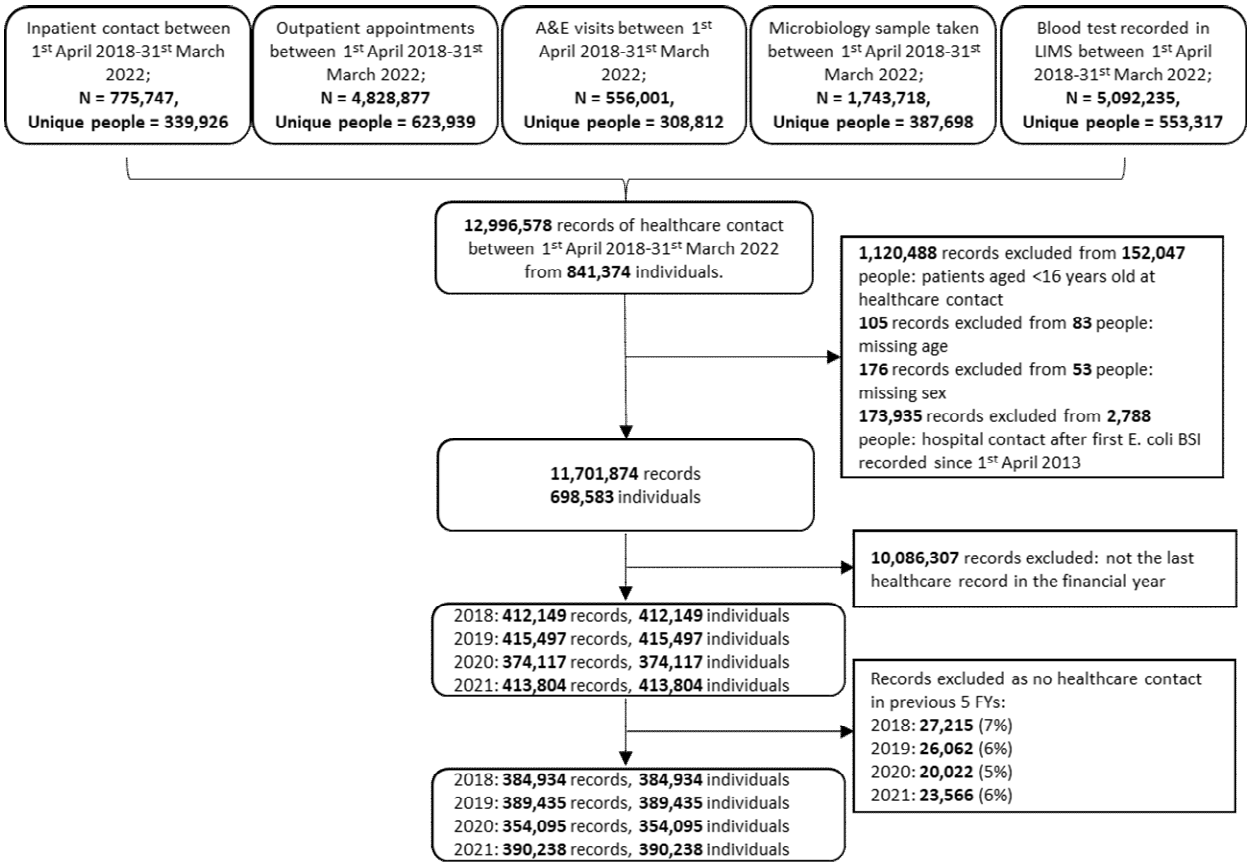
**Figure 4.2: Flowchart of the case population.**




From 1st April 2018 to 31st March 2022 there were 12,996,578 records from 841,374 individuals in the IORD database from inpatient admissions, outpatient appointments, A&E visits, microbiology samples, and blood tests (**Figure 4.3**). 1,120,488 (9%) records from 152,047 (18%) individuals were from those aged <16 years at the time of contact and were hence dropped from all subsequent

analyses. A very small number of individuals were missing age (83 [<0.01%]) or sex (53 [<0.01%]) – these individuals were also dropped from the analysis. From the potential control group population, 173,935 (1%) records from 2,788 (0.3%) individuals were excluded as they had had a previous *E. coli* BSI in IORD since 1st April 2013. This left 11,701,874 records from 698,583 individuals for control group consideration. Selecting the last observation per person per FY dropped 10,086,307 records. Between 5-7% of potential controls had no healthcare contact in the previous 5FYs in IORD and were dropped from all following analyses. As with cases, the highest number of potential controls was in FY2019 (n=389,435) and the lowest in FY2020 (n=354,095).

**Figure 4.3: Flowchart of the potential control group population.**



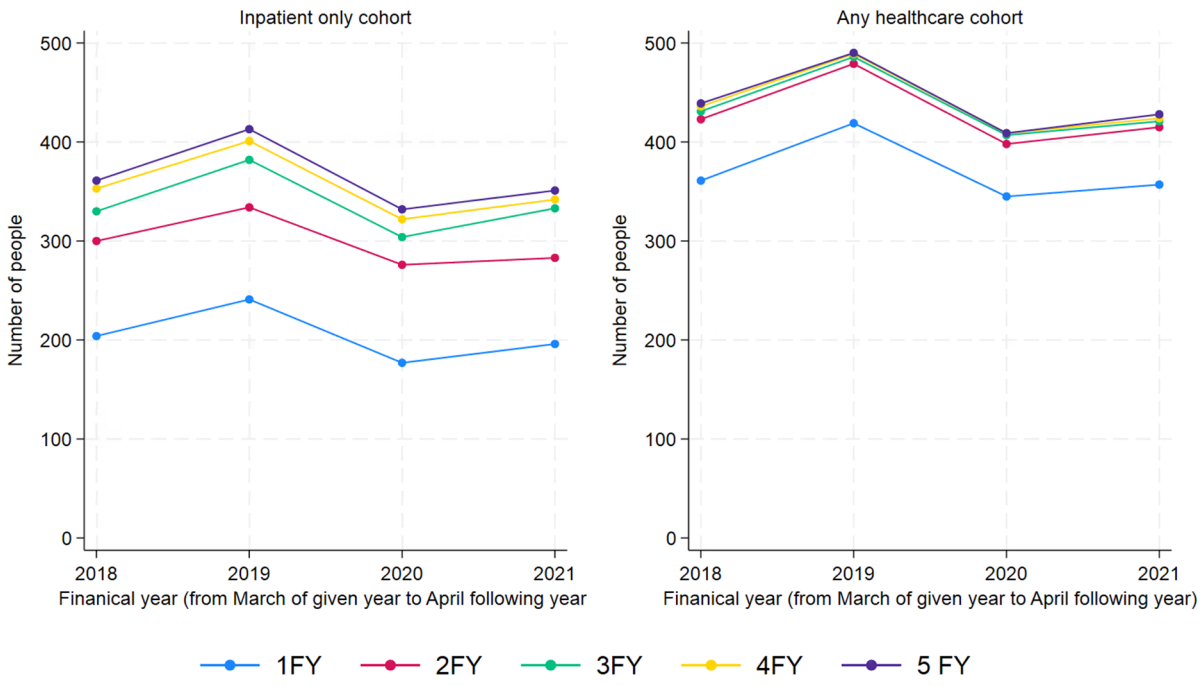

### 4.3.2 Comparing control groups

*Summary of control group characteristics*

The number of cases that could be included based on previous healthcare contact varied across the FYs included in the study period but the trend across the years was similar for all lookback lengths (**Figure 4.4**). For all cohorts and all lookback lengths, the number of cases increased between FY2018 to FY2019 (e.g. 14% increase using 5FY lookback and the "inpatient only" cohort) and then decreased in FY2020 (by 20% using 5FY lookback and the "inpatient only" cohort, onset of the

COVID-19 epidemic) before slightly increasing in FY2021 (by 6% using 5FY lookback and the "inpatient only" cohort). The number of *E. coli* cases included was lowest in the "inpatient only" cohort when only using those with inpatient contact in the previous 1FY. Looking back an extra FY (from 1FY to 2FY) increased the sample size of cases by approximately 50% in all FYs 2018-2022. The proportion of additional people was reduced as lookback length successively increased. The number of cases included in the "any healthcare" cohort with 1FY lookback was similar to the number of cases included in the "inpatient only" cohort with 5FYs lookback. An additional 62 cases were added in FY2018 when looking back 2FYs versus 1FY (relative percentage increase of 17%), with this pattern consistent across all the other FYs. In the "any healthcare" cohort, the number of cases added for each additional increase in 1FY lookback was small when increasing from 2FY-5FY lookback with 8, 5, and 3 additional cases added in 2018 when increasing to 3FY, 4FY, and 5FY, respectively.

**Figure 4.4: Number of people in the case group from FY2018-2021, stratified by different amounts of lookback and for the "inpatient only" cohort (left) and "any healthcare" cohort (right).**

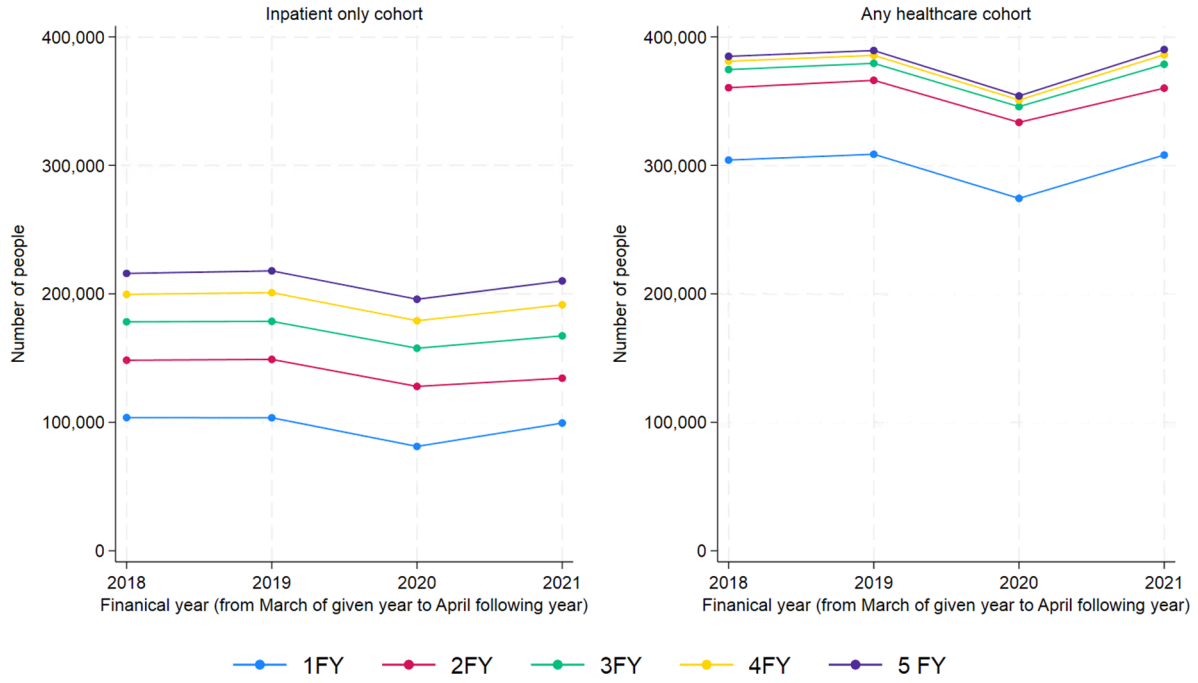



Note: FY = financial year.

The number of people in the control group that could be included based on previous healthcare contact was stable between FY2018 and FY2019, before decreasing in FY2020 and increasing back to similar levels in FY2021 (**Figure 4.5**). The impact of the reduction in the number of people in the control groups in FY2020 was larger when only considering 1FY compared with 5FY lookback. Using 1FY of lookback there were 103,602 individuals in FY2019 reducing to 81,306 in FY2020 (22,296 fewer people). Using 5FYs of lookback there were 217,934 individuals in FY2019 reducing to 195,881 in FY2020 (22,053 fewer people). Therefore, while the absolute reduction in the number of people

was similar, the relative decrease was larger using 1FY lookback (22%) versus using 5FYs lookback (10%). Similar to the case group, additional lookback length made a larger difference in the "inpatient only" cohort compared with the "any healthcare" cohort; for example, in FY2019, the control group size increased from 103,602 to 217,934 (absolute increase=114,332; 110% relative increase) when increasing the lookback length from 1FY to 5FY in the "inpatient only" cohort and increased from 308,722 to 389,435 (absolute increase=80,713; 26% relative increase) in the same year for the "any healthcare" cohort. The relative difference in the number of people in the "any healthcare" cohort compared with the "inpatient only" cohort was larger in the control group compared with the cases.

**Figure 4.5: Number of people in the potential control group from FY2018-2021, stratified by different amounts of lookback and for the "inpatient only" cohort (left) and "any healthcare" cohort (right).**




Note: FY = financial year.

There was little change in core characteristics across FYs, varying lookback lengths, or the "inpatient only" or "any healthcare" cohorts (**Figure 4.6**). In each financial year, those with *E. coli* BSIs were generally older than those in all control groups; for example, considering FY2021 the median (IQR) age in those with *E. coli* BSI was 77y (67y-85y) versus 56y (38y-72y) in controls for the "any healthcare" cohort with 1FY lookback. The median age was slightly higher in FY2020 in the "inpatient only" cohort with 1FY lookback compared with FY2018, however IQRs were broadly similar (median (IQR) age 80y (66y-85y) in FY2020, 74y (62y-85y) in FY2018).

The distribution of catchment percentage was right-skewed, with most individuals' home addresses being in an area with a high catchment percentage of >90% (**Figure 4.7**). Those with *E. coli* BSIs had smaller interquartile ranges with higher values of the 25th percentile compared with all control groups. The "any healthcare" cohorts generally had a slightly larger interquartile range including lower catchment percentages; for example, the median (IQR) catchment percentage for the "inpatient only" cohort control group with 2FYs of lookback in FY2019 was 94 (85-96) compared with 94 (82-96) in the equivalent "any healthcare" cohort control group.

**Figure 4.6: Median (IQR) age from FY2018 to FY2021 for the "inpatient only" and "any healthcare" cohorts with 1FY to 5FYs lookback.**

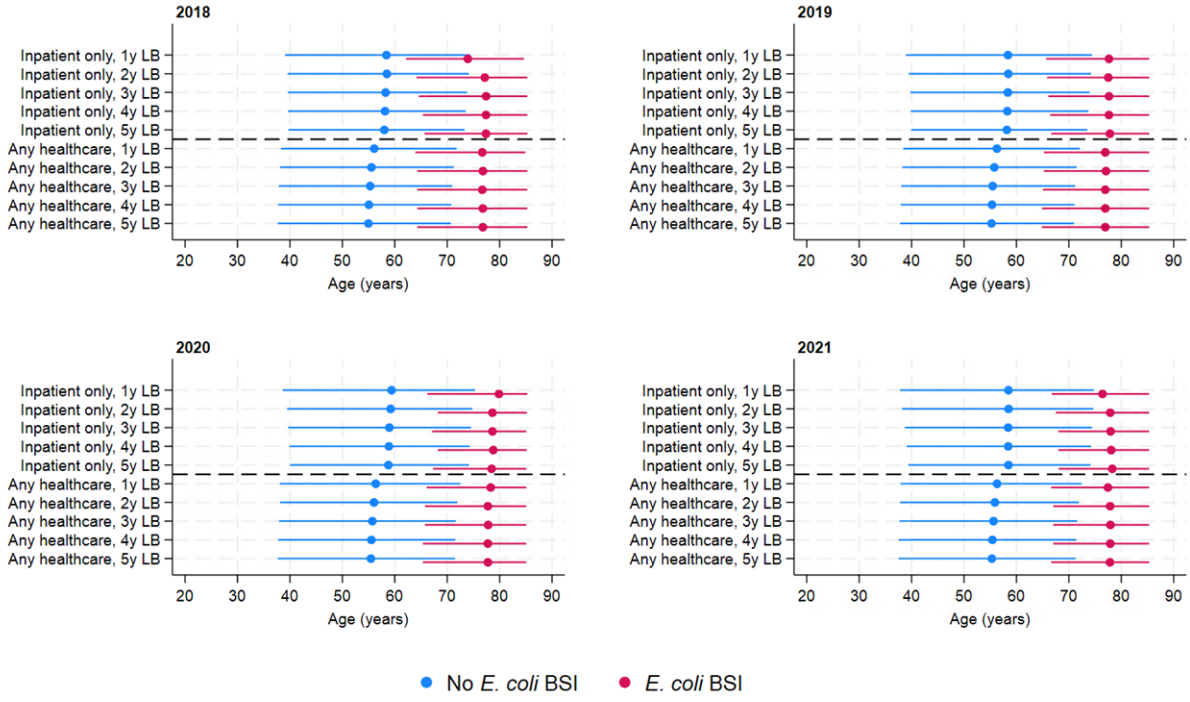
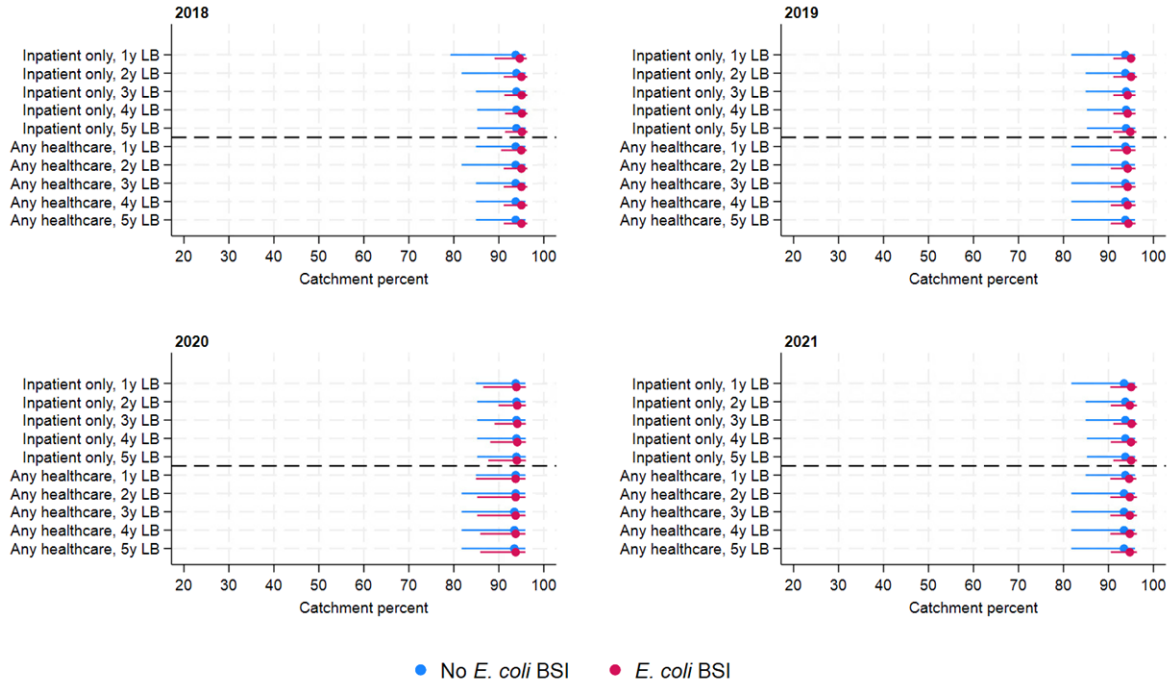
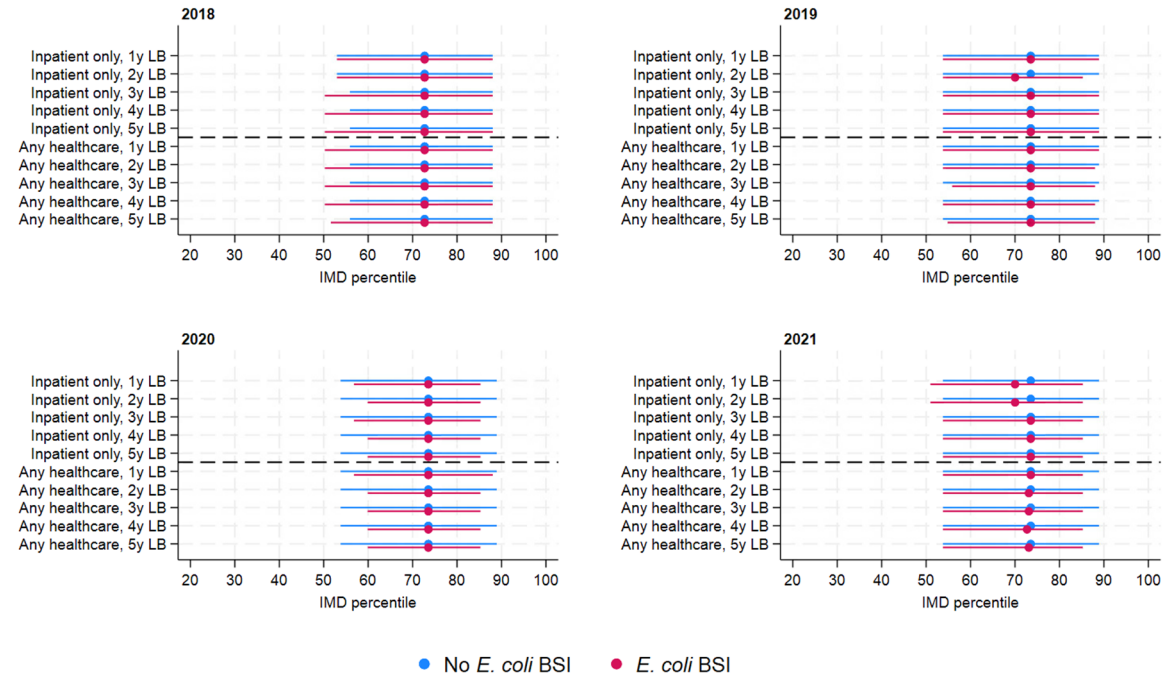


**Figure 4.7: Median (IQR) catchment percent from FY2018 to FY2021 for the "inpatient only" and "any healthcare" cohorts with 1FY to 5FYs lookback.**




Deprivation, measured by IMD percentile, was similar across all groups and varied very little between cases and controls, with a median deprivation ranging between 70-74 (25th percentile between 50-60; 75th percentile between 85-89) (higher scores indicate lower levels of deprivation) (**Figure 4.8**).

**Figure 4.8: Median (IQR) IMD percentile from FY2018 to FY2021 for the "inpatient only" and "any healthcare" cohorts with 1FY to 5FYs lookback.**



The percentage of females in the population was consistently higher in the control groups than it was in the cases, with between 56%-59% females in controls versus 41-52% in cases (**Figure 4.9**). In FY2019 and FY2020 the percentage of females was very consistent in the cases and controls across different lookback periods, while there were slight variations in cases in FY2018 and FY2021. In FY2021, the percentage of females increased from 41% using 1FY lookback to 46% using 5FY lookback in "inpatient only" cohort cases. In FY2021, the percentage of females was lower in the "inpatient only" versus the "any healthcare" cohort, with between 45-46% females in the "inpatient only" cohort cases and 49-52% in the "any healthcare" cohort cases.

The percentage of individuals of non-white ethnicities was low in all FYs, with varying lookback lengths, and in both cases and controls, constituting between 2%-9% of the populations (**Figure 4.10**). There was a consistently lower percentage of non-white ethnicities in cases than in controls, with a larger difference in FY2020. This was due to a lower proportion of cases with non-white ethnicities in FY2020 compared with other FYs; median (IQR) percentage of cases of non-white ethnicity across all lookbacks in FY2020: 2.1% (2.1%-2.2%) versus 4.9% (3.3%-5.4%) in all other FYs.

**Figure 4.9: Percentage female from FY2018 to FY2021 for the "inpatient only" and "any healthcare" cohorts with 1FY to 5FYs lookback.**

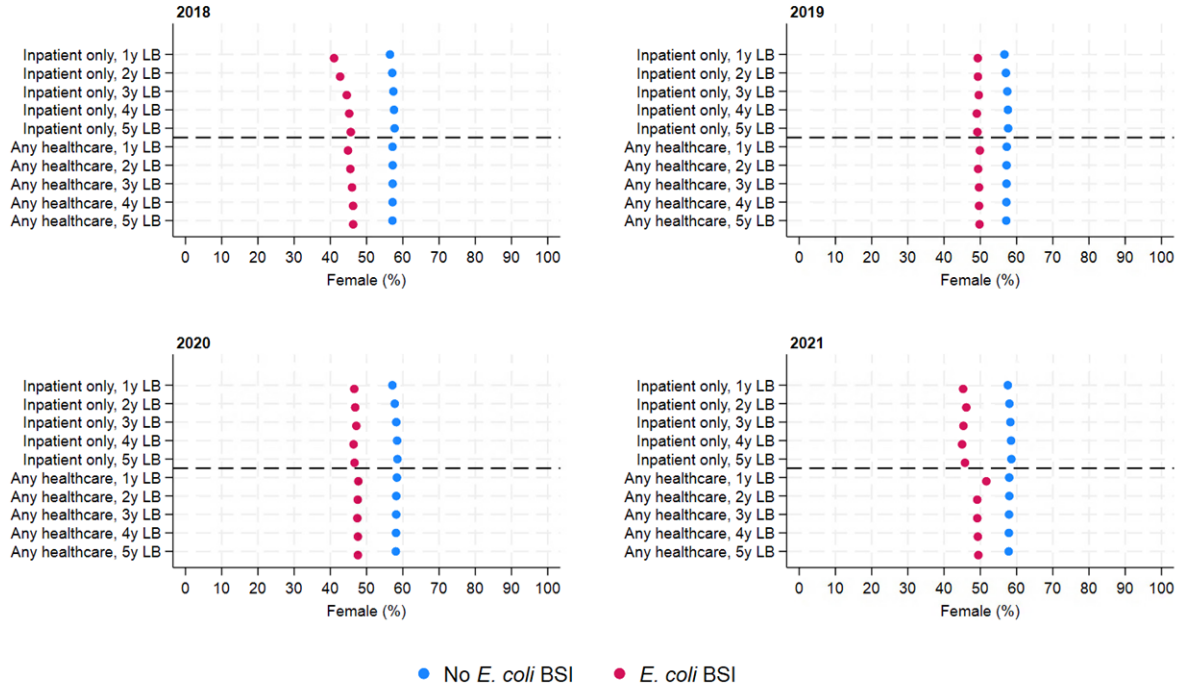
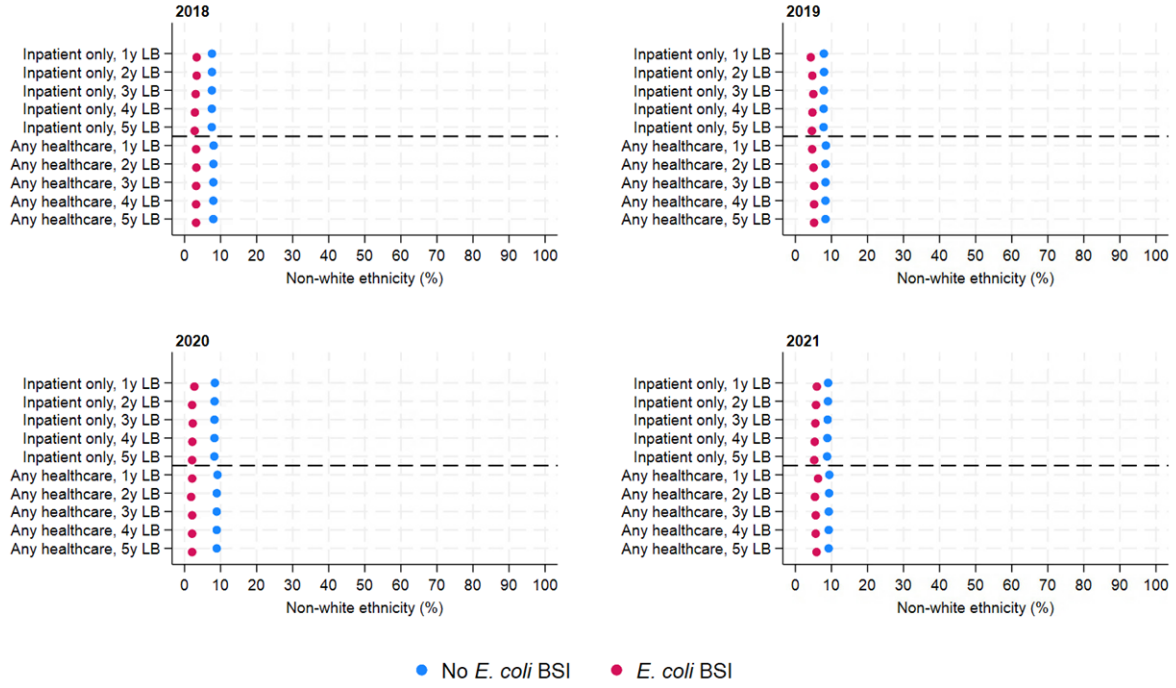
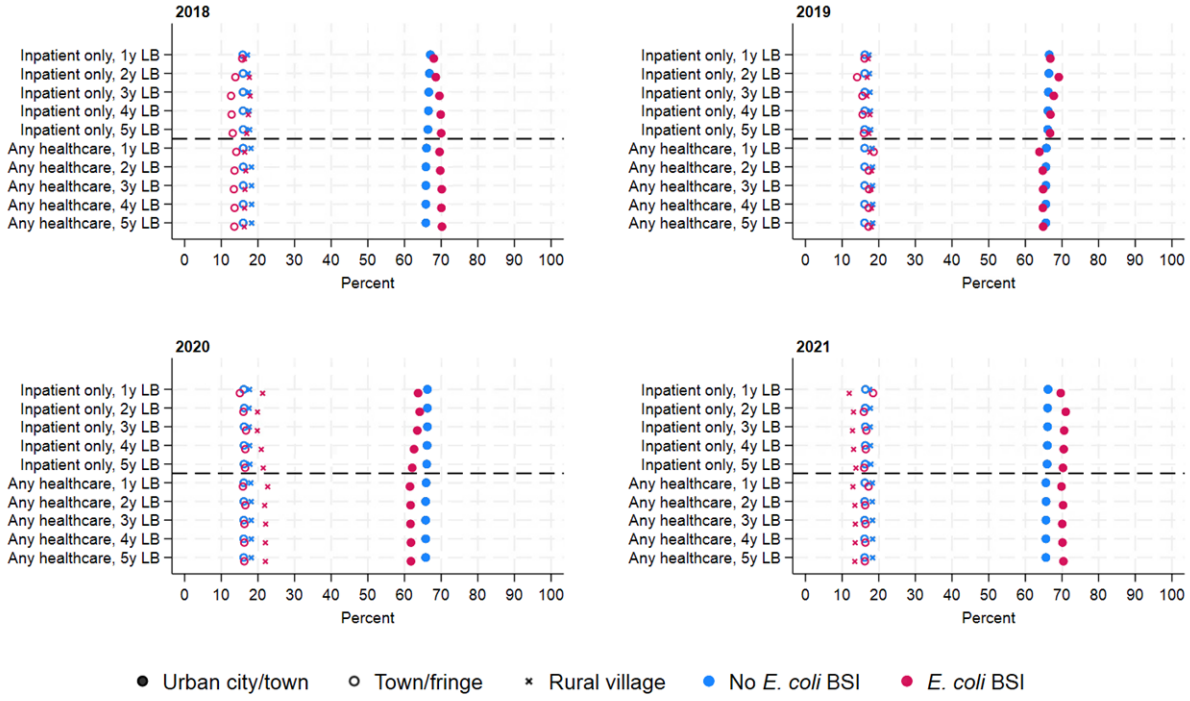


**Figure 4.10: Percentage of non-white ethnicities from FY2018 to FY2021 for the "inpatient only" and "any healthcare" cohorts with 1FY to 5FYs lookback.**



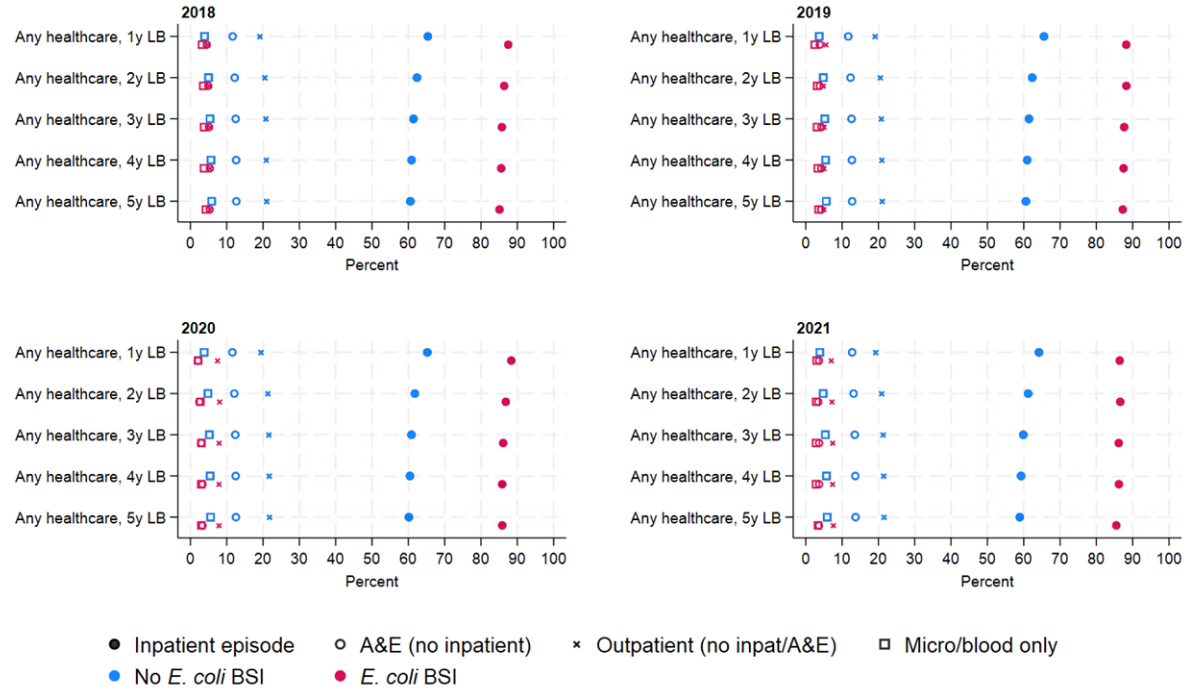

The large majority of both cases and controls were living in urban cities/towns, constituting between 61-71% across both cases and controls for all financial years and lookback lengths (**Figure 4.11**). The percentage of controls living in towns/fringes and rural villages was very similar, while cases varied slightly across the years. For example, in FY2020 between 20-23% of *E. coli* cases resided in rural villages, compared with between 12-14% of *E. coli* cases in FY2021. Increased lookback length had little to no impact on the percentage of cases and controls from different rural/urban areas.

**Figure 4.11: Percentage in rural/urban categories from FY2018 to FY2021 for the "inpatient only" and "any healthcare" cohorts with 1FY to 5FYs lookback.**



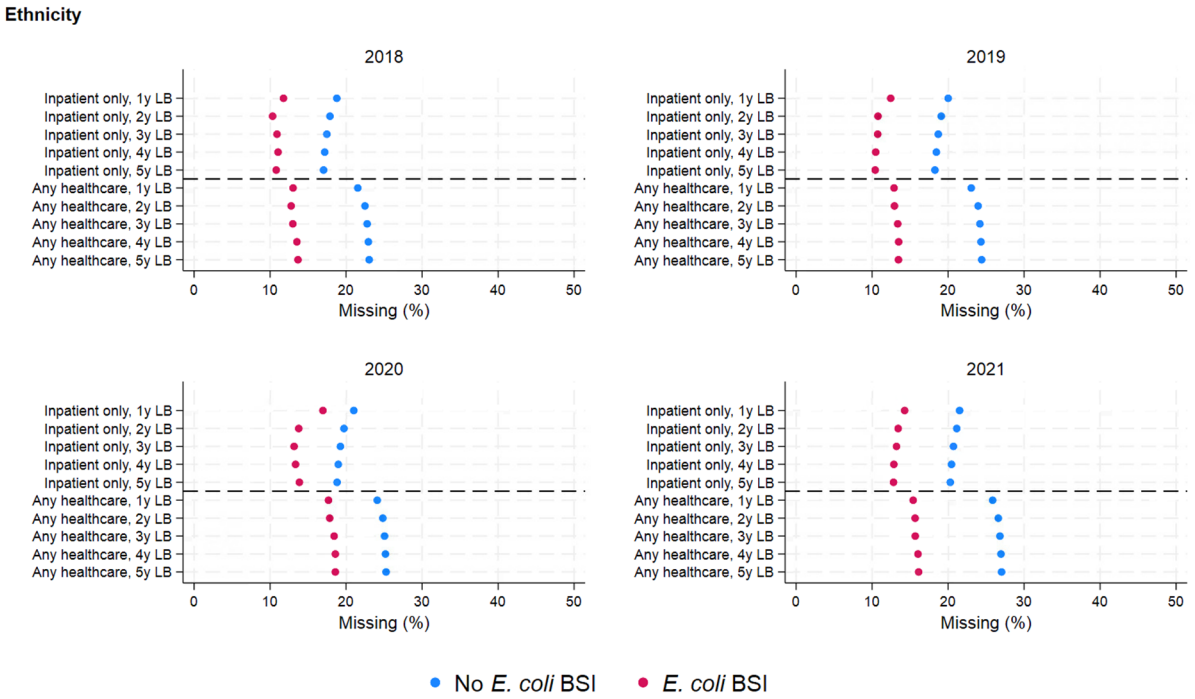

In the "any healthcare" cohort, a higher percentage of *E. coli* cases had previous inpatient contact compared with controls (85-88% of cases versus 59-66% of controls) (**Figure 4.12**). For controls, after previous inpatient contact, having previous outpatient contact but without inpatient or A&E contact was the second most common previous healthcare exposure, constituting 19-22%. Around 12-14% of controls had a previous A&E visit without an inpatient admission compared with around 2-5% of cases. Only between 2-4% of cases and 4-6% of controls had a previous microbiology or blood test but without a previous inpatient admission, A&E visit, or outpatient attendance. The percentage from different hospital exposure groupings remained similar across different lookback lengths with a slightly higher percentage of individuals with previous inpatient contact using 1FY of lookback compared with 2FYs-5FYs.

**Figure 4.12: Percentage of different previous hospital exposures from FY2018 to FY2021 for the "any healthcare" cohort with 1FY to 5FYs lookback.**




More data was missing for the core variables in the “any healthcare” cohort, particularly for ethnicity in controls (**Figure 4.13**). The “any healthcare” cohort control group with 1Y lookback had 22-26% missing ethnicity across all FYs, compared with 13-18% in *E. coli* cases. The proportion of people missing ethnicity remained relatively stable as the length of lookback increased in the “any healthcare” cohort; for example, in FY2018 in the “any healthcare” cohorts, the percentage of people missing ethnicity increased only slightly from 13.0% to 13.7% and 21.5% to 23.0% in cases and controls, respectively. Those with an inpatient episode in the previous 1FY had a higher proportion of missing data than those with an inpatient episode in the previous 5FYs, most evidently in FY2020. A small number of individuals were missing IMD percentile, catchment percentage, or rural/urban classification, ranging between 435-2,312 (0.3%-0.8%) individuals missing values in controls and 0-4 (0-1.5%) missing in cases. This small amount of missing data was similar across all these three variables as they were matched to each individual’s underlying residential address in the underlying NHS data using Lower Super Output Areas, hence generally either all present or all were missing.

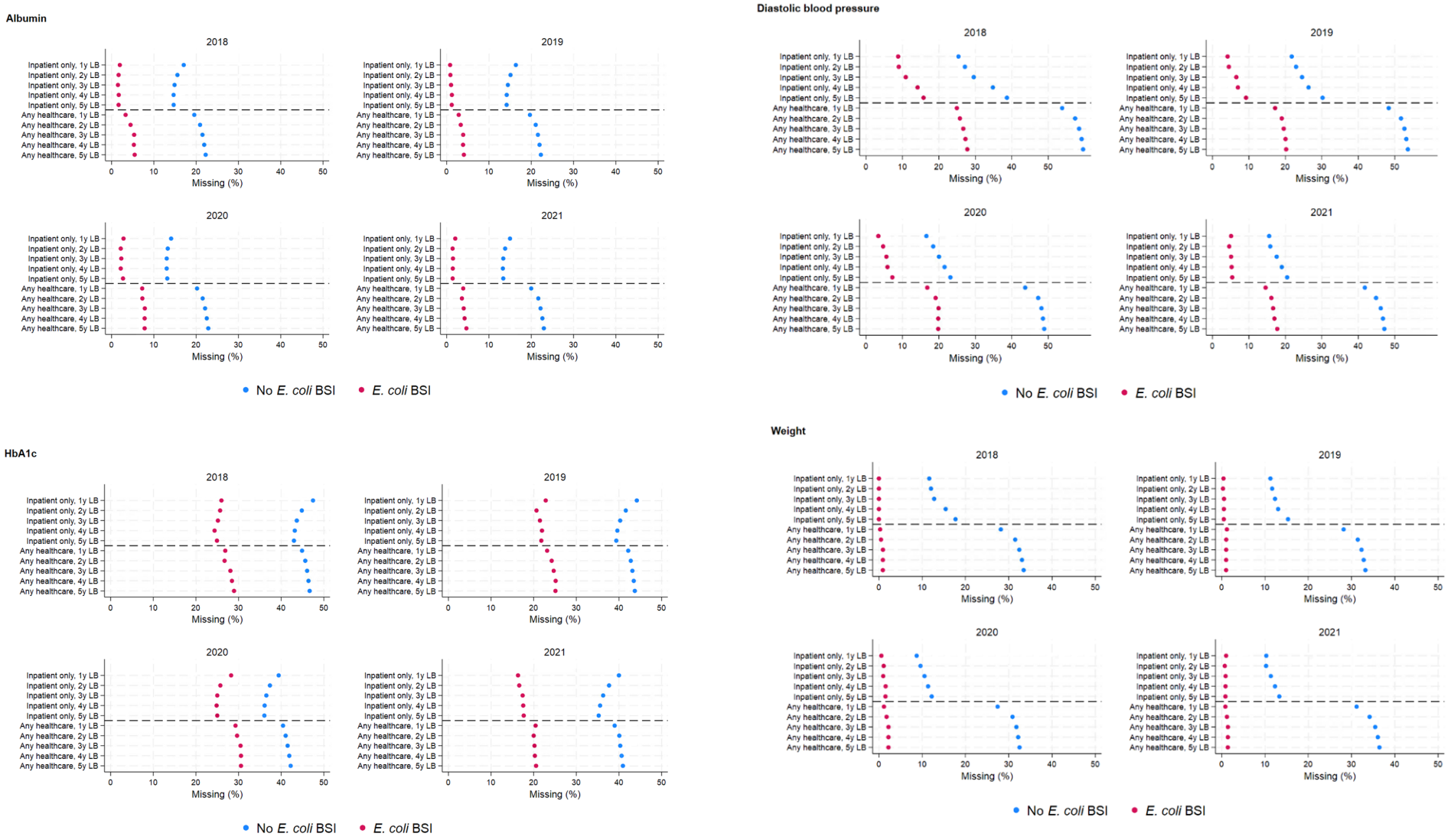
**Figure 4.13: Percentage of missing data for ethnicity for potential controls and *E. coli* BSI cases from FY2018-2021 with varying amounts of lookback.**




There was also more missing data for the screening variables in the "any healthcare" cohort compared with the "inpatient only" cohort, specifically for blood test results, vital signs, and measurements of weight and height (examples given in **Figure 4.14**). This was the case for both cases and controls. For many blood tests, as shown for albumin and HbA1c, increasing the length of lookback decreased the percentage of missing data in both cases and controls in the "inpatient only" cohorts, while increased lookback in the "any healthcare" cohorts increased the percentage of individuals with missing data, suggesting that the extra individuals being added into the control groups with increasing lookback periods were more and less likely to have these variables measured, respectively. This may be because these blood test measurements are more likely to be recorded within inpatient admissions and therefore adding in individuals with previous inpatient admissions may include more people with these measurements. In contrast, adding individuals with only previous outpatient or A&E visits may increase missing data as these tests are less likely to have been done in their previous, non-inpatient, contact. The amount of missing data varied between blood tests as some blood tests, for example, HbA1c, may not be as routinely done as other tests. The percentage of individuals missing data increased as lookback length increased for vital signs, such as diastolic blood pressure, in FY2018, FY2019, and FY2020; however, this effect was not present in FY2021. This is likely because vital signs were not consistently electronically recorded in IORD until 2016; hence, additional years of lookback before and including 2016 were more likely not to have measurements recorded. The percentage of individuals missing weight was very small in

cases in both the "inpatient only" and "any healthcare" cohorts, and larger in the potential controls. The "inpatient only" cohort had between 10-18% missing weight, while the "any healthcare" cohort had 28-34% missing.

**Figure 4.14: Percentage of missing data for four example screening variables in potential controls and *E. coli* BSI cases from FY2018-2021 with varying lookback.**



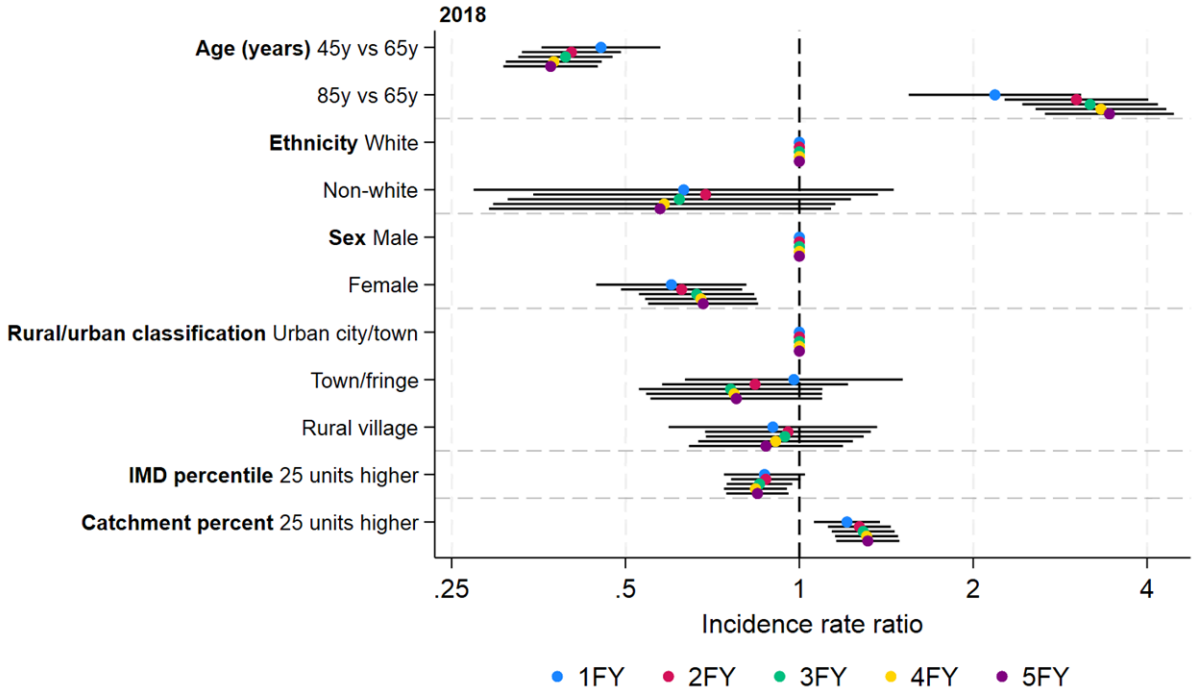
*Comparison of model estimates*

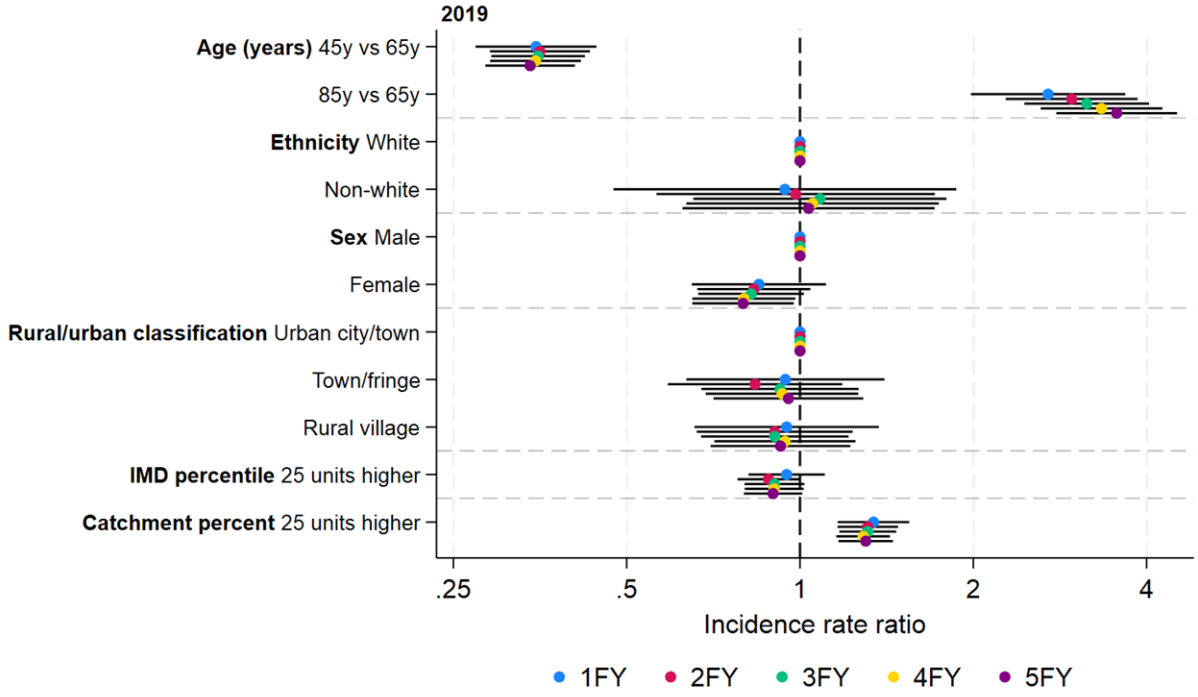
Estimates from the core model using different lengths of lookback for the "inpatient only" cohort varied; however, interpretations from the models remained similar (**Figure 4.15**). In all FYs and for all lookback lengths, older ages versus the median were associated with higher risk of *E. coli* BSIs while younger ages versus the median were associated with lower risk. Estimates for ethnicity varied across FY but were broadly similar across different lookback periods. However, confidence intervals were large due to small numbers, perhaps influencing the stability of results. In FY2018, the effect of sex attenuated slightly as lookback length increased; however, this was not consistent across all FYs, with the opposite effect in FY2019. In FY2020, there was no evidence of an effect of sex at the $p<0.05$ threshold using the 1FY lookback population but there was evidence at $p<0.05$ for 4FYs and 5FYs lookback. There was evidence of an effect at $p<0.05$ for catchment percentage and IMD percentile in all FYs apart from FY2020 where estimates were close to the null. In all other FYs, lower deprivation (indicated through higher IMD percentiles) was associated with lower risk of *E. coli* BSIs and higher catchment percentage with higher risk of *E. coli* BSIs. In almost all scenarios, these effects were stronger in longer lookback populations.

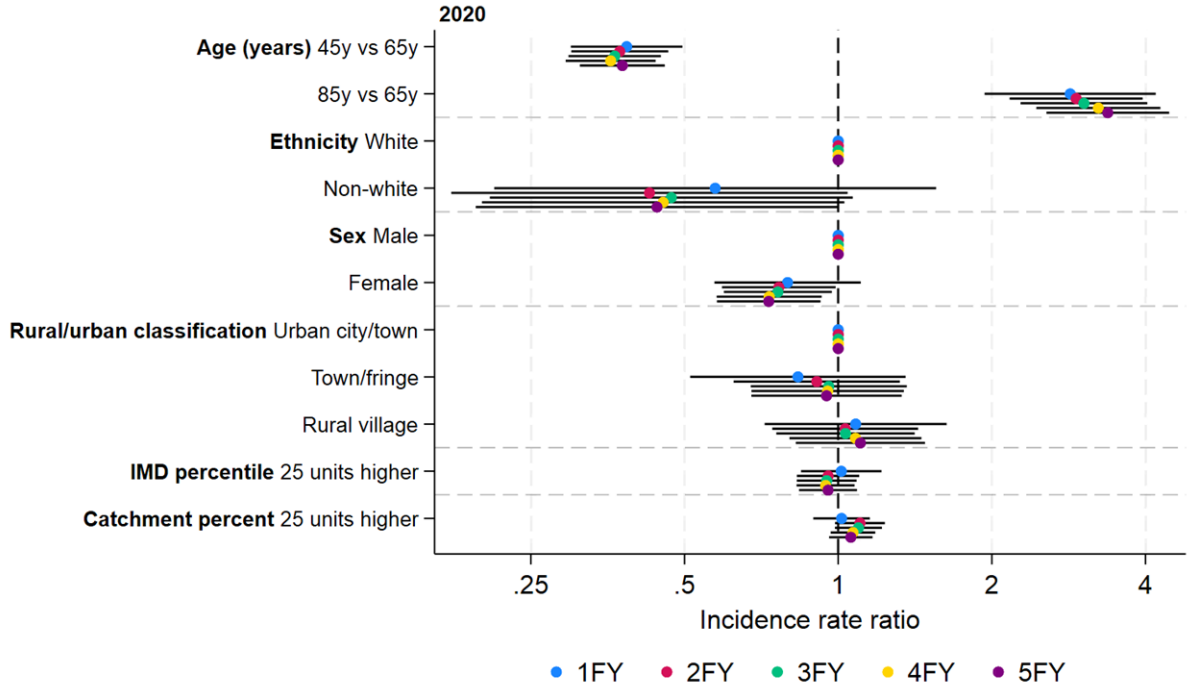
In the "any healthcare" cohort, there was evidence of an effect of ethnicity of $p<0.05$ using all lookback lengths in FY2020 with those reporting non-white ethnicities associated with lower risk of *E. coli* BSIs (**Figure 4.16**). There was no evidence of an effect of ethnicity in any other FY. Increased lookback length made very little difference to core model estimates in this cohort. The risk of *E. coli* BSIs was lower in all groups of individuals with no previous inpatient contact apart from individuals with a prior microbiology or blood collection, particularly for shorter lookback lengths. Associations between all core variables included in both the "inpatient only" and "any healthcare" core models were similar.

As differences in model estimates between "inpatient only" and "any healthcare" cohorts were small but there was more missing data in the "any healthcare" cohort, the "inpatient only" cohort was selected for further investigation. As the length of lookback period made little difference to the distribution of the core variables or the core model estimates, a lookback period of 5FYs was selected to maximise the number of people included in analyses.

**Figure 4.15: Model estimates comparing different lengths of lookback in the "inpatient only" cohort between 2018-2022.**




2020
Age (years) 45y vs 65y
85y vs 65y
Ethnicity White
Non-white
Sex Male
Female
Rural/urban classification Urban city/town
Town/fringe
Rural village
IMD percentile 25 units higher
Catchment percent 25 units higher
.25
.5
1
2
4
Incidence rate ratio
1FY
2FY
3FY
4FY
5FY

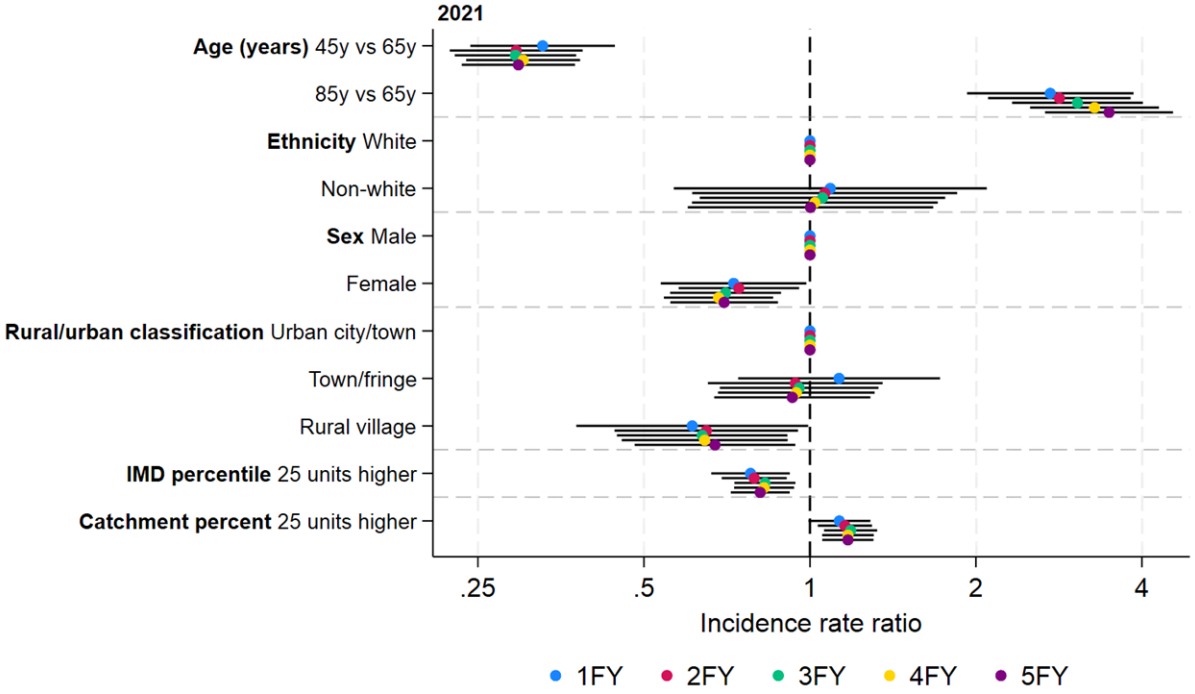
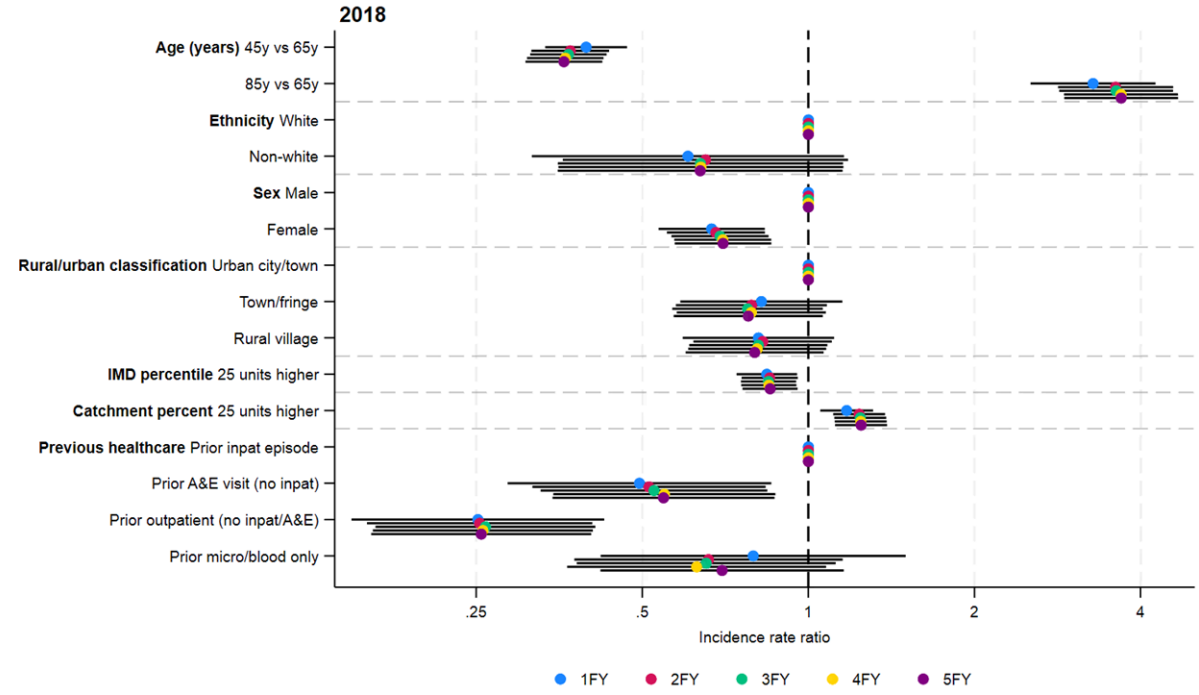
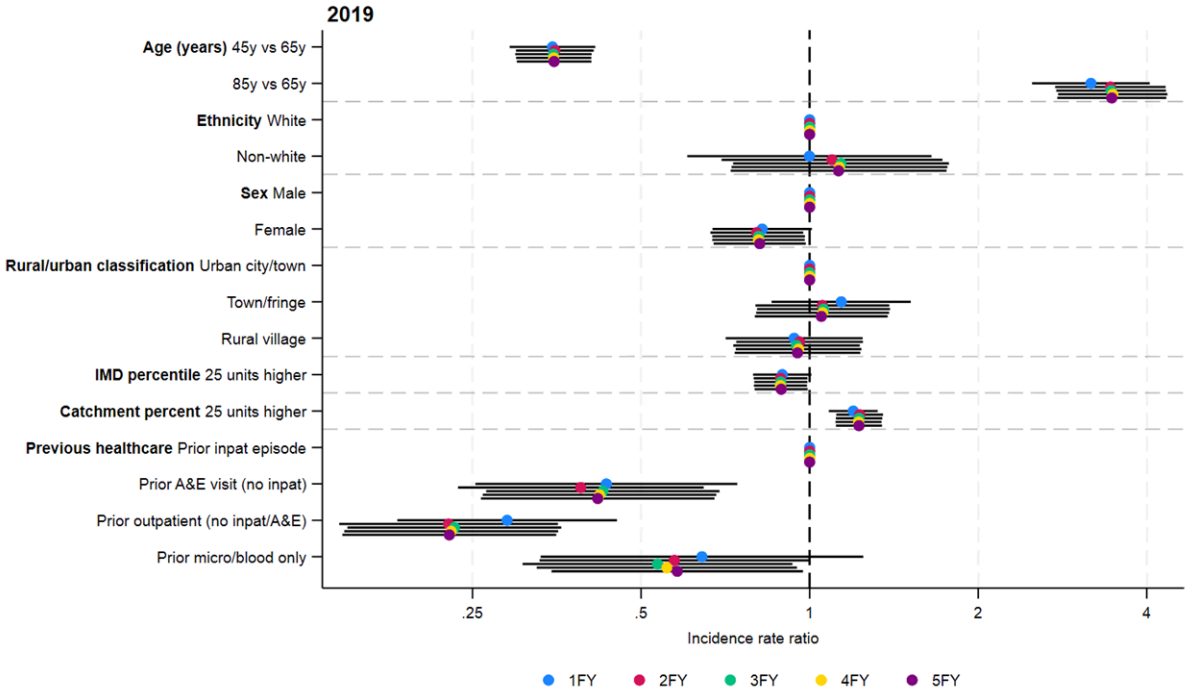
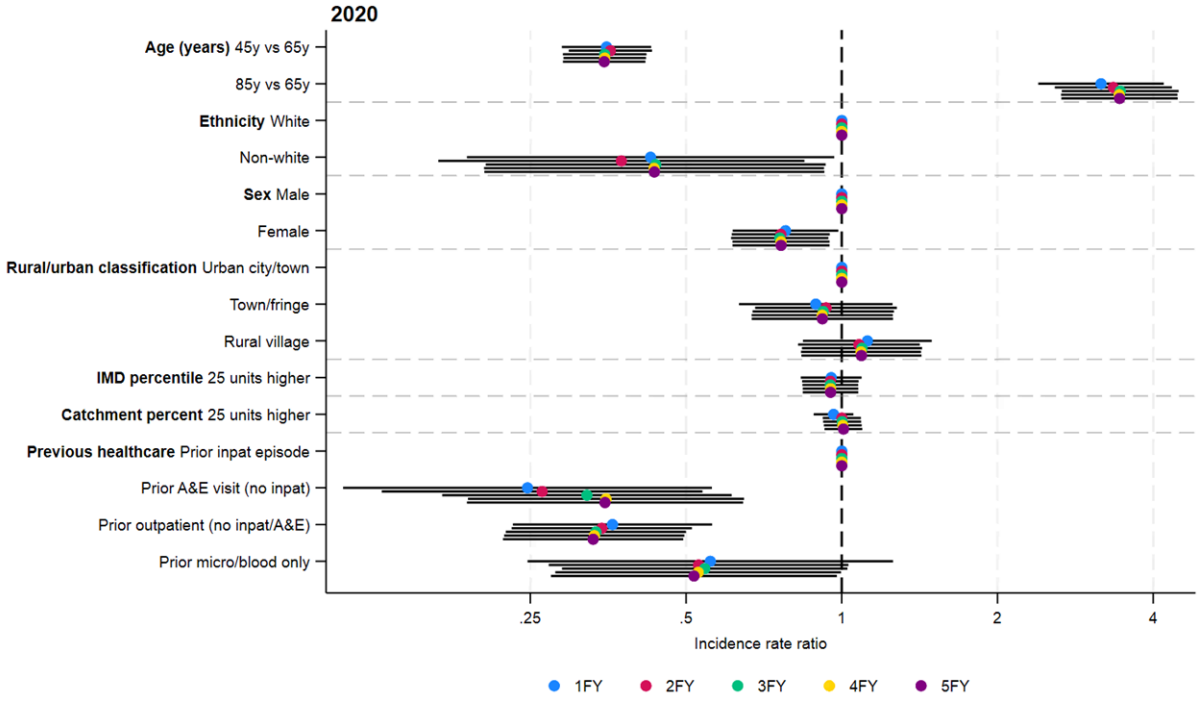
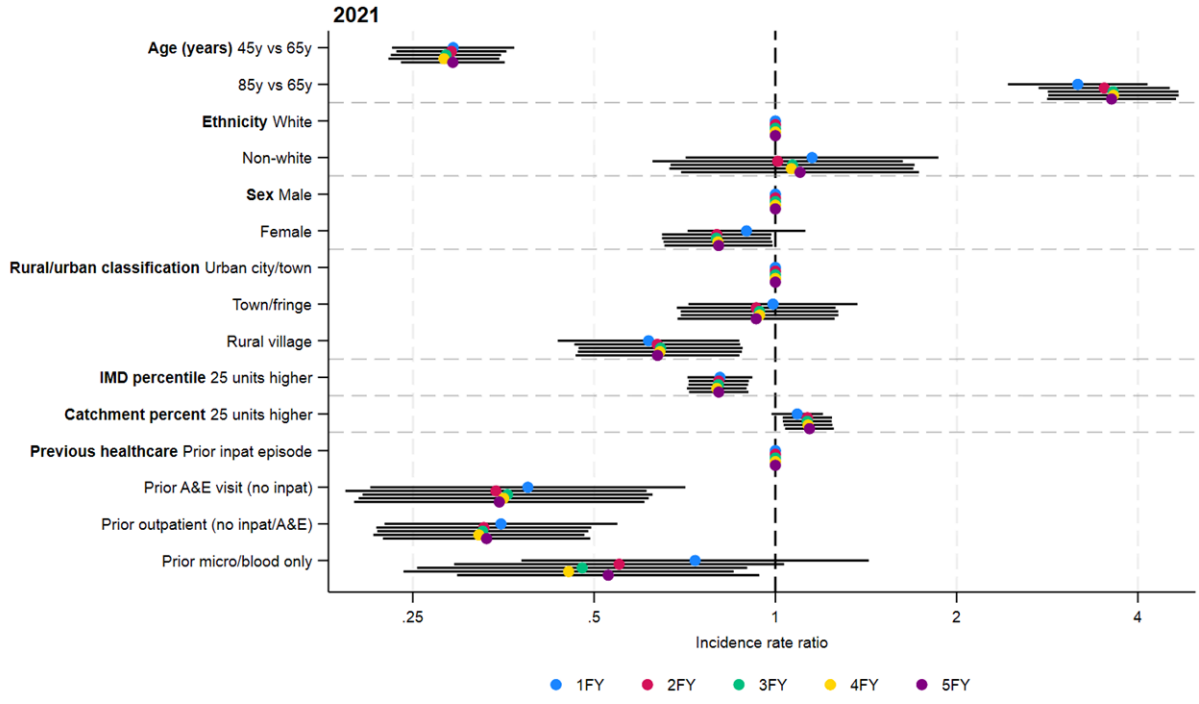

2021
Age (years) 45y vs 65y
85y vs 65y
Ethnicity White
Non-white
Sex Male
Female
Rural/urban classification Urban city/town
Town/fringe
Rural village
IMD percentile 25 units higher
Catchment percent 25 units higher
.25
.5
1
2
4
Incidence rate ratio
1FY
2FY
3FY
4FY
5FY

**Figure 4.16: Model estimates comparing different lengths of lookback in the "any healthcare" cohort between 2018-2022.**





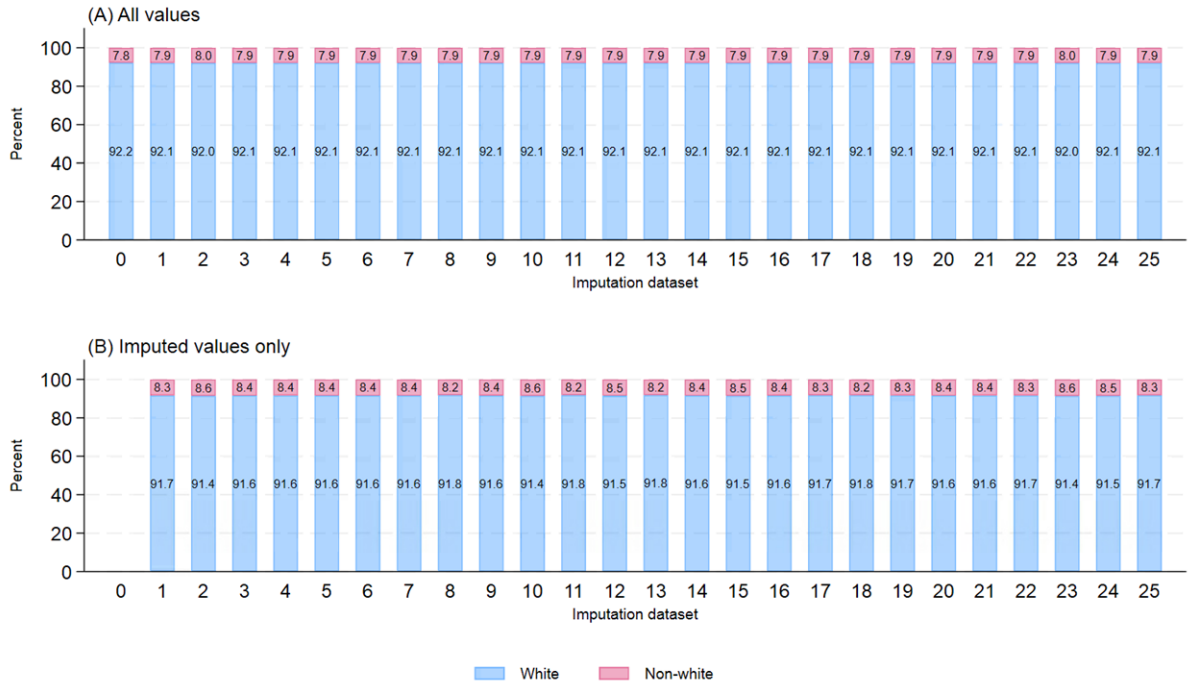

### 4.3.3 Sensitivity analyses for control group definitions

*Multiple imputation of ethnicity*

The relatively high proportion of individuals missing ethnicity in both the cases and controls with varying lengths of lookback (~20%) could have impacted model results in the complete cases analysis presented above. Therefore, ethnicity was imputed for FY2019 for the "inpatient only" cohort using

5FYs of lookback. In FY2019, 39,859 (18%) of individuals were missing ethnicity: 39,816 (18%) controls and 43 (10%) cases. A small number of people were missing either IMD percentile, rural/urban classification, or catchment percentage (974 (0.4%) controls and 0 (0%) cases) so these variables were not imputed; instead, these individuals were dropped from the imputation analysis. Imputed data had a slightly higher proportion of non-white ethnicities compared with the complete case analysis (**Figure 4.17**).

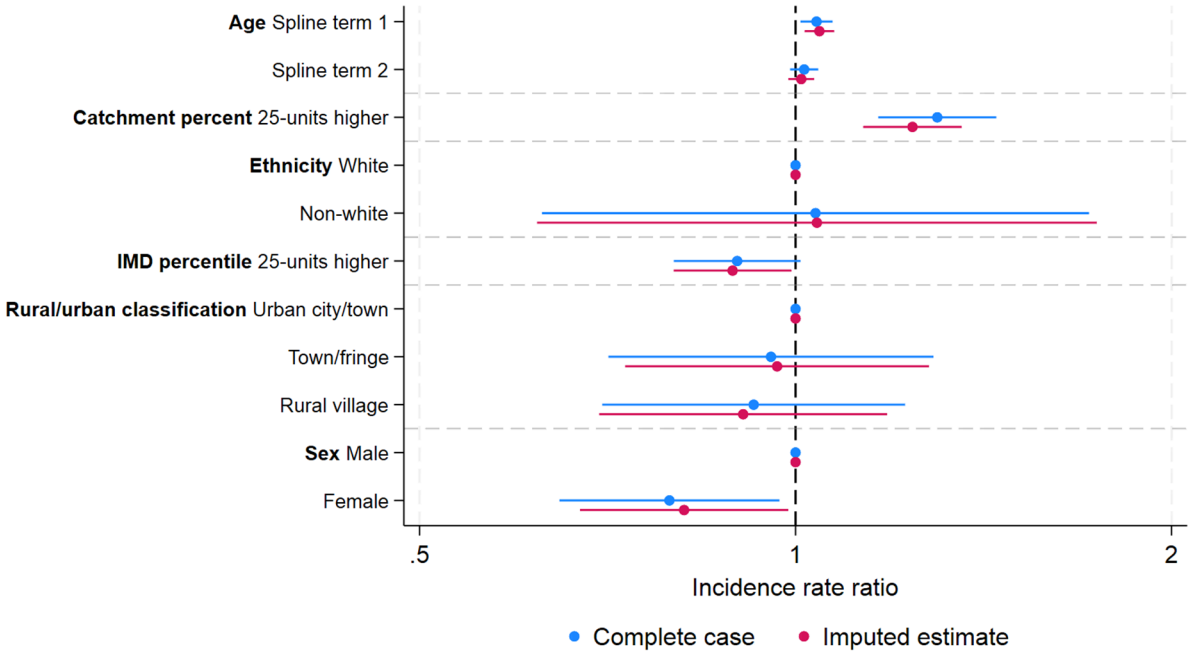
**Figure 4.17: Summary of imputed values for ethnicity.**




*Note: Imputation dataset 0 is the complete case data.*

Multiple imputation of missing ethnicity made little difference to estimates from the core model (**Figure 4.18**). The effect of catchment percentage slightly shifted towards zero in the imputed datasets analysed, however, the effect remained with $p<0.05$ and hence similar conclusions would be drawn. All other effects remained stable in the combined estimate from the imputed datasets compared with the complete-case analysis. The multiple imputation and estimating the combined effect took approximately five minutes of computational time.

**Figure 4.18: Model results from the core model run on complete cases (blue) and the combined estimate of the 25 datasets after multiple imputations.**




Note: Estimates of the effect of catchment percentage and IMD percentile are per 25 units higher.

*Comparison to Census data*

I observed large but expected differences when comparing the characteristics of all those in the "any healthcare" and "inpatient only" cohorts (cases and controls) with contact in FY2021 in IORD to those in the ONS census data for Oxfordshire from 2021. The "inpatient only" cohort had the highest proportion of individuals aged 61y-70y, 71y-80y, and 81y+ for both males and females, followed by the "any healthcare" cohort, and then the census (**Figure 4.19**) There was a corresponding under-representation of age groups between 16y-50y in both the IORD cohorts, particularly males between the ages of 21y-40y. Women between the ages of 31y-40y made up a similar percentage of the population in both the IORD cohorts and the census, likely relating to pregnancy-based attendances. Ethnicity was generally similar across the two data sources with both IORD cohorts having a slightly higher percentage of individuals of white ethnicity (91% versus 87%) (**Figure 4.20**). Compared with the latest 2019 population level deprivation statistics, those within the "inpatient only" and "any healthcare" cohorts in IORD in FY2019 had a higher proportion of people from more deprived areas (lower IMD deciles) and a lower proportion of people from less deprived areas (higher IMD deciles) (**Figure 4.21**), consistent with deprivation being associated with poorer health status.

**Figure 4.19: Age-sex distribution for IORD and the Office for National Statistics census data for 2021.**

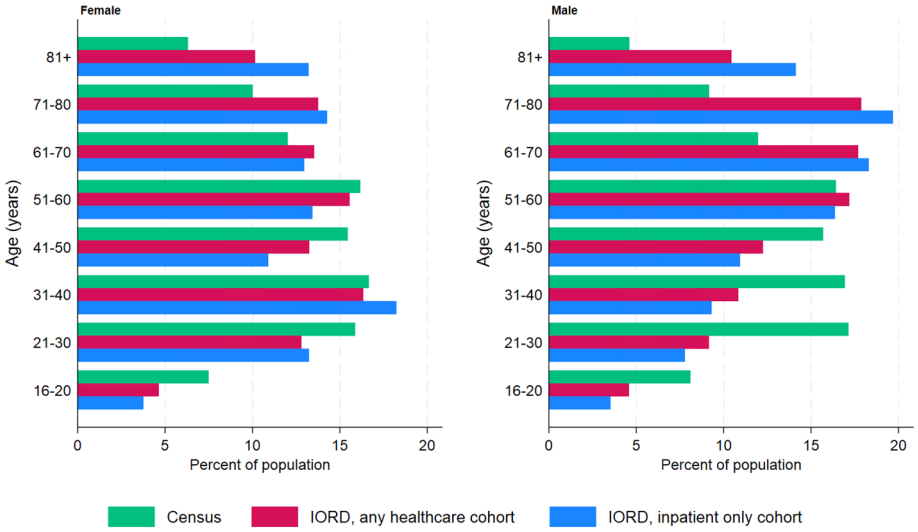



**Figure 4.20: Ethnicity distribution for IORD and the Office for National Statistics census data for 2021.**

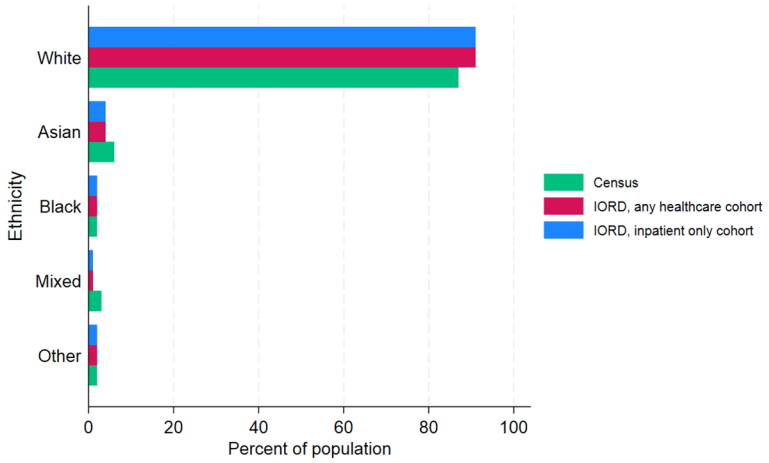
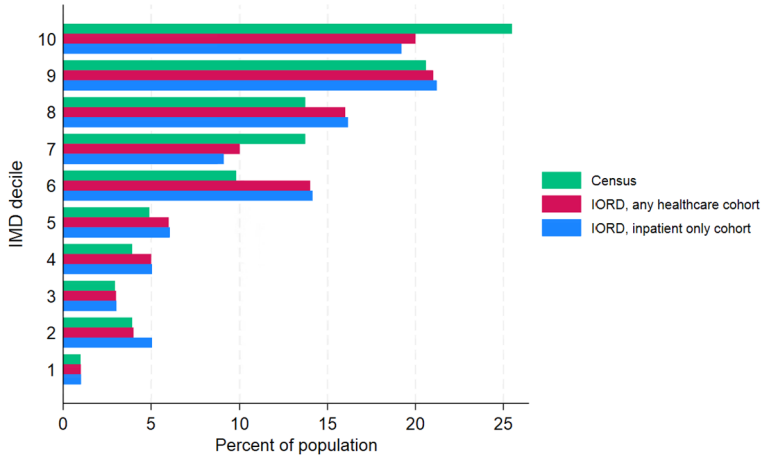



**Figure 4.21: Index of Multiple Deprivation distribution for IORD and the Office for National Statistics census data for 2019.**




Note: 1 = most deprived, 10 = least deprived.

### 4.3.4 Defining exposures

Due to the lower amount of missing data and little difference in model estimates, as described above, the "inpatient only" cohort using a 5FY lookback was chosen for further investigation of screening variables.

228 potential risk factors were defined across all available datasets, as described in Methods. There were 410 variables after considering both the time since the most recent exposure in that last 365d and never/ever having the characteristics. While 365d is an arbitrary time window, it was selected as I assumed that more recent exposure may increase risk and this would attenuate as time since the most recent occurrence got further away. These variables could broadly be split into seven groups dependent on where they were derived from:

- 102 from diagnostic codes,
- 140 from procedure codes,
- 54 from microbiology isolates,
- 81 from previous healthcare attendance (inpatient admissions, outpatient appointments, and A&E visits),
- 24 from blood tests,
- 6 vital signs,
- 3 personal traits.

A full list of variables with definitions is available in **Appendix A**.

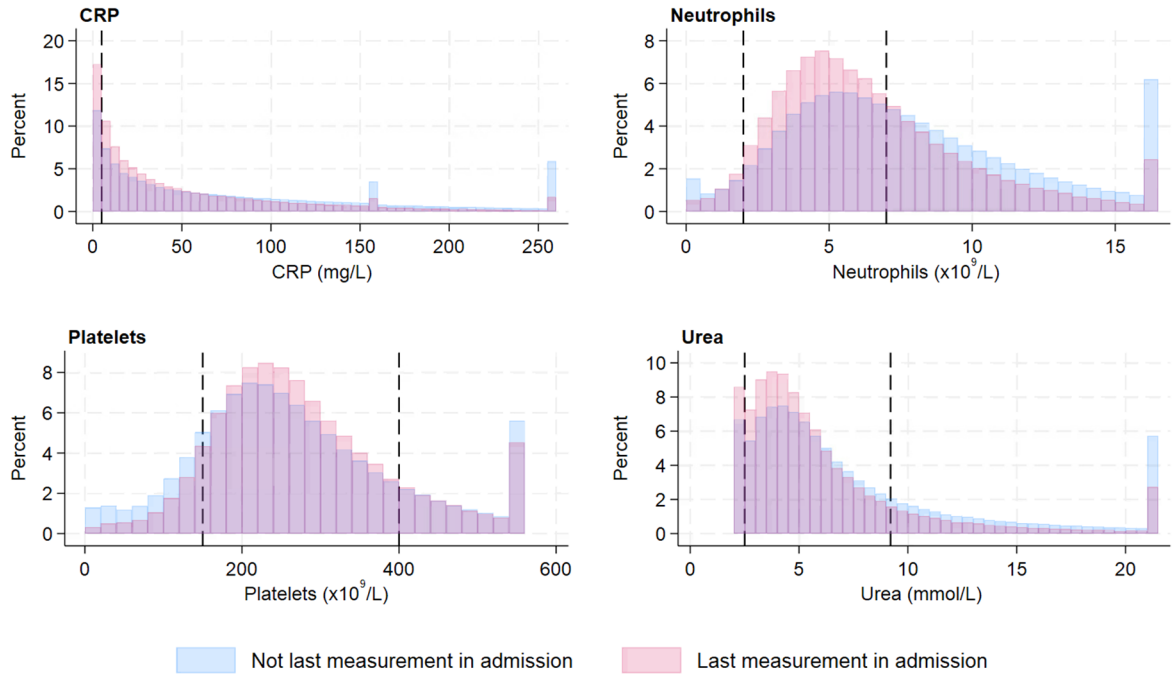
Before summarising the presence of these variables in cases and controls, I first had to decide which values of each variable to consider in each FY dependent on the proximity of the value to the "most recent contact" date, and the location in which the risk factor was collected. I therefore explored the timing and location of risk factors with respect to the "most recent contact" date for both cases and controls.

*Timing of measurements had an impact on exposure value*

To investigate whether certain values of different blood test results and vital signs were risk factors for *E. coli*, I wanted to get measurements as close to a "baseline" as possible rather than using measurements taken during acute illness. I first assessed the variability of measurements taken at different times within inpatient admissions for cases and controls in FY2019. As inpatient episodes all strictly preceded the "most recent contact date" they, by definition, did not include admissions which ended in death.

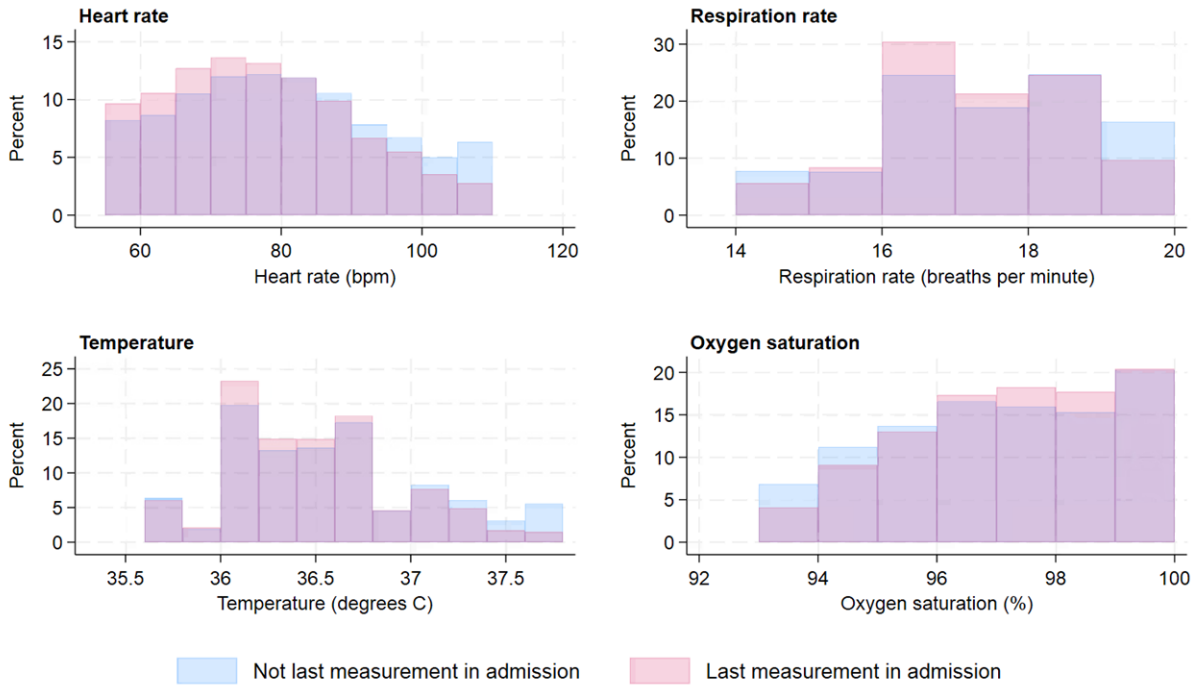
The distribution of test results was closer to the normal range if measurements were taken closer to hospital discharge, compared with measurements taken at the start or during an inpatient admission (**Figure 4.22**). While abnormal test results did exist in the last measurements in an inpatient admission, measurements taken elsewhere in admissions had a higher percentage of tests above the 95th percentile of the distribution. A similar, but not as pronounced, pattern was also observed for vital signs (**Figure 4.23**). For example, the distribution of heart rate contained slightly more people with higher heart rates (>85bpm) and higher temperatures if the measurement was not the last within an inpatient admission. While some of these differences were subtle, it made sense to prioritise the last recorded blood test or vital signs within an inpatient admission over measurements taken at other times during an admission.

**Figure 4.22: Distribution of four blood test results split by whether the test was the last available measurement in the inpatient admission.**




Note: Black dashed lines indicate reference ranges for tests, as recorded in IORD. Results were truncated at the 95th percentile. The higher percentage of CRP values at 160mg/L was due to some results being reported as CRP>160mg/L if any value >160mg/L was detected in the database.

**Figure 4.23: Distribution of four vital signs split by whether the result was the last measurement available in the inpatient admission.**




As the long-term goal of the analysis was to find risk factors for *E. coli* BSI, it was important that variables did not reflect symptoms of the *E. coli* BSI itself. I therefore explored the impact of removing values close to the "most recent contact" date from both the case and control groups.

There was evidence of reverse causality when including variables taken near the *E. coli* BSI sample collection. This was particularly evident for variables such as A&E visits, with 68% of individuals having an A&E visit ≤365d before *E. coli* sample collection when including information in the 72 hours before this, reducing to 45% if excluding all visits in the 72 hours before *E. coli* BSI sample collection (**Figure 4.24**). The impact of removing information in the 72 hours before the "most recent contact" date was much reduced for the control group, with 25% versus 22% of individuals having an A&E visit in previous ≤365d with and without including information in the 72 hours before the "most recent contact" date, respectively. A similar pattern was also observed when considering blood test results, with results being closer to normal ranges when excluding the 72 hours before the current contact in cases (e.g. for neutrophils in **Figure 4.25**).

To reduce the impact of reverse causality, all variables were calculated excluding the 72 hours before *E. coli* BSI collection for cases but such values were not excluded for the control group due to the little impact observed on the observed values of the risk factors.

**Figure 4.24: The percentage of cases and controls that had an A&E visit in the previous 365d ago split by inclusion/exclusion of information in the 72 hours before the "most recent contact date".**

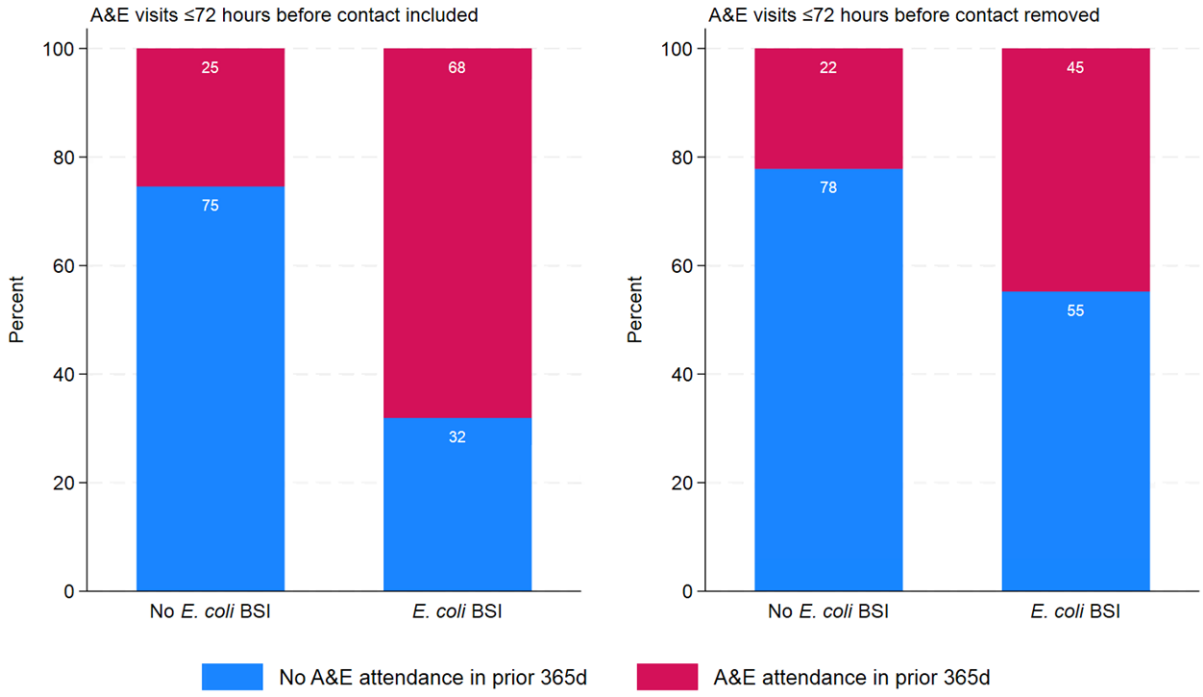
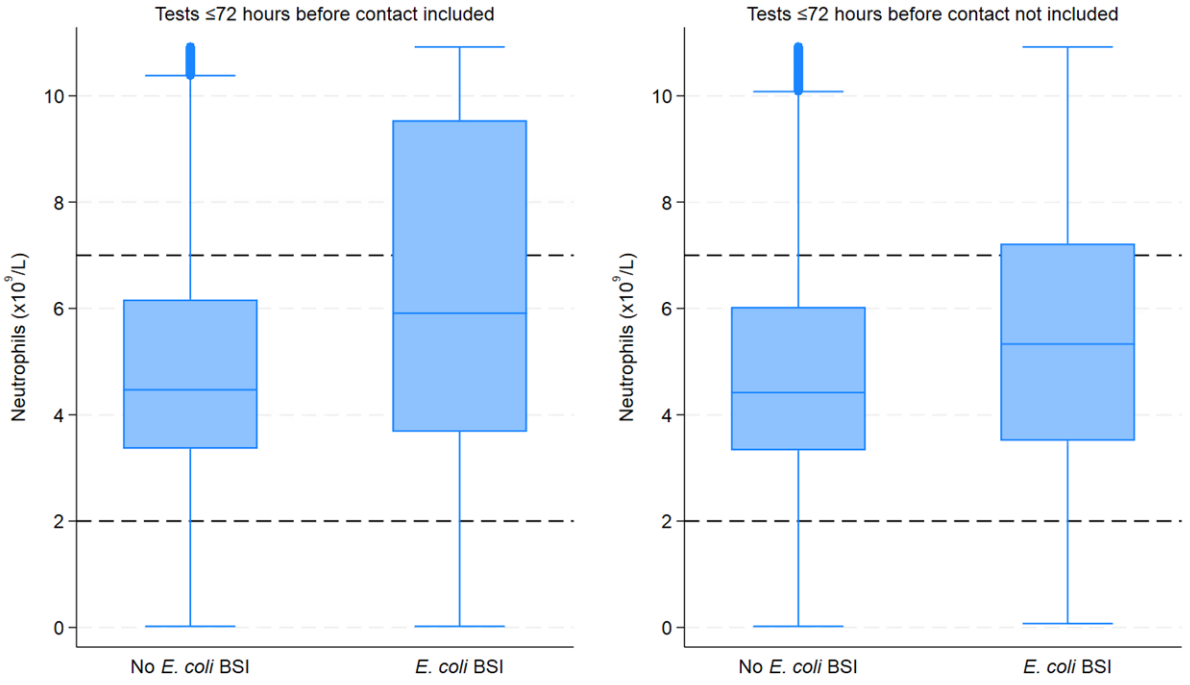



**Figure 4.25: The median (IQR) neutrophil value for cases and controls split by inclusion/exclusion of information in the 72 hours before the "most recent contact date".**

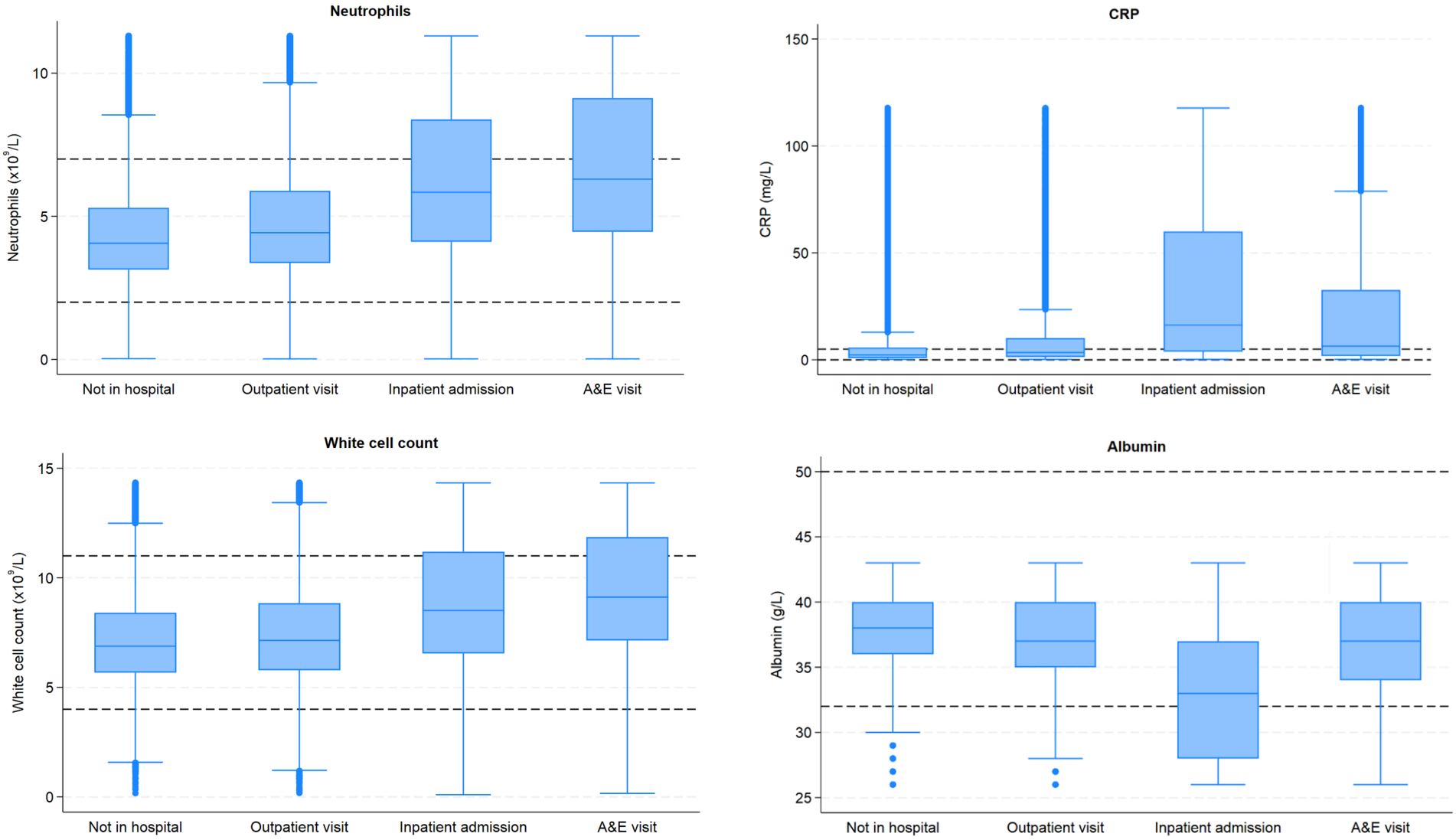



*Location where measurements were taken*

When taking the closest measurement to the “most recent contact” date, there were differences in blood test results dependent on the location where the blood was sampled (**Figure 4.26**). In general,

blood tests from samples taken outside of a hospital attendance were more likely to be within the normal range. Measurements from outpatient appointments showed a similar distribution to samples taken outside of hospital for most tests. Blood test results from samples taken within inpatient admissions and A&E visits were more skewed to distributions outside the normal ranges, for example for neutrophils, white cell count, and CRP. In contrast, albumin results from samples taken within A&E were more similar to those taken outside of hospital or at outpatient appointments. CRP had a wider IQR including higher CRP values for measurements taken in inpatient admissions compared with tests taken in other locations. Similar patterns were observed across all other blood tests (data not shown).

**Figure 4.26: Distribution of results from the closest blood test measurement to the "most recent contact date" in FY2019 for cases and controls for four blood tests, split by the collection location.**



After removing all tests in the 72 hours before an *E. coli* BSI sample collection, and preferentially choosing previous measurements taken outside of hospital, followed by those in outpatient appointments, the last measurement in an inpatient admission, and lastly A&E, the majority of blood tests selected to define risk factors were taken outside of hospital (**Figure 4.27**). Around 85% and 80% of cases and controls respectively used measurements outside of hospital contact, 10% from outpatient visits, ~3% and ~7% from inpatient admissions, and <1% and ~3% from A&E visits. The largest deviation from this was for urea, with only 54% and 29% of results coming from samples taken outside of hospital for cases and controls, respectively, and therefore a higher proportion of results came from outpatient visits, inpatient admissions, or A&E visits. Some tests including PSA, HbA1c, serum folate, and transferrin were seldom taken at A&E visits.

Vital signs recorded in IORD were all taken within hospital, and hence could not be taken outside of hospital attendances like blood tests. Most measurements used for the screening variables were taken within, or within ±72 hours, of inpatient admissions (**Figure 4.28**). Slightly more measurements were taken from outpatient visits for *E. coli* BSI cases compared with controls. The distribution was consistent across all vital signs. Whilst **Figure 4.27** and **Figure 4.28** only show results from FY2019, the patterns remained similar across all FYs (data not shown).

**Figure 4.27: Location where selected samples for blood tests were taken in FY2019.**

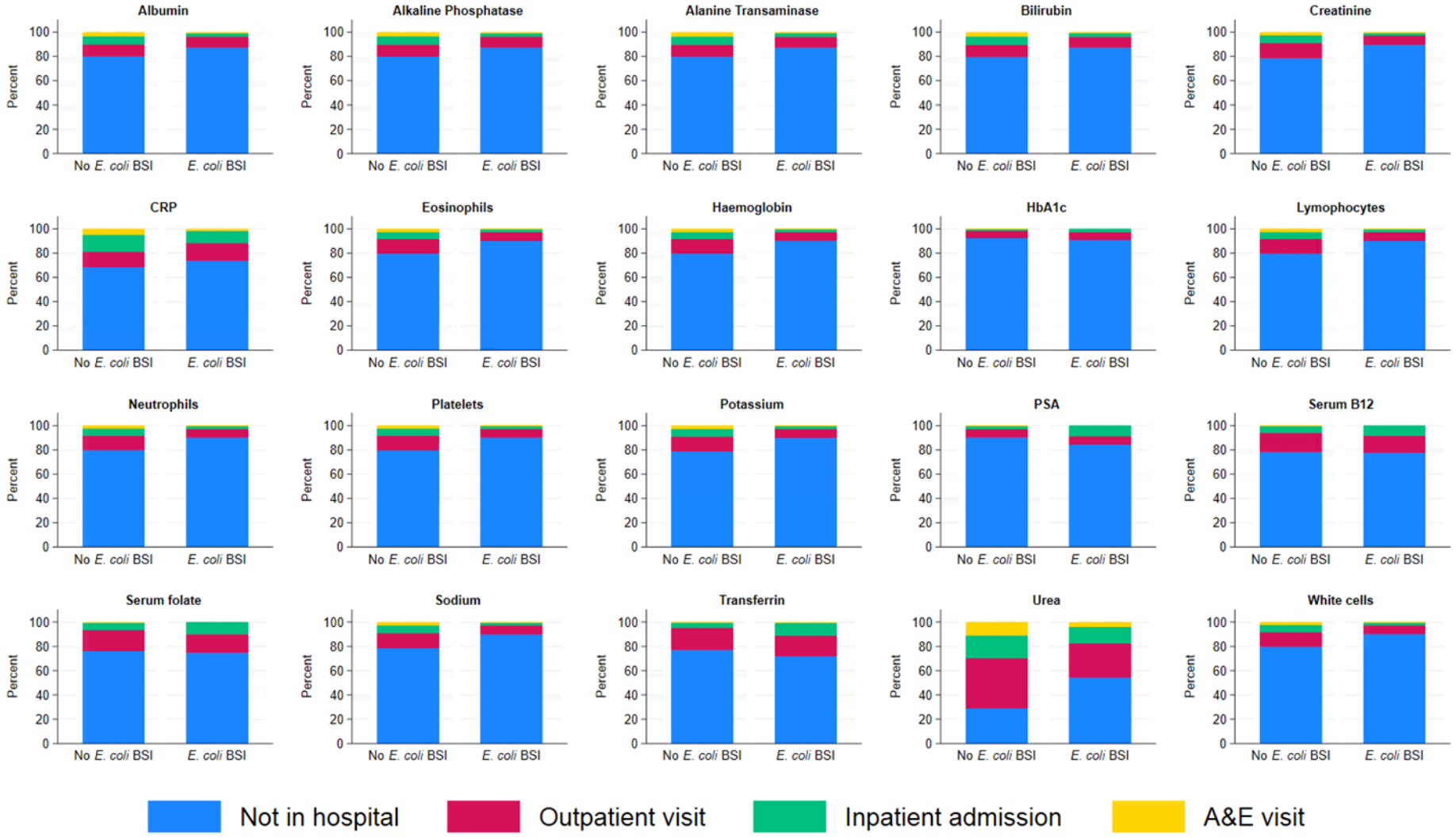
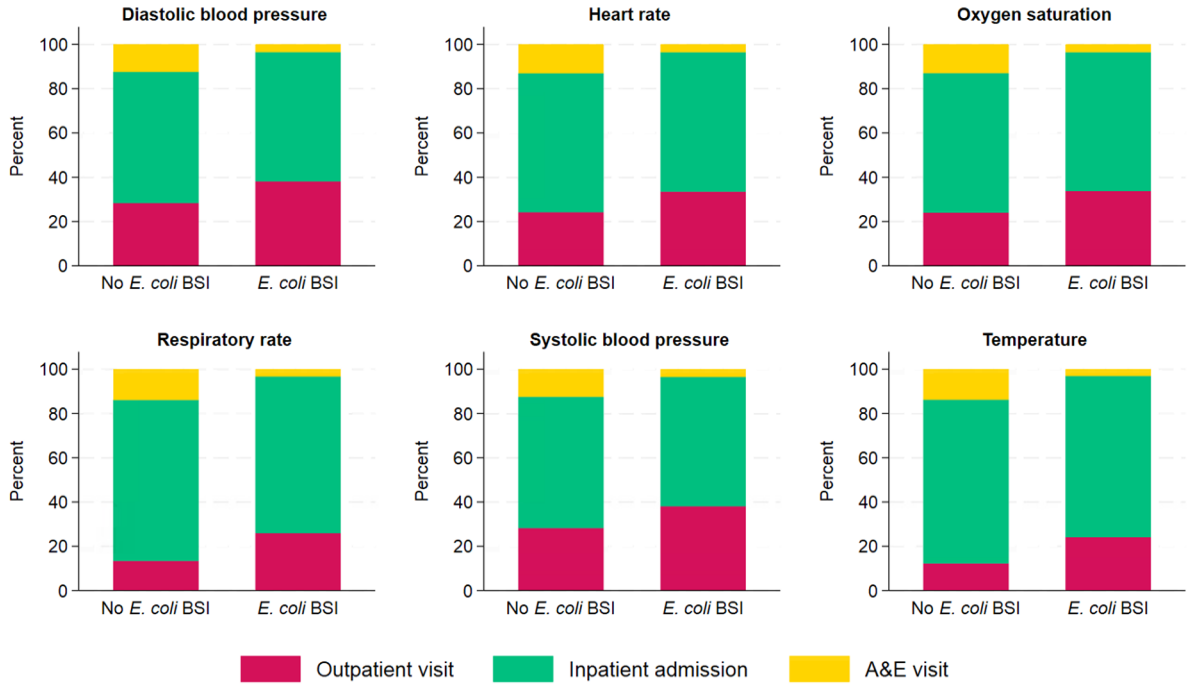


**Figure 4.28: Location where selected vital signs were taken in FY2019.**




*Summary of exposures*

Based on the timings and locations above, variables were created for all the case and control groups from FY2018 to FY2021.

As described above, by definition variables based on procedure and diagnosis codes reflect whether the code was recorded in the EHR and hence do not have missing data. 29 other variables (7%) had data missing in either cases or controls across at least one of the FYs (**Figure 4.29**). Controls had a higher percentage of missing data for all variables across all FYs. Blood tests had a varying amount of missing data with some tests having a lower amount of missingness (e.g. lymphocytes, neutrophils, platelets) and some tests having a very high amount of missingness (e.g. serum B12, serum folate). PSA had the highest amount of missingness, with the test only done in men being a contributing factor. Height had about 40% missingness in controls, contributing to a high number of people missing BMI. Weight was missing in around 15-20% of controls, but <5% in cases.

Year-on-year changes in the percentage of missingness for each variable were between -10% to +5% across all variables (**Figure 4.30**). Looking across the four years, the percentage of cases and controls missing vital measurements reduced the most, as above likely reflecting inclusion in the underlying data only from 2016 (**Figure 4.29**). There was a slight increase in the amount of missing data from FY2019 to FY2020 for the cases (this being evident for the majority of blood tests in **Figure 4.29** but not for vital signs measurements); however, the percentage of missingness decreased again from FY2020-FY2021.

**Figure 4.29: Percentage of missing data for screening variables with any missing data split by cases and controls and for each FY.**

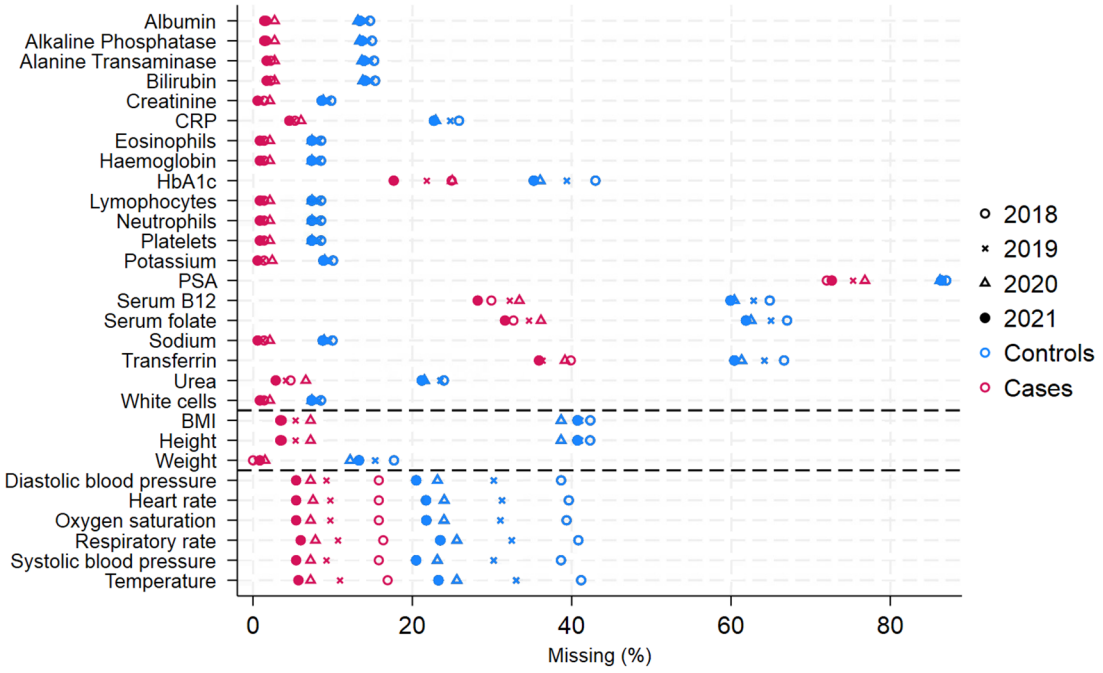
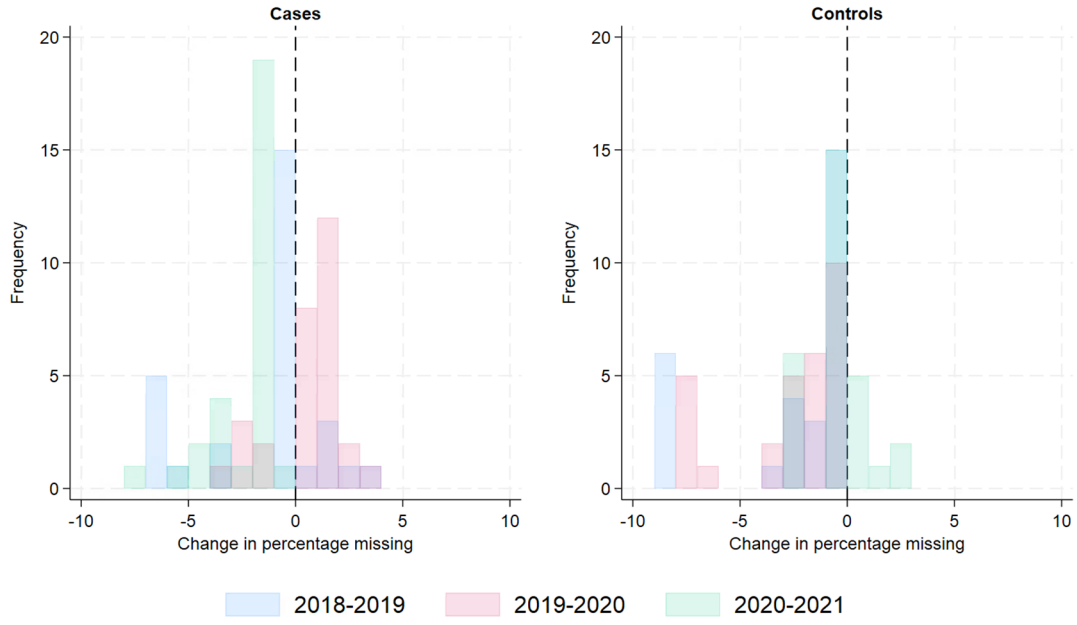



**Figure 4.30: Change in percentage missing for all variables in subsequent FYs.**

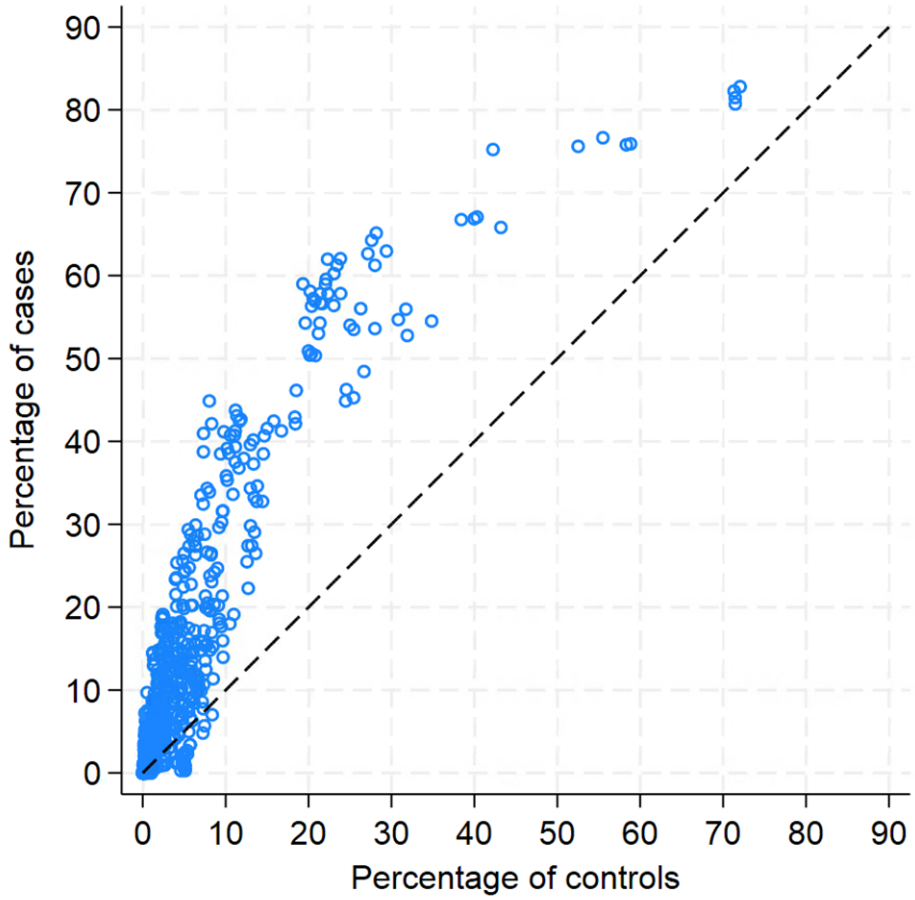



Note: A positive percentage change indicates that missing data increases from one year to the next. A negative percentage change indicates that missing data decreases from one year to the next.

Almost all characteristics were more common in the ≤365d before the "most recent contact" date in cases than controls (**Figure 4.31**). This was particularly evident for characteristics with >10% of cases having the characteristic ≤365d ago. The majority of variables had a prevalence of <20% in cases and

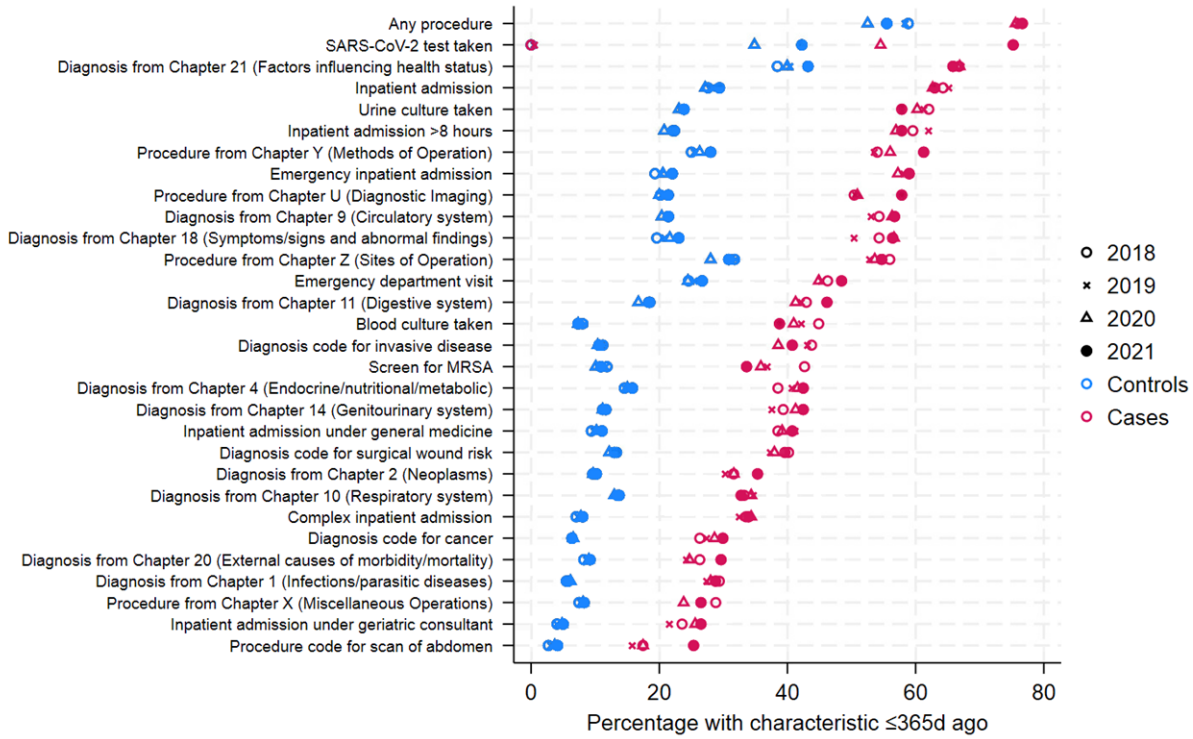
<10% in controls. The median (IQR) difference in percentage prevalence between cases and controls was 4% (0.6%, 12%). Characteristics were subsequently summarised by the largest percentage difference between cases and controls across all FYs.

**Figure 4.31: The percentage of cases and controls with each characteristic (n=176) ≤365d before the "most recent contact" date across all FYs (N=704).**



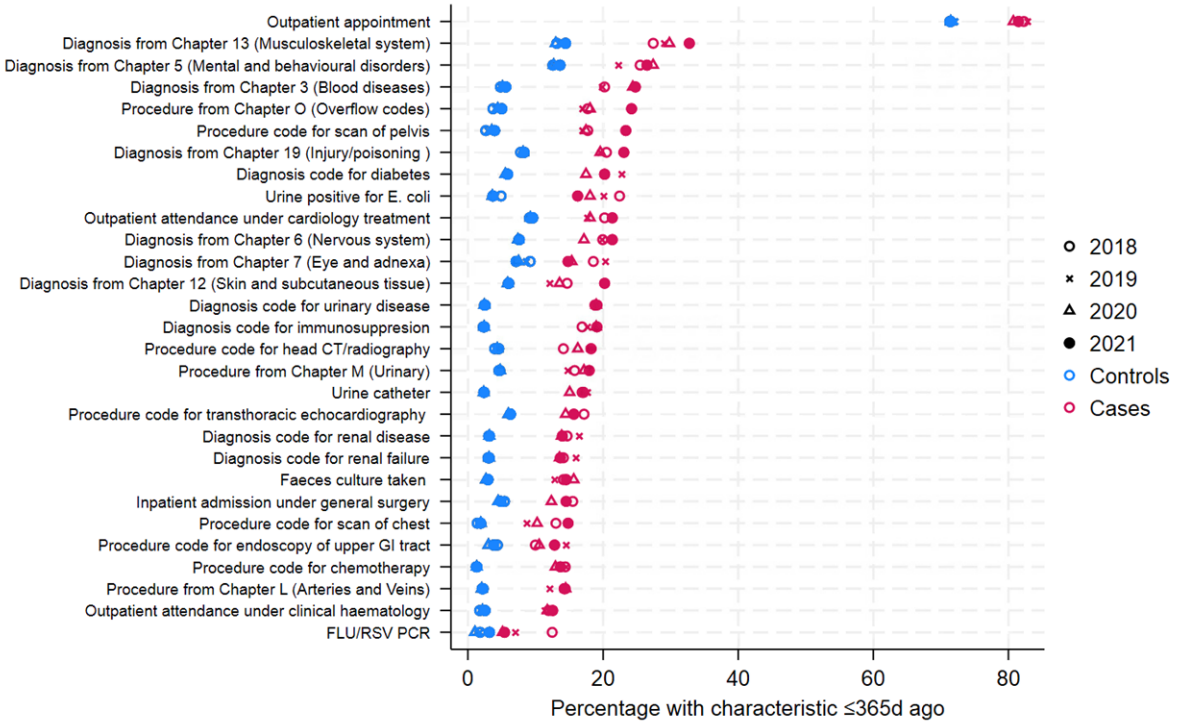

Thirty characteristics (17%) had a relative percentage prevalence in cases which was at least >20% higher than the prevalence in controls (**Figure 4.32**). Some characteristics referred generally to any hospital contact, such as any procedure code, admission >8 hours, or A&E visit. Some characteristics that were more common in cases referred to specific diseases, such as diagnosis codes for cancer or diagnosis codes from Chapter 1 (infections/parasitic diseases). Indicators of frailty were also more common in cases than controls, for example, inpatient admissions under geriatric consultants and diagnosis codes from Chapter 9 (circulatory system). Urinary disease indicators were common in cases with variables such as any urine culture taken (~60% cases vs ~25% controls) and diagnosis codes from Chapter 14 (genitourinary system) (~40% cases vs ~10% controls). Other potential risk factors, including a diagnosis code for surgical wound risk and procedure codes for abdomen scans, were also >20% more common in cases than controls. There was little variation between FYs apart from for SARS-CoV-2 tests, with no tests done in cases or controls in FYs 2018 and 2019 as expected.

**Figure 4.32: Variables with a percentage occurrence in the cases of >20% (absolute) more than in the controls.**




Twenty-nine characteristics (16%) had a prevalence between 10-20% greater in cases than controls (**Figure 4.33**). Having an outpatient appointment in the ≤365d ago was the most prevalent characteristic within this category, being prevalent in ~80% of cases and ~70% of controls. Again, many proxies for urinary disease were within this category, including urine positive for *E. coli*, diagnosis codes for urinary disease, procedure code from Chapter M (urinary), and use of urinary catheters. Specific diseases such as diabetes and diagnosis codes for renal failure were also in this group, alongside specific procedures such as chemotherapy and pelvis scans. Procedures indicative of frailty were also present such as computed tomography (CT) scan/radiography of the head (common after a fall). Diagnosis codes from multiple chapters were also in this group including Chapter 13 (musculoskeletal system), Chapter 3 (blood diseases), Chapter O (overflow codes) and Chapter 19 (injury/poisoning).

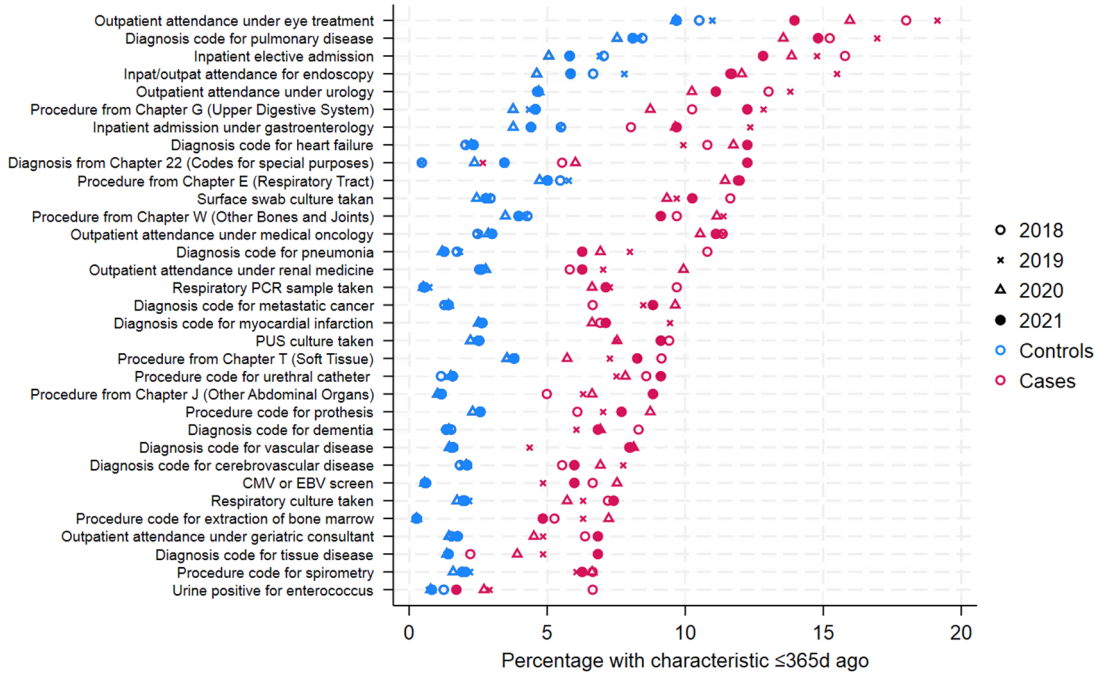
**Figure 4.33: Variables with a percentage occurrence in the cases of 10-20% (absolute) more than in the controls.**




Thirty-three characteristics (19%) were between 5-10% more common in cases than controls in at least 1FY (**Figure 4.34**). These characteristics all had <20% prevalence in cases and <12% prevalence in controls. A broad variety of characteristics were included in this group. Similar to the characteristics with higher prevalence presented above, there were indicators of cancer through outpatient attendance under medical oncology or metastatic cancer diagnosis codes and frailty identified through diagnosis codes for pneumonia or dementia.

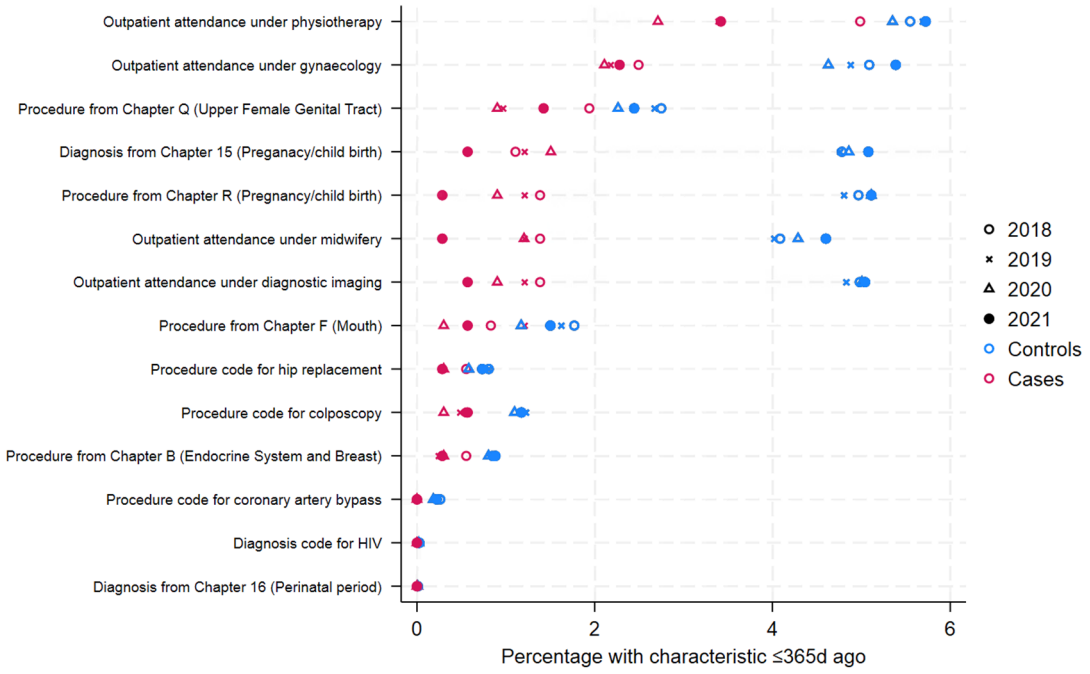
70 characteristics (40%) were at most 0-5% more common in cases than controls. These variables had median (IQR) 1.7% (0.7, 3.6) and 0.5% (0.2, 1.5) prevalence in cases and controls, and hence were not very common in either population. Whilst I have not plotted these differences, these variables are included in the screening model moving forward to the next Chapter.

**Figure 4.34: Variables with a percentage occurrence in the cases of 5-10% (absolute) more than in the controls.**



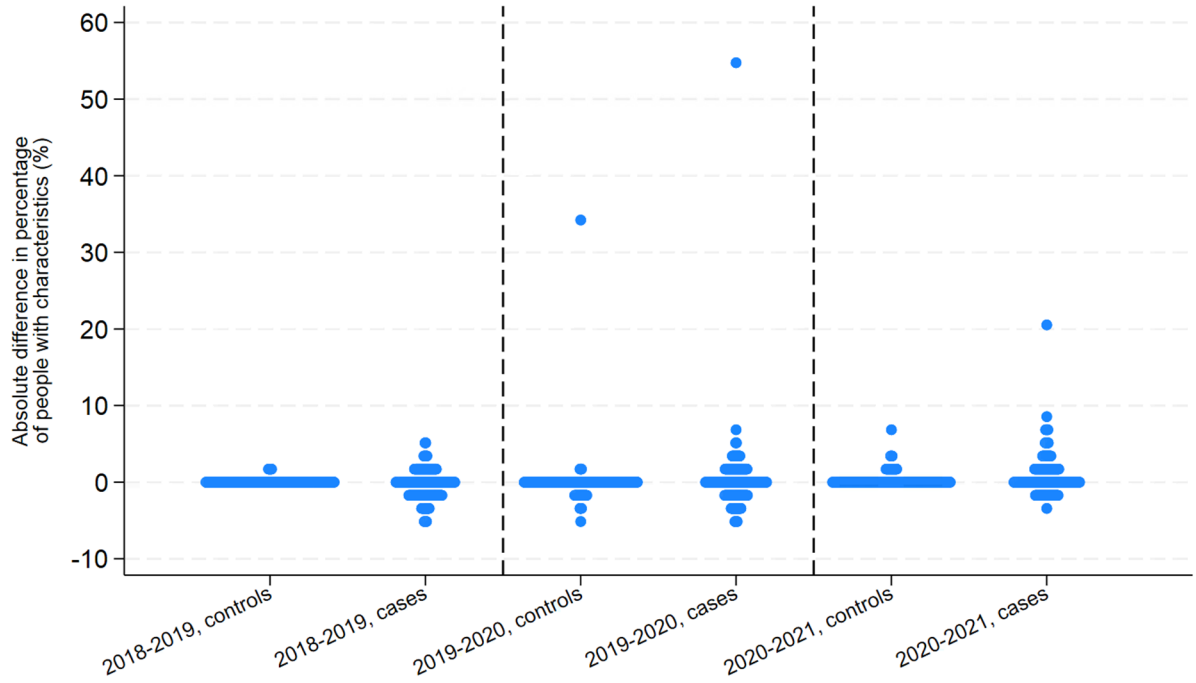

Fourteen characteristics (8%) were consistently more common in controls than cases across all the FYs (**Figure 4.35**). Many of these characteristics were associated with pregnancy and/or childbirth, including diagnosis codes from Chapter 15 (pregnancy/childbirth), procedure codes from Chapter R (pregnancy/childbirth) and Chapter Q (upper female genital tract), or outpatient attendances under midwifery. All these characteristics were <6% prevalent in cases and controls.

**Figure 4.35: Variables with a higher percentage (absolute) occurrence in the controls than cases.**



Prevalence of most exposures remained stable over the four FYs years studied, with mostly between ±10% change in prevalence in cases and controls (**Figure 4.36**). There were three occurrences of a difference >10% between consecutive FYs. This constituted COVID-19 tests having a 54% and 34% increase between FY2019 to FY2020 in cases and controls, respectively, and a 21% rise for cases between FY2020 to FY2021 (compared with a 7% increase in controls from FY2020-2021). The distribution of percentage change was generally tighter in controls than in cases due to changes in the smaller number of cases having a larger effect on the absolute percentage prevalence compared with the larger number of controls.

**Figure 4.36: The absolute change in percentage prevalence in characteristics recorded ≤365d ago between consecutive FYs.**

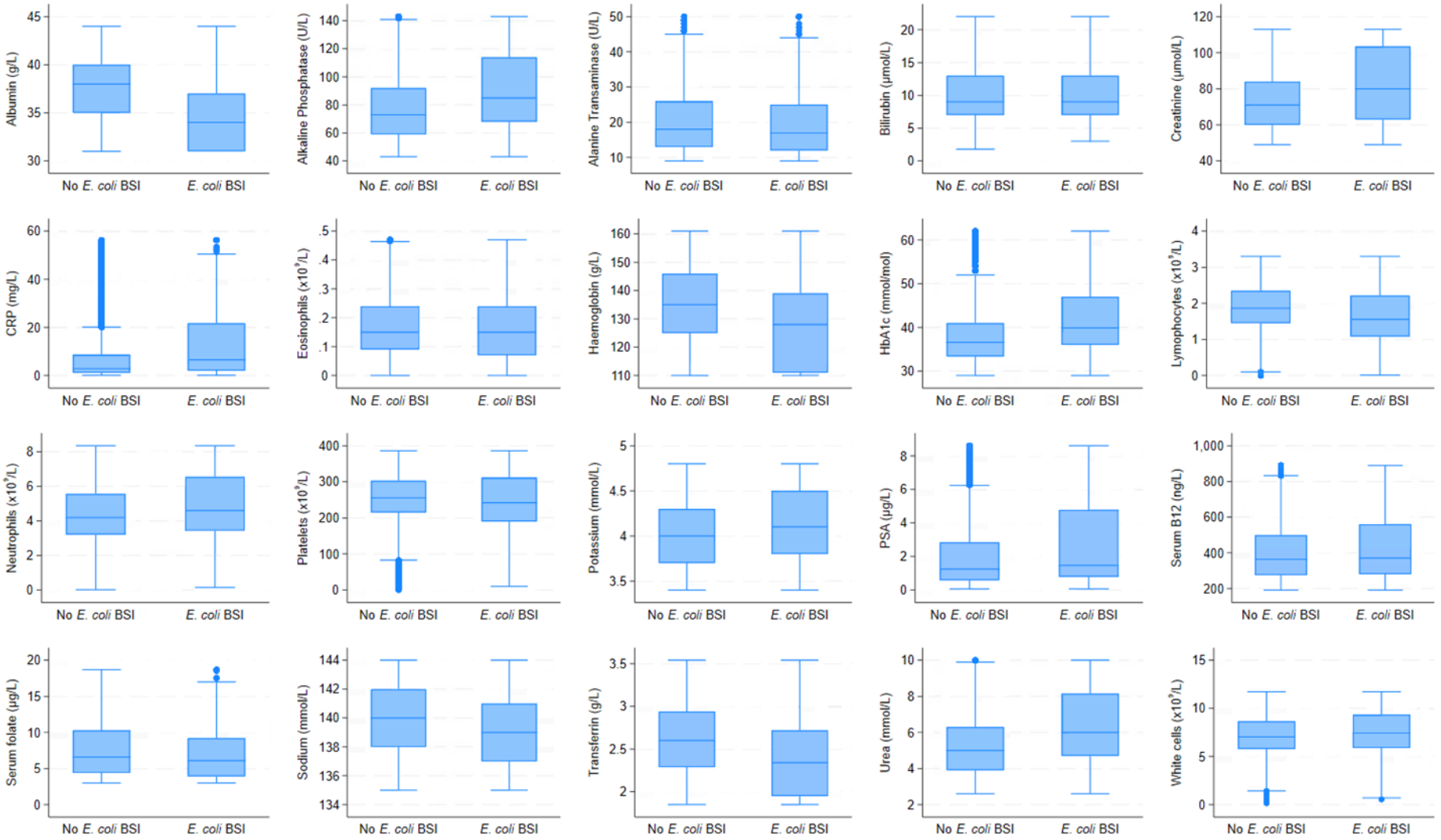



All variables summarised so far were included as time since the most recent occurrence ≤365d ago (continuous) and ever/never present (categorical). In contrast, the values of blood tests, vital signs, and height/weight/BMI were only included as continuous variables.

For cases and controls with blood test results, values from cases were sometimes more skewed toward abnormal values; however, often the distributions of variables summarised by medians and IQRs looked quite similar (**Figure 4.37**). The median and 75th percentile were higher for CRP in cases than controls and above the normal range of 0-5mg/L. Haemoglobin was, on average, lower in cases than in controls, as were albumin and transferrin. The following tests were, on average, higher in cases than controls: alkaline phosphatase, creatinine, HbA1c, PSA, and urea. The distribution of

eosinophils, bilirubin, and white cell count was similar between cases and controls. The distribution was similar across all years and hence only presented for FY2019 for simplicity.

**Figure 4.37: Distribution of blood test results for FY2019, split by cases and controls.**



Vital signs were relatively similar between cases and controls, with the distribution of heart rate being slightly higher in cases than controls and oxygen saturation being distributed, on average, slightly lower in cases than controls (**Figure 4.38**). Height, weight, and BMI had similar distributions between cases and controls (**Figure 4.39**).

**Figure 4.38: Distribution of vital sign results for FY2019, split by cases and controls.**

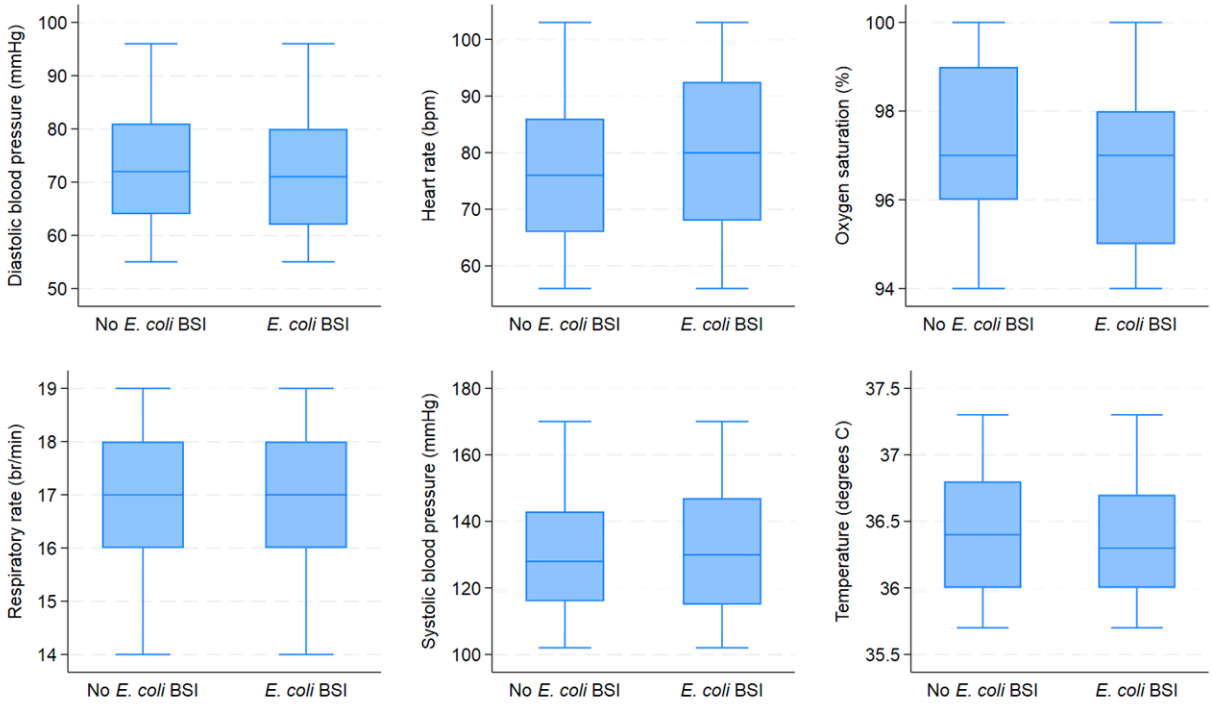
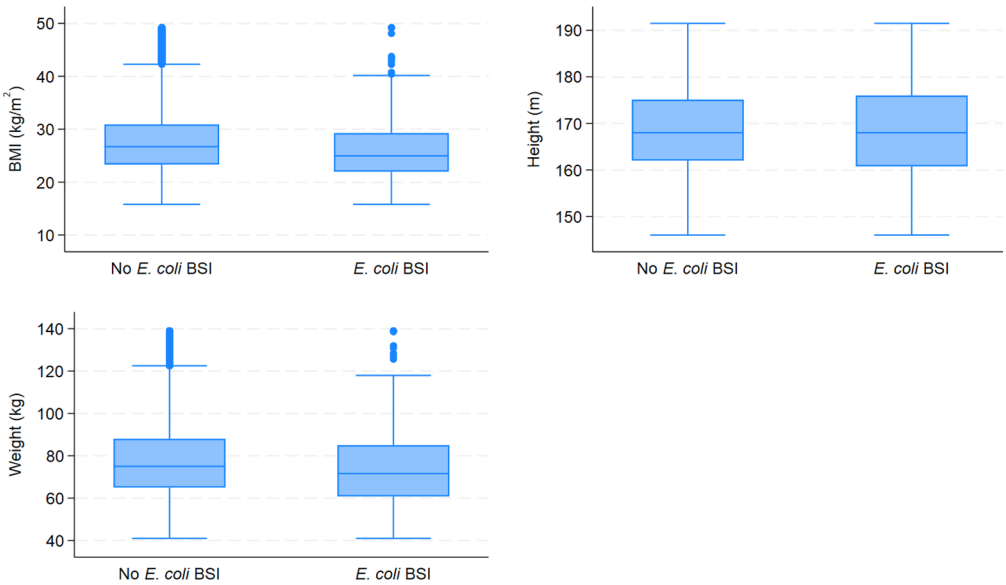



**Figure 4.39: Distribution of personal trait results for FY2019, split by cases and controls.**



## 4.4 Discussion

In this Chapter, I investigated how differences in control group and exposure definitions impacted estimated associations between *E. coli* BSIs and risk factors. When defining a control group, I found that including individuals with any type of previous healthcare contact (the "any healthcare" cohort, including outpatient, A&E, blood test and microbiology samples) yielded more missing data for core characteristics, such as ethnicity. Missingness was also systemic for many potential risk factors in the "any healthcare" cohort. For example, only vital signs recorded in inpatient admissions are available in IORD; diagnosis and procedure codes were well documented in inpatient admissions but inconsistently recorded in outpatient appointments and not recorded in A&E visits. Missingness of variables also varied by length of lookback, with the proportion of missing data decreasing with increased lookback for some variables such as HbA1c, and missing data increasing with increased lookback for other variables such as vital sign measurements which were only consistently recorded from 2016 onwards in IORD. This was most evident for vital sign results where measurements were electronically recorded from 2016 onwards, hence longer lookback periods reduced missingness substantially in later FYs. There were small but insubstantial shifts in model estimates when increasing the length of lookback from one-five FYs. Model estimates for "core" variables were also broadly similar between the "inpatient only" and "any healthcare" cohorts, with similar interpretations for all variables across all FYs. Due to the lower amounts of missing data, I decided to focus on the "inpatient only" cohort for the primary analysis in the next Chapter and use the "any healthcare" cohort as a secondary analysis. I decided to use a 5FY lookback to allow a longer amount of time to capture all variables and maximise the number of individuals included.

When deriving all the risk factors for analysis, I found that the proximity of variables to the inclusion date was important to consider. Taking measurements in the 72 hours before *E. coli* BSI sample collection resulted in reverse causality with variables known to be associated with *E. coli* BSIs, such as attendances at A&E being over-represented in the case group and higher CRP levels. Removing results in the 72 hours before the current contact date made no difference to controls. I therefore decided not to consider any characteristic recorded in the 72 hours before the current contact (index blood specimen collection date/time) in cases in the next Chapter but not to apply this exclusion to controls. For blood tests, the timing of measurements within an inpatient admission somewhat impacted blood test results, with results falling closer to the "normal range" if they were the last result in an admission, compared with any other time in an admission. For blood tests, the location where the test was carried out more strongly impacted the value of the measurement. Blood test results from samples taken in outpatient settings or sent from GPs were more likely distributed within the "normal range", compared with those measured within inpatient admissions or A&E.

Similarly, for vital signs, measurements taken at outpatient appointments were more within the "normal range" than the measurements taken in A&E and inpatient admissions. I therefore prioritised measurements from GP and outpatient settings where available and used the last measurement in a previous inpatient admission or a measurement from A&E otherwise.

### 4.4.1 Control group choice

Choosing who to include in the control group was a balance between trying to get as close to the target population of Oxfordshire while reducing the amount of missing data in risk factors and not including people who would not have come to an Oxfordshire hospital had they got an *E. coli* BSI. Only having access to data from one hospital trust in England risked missing hospital attendances at other hospitals. It would be reasonable to assume that some patients, particularly those living at the border of Oxfordshire, might attend multiple NHS trusts for their healthcare and therefore their records might not be in IORD if they attended elsewhere. In a previous study, an estimated 25% of individuals who had at least two inpatient, outpatient, or A&E encounters in England in 2017-2018 attended two or more NHS trusts; however, there were large regional variations in this estimate.[191] This study did not provide separate percentages for each NHS Trust, but it noted that 54% of individuals who attended multiple Trusts visited 20 pairs of hospitals. Oxford University Hospitals NHS Foundation Trust was not one of these, so while patient movement between Trusts affects data, the impact of this is IORD may be limited. It may also be expected that some patients from outside of Oxfordshire attend one of the Oxford University Hospitals, for example when travelling into the area for a holiday, or to attend for specialist care such as a cardiac specialist service, or again those living at the border of Oxfordshire.[192] *E. coli* BSIs from these patients may therefore be missed. I attempted to account for this variation in healthcare attendance by requiring individuals to have contact in the current year and I also adjusted for catchment percentage in the core model to help account for this variation.

There were differences in the distribution of characteristics between IORD and Census data, particularly with a higher proportion of older age groups in IORD. This would be expected as the risk of many illnesses increases with age; in 2016-2017 in England, 41% of admissions were in those aged >65y, compared with 24% in those aged 15-44y.[193] There was also a higher proportion of people from deprived backgrounds in IORD compared with the census data. Higher healthcare attendance in those experiencing higher levels of deprivation has been observed elsewhere, such as higher levels of deprivation being associated with higher odds of attending A&E in England in FY2021,[194] perhaps caused by limited access to primary care services in more deprived areas.[195] The proportion of people in IORD aged 51-60y was very similar to the proportion in the census, perhaps due to many of these individuals attending the GP every 5 years for the recommended NHS Health Check

for all those aged 40-70y.[196] The proportion of women aged 31-40y was also very similar, perhaps indicative of women visiting healthcare services for pregnancy and childbirth.

While proving that collider bias has not impacted model results is generally not possible,[168] I attempted to assess its potential impact on the results of this study through comparisons of the distribution of available core variables with the target population of the whole of Oxfordshire, as extracted from census data. Having an unrepresentative sample and then drawing inferences to a larger population can introduce bias.[197] It is challenging to assess whether collider bias impacts the data further as no information is available from EHRs on the individuals who have no contact with the healthcare system. I compared model estimates between the "inpatient only" and "any healthcare" cohorts to assess whether selection bias could influence the "inpatient only" cohort results, showing that there was little difference between model estimates of the core variables. As most *E. coli* BSI cases result in hospitalisation, there will be few individuals with *E. coli* BSIs who do not have contact with the healthcare system and therefore selection into the study for cases is primarily based on the occurrence of *E. coli* BSIs. This may reduce, but not eliminate, the risk of collider bias. Studying a disease which does not almost exclusively result in hospitalisation may result in a larger impact on model results and introduce more selection bias. For example, if studying risk factors for COVID-19 in EHRs, the analysis will be restricted to only those hospitalised with COVID-19 which will be a much smaller subset of all those with COVID-19[168] (an estimated 2.1% of people with infection, as estimated by one study in the US[198] but likely to vary by variant, background prevalence, and individual characteristics such as age and frailty). This may need to be considered further if expanding the methodology in this study to other diseases. When acknowledging the potential impact of selection bias in this study, it may be more appropriate to interpret risk factors conditional on the cohort used for analyses. For example, risk factors found in the "inpatient only" cohort should be considered risk factors within the sub-population of those who have had an inpatient admission in the last 5 years.

### 4.4.2 Risk factor definitions

In this study, I found that the timing of measurements relative to blood culture collection/last contact date impacted values, as well as the location where the measurements were taken.

I found that the proximity of measurements to blood culture collection was an important consideration to avoid reverse causality for cases, whereby abnormal values are a consequence of being a case, rather than a cause. A 72-hour window before blood culture collection was established based on clinical advice and the knowledge that blood cultures will be requested based on abnormal clinical observations therefore inducing reverse causality. For different diseases, different cut-offs

may have been more appropriate. I included risk factors from inpatient episodes in the same spell (a continuous hospital stay made up of contiguous consultant episodes) as a positive *E. coli* blood culture was collected, provided the episode ended >72 hours before the *E. coli* BSI collection. As inpatient admissions are coded after a patient has been discharged, the *E. coli* BSI may impact what diagnoses were coded in earlier episodes. However, I deemed it more important to include these risk factors close to the *E. coli* BSIs to avoid missing potentially important signals. A sensitivity analysis could compare risk factors found with and without including codes from these inpatient episodes in the same admission as the *E. coli* BSI, but finishing at least 72 hours before the index sample collection.

I found that blood test measurements taken at the GP or outpatient appointments were more closely distributed to normal ranges, compared with blood tests taken at inpatient admissions or A&E visits. This was expected as A&E is primarily for those with serious and/or life-threatening injuries, while GP services are for less severe health concerns. There was a particularly high distribution for CRP levels in inpatient admissions, compared with any CRP measurement taken outside of hospital or in outpatient appointments. As a marker for inflammation, high CRP can be present in many illnesses, predominantly infection, but also including rheumatologic diseases, malignancy, and drug reactions;[199] common reasons for inpatient care. While all measurements were taken in healthcare settings, I prioritised measurements taken in locations where patients are usually in a healthier state (GPs, outpatients rather than inpatients, A&E) to avoid including measurements taken at the time of acute illness and hence be a better proxy for a "baseline" background level. Some patients were also missing blood test values, particularly for blood tests that are generally taken when a doctor is concerned about a particular illness (e.g. PSA) rather than tests which are routinely done. Machine learning and traditional statistical methods have been evaluated for use in imputing unmeasured values for blood test measurements in EHRs; however validating these methods is challenging and would need further work, as well as considerable computing resource/time.[200]

While associations between risk factors and *E. coli* BSIs will be modelled in the following Chapter, initial summaries of the raw data indicate large differences in the proportion of individuals with risk factors between cases and controls. The majority of risk factors were more common in cases than controls. This was expected as I defined many risk factors from previously published research considering risk factors for infections, for example, risk factors for mortality associated with *E. coli* BSIs[174] and risk factors for healthcare-associated infections.[62] Many of the risk factors which were more common in cases than controls related to increased hospital contact through inpatient admissions, including complex and ordinary admissions, as well as more procedures. Many of these

risk factors are likely to be confounded by age to varying degrees and therefore multivariable modelling in the following Chapter will be essential. A small number of variables were more common in the control group, with the majority of these being related to pregnancy and childbirth. The control group was likely to include a high proportion of these individuals as, for example, for females, those aged 30-34y had the highest number of attendances for outpatient appointments in England in 2020-2021, with 43% of female outpatient attendances aged 20-39y attending maternity services.[201]

The proportion of cases with a urine culture taken regardless of the result was >20% higher in cases than in controls. Having access to negative microbiology tests, meaning that risk factors could be created including having any urine sample or blood sample collected regardless of result, was a strength of this study. Presence of a test result has previously been found to be predictive of ill health for blood tests, irrespective of the result of the test,[172] but not for microbiology (to my knowledge) and therefore exploring their impact in the following Chapter will be interesting.

### 4.4.3 Further considerations

The COVID-19 pandemic may have impacted model results in FY2020. In FY2020, there was a lower risk of *E. coli* BSI in those of non-white ethnicities while in FY2019 and FY2021 there was no evidence of any effect of ethnicity on the outcome. This could have been due to competing risks, with those of non-white ethnicities more likely to suffer COVID-19 mortality. Increased risk of severe COVID-19 disease in non-white ethnicities has been shown globally, with those of Black and Hispanic ethnicities at a higher risk of COVID-19 hospitalisation and intensive care unit (ICU) admissions being highest among Black, Hispanic, South Asian, East Asian and Mixed ethnic groups and Indigenous peoples.[202] Due to the relatively small number of *E. coli* BSIs and the small proportion of individuals of non-white ethnicities in Oxfordshire, I was not able to explore associations between ethnicity and *E. coli* BSIs in more granular detail.

In this Chapter, the main focus has been on how to code exposures when doing a risk-based analysis; however, how to code the outcome is also an important consideration. I defined the presence of *E. coli* BSIs using microbiology isolations. Blood cultures usually have a high positive predictive value of infection (truly positive, given a positive test result) due to blood being a sterile site,[203] with contamination risk through bacteria on the skin entering the culture during extraction being low for *E. coli*.[204] If access to microbiology samples is not available, ICD-10 codes alone may not be good enough to identify positive BSIs as ICD-10 codes are used for billing purposes and therefore may not accurately record specific patient information.[172] Further, variations of the outcome may be important for future studies. In the following Chapter, I will consider *E. coli* BSIs split by where they

were acquired (e.g. nosocomial, community infections), as well as look for risk factors specific to antibiotic-resistant infections.

### 4.4.4 Limitations

The main specific limitation of my study is that it was conducted using a single population from Oxfordshire. While the sample was large and accounted for around 1% of the UK population, generalisability may be limited. Furthermore, ICD-10 and procedure coding practices can vary between hospital trusts and therefore exposure definitions may need to be considered further when expanding to different settings. Different hospitals may also use different laboratory equipment over time so care would have to be taken when using continuous laboratory test results in other settings. Only a small number of people experienced the outcome in this study, reducing power. Expanding the work to national-level data would increase both the power and generalisability of the results. This study was limited by using secondary care data with no access to primary care records, meaning for example only vital signs taken in inpatient admissions, outpatient attendances and A&E visits could be included. Access to GP records may have allowed a larger proportion of the Oxfordshire population to be studied and also allowed more risk factors to be defined. The OpenSAFELY platform could be an alternative source of community data however does not include hospital-requested microbiology data and hence would miss microbiologically confirmed *E. coli* BSIs.[49] Lastly, the definitions of cohorts and exposures established in this study have been investigated specific to the outcome of *E. coli* BSIs; however, with the very large number of controls, distributions of exposures are likely to be similar to other control populations. Assessing the impact of the different cohorts would be important if studying a different disease; however, the framework established in this Chapter may offer a pragmatic way to do so. While the exposures defined here are *E. coli* BSI specific, many would be relevant risk factors for other diseases and hence may be useful for future studies.

### 4.4.5 Conclusions

Overall, care and consideration had to be taken in defining control groups and exposures to avoid introducing bias into this study. Future work focusing on the impact of the control group choice on other diseases may be helpful going forward, as well as studies assessing the difference between pre-defined risk factors versus taking a more general approach.

# Chapter 5 Monitoring populations at an increased risk of E. coli bloodstream infections using Electronic Health Records

## 5.1 Introduction

*E. coli* bloodstream infections (BSIs) are the leading Gram-negative bloodstream infection in the UK, with 38,380 cases reported from mandatory surveillance reports in 2022.[61] While *E. coli* BSIs have been increasing over the last decade,[140,205] in the UK there was a reduction in the number of cases from 76 to 68 cases per 100,000 population from 2018 to 2022,[61] coincident with the COVID-19 epidemic. However the number of infections in the population remains high, and numbers have subsequently increased again. *E. coli* BSIs can cause severe illness and death, with an estimated case fatality rate of 15.9% in England during FY2022 to 2023.[206] Reducing the number of *E. coli* BSIs could therefore reduce illness and potentially save lives.

A further motivation to reduce the number of *E. coli* BSIs is to help reduce the burden of antimicrobial resistance (AMR). *E. coli* BSIs are a key target for the reduction of AMR as 82% of antibiotic-resistant BSIs in the UK in 2022 were *E. coli* BSIs.[61] One study estimated that, in high-income countries, around 23% (95% CI: 20%-28%) of deaths from AMR were linked to *E. coli* BSIs.[207] While an increased mortality rate has been reported for AMR *E. coli* BSIs compared with non-AMR *E. coli* BSIs,[208,209] others have found no evidence of this[210] but have instead found associations with other adverse outcomes such as an increased length of hospital stay.[211] To combat the increasing threat of AMR, in 2019, the UK Government provided a 5-year[212] and 20-year[213] national action plan (NAP), and most recently a second 5-year NAP covering 2024-2029,[214] aiming to reduce infections and optimise antimicrobial use. The 20-year vision outlines nine ambitions for change, one of which is to reduce the number of infections overall. With fewer infections to treat, antibiotics can be used more sparingly and therefore reduce the risk of emergence of new resistant strains.[215]

*E. coli* BSIs are currently monitored through the UKHSA mandatory surveillance programme in England. Surveillance of *E. coli* BSIs has been mandatory in England since June 2011 after increases were observed through voluntary surveillance programs.[56] Annual reports have been produced since 2013 covering the number of *E. coli* BSIs stratified by age and sex, and also whether the BSIs were resistant to a range of antibiotics.[61,216]

One gap in the current surveillance programme is the small number of risk factors collected alongside infection data. Beyond age and sex, the populations who are most at risk of having *E. coli* BSIs are not routinely monitored. Ad hoc studies have assessed risk factors for *E. coli* BSIs in a general hospital population.[217,218,219] Two of these studies, based in the US and Wales (UK), used a

control population,[218,219] while the third compared the proportion of risk factors in *E. coli* BSIs with the general population using registry data.[217] All these studies found the populations at highest risk included individuals on dialysis or having experienced renal disease or failure. Cancer was also found to be a risk factor[217,218], as well as urinary catheterisation and urinary incontinence,[218] urinary tract infections,[219] and higher comorbidity scores.[219] However, none of these studies considered how to continuously monitor these risk factors over time. While some risk factors may be consistent, it is also reasonable to assume that risk factors could change over time as population composition changes and new diseases, such as COVID-19, are present in the population.

As discussed in the previous Chapter, electronic health records (EHRs) offer a wealth of data in which to find populations at risk of infection. The large majority (90%) of hospitals in the UK now have electronic patient records,[24] and approximately 80.5% of hospitals in the USA,[220] for example. As laboratory results, vital sign measurements, and blood test results are now often automatically uploaded to the electronic system, and reasons for inpatient admissions are coded by teams of coders after hospital discharge, data in EHR should be relatively up-to-date. Using up-to-date data should allow risk factors to be monitored regularly, for example running a screening process annually to assess at-risk populations in the previous year. A more extensive overview of the use of EHRs for surveillance can be found in the overall Introduction to this thesis.

This Chapter aims to implement the learning in Chapter 2 and Chapter 4 to identify populations at increased risk of *E. coli* BSIs in an EHR population. The screening process developed in Chapter 2 will be implemented and the control groups and risk factors developed in Chapter 4 will be used to investigate those most at risk of *E. coli* BSI. I will train the screening process on one year of data and then test the process on multiple years of data to assess whether it could be used in a near real-time scenario. I will also consider the location of *E. coli* BSIs onset (e.g. nosocomial or community) on risk factors found. Risk factors for antibiotic-resistant *E. coli* BSIs will also be considered separately. Overall, I will report risk factors for *E. coli* BSIs across four years of data from 2018-2022.

## 5.2 Methods

As in Chapter 4, I used data from the Infections in Oxfordshire Research Database (IORD): a large dataset including inpatient admissions, outpatient appointments, and accident and emergency (A&E) attendances, as well as microbiology and biochemistry/haematology test results from samples sent from both within hospital and general practice. Inpatient admissions and outpatient appointments can be linked with diagnosis codes, procedure codes, and vital sign measurements taken during hospital attendance. IORD includes four large teaching hospitals covering a catchment area of around 660,000 individuals. The dataset goes back to 1997 with electronic recording of most variables, for example, diagnosis codes, from 2007 onwards. Electronic recording of vital sign measurements and medications began in 2016.

The screening process was tested and developed on data from the financial year 2019 (FY2019; 1st April 2019 – 31st March 2020) and then expanded to run on FY2018, FY2020, and FY2021. The risk factors selected from the process were then compared across the years.

### 5.2.1 Case definition

As described in detail in Chapter 4, cases were defined using microbiological isolations and de-duplicated within 90 days from the index positive. The first *E. coli* BSI in each financial year was selected if individuals had multiple BSIs recorded.

### 5.2.2 Defining the "most recent contact" date

To be considered in subsequent analyses, individuals had to have healthcare contact (inpatient admission, outpatient appointment, A&E visit, or any microbiology sample or blood test taken) recorded in IORD in the current financial year. For cases, this contact would be the collection of the index blood sample which tested positive for *E. coli*. For controls, this was one of the following: a discharge date from an inpatient admission, an appointment date from an outpatient appointment, an arrival date from an A&E visit, or a collection date from a blood test or microbiology sample collection.

One record was selected per person per financial year, this being the first positive *E. coli* BSI if present, or the last healthcare contact otherwise. The date of this contact is subsequently referred to as the "most recent contact" and defines the dates on which cases and controls are considered to be comparable.

### 5.2.3 Defining the analysis cohorts

Ideally, the control group would include everyone in Oxfordshire; however, IORD only includes a subset of individuals who have had healthcare contact within Oxford University Hospitals (as defined

above). Further, many risk factors of interest can only be calculated from data recorded in inpatient admissions, for example, diagnosis codes, or vital sign measurements. To account for the fact that individuals with no previous or current inpatient admission would have missing data for many characteristics, two analysis cohorts were subsequently defined based on previous and current healthcare contact recorded in IORD, as discussed in Chapter 4 above.

One cohort was based on previous inpatient contact, subsequently referred to as the "inpatient only" cohort, to minimise the amount of missing data for risk factors. The second cohort was defined conditional on having any previous healthcare contact (not just inpatient admissions), subsequently referred to as the "any healthcare" cohort, to allow the retention of as many individuals as possible for analysis. A 5FY lookback period was used for both cohorts. A detailed description of these two cohorts is described above in the Methods of Chapter 4 and a summary of the two analysis cohorts is provided in **Table 5.1**.

**Table 5.1: Summary of the populations included in the two analysis cohorts.**

| | **Inpatient only cohort** | **Any healthcare cohort** |
|---|---|---|
| **Cases** | All individuals with any inpatient episode ending >72 hours before the collection date/time of an *E. coli* BSI, and ending within the previous 5FYs. | All individuals with any previous inpatient episode (as defined for the "inpatient only" cohort), or any previous outpatient appointment, A&E visit, blood test, or microbiology sample collection >72 hours before the collection date/time of an *E. coli* BSI, and within the previous 5FYs. |
| **Controls** | All individuals with any inpatient episode within the last 5 FYs ending before or at the "most recent contact" date. | All individuals with any inpatient episode (as defined for the "inpatient only" cohort), or any outpatient appointment, A&E visit, blood test, or microbiology sample collection strictly before the "most recent contact" date, and within the previous 5FYs. |

### 5.2.4 Defining risk factors

The same set of six key potential confounders outlined in Chapter 4 that I considered should be adjusted for in all models regardless of the magnitude of association were defined ("core" variables):

- Age (years),
- Sex (female vs male),
- Ethnicity (white vs non-white as small numbers in the latter group),

- Deprivation (Index of Multiple Deprivation (IMD) percentile[190]),
- Rural/urban classification (urban town/city, rural town, and rural village), and
- Catchment percentage (percentage of individuals in the local area visiting an Oxfordshire hospital as defined by the Office of Health Disparities; 0 = none, 100 = all[186]).

An additional core variable capturing the type of previous hospital exposure was also included for all models using the "any healthcare" cohort (any previous inpatient episode, any previous A&E visits with no previous inpatient episode, any previous outpatient appointment with no previous inpatient episode or A&E visit, or any blood test or microbiology sample collection with no previous inpatient admission, A&E visit or outpatient appointment).

I next defined 228 non-core "screening" characteristics from various data sources available using EHR data. These could be broadly split into the following categories:

- Diagnosis codes from inpatient admissions and outpatient appointments (not recorded at A&E visits). International Classification of Diseases, Tenth Revision (ICD-10) were used for all definitions. ICD-10 codes from outpatient appointments are not widely used as reimbursement for appointments is through clinic codes, rather than ICD-10 codes as is the case for inpatient admissions. N=53*.
- Procedure codes from inpatient admissions and outpatient appointments (not recorded at A&E visits). OPCS Classification of Interventions and Procedures (OPCS-4) were used for all definitions. N=70*.
- Healthcare contact history based on inpatient admissions, outpatient appointments, and A&E visits. N=44.
- Positive and negative results from microbiology isolations. N=28.
- Results from blood tests. N=24.
- Results from vital sign measurements. N=6.
- Patient characteristics of weight, height, and BMI. N=3.

*While there were many more ICD-10 codes and procedure codes available, I focused on risk factors coded from previously published research, as well as selecting the most common diagnoses and procedures in inpatient admissions. How these risk factors were selected and defined is explored in more detail in Chapter 4. A full list of all variables with definitions is available in **Appendix A**.

Whether individuals had each of the risk factors was based on the "most recent contact" date. Definitions of the timing of risk factors are provided in **Table 5.2**. In summary, risk factors were not considered if they were recorded in the 72 hours directly before the blood culture collection date for

*E. coli* cases to avoid issues of reverse causality. Risk factors calculated for the control group were not subject to this exclusion and hence all risk factors recorded in the previous 5 FYs up to and including the current contact were considered.

All risk factors were included as both categorical and continuous parameterisations to capture the impact of both ever/never having a characteristic (within the last 5 FYs) as well as the proximity of the most recent record of this characteristic to the current contact to capture more recent exposure and severity in long-term conditions (such as diabetes). The categorical effect included three levels denoting (i) risk factor recorded in the 365 days (365d) before the "most recent contact" date; (ii) risk factor recorded >365d before the "most recent contact" date; (iii) risk factor not recorded in IORD in the previous 5FYs. The continuous effect denoted days since the most recent record of the characteristic within the last 365d. For inpatient admissions, outpatient appointments, A&E visits, blood cultures, and urine tests, the number of occurrences within the previous 365d were also considered as variables, as well as the cumulative length of stay within the previous 365d in inpatient admissions.

**Table 5.2: Risk factor timing definitions**

| Risk factor dataset | Definition for cases | Definition for controls |
|---|---|---|
| **Diagnosis codes** | Any diagnosis from an episode which started >72 hours before and ended before the blood culture (BLC) collection date/time and started within the previous 5FYs. The time of diagnosis is calculated based on the episode start date. | Any diagnosis from the current episode, or any episode which started within the previous 5FYs. The time of diagnosis is calculated based on the episode start date. |
| **Procedure codes** | Any procedure with a procedure date/time >72 hours before the BLC collection date/time and within the previous 5FYs. | Any procedure with a procedure date/time at, or within the previous 5FYs before, the "most recent contact" date. |
| **Previous inpatient admissions** | Any admission which started >72 hours before the BLC collection date/time and started within the previous 5FYs, and ended strictly before the most recent contact date, based on admission date/time. | Any admission which started and ended before the "most recent contact" date and within the previous 5FYs. |

| Risk factor dataset | Definition for cases | Definition for controls |
|---|---|---|
| | Length of stay was calculated from these admissions i.e. admissions which were strictly before the most recent contact date. | |
| **Previous outpatient appointments & previous A&E visits** | Any appointment/visit which happened >72 hours before the BLC collection date/time and within the previous 5FYs. | Any appointment/visit which happened before the "most recent contact" date and within the previous 5FYs. |
| **Microbiology sample collection** | Any microbiology sample with a collection date/time >72 hours before the BLC collection date/time and within the previous 5 FYs. | Any microbiology sample with a collection date/time at, or within the previous 5FYs before, the "most recent contact" date. |
| **Blood test collection** | All blood test results taken in the 72 hours preceding BLC collection were excluded. Measurements taken outside of hospital were prioritised because they were more likely to reflect a "steady state" than acute illness, followed by measurements at outpatient appointments, measurements closest to discharge from an inpatient admission, and lastly, measurements taken at A&E. For each, the closest blood test measurement within the previous 5FYs before was subsequently taken. | Measurements taken outside of hospital were prioritised, followed by measurements at outpatient appointments, measurements closest to discharge from an inpatient admission, and lastly, measurements taken at A&E. The closest blood test measurement at, or within the 5FYs before, the "most recent contact" date. |
| **Vital signs** | All vital sign results taken in the 72 hours preceding BLC collection were excluded. Measurements taken at outpatient appointments were prioritised, followed by measurements closest to discharge from an inpatient admission, and | Measurements taken at outpatient appointments were prioritised, followed by measurements closest to discharge from an inpatient admission, and lastly, measurements taken at A&E. The closest vital sign measurements taken at or within |

| Risk factor dataset | Definition for cases | Definition for controls |
|---|---|---|
| | lastly, measurements taken at A&E. The closest vital sign measurement within the previous 5FYs before BLC collection was then used. | the 5FYs before the "most recent contact" data were used. |
| **Personal traits (height, weight, BMI)** | The closest measurement recorded in the previous 5FYs was used. If no previous measurement was available, results in the future 5FYs were used (closest first). | The closest measurement recorded in the previous 5FYs was used. If no previous measurement was available, results in the future 5FYs were used (closest first). |

Variables based on vital sign measurements were not considered in the "any healthcare" cohort as vital sign measurements available in IORD are only recorded within hospital and hence would contribute large amounts of missing data if included. For diagnosis and procedure code-based risk factors, variables were considered not present if not recorded and hence included in the "risk factor not recorded in IORD in the previous 5FYs" level of the categorical variable. This means that for the "any healthcare" cohort, all individuals with no previous inpatient admissions or outpatient attendance were included in this "never" group if they only had previous A&E visits or blood and microbiology tests for all variables based on diagnosis and procedure codes. The interpretation of these variables as ever being recorded (as opposed to ever being present) was therefore important as the absence of code did not necessarily equate to the absence of the characteristic in an individual. For example, diagnosis codes for diabetes would not include everyone with diabetes as many people with diabetes would not have inpatient admissions or outpatient attendances.

### 5.2.5 Training the screening process on data from FY2019

*Statistical analyses*

The screening process outlined below was mainly based on the process developed in Chapter 2. There were some changes and hence it is outlined again below. Summary statistics were presented as medians with interquartile ranges for continuous variables and proportions for categorical variables. All analyses used complete cases.

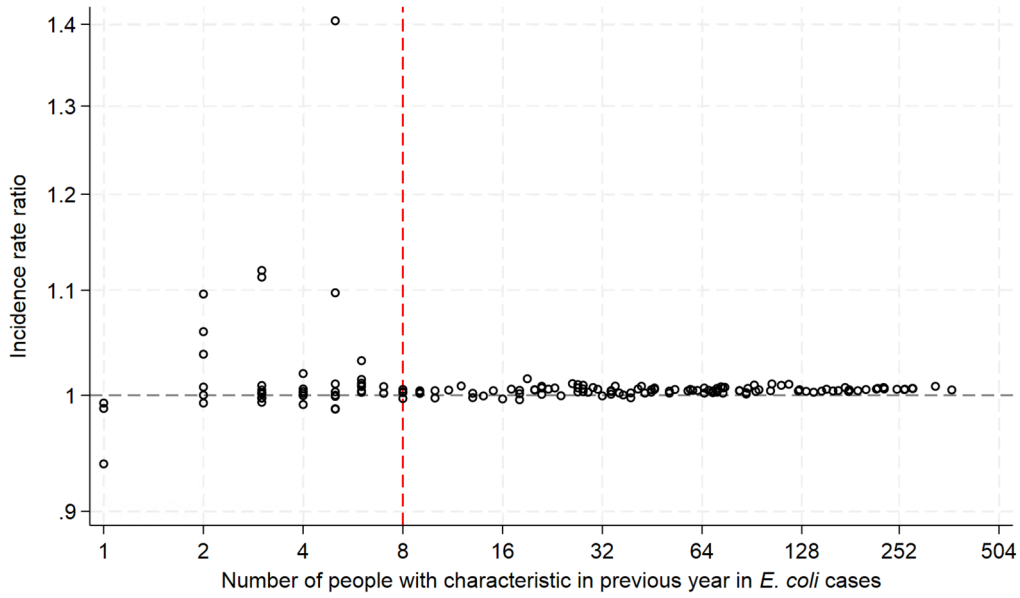
The screening process outlined below was initially run on the "inpatient only" cohort for the primary analysis.

For each financial year, associations between *E. coli* BSIs (binary yes/no, cases and controls) and the "core" variables were estimated using Poisson regression (log link) with cluster robust standard

errors (requiring complete data for “core” variables). The “core” variables were included in all subsequent models regardless of significance. Restricted cubic splines with between one to five internal knots were considered for all continuous core variables, with a minimum of one internal knot included for age due to expected variation and linear effects allowed for deprivation and catchment percentage. Internal knots were placed at even percentiles throughout the range of values for each variable (after truncation at the 5th and 95th percentiles) with boundary knots included at the 10th and 90th percentiles of the range of values for each variable. Non-linearity was tested in univariate Poisson models for each continuous characteristic, with a Bayesian Information Criterion (BIC) difference greater than 10 resulting in non-linear parameterisation with more knots being selected. A BIC difference of >10 provides very strong evidence in favour of the model with the smaller BIC according to guidelines and was used to favour parsimony given relatively small numbers of BSIs.[221] Pairwise interactions between all core variables were tested, conducting backwards elimination on all interactions which individually had global heterogeneity p-value <0.002 (Bonferroni adjustment, 0.05/21 (number of interaction tests)). The “core model” was subsequently created. Estimates were presented as incidence rate ratios (IRRs) and 95% confidence intervals (CIs); however, the core model aims to adjust for differences in the selected cohorts rather than for interpretation.

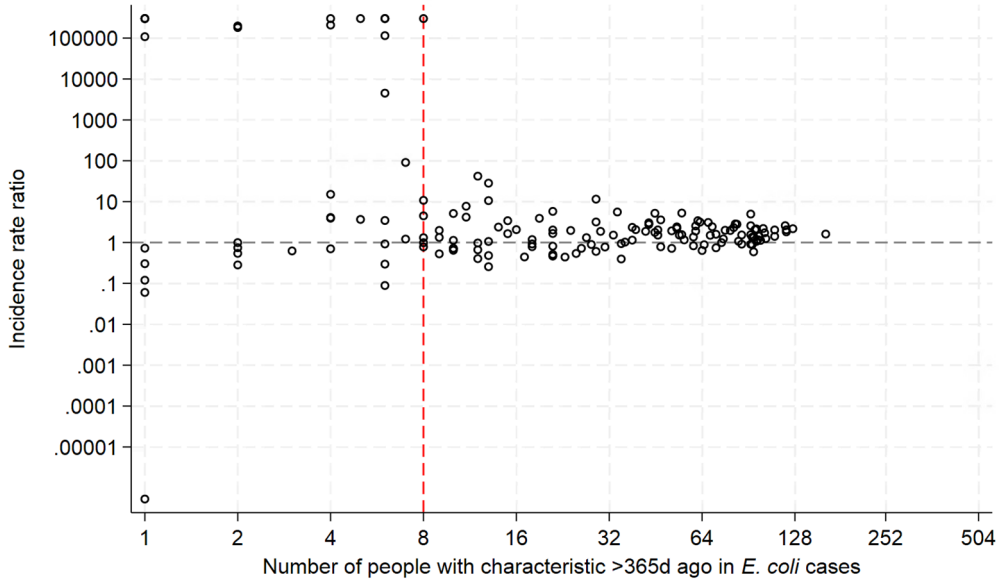
Before the initial screening step, variables with few observations in the 365d before current contact were dropped from the screening process. An arbitrary but pragmatic cut-off was selected based on the size of the incidence rate ratios when all variables were considered non-linearly with the core variables. Variables with ≤8 occurrences in the 365d before current contact in cases were therefore dropped from analysis (**Figure 5.1**). If there were ≤8 occurrences >365d ago, individuals with the characteristics>365d ago were grouped in with those who never had the characteristic (**Figure 5.2**). If there were ≤8 occurrences after combining those with the characteristic >365d ago and those who never had the characteristic, the categorical variable was dropped from the analysis, but the continuous variable explaining time since the most recent occurrence in the last year was retained.

**Figure 5.1: Incidence rate ratio for a one day greater time since closest characteristic (continuous effect) by number of occurrences in cases in the previous year.**




Note: All variables parameterised as days since closest characteristic, including diagnostic code, procedure code, previous admissions, and microbiology variables. Large values indicate implausible effects due to over-fitting.

**Figure 5.2: Incidence rate ratio for having the characteristic >365d ago versus ≤365d ago (categorical effect) by number of occurrences in cases >365d ago.**




Note: IRRs are truncated at the 95th percentile.

I next added each of the "screening" variables individually on top of the core model. All continuous variables were considered for non-linearity in univariate models as described above. Variables were considered with only the core variables first rather than using a large multivariate model from the

beginning as with many variables (>400) the multivariate model may not fit and/or be subject to substantial collinearity. Further, with large amounts of missing data in some of the screening variables, running an analysis using complete cases for all variables would result in a massive loss of data.

The global heterogeneity p-value for each screening variable was extracted. Backwards elimination was subsequently run on all screening variables with a global p-value<0.25. This threshold was selected as it has previously been shown that thresholds which are too strict can fail to identify variables known to be important in forward selection, and p<0.25 has been recommended as a reasonable level to account for this.[222]

Before backwards elimination, pairwise correlations between all variables selected at the p<0.25 threshold were summarised using the Pearson correlation coefficient. For pairs of variables with a correlation coefficient >0.95, one variable from each pair was excluded to reduce collinearity. The variable removed was selected on a case-by-case basis.

Backwards elimination was then run on including all remaining variables with p<0.25. An exit p-value of 0.05 was used for backwards elimination. After backwards elimination, the final model was refitted on complete cases for all selected variables. Collinearity was assessed by comparing the direction of the effect between the univariate and multivariate estimates. Variables with evidence of collinearity were assessed on an individual basis. Final model results were then presented.

Changes in McFadden's pseudo R-squared value, as defined in equation **(1)**,[223] were used to assess the importance of the variables selected for the final model. Each variable included in the final model was removed from the full model one at a time with the R-squared extracted from each model. The same population was used for each model to ensure that the populations compared were equivalent and that individuals were not added into a sub-model if a variable with larger amounts of missing data was removed. The difference between the R-squared value from the full model and the R-squared values from the models with each variable removed was calculated and plotted. The largest difference represented the variable which explained the most variation in the observed data, while the smallest difference represented the variable explaining the least variation.

$$R^2_{McFadden} = 1 - \frac{\ln \hat{L}(M_{full})}{\ln \hat{L}(M_{intercept})}$$

*Where $M_{full}$ is the fitted model and $M_{intercept}$ is the null model with the intercept only.*

**(1)**

*Sensitivity analysis: Using different entry p-value thresholds*

With >400 variables and many tests being carried out, there was a risk of finding spurious results through conducting multiple tests. To assess the impact of multiple testing on my findings, I compared two additional p-value thresholds with the original threshold of global p-value<0.25. First, after the initial screen, all variables with a global p-value less than the Benjamini-Hochberg adjusted p-value threshold were selected for backwards elimination. Second, after the initial screen, all variables with a global p-value less than the Bonferroni adjusted p-value threshold were selected for backwards elimination. These two thresholds were selected as they both account for multiple testing but vary in how conservative they are, with the Bonferroni adjustment being stricter. Backwards elimination was run separately using the two sets of variables. The final variables selected and the size of the effects were compared between the original threshold of p<0.25, Benjamini-Hochberg adjustment, and Bonferroni adjustment.

*Subgroup analysis: Screening on different populations*

As described above (and in Chapter 4), a broader cohort was defined, referred to as the "any healthcare" cohort, which included all individuals in the "inpatient only" cohort and additionally included individuals with any outpatient appointment, A&E visit, microbiology sample collected, or blood test result in the 5FYs strictly before the most recent contact date. The screening process outlined above was carried out on this population with the final variables selected after backwards elimination compared with those selected in the "inpatient only" cohort.

A further cohort was defined restricting both the cases and controls to only individuals aged>65y in the "inpatient only" cohort. As the majority of *E. coli* BSIs are in older individuals, restricting to only those aged>65y in both the cases and controls could balance strong predictors of *E. coli* BSIs such as age and frailty and expose different risk factors. The age>65y cohort was selected from the "inpatient only" cohort to reduce the amount of missing data. Again, the screening process was run in full with variables selected after backwards elimination compared with the "inpatient-only" cohort.

*Subgroup analysis: Using different outcomes*

In all the analyses presented so far, all *E. coli* BSIs were included as the outcome of the analyses. Different outcomes of interest based on prior hospital exposure and antibiotic resistance were defined to assess whether different risk factors were found in these subgroups.

Different risk factors for *E. coli* BSIs may be expected based on differing previous exposure to hospital. *E. coli* BSIs were therefore categorised into the following groups dependent on the recency of hospital exposure:

- Nosocomial: blood sample for culture taken between 48h after hospital admission and hospital discharge.
- Quasi-nosocomial: not nosocomial and last discharged from hospital 0-30 days previously to when the blood sample was taken.
- Quasi-community: not nosocomial and last discharged 31-365 days previously.
- Community: not nosocomial and last discharged >1 year previously or never previously admitted to OUH.

Admission date/time and discharge date/time from inpatient admissions and collection date/time from bacteraemias were used to define and calculate the groupings. The "inpatient only" analysis cohort and controls were used for the nosocomial, quasi-nosocomial, and quasi-community cases as these groupings are based on previous and current inpatient exposure. The "any healthcare" analysis cohort and controls were used for the community *E. coli* BSIs as, by definition, the community-onset *E. coli* BSIs do not have any inpatient contact in the 365d before the date of collection of the *E. coli* BSI. Again, variables selected for the final model after backwards elimination were compared with the all case analysis with the "inpatient only" cohort for the nosocomial, quasi-nosocomial, or quasi-community groupings, or the all case analysis with the "any healthcare" cohort for the community group.

A second subgroup analysis considered only *E. coli* BSIs that were resistant to third-generation cephalosporins as cases (excluding all other cases from analysis). Consistent with the 2022/23 ESPAUR report, third-generation cephalosporins were defined as resistance to any of the following drugs: cefotaxime, ceftazidime, cefpodoxime and ceftriaxone.[61] All *E. coli* BSIs reported as resistant from the diagnostic laboratory to any of these four drugs were flagged as resistant. As well as susceptible and resistant results, there is a third result of "intermediate" if the MIC was between susceptible and resistant threshold in the European Committee on Antimicrobial Susceptibility Testing (EUCAST) guidelines.[224] For this sensitivity analysis, intermediate results were not classed as resistant as they are susceptible in a dose-dependent way. The "inpatient only" cohort and controls were used for this analysis to reduce the amount of missing data across characteristics. The final model results were compared with results from all cases and the "inpatient only" cohort.

### 5.2.6 Testing the screening process on different years of data

After completing the analysis on the training dataset of FY2019, the screening process was run on FYs 2018, 2020, and 2021 to assess whether the risk factors selected varied across the years. All the same risk factors were considered and the screening process was carried out as before. All cases and the "inpatient only" cohort were used across all FYs to reduce the amount of missing data. Which risk factors and the sizes and direction of effects from the final model after backwards elimination were compared.

All analysis was carried out in Stata 17 and Stata 18.

## 5.3 Results

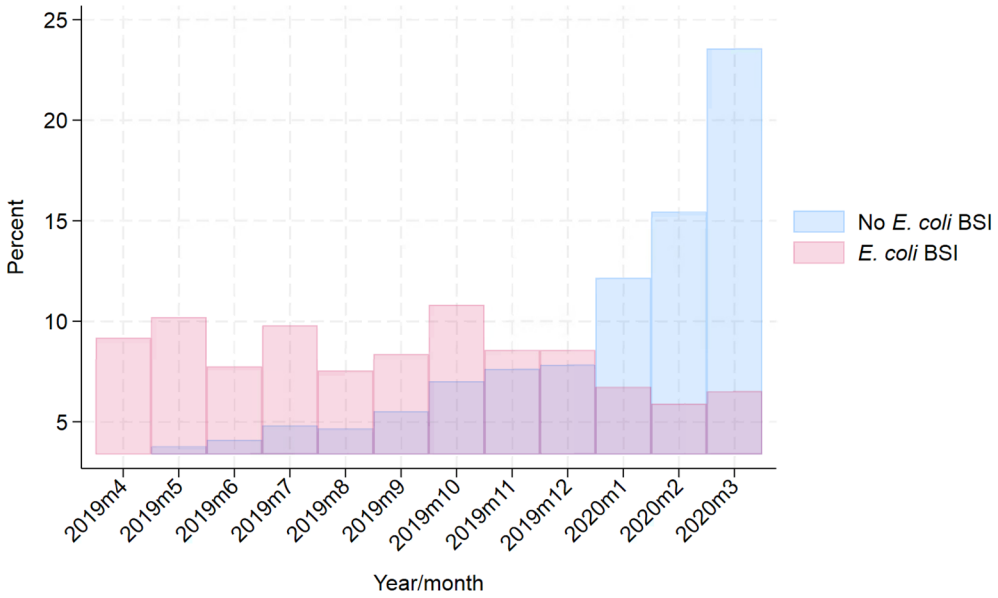
### 5.3.1 Summary of data

Between 1st April 2018 and 31st March 2022, there were 2,356 *E. coli* isolations from blood from 1,830 individuals. After de-duplication, removing individuals aged <16 years old, dropping those missing age and sex, and excluding all records which were not the first *E. coli* BSI in the financial year there were 460, 512, 421, and 450 *E. coli* BSIs in FYs 2018, 2019, 2020, and 2021, respectively (flowchart provided in Chapter 4, **Figure 4.2**).

From 1st April 2018 to 31st March 2022 there were 12,996,578 records from 841,374 individuals in the IORD database from inpatient admissions, outpatient appointments, A&E visits, microbiology samples, and blood tests. After removing all individuals aged <16 years, those missing age and sex, those with an *E. coli* BSI since 1st April 2013 in IORD, and selecting only the last contact in each FY, there were 412,149, 415,497, 374,117, and 413,804 potential controls in FYs 2018, 2019, 2020, and 2021, respectively (flowchart provided in Chapter 4, **Figure 4.3**)

As the first date in each FY was selected for the *E. coli* BSIs and the last date in the FY was selected for potential controls, the distribution of the "most recent contact" date was skewed to the latter end of the FY for potential controls while the date of *E. coli* BSI collection was more evenly distributed over the FY, as shown for the example of FY2019 (**Figure 5.3**).

**Figure 5.3: Distribution of the month of the most recent contact taken for those without/with E. coli BSIs from FY2019.**



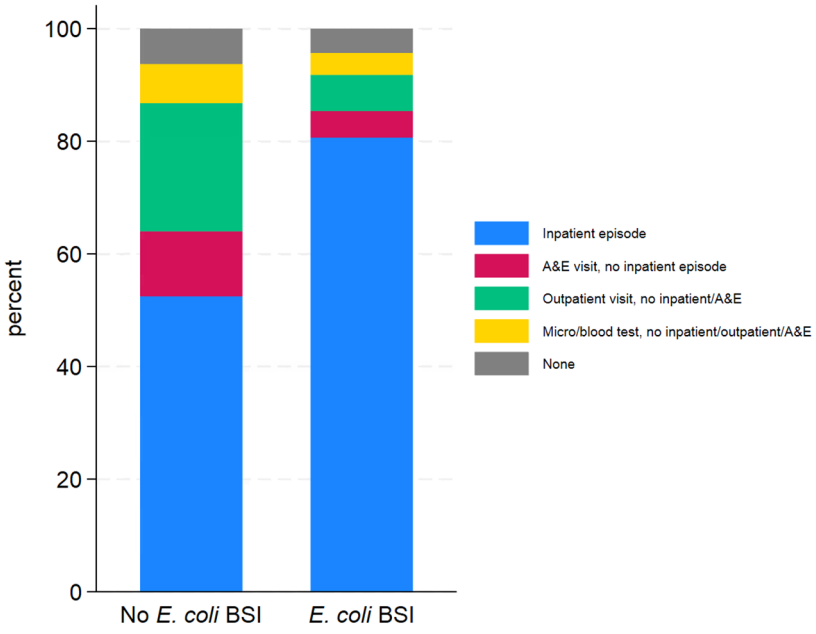
### 5.3.2 Training the screening process on FY2019 data

The screening process was first trained on data from FY2019, helping to update the process from its application in Chapter 2 on COVID-19 data before expanding the process to other subgroups and years of data.

*Summary of the FY2019 data*

416,009 people had any type of healthcare contact in FY2019. 512 (0.1%) had an *E. coli* BSI, while 415,497 (99.9%) did not (potential controls). 413 (81%) *E. coli* cases had at least one inpatient episode in the previous 5FYs, compared with 217,934 (52%) potential controls (**Figure 5.4**). 48,027 (12%) individuals in the potential control group had any previous A&E visits without an inpatient episode, compared with 24 (5%) cases. A large proportion of potential controls (94,672, 23%) had an outpatient appointment in the previous 5FYs without an inpatient episode or A&E visit, compared with 33 (7%) cases. 22 (4%) *E. coli* BSIs and 26,062 (6%) potential controls were dropped from all subsequent analyses (i.e. both cohorts) as they had no healthcare contact of any kind recorded in IORD in the previous 5FYs. The median (IQR) catchment percentage for these individuals was very low, at 2% (0.2%, 90%) and 9% (4%, 93%), with 23% (5/22) and 6% (1,436/26,062) missing for cases and controls respectively. In comparison, for all other individuals, the median (IQR) catchment percentage was 97% (90%, 96%) and 93% (71%, 96%), with 0% (0/490) and 0.6% (2,309/389,435) missing for cases and controls respectively.

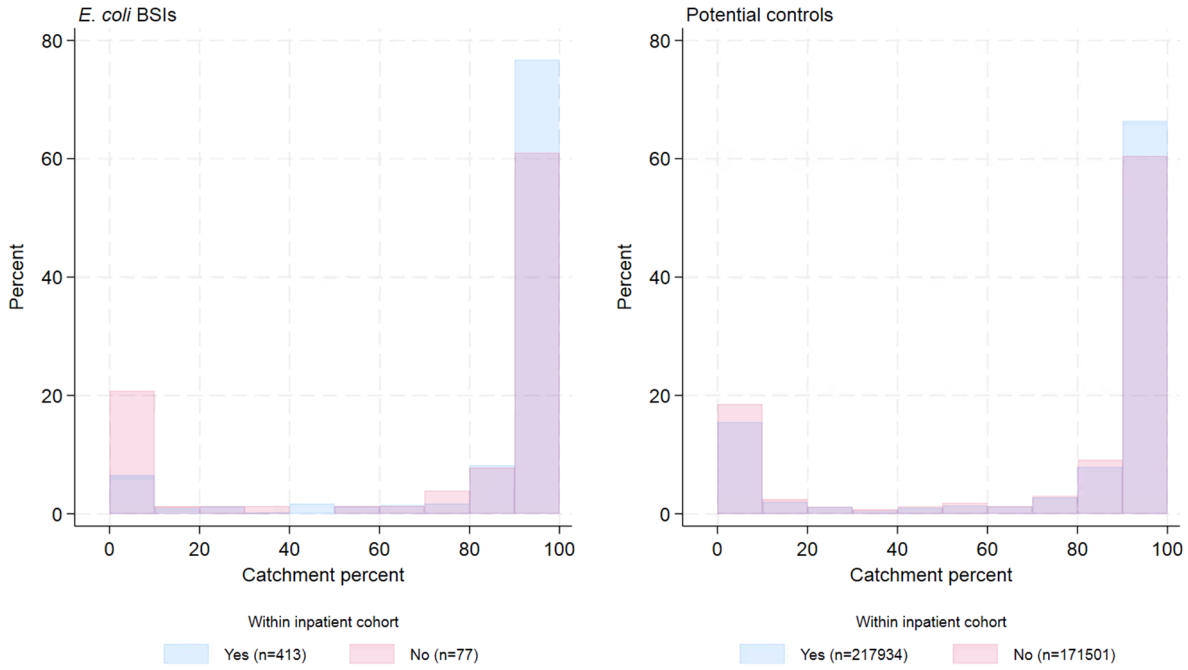
**Figure 5.4: Health contact history in the previous five financial years split by the presence of *E. coli* BSI.**




The primary analysis was conducted on the "inpatient only" cohort, just considering those with at least one inpatient episode in the previous 5FYs (the blue bars in **Figure 5.4**, comprising 413/490

(84%) *E. coli* BSIs with any healthcare contact and 217,934/389,435 (56%) controls with any healthcare contact). The remaining 77 (16%) cases and 171,501 (44%) controls were therefore dropped from all subsequent analyses using the "inpatient only" cohort. A higher proportion of *E. coli* cases with any healthcare contact but without a previous inpatient episode were out of Oxfordshire catchment, with 7% of cases within the "inpatient only" cohort having a catchment percentage <10% compared with 21% of cases dropped from the "inpatient only" cohort but included in the any healthcare contact cohort (**Figure 5.5**). The difference in catchment percentage was not as large for the control group.

**Figure 5.5: Distribution of catchment percentage for those within and not within the "inpatient only" cohort, but in the "any healthcare" cohort, split by *E. coli* cases (left) and potential controls (right).**




In complete cases for "core" variables, comprising 370/413 (90%) *E. coli* BSIs with inpatient contact and 177,407/217,934 (81%) controls with inpatient contact (explored in more detail below), those with *E. coli* BSIs were generally older than controls as expected: median (IQR) 78y (67y-86y) for cases versus 58y (40y-74y) for controls (**Table 5.3**). There was a higher proportion of females in controls compared with cases (58% vs 49% for controls versus cases), again as expected given maternity admissions (see Chapter 4), and also a higher proportion of those reporting non-white ethnicities (8% vs 5% for controls versus cases). Deprivation and rural/urban classification were very similar across cases and controls.

**Table 5.3: Summary of core variables for complete cases only.**

| | Inpatient cohort, n (%) | | |
|---|---|---|---|
| | **Controls** | **Cases** | **Total** |
| **N** | **177,407 (100)** | **370 (100)** | **177,777 (100)** |
| **Age (years)** | 58 (40-74) | 78 (67-86) | 58 (40-74) |
| **Sex** | | | |
| Male | 75,075 (42) | 188 (51) | 75,263 (42) |
| Female | 102,332 (58) | 182 (49) | 102,514 (58) |
| **Ethnicity** | | | |
| White | 163,529 (92) | 353 (95) | 163,882 (92) |
| Non-white | 13,878 (8) | 17 (5) | 13,895 (8) |
| **IMD percentile** | 74 (54-89) | 74 (54-89) | 74 (54-89) |
| **Catchment percentage** | 94 (85-96) | 95 (91-96) | 94 (85-96) |
| **Rural/urban classification** | | | |
| Urban city/town | 117,284 (66) | 247 (67) | 117,531 (66) |
| Town/fringe | 28,643 (16) | 59 (16) | 28,702 (16) |
| Rural village | 31,480 (18) | 64 (17) | 31,544 (18) |

Note: showing n (%) or median (IQR). Higher IMD percentiles indicate less deprivation.

The amount of missing data in core variables was larger in controls than in cases, with 19% (n=40,527) of controls missing data for at least one core variable compared with 10% (n=43) of cases (**Table 5.4**). All missing data in cases was due to missing ethnicity. Almost all missingness was attributable to missing ethnicity in the control group, with a very small number of people missing IMD percentile, catchment percentage, and rural/urban classification.

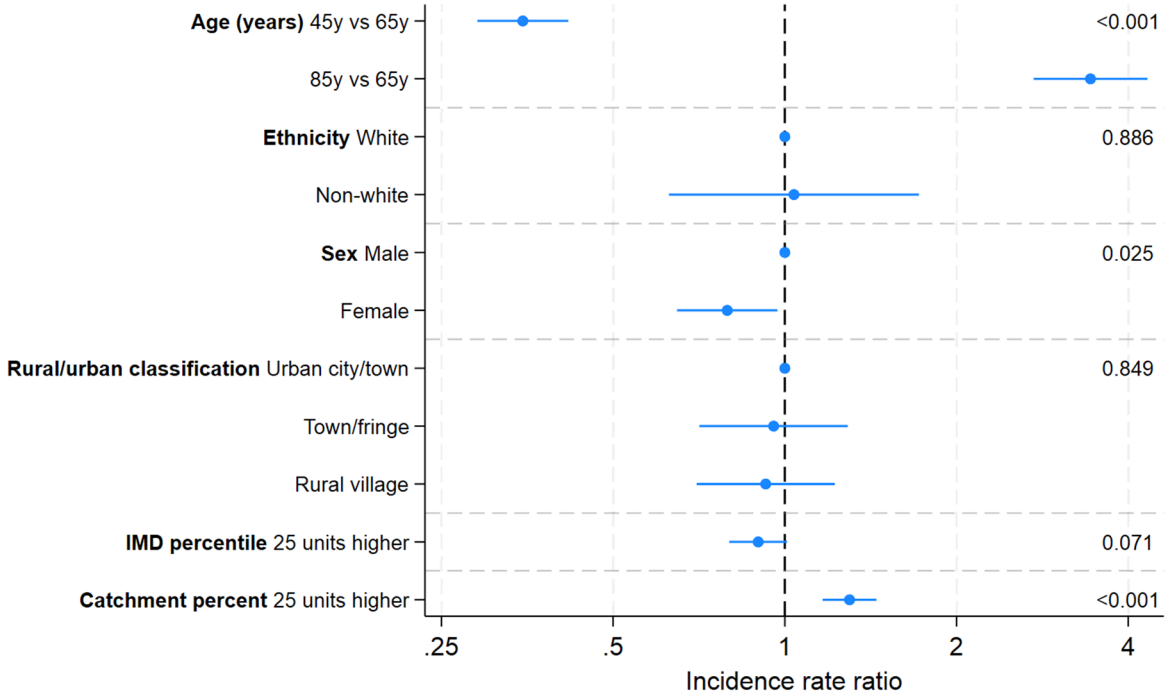
**Table 5.4: Number of individuals missing core variables for cases and controls.**

| | Inpatient cohort, n (%) | | |
|---|---|---|---|
| | **Controls** | **Cases** | **Total** |
| **N** | **217,934 (100)** | **413 (100)** | **218,347 (100)** |
| Ethnicity | 39,816 (18) | 43 (10) | 39,859 (18) |
| IMD percentile | 971 (0) | 0 (0) | 971 (0) |
| Catchment percentage | 971 (0) | 0 (0) | 971 (0) |
| Rural/urban class | 974 (0) | 0 (0) | 971 (0) |
| Total missing | 40,527* (19) | 43 (10) | 40,570* (19) |

*Total number of individuals missing at least one core variable. Individuals may be missing more than one core variable.

When fitting the "core model", all continuous variables were tested for non-linearity, with age included with one internal knot and IMD percentile and catchment percentage included as linear. All pairwise interactions were tested between all core variables, with no interactions being selected based on a global p-value threshold of <0.002 to account for multiple testing (Bonferroni threshold). The incidence rate ratios from the core model are presented below in **Figure 5.6**. These factors were included in all subsequent models to adjust for the underlying differences in population structure rather than to explore actual risk in these variables but are presented below for completeness.

**Figure 5.6: IRRs with 95% confidence intervals for core variables for the "inpatient-only" cohort.**




Note: Global p-values testing heterogeneity for each variable are presented on the right of the graph. Age was included as a restricted cubic spline (one internal knot) with the values presented predicted from the fitted spline. Higher IMD values indicate lower deprivation.

*Screening on the inpatient cohort with entry p-value<0.25*

A total of 228 characteristics were considered, leading to 410 variables when considering both the time since the closest occurrence of a characteristic and never/ever having had the characteristic (**Table 5.5**). The majority of characteristics were calculated from procedure codes (n=70, 31%), followed by diagnosis codes (n=53, 23%), and characteristics derived from previous hospital attendances (44, 19%).

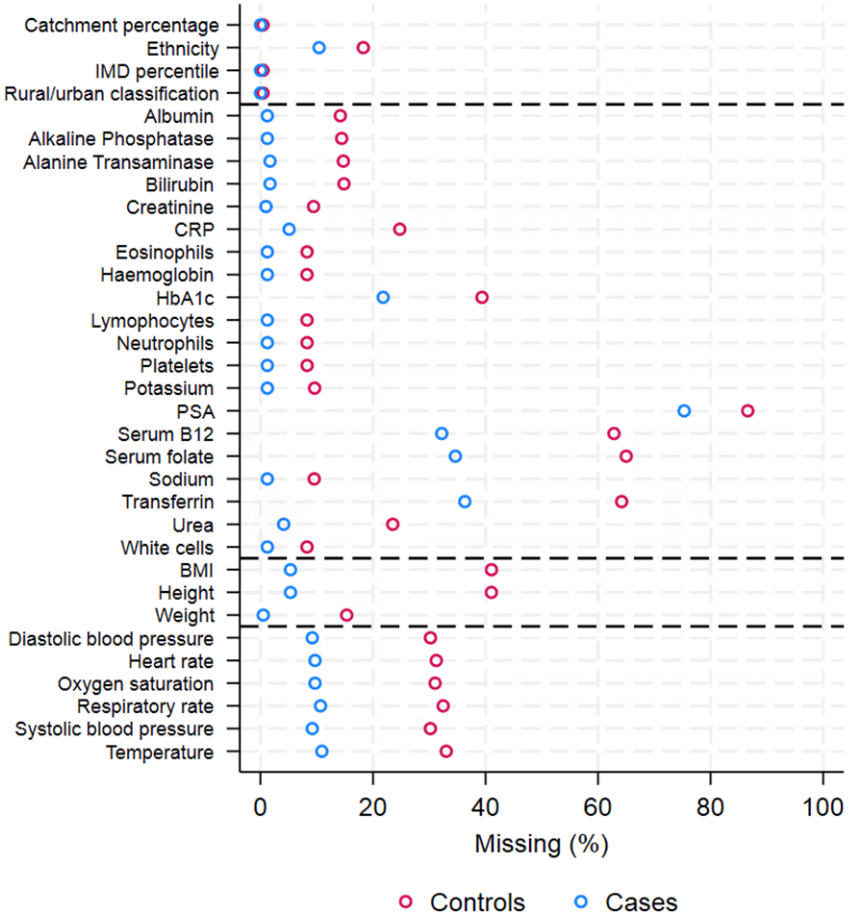
**Table 5.5: Number of characteristics and variables from each data source.**

| Dataset | Number of characteristics (N=228) | Number of variables (N=410) |
|---|---|---|
| Diagnosis codes | 53 | 102 |
| Procedure codes | 70 | 140 |
| Previous healthcare attendance | 44 | 81 |
| Microbiology | 28 | 54 |
| Blood tests | 24 | 24 |
| Vital signs | 6 | 6 |
| Personal traits | 3 | 3 |

The amount of missing data varied between the different characteristics as well as between cases and controls (**Figure 5.7**). Cases had less missing data than controls for all characteristics. Missing

data was present for all blood tests but the amount of missing data varied between tests. The prostate-specific antigen (PSA) test was missing the most data with 75% of cases missing measurements and 87% of controls. Tests for serum B12, serum folate, and transferrin all had high missingness: between 32-36% in cases and 63-65% in controls. As these four blood tests had very high missingness, the values of these tests were not subsequently considered in the analysis and instead, the presence/absence of these test results were considered since the presence/absence of test results may also be informative about risk.[172] All vital sign measurements were around 10% missing in cases and 30% missing in controls. Height was 41% missing in controls, contributing to 41% of individuals missing BMI measurements. Weight was more consistently recorded with 15% missing in the controls and 0.5% missing in cases.

**Figure 5.7: Percentage of missing data for all variables split by presence of *E. coli* BSI ("inpatient only" cohort).**




*Initial Screen*

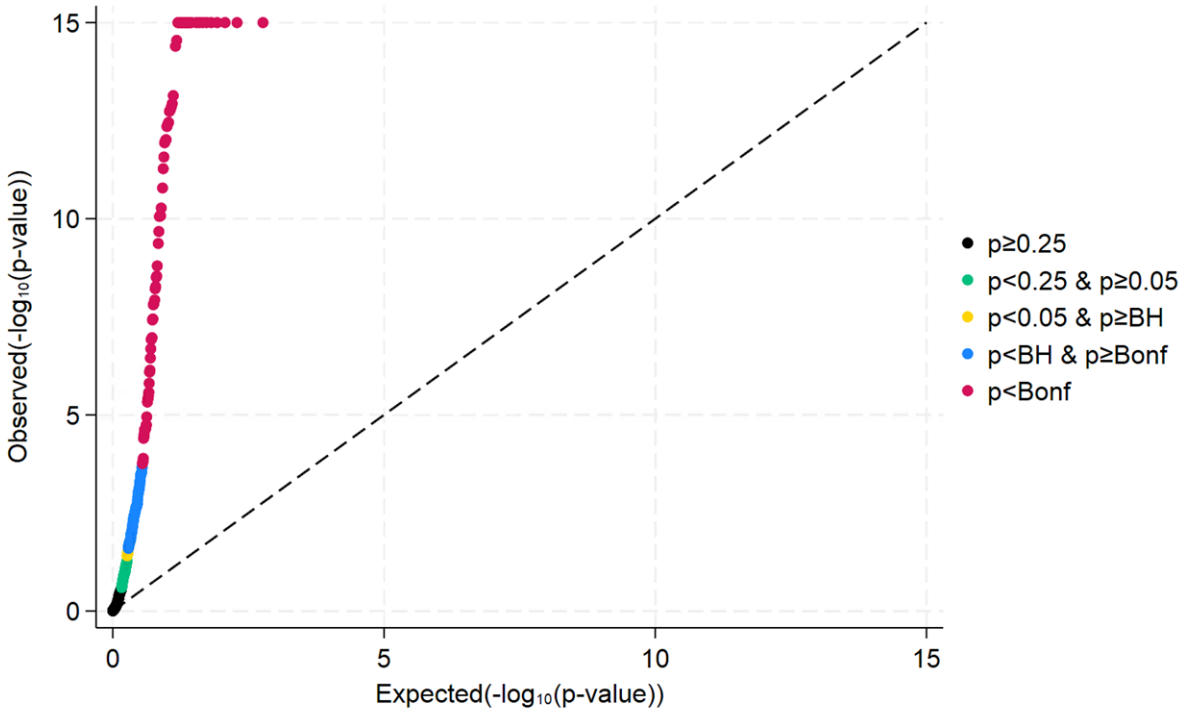
122 (30%) variables from 61 characteristics were dropped because they had ≤8 occurrences in the 365 days before *E. coli* BSI collection (median [IQR]: 3 (1-5) occurrences in the previous year). The majority of these (n=72, 59%) were variables derived from procedure codes, followed by 24 (20%)

from microbiology results and 18 (15%) from diagnosis codes. 11 (3%) categorical variables were collapsed from having three levels (≤365d ago vs >365d ago vs never in previous 5FYs) to two levels (≤365d ago or >365d ago vs never in previous 5FYs) as they had ≤8 occurrences >365d ago. Four (1%) variables were dropped as >50% of measurements were missing in either the cases or controls. These were blood test results for PSA, serum B12, serum folate, and transferrin, as described above. Flags for whether these tests were ever requested were created to include in models instead. In total, 370 (90%) cases and 177,407 (81%) controls in the inpatient cohort had complete data for core variables and only 277 (67%) cases and 60,157 (28%) controls had complete data for all 281 variables subsequently screened.

Non-linearity was considered for all 161 continuous variables which were included in the initial screen. 152 (94%) variables were selected as linear effects, with 8 (5%) and 1 (1%) selected with one and two internal knots respectively.

Observed global p-values were more significant than would be expected if the p-values were randomly selected from a uniform(0,1) distribution (the null hypothesis if no true associations exist), with very large deviations from the uniform distribution (**Figure 5.8**). Separation between observed and expected p-values occurred early in the distribution suggesting that many moderately significant p-values were more significant than expected.

**Figure 5.8: Expected -log10 global p-values from the initial screen compared with expected -log10 p-values from the uniform(0,1) distribution.**



Note: BH = Benjamini-Hochberg threshold, Bonf = Bonferroni threshold. Observed p-values were truncated to -$\log_{10}(15)$ for plotting purposes only.

Of a total of 281 p-values from the variables screened, 197 (70%) had a p-value<0.25. Pairwise correlations between all variables below the p<0.25 threshold were generally normally distributed, however, there was some deviation away from normality at the very low (close to -1) and very high (close to 1) correlations (**Figure 5.9**). Fifteen pairs of variables had an absolute correlation coefficient >0.9 (**Table 5.6**). For the eight pairs of variables with an absolute correlation >0.95, one variable was removed from consideration for backwards elimination. Where both admission to and discharge from inpatient admissions were available as variables, the variables denoting discharge from admission were dropped. Both admission to and discharge from admissions were included in the initial screening step as the number of admissions was calculated based on admission date while length of stay included time from admissions which started >365d ago but ended ≤365d ago. Length of stay (LOS) from inpatient admissions >8 hours was selected over the LOS from all inpatient admissions to avoid the inclusion of day-case admissions and restrict LOS to ordinary admissions. A diagnosis code for renal failure was selected over a diagnosis code for renal disease to keep the most severe outcome. Any remaining pairs with a correlation >0.9 were included together in the multivariate models and checked after backwards elimination for any impact of correlation on model estimates.

**Figure 5.9: Distribution of pairwise correlations between variables significant at the p<0.25 threshold.**

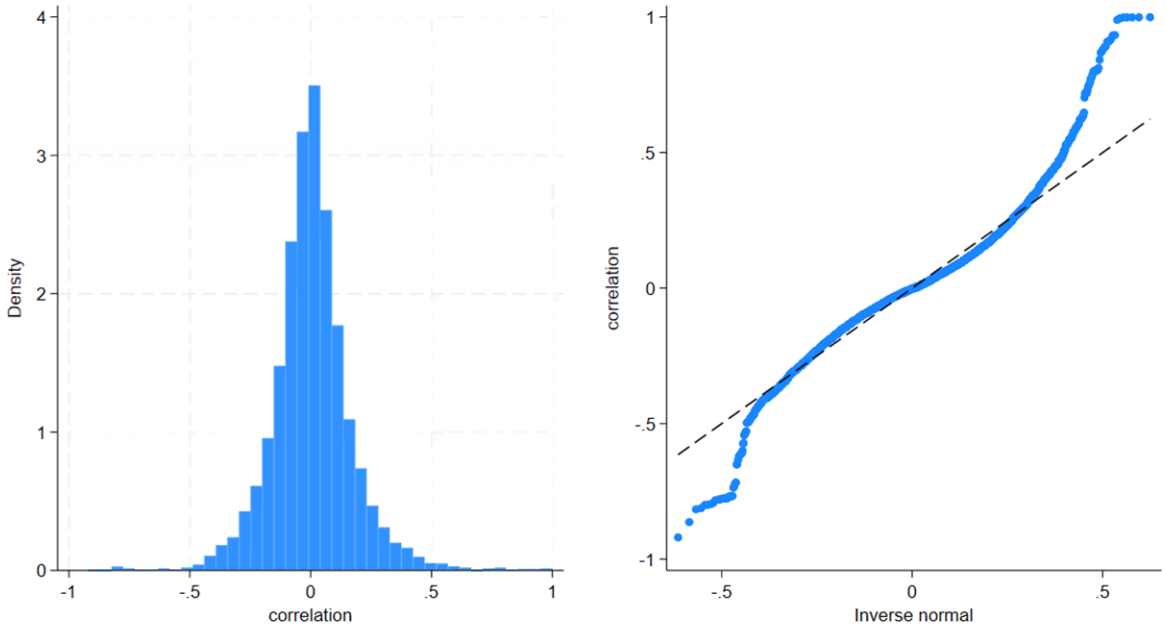


**Table 5.6: Correlation between variables with absolute correlation > 0.90**

| Variable 1 | Variable 2 | Correlation |
|---|---|---|
| Admission to emergency inpatient admission (categorical) | Discharge from emergency inpatient admission (categorical) | 1.00 |
| Admission to elective inpatient admission (categorical) | Discharge from elective inpatient admission (categorical) | 1.00 |
| Admission to complex inpatient admission (categorical) | Discharge from complex inpatient admission (categorical) | 1.00 |
| Admission to any inpatient admission (categorical) | Discharge from any inpatient admission (categorical) | 1.00 |
| Admission to any inpatient admission lasting > 8 hours (categorical) | Discharge from any inpatient admission lasting >8 hours (categorical) | 1.00 |
| Cumulative length of stay from inpatient admission lasting > 8 hours in previous 365d (continuous) | Cumulative length of stay from all inpatient admissions in previous 365d (continuous) | 1.00 |
| Diagnosis code for renal failure (categorical) | Diagnosis code for renal disease (categorical) | 0.99 |
| Outpatient treatment under the Clinical Haematology Service treatment function code (categorical) | Outpatient treatment under the Clinical Haematology Service treatment function code (continuous) | -0.95 |
| Procedure code for endoscopy of lower GI tract (categorical) | Procedure code for Chapter H (Lower Digestive system) (categorical) | 0.94 |
| Any serum B12 blood test requested | Any serum folate blood test requested | 0.93 |
| Procedure code for endoscopy of upper GI tract (categorical) | Procedure code from Chapter G (Upper Digestive System) (categorical) | 0.93 |
| Diagnosis code for paraplegia (categorical) | Diagnosis code for paraplegia (continuous) | -0.92 |
| White cell count (continuous) | Neutrophils (continuous) | 0.92 |
| Procedure code for scan of pelvis (categorical) | Procedure code for scan of abdomen (categorical) | 0.91 |
| Scan of pelvis (continuous) | Scan of abdomen (continuous) | 0.91 |

Note: Variables highlighted in green were selected for backwards elimination, variables highlighted in orange were excluded from backwards elimination.

A total of 35 variables were selected after backwards elimination (exit p>0.05). There was evidence of collinearity for some variables as measured by IRRs swapping signs between univariate ("core adjusted") and multivariate (fully adjusted) analyses (**Table 5.7**). Investigating the mechanism of action of collinearity, I found that this seemed to fall within two broad categories: potential health-seeking behaviour and potential competing risks. For example, a higher number of outpatient appointments was associated with higher risk in univariate models, but a lower risk after adjusting for other confounders (**Figure 5.10**). This effect attenuated towards zero after removing chemotherapy and hence, after adjusting for a characteristic which strongly increased the risk of *E. coli* BSI and was also a common reason for outpatient appointments, the effect of having more outpatient appointments may instead reflect individuals who more willingly seek appointments when needed, i.e. had better health-seeking behaviour. A similar pattern was observed for both

diagnosis codes from Chapter 18 (Symptoms/signs and abnormal findings) and Chapter 21 (Factors influencing health status) after removing time since the closest inpatient admission lasting >8 hours from the model (**Figure 5.11**, **Figure 5.12**).

The other variables with evidence of collinearity may be examples of competing risks demonstrating other reasons why individuals may be in hospital and hence in the control group. For example, the risk of *E. coli* BSI was higher for those with a diagnosis code for pneumonia in univariate analyses but lower after adjustment in multivariate models (**Table 5.7**). This may be an example of a competing risk as the selection criteria for controls in the study required hospital contact. Therefore, after adjusting for other confounders, being in hospital with pneumonia unrelated to an *E. coli* BSI might be associated with a lower risk as these patients are, by definition, part of the control group. These variables were not as strongly influenced by removing other variables from the multivariate model, for example in **Figure 5.13**. Collinear variables were kept in the final model.

**Table 5.7: Variables with evidence of collinearity selected after backwards elimination (exit p>0.05) and using univariable p<0.25 as the entry threshold.**

| Characteristic | Level | Multivariate IRR (95% CI) [p-value] | Multivariate global p-value | Univariate IRR (95% CI) [p-value] | Univariate global p-value |
|---|---|---|---|---|---|
| **Potentially related to health-seeking behaviour** | | | | | |
| Number of outpatient attendances | Per 2 units higher | **0.93 (0.88, 1.00) [0.038]** | **0.038** | **1.22 (1.16, 1.27) [<0.001]** | **<0.001** |
| Diagnosis from Chapter 18 (Symptoms/signs and abnormal findings) | Per 3 months closer | **0.89 (0.81, 0.98)** | **0.013** | **1.37 (1.29, 1.46) [<0.001]** | **<0.001** |
| Diagnosis from Chapter 21 (Factors influencing health status) | Per 3 months closer | **0.90 (0.83, 0.99)** | **0.023** | **1.29 (1.22, 1.37) [<0.001]** | **<0.001** |
| **Potentially related to competing risks** | | | | | |
| Diagnosis from Chapter 5 (Mental and behavioural disorders) (categorical) | ≤365d ago | 1 | 0.001 | 1 | <0.001 |
| | >365d ago | **1.86 (1.33, 2.61) [<0.001]** | **0.001** | **0.89 (0.64, 1.24) [0.498]** | **<0.001** |
| | Never | **1.50 (1.12, 2.01) [0.007]** | **0.001** | **0.52 (0.40, 0.67) [<0.001]** | **<0.001** |
| Diagnosis from Chapter 12 (Skin and subcutaneous tissue) (categorical) | ≤365d ago | 1 | 0.008 | 1 | <0.001 |
| | >365d ago | 1.80 (1.20, 2.69) [0.004] | 0.008 | 1.02 (0.69, 1.50) [0.923] | <0.001 |
| | Never | **1.64 (1.17, 2.31) [0.004]** | **0.008** | **0.64 (0.46, 0.88) [0.005]** | **<0.001** |
| Diagnosis code for pneumonia | Per 3 months closer | **0.84 (0.74, 0.96) [0.008]** | **0.008** | **1.15 (1.02, 1.29) [0.023]** | **0.023** |

Note: bold shows p<0.05 in opposite directions in univariable and multivariable models.

**Figure 5.10: IRR (95%) CI per 2-unit higher number of outpatient attendances in models removing each variable selected after backwards elimination one at a time.**

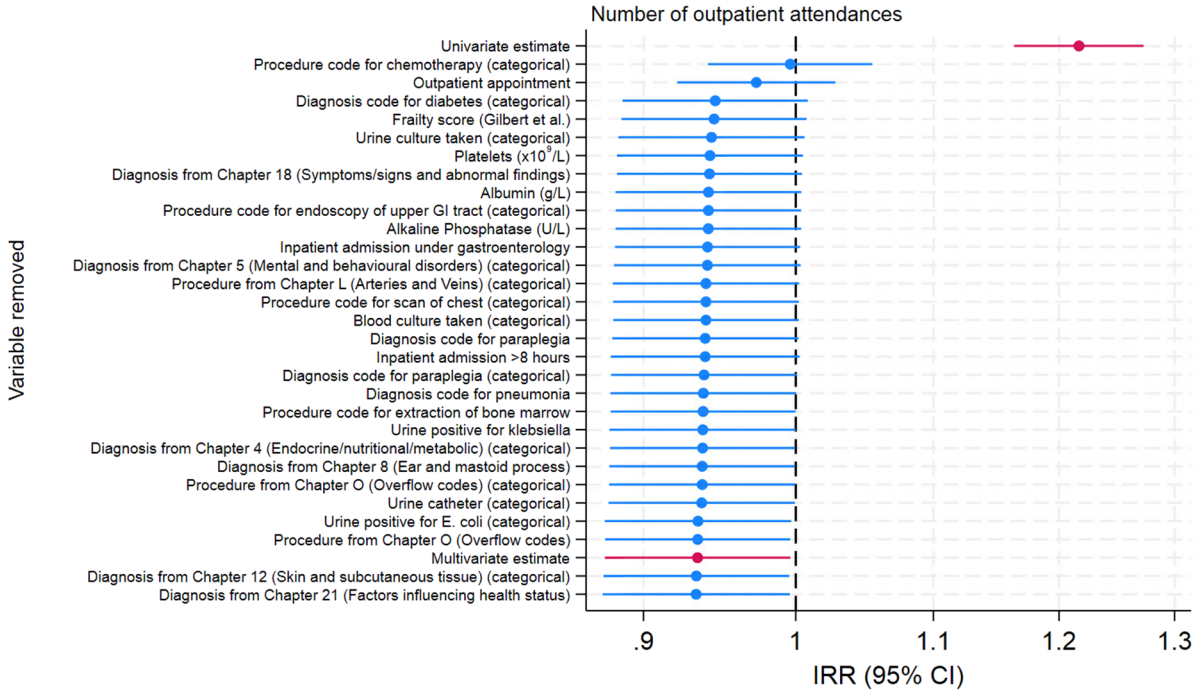



**Figure 5.11: IRR (95%) CI per 3 months closer diagnosis code from Chapter 21 in models removing each variable selected after backwards elimination one at a time.**

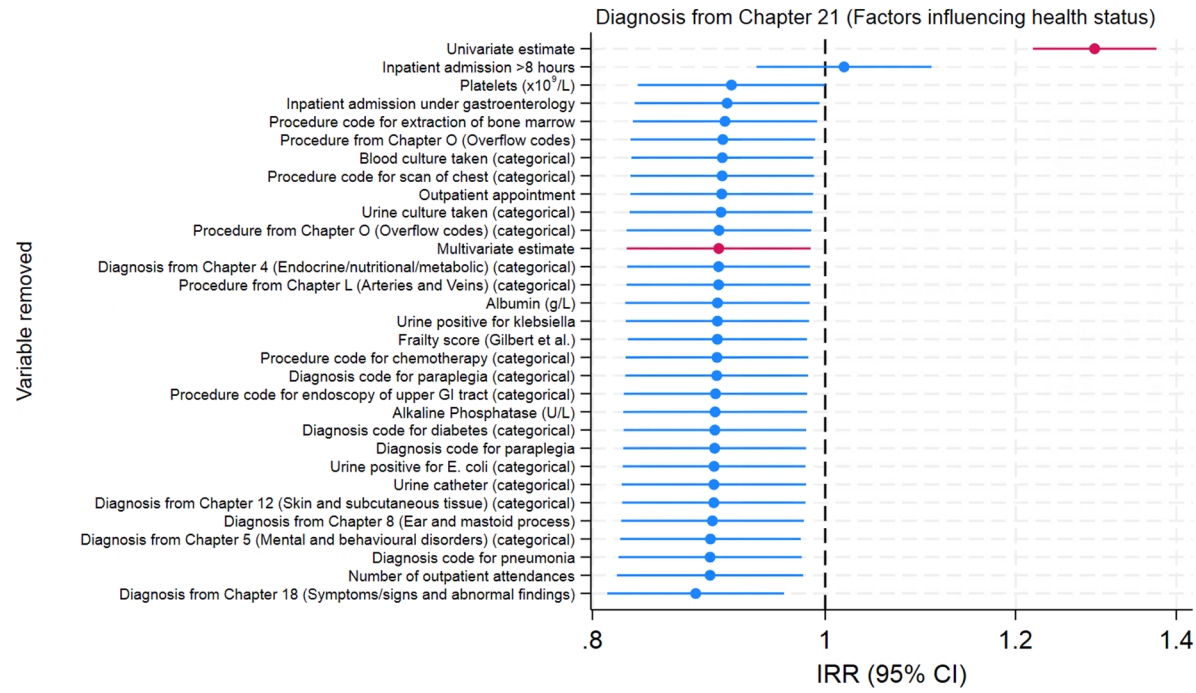


**Figure 5.12: IRR (95%) CI per 3 months closer diagnosis code from Chapter 18 in models removing each variable selected after backwards elimination one at a time.**

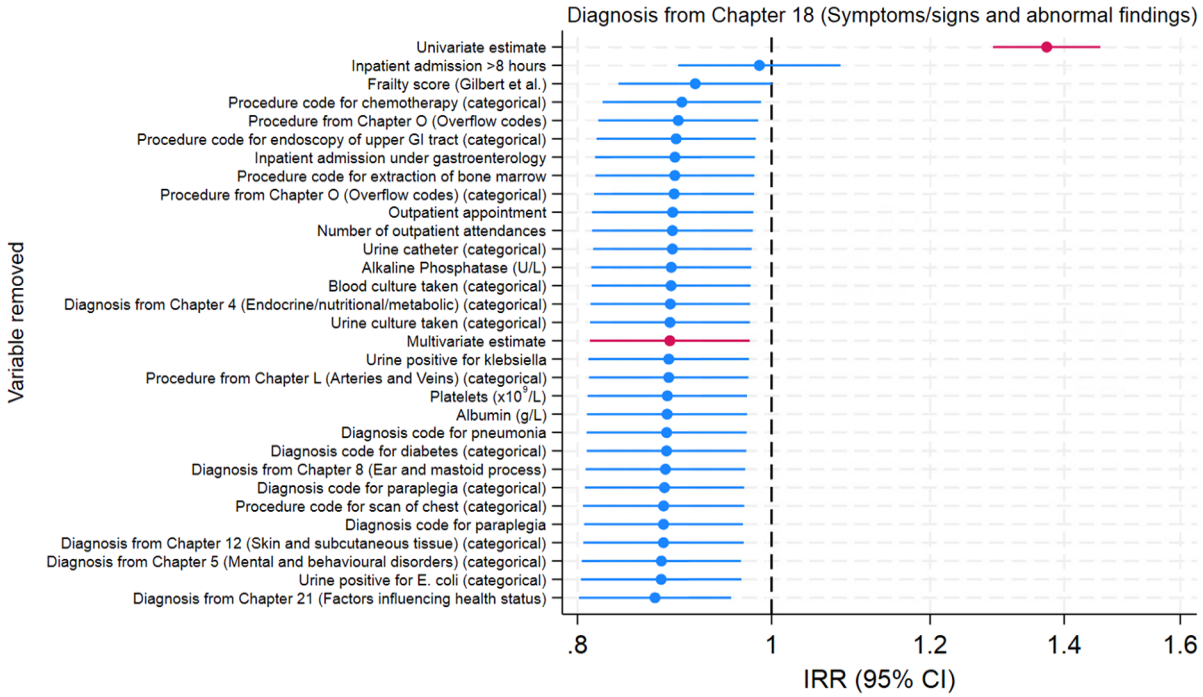
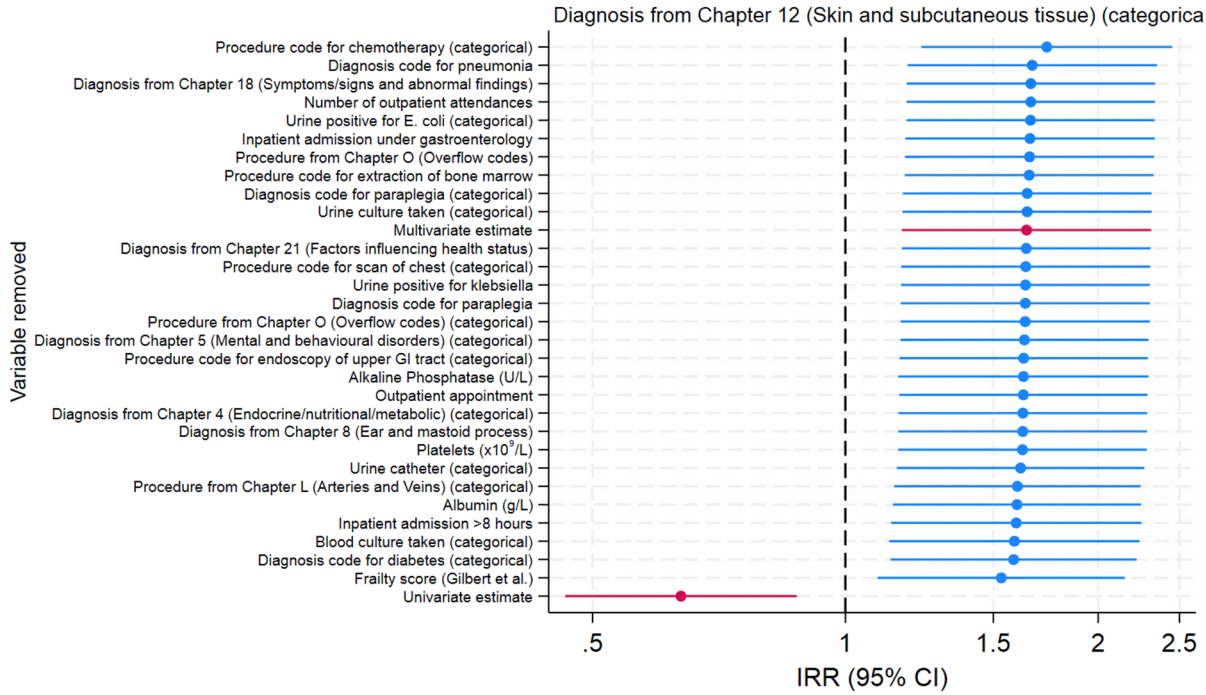



**Figure 5.13: IRR (95% CI) for the effect of never having had a diagnosis code from Chapter 12 versus having a diagnosis code from Chapter 12 ≤365d ago in models removing each variable selected after backwards elimination one at a time.**



*Final model for E. coli BSIs in FY2019*

Receiving chemotherapy ≤365d ago, compared with those never receiving chemotherapy in the previous 5FYs, was associated with a higher risk of *E. coli* BSI (IRR (95% CI): 0.29 (0.19, 0.44) for never versus ≤365d ago) (**Figure 5.14**). The risk was also higher in those receiving chemotherapy >365d ago compared with those never receiving it in the previous 5FYs, but the risk was reduced compared with those receiving chemotherapy more recently (IRR (95% CI): 0.56, 0.31, 1.01 for >365d ago versus ≤365d ago). Additionally, the closer the most recent diagnostic extraction of bone marrow procedure was to the current contact was associated with increased risk. Having a procedure code for endoscopy of the upper GI tract, a code for diabetes, or a blood culture taken was associated with a higher risk of *E. coli* BSI compared to those without these codes in the previous 5FYs. A higher frailty score and procedure codes for Chapter L (arteries and veins) were also associated with increased risk of *E. coli* BSI. The presence of procedures recorded in the Overflow codes chapter was associated with increased risk of *E. coli* BSI the closer they were to the current contact if they were present ≤365d ago, while the risk of *E. coli* BSI was higher if an overflow procedure was present >365d ago versus 365d ago (IRR (95% CI: 2.97, 1.13, 7.82)) (i.e. a “U”-shaped relationship with time across the previous 5FY).

Previous hospital exposure was associated with both increased and decreased risk of *E. coli* BSIs. The risk of *E. coli* BSI was 1.53 times higher (95% CI: 1.38, 1.69) for every 3 months closer an inpatient admission lasting >8 hours was to the current contact (**Figure 5.14**). In contrast, higher numbers of outpatient appointments were associated with decreased risk of *E. coli* BSIs (**Figure 5.15**) and risk was independently 32% lower for individuals with outpatient appointments 3 months versus 6 months ago (95% CI: 11%-48% lower) (**Figure 5.15**). However, the risk of *E. coli* BSI was significantly higher for those with outpatient appointments 0-20 days before the "most recent contact" date compared with appointments 6 months before the "most recent contact" date (**Figure 5.16**).

Having had microbiology tests, specifically those focused on urinary samples, was associated with a higher risk of *E. coli* BSI (**Figure 5.14**). First, any record of a urinary catheter as the source of a microbiological specimen ≤365d ago was associated with a higher risk of *E. coli* BSI, compared to those who had no such record of a urine catheter in the previous 5FYs. The collection of any urine sample for microbiology, regardless of the result, was independently associated with higher risk both for those with a urine culture taken >365d ago and ≤365d ago versus never in the previous 5FYs. In addition, the risk of *E. coli* BSI was 60% lower (95% CI: 46%-70% lower) in those who never had a urine positive for *E. coli* compared with those who had a urine positive for *E. coli* ≤365d ago.

Results from the most recent previous blood tests were also associated with risk of *E. coli* BSIs (**Figure 5.14**, **Figure 5.15**). Lower albumin levels and higher alkaline phosphatase levels (both potentially reflecting poorer underlying status, liver/biliary disease, or for alkaline phosphatase also bone disease including metastatic cancer) were associated with a higher risk of *E. coli* BSI. Higher platelet levels were associated with lower risk and lower platelets with a higher risk of *E. coli* BSI versus the median platelet level (e.g. relating lower blood counts with chemotherapy or blood cancers).

Some characteristics which were associated with decreased risk of *E. coli* BSI reflected other reasons for coming into hospital and therefore potentially competing risks (**Figure 5.15**). A more recent urine positive for *Klebsiella* was associated with a decreased risk of *E. coli* BSI, with risk 26% (95% CI: 1%-44%) lower per three months closer. A diagnosis code for pneumonia and a chest scan were also both associated with a decreased risk of *E. coli* BSI. The risk of *E. coli* BSI was lower if a chest scan was done ≤365d ago vs >365d ago. Within the most recent 365d, the risk of *E. coli* BSI was 16% lower (95% CI: 4%-26%) per 3 months closer a diagnosis code for pneumonia was. These characteristics likely reflected other reasons for attending hospital which did not reflect underlying disorders that would also increase *E. coli* BSI risk.

**Figure 5.14: Final variables selected in FY2019 using the "inpatient only" cohort where the presence of the characteristic, or having had the characteristic more recently, increases the risk of having an *E. coli* BSI**

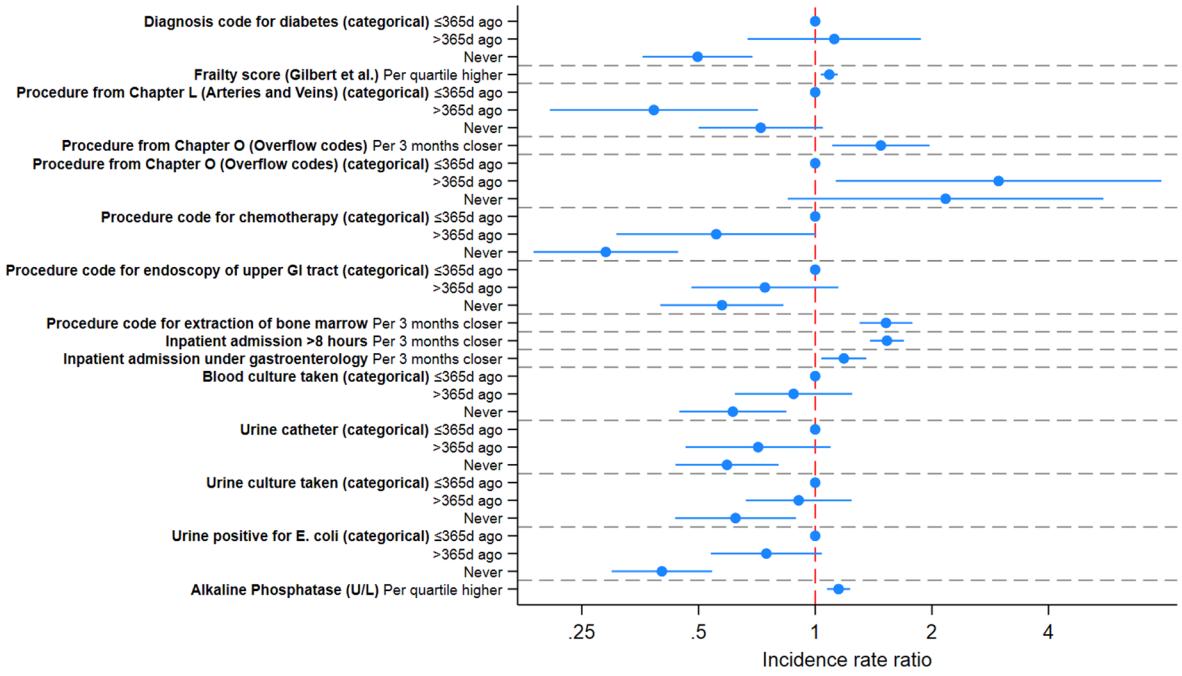


**Figure 5.15: Final variables selected in FY2019 using the "inpatient only" cohort where the presence of the characteristic decreases the risk of having an *E. coli* BSI**

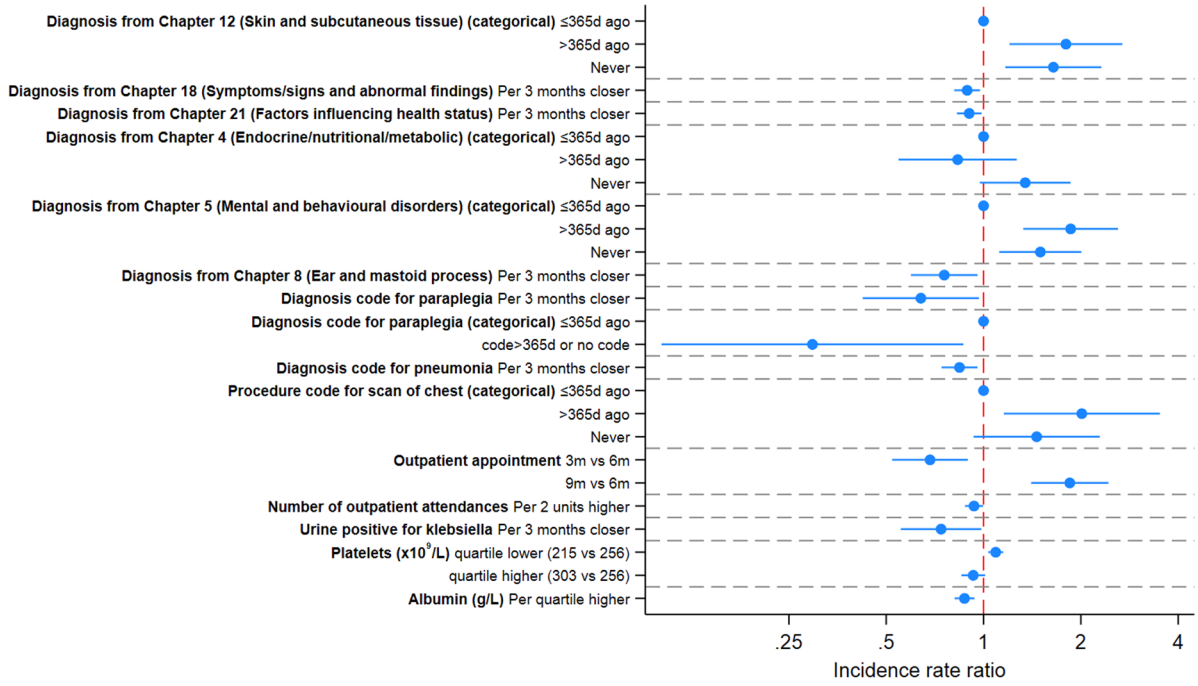
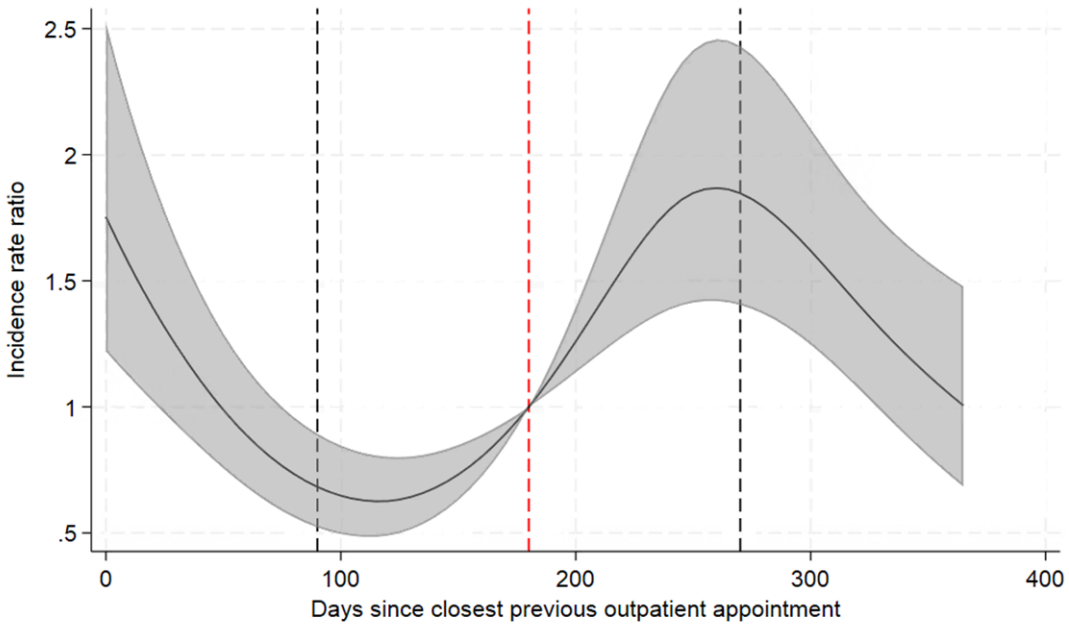



**Figure 5.16: Predicted IRR (95% CI) for the effect of time since the most recent outpatient appointment, as predicted from the multivariable model.**

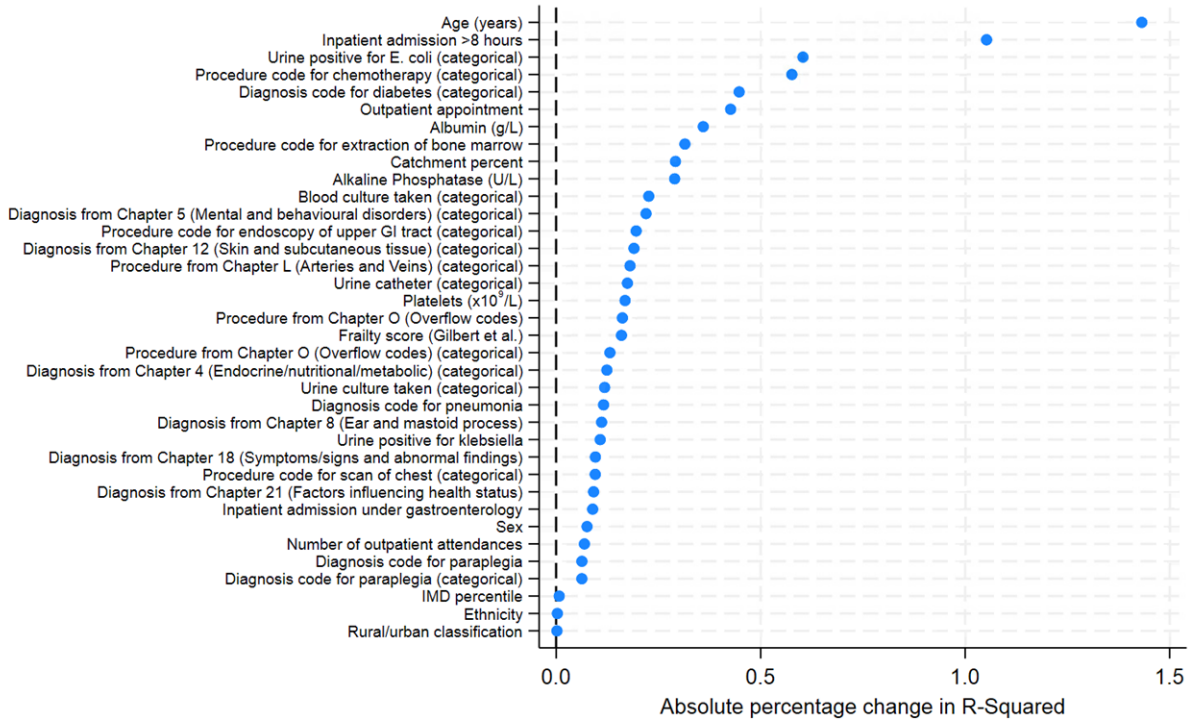



Note: The red dashed line shows the reference level for predictions (6 months). Black dashed lines show where the predictions at 3 months and 6 months, as reported in **Figure 5.15**.

Variables selected for the final model varied in the amount they contributed to explain the variation in the observed data, as measured by the change in R-squared value (**Figure 5.17**). The total R-

squared for the final model was 20.9%, meaning approximately 20.9% of the variability observed in the binary outcome (*E. coli* BSI) could be explained by the variables selected for the final model. Comparing the R-squared value from the final model to the R-squared values from models removing each explanatory variable in turn, removing age had the biggest impact on R-squared, causing an absolute reduction in percentage R-squared of 1.43%. The second biggest impact was removing inpatient admission >8 hours, reducing R-squared by 1.05%, then removing the variable for urine positive for *E. coli* (0.60%), procedure code for chemotherapy (0.58%), diagnosis code for diabetes (0.45%), outpatient appointment (0.43%), albumin (0.36%), procedure code for extraction of bone marrow (0.31%), catchment percentage (0.29%), and alkaline phosphatase (0.29%). Removing the diagnosis code for paraplegia from the model had a very small impact on the R-squared (0.06% absolute percentage change), along with the number of outpatient attendances (0.07% absolute percentage change).

**Figure 5.17: Absolute percentage change in R-squared when removing each variable on the y-axis from the final model.**




Microbiology data recording negative test results may not always be available and hence some risk factors shown in the model above, for example use of urinary catheters as defined from microbiology samples, would not be definable in the same way in nationally available data. However, urinary catheter use can also be obtained from procedure code data. As an aside, I swapped the effect of urinary catheters included in the final model above with urinary catheter use as defined by procedure code variables. In the original model, urinary catheter use was strongly

significantly associated with a higher risk if the catheter had been present ≤365d ago compared with never (IRR = 0.59 [95% CI: 0.44, 0.80 for never versus ≤365d ago; p-value = 0.0008]). Using the procedure code variable for urinary catheter there was no evidence of an association between catheter use and the risk of *E. coli* BSI with an odds ratio of 1.23 (0.72,2.09; p-value=0.448) for catheter use >365d ago versus ≤365d ago and 0.84 (0.57, 1.25; p-value=0.399) for catheter use never versus ≤365d ago.

Considering the raw data, there was little agreement between the coding of urinary catheters between the microbiology and the procedure code data (**Table 5.8**). In general, a higher proportion of urinary catheters were recorded in the microbiology data in both cases and controls. For cases with a microbiology-coded catheter ≤365d ago, only 19/73 (26%) individuals had a catheter similarly coded in procedure codes (22% for the controls). Conversely, 19/31 (61%) of catheters recorded ≤365d ago in procedure code data were also coded ≤365d ago in microbiology data, while this occurred in only 37% of controls (1,080/2,942). The difference in model results was therefore supported by the differences in the underlying data.

**Table 5.8: Comparison of the number and percentage of cases and controls in different urinary catheter groups as recorded in microbiology data and procedure codes data.**

| | | **Cases (N=413)** | | | | **Controls (N=217,934)** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Microbiology data | | | | Microbiology data | | | |
| | | ≤365d ago | >365d ago | Never | **Total** | ≤365d ago | >365d ago | Never | **Total** |
| Procedure codes data | ≤365d ago | 19 (5) | 4 (1) | 8 (1) | **31 (8)** | 1,080 (0.5) | 127 (0.1) | 1,735 (0.8) | **2,942 (1)** |
| | >365d ago | 13 (3) | 11 (3) | 7 (2) | **31 (8)** | 345 (0.2) | 1,094 (0.5) | 2,046 (0.9) | **3,485 (2)** |
| | Never | 41 (10) | 25 (6) | 285 (69) | **351 (85)** | 3,408 (2) | 4,290 (2) | 203,809 (94) | **211,507 (97)** |
| | **Total** | **73 (18)** | **40 (7)** | **300 (72)** | **413 (100)** | **4,833 (2)** | **5,511 (3)** | **207,590 (95)** | **217,934 (100)** |

Note: Percentages are of the total number of cases and controls, i.e. are cell percentages.

*Sensitivity analysis - different entry thresholds*

After the initial univariate screening step, variables needed to be selected to carry forward to the backwards elimination step. The p-value threshold at which to select variables is unclear. Variables with p<0.25 were selected in the primary analysis to avoid missing potentially important effects; however, the impact of multiple testing was not taken into account with this threshold, particularly given the very large deviations from the uniform distribution shown in **Figure 5.8**. Two additional analyses were therefore run including variables with a p-value less than the Benjamini-Hochberg and the Bonferroni adjusted p-value thresholds in the screening step to assess the impact of these thresholds on the final variables selected after backwards elimination.

In the primary analysis, 197 (70%; total=281) of variables had a p-value<0.25. 142 (51%) and 78 (28%) variables had p-values below the Benjamini-Hochberg (BH) and Bonferroni adjusted p-value thresholds after the initial screening step respectively; an absolute percentage reduction in the number of variables of 19% and 42%, respectively. As the groups of variables are all nested within the p<0.25 threshold, the same variables were removed before backwards elimination due to correlation if present at the stricter thresholds. An exit p-value of 0.05 was consistently used for the backwards elimination.

A total of 34 variables were selected after backwards elimination using one or more of the three thresholds (**Table 5.9**). Twelve (35%) variables were selected by all three thresholds. One-fifth of variables (n=7; 21%) were only selected when using the p<0.25 threshold. Four (12%) variables were not selected by the p<0.25 threshold, compared with 10 (29%) and 17 (50%) not selected by the Benjamini-Hochberg and Bonferroni thresholds, respectively.

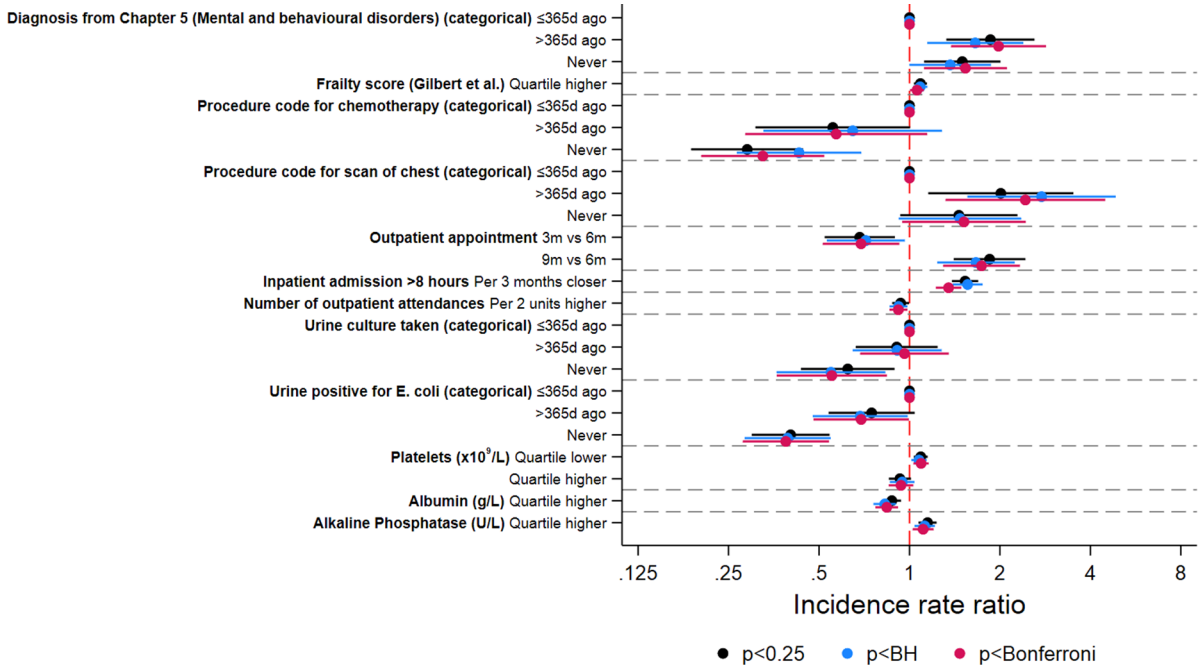
**Table 5.9: Number of variables selected after backwards elimination for different entry p-value thresholds.**

| Category | N selected after backwards elimination (col %) [total = 34] | N considered by relevant models* |
|---|---|---|
| Selected by all 3 p-value thresholds | 12 (35) | 78 |
| Selected by p<0.25 & p<Benjamini-Hochberg | 9 (26) | 142 |
| Selected by p<0.25 & p<Bonferroni | 2 (6) | 78 |
| Selected at p< Benjamini-Hochberg & p<Bonferroni | 2 (6) | 78 |
| Selected at p<0.25 only | 7 (21) | 197 |
| Selected at p<Benjamini-Hochberg only | 1 (3) | 142 |
| Selected at p<Bonferroni | 1 (3) | 78 |

*Number of variables that theoretically could have been selected, i.e. were offered to all relevant models

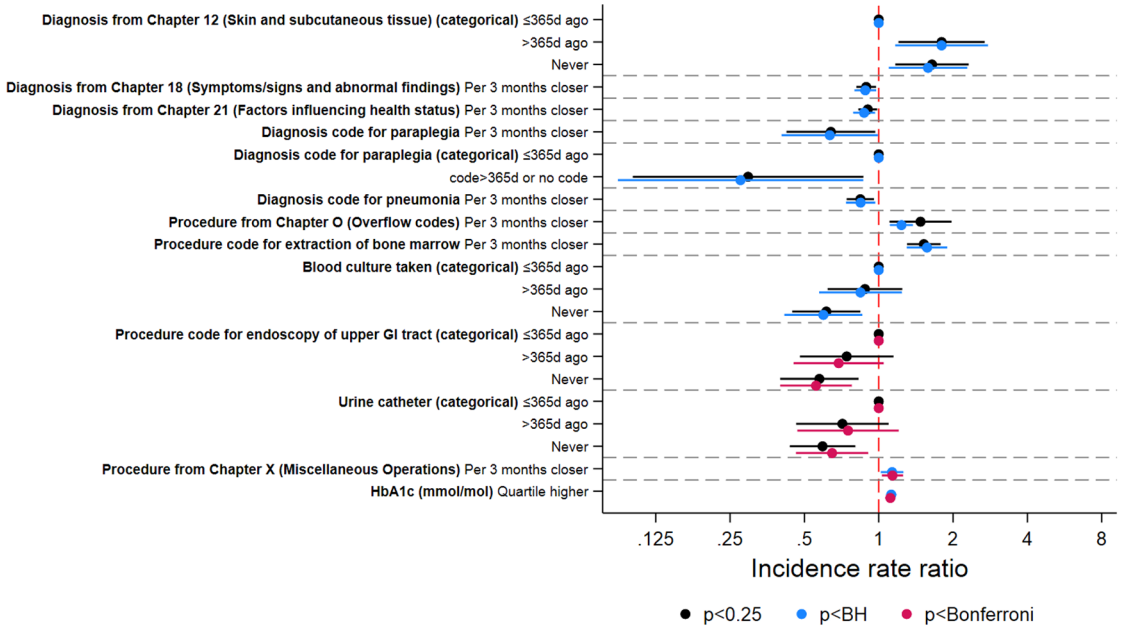
The effect sizes were very similar for variables which were selected by all three thresholds (**Figure 5.18**). A broad variety of characteristics were selected by all three thresholds. Time previously spent in hospital was captured through higher risk for individuals with more recent inpatient admissions lasting >8 hours and for those with outpatient appointments 9 versus 6 months previously. The number of outpatient appointments was associated with reduced risk in all models. The frailty score was identified by all thresholds, as well as higher risk in those receiving chemotherapy in the last year, compared with those never receiving chemotherapy. There were strong effects of having a urine positive for *E. coli* and any urine culture taken. Blood test results for platelets, albumin, and alkaline phosphatase were identified in all models.

**Figure 5.18: IRRs from multivariate models with variables selected using all three p-value thresholds.**



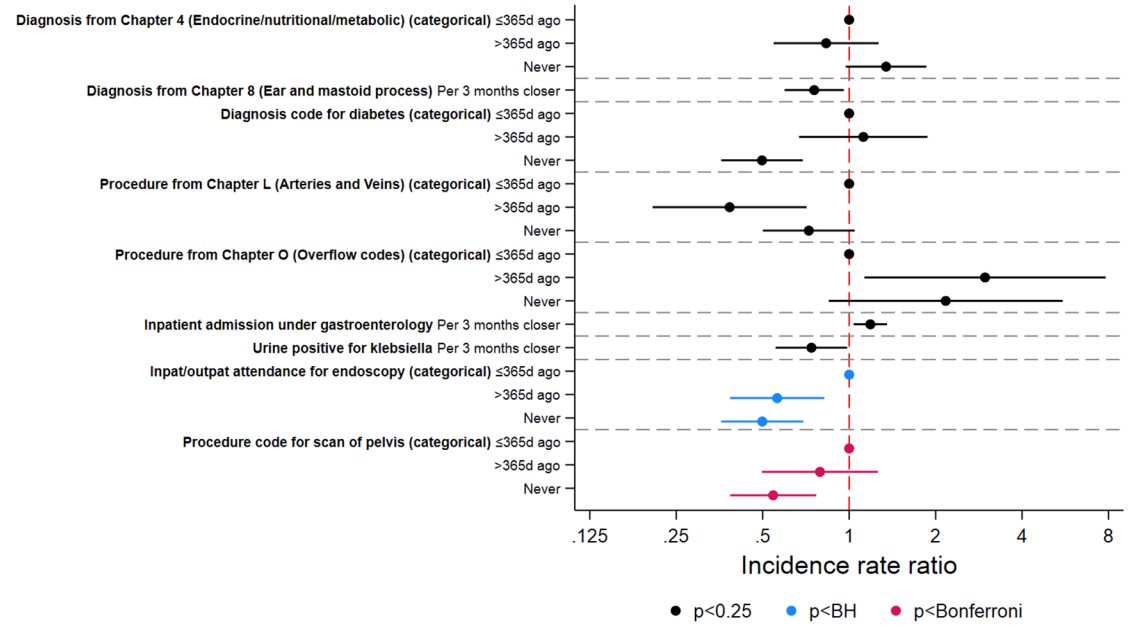

Variables selected after using the p<0.25 threshold and the p<BH threshold captured some specific risk factors not found using the Bonferroni threshold, specifically a higher risk associated with diagnosis codes for paraplegia in the 365d before most recent contact versus a diagnosis code for paraplegia >365d ago or never, and a higher risk in those with a more recent procedure for diagnostic extraction of the bone marrow (**Figure 5.19**). This combination of p-values also captured effects of health-seeking behaviour – for example, diagnosis codes from Chapter 18 (symptoms/signs and abnormal findings) and Chapter 21 (Factors influencing health status). The higher risk associated with recent urinary catheter use was missed by the p<BH threshold but identified by both the p<0.25 and the p<Bonferroni threshold. A procedure code from Chapter X (miscellaneous operations) was missed by the p<0.25 threshold; however, this procedure code chapter mainly contains chemotherapy codes which were already identified by the p<0.25 threshold. Similarly, HBA1c was not identified at the p<0.25 threshold but a diagnosis code for diabetes was instead.

**Figure 5.19: IRRs from multivariate models with variables selected using two p-value thresholds.**




The variables selected only under the p<0.25 threshold were potentially useful and interesting risk factors (**Figure 5.20**). Procedure from Chapter L (arteries and veins) were selected, perhaps indicative of frailty. Closer inpatient admissions under gastroenterology were also associated with a higher risk of *E. coli* BSI. The variables missed by the p<0.25 threshold but selected by the Benjamini-Hochberg and Bonferroni thresholds were endoscopy attendances in the hospital; however endoscopy of the upper GI tract was selected at the p<0.25 threshold (**Figure 5.19**), and procedure codes for scan of the pelvis.

**Figure 5.20: IRRs from multivariate models with variables selected using one p-value threshold only.**



Overall, potentially important risk factors were missed when running the analysis at the stricter BH and Bonferroni thresholds, and risk factors identified from these models but not the p<0.25 backwards elimination were generally reflected in other factors selected using the p<0.25 threshold. Hence selecting variables for the backwards elimination with p<0.25 after the initial screen seemed suitable for all models moving forward. There was no evidence that multiple testing was impacting the model results in the final model, potentially because most factors considered had at least some clinical support.

*Screening on different populations*

In the "any healthcare" cohort, there were 490 (0.1%) cases and 389,435 (99.9%) controls. This cohort included 77 (16%) more cases and 171,501 (44%) more controls than the "inpatient-only" cohort. In the age>65y cohort, subsetted from the "inpatient only" cohort, there were 314 cases and 83,373 controls – 76% and 38% of the "inpatient only" cohort, respectively.

A total of 45 variables were selected after backwards elimination across all three cohorts (**Table 5.10**). While 12 (27%) variables were found in all these three cohorts, 18 variables (40%) were selected in only one of the three cohorts.

**Table 5.10: Number of variables selected after backwards elimination for cohorts.**

| Category | N (col %) [total = 45] | N considered by relevant models* |
|---|---|---|
| Selected in all 3 cohorts | 12 (27) | 167 |
| Selected in the "inpatient only" and the "any healthcare" cohort | 11 (24) | 193 |
| Selected in the "inpatient only" and the "age>65y" cohort | 1 (2) | 167 |
| Selected in the "any healthcare" and the "age>65y" cohort | 3 (7) | 167 |
| Selected in the "inpatient only" cohort only | 6 (13) | 197 |
| Selected in the "any healthcare" cohort only | 6 (13) | 193 |
| Selected in the "age>65y" cohort only | 6 (13) | 167 |

*Number of variables that theoretically could have been selected, i.e. were offered to all relevant models

When variables were selected in multiple cohorts, the estimated effects of risk factors were similar between cohorts (**Figure 5.21**; **Figure 5.22**). Higher frailty scores were associated with a higher risk of *E. coli* BSIs in both the "inpatient only" and "any healthcare " cohorts, however when subsetting the data to only those aged>65y, the effect of frailty was not significant at the p<0.05 threshold (**Figure 5.22**). When adding the frailty score on top of the selected variables for the age>65y cohort, risk was increased 1.56 times in cases versus controls per quartile increase in frailty score, however confidence intervals were large and overlapped one (95% CI: 0.77, 3.18; p=0.219). In the age>65y

cohort, some specific risk factors related to age were identified, specifically outpatient attendances under geriatric consultants being associated with increased risk of *E. coli* BSIs (**Figure 5.23**). Procedure codes from Chapter X (miscellaneous) were also associated with a higher risk of *E. coli* BSIs in those aged >65y with the majority of codes within this chapter recording chemotherapy. A higher risk of *E. coli* BSI was also seen at taller heights in the age>65y cohort. This could be a collinearity issue with men being at a higher risk of *E. coli* BSI and men being, on average, taller. This was demonstrated through the attenuation of the effect of being female versus male between the core model (IRR 0.79 (95% CI: 0.63, 1.00; p-value = 0.051)) and the fully adjusted model (1.06 (95% CI 0.72, 1.56; p-value = 0.775)). Interestingly, increases in risk were associated with higher CRP and closer outpatient attendances under renal medicine in the "any healthcare" cohort only. The "inpatient only" cohort identified any blood and urine culture collection (irrespective of result) in the 365d before the most recent contact to be at a higher risk for *E. coli* BSI compared with never having had a culture taken while the other cohorts did not select these as risk factors.

**Figure 5.21: Variables selected in all three cohorts.**

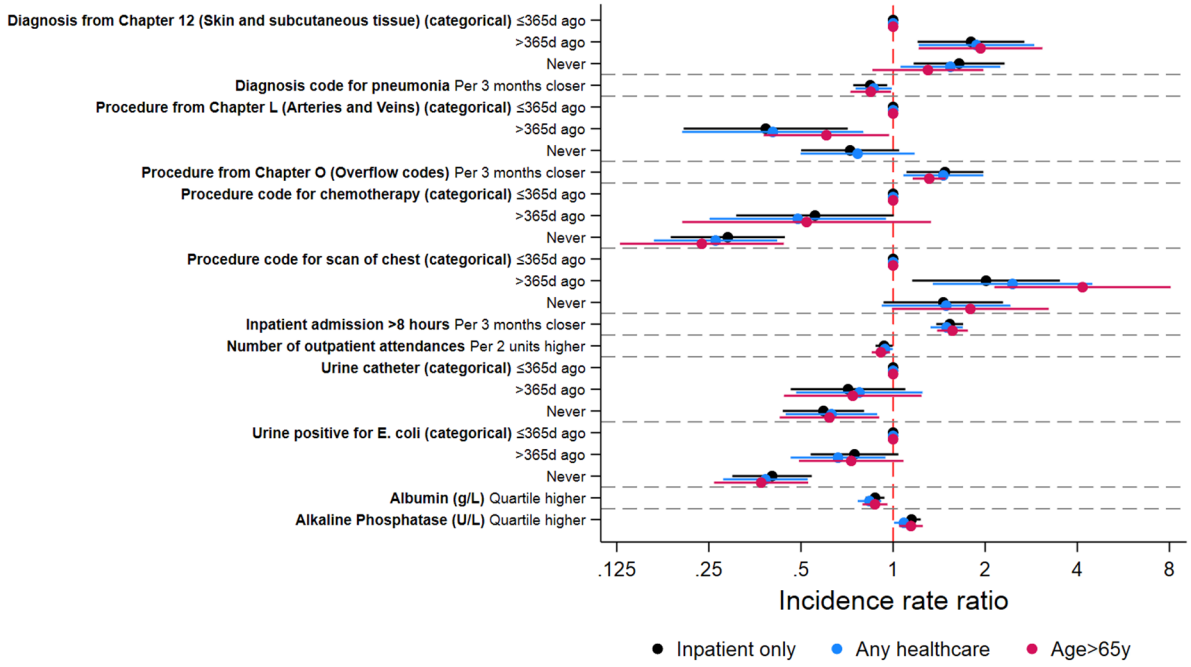


**Figure 5.22: Variables selected in two of the three cohorts.**

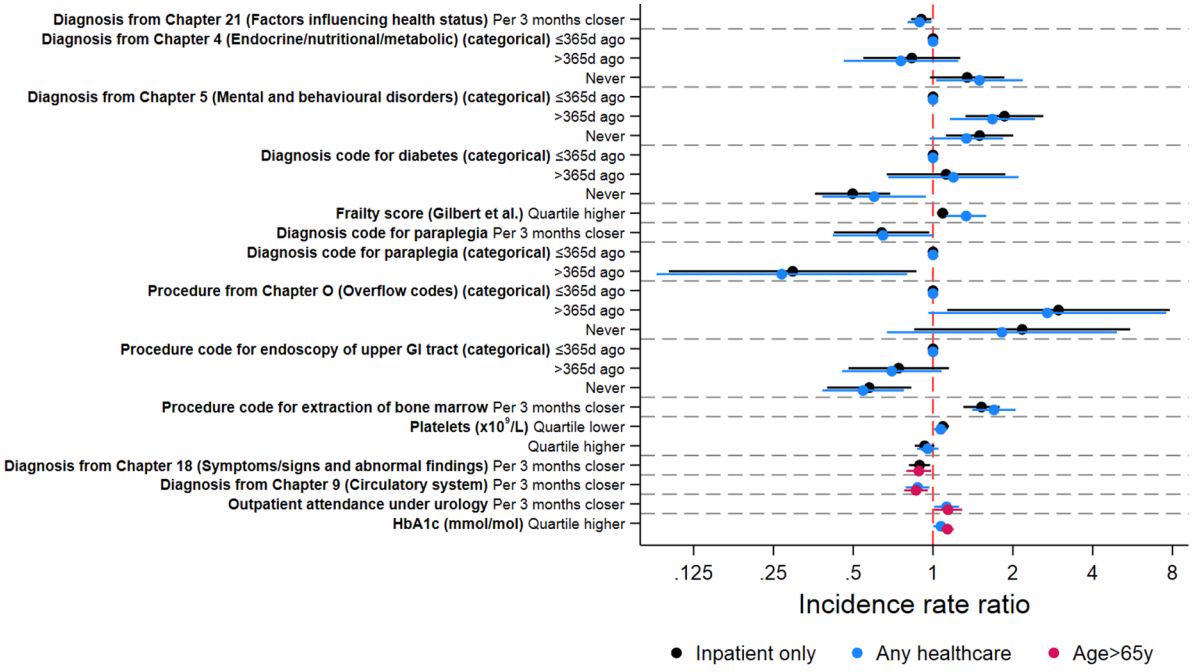
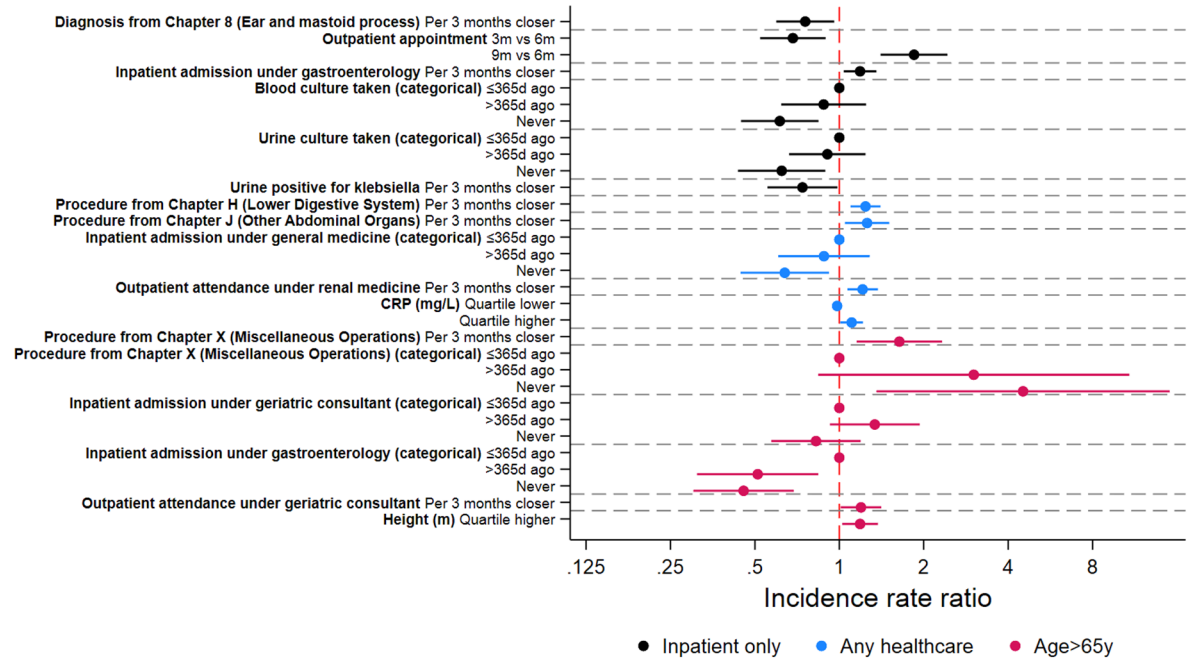



**Figure 5.23: Variables selected in one of the three cohorts only.**




*Using different outcomes*

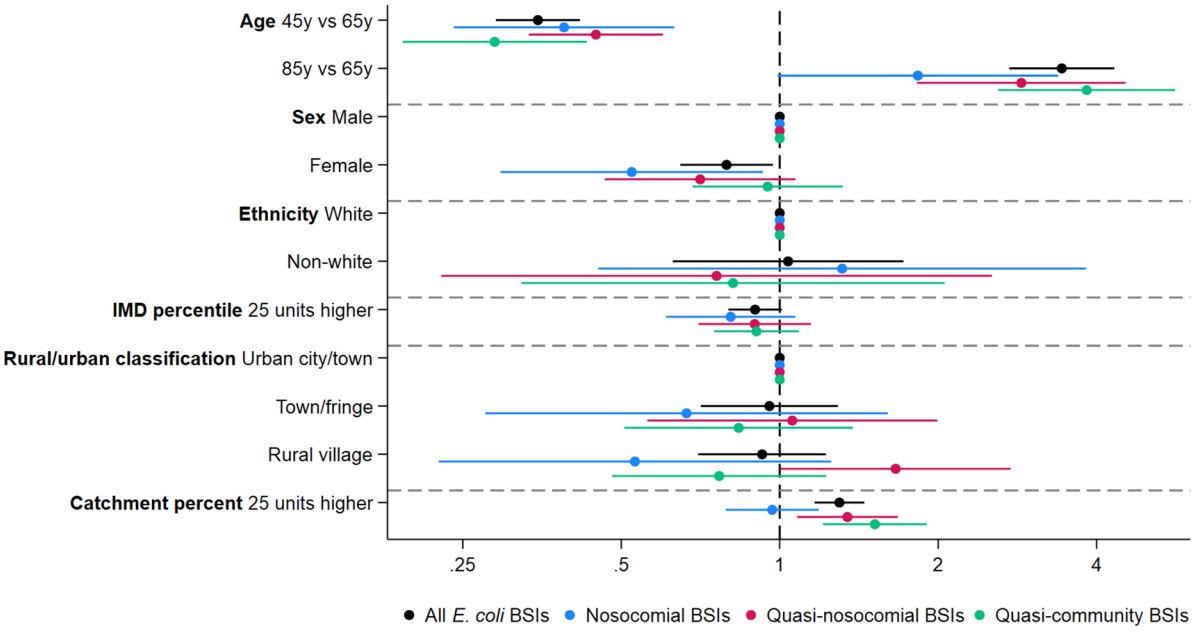
**Nosocomial and community infections**

Of the 413 *E. coli* BSIs in the "inpatient only" cohort in FY2019, 66 (16%) infections were nosocomial, 98 (24%) were quasi-nosocomial, and 150 (36%) were quasi-community. The remaining 99 (24%) were community-associated. Due to the nature of the risk factors defined in this study looking for

associations between *E. coli* BSIs and characteristics in the previous 365d, almost all risk factors had to be excluded for community *E. coli* BSIs, including all diagnosis code and procedure code variables, as by definition community-associated *E. coli* BSIs have no contact as an inpatient in the previous 365d (total of 359/410 (88%) of variables dropped). Due to the small number of risk factors available to screen (51/410, 12%), risk factors for community-acquired *E. coli* BSIs are therefore not presented.

Estimates from the core model varied across the different outcome groups using the "inpatient only" cohort (nosocomial, quasi-nosocomial, and quasi-community), particularly for age and rural/urban classification (**Figure 5.24**). The effect of age attenuated at higher ages for nosocomial infections, compared with all other sub-groups; for example, IRR 1.83 (95% CI: 0.99, 3.38) versus 3.83 (95% CI: 2.60, 5.64) for those aged 85y versus 65y in nosocomial BSIs compared with quasi-community BSI, respectively. The effect of sex was larger for nosocomial *E. coli* BSIs compared with the estimate using all *E. coli* BSIs, however confidence intervals were wide. Small numbers of individuals of non-white ethnicity contributed to wide confidence intervals in the nosocomial, quasi-nosocomial, and quasi-community subgroups. Considering the primary outcome of all *E. coli* BSIs, there was no evidence of an effect of rural-urban classification on the risk of *E. coli* BSI (global p-value = 0.849, individual p-values>0.588). While there was no overall evidence of an effect of rural-urban classification for the quasi-nosocomial group, the risk of quasi-community *E. coli* BSIs was higher in those living in rural areas versus urban cities/towns (1.66 [95% CI: 1.00, 2.75]; p=0.048). There was no evidence of an effect of catchment percentage for nosocomial infections versus controls (as expected since these infections are acquired after admission), while there was a higher risk associated with higher catchment areas for all other outcome groups consistent with the estimate including all *E. coli* BSIs.

**Figure 5.24: Estimates from core models with different outcomes based on nosocomial, quasi-nosocomial, and quasi-community *E. coli* BSIs.**




Note: "Inpatient only" cohort used for all models.

Due to the reduced number of cases in all the nosocomial, quasi-nosocomial, and quasi-community groups, a higher number of variables were dropped before the initial screening step as they had ≤8 occurrences in the 365d before the *E. coli* BSI (**Table 5.11**). The number of variables dropped was proportional to the number of cases; using the outcome of nosocomial BSIs removed the largest number of variables. Variables with few observations >365d ago were combined with the never category, affecting 11 variables when using all *E. coli* BSIs, increasing to 60 variables using nosocomial BSIs. An additional small number of categorical variables had to be dropped as there were ≤8 occurrences after combining the >365d ago and never groups. This was the case for 2 variables when using nosocomial BSIs and 6 variables using quasi-nosocomial BSIs; these were variables based on previous hospital exposure as, by definition, cases had to have an inpatient admission close to the *E. coli* BSI collection.

**Table 5.11: The number of variables dropped from analyses due to small numbers.**

| Population (number of cases) | Number of variables dropped as ≤8 occurrences ≤365d ago | Number of variables collapsed as ≤8 occurrences >365d ago | Number of variables dropped as ≤8 occurrences ≤365d ago |
|---|---|---|---|
| All *E. coli* BSIs (413) | 122 | 11 | 0 |
| Nosocomial BSIs (66) | 226 | 60 | 2 |
| Quasi-nosocomial BSIs (98) | 194 | 35 | 6 |
| Quasi-community BSIs (150) | 166 | 19 | 0 |

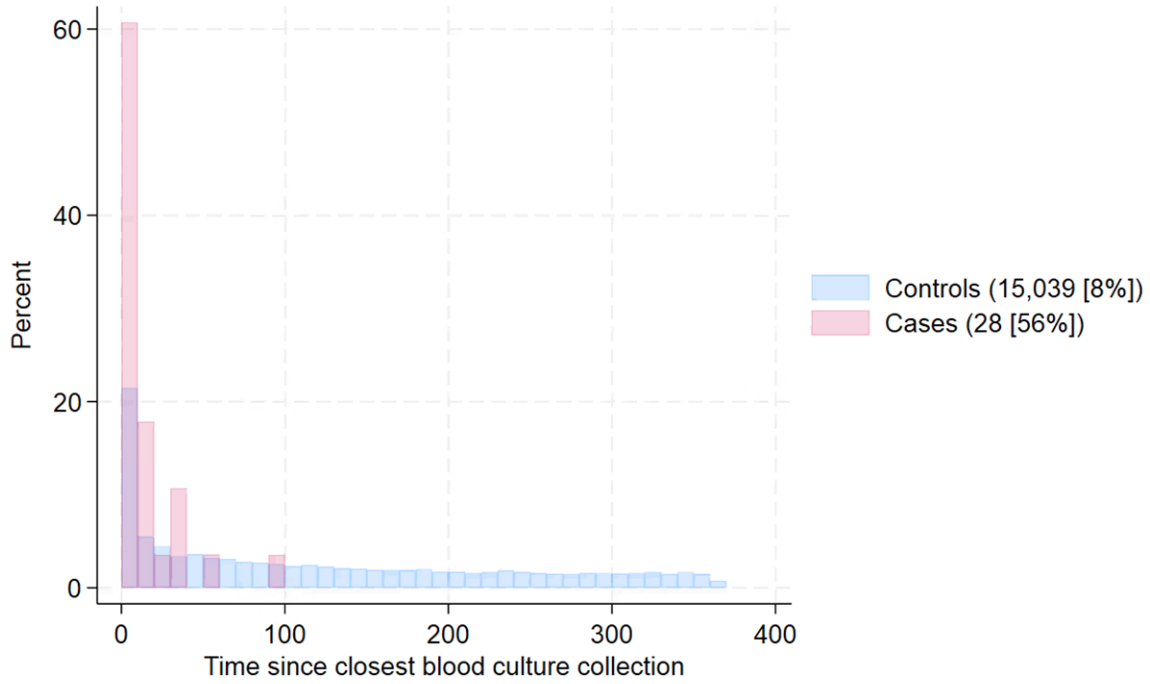
The initial screening step was carried out followed by backwards elimination (exit p>0.05) on all variables from the screening step with p<0.25.

After backwards elimination, there were large IRRs in the multivariable models for the nosocomial and quasi-nosocomial infections. When using the nosocomial *E. coli* BSIs as an outcome, there was a very large effect of having had a previous blood culture taken (**Table 5.12**). Upon closer investigation into the raw distribution of this variable, almost all the *E. coli* cases with a previous blood culture had this taken within 3-20 days before the blood culture which tested positive for *E. coli* was collected (**Figure 5.25**). Similarly, for quasi-nosocomial BSIs, there was a similar effect and distribution for a procedure code for prosthesis (**Table 5.12**; **Figure 5.26**). For ease of interpretation, the time since the closest blood culture was taken and the procedure code for prosthesis were categorised for presentation of the final model results with categories based on the distributions in **Figure 5.25** and **Figure 5.26**.

**Table 5.12: Incidence rate ratios for large effects from nosocomial and quasi-nosocomial multivariate models.**

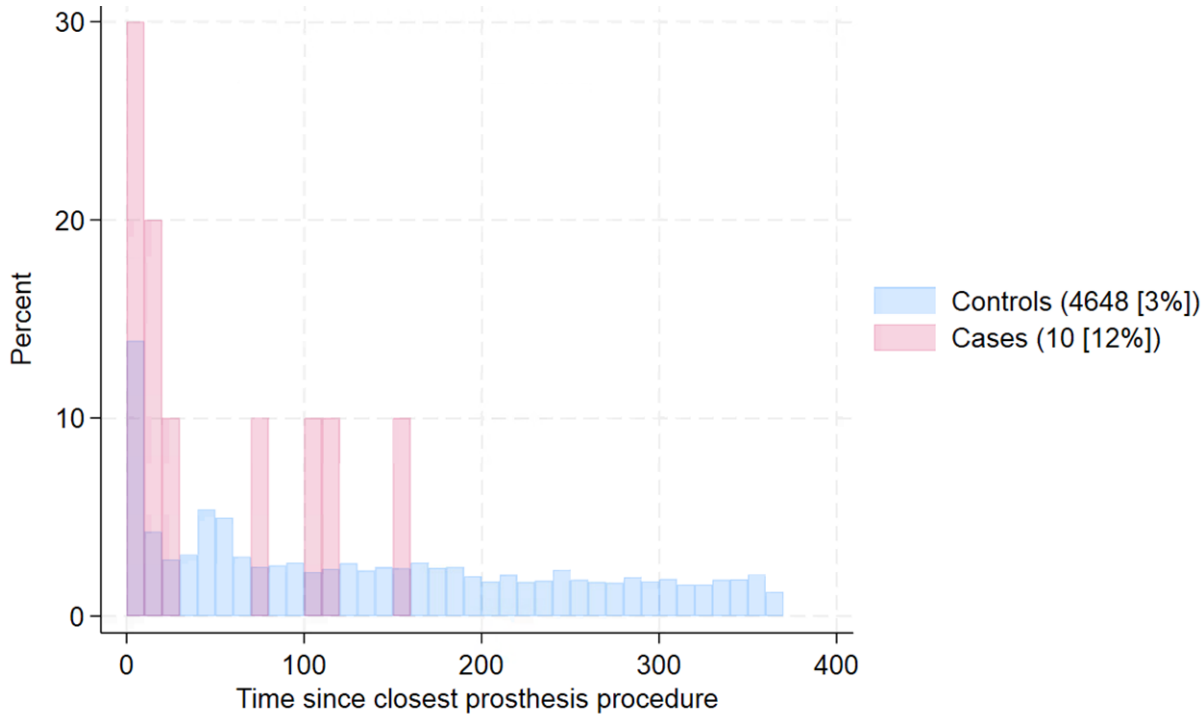
| Characteristic | Level | IRR (95% CI) | Global p-value |
|---|---|---|---|
| **Nosocomial** | | | |
| Blood culture taken | Per 3 months closer | 31 (3.5, 280) | 0.002 |
| Blood culture taken (categorial) | ≤365d ago<br>>365d ago or never | 1<br>257,552 (57, $1.172 \times 10^9$) | 0.004 |
| **Quasi-nosocomial** | | | |
| Procedure code for prosthesis | Per 3 months closer | 297 (6.8, 13,031) | 0.003 |
| Procedure code for prosthesis (categorical) | ≤365d ago<br>>365d ago or never | 1<br>$1.5 \times 10^9$ (514, $4.4 \times 10^{15}$) | 0.005 |

**Figure 5.25: Distribution of days since the closest blood culture collection in the 365d before the most recent contact for cases (nosocomial *E. coli* BSIs) and controls.**




Note: Bin width = 10 days

**Figure 5.26: Distribution of days since the closest procedure code for prosthesis in the 365d before the most recent contact for cases (quasi-nosocomial E. coli BSIs) and controls.**




Note: Bin width = 10 days

A total of 48 variables were found across the three multivariate models run with the outcomes of nosocomial, quasi-nosocomial, and quasi-community *E. coli* BSIs. 15 (31%) variables were found in at least one of the nosocomial, quasi-nosocomial, and quasi-community sub-groups and were previously identified using all *E. coli* BSIs as an outcome (**Table 5.13**). Only two variables (albumin levels and urine positive for *E. coli*) were selected using all three subgroups and all *E. coli* BSIs as outcomes. The remaining variables (33/48; 69%) found were unique to each outcome and were not also found using all *E. coli* BSIs as an outcome with no direct overlap between the risk factors using these different outcomes.

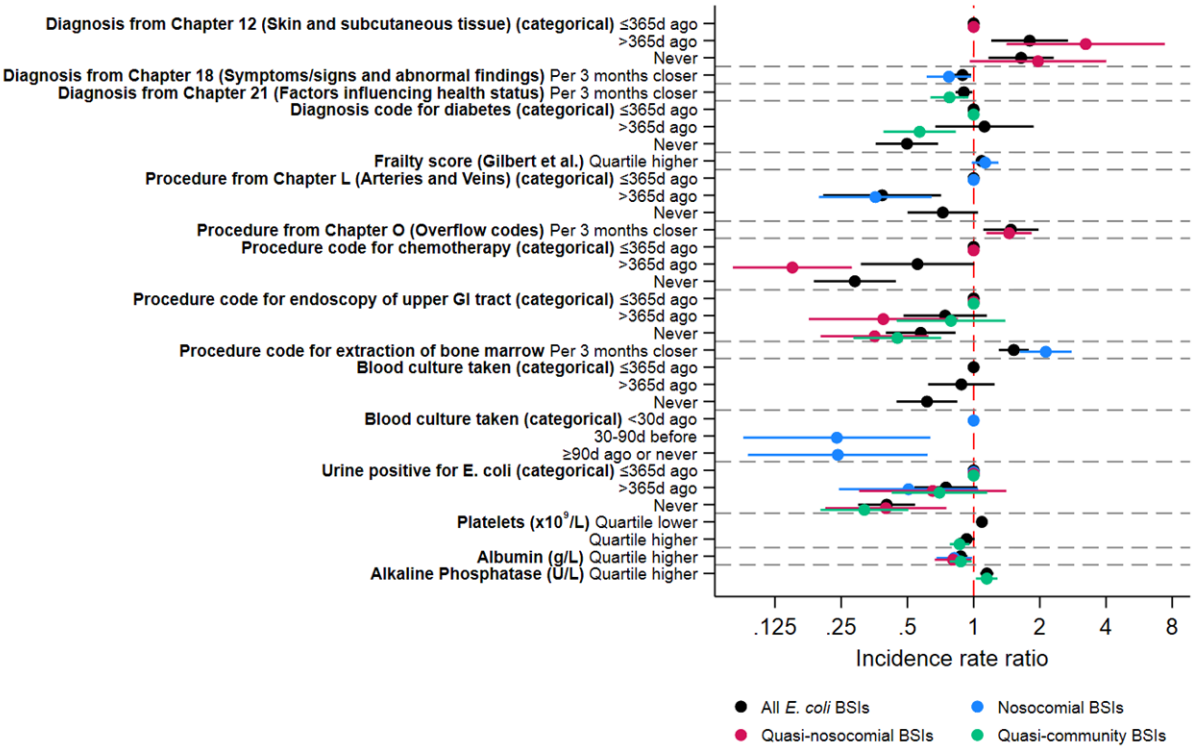
**Table 5.13: Number of variables selected after backwards elimination using nosocomial, quasi-nosocomial, and quasi-community *E. coli* BSIs as outcomes.**

| Previously identified in the model using all *E. coli* BSIs as an outcome? | Category | n (%) (N=48) |
|---|---|---|
| **Yes** | Nosocomial, quasi-nosocomial, and quasi-community | 2 (4) |
| | Nosocomial and quasi-nosocomial | 0 (0) |
| | Nosocomial and quasi-community | 0 (0) |
| | Quasi-nosocomial and quasi-community | 1 (2) |
| | Nosocomial only | 5 (10) |
| | Quasi-nosocomial only | 3 (6) |
| | Quasi-community only | 4 (8) |
| **No** | Nosocomial, quasi-nosocomial, and quasi-community | 0 (0) |
| | Nosocomial and quasi-nosocomial | 0 (0) |
| | Nosocomial and quasi-community | 0 (0) |
| | Quasi-nosocomial and quasi-community | 0 (0) |
| | Nosocomial only | 9 (19) |
| | Quasi-nosocomial only | 10 (21) |
| | Quasi-community only | 14 (29) |

Note: All models used the inpatient-only cohort.

Estimates from the *E. coli* BSI subgroups were broadly similar to those found by the model including all *E. coli* BSIs if the variables were included in both final multivariable models (**Figure 5.27**), although tended to shift slightly further from the null. Some estimates slightly shifted due to levels of the variables having to be grouped in the subgroups because of small numbers; for example, the categorical effects of diabetes or a chemotherapy code >365d ago and/or never. Both of these variables still showed higher risk if the characteristic was present ≤365d ago however. The effect of having a closer procedure for the extraction of bone marrow was slightly stronger for nosocomial BSIs versus all BSIs, but 95% CI confidence intervals were overlapping.

**Figure 5.27: IRRs and 95% CIs for all estimates in multivariable models which were previously found using all E. coli BSIs as the model outcome.**

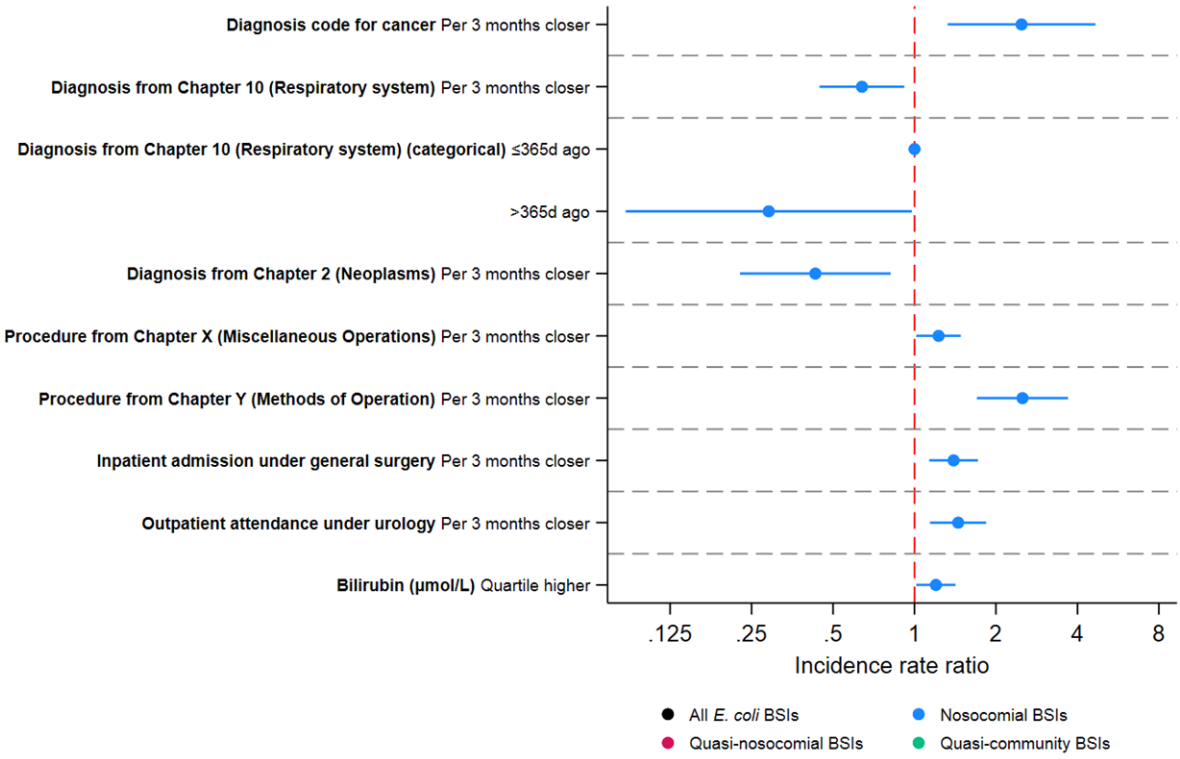



Note: The estimates from the multivariable model using all E. coli BSIs as an outcome are adjusted for all variables shown in **Figure 5.14** and **Figure 5.15**. All models are adjusted for the core variables.

Characteristics selected only when using the nosocomial BSI outcome included additional variables related to cancer, the respiratory system, and surgery (**Figure 5.28**). Interestingly, having a recent inpatient admission under general surgery was associated with an increased risk of nosocomial *E. coli* BSI. Surgery was not explicitly found using all *E. coli* BSIs. A diagnosis code for cancer was associated with an increased risk of nosocomial BSI, while a diagnosis code from Chapter 2 (Neoplasms) was associated with a decreased risk, perhaps indicating some residual collinearity. A close diagnosis code for respiratory system diagnoses was associated with a decreased risk of nosocomial *E. coli* BSI, however the risk was higher at 365d ago compared with those who had a

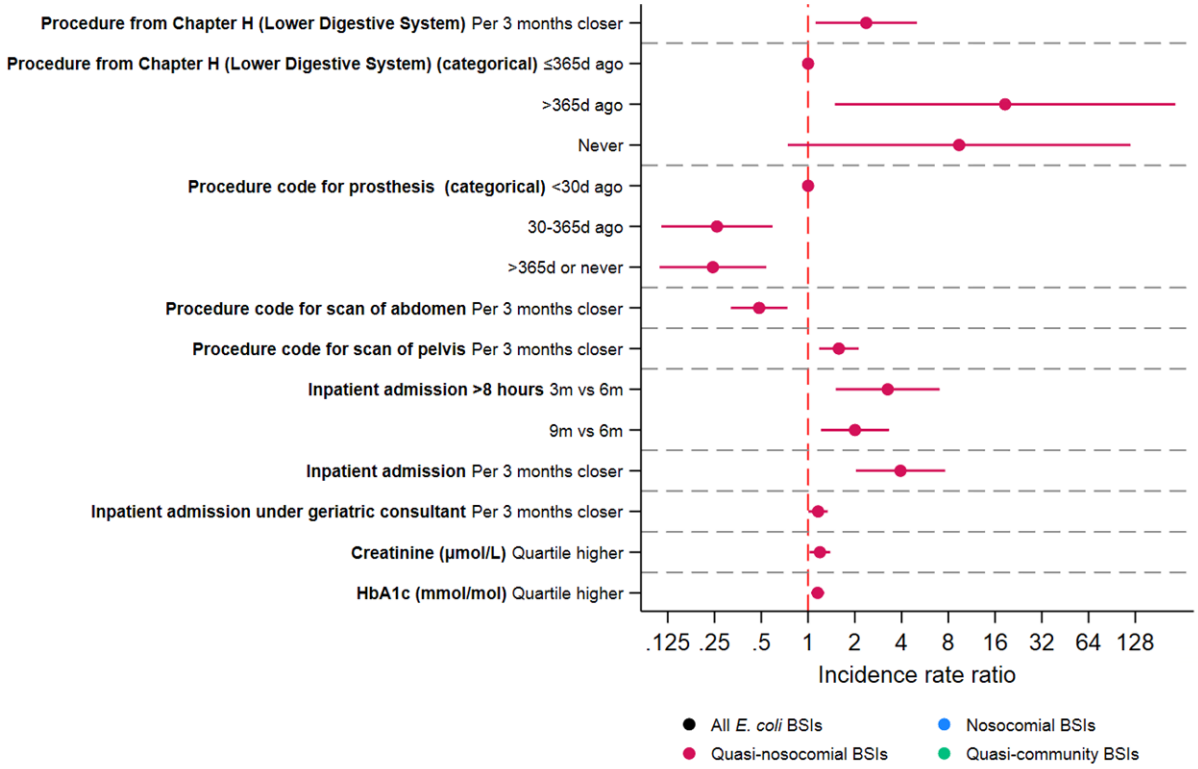
diagnosis code >365d ago or never. Higher bilirubin levels were associated with a higher risk of nosocomial *E. coli* BSI.

**Figure 5.28: IRRs and 95% CIs from models using nosocomial BSIs as an outcome for variables which were not found using other *E. coli* outcomes.**



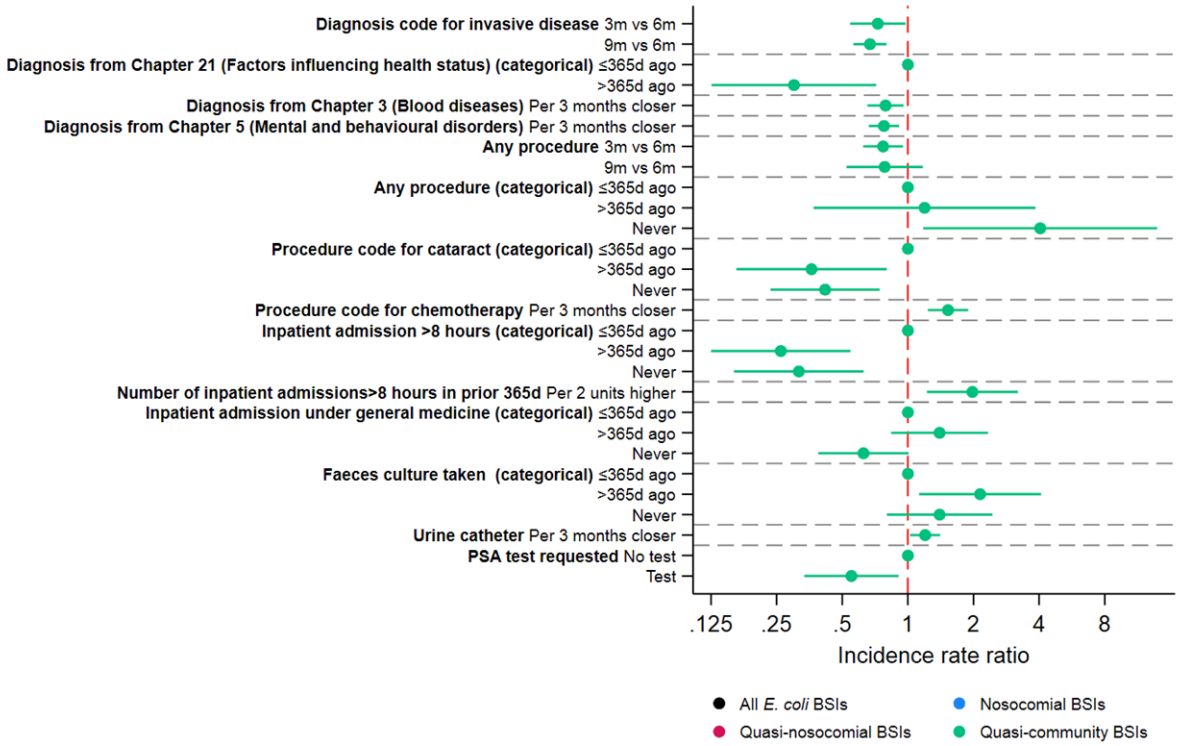

Characteristics selected only for quasi-nosocomial *E. coli* BSIs included a variety of different risk factors (**Figure 5.29**). Having any inpatient admission or an inpatient admission lasting >8 hours was associated with an increased risk if that admission was closer to the current contact, perhaps indicative of the definition of quasi-nosocomial BSIs. Risk was also higher at 9 versus 6 months or those with an inpatient admission lasting >8 hours, suggesting a "U-shaped" distribution of risk. Variables for scan of the abdomen and scan of the pelvis were selected but with the effects going in opposite directions in the multivariate model, likely reflecting correlation between these variables and collinearity. Having a recent procedure code for prosthesis in the previous 30d was associated with an increased risk of quasi-nosocomial BSI compared with those having a code for prosthesis 30-365d ago and >365d or never. Having a recent procedure code for the lower digestive system was also associated with increased risk of quasi-nosocomial *E. coli* BSIs; however this appeared to have attenuated by 365d with those having a code >365d ago at a higher risk than those 365d ago (i.e. a "U"-shaped risk). Seeing a geriatric consultant recently increased the risk of quasi-nosocomial *E. coli* BSI, along with higher creatinine and higher HbA1c.

**Figure 5.29: IRRs and 95% CIs from models using quasi-nosocomial BSIs as an outcome for variables which were not found using other *E. coli* outcomes.**




Risk factors only for quasi-community *E. coli* BSIs again included a range of characteristics (**Figure 5.30**). Similarly to quasi-nosocomial BSIs, previous inpatient admissions were associated with a higher risk of quasi-community *E. coli* BSI. Never having had a previous procedure recorded in IORD in the last 5FYs was associated with a higher risk of quasi-community *E. coli* BSI compared to a procedure 365d ago. Closer proximity of codes reflecting invasive diseases, blood diseases, and mental and behavioural disorders were associated with decreased risk of quasi-community *E. coli* BSI perhaps due to the quasi-community definition applied to the cases (although controls still had recent hospital contact). A faeces culture >365d ago was associated with a higher risk of quasi-community *E. coli* BSI compared to those with a faeces culture ≤365d ago. The closer proximity of a urinary catheter code from a microbiological specimen was associated with a higher risk of quasi-community *E. coli* BSI. Having a PSA test taken was associated with a decreased risk of quasi-community *E. coli* BSI.

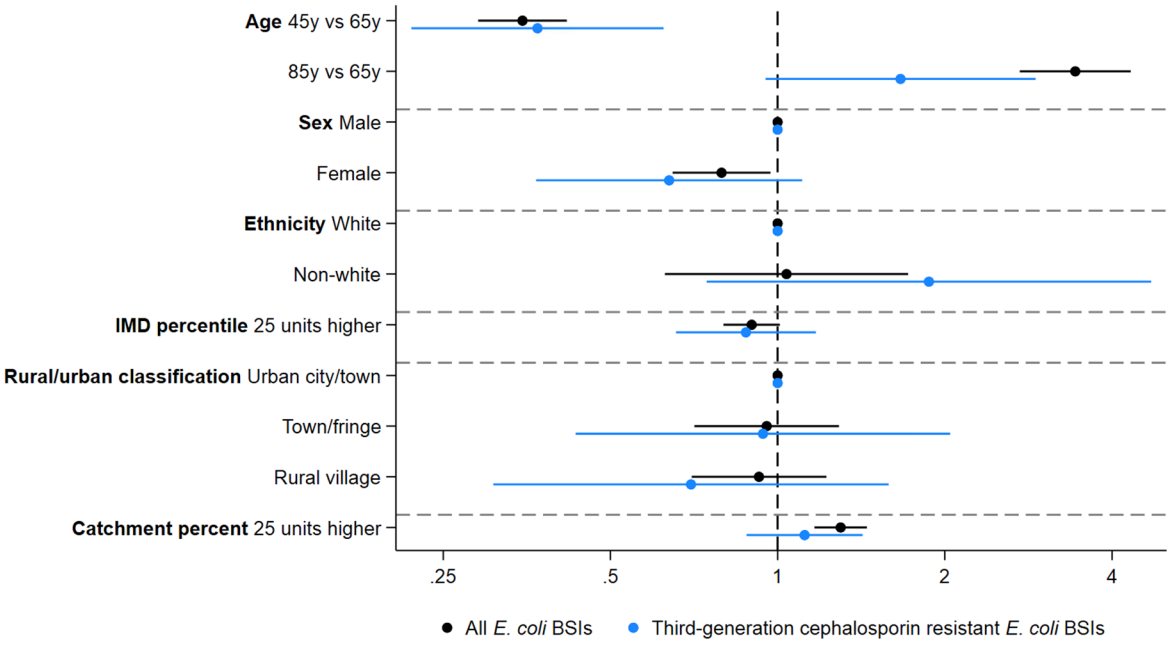
**Figure 5.30: IRRs and 95% CIs from models using quasi-community BSIs as an outcome for variables which were not found using other *E. coli* outcomes.**




### Antibiotic-resistant infections

In FY2019, in the inpatient-only cohort, there were 51 *E. coli* BSIs resistant to third-generation cephalosporins. This constituted 12% of all *E. coli* BSIs within the year. The effect of age on risk of third-generation cephalosporin resistant BSIs attenuated for higher ages compared with the outcome of all *E. coli* BSIs (**Figure 5.31**). All other estimates from the core model were broadly similar to that of the inpatient-only cohort given the larger confidence intervals due to the small number of *E. coli* BSIs resistant to third-generation cephalosporins.

**Figure 5.31: Comparison of core model estimates using the outcomes of all *E. coli* BSIs and third-generation cephalosporin-resistant *E. coli* BSIs.**

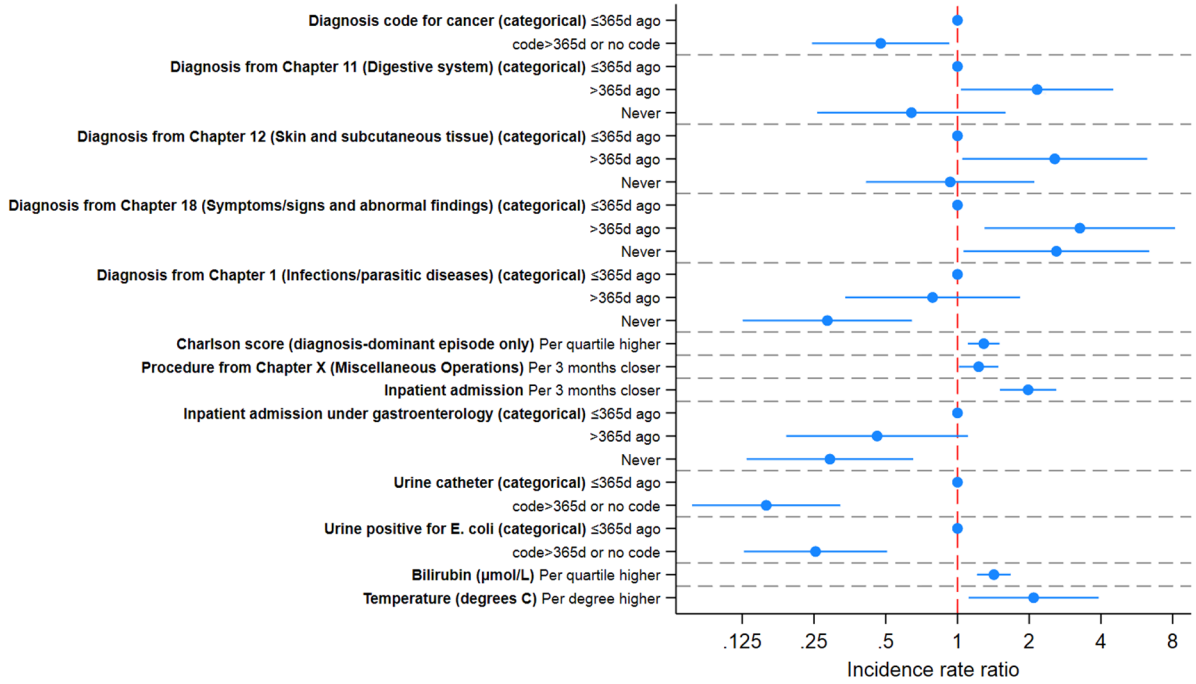



Note: Both populations were using the "inpatient-only" cohort.

A large proportion of variables had to be either dropped or have categorical levels collapsed due to the small number of cases in FY2019. 236 variables were not considered in the univariate screen as there were ≤8 occurrences in the 365d before *E. coli* BSI. A further 46 categorical variables had to be collapsed due to ≤8 occurrences >365d ago. 165 variables were taken forward to the univariate screen, with 117 (71%) having a global p-value<0.25 and thus carried forward to the backwards elimination step (compared with 281 and 197 (70%) respectively for the all *E. coli* cases model).

The majority of the 13 variables selected after backwards elimination when using resistant *E. coli* BSIs as cases were similar to variables selected for all *E. coli* cases but with a couple of additional variables perhaps more indicative of frailty and/or acute illness (**Figure 5.32**). Again, cancer and previous inpatient admissions were strongly associated with risk of resistant *E. coli* BSIs, as well as infection-specific markers, such as diagnosis codes from Chapter 1 (Infections/parasitic diseases) and previous urine cultures positive for *E. coli*. The Charlson score using diagnostic codes only from the closest previous admission and higher temperature were associated with higher risk, perhaps indicating an impact of more background comorbidity (potentially leading to more antibiotic use in the community). Additional markers of frailty including higher bilirubin levels were also associated with resistant *E. coli* BSIs.

**Figure 5.32: IRRs (95% CI) from the final model using E. coli BSIs resistant to third-generation cephalosporins at the outcome.**

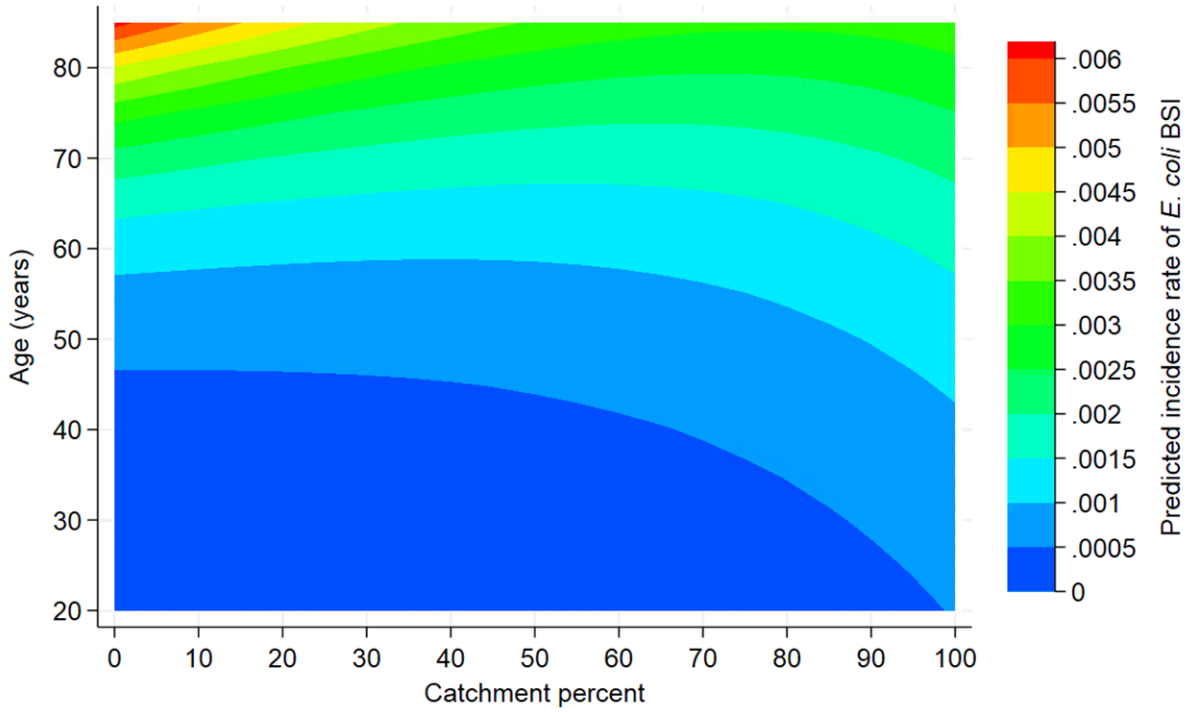



*Note: The model was adjusted for the core variables. This model was run on the inpatient-only cohort.*

### 5.3.3 Expanding the screening process to FYs 2018, 2020, and 2021

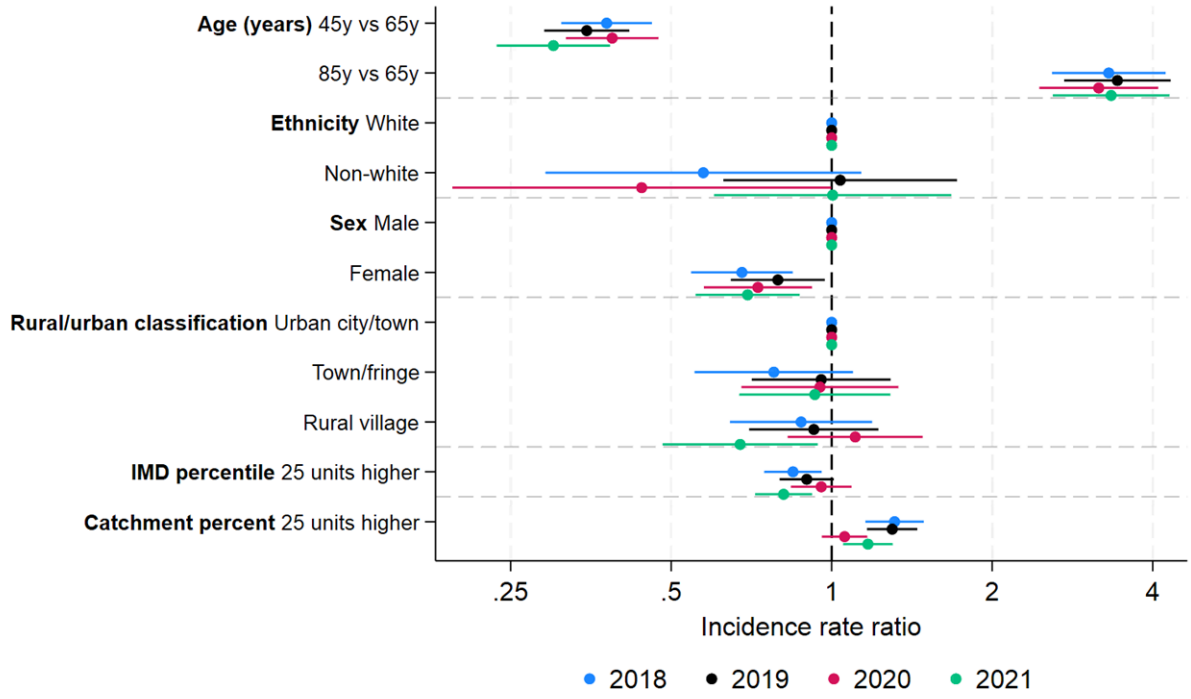
Across all the years of data, only one interaction was significant at the p-value<0.002 threshold in the core model. This was an interaction between age and catchment percentage in FY2018. Considering the model including this interaction, the risk of *E. coli* BSI was higher at higher catchment areas for younger ages, but higher at lower catchment areas for older ages (**Figure 5.33**, p-value for interaction=0.00013). Risk was always higher at older than younger ages regardless of catchment. This was the only significant interaction in the core model across all FYs, and there was no evidence of this interaction between age and catchment percentage in other FYs at the adjusted p-value threshold of 0.002 (interaction p-value 0.032, 0.035, 0.081 in 2019, 2020, and 2021, respectively). Given potential for false-positives from multiple testing and for model comparability, no interactions were therefore included in the core model for any of the FYs.

**Figure 5.33: Predicted incidence rate for the interaction between age and catchment percentage.**




While the core estimates varied somewhat across the FYs, interpretation of estimates remained relatively stable (**Figure 5.34**). The effect of age was strong across all years, and females had lower risk across all years, compared with males. Those of non-white ethnicities had a lower risk in 2018 and 2020 compared with those of white ethnicities, however, confidence intervals were large. There was no evidence of an effect of ethnicity in 2019 or 2021. Those living in rural villages were at a lower risk in FY2021 compared with those in urban cities/towns in 2021, but there was no other evidence of an effect of rural/urban classification in other FYs. The risk of *E. coli* BSI was higher in higher catchment percentages for all FYs but this effect attenuated in 2020, perhaps due to fewer individuals travelling for specialised care in hospitals out-of-catchment during COVID-19 lockdowns.

**Figure 5.34: Estimates from core models for models run on data from FY2018, FY2019, FY2020, and FY2021.**




*Note: The "inpatient-only" cohort was used for all estimates.*

Across all four FYs, 106 variables from 85 characteristics (combining categorical/continuous parameterisations) were ever selected after backwards elimination.

The majority of characteristics selected did not overlap between FYs, with 60 characteristics (70%) being found in one financial year only (**Table 5.14**). Only three characteristics (4%) were selected in all four FYs: albumin, a procedure code for chemotherapy, and urine positive for *E. coli*. An additional five characteristics were selected across three financial years. 17 (20%) characteristics were selected in two financial years. In both FYs 2020 and 2021, SARS-CoV-2 tests were selected, however, this would not have had the opportunity to be selected earlier in FYs 2018 and 2019 as these periods predated COVID-19. There were a similar number of variables selected in models from each FY individually (but not in any other FY).

**Table 5.14: Summary of the number of variables selected after backward elimination across all FYs.**

| Total number of financial years | Financial years present | n (%) (N=85) | Characteristics |
|---|---|---|---|
| Four (N=3, 4%) | 2018, 2019, 2020, 2021 | 3 (4) | Albumin, procedure code for chemotherapy, urine positive for *E. coli* |
| Three (N=5, 6%) | 2018, 2019, 2020 | 1 (1) | Procedure from Chapter L (Arteries and Veins) |
| | 2019, 2020, 2021 | 3 (4) | Alkaline phosphatase, blood culture taken, procedure for extraction of bone marrow |

| Total number of financial years | Financial years present | n (%) (N=85) | Characteristics |
|---|---|---|---|
| | 2018, 2020, 2021 | 1 (1) | Procedure from Chapter J (Other Abdominal Organs) |
| | 2018, 2019, 2021 | 0 (0) | |
| Two (N=17, 20%) | 2018, 2019 | 2 (2) | Diagnosis from Chapter 12 (Skin and subcutaneous tissue), urine catheter |
| | 2019, 2020 | 1 (1) | Diagnosis code for diabetes |
| | 2020, 2021 | 3 (4) | Diagnosis code for palliative care, Diagnosis from Chapter 14 (Genitourinary system), SARS-CoV-2 test taken |
| | 2018, 2020 | 5 (6) | Diagnosis code for urinary disease, Inpatient admission, Inpatient admission under general surgery, Lymphocytes, white cells |
| | 2019, 2021 | 3 (4) | Diagnosis code for paraplegia, frailty score, platelets |
| | 2018, 2021 | 3 (4) | CMV/EBV screen, procedure code for scan of abdomen, procedure code for scan of pelvis |
| One (N=60, 71%) | 2018 | 15 (18) | Presented in **Figure 5.39** and **Figure 5.40** |
| | 2019 | 15 (18) | Presented in **Figure 5.14** and **Figure 5.15** |
| | 2020 | 13 (15) | Presented in **Figure 5.41** and **Figure 5.42** |
| | 2021 | 17 (20) | Presented in **Figure 5.43** and **Figure 5.44** |

There was evidence of collinearity in all FYs with similar patterns of collinearity evident across all four FYs (**Table 5.15**). As with the FY2019 dataset, there was evidence of health-seeking behaviour potentially leading to swapping of signs between univariate and multivariate estimates; for example, the presence of any procedure in 2018 and the number of urine cultures taken in 2021. For the number of urine cultures taken, the risk attenuated towards zero after removing urine culture positive for *E. coli* from the model, suggesting that after adjustment for higher risk culture results, having more urine cultures done did not increase risk (**Figure 5.35**). There were also examples of potential competing risks, such as diagnosis codes for palliative care in FY2020, whereby having a palliative care code ≤365d ago was univariately associated with higher risk, but after adjustment for other factors associated with risk, its independent effect was an association with decreased risk. A palliative care code within ≤365d could demonstrate competing risk as those with a code are at a higher risk of death and hence unable to have an *E. coli* BSI.

Some of the collinearity present appeared to be due to high correlation between selected variables which were not removed before backwards elimination using the (arbitrary) 0.95 correlation threshold. This was evident in the swapping of signs for the effect of neutrophils in FY2018, going from an increased risk to a decreased risk for higher neutrophils in univariate versus multivariate models, respectively, following inclusion of white cell count in the multivariate model (**Table 5.15**). Blood test results for white cells and lymphocytes were also included in the 2018 final model and, when removing these variables one at a time, the effect of neutrophils moved back towards the univariate direction (**Figure 5.36**). The pairwise correlation between neutrophils and white cells was 0.915 suggesting that the 0.95 correlation threshold may have been too low. A similar pattern was also observed in FYs 2018 and 2021 where the inclusion of scans of the abdomen and pelvis together likely caused the signs to swap in the multivariable model (**Figure 5.37** for 2018; 2021 not presented but the same pattern was present). Again, the correlation between the scan of the abdomen and pelvis was high, being 0.920 in 2018. A similar pattern was observed in FY2021 for the effect of a requested transferrin test and the removal of a test for serum B12 from the multivariable model; however, the removal of this variable had a smaller effect than the other examples above (**Figure 5.38**). Overall this suggests that, with this number of cases and controls, 0.9 may be a better correlation threshold to consider removing variables to avoid collinearity.

**Table 5.15: Variables with evidence of collinearity selected after backwards elimination and using p<0.25 as the entry threshold for FYs 2018, 2019, 2020, and 2021.**

| Financial year | Characteristic | Level | Multivariate IRR (95% CI) [p-value] | Multivariate global p-value | Univariate IRR (95% CI) [p-value] | Univariate global p-value |
|---|---|---|---|---|---|---|
| **2018** | Diagnosis from Chapter 12 (Skin and subcutaneous tissue) | Per 3 months closer | **0.82 (0.72, 0.94) [0.004]** | **0.004** | **1.18 (1.07, 1.31) [0.001]** | **0.001** |
| | Neutrophils (x$10^9$/L) | Per quartile higher | **0.74 (0.63, 0.87) [<0.001]** | **<0.001** | **1.14 (1.06, 1.23) [0.001]** | **0.001** |
| | Any procedure (categorical) | ≤365d ago | 1 | <0.001 | 1 | <0.001 |
| | | **>365d ago** | **1.65 (1.02, 2.67) [0.042]** | | **0.41 (0.31, 0.55) [<0.001]** | |
| | | Never | 4.45 (2.36, 8.40) [<0.001] | | 1.39 (0.88, 2.19) [0.161] | |
| | Procedure code for scan of abdomen (categorical) | ≤365d ago | 1 | **0.003** | 1 | <0.001 |
| | | **>365d ago** | **5.21 (1.92, 14.12) [0.001]** | | **0.52 (0.34, 0.80) [0.003]** | |
| | | Never | 1.99 (0.80, 4.96) [0.141] | | 0.18 (0.13, 0.25) [<0.001] | |
| **2019** | Number of outpatient attendances | Per 2 units higher | **0.93 (0.88, 1.00) [0.038]** | **0.038** | **1.22 (1.16, 1.27) [<0.001]** | **<0.001** |
| | Diagnosis from Chapter 18 (Symptoms/signs and abnormal findings) | Per 3 months closer | **0.89 (0.81, 0.98)** | **0.013** | **1.37 (1.29, 1.46) [<0.001]** | **<0.001** |
| | Diagnosis from Chapter 21 (Factors influencing health status) | Per 3 months closer | **0.90 (0.83, 0.99)** | **0.023** | **1.29 (1.22, 1.37) [<0.001]** | **<0.001** |
| | Diagnosis from Chapter 5 (Mental and behavioural disorders) (categorical) | ≤365d ago | 1 | 0.001 | 1 | <0.001 |
| | | >365d ago | 1.86 (1.33, 2.61) [<0.001] | | 0.89 (0.64, 1.24) [0.498] | |
| | | **Never** | **1.50 (1.12, 2.01) [0.007]** | | **0.52 (0.40, 0.67) [<0.001]** | |
| | Diagnosis from Chapter 12 (Skin and subcutaneous tissue) (categorical) | ≤365d ago | 1 | 0.008 | 1 | <0.001 |
| | | >365d ago | 1.80 (1.20, 2.69) [0.004] | | 1.02 (0.69, 1.50) [0.923] | |
| | | **Never** | **1.64 (1.17, 2.31) [0.004]** | | **0.64 (0.46, 0.88) [0.005]** | |
| | Diagnosis code for pneumonia | Per 3 months closer | **0.84 (0.74, 0.96) [0.008]** | **0.008** | **1.15 (1.02, 1.29) [0.023]** | **0.023** |

| Financial year | Characteristic | Level | Multivariate IRR (95% CI) [p-value] | Multivariate global p-value | Univariate IRR (95% CI) [p-value] | Univariate global p-value |
|---|---|---|---|---|---|---|
| **2020** | Diagnosis from Chapter 22 (Codes for special purposes) (categorical) | ≤365d ago | 1 | <0.001 | 1 | <0.001 |
| | | >365d ago | 3.97 (1.98, 7.98) [<0.001] | | 3.40 (1.74, 6.65) [<0.001] | |
| | | **Never** | **1.77 (1.05, 2.97) [0.031]** | | **0.47 (0.29, 0.77) [0.003]** | |
| | Diagnosis code for palliative care (categorical) | ≤365d ago | 1 | <0.001 | 1 | 0.011 |
| | | **>365d ago or never** | **2.98 (1.76, 5.02) [<0.001]** | | **0.48 (0.28, 0.85) [0.011]** | |
| **2021** | Diagnosis from Chapter 10 (Respiratory system) | **Per 3 months closer** | **0.88 (0.81, 0.95) [0.002]** | **0.002** | **1.20 (1.12, 1.29) [<0.001]** | **<0.001** |
| | Diagnosis from Chapter 9 (Circulatory system) (categorical) | ≤365d ago | 1 | 0.003 | 1 | <0.001 |
| | | **>365d ago** | **1.81 (1.21, 2.72) [0.004]** | | **0.51 (0.39, 0.66)** [<0.001] | |
| | | Never | 0.90 (0.62, 1.31) [0.585] | | 0.24 (0.17, 0.34) [<0.001] | |
| | Transferrin test requested | No test | 1 | 0.015 | 1 | <0.001 |
| | | **Test** | **0.71 (0.54, 0.94) [0.015]** | **0.015** | **1.86 (1.45, 2.38)** [<0.001] | **<0.001** |
| | Number of urine cultures | **Per 2 units higher** | **0.74 (0.59, 0.93) [0.009]** | **0.009** | **2.15 (1.84, 2.53)** [<0.001] | **<0.001** |
| | Procedure code for scan of pelvis (categorical) | ≤365d ago | 1 | 0.012 | 1 | <0.001 |
| | | >365d ago | 2.72 (0.90, 8.21) [0.075] | | 0.51 (0.34, 0.78) [0.001] | |
| | | **Never** | **2.59 (1.38, 4.87) [0.003]** | | **0.21 (0.16, 0.28)** [<0.001] | |

**Figure 5.35: IRR (95%) CI per 2 unit higher number of urine cultures taken in models removing each variable selected after backwards elimination from one at a time for FY2021.**

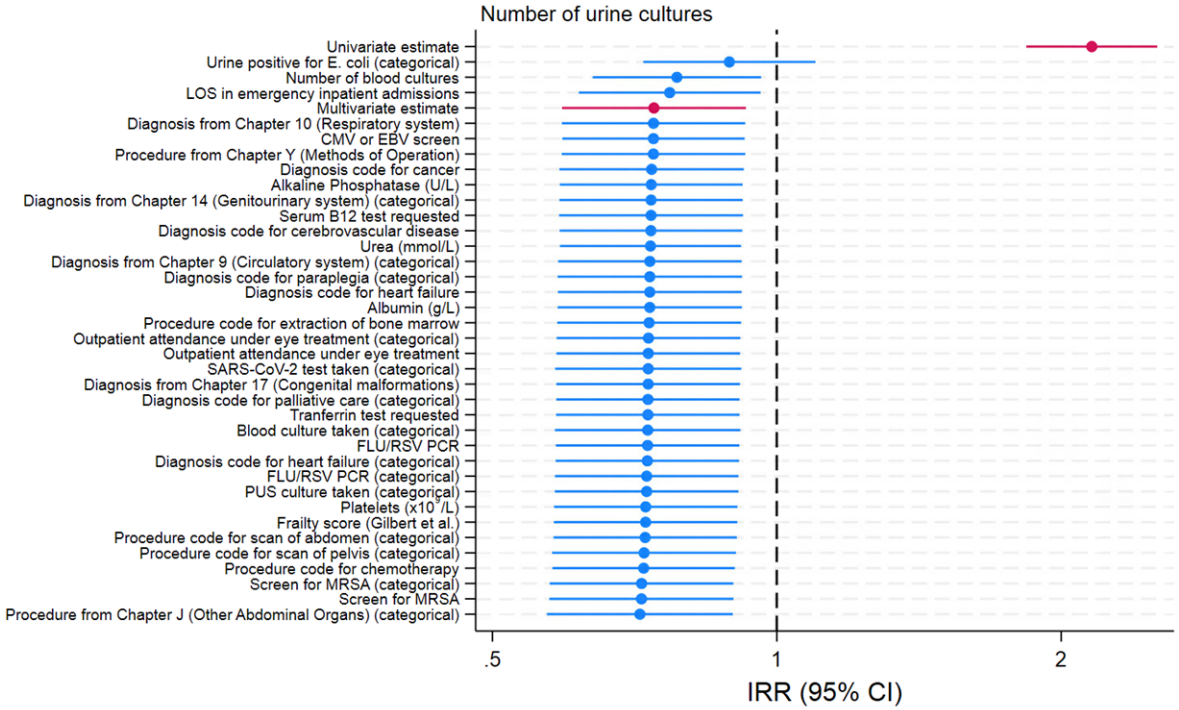
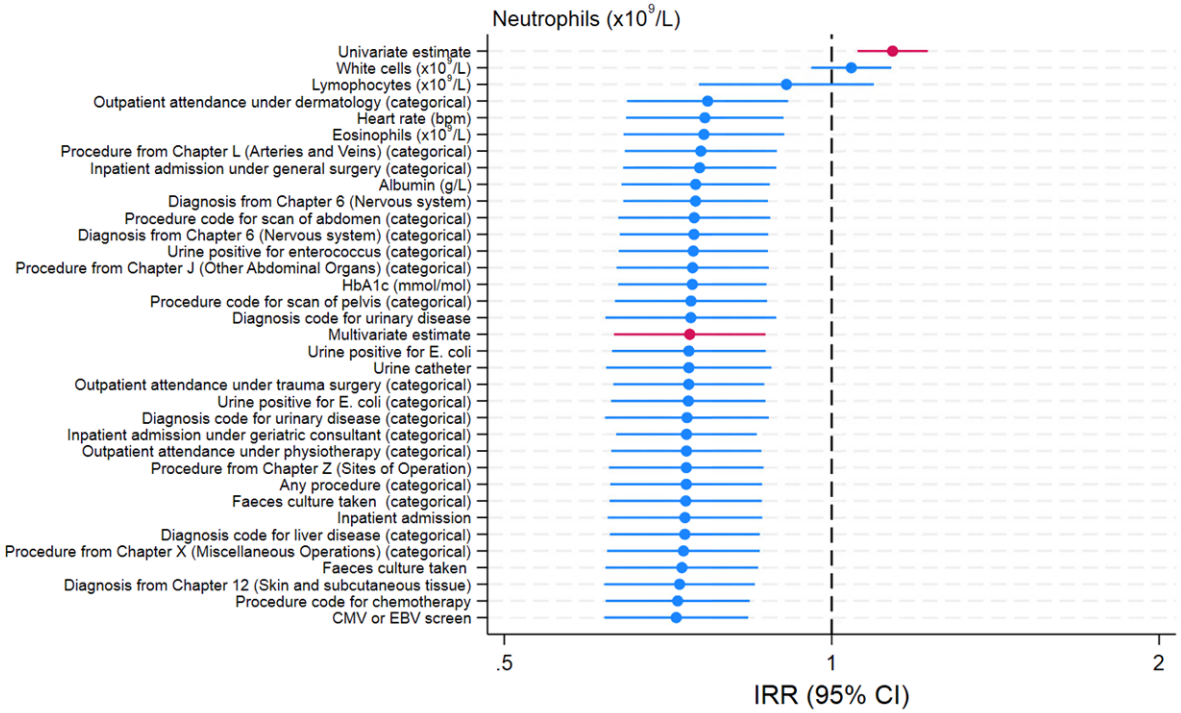
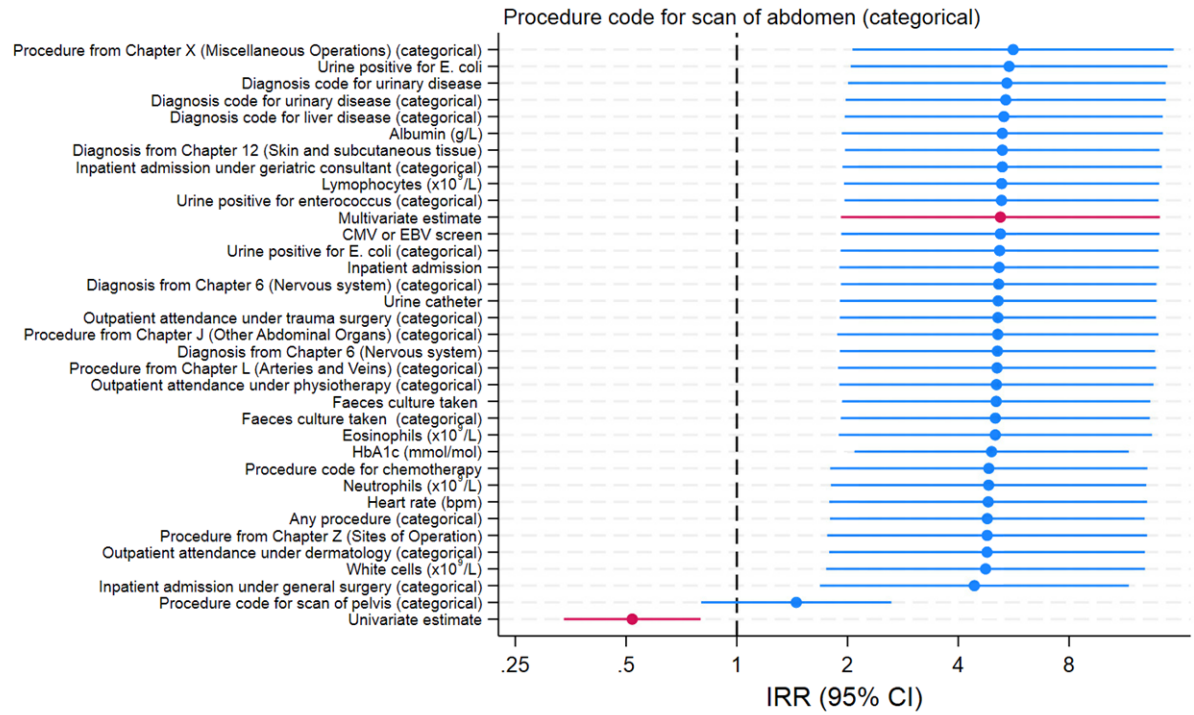



**Figure 5.36: IRR (95%) CI per quartile higher neutrophil levels taken in models removing each variable selected after backwards elimination from one at a time for FY2018.**



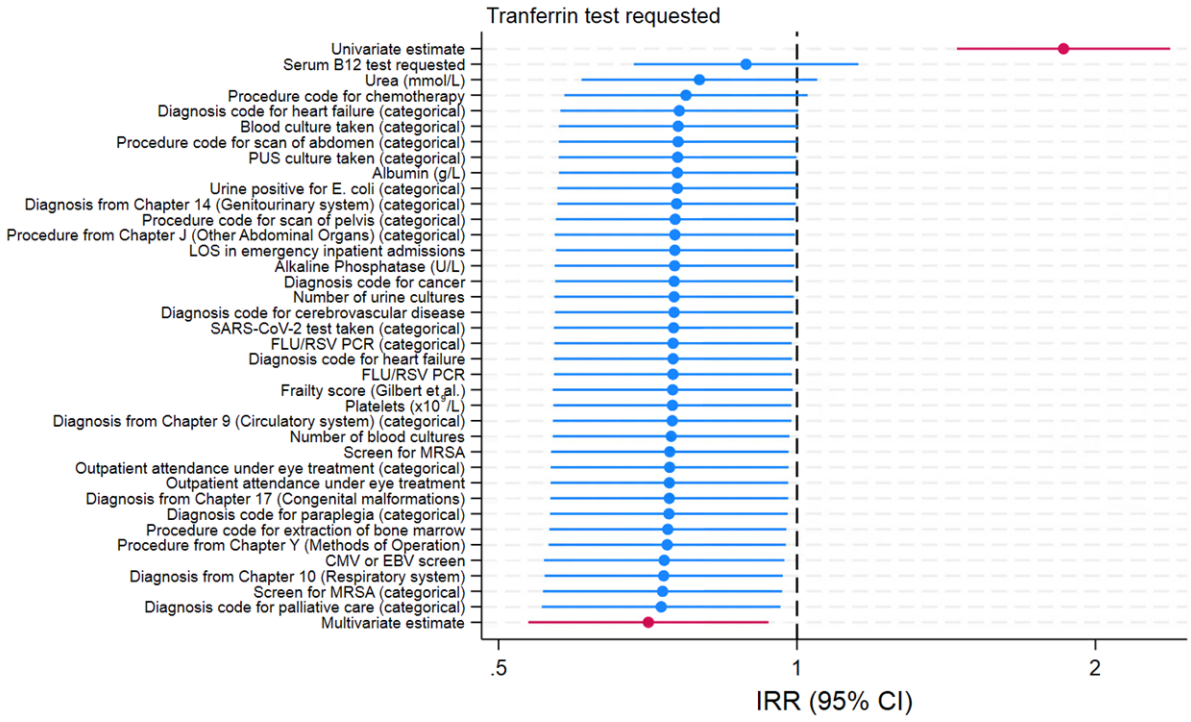

Note: Pairwise correlation for neutrophils with white cells is 0.915. Pairwise correlation for lymphocytes and white cells is 0.403.

**Figure 5.37: IRR (95%) CI for scan of abdomen >365d ago versus ≤365d ago from models removing each variable selected after backwards elimination from one at a time for FY2018.**




Note: Pairwise correlation between procedure code for scan of pelvis and scan of abdomen is 0.920.

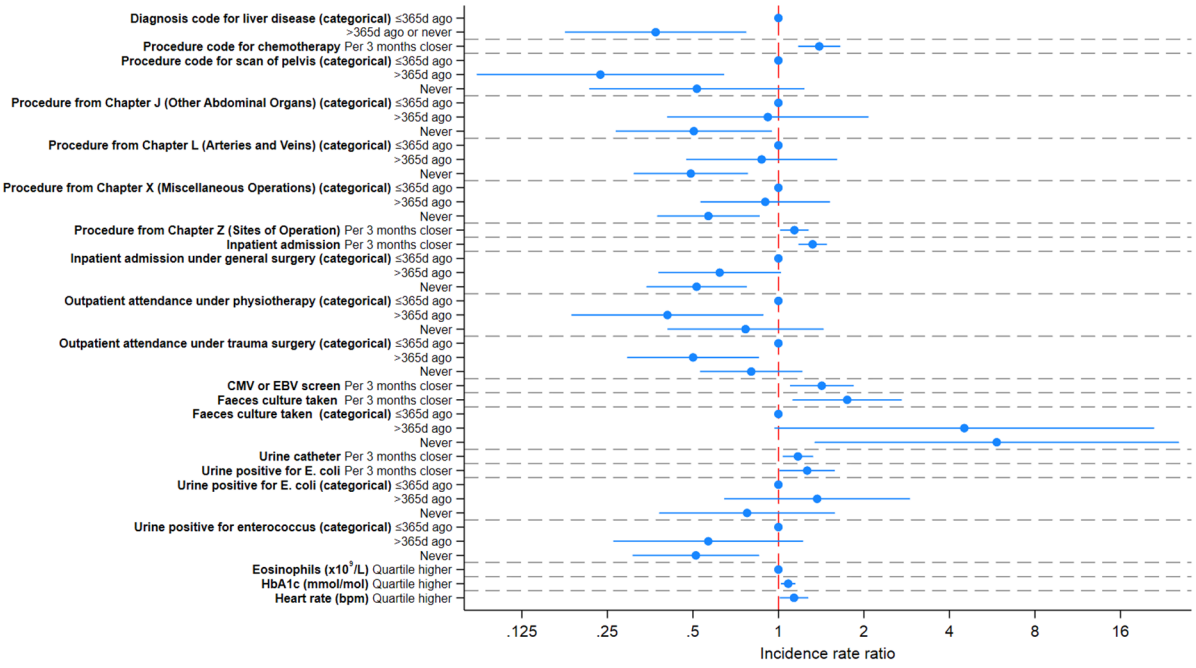
**Figure 5.38: IRR (95%) CI for having a transferrin test requested versus no transferrin test requested from models removing each variable selected after backwards elimination from one at a time for FY2021.**



The full model estimates are presented in **Figure 5.39** to **Figure 5.44** below for FY2018, FY2020, and FY2021. Each FY has been plotted separately as there is very little overlap in specific factors included between each FY and therefore model adjustment was very different so comparing the effect of the same variables across different FYs is not informative.

In FY2018, a combination of specific procedures and microbiology tests alongside variables denoting high hospital exposure were associated with a higher risk of *E. coli* BSI compared with the control group (**Figure 5.39**). Specific procedures/diagnoses which were associated with increased *E. coli* BSI risk included diagnosis codes for liver disease, procedures from Chapter J (abdominal organs), and chemotherapy. Further, outpatient attendances under physiotherapy treatment or trauma surgery treatment ≤365d ago were associated with a higher risk of *E. coli* BSI, compared with those with these attendances >365d ago. Specific microbiology results which were associated with increased risk the closer the test was taken in the previous 365d included CMV/EBV screens, evidence of a urinary catheter (from a microbiological specimen), and urine culture positive for *E. coli*. Higher HbA1c was associated with increased risk of *E. coli* BSI, alongside higher heart rate. More generally, closer inpatient admissions in the previous 365d were associated with higher risk.

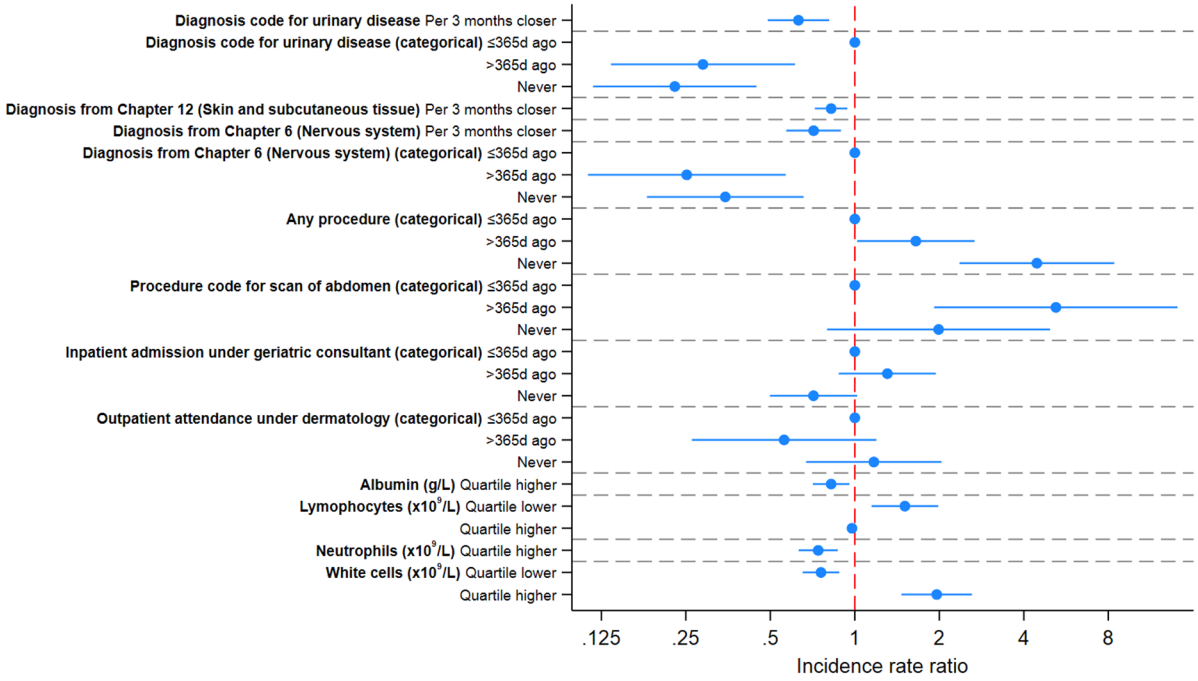
**Figure 5.39: Final variables selected in FY2018 using the "inpatient only" cohort where the presence of the characteristic was associated with increased risk of having an E. coli BSI.**




Some of the variables which were associated with reduced risk of *E. coli* BSI may be due to competing risks as observed for the FY2019 data (**Figure 5.40**). For example, the presence of more

recent urinary disease codes or codes from Chapter 6 (nervous system) ≤365d ago was associated with reduced risk of *E. coli* BSI, however risk was still higher 365d ago versus those who had a code either >365d ago or never ("N"-shaped risk over the full 5 FYs). Never having had a procedure or whose closest previous procedure was >365d ago was associated with a higher risk compared with those having had a procedure ≤365d ago, perhaps indicative of health-seeking behaviour being protective. Some of the variables which were associated with reduced risk were likely due to high correlation with other variables (e.g. scan of abdomen with scan of pelvis, and neutrophils with white cells) as discussed above.

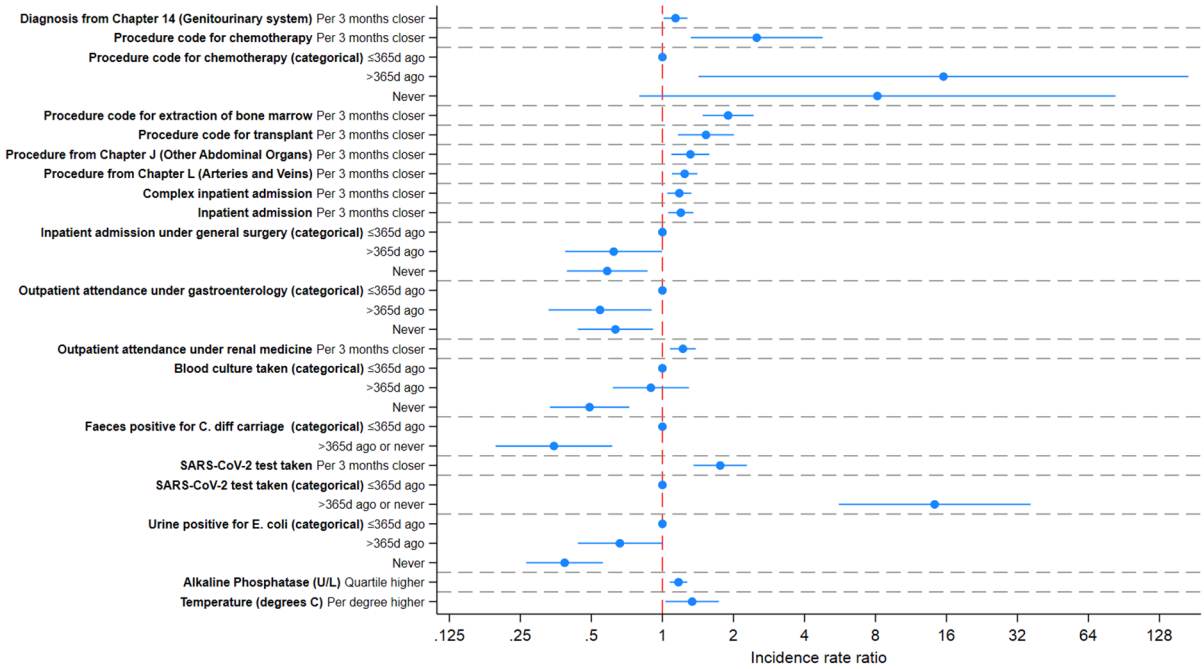
**Figure 5.40: Final variables selected in FY2018 using the "inpatient only" cohort where the presence of the characteristic was associated with decreased risk of having an E. coli BSI.**




In FY2020, the presence of SARS-CoV-2 may have influenced the risk factors found while some similar characteristics to FY2018 and FY2019 were also identified (**Figure 5.41**). A recent SARS-CoV-2 test ≤365d ago was associated with a higher risk of an *E. coli* BSI. Compared with those having a SARS-CoV-2 test 365d ago, those having a SARS-CoV-2 test >365d ago or never were at a higher risk of *E. coli* BSI ("U"-shaped risk over the last 5FY), perhaps reflecting previous hospital exposure before the onset of the COVID-19 pandemic in the cases, and more people coming to hospital in the last year for the first time with COVID-19 in the control group. Specific procedures which were associated with increased risk of *E. coli* BSI in FY2020 included diagnostic extraction of bone marrow, transplant, and outpatient attendances under renal medicine. Those with an inpatient admission for general surgery in the previous year were at a higher risk compared with those having general

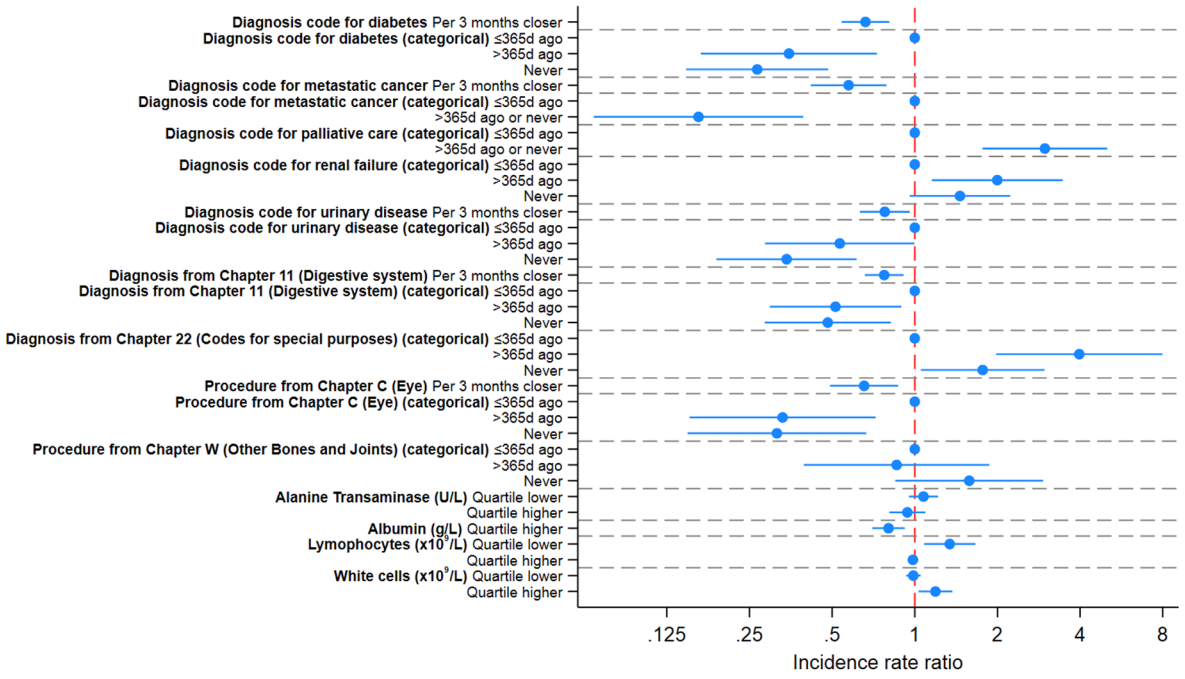
surgery >365d ago or never. Higher alkaline phosphatase and higher (background) temperature were both associated with increased risk of *E. coli* BSI.

**Figure 5.41: Final variables selected in FY2020 using the "inpatient only" cohort where the presence of the characteristic was associated with increased risk of having an *E. coli* BSI.**



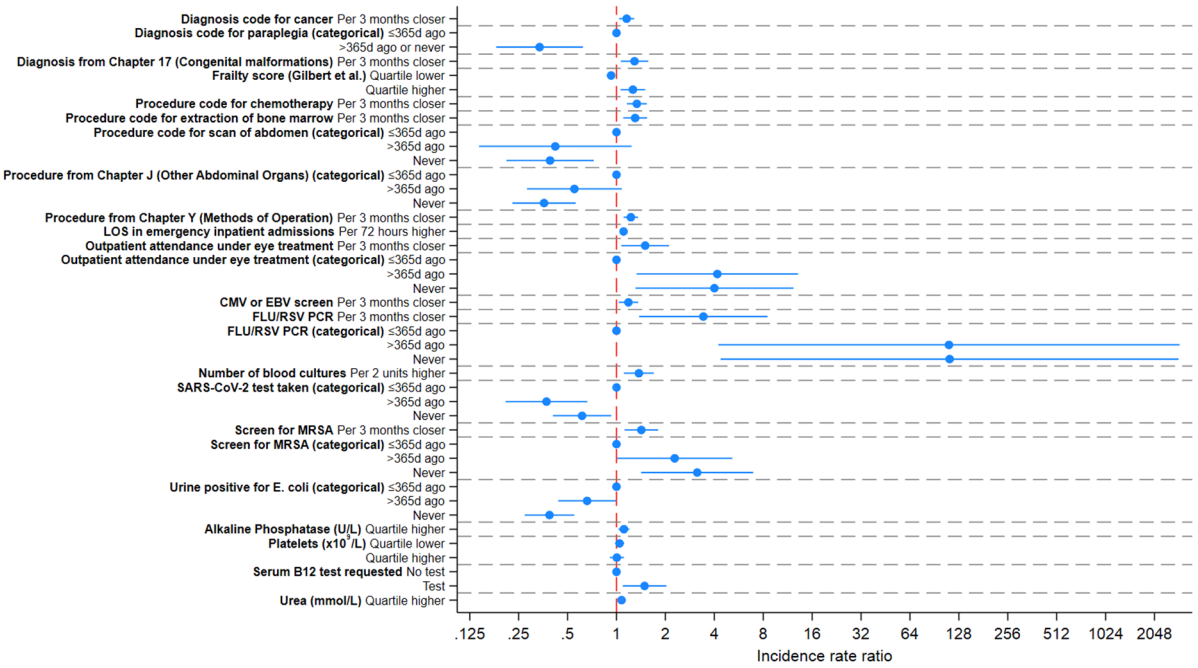

In 2020, the risk of *E. coli* BSI was lower the closer some characteristics were to the current contact but the risk of *E. coli* BSI was still higher 365d ago versus having the characteristic >365d or never ("N"-shaped risk over the previous 5FY) (**Figure 5.42**). This was evident for diagnosis codes for diabetes, metastatic cancer, digestive system diagnoses, urinary diseases, and eye procedures. Those with a palliative care code ≤365d ago were at a lower risk of *E. coli* BSI than those with a palliative care code >365d or never, likely to reflect the competing risk of death versus getting an *E. coli* BSI. As seen in all other FYs, higher albumin levels were associated with reduced risk of *E. coli* BSIs compared with controls.

**Figure 5.42: Final variables selected in FY2020 using the "inpatient only" cohort where the presence of the characteristic was associated with decreased risk of having an *E. coli* BSI.**



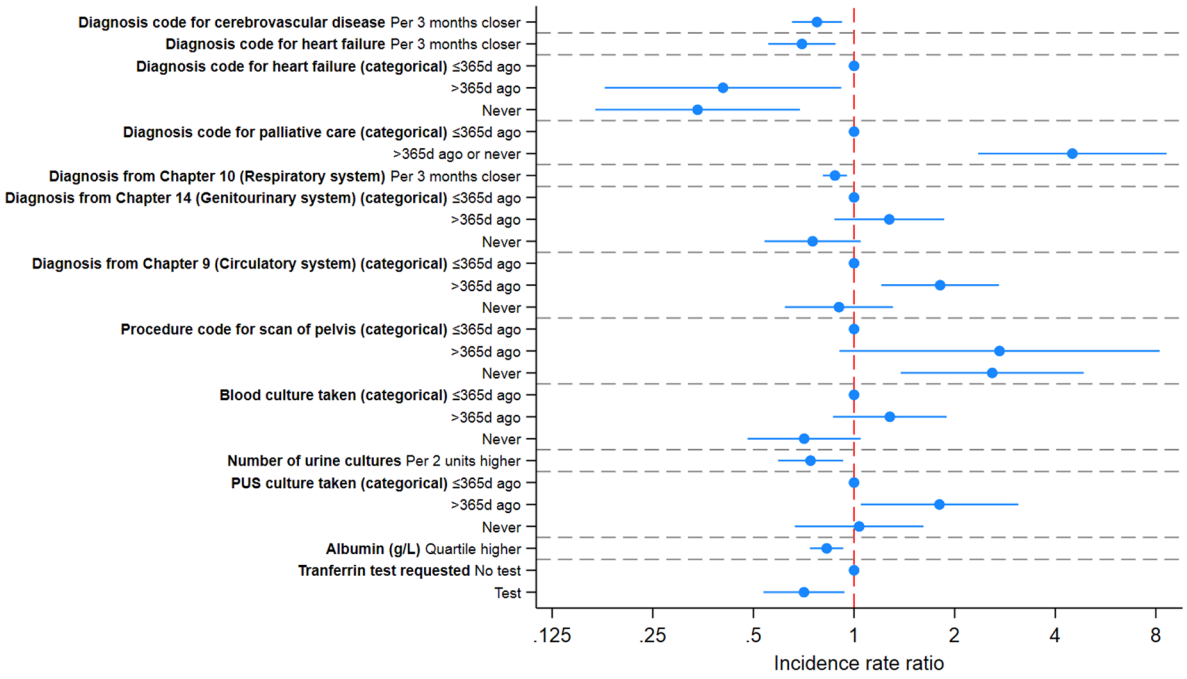

As with FY2020, in FY2021, the impact of COVID-19 on the risk factors was still evident (**Figure 5.43**). Having recent influenza/RSV PCRs or SARS-CoV-2 tests were both associated with a higher risk of *E. coli* BSI while having an influenza/RSV PCR >365d ago or never was also associated with a higher risk of *E. coli* BSI versus having an influenza/RSV PCR 365d ago ("U"-shaped risk over the previous 5FYs). Similarly to other FYs, having chemotherapy, extraction of bone marrow, urine culture positive for *E. coli*, more previous blood cultures taken, or a higher frailty score was associated with increased risk of *E. coli* BSI. More time in hospital in the previous year was also associated with higher risk, as demonstrated through increases in length of stay in emergency inpatient admissions being associated with a higher risk of *E. coli* BSI. For the first time across all the FYs, higher levels of urea were associated with a higher risk of *E. coli* BSI, and also more recent MRSA screening tests were associated with higher risk.

**Figure 5.43: Final variables selected in FY2021 using the "inpatient only" cohort where the presence of the characteristic was associated with increased risk of having an *E. coli* BSI.**




Similar to other FYs, in FY2021 there was evidence of competing risks in variables whose recent presence was associated with reduced risk of *E. coli* BSIs (**Figure 5.44**). While this pattern was consistent across all the FYs, the variables which may have been indicating competing risks changed each year. In FY2021, the closer presence in the 365d before contact of a diagnosis code for heart failure or cerebrovascular disease was associated with decreased risk of *E. coli* BSI. The presence of other variables, such as diagnosis codes from the genitourinary system or circulatory system, or having a previous blood culture taken, were associated with a higher risk if the code was present >365d ago, but a slightly lower risk of *E. coli* BSI if the code was never present, compared with having a code in the last year.

**Figure 5.44: Final variables selected in FY2021 using the "inpatient only" cohort where the presence of the characteristic was associated with a decreased risk of having an *E. coli* BSI.**



## 5.4 Discussion

Using one year of data from FY2019, I updated the screening process from Chapter 2 to find risk factors for *E. coli* BSIs using electronic health records. In FY2019, I found that the presence of chemotherapy, diabetes, and characteristics associated with urinary disease and frailty were associated with increased risk of *E. coli* BSI. The risk of *E. coli* BSI was associated with a variety of blood test results, such as higher risk for individuals with lower albumin levels and higher alkaline phosphatase levels. I found characteristics associated with good health-seeking behaviour to be associated with reduced risk of *E. coli* BSI after adjusting for riskier diagnoses/procedures, such as having more outpatient appointments. Using different control groups based on all healthcare contact, I found similar risk factors compared with restricting the cohort to those with current or previous inpatient contact only. Some differences in risk factors found were due to similar variables explaining the same variation e.g. diabetes as a risk factor being selected as either a diagnosis code for diabetes or elevated levels of HbA1c. Risk factors across nosocomial, quasi-nosocomial, and quasi-community *E. coli* BSIs were generally quite similar with a few different risk factors, for example, a high risk of quasi-nosocomial infections in those with a recent procedure code for prosthesis. Risk factors found for *E. coli* BSIs with resistance to third-generation cephalosporins were also broadly similar but with a few risk factors perhaps indicating more background comorbidity (potentially leading to more antibiotic use in the community) as higher temperature and recent comorbidity scores were found as risk factors. Risk factors were much more substantially different when running the analysis over FYs 2018, 2020, and 2021, possibly due to the influence of COVID-19 on hospitalisation in FYs 2020 and 2021. Issues due to collinearity persisted throughout all the models and future work should find better ways to deal with this.

### 5.4.1 Comparison of risk factors to other studies

Some of the risk factors I found in this study have previously been found in different populations and settings. I found urinary catheter use (as identified from microbiological specimens) to be associated with a higher risk of *E. coli* BSI; a finding previously documented elsewhere for *E. coli* BSIs,[218] ESBL-producing *E. coli* BSIs,[225] and other bloodstream infections.[226] I also consistently found chemotherapy to be a risk factor for *E. coli* BSI, accounting for 3% of the total 21% of variance explained in the "inpatient only" model in FY2019. It is well known that individuals receiving chemotherapy are at a higher risk of infection for multiple reasons, from having a low white cell count (neutropenia) contributing to a weakened immune system, to a compromised host response due to the underlying cancer, and increased intensity of hospital exposure increasing risk of hospital-acquired infections.[227] One study showed chemotherapy to be a direct risk factor for *E. coli* BSI in pancreatic cancer patients,[228] while another showed chemotherapy to be a risk factor for ESBL-

producing *E. coli* BSIs compared with third-generation cephalosporin-susceptible strains.[229] I also found diabetes to be a risk factor, through both diagnosis codes for diabetes from inpatient admissions and through elevated HbA1c levels. One study showed the incidence of any BSI to be 4.4 times higher in diabetic versus non-diabetic patients,[230] and another hospital-based study showed diabetes to be a direct risk factor for *E. coli* BSIs.[217] Finally, I also found characteristics related to urinary disease to be associated with increased risk of *E. coli* BSIs, for example; outpatient attendances under urology were associated with higher risk in FY2019 in nosocomial BSIs and those in the "any healthcare" and "age>65y" cohort while having any urine culture taken were associated with higher risk in the "inpatient only" cohort in FY2019. It is estimated that >50% of *E. coli* BSIs are caused by urinary tract infections[231] and hence urinary disease related risk factors would be expected.

Being able to consider results from a variety of blood tests as risk factors yielded some interesting findings. Higher albumin levels were consistently associated with a lower risk of *E. coli* BSI across many of the different subgroups and all FYs. Low albumin levels could be associated with a higher risk of *E. coli* BSIs in several ways. First, low albumin levels are part of the pathophysiological process of certain conditions, for example, decompensated liver cirrhosis,[232] and cirrhosis has been shown to increase the risk of bacterial infections.[233] Low albumin levels may therefore be a proxy for other diseases in which in risk of BSIs is higher. Second, reduction in albumin levels is directly related to the body's response to bacteraemia e.g. through leakage of albumin from the plasma to the interstitial space.[234] While I did exclude albumin measurements in the 72 hours directly before an *E. coli* BSI, there could potentially be some reverse causality in this association dependent on when individuals had a blood test for albumin and a blood sample taken for culture. However, I did prioritise albumin results collected outside of inpatient admissions and A&E. Contrary to the lower risk with higher albumin in this study, high albumin can be a marker for dehydration; a potential risk factor for BSIs;[235] however, increased urea may likely be a better marker of dehydration.

Some risk factors found in this study after adjusting for these characteristics associated with high risk may have been demonstrating a lower risk of *E. coli* BSI for those with good health-seeking behaviour. For example, some risk factors, including the number of previous urine cultures and outpatient attendances or having any prior procedure, were associated with a higher risk of *E. coli* BSIs in univariate analysis but after adjustment for many other risk factors became associated with a lower risk of *E. coli* BSIs; that is, like for like, once high-risk factors have been accounted for, those under closer clinical care or seeking more regular care may have a lower risk by virtue of their behaviour, rather than anything to do with the factor itself. Adjusting for low-value procedures as an indicator for good health-seeking behaviour has been suggested elsewhere;[236] therefore after

adjustment for riskier procedures in the multivariate models presented here, any prior procedure may be closer to low-value, lower-risk procedures. Further, a higher number of outpatient attendances may also be associated with a lower risk of infection as more regular healthcare contact may mean diseases/sub-clinical infections are caught earlier. For example, one study that focused on patients with heart failure and chronic obstructive pulmonary disease found that missed outpatient appointments were associated with worse outcomes, such as a higher risk of both mortality and readmission.[237]

There was also evidence of competing risks in some risk factors found by the screening process. A competing risk can be defined as an event which likely precludes the event of interest being observed e.g. death or another major health event occurring before an *E. coli* BSI.[238] For example, a recent diagnosis code for pneumonia was associated with a lower risk of *E. coli* BSI in the "inpatient only" cohort in FY2019, suggesting that these individuals may be less vulnerable to *E. coli* BSIs soon after, or during, pneumonia treatment. Further, as this control group constituted individuals who had current or previous inpatient contact, risk factors associated negatively with the risk of *E. coli* BSIs may reflect other reasons individuals frequently attend hospital. Never having had a diagnosis under the ICD-10 code Chapter for skin and subcutaneous tissue disease was associated with a higher risk of *E. coli* BSIs, compared with those with a diagnosis code from this Chapter ≤365d ago. Those with an ICD-10 code for skin and subcutaneous tissue disease in the prior year may have instead gone on to have, for example, *Staphylococcus aureus* bacteraemia, with skin and soft tissue infections being the most commonly reported source of MRSA bacteraemia.[239]

I found similar risk factors for *E. coli* BSIs with resistance to third-generation cephalosporins as I did for all *E. coli* BSIs combined, with two notable exceptions. Specifically, I found a higher Charlson score was found to be associated with third-generation cephalosporin-resistant *E. coli* BSIs; this has also been reported previously when compared with a control group of susceptible *E. coli* BSIs.[240] A higher number of previous antibiotic prescriptions in the previous year was also found to be a risk factor for *E. coli* BSIs resistant to third-generation cephalosporins compared with susceptible controls in a variety of different populations;[240,241,242] a risk factor which I did not use in this study as I did not have access to community prescribing (only hospital prescribing) but could be mediated through higher Charlson comorbidity score.

### 5.4.2 Defining variables

There were many different ways I could have defined the characteristics and variables included in this analysis, and additionally, other variables which could have been derived from the database.

This section of the Discussion will discuss the limitations of the variables used in this study and how they were defined, as well as variables which could have additionally been included.

The variables used in this study were a balance of individual risk factors selected through clinical judgement and higher-level categorisations. Higher-level categorisations of variables often yielded risk factors which were harder to interpret. For example, I split procedure codes and diagnosis codes at the Chapter level, creating 24 variables and 22 variables, respectively. Procedure codes from Chapter O (overflow codes) and Chapter X (miscellaneous operations) were often selected after backwards elimination. Without further investigation, the Chapter names themselves are not informative. Upon further investigation, Chapter X contains mainly procedure codes for chemotherapy – a risk factor found across multiple populations and years during this analysis as well as in other studies as highlighted above. Chapter X therefore appeared to be a proxy for chemotherapy and so it is debatable whether including Chapter X is useful at all if chemotherapy would be able to explain the increase in risk more informatively. However, creating variables using high-level categorisations can save time, being a useful broad approach and can avoid missing potentially important variables not thought of before analysis. Overflow codes belong to a variety of different Chapters and are therefore not informative as a group; were I starting this analysis again I would not use them.

There were many different ways to define the risk factors used in this study with it being unclear which way was best. In my study, frailty was defined using one combined validated frailty score,[180] however there are numerous other frailty scores available.[243,244] For example, Luo et al.[245] used electronic health records from primary care to identify many possible indicators for frailty, some of which could also be calculated from the secondary care data available to me, such as weight loss. These could additionally be used as individual markers for frailty in the screening model, alongside a combined frailty score. Furthermore, I defined renal dialysis using procedure codes; however, this is also coded in outpatient specialty codes which I did not use in this study but may be useful to consider in future work. Some variables required more nuanced decisions in their definitions, such as whether to include the current time in an inpatient admission to calculate inpatient length of stay. In this study, only strictly previous inpatient admissions were used for this calculation (i.e. discharged before the current contact) to reduce the risk of reverse causality by not including time in hospital before an *E. coli* BSI in the calculation. However, one could additionally include all the time in hospital to up the current contact in the length of stay as this time may vary and influence risk, particularly of nosocomial *E. coli* BSIs. Both options are justifiable, and neither are likely to be uniformly correct or even preferable. Similarly, I defined time since the most recent diagnoses using the episode start date in which diagnoses were recorded, however, episode end dates may have

been more suitable and more consistent with the definitions of prior contact used to define the "inpatient only" cohort. However decisions on how to define risk factors had to be made and sensitivity analyses had to be selected to show the robustness of the most important parts of the screening process.

All risk factors used in this analysis were calculated from electronic health records, however, environmental factors could also have been added to the dataset from elsewhere. Some studies have shown an effect of outdoor temperature on the incidence of *E. coli* BSIs[246] with one study showing incidence of community-onset *E. coli* BSIs increasing by 6.2% for each 5.5C increase in outside temperature.[247] In Oxford, the mean monthly temperature can be downloaded and this could have been added to the dataset as a screening variable.[248] The month of year could also have been added in to capture some of the seasonal variation which has previously been shown to have an impact on community-onset *E. coli* BSIs but not hospital-onset infections.[249] Future work could include these risk factors in the screening process. However, for them to affect estimates of risk associated with other factors, they would need to be confounders and it is not clear that this is the case for the risk factors I considered which should be relatively seasonally invariant.

One important consideration in defining variables in risk-based analysis using EHRs is the availability of data. In this analysis using IORD, I had access to diagnosis and procedure codes as are available through Hospital Episode Statistics (HES),[45] but I additionally had access to microbiology and blood test results from samples taken both in hospital and at the GP and tested in the hospital pathology laboratories. This allowed me to assess a wider variety of risk factors than would be available in HES alone, finding interesting associations such as a higher risk in individuals with a urinary culture requested irrespective of the results, and blood test results such as higher HbA1c being associated with a higher risk of *E. coli* BSI. If these additional fields were not available, more time may have to be spent defining proxies for these variables using diagnosis and procedure codes. However, as observed in this study, when swapping out urinary catheter use as defined in microbiology data for the equivalent defined by procedure codes no effect of catheterisation was found, hence risking missing important associations. Procedure code recordings of urinary catheterisation may be more incomplete as smaller, more routine procedures may not be specifically recorded in outpatient attendances. In the NHS, while inpatient admissions are billed via diagnosis and procedure codes, outpatient appointments are billed by clinic and hence there is no financial incentive to record small procedures as the amount the hospital is reimbursed will not change.[13] Microbiology test results in IORD should contain all samples which passed through the laboratory and therefore should be more complete, although will miss catheters from which no specimen is ever taken for culture. Studies using only data available in HES may therefore miss important risk factors. While I had access to a

large number of risk factors, I did not consider medication prescriptions as I only had access to in-hospital prescribing as there is currently no linkage between community prescribing and IORD. Looking at medications prescribed both in the community and in hospital could be very informative.

Keeping variables up-to-date and relevant will have to be considered when running the screening process in the future. For example, if running this analysis in real-time from 2018 to 2022, COVID-19 would need to have been added in 2020 as it had such a large impact on the hospitalised population and the incidence of bloodstream infections.[61,216] Another potentially important variable to consider in the future would be virtual wards, also known as hospitals at home, as they may influence the population in hospital. Virtual wards are still scaling up in England currently[250] so, while not relevant yet, may be added in the future. New variables would have to be considered for each annual iteration of the screening process to ensure the list of variables being screened is up-to-date and to add new variables into the process which may impact the risk of *E. coli* BSIs.

The time since the most recent occurrence of a characteristic in the previous 365d as a continuous effect was considered for almost all characteristics. While 365d is an arbitrary time window, it was selected as I assumed that more recent exposure may increase risk and this would attenuate as time since the most recent occurrence got further away. This 365d cut-off may not have been ideal for all variables, for example, chemotherapy where the risk of infection may remain higher long after treatment, especially in individuals who experience complications such as febrile neutropenia during chemotherapy.[251] Including a categorical variable and splitting out those who have never experienced the characteristic in the previous five financial years, and those who had experienced the characteristics >365d ago was a way to capture this variation, however inflexibly. Further, modelling any occurrence in the previous 365d was impractical for looking at community BSIs as, by definition, they had no inpatient exposure in the previous 365d. Properly investigating risk factors for community-onset *E. coli* BSIs would require primary care data to capture recent health conditions.

In the above section, I have outlined limitations to do with the way I defined variables for the screening process. One future direction this work could take is to define variables in a more automated way taking inspiration from genome-wide association studies (GWAS). Variables could be created from EHRs by aggregating the data at different levels. For example, ICD-10 codes could first be split into Chapters, e.g. codes A00-B99 are all infectious and parasitic diseases. They could then be split into categories based on the first three characters, for example, A00 denotes cholera. This code is then followed by a decimal point with up to two further subclassifications which describe the severity, location, or cause of disease.[252] Taking advantage of the structured nature of these codes

may help capture more variables in the future, and keep variables up-to-date. Alternatively, this could be done for Chapters identified through the initial screen.

### 5.4.3 Making statistical decisions

Throughout the screening process, decisions had to be made for the process to move forward in as automated fashion as possible. For example, after the initial screen, which variables to select to take forward to the backwards elimination process based on a p-value threshold was unclear. In this example, I used sensitivity analysis to test including variables below the following three thresholds separately: Bonferroni adjusted p-value, Benjamini-Hochberg adjusted p-value, and $p<0.25$. There were other decisions made where a sensitivity analysis was not run, two of which are addressed below.

Choosing a different exit p-value for backwards elimination could have altered the results in the final model but it is not clear which exit p-value is best. A p-value threshold was set at 0.05 to remove variables during backwards elimination in this study, but thresholds such as $p<0.20$ have been suggested to reduce bias in estimates.[253,254] After backwards elimination, an additional step can be taken to add all variables eliminated back into the final model one at a time, keeping them if they are significant. Royston & Sauerbrei suggest that this re-inclusion after exclusion rarely occurs and therefore by not doing this step, the risk of missing important variables in the screening process may have been low.[255] However, it is unclear whether some of the lack of stability in model estimates across financial years could have been affected by this. To further investigate this, the exit p-value for backwards elimination could be increased from 0.05 to 0.10, only interpreting risk factors if $p<0.05$, but counting as a match if the p-value was between 0.05-0.10. This would allow more flexibility in assessing the lack of stability due to small numbers in models.

A minimum number of nine occurrences in *E. coli* cases for each variable was selected to allow the inclusion of the characteristic in the screening process. This decision was made based on the size of the model estimates when including variables with eight or fewer occurrences, with model estimates stabilising from around nine onwards. Excluding these variables risked missing true signals, however these estimates would have also had large variances. A general rule of thumb of 10 events per predictor variable (EPV) is often advised based on two simulation studies,[256,257] but this is arbitrary and can be relaxed without greatly increasing bias.[258] Vittinghoff & McCulloch (2006) argue that systematically discounting variables with 5-9 EPV is not justified in their simulation study based on the confidence interval coverage and type I error rate.[258] All these simulation studies argue that results should be interpreted cautiously when the number of EPV is small. A more sophisticated approach to the problem of sparse data could be penalisation.[259] This method adds artificial data

records that penalise (or shrink) coefficient estimates in proportion to their size, hence improving the accuracy of risk prediction. This is particularly useful when sparse data can cause model coefficient estimates to move further from the null as more variables are added to the model,[259] as occurred in the screening process presented in this Chapter. The impact of penalisation on coefficient estimates would be a useful additional analysis to do in the future.

The two examples above provide insight into some of the decisions made for the analysis in this Chapter and raise the question of how to make sensible decisions in statistical analyses. Morris et al. (2019)[125] outlines how simulation studies can be used to evaluate statistical methods, for example, a comparative evaluation of two or more statistical methods. Generating data complex enough to emulative EHRs may be challenging, but might be a future direction for making these statistical decisions in the future.

### 5.4.4 Discussion of the statistical analyses

Poisson regression models were used in this Chapter, however, they can overfit, especially with large numbers of variables.[260] Regularisation techniques, such as lasso, ridge regression, and elastic net, can help reduce model complexity and overfitting. Regularisation techniques work by adding a penalty term to the loss function, sending some coefficients towards zero to reduce spurious results. Elastic net uses a combination of both R1 (lasso) and R2 (ridge) regularisation and has been used in the analysis for COVID-19 risk factors, improving the prediction of COVID-19 cases when compared to multiple regression.[261] Another study looked at using elastic net to find risk factors for blunt cerebrovascular injury in trauma patients.[262] They found that using elastic net identified risk factors missed by the alternative multiple logistic regression approach; for example, they found higher odds ratios with $p>0.05$ in the logistic regression and significant results with lower odds ratios in the elastic net reflecting shrinkage of estimates and allowing important risk factors to be identified in their small sample size. Further, these techniques can be effective when covariates are correlated,[263] a challenge encountered in this Chapter's analysis.

However, these regularisation techniques also have disadvantages compared with classical statistical approaches. Lasso penalties can delete important confounders from a regression, therefore adding a source of bias into the model.[259] Further, estimates from lasso models are biased, increasing as the amount of shrinkage increases which in turn decreases the variance. If a smaller amount of shrinkage is selected to limit the amount of bias, the number of false positive variables in the selected models increases and can lead to inconsistent models.[264] This issue in interpretability is also present if using splines to model continuous variables. As the shrinkage parameter may reduce part of the spline to zero, the spline no longer makes sense as a continuous function which can make it

hard to interpret. Further, machine learning methods will intrinsically include these collinear variables in predictions, assuming generalisability of collinearity. While these methods may be worth exploring further in future work, as the aim of this analysis is to find risk factors which can then be communicated and interpreted, the methods originally used may be suitable due to the increase in biased estimates in regularisation techniques.

The large amount of missing data in screening variables led to the use of an initial screening step using a forward selection style approach before backwards elimination. If all variables were included in one model for backwards elimination, around one-third of cases and above two-thirds of controls would have been dropped from the model. An alternative approach could have been to impute the missing data. Multiple imputation is a common statistical technique to impute missing data, however can produce biased results when data is non-normally distributed without special consideration and assumes data is "missing at random", even worse than complete case analyses.[90] The complexity of my dataset and the large amounts of missing data suggest that multiple imputation may not be an optimal technique for this study. In particular, some blood test results will be far from normally distributed and, more importantly, may not be missing at random as they may be more likely to be taken in those who are more unwell. Some newer methods suggest using random forest imputation to tackle some of these issues, as well as being able to address interactions and non-linearity, and scale to high dimensions while avoiding overfitting.[265] However, these methods are computationally expensive, especially when there are many variables to impute.[265]

Checking for interactions between the variables selected for the final model could have found interesting relationships between risk factors and *E. coli* BSIs. For example, the associated higher risk of infection in individuals having previously received chemotherapy may differ by different levels of frailty. Frailer individuals may be at an increased risk of infection during chemotherapy treatment compared with less frail individuals due to increased intolerance to treatment and a higher risk of disease progression dependent on the cause of frailty.[266] Future work could look into whether including interactions between screening variables finds new at-risk groups; however sparse data and multiple testing would have to be accounted for. I did consider interactions between core variables however there was limited evidence that any were important. Interestingly, interactions between age and deprivation have previously shown different effects of deprivation for those <65y versus ≥65y;[267] however I did not find this interaction to be significant in this study.

Model stability was not assessed for the analysis presented in this Chapter and could be done in the future via bootstrapping. Model instability can occur when selected predictions are sensitive to small

changes in the data.[268] Due to the large number of variables assessed and the small number of cases as an outcome, model stability may be low and hence assessing the stability would be a useful exercise to ensure confidence in the variables selected in the final models and their effect sizes. One type of methodology for conducting bootstrapping to evaluate model stability is provided by Royston & Sauerbrei (2009).[268] This could potentially explain differences in risk factors identified across years. To explore whether changes in risk factors across years were likely to be true changes or artefacts caused by the relatively small numbers, the screening process could be run on periods of two years to increase power. Risk factors found could be compared to the screening process run on the yearly data. The number of observations retained in the core and final models differed across the years due to complete cases being used and varying amounts of missing data, potentially contributing to unstable risk factors. Around 80-90% of cases were retained in the final models for FY2019-FY2021 while only 57% of cases were used in FY2018 (**Table 5.16**). This was starker for controls with only 31% of controls being used in the final model in FY2018, compared with 71% in FY2019. The selection of HbA1c in the final model in FY2018 contributed most to the number of observations dropped in FY2018. Focussing multiple imputation of potentially important variables, such as HBA1c, with high missingness may be useful in future work. Furthermore, the lack of consistency I identified over calendar time in associated risk factors in automated, but nevertheless relatively carefully built models may explain why so few machine learning models generalise.

**Table 5.16: The number of cases and controls retained in the core and final models in each financial year in the "inpatient only" cohort.**

| Financial year | Cases/controls | Core model, n (%) | Final model, n (%) |
|---|---|---|---|
| 2018 | Cases (n=361) | 320 (89) | 206 (57) |
| | Controls (n=215,967) | 178,478 (83) | 66,175 (31) |
| 2019 | Cases (n=413) | 370 (90) | 364 (88) |
| | Controls (n=217,934) | 177,407 (81) | 154,540 (71) |
| 2020 | Cases (n=332) | 285 (86) | 266 (80) |
| | Controls (n=195,881) | 158,683 (81) | 110,384 (56) |
| 2021 | Cases (n=351) | 306 (87) | 297 (85) |
| | Controls (n=210,163) | 166,925 (79) | 127,618 (61) |

Note: Row percentages are given.

To assess whether different risk factors were found for nosocomial, quasi-nosocomial, and quasi-community cases, I ran the screening process an additional three times and compared risk factors found at the end of analyses. Using this method, I could compare which variables were present or absent, and summarise how the effect sizes differed across the groups. I could not, however, formally test the differences in these estimates as they were obtained from different models. To test whether the estimates were statistically different between the different outcomes, a multinomial

logistic regression could be used with the outcome being nosocomial, quasi-nosocomial, quasi-community *E. coli* BSIs and controls. This would however require formatting the screening process to allow for a non-binary outcome. The advantage of the current screening process is that it can be given any binary outcome and will output risk factors so different outcomes could be assessed easily if provided as binary indicators.

For this analysis, the control group consisted of all individuals with healthcare contact in Oxfordshire. If expanding this analysis to different settings, for example to national-level data, having access to all individuals may not be possible, for example due to information governance laws not allowing broad access. The dataset may also be very large if using national-level data and models may become harder to run depending on available computational capacity. One alternative may be to use a matched case-control study design selecting, for example, five controls per case. While using all data would be more efficient, matching may be a suitable alternative in scenarios where this is not possible. Matching on the core variables may be suitable. Precision of estimates may improve upon matching, particularly for age as the distribution of ages between cases and controls was quite different, with younger ages including mainly controls and very few cases. Matching on age could ensure a similar number of people across the age distribution in cases and controls, thus improving precision.[269]

### 5.4.5 Conclusion

Overall, I adapted the screening process and applied it to an EHR setting, finding interesting risk factors for *E. coli* BSIs across different years, population subgroups, and *E. coli* subgroups. Training the analysis on data from FY2019 had limitations observed when expanding the analysis to FYs 2018, 2020 and 2021, with residual issues of collinearity. Future work to further explore these remaining issues could improve the screening process and make it better for application to different datasets and diseases.

# Chapter 6 Conclusions and Future Work

This thesis aimed to identify populations at increased risk of infections using large datasets. This Chapter summarises the main findings of this thesis and discusses the direction of future work.

## 6.1 Main findings

When I began work on this thesis in October 2020 the world was at the height of the COVID-19 pandemic. With new variants emerging, a variety of control policies in place, and new interventions such as a vaccine, continuously monitoring those at an increased risk of SARS-CoV-2 positivity was important to understand the virus and control its spread. In Chapter 2 of this thesis, I therefore developed a statistical workflow denoted the "screening process" which identified populations most at risk of infections using large data. This process used standard statistical methods available in all statistical programs and was designed to be run in near real-time, coping with large numbers of potential risk factors and large amounts of missing data. I used the process to retrospectively monitor populations at risk of SARS-CoV-2 infection between July 2020 and July 2021 using data from the Office for National Statistics (ONS) COVID-19 Infection Survey (CIS). Over this period, I found that few risk factors were consistently associated with SARS-CoV-2 positivity. For example, before the Alpha variant emerged (September-November 2020), I found higher positivity in those who worked in healthcare roles or worked outside of the home, while between February-May 2021, as positivity decreased, the impact of work role/location on positivity also decreased. After the vaccine was rolled out, I consistently found reduced risk in those who had received the vaccine.

From my search of published literature, I was not able to find examples of processes to identify changing risk factors in near real-time and therefore this work contributes to the literature and helps address this gap. This screening process also addressed major challenges with real-world, large-scale data. Data were often incomplete and messy with imputation not being feasible due to the large size of the dataset. One strength of the screening process was its design for use in near real-time. This process was used numerous times by the ONS between 2020-2022 as reported in public Statistical Bulletins entitled "Characteristics of people testing positive for COVID-19" (full list on https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/results/results-longer-articles-and-adhoc-publications-from-ons).

Another aspect of identifying populations at high risk of infection was monitoring changes in the trends of the proportion testing positive. Understanding individual characteristics that increased risk from the screening model in Chapter 2 could help target interventions to specific groups; however, it was also important to get an overview of whether infections were rising or falling across the whole population and in key population sub-groups. Many methods that look for changes in trend identify

step changes (such as a change in the mean in a time series) but identifying points where the trends were changing would be more useful for infectious disease time series. In Chapter 3, I therefore compared two methods for detecting change in trends: iterative sequential regression (ISR) and second derivatives of generalised additive models (GAMs). I compared their ability to identify changes in trends in SARS-CoV-2 positivity retrospectively and in near real-time using data from CIS. Both methods identified changes in the major COVID-19 variants at similar dates and also identified smaller growth fluctuations in between these increases/decreases associated with emergence of major variants. In real-time scenarios, GAMs tended to detect change-points slightly earlier in large geographical regions; however, this was not consistent across smaller regions or subgroups. Through this Chapter's analysis, I found that both methods were suitable for detecting changes and that running both in parallel could increase confidence in the identified change-points being genuine. When I began work on this Chapter, it was unclear what methods would be most suitable for detecting changes in trends and therefore this work adds to the existing literature on change-point detection, demonstrating the use of two methods on a practical and important example.

I accompanied this work on change-points with key code I made publicly available on GitHub. This code provided an example of how to use both ISR and second derivatives of GAMs for finding change-points. This code was written in R which is a free statistical software package making the methods more accessible for people to use. This could save others time in future if they want to use these methods and provides a good starting point to understand how these methods work. This analysis code is available at https://github.com/EmmaPritchard.

The work presented in Chapters 2 and 3 of this thesis focused solely on identifying populations at risk of SARS-CoV-2. However, I wanted to use the skills and knowledge acquired during the pandemic to consider populations at risk of other infections and investigate whether the developed methods could be applied to different data sources. During my work on COVID-19, I used CIS which, as described in this thesis, was a large-scale cohort study. Questionnaires were designed for statistical analyses and the population included was randomly selected across the population of the UK, reducing the impact of bias and making results more generalisable. While these surveys have obvious benefits, they are also costly and therefore can only be implemented for limited amounts of time. Electronic health records (EHRs) offer an opportunity to assess populations at risk of other infections due to this data already being collected continuously for reimbursement purposes. However, using EHR data to identify risk factors has challenges, such as how to define a control group and how best to define risk factors. Chapter 4 of this thesis therefore bridged the gap between using the screening process on purposefully collected cohort data to using it on EHR data

collected for reimbursement purposes under a case-control design, where the conditioning outcome (defining cases/controls) was *E. coli* bloodstream infection (BSI).

I considered a variety of different control groups, differing in both the type of previous hospital contact and the length of lookback. I assessed the impact of missing data and selection bias on model results. I also compared varying definitions of risk factors derived from EHRs, considering reverse causality and potential bias in hospital populations. I found that control group choice had little impact on model estimates but a larger impact on the amount of missing data, with control groups based on inpatient admissions reducing missing data for risk factors defined from vital signs and diagnosis/procedure codes. I found that risk factors defined from blood tests recorded in the 72 hours before *E. coli* BSI collection were less likely to be distributed within the physiologically normal range versus measurements taken outside >72 hours before blood collection. I therefore decided a control group based on inpatient admissions would be suitable for my primary analyses, and I did not consider risk factors recorded in the 72 hours before *E. coli* BSI to reduce the impact of reverse causality. The suitability of a control group will vary based on the research question and the available data; however, in this thesis, I suggested potential control groups and demonstrated aspects which could be checked before analysis begins.

One output from this thesis was a spreadsheet including definitions of different risk factors which could be calculated from EHRs depending on data availability. This is something I would have found useful at the beginning of the thesis so I have aimed to make something practical that others could use and made it publicly available on my GitHub page to make access easy. I collated definitions from various pre-published coding lists, for example, procedure codes for chemotherapy[181] and different surgery groupings.[184] I tried to use published definitions where available, including references to make the document more useful to others; however, this was not always possible. I also included risk factors defined especially for this thesis in the spreadsheet, flagging where this is the case. These risk factors, such as any procedure code for urine catheterisation, will likely be relevant to other studies.

Chapter 5 of this thesis used the control groups and risk factors developed in Chapter 4, along with the screening process developed in Chapter 2, to identify individuals at higher risk of *E. coli* BSIs. I found that the presence of chemotherapy, diabetes, characteristics associated with urinary disease, and frailty were associated with increased risk of *E. coli* BSI in each financial year from 2018-2022. I also found results from a variety of blood tests to be associated with *E. coli* BSI risk; for example, higher risk with lower albumin levels and higher alkaline phosphatase levels. As the liver produces albumin, low albumin levels can be a marker for liver disease such as cirrhosis.[270] It is also often

considered a general marker of poor health with low albumin associated with malnutrition[271] and inflammation.[272] High alkaline phosphatase levels can be a marker of liver problems, such as an unresolved biliary obstruction, or bone disorders.[273] Some risk factors were associated with lower risks of *E. coli* BSI in multivariable models, for example, having more outpatient attendances was associated with a lower risk of *E. coli* BSI after adjusting for riskier procedures such as chemotherapy. Risk factors varied across the years studied, however some differences were likely due to similar variables explaining the same underlying risk factors e.g. diabetes as a risk factor being selected as either a diagnosis code for diabetes or elevated levels of HbA1c. Overall, screening characteristics in EHRs were able to identify populations at increased risk of *E. coli* BSIs.

The overall aim of identifying risk factors was to be able to use selected characteristics to inform interventions to reduce *E. coli* BSIs. Urinary catheters, which I consistently identified as associated with an increased risk of *E. coli* BSIs, would be one potential area for interventions to reduce the risk of infection. Avoiding unnecessary catheter use and removing catheters sooner when no longer needed through implementing a "stop order" (removing catheters after a certain time has elapsed) are two suggestions on how to reduce infection due to catheterisation.[274] However, where removal of a catheter is not an option, for example in bladder outlet obstruction due to enlarged prostates, addressing backlogs in non-urgent elective care may contribute to reduced catheter need. Previous chemotherapy was also consistently associated with a higher risk of *E. coli* BSI and has the potential for interventions. Patients undergoing chemotherapy have a weakened immune system making them more susceptible to infections. Gastrointestinal mucositis, the translocation of gut bacteria through the disrupted epithelial barrier into the bloodstream, also increases the risk of BSIs in chemotherapy patients.[275,276] Some suggested treatment strategies include prophylactic probiotics or antibiotics to prepare the GI tract before chemotherapy and treatments such as growth hormones after chemotherapy dosing to stimulate gut proliferation after chemotherapy-induced injury.[277] For example, a clinical trial in 2019 showed that prophylactic levofloxacin reduced febrile episodes in newly diagnosed myeloma.[278] Some identified risk factors such as frailty offer less opportunity for interventions as frailty is unlikely to be modifiable and therefore concentrating on risk factors where interventions are more possible would be a better place to start. Discussions with clinicians would help solidify optimal and feasible interventions to reduce the risk of BSIs.

One strength of this thesis was testing the screening process on data from two differing sources. I developed the process on data from CIS, a cohort study designed for research and where statistical analyses of risk factors were a pre-defined secondary objective. Risk factors were defined using the questions designed for the survey. This contrasted with using the screening process on EHRs using a case-control design, where risk factors had to be defined using pre-collected data that was not

designed for research. In both contexts, the screening process was able to identify populations with an increased risk of infections. I also applied the process to two different outcomes of SARS-CoV-2 positivity and *E. coli* BSIs and therefore believe that it could be applied to multiple other outcomes in future, either defined using specially designed cohort studies or from EHRs. Case ascertainment would have to be considered if using EHRs for different outcomes. *E. coli* BSIs worked well as an outcome as, due to the severity of BSIs, it is expected that almost all individuals with *E. coli* BSIs would be admitted to hospital and hence very few cases would be missing from analyses based on hospital EHRs. This may not be true for other, less severe, infections.

Further, CIS offered a unique opportunity to run analyses in near real-time due to the fast input of data collected from participants and having a team that processed and cleaned the data so it was available to researchers twice a week with updated swab results collected only days previously. While the screening process could theoretically be run on EHR data in near real-time, it would likely take longer for hospital records to be available and they would then have to be cleaned by the researcher before the screening process was run. While much of this data cleaning could be automated, it would likely take more time than cohort data which is purposefully collected for research and therefore more complete.

Further, when using the screening process in near real-time, it is important to decide how often to run it to get useful results. This would depend on how quickly the risk factors for an infection are expected to change. During the COVID-19 pandemic, risk factors changed quickly for various reasons including the virus mutating, the introduction of interventions such as vaccination, and changing widespread public health policies such as nationwide lockdowns. All these factors would influence the likelihood of SARS-CoV-2 infection and thus impact potential risk factors. In contrast, risk factors for *E. coli* BSIs are likely to be more stable due to fewer external factors directly impacting infection risk and a stable cause of infection. Increases in antimicrobial resistance may influence *E. coli* BSI risk factors more in the future; however, this is unlikely to be as fast evolving as aspects of the COVID-19 pandemic.

The COVID-19 pandemic highlighted the need for good pandemic preparedness. While the UK had a strategy in place for an influenza pandemic, and there had been a lot of discussion about “Disease X” as a potential unknown future threat, it did not have a specific strategy for a coronavirus pandemic.[279] A key part of the pandemic preparedness plans was surveillance and modelling, specifically detecting and assessing any new virus's impact. The methods presented in this thesis add to the knowledge base on monitoring infectious diseases in near real-time and were developed and tested during a pandemic. The screening process and comparison of ISR and second derivatives of

GAMs have been published as open-access articles in academic journals, making them available for others to use if needed.

## 6.2 Future work

### 6.2.1 Expanding to national-level data

One limitation of Chapters 4 and 5 of this thesis was that data from one hospital Trust was used, albeit covering 4 hospitals including a district general hospital in Banbury and major cancer/orthopaedic centres, and with a catchment area covering 1% of the UK. Nevertheless, this potentially limits the generalisability of the risk factors found. Future work could use national-level linked datasets held by UKHSA to assess risk factors for *E. coli* BSIs across the population of England. This would increase generalisability and reduce the risk of missed BSIs, for example, using Oxfordshire data where individuals may visit a hospital within the Trust for specialist care but attend a hospital outside the Trust for more acute illnesses. As reporting of *E. coli* infections isolated from blood is mandatory, microbiological isolations could again be used to define cases, as in this thesis. These mandatory isolations are stored in the UKHSA Second Generation Surveillance System (SGSS) and have been linked with national data from Hospital Episode Statistics (HES) by UKHSA. Therefore all the risk factors calculated from diagnoses and procedure codes could be recreated in SGSS+HES; however, there are no blood tests or vital signs recorded, some of which were significantly associated with *E. coli* BSIs in Chapter 5 of this thesis.

However, there are several considerations when expanding the screening model to national-level data. Firstly, test results from negative cultures are not recorded in SGSS. In Chapter 5, I consistently showed that having any previous urine culture taken irrespective of the result was associated with an increased risk of *E. coli* BSIs. These potential risk factors could not be defined from negative test results in SGSS+HES and therefore some important risk factors may be missed. While some risk factors defined using microbiological isolations are also sometimes recorded in procedure codes, I found little agreement between the recording of urinary catheter use in microbiology versus procedure code data. Urinary catheter use was associated with a higher risk of *E. coli* BSIs with evidence at $p<0.05$ in many multivariate models; however, when swapping this variable to urinary catheter use as recorded by procedure codes, there was no evidence of an association at $p<0.05$ or even higher thresholds. Therefore, relying on recordings of risk factors from procedure codes may miss important associations. Secondly, in Chapter 5 I used all individuals with an inpatient admission, outpatient or A&E visit, or microbiology or blood test done in Oxfordshire as the control group. Gaining access to data from all these individuals using national-level data, i.e. most of the UK population, may be difficult due to data governance rules. A matched case-control study using, for

example, five controls per case may be more suited for this future analysis. The impact of a matched approach could first be assessed using data from the Infections in Oxfordshire Research Database before implementation on national-level data.

### 6.2.2 The screening process

While the screening process presented in this thesis was useful for identifying risk factors, it did have limitations. Forward selection and backward elimination formed the basis for variable selection; however these methods have disadvantages related to the rigidity of the approach, with forward selection potentially missing confounders if the confounder has no association univariably and backward elimination not allowing variable re-entry after elimination. This approach was selected due to large amounts of missing data. Including all variables in a backwards elimination dropped around one-third of cases and two-thirds of controls. The initial forward selection step removed variables with large amounts of missing data that did not appear statistically important. Multiple imputation could have been used, however the complexity of the data presents challenges when attempting to get multiple imputation models with complex non-linearity correct. Newer methods using random forest imputation may have been more suitable but these methods are computationally expensive. Regression using penalisation techniques could have been used instead of forward/backward selection/elimination; however, estimates from these methods can be harder to interpret as they shrink estimates towards zero, making continuous parameterisations using splines particularly hard to understand.

In Chapter 5, the screening process occasionally identified different variables for the same underlying risk factor when comparing across different years and subgroups. For example, one year might select a chest scan while another year identified pneumonia, or HbA1c rather than a diagnosis code for diabetes. A new method developed by colleagues at the University of Oxford in 2023/2024 which aimed to tackle this was “Doublethink”.[280,281] Similar to the screening process, this method was also designed to consider a large number of variables agnostically. In contrast with the screening process, Doublethink also considered groups of variables clustered together based on correlation. This approach was demonstrated on data from the UK Biobank, considering 1,912 potential risk factors on the outcome of COVID-19 hospitalisation. Some groups contained clustered variables which were very highly correlated; for example, one group consisted of diagnosis codes for unspecified dementia, Alzheimer's disease (unspecified), and Dementia in Alzheimer's disease (unspecified). Some groups also highlighted interesting underlying phenotypes, for example, a grouping of constipation with urinary tract infections, which may reflect underlying kidney disease or incomplete bladder emptying associated with UTI.[282] As I observed in Chapter 5, many variables defined using EHRs were highly correlated, leading to nonsensical results when including correlated

variables in the same model. Using the dataset curated in Chapter 4, I would therefore like to use the Doublethink process and compare the risk factors/groups identified to those selected in Chapter 5 in this thesis.

Some potential interventions after establishing risk factors were discussed earlier in this conclusion, for example avoiding unnecessary catheter use; however, further steps would be needed to implement interventions. Firstly, it would be important to decide which risk factors to intervene with initially. In Chapter 5, I looked at the variance explained by risk factors selected in the final multivariate models. Beginning with targeting interventions at modifiable risk factors which explain the most variation would be sensible. The population attributable fraction would also be useful to calculate as it considers both the estimated risk and the prevalence of the exposure among cases, hence helping to optimise interventions which not only highly increase risk but also impact a higher number of cases therefore resulting in a larger reduction of BSIs.[283] Investigations into individual risk factors within cases may also be useful in establishing possible interventions. For example, I found a higher risk of quasi-nosocomial *E. coli* BSIs in individuals with recent prosthesis procedures. A deeper dive into which prostheses were recorded for these individuals with *E. coli* BSIs could help to establish useful interventions. Further, focussing on risk factors identified in nosocomial, quasi-nosocomial, or quasi-community cases may be valuable as, by definition, these individuals have had contact as an inpatient either currently, 0-30 days ago, and 31-365 days ago and are therefore a group where intervention would be possible. In contrast, community-acquired BSIs may be harder to target using results from this study as they have no contact as an inpatient in the previous year and so healthcare interventions are less likely to impact them. Assessing risk factors for community-required BSIs would require data from the community, whether that be through primary care (GP) records or using population-level data such as the census. Some risk factors such as age had a large impact on *E. coli* BSI risk, however cannot be changed and therefore interventions targeting them are not possible. For unmodifiable risk factors, such as chemotherapy and frailty, more intensive awareness may be useful. Focussing on risk factors which can be targeted whilst adjusting for key confounders such as age and frailty would be a useful approach moving forward.

Future work could also consider clustering risk factors to identify at-risk “phenotypes”. Clustering could be done either before screening with the clusters identified then modelled as factors, or after risk factors had been identified to summarise the specific factors in different subgroups most as risk. Many of the risk factors identified in Chapter 5 could frequently occur together in individuals, for example, pneumonia and higher frailty scores. Establishing groups of individuals may help target interventions to groups more effectively, particularly if the occurrence of specific characteristics combined particularly increases *E. coli* BSI risk. Other studies have used clustering to group

individuals and assess risk factors in those groups, for example, clustering cardiovascular patients identified younger individuals with an unhealthy lifestyle at higher risk of poor medication adherence suggesting that these patients should be offered more guidance, while interventions were not as important in other clusters.[284] As there are many different clustering methods, future work could start with a literature search to establish a suitable approach, starting by looking at studies which have previously combined risk factor identification with cluster analysis.

In Chapter 5, I observed variation in the risk factors identified across financial years from 2018 to 2022. These changes could either be genuine variations in risk factors between years or a product of model instability due to the small number of *E. coli* BSI cases each year in the dataset. The total number of *E. coli* BSI cases did decrease over the pandemic, likely due to the competing risks with the oldest individuals dying from COVID-19 and therefore no longer at risk for *E. coli* BSIs.[285] Further work could assess whether changes in risk factors were a product of true variation in multiple ways. Firstly, I could run the screening process on data from FY2022 and FY2023 and observe whether risk factors were stable in these years. If risk factors stabilised from FY2022 onwards, it could be plausible that the COVID-19 pandemic impacted the variation in risk factors in FY2020 and FY2021. Secondly, I could increase statistical power by (i) re-running the analysis in Chapter 5 but grouping FYs into two-year periods or (ii) running the analysis on national-level data as discussed above. With more power, if risk factors still varied each year, it would be more credible that this was true variation. If risk factors instead stabilised, it would be evidence that the small number of cases was impacting model stability and wider time intervals to increase dataset size, or using national-level data, may be more appropriate moving forward.

Finally, the screening process could be used to identify risk factors for other bloodstream infections. Identified risk factors could be compared with those found in *E. coli* BSIs. As the second most common BSI, *Staphylococcus aureus* would be an obvious place to start as it is relatively frequent and can be identified from microbiological isolations. Other key pathogens identified in the ESPAUR report could also be targets for the screening process, including *Klebsiella*, *Pseudomonas*, and *Enterococcus* species.[61] Using EHRs, risk factors for diseases identified using ICD-10 codes could also be studied, such as endocarditis, acknowledging that ICD-10 codes are often not as reliable as microbiological isolations and using them without assessing their sensitivity to predictive ability can lead to missed cases and inaccurate results.[171] More generally, this study had wider importance than its specific application to finding risk factors of *E. coli* BSIs as EHRs are increasingly used for research.

## 6.3 Concluding remarks

This thesis has demonstrated methods which can identify populations at increased risk of infectious diseases using large datasets. The screening process developed in Chapter 2 was able to identify populations at increased risk of SARS-CoV-2 positivity in near real-time, with it being successfully implemented by the ONS during the pandemic to monitor characteristics of those at increased risk. I also compared methods for detecting changes in infection trends, being able to identify when trends were increasing and decreasing using two methods, ISR and second derivatives of GAMs, which generally performed equally well and could be used together to increase confidence in findings. I was able to use this learning from throughout the COVID-19 pandemic to apply the screening process to identify risk factors for *E. coli* BSIs in EHR data. Future work could use these methods to identify and monitor populations at increased risk of various infectious diseases using large-scale data.

# References


1. Snow J. On the Mode of Communication of Cholera. 2 ed: John Churchill; 1855.
2. Begum F. Mapping disease: John Snow and Cholera. 2016. https://www.rcseng.ac.uk/library-and-publications/library/blog/mapping-disease-john-snow-and-cholera/ (accessed 10 June 2024).
3. Bradshaw NA. Florence Nightingale (1820-1910): An Unexpected Master of Data. *Patterns (N Y)* 2020; **1**(2): 100036.
4. Framingham Heart Study. About the Framingham Heart Study. 2024. https://www.framinghamheartstudy.org/fhs-about/ (accessed 10 June 2024).
5. Bitton A, Gaziano TA. The Framingham Heart Study's impact on global risk assessment. *Prog Cardiovasc Dis* 2010; **53**(1): 68-78.
6. Brittain E. Probability of coronary heart disease developing. *West J Med* 1982; **136**(1): 86-9.
7. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. *Am J Cardiol* 1976; **38**(1): 46-51.
8. Di Cicco ME, Ragazzo V, Jacinto T. Mortality in relation to smoking: the British Doctors Study. *Breathe (Sheff)* 2016; **12**(3): 275-6.
9. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits; a preliminary report. *Br Med J* 1954; **1**(4877): 1451-5.
10. Doll R, Peto R. Mortality in relation to smoking: 20 years' observations on male British doctors. *Br Med J* 1976; **2**(6051): 1525-36.
11. Opazo Breton M, Gillespie D, Pryce R, et al. Understanding long-term trends in smoking in England, 1972-2019: an age-period-cohort approach. *Addiction* 2022; **117**(5): 1392-403.
12. National Cancer Institute. electronic health record. 2024. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/electronic-health-record (accessed 10 June 2024).
13. NHS. Understanding and using the national tariff, 2020.
14. Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL. The Evolving Use of Electronic Health Records (EHR) for Research. *Semin Radiat Oncol* 2019; **29**(4): 354-61.
15. Manca DP. Do electronic medical records improve quality of care? Yes. *Can Fam Physician* 2015; **61**(10): 846-7, 50-1.
16. Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP system. *J Med Syst* 1983; **7**(2): 87-102.
17. Barnett GO. Computer-stored ambulatory record (COSTAR): US Department of Health, Education, and Welfare, Public Health Service …; 1976.
18. Evans RS. Electronic health records: then, now, and in the future. *Yearbook of Medical Informatics* 2016; **25**(S 01): S48-S61.
19. Ornstein SM, Oates RB, Fox GN. The computer-based medical record: current status. *J Fam Pract* 1992; **35**(5): 556-65.
20. Burns F. Information for Health: NHS Executive, 1998.
21. Department of Health. Delivering 21st Century IT Support for the NHS, 2002.
22. Robertson A, Cresswell K, Takian A, et al. Implementation and adoption of nationwide electronic health records in secondary care in England: qualitative analysis of interim results from a prospective national evaluation. *BMJ* 2010; **341**: c4564.
23. Jiang JX, Qi K, Bai G, Schulman K. Pre-pandemic assessment: a decade of progress in electronic health record adoption among U.S. hospitals. *Health Aff Sch* 2023; **1**(5): qxad056.
24. NHS England. 90% of NHS trusts now have electronic patient records. 2023. https://digital.nhs.uk/news/2023/90-of-nhs-trusts-now-have-electronic-patient-records (accessed 9 May 2024).

25. Woldemariam MT, Jimma W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. *BMJ Health & Care Informatics* 2023; **30**(1): e100704.
26. Liang L, Wiens MO, Lubega P, Spillman I, Mugisha S. A Locally Developed Electronic Health Platform in Uganda: Development and Implementation of Stre@mline. *JMIR Form Res* 2018; **2**(2): e20.
27. Fraser HSF, Mugisha M, Remera E, et al. User Perceptions and Use of an Enhanced Electronic Health Record in Rwanda With and Without Clinical Alerts: Cross-sectional Survey. *JMIR Med Inform* 2022; **10**(5): e32305.
28. Zharima C, Griffiths F, Goudge J. Exploring the barriers and facilitators to implementing electronic health records in a middle-income country: a qualitative study from South Africa. *Front Digit Health* 2023; **5**: 1207602.
29. World Health Organization. Global strategy on digital health 2020-2025, 2021.
30. Clarke GM, Conti S, Wolters AT, Steventon A. Evaluating the impact of healthcare interventions using routine data. *BMJ* 2019; **365**: l2239.
31. Pennington M, Grieve R, Sekhon JS, Gregg P, Black N, van der Meulen JH. Cemented, cementless, and hybrid prostheses for total hip replacement: cost effectiveness analysis. *BMJ* 2013; **346**: f1026.
32. Grout R, Gupta R, Bryant R, et al. Predicting disease onset from electronic health records for population health management: a scalable and explainable Deep Learning approach. *Front Artif Intell* 2023; **6**: 1287541.
33. Strongman H, Gadd S, Matthews A, et al. Medium and long-term risks of specific cardiovascular diseases in survivors of 20 adult cancers: a population-based cohort study using multiple linked UK electronic health records databases. *Lancet* 2019; **394**(10203): 1041-54.
34. Qizilbash N, Gregson J, Johnson ME, et al. BMI and risk of dementia in two million people over two decades: a retrospective cohort study. *Lancet Diabetes Endocrinol* 2015; **3**(6): 431-6.
35. Sudat SEK, Robinson SC, Mudiganti S, Mani A, Pressman AR. Mind the clinical-analytic gap: Electronic health records and COVID-19 pandemic response. *J Biomed Inform* 2021; **116**: 103715.
36. Quan TP, Lacey B, Peto TEA, Walker AS. Health record hiccups—5,526 real-world time series with change points labelled by crowdsourced visual inspection. *GigaScience* 2023; **12**.
37. Kass RE, Caffo BS, Davidian M, Meng XL, Yu B, Reid N. Ten Simple Rules for Effective Statistical Practice. *PLoS Comput Biol* 2016; **12**(6): e1004961.
38. Schwab P, Mehrjou A, Parbhoo S, et al. Real-time prediction of COVID-19 related mortality using electronic health records. *Nat Commun* 2021; **12**(1): 1058.
39. Sheikhtaheri A, Tabatabaee Jabali SM, Bitaraf E, TehraniYazdi A, Kabir A. A near real-time electronic health record-based COVID-19 surveillance system: An experience from a developing country. *Health Inf Manag* 2024; **53**(2): 145-54.
40. Harron K. Data linkage in medical research. *BMJ Med* 2022; **1**(1): e000087.
41. Wellcome Trust. Enabling Data Linkage to Maximise the Value of Public Health Research Data: Summary, 2015.
42. Moorthie S, Hayat S, Zhang Y, et al. Rapid systematic review to identify key barriers to access, linkage, and use of local authority administrative data for population health research, practice, and policy in the United Kingdom. *BMC Public Health* 2022; **22**(1): 1263.
43. Wood A, Denholm R, Hollings S, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021; **373**: n826.

44. NIHR Oxford Biomedial Research Centre. What is the Infections in Oxfordshire Research Database (IORD)? 2024. https://oxfordbrc.nihr.ac.uk/research-themes/modernising-medical-microbiology-and-big-infection-diagnostics/iord-about/ (accessed 14 June 2024).
45. NHS Digital. Hospital Episode Statistics (HES). 2024. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics (accessed 24 April 2024).
46. CPRD. Clinical Practice Research Datalink. 2024. https://www.cprd.com/ (accessed 14 June 2024).
47. Banerjee A, Dashtban A, Chen S, et al. Identifying subtypes of heart failure from three electronic health record sources with machine learning: an external, prognostic, and genetic validation study. *Lancet Digit Health* 2023; **5**(6): e370-e9.
48. McDonald HI, Thomas SL, Millett ER, Nitsch D. CKD and the risk of acute, community-acquired infections among older people with diabetes mellitus: a retrospective cohort study using electronic health records. *Am J Kidney Dis* 2015; **66**(1): 60-8.
49. OpenSAFELY. About OpenSAFELY. 2024. https://www.opensafely.org/about/ (accessed 17 June 2024).
50. Nab L, Schaffer AL, Hulme W, et al. OpenSAFELY: A platform for analysing electronic health records designed for reproducible research. *Pharmacoepidemiology and Drug Safety* 2024; **33**(6): e5815.
51. Cavallaro F, Lugg-Widger F, Cannings-John R, Harron K. Reducing barriers to data access for research in the public interest—lessons from covid-19. *BMJ Opinion* 2020; **6**.
52. NHS England. Control of patient information (COPI) notice. 2020. https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/control-of-patient-information-copi-notice (accessed 8 July 2024).
53. NHS England. Data Access Request Service (DARS): process. 2024. https://digital.nhs.uk/services/data-access-request-service-dars/process (accessed 17 June 2024).
54. Office for National Statistics. Secure Research Service. 2024. https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice (accessed 17 June 2024).
55. Department of Health and Social Care. Better, broader, safer: using health data for research and analysis, 2022.
56. UK Health Security Agency. Mandatory healthcare associated infection surveillance: data quality statement for April 2019 to March 2020, 2022.
57. Johnson AP, Davies J, Guy R, et al. Mandatory surveillance of methicillin-resistant Staphylococcus aureus (MRSA) bacteraemia in England: the first 10 years. *J Antimicrob Chemother* 2012; **67**(4): 802-9.
58. Ashiru-Oredope D, Hopkins S. Antimicrobial stewardship: English Surveillance Programme for Antimicrobial Utilization and Resistance (ESPAUR). *J Antimicrob Chemother* 2013; **68**(11): 2421-3.
59. Public Health England. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR), 2014.
60. UK Health Security Agency. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR) report. 2024. https://www.gov.uk/government/publications/english-surveillance-programme-antimicrobial-utilisation-and-resistance-espaur-report (accessed 12 June 2024).
61. UK Health Security Agency. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR). Report 2022 to 2023., 2023.
62. Quan TP, Hope R, Clarke T, et al. Using linked electronic health records to report healthcare-associated infections. *PLoS One* 2018; **13**(11): e0206860.

63. Spiteri G, Fielding J, Diercke M, et al. First cases of coronavirus disease 2019 (COVID-19) in the WHO European Region, 24 January to 21 February 2020. *Euro Surveill* 2020; **25**(9).
64. World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. 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 12 June 2024).
65. World Health Organization. Number of COVID-19 cases reported to WHO (cumulative total). 2024. https://data.who.int/dashboards/covid19/cases?n=c (accessed 8 July 2024).
66. World Health Organization. Number of COVID-19 deaths reported to WHO (cumulative total). 2024. https://data.who.int/dashboards/covid19/deaths?n=c (accessed 8 July 2024).
67. GOV.UK. UK marks one year since deploying world's first COVID-19 vaccine. 2021.
68. Wellcome. 2021. https://wellcome.org/news/quick-safe-covid-vaccine-development (accessed 12 June 2024).
69. Department of Health and Social Care. Experimental statistics Weekly NHS Test and Trace bulletin, England: 28 May – 3 June 2020, 2020.
70. Centers for Disease Control and Prevention. SARS-CoV-2 Variant Classifications and Definitions. 2023. https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html (accessed 14 June 2024).
71. Institute for Government. Timeline of UK government coronavirus lockdowns. 2021.
72. Office for National Statistics. Coronavirus (COVID-19) Infection Survey: quality and methodology information (QMI). 2023. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/methodologies/coronaviruscovid19infectionsurveyqmi (accessed 12 June 2024).
73. Imperial. About the REACT programme. 2024. https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/ (accessed 12 June 2024).
74. Zoe Health Study. 2024. https://health-study.zoe.com/data (accessed 12 June 2024).
75. GOV.UK. SIREN study. 2022. https://www.gov.uk/guidance/siren-study (accessed 12 June 2024).
76. Bhattacharya A, Collin SM, Stimson J, et al. Healthcare-associated COVID-19 in England: a national data linkage study. *Journal of Infection* 2021; **83**(5): 565-72.
77. Nyberg T, Ferguson NM, Nash SG, et al. Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 omicron (B.1.1.529) and delta (B.1.617.2) variants in England: a cohort study. *Lancet* 2022; **399**(10332): 1303-12.
78. Andrews N, Stowe J, Kirsebom F, et al. Covid-19 Vaccine Effectiveness against the Omicron (B.1.1.529) Variant. *N Engl J Med* 2022; **386**(16): 1532-46.
79. Ma Q, Liu J, Liu Q, et al. Global Percentage of Asymptomatic SARS-CoV-2 Infections Among the Tested Population and Individuals With Confirmed COVID-19 Diagnosis: A Systematic Review and Meta-analysis. *JAMA Netw Open* 2021; **4**(12): e2137257.
80. Bajaj S, Chen S, Creswell R, et al. Understanding COVID-19 testing behaviour in England through a sociodemographic lens. *medRxiv* 2023: 2023.10. 26.23297608.
81. Pritchard E, Jones J, Vihta KD, et al. Monitoring populations at increased risk for SARS-CoV-2 infection in the community using population-level demographic and behavioural surveillance. *Lancet Reg Health Eur* 2022; **13**: 100282.
82. Public Health England. Disparities in the risk and outcomes of COVID-19. *Public Health England* 2020.
83. de Lusignan S, Dorward J, Correa A, et al. 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* 2020; **20**(9): 1034-42.
84. Zheng Z, Peng F, Xu B, et al. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. *J Infect* 2020; **81**(2): e16-e25.

85. Elimian KO, Ochu CL, Ebhodaghe B, et al. Patient characteristics associated with COVID-19 positivity and fatality in Nigeria: retrospective cohort study. *BMJ open* 2020; **10**(12): e044079.
86. World Health Organization. Tracking SARS-CoV-2 variants. 2021. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ (accessed 29 July 2021).
87. GOV.UK. Full list of local restriction tiers by area. 2021. https://www.gov.uk/guidance/full-list-of-local-restriction-tiers-by-area (accessed 29 July 2021).
88. Sah P, Fitzpatrick MC, Zimmer CF, et al. Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. *Proceedings of the National Academy of Sciences* 2021; **118**(34).
89. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; **10**(2): e1001381.
90. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; **338**.
91. Hughes RA, Heron J, Sterne JA, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International journal of epidemiology* 2019; **48**(4): 1294-304.
92. Johnston R, Jones K, Manley D. Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Qual Quant* 2018; **52**(4): 1957-76.
93. Riley S, Atchison C, Ashby D, et al. REal-time Assessment of Community Transmission (REACT) of SARS-CoV-2 virus: Study protocol. *Wellcome Open Res* 2020; **5**: 200.
94. Ministry of Housing, Communities & Local Government,. English indices of deprivation 2019. 2019. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 (accessed 12 April 2024).
95. Ministry of Housing, Communities & Local Government,. English indices of deprivation 2019, 2019.
96. Northern Ireland Statistics and Research Agency. Northern Ireland Multiple Deprivation Measure 2017 (NIMDM2017), 2017.
97. Scottish Government. Scottish Index of Multiple Deprivation 2020, 2020.
98. Statistics for Wales. Welsh Index of Multiple Deprivation (full Index update with ranks): 2019, 2019.
99. Chang BH, Hoaglin DC. Meta-Analysis of Odds Ratios: Current Good Practices. *Med Care* 2017; **55**(4): 328-35.
100. Doerken S, Avalos M, Lagarde E, Schumacher M. Penalized logistic regression with low prevalence exposures beyond high dimensional settings. *PLoS One* 2019; **14**(5): e0217057.
101. Office for National Statistics. COVID-19 Infection Survey: methods and further information. 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/methodologies/covid19infectionsurveypilotmethodsandfurtherinformation (accessed 29 July 2021).
102. Vihta KD, Pouwels KB, Peto TEA, et al. Symptoms and Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Positivity in the General Population in the United Kingdom. *Clin Infect Dis* 2022; **75**(1): e329-e37.
103. Walker AS, Vihta KD, Gethings O, et al. Increased infections, but not viral burden, with a new SARS-CoV-2 variant. *medRxiv* 2021.
104. Public Health England. SARS-CoV-2 variants of concern and variants under investigation in England, 2021.
105. Understanding Herefordshire. Living environment deprivation - Indoor living environment sub-domain. 2021. https://understanding.herefordshire.gov.uk/inequalities/index-of-multiple-deprivation-imd/living-environment-deprivation-indoor-living-environment-sub-domain/ (accessed 19 November 2021).

106. Riley S, Ainslie KE, Eales O, et al. High and increasing prevalence of SARS-CoV-2 swab positivity in England during end September beginning October 2020: REACT-1 round 5 updated report. *medRxiv* 2020.
107. Riley S, Walters CE, Wang H, et al. REACT-1 round 7 updated report: regional heterogeneity in changes in prevalence of SARS-CoV-2 infection during the second national COVID-19 lockdown in England. *medRxiv* 2020.
108. Riley S, Eales O, Walters CE, et al. REACT-1 round 8 final report: high average prevalence with regional heterogeneity of trends in SARS-CoV-2 infection in the community in England during January 2021. *medRxiv* 2021.
109. Riley S, Walters CE, Wang H, et al. REACT-1 round 12 report: resurgence of SARS-CoV-2 infections in England associated with increased frequency of the Delta variant. *medRxiv* 2021.
110. Dyer O. Covid-19: Unvaccinated face 11 times risk of death from delta variant, CDC data show. *Bmj* 2021; **374**: n2282.
111. Scobie HM, Johnson AG, Suthar AB, et al. Monitoring Incidence of COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Status - 13 U.S. Jurisdictions, April 4-July 17, 2021. *MMWR Morb Mortal Wkly Rep* 2021; **70**(37): 1284-90.
112. Thompson HA, Mousa A, Dighe A, et al. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) setting-specific transmission rates: a systematic review and meta-analysis. *Clinical Infectious Diseases* 2021.
113. Timmins N. Schools and coronavirus The government's handling of education during the pandemic: Institute for Government 2021.
114. Office for National Statistics. Coronavirus (COVID-19) latest insights: Hospitals. 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19latestinsights/hospitals (accessed 07 October 2021).
115. Ueki H, Furusawa Y, Iwatsuki-Horimoto K, et al. Effectiveness of Face Masks in Preventing Airborne Transmission of SARS-CoV-2. *mSphere* 2020; **5**(5).
116. Usman MS, Siddiqi TJ, Khan MS, et al. Is there a smoker's paradox in COVID-19? *BMJ evidence-based medicine* 2020.
117. Elliott P, Haw D, Wang H, et al. REACT-1 round 13 final report: exponential growth, high prevalence of SARS-CoV-2 and vaccine effectiveness associated with Delta variant in England during May to July 2021. *medRxiv* 2021.
118. Krzywinski M, Altman N. Power and sample size. *Nature Methods* 2013; **10**(12): 1139-40.
119. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med Res Methodol* 2009; **9**: 56.
120. van Smeden M, de Groot JA, Moons KG, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol* 2016; **16**(1): 163.
121. Albers C. The problem with unadjusted multiple and sequential statistical testing. *Nat Commun* 2019; **10**(1): 1921.
122. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995; **57**(1): 289-300.
123. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* 2000; **25**(1): 60-83.
124. Riley S, Eales O, Haw D, et al. REACT-1 round 13 interim report: acceleration of SARS-CoV-2 Delta epidemic in the community in England during late June and early July 2021. *medRxiv* 2021.
125. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med* 2019; **38**(11): 2074-102.

126. UK Department of Health & Social Care. NHS Test and Trace Statistics (England): Methodology. 2021. https://www.gov.uk/government/publications/nhs-test-and-trace-statistics-england-methodology/nhs-test-and-trace-statistics-england-methodology (accessed 29 July 2021).
127. Pan American Health Organisation. Interim guidelines for detecting cases of reinfection by SARS-CoV-2, 2020.
128. Pritchard E, Vihta KD, Eyre DW, et al. Detecting changes in population trends in infection surveillance using community SARS-CoV-2 prevalence as an exemplar. *Am J Epidemiol* 2024.
129. Venkatesan P. Rise in group A streptococcal infections in England. *Lancet Respir Med* 2023; **11**(2): e16.
130. McDonald LC, Killgore GE, Thompson A, et al. An epidemic, toxin gene-variant strain of Clostridium difficile. *N Engl J Med* 2005; **353**(23): 2433-41.
131. Davies NG, Abbott S, Barnard RC, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* 2021; **372**(6538).
132. World Health Organization. Methods for the detection and characterisation of SARS-CoV-2 variants: second update, 21 June 2022: World Health Organization. Regional Office for Europe, 2022.
133. Fearnhead P, Rigaill G. Changepoint detection in the presence of outliers. *Journal of the American Statistical Association* 2019; **114**(525): 169-83.
134. Aminikhanghahi S, Cook DJ. A Survey of Methods for Time Series Change Point Detection. *Knowl Inf Syst* 2017; **51**(2): 339-67.
135. Dehning J, Zierenberg J, Spitzner FP, et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science* 2020; **369**(6500).
136. Dass SC, Kwok WM, Gibson GJ, Gill BS, Sundram BM, Singh S. A data driven change-point epidemic model for assessing the impact of large gathering and subsequent movement control order on COVID-19 spread in Malaysia. *PLoS One* 2021; **16**(5): e0252136.
137. Jiang F, Zhao Z, Shao X. Time series analysis of COVID-19 infection curve: A change-point perspective. *J Econom* 2023; **232**(1): 1-17.
138. Coughlin SS, Yiğiter A, Xu H, Berman AE, Chen J. Early detection of change patterns in COVID-19 incidence and the implementation of public health policies: A multi-national study. *Public Health Pract (Oxf)* 2021; **2**: 100064.
139. Schlackow I, Walker AS, Dingle K, et al. Surveillance of infection severity: a registry study of laboratory diagnosed Clostridium difficile. *PLoS Med* 2012; **9**(7): e1001279.
140. Vihta KD, Stoesser N, Llewelyn MJ, et al. Trends over time in Escherichia coli bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998-2016: a study of electronic health records. *Lancet Infect Dis* 2018; **18**(10): 1138-49.
141. Fewster RM, Buckland ST, Siriwardena GM, Baillie SR, Wilson JD. Analysis of population trends for farmland birds using generalized additive models. *Ecology* 2000; **81**(7): 1970-84.
142. Simpson GL. Modelling palaeoecological time series using generalised additive models. *Frontiers in Ecology and Evolution* 2018: 149.
143. Mair C, Wulaningsih W, Jeyam A, et al. Glycaemic control trends in people with type 1 diabetes in Scotland 2004-2016. *Diabetologia* 2019; **62**(8): 1375-84.
144. Wood SN. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2003; **65**(1): 95-114.
145. Pedersen EJ, Miller DL, Simpson GL, Ross N. Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ* 2019; **7**: e6876.
146. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2011; **73**(1): 3-36.
147. Wood SN. Generalized additive models: an introduction with R: chapman and hall/CRC; 2006.

148. Simpson GL. *gratia*: Graceful *ggplot*-Based Graphics and Other Functions for *GAMs* Fitted using *mgcv*. 2022. https://gavinsimpson.github.io/gratia/ (accessed 4 April 2022).
149. Pouwels KB, House T, Pritchard E, et al. Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. *Lancet Public Health* 2021; **6**(1): e30-e8.
150. UK Government. Prime Minister confirms move to Plan B in England. 2021.
151. Macartney K, Quinn HE, Pillsbury AJ, et al. Transmission of SARS-CoV-2 in Australian educational settings: a prospective cohort study. *Lancet Child Adolesc Health* 2020; **4**(11): 807-16.
152. Mensah AA, Sinnathamby M, Zaidi A, et al. SARS-CoV-2 infections in children following the full re-opening of schools and the impact of national lockdown: Prospective, national observational cohort surveillance, July-December 2020, England. *J Infect* 2021; **82**(4): 67-74.
153. Ismail SA, Saliba V, Lopez Bernal J, Ramsay ME, Ladhani SN. SARS-CoV-2 infection and transmission in educational settings: a prospective, cross-sectional analysis of infection clusters and outbreaks in England. *Lancet Infect Dis* 2021; **21**(3): 344-53.
154. Ladhani SN, Ireland G, Baawuah F, et al. SARS-CoV-2 infection, antibody positivity and seroconversion rates in staff and students following full reopening of secondary schools in England: A prospective cohort study, September-December 2020. *EClinicalMedicine* 2021; **37**: 100948.
155. Li Y, Campbell H, Kulkarni D, et al. The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: a modelling study across 131 countries. *Lancet Infect Dis* 2021; **21**(2): 193-202.
156. Romero Starke K, Reissig D, Petereit-Haack G, Schmauder S, Nienhaus A, Seidler A. The isolated effect of age on the risk of COVID-19 severe outcomes: a systematic review with meta-analysis. *BMJ Glob Health* 2021; **6**(12).
157. Wyllie DH, Walker AS, Miller R, et al. Decline of meticillin-resistant Staphylococcus aureus in Oxfordshire hospitals is strain-specific and preceded infection-control intensification. *BMJ Open* 2011; **1**(1): e000160.
158. Rambaut AL, N. Pybus, O. Barclay, W. Barrett, J. Carabelli, A. Connor, R. Peacock, T. Robertson, D L. Volz, E. COVID-19 Genomics Consortium UK (CoG-UK) Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations, 2020.
159. Public Health England. Investigation of novel SARS-CoV-2 variant Variant of Concern 202012/01 2021.
160. BBC. Covid-19: Six million more people to enter tier 4 on Boxing Day. BBC. 2020.
161. Public Health England. SARS-CoV-2 variants of concern and variants under investigation in England. Technical briefing 10. , 2021.
162. UK Health Security Agency. SARS-CoV-2 variants of concern and variants under investigation in England. Technical briefing 39. , 2022.
163. Public Health England. COVID-19 surge testing outcomes reports: management information 2021. https://www.gov.uk/government/statistical-data-sets/covid-19-surge-testing-outcomes-reports-management-information (accessed 19 April 2022).
164. UK Health Security Agency. Surge testing for new coronavirus (COVID-19) variants. 2021. https://www.gov.uk/guidance/surge-testing-for-new-coronavirus-covid-19-variants (accessed 19 April 2022).
165. Sauer CM, Chen LC, Hyland SL, Girbes A, Elbers P, Celi LA. Leveraging electronic health records for data science: common pitfalls and how to avoid them. *Lancet Digit Health* 2022; **4**(12): e893-e8.
166. Mann CJ. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emerg Med J* 2003; **20**(1): 54-60.

167. Lutsey PL. Case-control studies: Increasing scientific rigor in control selection. *Res Pract Thromb Haemost* 2023; **7**(2): 100090.
168. Griffith GJ, Morris TT, Tudball MJ, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun* 2020; **11**(1): 5749.
169. Casas JP, Shah T, Cooper J, et al. Insight into the nature of the CRP-coronary event association using Mendelian randomization. *Int J Epidemiol* 2006; **35**(4): 922-31.
170. Hripcsak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. *J Am Med Inform Assoc* 2013; **20**(e2): e311-8.
171. Fawcett N, Young B, Peto L, et al. 'Caveat emptor': the cautionary tale of endocarditis and the potential pitfalls of clinical coding data-an electronic health records study. *BMC Med* 2019; **17**(1): 169.
172. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 2018; **361**: k1479.
173. UK Biobank. Hospital inpatient data, 2023.
174. Yoon CH, Bartlett S, Stoesser N, et al. Mortality risks associated with empirical antibiotic activity in Escherichia coli bacteraemia: an analysis of electronic health records. *J Antimicrob Chemother* 2022; **77**(9): 2536-45.
175. Le HV, Poole C, Brookhart MA, et al. Effects of aggregation of drug and diagnostic codes on the performance of the high-dimensional propensity score algorithm: an empirical example. *BMC Med Res Methodol* 2013; **13**: 142.
176. Tazare J, Smeeth L, Evans SJ, Douglas IJ, Williamson EJ. hdps: A suite of commands for applying high-dimensional propensity-score approaches. *The Stata Journal* 2023; **23**(3): 683-708.
177. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987; **40**(5): 373-83.
178. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005; **43**(11): 1130-9.
179. Dr Foster Intelligence. Understanding HSMRs. A Toolkit on Hospital Standardised Mortality Ratios. 2012.
180. Gilbert T, Neuburger J, Kraindler J, et al. Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records: an observational study. *Lancet* 2018; **391**(10132): 1775-82.
181. NHS. The NHS Classifications Browser. 2024. https://classbrowser.nhs.uk/#/ (accessed 30 May 2024).
182. OpenCodelists. Dialysis OPCS-4. 2022. https://www.opencodelists.org/codelist/ukrr/dialysis-opcs-4/505bef04/ (accessed 30 May 2024).
183. OpenCodelists. Transplant (SPL-HES) (OPCS4). 2022. https://www.opencodelists.org/codelist/nhsd/transplant-spl-hes-opcs4/1b47de52/ (accessed 30 May 2024).
184. UK Health Security Agency. Operating Procedure Codes Supplement (OPCS). Surgical Site Infection Surveillance Service (SSISS). 2019.
185. NHS inform. Common blood tests. 2023. https://www.nhsinform.scot/tests-and-treatments/blood-tests/common-blood-tests/ (accessed 31 May 2024).
186. Office for Health Improvement & Disparities. NHS Acute (Hospital) Trust Catchment Populations. 2022. https://app.powerbi.com/view?r=eyJrIjoiODZmNGQ0YzItZDAwZi00MzFiLWE4NzAtMzVmNTUwMThmMTVlIiwidCI6ImVlNGUxNDk5LTRhMzUtNGIyZS1hZDQ3LTVmM2NmOWRlODY2NiIsImMiOjh9 (accessed 31 Jan 2024).
187. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011; **30**(4): 377-99.

188. Office for National Statistics. Population and household estimates, England and Wales: Census 2021, unrounded data. 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/populationandhouseholdestimatesenglandandwales/census2021unroundeddata (accessed 5 June 2024).
189. GOV.UK. Regional ethnic diversity. 2022. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/regional-ethnic-diversity/latest/#download-the-data (accessed 5 June 2024).
190. Ministry of Housing, Communities & Local Government,. The English Indices of Deprivation 2019 (IoD2019) 2019.
191. Warren LR, Clarke J, Arora S, Darzi A. Improving data sharing between acute hospitals in England: an overview of health record system distribution and retrospective observational analysis of inter-hospital transitions of care. *BMJ open* 2019; **9**(12): e031637.
192. NHS Oxfordshire University Hosptials. ABOUT US. 2024. https://www.ouh.nhs.uk/about/ (accessed 5 June 2024).
193. National Statistics. Hospital Admitted Patient Care Activity 2017.
194. Office for National Statistics. Inequalities in Accident and Emergency department attendance, England: March 2021 to March 2022. 2023. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/articles/inequalitiesinaccidentandemergencydepartmentattendanceengland/march2021tomarch2022 (accessed 5 June 2024).
195. Wise J. Poor GP access may be driving people in deprived areas in England to use emergency departments, analysis suggests. *BMJ* 2023; **383**: 2323.
196. NHS. NHS Health Check. 2023. https://www.nhs.uk/conditions/nhs-health-check/ (accessed 5 June 2024).
197. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol* 2018; **47**(1): 226-35.
198. Menachemi N, Dixon BE, Wools-Kaloustian KK, Yiannoutsos CT, Halverson PK. How Many SARS-CoV-2-Infected People Require Hospitalization? Using Random Sample Testing to Better Inform Preparedness Efforts. *J Public Health Manag Pract* 2021; **27**(3): 246-50.
199. Landry A, Docherty P, Ouellette S, Cartier LJ. Causes and outcomes of markedly elevated C-reactive protein levels. *Can Fam Physician* 2017; **63**(6): e316-e23.
200. Luo Y. Evaluating the state of the art in missing data imputation for clinical data. *Brief Bioinform* 2022; **23**(1).
201. NHS Digital. Hospital Outpatient Activity 2020-21. 2021. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/2020-21/summary-report---gender (accessed 5 June 2024).
202. Irizar P, Pan D, Kapadia D, et al. Ethnic inequalities in COVID-19 infection, hospitalisation, intensive care admission, and death: a global systematic review and meta-analysis of over 200 million study participants. *EClinicalMedicine* 2023; **57**: 101877.
203. Centers for Disease Control and Prevention. Blood Culture Contamination: An Overview for Infection Control and Antibiotic Stewardship Programs Working with the Clinical Laboratory, 2022.
204. Doern GV, Carroll KC, Diekema DJ, et al. Practical Guidance for Clinical Microbiology Laboratories: A Comprehensive Update on the Problem of Blood Culture Contamination and a Discussion of Methods for Addressing the Problem. *Clin Microbiol Rev* 2019; **33**(1).
205. Buetti N, Atkinson A, Marschall J, Kronenberg A. Incidence of bloodstream infections: a nationwide surveillance of acute care hospitals in Switzerland 2008-2014. *BMJ Open* 2017; **7**(3): e013665.
206. UK Health Security Agency. 30 day all-cause mortality following MRSA, MSSA and Gram-negative bacteraemia and C. difficile infections: 2022 to 2023 report. 2023.

https://www.gov.uk/government/statistics/mrsa-mssa-and-e-coli-bacteraemia-and-c-difficile-infection-30-day-all-cause-fatality/30-day-all-cause-mortality-following-mrsa-mssa-and-gram-negative-bacteraemia-and-c-difficile-infections-2022-to-2023-report (accessed 2 February 2024).
207. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 2022; **399**(10325): 629-55.
208. Feldman SF, Temkin E, Wullfhart L, et al. A nationwide population-based study of Escherichia coli bloodstream infections: incidence, antimicrobial resistance and mortality. *Clin Microbiol Infect* 2022; **28**(6): 879.e1-.e7.
209. Schwaber MJ, Carmeli Y. Mortality and delay in effective therapy associated with extended-spectrum beta-lactamase production in Enterobacteriaceae bacteraemia: a systematic review and meta-analysis. *J Antimicrob Chemother* 2007; **60**(5): 913-20.
210. de Lastours V, Laouénan C, Royer G, et al. Mortality in Escherichia coli bloodstream infections: antibiotic resistance still does not make it. *Journal of Antimicrobial Chemotherapy* 2020; **75**(8): 2334-43.
211. Handal N, Whitworth J, Lyngbakken MN, Berdal JE, Dalgard O, Bakken Jørgensen S. Mortality and length of hospital stay after bloodstream infections caused by ESBL-producing compared to non-ESBL-producing E. coli. *Infect Dis (Lond)* 2024; **56**(1): 19-31.
212. HM Government. Tackling antimicrobial resistance 2019–2024. The UK's five-year national action plan., 2019.
213. HM Government. Contained and controlled. The UK's 20-year vision for antimicrobial resistance., 2019.
214. HM Government. Confronting antimicrobial resistance 2024 to 2029, 2024.
215. Velazquez-Meza ME, Galarde-López M, Carrillo-Quiróz B, Alpuche-Aranda CM. Antimicrobial resistance: One Health approach. *Vet World* 2022; **15**(3): 743-9.
216. UK Health Security Agency. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR). Report 2021 to 2022., 2022.
217. Laupland KB, Gregson DB, Church DL, Ross T, Pitout JD. Incidence, risk factors and outcomes of Escherichia coli bloodstream infections in a large Canadian region. *Clin Microbiol Infect* 2008; **14**(11): 1041-7.
218. Jackson LA, Benson P, Neuzil KM, Grandjean M, Marino JL. Burden of community-onset Escherichia coli bacteremia in seniors. *J Infect Dis* 2005; **191**(9): 1523-9.
219. Song J, Walters A, Berridge D, Akbari A, Evans M, Lyons RA. Risk factors for Escherichia coli bacteraemia: a population-based case-control study. *The Lancet* 2017; **390**: S85.
220. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. *J Am Med Inform Assoc* 2017; **24**(6): 1142-8.
221. Raftery A. Sociological methodology, chapter Bayesian Model Selection in Social Research Cambridge, MA: Wiley-Blackwell; 1998.
222. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 1989; **129**(1): 125-37.
223. McFadden D. Conditional Logit Analysis of Qualitative Choice Behavior: Institute of Urban and Regional Development, University of California; 1973.
224. European Committee on Antimicrobial Susceptibility Testing (EUCAST). Clinical breakpoints - breakpoints and guidance. 2024. https://www.eucast.org/clinical_breakpoints (accessed 16 April 2024).
225. Rodríguez-Baño J, Picón E, Gijón P, et al. Community-onset bacteremia due to extended-spectrum beta-lactamase-producing Escherichia coli: risk factors and prognosis. *Clin Infect Dis* 2010; **50**(1): 40-8.
226. Trautner BW, Darouiche RO. Catheter-associated infections: pathogenesis affects prevention. *Arch Intern Med* 2004; **164**(8): 842-50.

227. Vento S, Cainelli F. Infections in patients with cancer undergoing chemotherapy: aetiology, prevention, and treatment. *Lancet Oncol* 2003; **4**(10): 595-604.
228. Bai C, Zhang X, Yang D, Li D, Feng H, Li Y. Clinical Analysis of Bloodstream Infection of Escherichia coli in Patients with Pancreatic Cancer from 2011 to 2019. *Can J Infect Dis Med Microbiol* 2022; **2022**: 1338188.
229. Zhang Q, Gao HY, Li D, et al. Clinical outcome of Escherichia coli bloodstream infection in cancer patients with/without biofilm formation: a single-center retrospective study. *Infect Drug Resist* 2019; **12**: 359-71.
230. Stoeckle M, Kaech C, Trampuz A, Zimmerli W. The role of diabetes mellitus in patients with bloodstream infections. *Swiss Med Wkly* 2008; **138**(35-36): 512-9.
231. Bonten M, Johnson JR, van den Biggelaar AH, et al. Epidemiology of Escherichia coli bacteremia: a systematic literature review. *Clinical Infectious Diseases* 2021; **72**(7): 1211-9.
232. Turcato G, Zaboli A, Kostic I, et al. Severity of SARS-CoV-2 infection and albumin levels recorded at the first emergency department evaluation: a multicentre retrospective observational study. *Emerg Med J* 2022; **39**(1): 63-9.
233. Piano S, Brocca A, Mareso S, Angeli P. Infections complicating cirrhosis. *Liver Int* 2018; **38 Suppl 1**: 126-33.
234. Gradel KO, Vinholt PJ, Magnussen B, et al. Hypoalbuminaemia as a marker of trans-capillary leakage in community-acquired bacteraemia patients. *Epidemiol Infect* 2018; **146**(5): 648-55.
235. Okamoto A, Kanda Y, Kimura SI, Oyake T, Tamura K. Predictive and risk factor analysis for bloodstream infection in high-risk hematological patients with febrile neutropenia: post-hoc analysis from a prospective, large-scale clinical study. *Int J Hematol* 2021; **114**(4): 472-82.
236. Graham S, Walker JL, Andrews N, Nitsch D, Parker PKE, McDonald HI. Identifying markers of health-seeking behaviour and healthcare access in UK electronic health records. *medRxiv* 2023: 2023.11.08.23298256.
237. Bottle A, Honeyford K, Chowdhury F, Bell D, Aylin P. Factors associated with hospital emergency readmission and mortality rates in patients with heart failure or chronic obstructive pulmonary disease: a national observational study. *Health Services and Delivery Research* 2018.
238. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? *Stat Med* 2012; **31**(11-12): 1089-97.
239. Wilson J, Guy R, Elgohari S, et al. Trends in sources of meticillin-resistant Staphylococcus aureus (MRSA) bacteraemia: data from the national mandatory surveillance of MRSA bacteraemia in England, 2006-2009. *J Hosp Infect* 2011; **79**(3): 211-7.
240. Camilleri S, Tsai D, Langham F, Ullah S, Chiong F. Epidemiology, clinical outcomes and risk factors of third-generation cephalosporin-resistant Escherichia coli hospitalized infections in remote Australia-a case-control study. *JAC Antimicrob Resist* 2023; **5**(6): dlad138.
241. Richelsen R, Smit J, Laxsen Anru P, Schønheyder HC, Nielsen H. Risk factors of community-onset extended-spectrum β-lactamase Escherichia coli and Klebsiella pneumoniae bacteraemia: an 11-year population-based case-control-control study in Denmark. *Clin Microbiol Infect* 2021; **27**(6): 871-7.
242. Chopra T, Marchaim D, Johnson PC, et al. Risk factors for bloodstream infection caused by extended-spectrum β-lactamase-producing Escherichia coli and Klebsiella pneumoniae: A focus on antimicrobials including cefepime. *Am J Infect Control* 2015; **43**(7): 719-23.
243. Fan J, Yu C, Guo Y, et al. Frailty index and all-cause and cause-specific mortality in Chinese adults: a prospective cohort study. *Lancet Public Health* 2020; **5**(12): e650-e60.
244. Clegg A, Bates C, Young J, et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age Ageing* 2016; **45**(3): 353-60.
245. Luo J, Liao X, Zou C, et al. Identifying Frail Patients by Using Electronic Health Records in Primary Care: Current Status and Future Directions. *Front Public Health* 2022; **10**: 901068.

246. Al-Hasan MN, Lahr BD, Eckel-Passow JE, Baddour LM. Seasonal variation in Escherichia coli bloodstream infection: a population-based study. *Clin Microbiol Infect* 2009; **15**(10): 947-50.
247. Feldman SF, Temkin E, Wulffhart L, et al. Effect of temperature on Escherichia coli bloodstream infection in a nationwide population-based study of incidence and resistance. *Antimicrob Resist Infect Control* 2022; **11**(1): 144.
248. School of Geography and the Environment. Monthly, Annual and Seasonal Data from the Radcliffe Observatory site in Oxford. 2024. https://www.geog.ox.ac.uk/research/climate/rms/monthly-annual.html (accessed 24 April 2024).
249. Deeny SR, van Kleef E, Bou-Antoun S, Hope RJ, Robotham JV. Seasonal changes in the incidence of Escherichia coli bloodstream infection: variation with region and place of onset. *Clin Microbiol Infect* 2015; **21**(10): 924-9.
250. NHS England. Virtual Ward. 2024. https://www.england.nhs.uk/statistics/statistical-work-areas/virtual-ward/ (accessed 24 April 2024).
251. Nordvig J, Aagaard T, Daugaard G, et al. Febrile Neutropenia and Long-term Risk of Infection Among Patients Treated With Chemotherapy for Malignant Diseases. *Open Forum Infect Dis* 2018; **5**(10): ofy255.
252. Medical Billing and Coding. ICD-10-CM. 2023. https://www.medicalbillingandcoding.org/icd-10-cm/ (accessed 10 May 2024).
253. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol* 1993; **138**(11): 923-36.
254. Dunkler D, Plischke M, Leffondré K, Heinze G. Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. *PLoS One* 2014; **9**(11): e113677.
255. Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables: John Wiley & Sons; 2008.
256. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *Journal of clinical epidemiology* 1995; **48**(12): 1503-10.
257. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology* 1996; **49**(12): 1373-9.
258. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *American journal of epidemiology* 2007; **165**(6): 710-8.
259. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ* 2016; **352**: i1981.
260. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004; **66**(3): 411-21.
261. Moxley TA, Johnson-Leung J, Seamon E, Williams C, Ridenhour BJ. Application of Elastic Net Regression for Modeling COVID-19 Sociodemographic Risk Factors. *medRxiv* 2023.
262. Cooper ME, Risk B, Corey A, Fountain AJ, Allen JW. Statistical learning of blunt cerebrovascular injury risk factors using the elastic net. *Emergency radiology* 2021; **28**(5): 929-37.
263. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2005; **67**(2): 301-20.
264. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 2010; **20**(1): 101.
265. Tang F, Ishwaran H. Random Forest Missing Data Algorithms. *Stat Anal Data Min* 2017; **10**(6): 363-77.

266. Ethun CG, Bilen MA, Jani AB, Maithel SK, Ogan K, Master VA. Frailty and cancer: implications for oncology surgery, medical oncology, and radiation oncology. *CA: a cancer journal for clinicians* 2017; **67**(5): 362-77.
267. Pereira Gray D, Henley W, Chenore T, Sidaway-Lee K, Evans P. What is the relationship between age and deprivation in influencing emergency hospital admissions? A model using data from a defined, comprehensive, all-age cohort in East Devon, UK. *BMJ Open* 2017; **7**(2): e014045.
268. Royston P, Sauerbrei W. Bootstrap assessment of the stability of multivariable models. *The Stata Journal* 2009; **9**(4): 547-70.
269. Pearce N. Analysis of matched case-control studies. *BMJ* 2016; **352**.
270. Gounden V, Vashisht R, Jialal I. Hypoalbuminemia. StatPearls. Treasure Island (FL): StatPearls Publishing Copyright © 2024, StatPearls Publishing LLC.; 2024.
271. Zhang Z, Pereira SL, Luo M, Matheson EM. Evaluation of Blood Biomarkers Associated with Risk of Malnutrition in Older Adults: A Systematic Review and Meta-Analysis. *Nutrients* 2017; **9**(8).
272. Gabay C, Kushner I. Acute-phase proteins and other systemic responses to inflammation. *N Engl J Med* 1999; **340**(6): 448-54.
273. MedlinePlus. Alkaline Phosphatase. 2022. https://medlineplus.gov/lab-tests/alkaline-phosphatase/ (accessed 26 June 2024).
274. Meddings J, Rogers MA, Krein SL, Fakih MG, Olmsted RN, Saint S. Reducing unnecessary urinary catheter use and other strategies to prevent catheter-associated urinary tract infection: an integrative review. *BMJ Qual Saf* 2014; **23**(4): 277-89.
275. Blijlevens NMA, de Mooij CEM. Mucositis and Infection in Hematology Patients. *Int J Mol Sci* 2023; **24**(11).
276. Weischendorff S, Rathe M, Petersen MJ, et al. Markers of intestinal mucositis to predict blood stream infections at the onset of fever during treatment for childhood acute leukemia. *Leukemia* 2024; **38**(1): 14-20.
277. Dahlgren D, Sjöblom M, Hellström PM, Lennernäs H. Chemotherapeutics-Induced Intestinal Mucositis: Pathophysiology and Potential Treatment Strategies. *Front Pharmacol* 2021; **12**: 681417.
278. Drayson MT, Bowcock S, Planche T, et al. Levofloxacin prophylaxis in patients with newly diagnosed myeloma (TEAMM): a multicentre, double-blind, placebo-controlled, randomised, phase 3 trial. *Lancet Oncol* 2019; **20**(12): 1760-72.
279. GOV.UK. UK pandemic preparedness. 2020. https://www.gov.uk/government/publications/uk-pandemic-preparedness/uk-pandemic-preparedness (accessed 20 June 2024).
280. Fryer HR, Arning N, Wilson DJ. Doublethink: simultaneous Bayesian-frequentist model-averaged hypothesis testing. *arXiv* 2023.
281. Arning N, Fryer HR, Wilson DJ. Identifying direct risk factors in UK Biobank with simultaneous Bayesian-frequentist model-averaged hypothesis testing using Doublethink. *medRxiv* 2024.
282. Alhababi N, Magnus MC, Drake MJ, Fraser A, Joinson C. The Association Between Constipation and Lower Urinary Tract Symptoms in Parous Middle-Aged Women: A Prospective Cohort Study. *J Womens Health (Larchmt)* 2021; **30**(8): 1171-81.
283. Mansournia MA, Altman DG. Population attributable fraction. *BMJ* 2018; **360**: k757.
284. Sieben A, van Onzenoort HAW, van Laarhoven KJHM, Bredie SJH, van Dulmen S. Identification of Cardiovascular Patient Groups at Risk for Poor Medication Adherence: A Cluster Analysis. *J Cardiovasc Nurs* 2021; **36**(5): 489-97.
285. Andrews A, Hope R, Muller-Pebody B, Hopkins S, Gerver S, Walker AS. Escherichia coli bacteraemia reductions during the COVID-19 pandemic in England, 2020-21. Federation of Infection Societies conference; 2023. p. 198.

# Appendix A: Definitions of Risk Factors from Electronic Health Records

A full Excel spreadsheet is available at: https://github.com/EmmaPritchard.

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| **Diagnosis codes** | diag_charlsondd | Charlson score calculated using secondary diagnosis codes from the diagnostic dominant episode of an inpatient spell. | ICD-10 codes and weightings based on the Dr Foster (2012) definition available on page 31 of the provided link. Use only secondary diagnostic codes from the diagnosis dominant episode of the current spell. Diagnosis dominant episode = first consultant episode in hospital spell, or the second if the first episode's primary diagnosis code is an "R" code and the second episode's diagnosis is not. | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_charlsonlb | Charlson score calculated using one-year lookback for primary and secondary diagnosis codes. | ICD-10 codes and weightings based on the Dr Foster (2012) definition available on page 31 of the provided link. Use primary and secondary diagnosis codes from the previous year and secondary diagnostic codes from the diagnosis-dominant episode of the current spell. | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_myocardinfarc | Acute myocardial infarction | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: I21*, I22*, I23*, I25.2*, I25.8* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_cerebrovasc | Cerebral vascular accident | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: G45.0*, G45.1*, G45.2*, G45.4*, G45.8*, G45.9*, G46*, I60*-I69* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_heartfail | Congestive heart failure | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: I50* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_tissuedis | Connective tissue disorder | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: M05*, M06.0*, M06.3*, M06.9*, M32*, M33.2*, M34*, M35.3* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | diag_dementia | Dementia | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: F00*, F01*, F02*, F03*, F05.1* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_liverdis | Liver disease | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: K70.2*, K70.3*, K71.7*, K73*, K74* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_pepticulcer | Peptic ulcer | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: K25*, K26*, K27*, K28* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_vascdis | Peripheral vascular disease | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: I71*, I73.9*, I79.0*, R02*, Z95.8*, Z95.9* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_pulmondis | Pulmonary disease | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: J40*-J47*, J60*-J76* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_cancer | Cancer | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: C00*-C76*, C80*-C97* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_paraplegia | Paraplegia | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: G04.1*, G81*, G82.0*, G82.1*, G82.2* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_renaldis | Renal disease | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: I12*, I13*, N01*, N03*, N05.2*-N05.6*, N07.2*-N07.4*, N18*, N19*, N25* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | diag_metcancer | Metastatic cancer | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: C77*, C78*, C79* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_sevliverdis | Severe liver disease | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: K72.1*, K72.9*, K76.6*, K76.7* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_hiv | HIV | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: B20*, B21*, B22*, B23*, B24*, O987* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_diabetes | Diabetes | Any of the following ICD-10 codes in primary or secondary diagnoses as defined in the Charlson score: E10*, E11*, E13*, E14* | Yes | https://beyondclinical.files.wordpress.com/2014/01/hsmr-toolkit-version-6-october-20111.pdf |
| | diag_renalfail | Renal failure | The following ICD-10 code in either primary or secondary diagnosis: N18* | No | |
| | diag_renaldialy | Renal dialysis | The following ICD-10 codes in either primary or secondary diagnoses: Z49*, Z99.2*, N18.6*, T82.4* | Yes | https://pubmed.ncbi.nlm.nih.gov/35723965/ |
| | diag_palliative | Palliative | The following ICD-10 codes in either primary or secondary diagnoses: Z51.5 | Yes | https://pubmed.ncbi.nlm.nih.gov/35723965/ |
| | diag_immunosuppress | Immunosuppressed | The following ICD-10 codes in either primary or secondary diagnoses: B20*-B24* (AIDS/HIV), 0987*, C77*-C96* (metastatic cancer, haematological malignancies), D80*-D84* (primary immunodeficiencies), K72.1*, K72.9*, K76.6*, K76.7* (end-stage liver disease). | Yes | https://pubmed.ncbi.nlm.nih.gov/35723965/ |
| | diag_invasdis | Invasive disease codes | Infection-related invasive disease ICD-10 codes in primary or secondary diagnoses defined previously (available on GitHub). | No | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | diag_urinedis | Urinary disease codes | Infection-related urinary disease ICD-10 codes in primary or secondary diagnoses defined previously (available on GitHub). | No | |
| | diag_surgsite | Surgical site infection | Infection-related surgical site ICD-10 codes in primary or secondary diagnoses defined previously (available on GitHub). | No | |
| | diag_covid19hosp | COVID-19 hospitalisation | Presence of the following ICD-10 codes in either primary or secondary diagnoses on or after January 2020 (based on episodestartdate): U07.1 (COVID-19 infection confirmed by laboratory testing); U07.2 (COVID-19 infection diagnosis clinically or epidemiologically where laboratory confirmation is negative, inconclusive, or not available/performed); B34.2 (coronavirus infection, site unspecified); B97.2 (the term “coronavirus” without any further specification). | No | |
| | diag_surgwoundrisk | Surgical wound risk | ICD-10 codes indicating a risk of surgical wound in primary or secondary diagnoses as defined in Table S3 of the provided link | Yes | https://pubmed.ncbi.nlm.nih.gov/30403746/ |
| | diag_pneumon | Pneumonia | The following ICD-10 codes in the primary or secondary diagnosis: A31.0, A42.0, A48.1, B01.2, B05.2, B25.0, B59, J10.0, J11.0, J12.8, J12.9, J13, J14, J15.0, J15.1, J15.2, J15.3, J15.4, J15.5, J15.6, J15.7, J15.8, J15.9, J16.8, J17.2, J18.0, J18.1, J18.2, J18.8, J18.9, J69.0, J85.0, J85.1 | Yes | https://pubmed.ncbi.nlm.nih.gov/30403746/ |
| | diag_allergy | Allergy | The following ICD-10 codes in primary or secondary diagnosis: T78, L50, J98.01, R06.1, I95.89, L27.2, T63.4, T88.6, T80.5, T88.2, L27.0, T65.811, T59.1, T80.5, T88.1, J01.70, J12.00 | Yes | https://pubmed.ncbi.nlm.nih.gov/31785369/ |
| | diag_chapter1 | Diagnostic code in Chapter 1 (Certain infectious and parasitic diseases) | Diagnostic code in range A00-B99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter2 | Diagnostic code in Chapter 2 (Neoplasms) | Diagnostic code in range C00-D48 | Yes | https://icd.who.int/browse10/2019/en |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | diag_chapter3 | Diagnostic code in Chapter 3 (Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism) | Diagnostic code in range D50-D89 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter4 | Diagnostic code in Chapter 4 (Endocrine, nutritional and metabolic diseases) | Diagnostic code in range E00-E90 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter5 | Diagnostic code in Chapter 5 (Mental and behavioural disorders) | Diagnostic code in range F00-F99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter6 | Diagnostic code in Chapter 6 (Diseases of the nervous system) | Diagnostic code in range G00-G99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter7 | Diagnostic code in Chapter 7 (Diseases of the eye and adnexa) | Diagnostic code in range H00-H59 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter8 | Diagnostic code in Chapter 8 (Diseases of the ear and mastoid process) | Diagnostic code in range H60-H95 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter9 | Diagnostic code in Chapter 9 (Diseases of the circulatory system) | Diagnostic code in range I00-I99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter10 | Diagnostic code in Chapter 10 (Diseases of the respiratory system) | Diagnostic code in range J00-J99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter11 | Diagnostic code in Chapter 11 (Diseases | Diagnostic code in range K00-K93 | Yes | https://icd.who.int/browse10/2019/en |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | | of the digestive system) | | | |
| | diag_chapter12 | Diagnostic code in Chapter 12 (Diseases of the skin and subcutaneous tissue) | Diagnostic code in range L00-L99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter13 | Diagnostic code in Chapter 13 (Diseases of the musculoskeletal system and connective tissue) | Diagnostic code in range M00-M99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter14 | Diagnostic code in Chapter 14 (Diseases of the genitourinary system) | Diagnostic code in range N00-N99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter15 | Diagnostic code in Chapter 15 (Pregnancy, childbirth and the puerperium) | Diagnostic code in range O00-O99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter16 | Diagnostic code in Chapter 16 (Certain conditions originating in the perinatal period) | Diagnostic code in range P00-P96 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter17 | Diagnostic code in Chapter 17 (Congenital malformations, deformations and chromosomal abnormalities) | Diagnostic code in range Q00-Q99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter18 | Diagnostic code in Chapter 18 (Symptoms, signs and abnormal clinical and | Diagnostic code in range R00-R99 | Yes | https://icd.who.int/browse10/2019/en |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | | laboratory findings, not elsewhere classified) | | | |
| | diag_chapter19 | Diagnostic code in Chapter 19 (Injury, poisoning and certain other consequences of external causes) | Diagnostic code in range S00-T98 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter20 | Diagnostic code in Chapter 20 (External causes of morbidity and mortality) | Diagnostic code in range V01-Y98 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter21 | Diagnostic code in Chapter 21 (Factors influencing health status and contact with health services) | Diagnostic code in range Z00-Z99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_chapter22 | Diagnostic code in Chapter 22 (Codes for special purposes) | Diagnostic code in range U00-U99 | Yes | https://icd.who.int/browse10/2019/en |
| | diag_frailtygil | Frailty score as defined by Gilbert et al. | Frailty score as defined by Gilbert et al. (2018). A two-year lookback was used as recommended in the paper. | Yes | https://pubmed.ncbi.nlm.nih.gov/29706364/ |
| **Procedure codes** | proc_chapter* | Inpatient and outpatient procedure chapters | 24 binary variables derived from the 24 chapters procedure codes are split into based on the first character of the code e.g. E49.2 = chapter E. There is no chapter "I" or "O". | | |
| | proc_urinecath | Urethral catheter | Any occurrence of "urethral catheter" in proclabel | No | |
| | proc_myocardscan | Myocardial perfusion scan | Any occurrence of "myocardial perfusion scan" in proclabel | No | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | proc_implant | Implant | Any occurrence of “implant” in procedure code label | No | |
| | proc_chemo | Chemotherapy | Any occurrence of “chemotherapy” in procedure code label OR any proccode in the chemotherapy regimens NHS Digital Classifications Browser: X70.1, X70.2, X70.3, X70.4, X70.5, X71.1, X71.2, X71.3, X71.4, X71.5, X71.6, X72.1, X72.2, X72.3, X73.1 | Yes | https://classbrowser.nhs.uk/#/ |
| | proc_echocardio | Transthoracic echocardiography | Any occurrence of “transthoracic echocardiography” in proclabel | No | |
| | proc_dialysis | Dialysis | Any procedure code from the dialysis code list available in the link provided | Yes | https://www.opencodelists.org/codelist/ukrr/dialysis-opcs-4/505bef04/ |
| | proc_transplant | Transplant | Any procedure code from the transplant code list in the link provided | Yes | https://www.opencodelists.org/codelist/nhsd/transplant-spl-hes-opcs4/1b47de52/ |
| | proc_spirometry | Spirometry | Any occurrence of "spirometry" in procedure code label | No | |
| | proc_extractbm | Diagnostic extraction of bone marrow | Any occurrence of "diagnostic extraction of bone marrow" in procedure code label | No | |
| | proc_spinalpuncture | Spinal puncture procedure | Any occurrence of “spinal puncture” in procedure code label | No | |
| | proc_invasventil | Invasive ventilation | procedure code label == "invasive ventilation" OR procedure code label == "bag valve mask ventilation" | No | |
| | proc_noninvasventil | Non-invasive ventilation | procedure code label == "non-invasive ventilation" | No | |
| | proc_endouppergi | Endoscopy of upper GI tract | (procedure code label == "endoscopic") AND (procedure code label == "oesophagus" OR "liver" OR "upper gastrointestinal" OR "bile duct" OR "pancreatic duct" OR "gastromy" OR "sphincter of oddi" OR "duodenum" OR "jejunum" OR "pancreas") | No | |
| | proc_endolowergi | Endoscopy of lower GI tract | (procedure code label == "endoscopic") AND (procedure code label == "ileum" OR "ileoanal" OR "cecum" OR "colon" OR "rectum" OR "anus" OR "bowel" | No | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | proc_endourine | Endoscopy of urethra | (procedure code label == "endoscopic") AND (procedure code label == "bladder" OR "urethra" OR "kidney" OR "ureter") | No | |
| | proc_endocolpos | Colposcopy | (proclabel == "diagnostic endoscopic examination or uterus") OR (proclabel == "colposcopy") | No | |
| | proc_cleansurg | Clean surgical procedure | Any procedure code marked as "clean" in the UKHSA groupings in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_cleancontamsurg | Clean-contaminated surgical procedure | Any procedure code marked as "clean-contaminated" in the UKHSA groupings in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_contamsurg | Contaminated surgical procedure | Any procedure code marked as "contaminated" in the UKHSA groupings in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_abdomhysterectomy | Abdominal hysterectomy surgery | Any procedure code in the "Abdominal Hysterectomy" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_bileductliverpancreat | Bile duct, liver, or pancreatic surgery | Any procedure code in the "Bile duct, liver, or pancreatic surgery" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/ |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | | | | | 1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_breastsurgery | Breast surgery | Any procedure code in the "Breast surgery" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_cardiacsurgery | Cardiac surgery | Any procedure code in the "Cardiac surgery (non-CABG)" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_cholecystectomy | Cholecystectomy (non-laparoscopic) | Any procedure code in the "Cholecystectomy (non-laparoscopic)" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_coronaryarterybypass | Coronary artery | Any procedure code in the "Coronary artery bypass graft" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_cranialsurgery | Cranial surgery | Any procedure code in the "Cranial surgery" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | proc_gastricsurgery | Gastric surgery | Any procedure code in the "Gastric surgery" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_hipreplacement | Hip replacement | Any procedure code in the "Hip replacement" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_kneereplacement | Knee replacement | Any procedure code in the "Knee replacement" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_largebowelsurgery | Large bowel surgery | Any procedure code in the "Large bowel surgery" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_limbamputation | Limb amputation | Any procedure code in the "Limb amputation" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_reductionbonefracture | Reduction of long bone fracture | Any procedure code in the "Reduction of long bone fracture" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploa |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | | | | | ds/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_repairneckfemur | Repair of neck of femur | Any procedure code in the "Repair of neck of femur" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_smallbowelsurgery | Small bowel surgery | Any procedure code in the "Small bowel surgery" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_spinalsurgery | Spinal surgery | Any procedure code in the "Spinal surgery" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_vascularsurgery | Vascular surgery | Any procedure code in the "vascular surgery" category in the link provided and in the tab "UKHSA procedure codes" | Yes | https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1069633/OPCS_codes_supplement_UKHSA.pdf |
| | proc_headctradio | Evidence of fall | procedure code label == “computed tomography of head” AND procedure code label == “radiology” at the same date and time | No | |
| | proc_scanabdomen | Any scan of abdomen | (procedure code label == “computed tomography” OR “radiology”) AND (procedure code label ==“abdomen”) at the same date and time | No | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | proc_scanpelvis | Any scan of pelvis | (procedure code label== “computed tomography” OR “radiology”) AND (procedure code label ==“pelvis”) at the same date and time | No | |
| | proc_scanchest | Any scan of chest | (procedure code label == “computed tomography” OR “radiology”) AND (procedure code label ==“chest”) at the same date time | No | |
| | proc_scanother | Any scan of other body part (NOT abdomen, pelvis or chest) | (procedure code label == “computed tomography” OR “radiology”) AND NOT (procedure code label == “abdomen” OR “pelvis” OR “chest”) at the same date and time | No | |
| | proc_prosthesis | Prosthetic/prosthesis | (procedure code label == “prosthetic” OR “prosthesis”) AND NOT (procedure code label == “prosthetic replacement for lens”) | No | |
| | proc_cataract | Cataract surgery | (procedure code label == “insertion of prosthetic replacement for lens”) AND (procedure code label == “Phacoemulsification of lens”) | No | |
| | proc_transrectal | Transrectal prostate biopsy or ultrasound | (proccode == "M70.3" AND proccode == "Y53.2") OR (proccode == "U21.6" AND proccode == "Z42.2") | No | |
| **inpatient admissions** | inpat_admiss | Ordinary inpatient admission | Any inpatient admission with a patient classification code == "1" (Ordinary admission ) | N/A | |
| | inpat_admiss8hours | Ordinary inpatient admission longer than 8 hours | Any inpatient admission with a patient classification code == "1" (Ordinary admission ) and the hours between admission date and discharge date > 8 | N/A | |
| | inpat_carehomeadmiss | Admitted from or to a care home | Any inpatient admission with admissionsource == 54, 85, 88 or dischargedestination == 54, 85, 88 | N/A | |
| | inpat_complexadmiss | Complex inpatient admission | Any inpatient admission with at least 2 episodes in the spell | N/A | |
| | inpat_consultgeri | Inpatient admission under a geriatric consultant | Any inpatient admission with consultantmainspecialtycode == 430 | N/A | |
| | inpat_electadmiss | Elective inpatient admission | Any inpatient admission with admissionmethodcode == 11, 12, 13 | N/A | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | inpat_emergadmiss | Emergency inpatient admission | Any inpatient admission with admissionmethodcode == 21, 22, 23, 24, 25, 2A, 2B, 2C, 2D, 28 | N/A | |
| | inpat_treatcardio | Cardiology treatment function code | Any inpatient admission with treatmentfunctioncode == 320 | N/A | |
| | inpat_treatgastroent | Gastroenterology treatment function code | Any inpatient admission with treatmentfunctioncode == 301 | N/A | |
| | inpat_treatgenmed | General medicine treatment function code | Any inpatient admission with treatmentfunctioncode == 300 | N/A | |
| | inpat_treatgensurg | General surgery treatment function code | Any inpatient admission with treatmentfunctioncode == 100 | N/A | |
| | inpat_treattraumasurg | Trauma surgery treatment function code | Any inpatient admission with treatmentfunctioncode == 110 | N/A | |
| | inpat_treaturology | Urology Service treatment function code | Any inpatient admission with treatmentfunctioncode == 101 | N/A | |
| | inpat_treatvascsurg | Vascular surgery treatment function code | Any inpatient admission with treatmentfunctioncode == 107 | N/A | |
| | inpat_admisslos365d | Cumulative length of stay in ordinary inpatient admission | Cumulative length of time in an ordinary inpatient admission in the previous 365 days | N/A | |
| | inpat_admissnum365d | Number of ordinary inpatient admissions | Number of ordinary inpatient admission in the previous 365 days | N/A | |
| | inpat_admiss8hourslos365d | Cumulative length of stay in ordinary inpatient admissions > 8 hours | Cumulative length of time in ordinary inpatient admissions >8 hours in the previous 365 days | N/A | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | inpat_admiss8hoursnum365d | Number of ordinary inpatient admissions >8 hours | Number of ordinary inpatient admissions >8 hours in the previous 365 days | N/A | |
| | inpat_complexadmisslos365d | Cumulative length of stay in ordinary complex inpatient admissions | Cumulative length of time in an ordinary complex inpatient admission in the previous 365 days | N/A | |
| | inpat_complexadmissnum365d | Number of ordinary complex inpatient admissions | Number of ordinary complex inpatient admission in the previous 365 days | N/A | |
| | inpat_electadmisslos365d | Cumulative length of stay in ordinary elective inpatient admissions | Cumulative length of time in an ordinary elective inpatient admission in the previous 365 days | N/A | |
| | inpat_electadmissnum365d | Number of ordinary elective inpatient admissions | Number of ordinary elective admission in the previous 365 days | N/A | |
| | inpat_emergadmisslos365d | Cumulative length of stay in ordinary emergency inpatient admissions | Cumulative length of time in an ordinary emergency inpatient admission in the previous 365 days | N/A | |
| | inpat_emergadmissnum365d | Number of ordinary emergency inpatient admissions | Number of ordinary emergency admission in the previous 365 days | N/A | |
| **outpatient attendances** | outpat_treatrenal | Renal medicine treatment function code | Any outpatient appointment with treatmentfunctioncode == 361 | N/A | |
| | outpat_treateye | Ophthalmology Service treatment function code | Any outpatient appointment with treatmentfunctioncode == 130 | N/A | |
| | outpat_treatphysio | Physiotherapy Service treatment function code | Any outpatient appointment with treatmentfunctioncode == 650 | N/A | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | outpat_treattraumasurg | Trauma surgery treatment function code | Any outpatient admission with treatmentfunctioncode == 110 | N/A | |
| | outpat_treatcardio | Cardiology treatment function code | Any outpatient admission with treatmentfunctioncode == 320 | N/A | |
| | outpat_treatdermat | Dermatology Service treatment function code | Any outpatient admission with treatmentfunctioncode == 330 | N/A | |
| | outpat_treatmedonc | Medical oncology treatment function code | Any outpatient admission with treatmentfunctioncode == 370 | N/A | |
| | outpat_treaturology | Urology Service treatment function code | Any outpatient admission with treatmentfunctioncode == 101 | N/A | |
| | outpat_treatgynae | Gynaecology Service treatment function code | Any outpatient admission with treatmentfunctioncode == 502 | N/A | |
| | outpat_treatneuro | Neurology Service treatment function code | Any outpatient admission with treatmentfunctioncode == 400 | N/A | |
| | outpat_treathaemo | Clinical Haematology Service treatment function code | Any outpatient admission with treatmentfunctioncode == 303 | N/A | |
| | outpat_treatgastroent | Gastroenterology Service treatment function code | Any outpatient admission with treatmentfunctioncode == 301 | N/A | |
| | outpat_treatdiagimage | Diagnostic Imaging Service treatment function code | Any outpatient admission with treatmentfunctioncode == 812 | N/A | |
| | outpat_treatmidwife | Midwifery Service treatment function code | Any outpatient admission with treatmentfunctioncode == 560 | N/A | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| **Inpatient admission or outpatient attendance** | admiss_endoscopy | Any endoscopy procedure | Any inpatient admission or outpatient attendance with a localsubspecialtycode of 10061, 10066, 10466, 30061, 30161, 30162, 30166 | N/A | |
| **A&E visits** | emerg_admiss | A&E attendance | Any recorded A&E visit | N/A | |
| **Microbiology** | micro_urinecath | Urinary catheterisation | specimencode == "CSU" OR "CATH" | N/A | |
| | micro_covidtest | SARS-CoV-2 test done | battestcode == "CV2P" OR battestname == "SARS CORONAVIRUS-2 PCR". Don't consider any tests taken before January 2020. | N/A | |
| | micro_covidpos | SARS-CoV-2 positive test result | (battestcode == "CV2P" OR battestname == "SARS CORONAVIRUS-2 PCR") AND (result == "DET"). Don't consider any tests taken before January 2020. If both "DET" and "NDET" are present at the same collection date and time for a person, take the positive result. | N/A | |
| | micro_urinecult | Urine culture taken | (battestcode = "UC", "UCG", "UCH", "URM", "UMI", "UMG", "UMH", "ANTE") OR (battestname including "URINE CULT" or "URINE MICROSCOPY") | N/A | |
| | micro_urineecoli | Urine positive for E. coli | micro_urinecult (as defined above) == 1 AND buggenus (as defined by Qingze in provided link) == "E COLI" | N/A | https://github.com/davideyre/ehr_tools/blob/main/bug_grouper.R |
| | micro_urineenterococ | Urine positive for enterococcus | micro_urinecult (as defined above) == 1 AND buggenus (as defined by Qingze in provided link) == "ENTERCOCCUS" | N/A | https://github.com/davideyre/ehr_tools/blob/main/bug_grouper.R |
| | micro_urineenterobact | Urine positive for other enterobacterales | micro_urinecult (as defined above) == 1 AND buggenus (as defined by Qingze in provided link) == "OTHER ENTEROBACTERALES" | N/A | https://github.com/davideyre/ehr_tools/blob/main/bug_grouper.R |
| | micro_urinekleb | Urine positive for klebsiella | micro_urinecult (as defined above) == 1 AND buggenus (as defined by Qingze in provided link) == "KLEBSIELLA" | N/A | https://github.com/davideyre/ehr_tools/blob/main/bug_grouper.R |
| | micro_urinestaph | Urine positive for any other pathogen | micro_urinecult (as defined above) == 1 AND buggenus (as defined by Qingze in provided link) == ("S AUREUS) | N/A | https://github.com/davideyre/ehr_tools/blob/main/bug_grouper.R |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | micro_urinecult_num | Number of urine cultures taken | Number of urine cultures (micro_urinecult == 1) requested in previous 30, 60, 180, and 365 days | N/A | |
| | micro_bloodcult | Blood culture taken | battestcode == “BLC” OR battestname == “BLOOD CULTURE” | N/A | |
| | micro_bloodstaph | Blood positive for S. Aureus | micro_bloodcult (as defined above) == 1 AND buggenus (as defined by Qingze in provided link) == "S AUREUS" | N/A | https://github.com/davideyre/ehr_tools/blob/main/bug_grouper.R |
| | micro_bloodcult_num | Number of blood cultures taken | Number of blood cultures (micro_bloodcult == 1) requested in previous 30, 60, 180, and 365 days | N/A | |
| | micro_bkvscreen | Screen for BK virus | battestname == "BKV VIRAL LOAD" | N/A | |
| | micro_cdiffcarr | Faeces positive for C. diff carriage | battestcode == "CDGD", "CDT" and result == "CDTV", "DET" | N/A | |
| | micro_cdifftox | Faeces positive for C. diff toxin | battestcode == "CDT" and result == "DET" | N/A | |
| | micro_cmvebvscreen | CMV or EBV screen | battestname == "CMV VIRAL LOAD", "EBV VIRAL LOAD" | N/A | |
| | micro_cnscult | CNS culture taken | battestcode == "CSFC" or battestname == "CSF CULT AND MICRO" | N/A | |
| | micro_faecescult | Faeces culture taken | battestname, "FAECES EXAMINATION", "FAECES MOLECULAR ASSAY", "MICROSCOPY OF FAECES" or battestcode == "FCU", "FPCR", "FMIC", "CDGD", "CDT" | N/A | |
| | micro_flursvpcr | FLU/RSV PCR | battestname == "FLU & RSV A/B PCR" or battestcode == "FRAB" | N/A | |
| | micro_mrsapos | Screen positive for MRSA | (battestname == "MRSA SCREEN" or battestcode == "MRS") and bugcode == "MRSA" | N/A | |
| | micro_mrsascreen | Screen for MRSA | battestname == "MRSA SCREEN" or battestcode == "MRS" | N/A | |
| | micro_otherscreen | Screen for viruses (not BK, CMV, or EBV) | battestname == "CPE SCREEN", "ESBL SCREEN" | N/A | |
| | micro_puscult | PUS culture taken | battestname == "PUS MICRO / CULTURE", "STERILE SITE CULT" or battestcode == "PMC", "STER" | N/A | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | micro_respcult | Respiratory culture taken | battestname == "RESP. CULT AND MICRO" or battestcode == "RESP" | N/A | |
| | micro_resppaerug | Respiratory sample positive for Pseudomonas Aeruginosa | (battestname == "RESP. CULT AND MICRO" or battestcode == "RESP") and (bugname == "PSEUDOMONAS AERUGINOSA" or bugcode == "PSAR") | N/A | |
| | micro_resppcr | Respiratory PCR sample taken | battestname == "RESPIRATORY PCR" or battestcode == "RPCR" | N/A | |
| | micro_surfswabcult | Surface swab culture taken | battestname == "SURFACE SWAB CULTURE" or battestcode == "SFC" | N/A | |
| **Blood tests** | lims_albumin | Albumin | testname == "ALBUMIN" | N/A | |
| | lims_alkphosphatase | Alkaline phosphatase | testname == "ALK.PHOSPHATASE" | N/A | |
| | lims_alt | Alanine transaminase (ALT) | testname == "ALT" | N/A | |
| | lims_bilirubin | Bilirubin | testname == "BILIRUBIN" | N/A | |
| | lims_creatinine | Creatinine | testname == "CREATININE" | N/A | |
| | lims_crp | C-reactive protein | testname == "CRP" | N/A | |
| | lims_eosinophils | Eosinophils | testname == "EOSINOPHILS", "M EOSINOPHILS" | N/A | |
| | lims_haemoglobin | Haemoglobin | testname == "HAEMOGLOBIN" | N/A | |
| | lims_hba1c | HbA1c | testname == "HBA1C (DCCT)", "HBA1C (IFCC)". Convert DCCT to mmol/mol units as follows: (value - 2.15)*10.929. | N/A | |
| | lims_lymphocytes | Lymphocytes | testname = "LYMPHOCYTES", "M LYMPHOCYTES" | N/A | |
| | lims_neutrophils | Neutrophils | testname == "NEUTROPHILS", "M NEUTROPHILS". Drop values out of range 0.02-150 X10*9/L or 0-100 %. | N/A | |
| | lims_platelets | Platelets | testname == "PLATELETS" | N/A | |
| | lims_potassium | Potassium | testname == "POTASSIUM" | N/A | |
| | lims_psa | Prostate-specific antigen (PSA) test | testname == "PROS SPEC AG" | N/A | |
| | lims_serumb12 | Serum B12 | testname == "SERUM B12" | N/A | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | lims_serumfolate | Serum folate | testname == "SERUM FOLATE" | N/A | |
| | lims_sodium | Sodium | testname == "SODIUM" | N/A | |
| | lims_transferrin | Transferrin | testname == "TRANSFERRIN" | N/A | |
| | lims_urea | Urea | testname == "UREA" | N/A | |
| | lims_whitecells | White cell count | testname == "WHITE CELLS". Drop values out of range 0.1-150 X10*9/L | N/A | |
| | limsflag_psa | Any PSA test requested | Binary variable for having had a previous PSA test in IORD in the past X years | N/A | |
| | limsflag_serumb12 | Any serum B12 test requested | Binary variable for having had a previous serum B12 test in IORD in the past X years | N/A | |
| | limsflag_serumfolate | Any serum folate test requested | Binary variable for having had a previous serum folate test in IORD in the past X years | N/A | |
| | limsflag_transferrin | Any transferrin test requested | Binary variable for having had a previous transferrin test in IORD in the past X years | N/A | |
| **Vital signs** | vital_dbp | Diastolic blood pressure | All results from eventname == “Diastolic Blood pressure”. Drop if below 10 mmHg or above 200 mmHg | N/A | |
| | vital_heartrate | Heart rate | All results from eventname == “Heart Rate”. Drop if below 20 bpm or above 350 bpm | N/A | |
| | vital_oxygensat | Oxygen saturation | All results from eventname == “Oxygen Saturation”. Drop if below 50% or above 100%. | N/A | |
| | vital_resprate | Respiratory rate | All results from eventname == “Respiratory Rate”. Drop if below 4 br/min or above 80 br/min | N/A | |
| | vital_sbp | Systolic blood pressure | All results from eventname == “Systolic Blood Pressure”. Drop if below 20 mmHg or above 300 mmHg. | N/A | |
| | vital_temperature | Temperature | All results from eventname == “Temperature Tympanic”. Drop if below 25C or above 45C. | N/A | |
| **Personal traits** | trait_weight | Weight (kg) | eventname == "weight_measured", "weight_working", "weight_estimate", "weight_unknown". | N/A | |

| Dataset | Variable name | Short variable description | Definition | Previously published flag | Previously published link |
|---|---|---|---|---|---|
| | trait_height | Height (cm) | eventname == "height" | N/A | |
| | trait_bmi | BMI (kg/m2) | Calculated from weight and height values as weight (kg)/(height (m)^2) | N/A | |