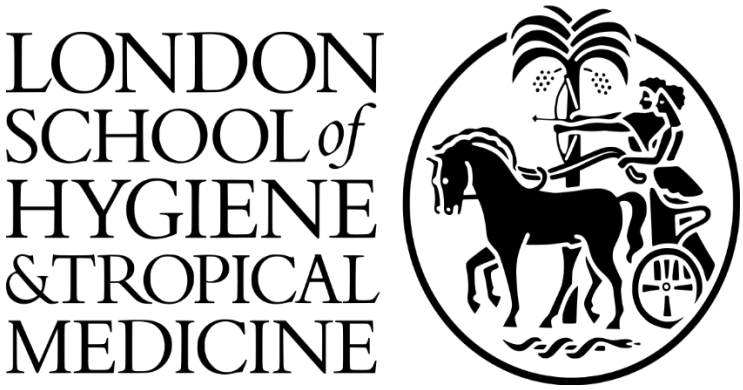


Advancing methods to account for biases in vaccine effectiveness research.

Sophie Graham

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

April 2024

Department of Clinical Research

Faculty of Infectious and Tropical Diseases

LONDON SCHOOL OF HYGIENE AND TROPICAL MEDICINE, UNIVERSITY OF LONDON

Funded by the National Institute of Health and Care Research (NIHR) Health Protection Research Unit in Vaccines and Immunisation

Declaration

I, Sophie Graham, confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed _____

Date _____

Plain language summary

Vaccines protect individuals from infections and reduce the burden of complications and death caused by the infections. Research into new vaccines occurs within the controlled setting of clinical trials. After vaccinations are approved, continued research in the 'real-world' is required to understand how vaccinations protect individuals in the community. However, research of vaccines in these settings has methodological challenges. For example, individuals that receive vaccines may be more likely to engage in other health-promoting activities than those who do not receive them. They also likely experience fewer barriers to accessing healthcare compared to those that do not receive vaccinations. When vaccinated and unvaccinated individuals are compared their health outcomes might appear to be better as these individuals are healthier in general. Therefore, methods are required to account for these differences between vaccinated and unvaccinated individuals in vaccine studies.

The first aim of this thesis was to investigate whether vaccine protection might be overestimated or underestimated by surveying the participants of an early test-negative study of COVID-19 vaccines in England. However, the test-negative design cannot be used in all situations, so the second aim was to identify markers that would indicate someone had healthier behaviours and better healthcare access in existing healthcare databases. The third aim was to use these markers to assess and account for overestimates or underestimates of vaccine protection in an influenza and COVID-19 vaccine study.

The first study identified no evidence of overestimates or underestimates of vaccine protection in the COVID-19 'real-world' study. The second study identified fourteen markers of healthier behaviours and healthcare access which could be measured using existing healthcare databases (e.g., breast cancer screening and pneumonia vaccination). In the third study, evidence of underestimates of vaccine protection were identified in the influenza study, until the markers were utilised. More accurate estimates of protection against influenza were obtained when differences in heath behaviours and healthcare access were accounted for. There was limited evidence of overestimates or underestimates in the COVID-19 vaccine study.

The study found that the test-negative design offers a robust way to estimate how well vaccines work. The markers offer an alternative approach for researchers to account for differences between vaccinated and unvaccinated people when the test-negative design cannot be used.

Abstract

Background

Observational studies are important for assessing vaccine effectiveness in the real-world, for example with new strains of pathogens or people changing their behaviour in response to vaccination. Influenza vaccine effectiveness estimates have previously been overestimated which may at least be partly due to unmeasured confounding from health-seeking behaviour and healthcare access. The test-negative-case-control design was developed to account for confounding from health-seeking behaviour and healthcare access in vaccine effectiveness research. This design requires test result data to be available and has a strong set of assumptions. It remains unclear whether confounding from health-seeking behaviour and healthcare access can be accounted for using alternative methods.

Aims

Overall, the aim of this thesis was to advance methods to account for biases in observational research. The first objective was to identify and quantify biases and alternative causal pathways in a COVID-19 vaccine effectiveness test-negative-case-control study which formed the basis of UK national monitoring by the UK Health Security Agency. As alternative methods are required to account for confounding from health-seeking behaviour in other study designs, the second objective was to systematically identify a set of markers of health-seeking behaviour and healthcare access in electronic health records (EHRs) that could potentially be used to quantify and account for this type of confounding. The third objective was to quantify and account for confounding from health-seeking behaviour and healthcare access in an influenza and COVID-19 vaccine effectiveness study with a cohort design using the markers from study two.

Methods

For the first objective, a questionnaire was sent to a sample of participants in one of the first UK COVID-19 test-negative-case-control vaccine effectiveness studies, which had used routinely-recorded data. Self-reported information on vaccination dates, symptomatic status, comorbidities and risk behaviours was used to explore potential biases and alternative causal pathways in the original study. For the second objective, markers of health-seeking behaviour and healthcare access were identified that were appropriate to a population aged ≥65 years. These were selected based on a health behavioral model known as the Theory of Planned Behaviour. These markers were then identified in the Clinical Practice Research Datalink (CPRD), a longitudinal dataset from

primary care practices, with linkages to hospital and mortality data. The prevalence of these markers in a population ≥66 years in England identified in the CPRD linked datasets were compared to national estimates. For the third objective, to quantify and account for confounding from health-seeking behaviour and healthcare access, a cohort study of COVID-19 and influenza vaccine effectiveness among older adults in England was conducted. Cox regression models were used to estimate vaccine effectiveness. The models were conducted in four sequential modelling steps – model one: adjusting for demographics, model two: additionally adjusting for ethnicity and deprivation, model three: additionally adjusting for comorbidities and model four: additionally adjusting for the health-seeking and healthcare access markers from study two. A negative control exposure cohort (history of influenza vaccination against early COVID-19 pandemic SARS-CoV2) was used to investigate the extent of residual confounding after adjustment for the markers.

Results

For the first objective, there was minimal evidence of bias, and accounting for multiple potential biases only changed the estimated vaccine effectiveness after two doses of BNT162b2 decreased from 88% (95% confidence interval [CI]: 79-94%) in the original study to 85% (95% CI: 68-94%). For the second objective, fourteen markers of health-seeking behaviour and healthcare access were systematically identified. These included preventative measures where the influence of underlying health need was minimal (e.g., bowel cancer screening). They had prevalence estimates that were comparable to national estimates e.g., 73.3% for influenza vaccination in the 2018/2019 season, compared to 72.4% in national estimates. For the third objective, adjusting for these markers in the influenza vaccine effectiveness study increased vaccine effectiveness estimates against influenza infections from -1.5% (95% CI: -3.2, 0.1%) in model three (adjusting for demographics, ethnicity, deprivation and comorbidities) to 7.1% (95% CI: 5.4, 8.7%) in model four (additionally adjusting for health-seeking and healthcare access markers). Similar trends were found for more severe endpoints. For COVID-19, vaccine effectiveness estimates minimally increased from 82.7% (95% CI: 78.3, 86.2%) in model one (adjusting for demographics) to 83.1% (95% CI: 78.7, 86.5%) in model four. Adjusting for these markers in the negative control exposure analysis, increased vaccine effectiveness estimates from nearer the null (model three: -7.5% [95% CI: -10.6 - -4.5%] to model four: -2.1% [95% CI: -6.0 - 1.7%]).

Conclusion

This thesis identified that when using the UK EHRs and the test-negative design the impact of potential biases on early pandemic COVID-19 observational vaccine effectiveness estimates was minimal. In instances where the test-negative-case-control design cannot be conducted, markers of health-seeking behaviour and healthcare access can be identified in EHRs. These markers can be used in other observational studies where health-seeking behaviour or healthcare access is relevant using study designs that are more broadly applicable (e.g., cohort). The effects of confounding from health-seeking behavior is context dependent with minimal impact during early COVID-19 pandemic implementation, but more pronounced for seasonal influenza estimates.

Acknowledgements

I would like to acknowledge the continued support and guidance of my supervisors, Dr Helen McDonald, Dr Edward Parker and Prof Dorothea Nitsch. Without them, the conduct of this thesis would not be possible. I would also like to thank Dr Jemma Walker, my original secondary supervisor, for her review of documents and statistical input at the start of this project.

I would like to thank the individuals at the UK Health Security Agency (UKHSA) for all their help, particularly with the data preparation and access for my first study. These individuals include Prof Nick Andrews, Dr Julia Stowe, Elise Tessier and Dr Charlotte Gower.

I would also like to thank my funders, NIHR, who financed the entirety of this project.

I am also grateful for the guidance that was received from my advisory committee which included Prof Nick Andrews, Prof Ian Douglas and Prof Anthony Scott, and from my upgrading examiners Prof Elizabeth Williamson and Dr Nick Davies.

Funding

This project was funded by NIHR Health Protection Research Unit in Vaccines and Immunisation (grant reference: NIHR200929).

Table of Contents

List of Tables

Tables in Appendices

List of figures

List of abbreviations

AAA: abdominal aortic aneurysm.

ACS: ambulatory care sensitive

APC: admitted patient care

ARI/ILI: acute respiratory infection or influenza like illness

BMI: body mass index

BMJ: British Medical Journal

CEV: clinically extremely vulnerable

CI: confidence interval

CPRD: Clinical Practice Research Datalink.

DAG: direct acyclic graphs

DEXA: dual-energy x ray absorptiometry

DID: diagnostic imaging datasets

DNA: did not attend.

EHR: electronic health record.

FIT: faecal immunochemical test

FOBT: faecal occult blood testing

GP: general practitioner.

HES: hospital episode statistics

HIV: human immunodeficiency virus

HRT: hormone replacement therapy

ICD-10: International Classification of Disease, 10th Revision

IMD: index of multiple deprivation

JCVI: Joint Committee on Vaccination and Immunisation

LSHTM: London School of Hygiene and Tropical Medicine

MESH: Medical Subject headings

MRI: Magnetic Resonance Imaging

NHS: National Health Service.

NICE: National Institutes of Health and Care Excellence

NIHR: National Institute for Health and Care Research

NIMS: National Immunisation Management Service

ONS: Office for National Statistics

OPCS: Operating Procedure Codes Supplement

PCR: polymerase chain reaction

PSA: prostate specific antigen.

QOF: quality outcomes framework

RCT: randomised controlled trial.

SGSS: Second Generation Surveillance System

SNOMED-CT: Systematised Nomenclature of Medicine Clinical Terms

TPP: The Phoenix Partnership.

UK: United Kingdom

UKHSA: UK Health Security Agency

US: United States

VAMP: Value Added Medical Products

1. Chapter 1: Thesis background

1.1 Introduction to the chapter

The aim of this thesis was to identify, quantify and account for biases in observational research using EHRs. The focus of this thesis was on vaccine effectiveness research, particularly for influenza and COVID-19.

Chapter 1 outlines the background and rationale for the thesis. The background and rationale informed the thesis overall objectives which are provided in **Chapter 2.**

1.2 Aim of chapter

To outline the background and rationale for the thesis.

1.3 Randomised clinical trials

Randomised controlled trials (RCTs) are the gold standard for clinical research and involve the randomisation of individuals to two or more interventions in a controlled setting. The randomisation of interventions is required to create randomness into the allocation process, which prevents systematic biases that could arise from non-random assignment[1,2]. However, despite their superiority to other designs, there are some key shortfalls that cannot be addressed through RCTs. One key shortfall is that it is not known whether the results hold for individuals that do not meet RCT strict inclusion criteria. In addition, their controlled nature does not allow for the study of drug-drug or drug-food interactions and often they cannot be used to study rarer endpoints due to lack of statistical power[3]. These key shortfalls can be addressed through observational research.

1.4 Observational research

Observational research is “non-experimental in nature, whereby the phenomenon of interest is observed without imposing experimental or controlled conditions”[4]. Thereby, patients are “observed” in the community clinical setting and certain clinical events are recorded as they occur. Observational research encompasses a variety of study designs, including cohort studies, case-control studies, cross-sectional studies, and others. Observational research can either be conducted using data collected prospectively, or using readily available datasets such as EHRs[5]. Historically these studies have been used to assess for potential associations between approved therapies and rare safety outcomes because of larger populations and longer follow-up times compared with RCTs. For example, in 1979 the association between post-menopausal oestrogens and endometrial cancer was studied in an observational study that used the Group

Health Cooperative of Puget Sound database[6]. Observational studies have been used in regulatory decision making for over 20 years and since the late 2010s interest has further increased with the publishing of regulatory guidelines on observational research best practices[3]. Furthermore, during the COVID-19 pandemic, these studies were critical to support the national and international response to the COVID-19 pandemic, including studies of vaccinations[7].

1.5 Electronic health records

EHRs are "the systematised collection of patient and population electronically stored health information in a digital format"[8].The contents of EHRs depends on the context, but can often include immunisations, laboratory results, radiological images, symptoms and clinical notes. In UK primary care, EHRs were first introduced in the late 1970s, with the aim of managing patients health[9]. Since the 1980s, EHRs have had multiple secondary purposes such their use in observational research[10]. For example, the Value Added Medical Products (VAMP) Research Databank was first created in 1987 for observational research purposes[11]. VAMP has since expanded to become the Clinical Practice Research Datalink (CPRD)[12]. The use of EHRs for observational research has advantages over prospective data collection due to the faster timelines and reduced costs[5].

1.6 Vaccine post-authorisation research

Vaccines stimulate the immune system to protect an individual from a harmful condition. It is currently estimated that vaccinations prevent 3.5 million to 5 million deaths annually, primarily from diseases like diphtheria, tetanus, pertussis, influenza and measles[13]. Vaccine efficacy is the term given to the protection afforded by vaccines in RCTs, whereas, vaccine effectiveness refers to protection from vaccinations in the 'real-world'[14]. Vaccine effectiveness estimates have many different uses. Each year influenza vaccine effectiveness estimates are generated using EHRs to inform vaccination recommendations for the current year and selection of vaccinations for the next season[15]. After vaccination marketing authorisation, manufacturers are often required to conduct post-authorisation studies using EHRs to assess the continued safety and effectiveness of the approved vaccination. For example, for COVID-19 vaccinations, vaccine effectiveness estimates using EHRs were mandated by the European Medicines Agency for the conditional approval of the Pfizer-BioNTech, Moderna, AstraZeneca, and Janssen COVID-19 vaccinations[16]. Vaccine effectiveness estimates using EHRs were also used during the COVID-19 pandemic to inform policy decisions in the UK. These informed the number needed to vaccinate to prevent hospitalisation and COVID-19 deaths to inform the booster programmes[17]. Generation of vaccine effectiveness estimates were particularly important during the COVID-19 pandemic as the

backdrop changed dramatically from the original RCT settings with the evolution of novel variants[18]. In future inevitable pandemics, it is likely that similar estimates using EHRs will be essential to ensure a rapid response is performed.

1.7 Causal inference

Causal effect for an individual is the outcome that would have been observed by an individual had they received treatment compared with the outcome that would have been observed were they untreated. This is also referred to as *counterfactual outcomes.* It is generally not possible to observe individual causal effects and therefore, average causal effects in a population are generally observed. For causal inference to hold, exchangeability, positivity and consistency are required. Exchangeability assumes there is no unmeasured confounding, enabling causal interpretation of treatment effects after adjusting for observed covariates. Positivity ensures that all individuals have a non-zero probability of receiving any treatment level, allowing for valid comparisons across treatment groups. Consistency assumes that the observed outcomes match the potential outcomes under the given treatment, with no interference between treatment groups. Interference between treatment groups, also known as spillover effects, occurs when the treatment or intervention applied to one group affects the outcomes of other units[19].

Average causal effects in vaccine effectiveness research typically compare the occurrence of the disease in vaccinated individuals compared to a group of unvaccinated individuals[20]. The causal estimand is the parameter that represents the true causal effect of vaccinations on the outcome of interest, whereas the statistical estimand is the parameter that a statistical model estimates from the dataset of interest[21]. Since this thesis aims to quantify and correct for structural bias, the causal estimand is a vaccine effectiveness estimate that is not impacted by residual bias.

1.8 Biases in observational research

Effectiveness estimates from observational studies do not always reflect estimates from RCTs because of the impact of bias. Systematic bias occurs when there is an “association between treatment and outcome that does not arise from the causal effect of treatment on outcome in the population of interest.”[22] Three main types of bias in observational research include selection bias, information bias and confounding bias. Selection bias is when errors are introduced through selection of the study population[22]. Information bias occurs when there are errors in data collection[23]. Confounding occurs when the observed association between an exposure and an outcome is influenced by the presence of an extraneous variable that is related to both the exposure and the outcome[19]. Inadequate control of confounding leads to confounding bias. In

observational research, there are known common confounders (e.g., age, gender) that are accounted for either in the study design or analysis, but typically there are other confounders that are not directly measurable in the data or are not known and therefore are not accounted for[24]. An example of a common potential confounder that is not directly measurable in EHRs is confounding from health-seeking behaviour and healthcare access.

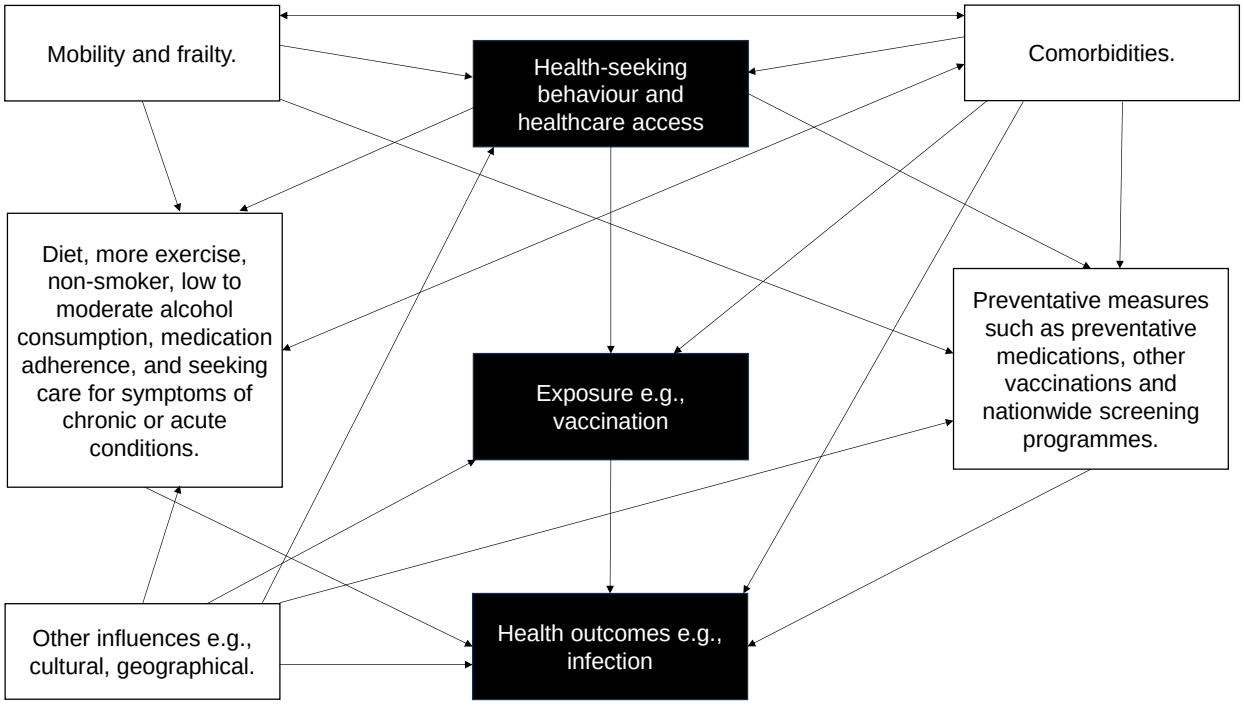
1.8.1 Confounding from health-seeking behaviour and healthcare access

Health-seeking behaviour can be defined as: “any activity undertaken by a person believing [themselves] to be healthy, for the purpose of preventing disease or detecting it in an asymptomatic stage”[25]. Healthcare access can be defined as: “the ability to obtain healthcare services such as prevention, diagnosis, treatment, and management of diseases, illness, disorders, and other health-impacting conditions”[26]. Individuals with better health-seeking behaviours and healthcare access are more likely to have better diets, exercise more, take up preventative measures such as cancer screenings or vaccinations. They are also more likely to adhere to their medications and preventative therapies and when they are presented with symptoms of a chronic or acute condition they are also more likely to seek care earlier, allowing for more effective treatment before their disease progresses compared with individuals not engaged in their health or with poor healthcare access[27]. In observational research, if health-seeking behaviours and healthcare access are associated with both the exposure and outcome of interest, this can lead to confounded estimates[28].

Confounding from health-seeking behaviour and healthcare access is a complex phenomenon that has many different influences. A conceptual diagram of the potential mechanism of confounding from health-seeking behaviour and healthcare access in observational research can be found in

Figure 1 below.

Figure 1 Conceptual diagram of the relationship between health-seeking behaviour or healthcare access and other variables in observational research



Note: Figure 1 from Eurich et al, 2012 was used to inform this diagram[27]

One of the first examples of potentially confounded estimates from health-seeking behaviour and healthcare access (herein shorthanded to: “confounding from health-seeking behaviour”, which will refer to both health-seeking behaviour and healthcare access, unless otherwise specified)

was the reported impacted of hormone replacement therapy (HRT). One study that was conducted in 1985, reported that women taking HRT had half the risk of cardiovascular disease as those not taking HRT29. Many other observational studies followed this with similar reported results30. On this basis, American guidelines recommended HRT for the prevention of cardiovascular disease, and by 2001 there were an estimated 15 million women in the United States (US) using this therapy31. RCTs have found the opposite impact, that HRT increase risk of cardiovascular disease32,33. Some of the overstated effects of HRT in observational research are not thought to be due to confounding from health-seeking behaviour34. According to the confounding structures shown in

Figure 1 above, women that take HRT as a preventative therapy are likely to be more engaged in their health and/or have better healthcare access, and therefore might be more likely to take up preventative services and have healthier lifestyles, all of which contribute to favourable outcomes. Other authors have reported other potential contributors to the overstated effects of HRT in observational research. These include selection bias introduced that is introduced through the selection of prevalent users35 and confounding from deprivation36.

Other examples of potential overstated effects due to confounding from health-seeking behaviour have occurred in statin use studies. These have consistently shown to reduce hip fracture risk, which has not been reflected in clinical trial data[37]. Observational studies of statins have also shown to reduce the risk of Alzheimer's disease[38], sepsis[39] and cancer[40]. However, more recently an observational study found associations between statin adherence and preventative therapies. They used a prospective cohort of 20,783 new users of statins between 1996 and 2004 and after adjustment for age, gender and comorbidites, they found that patients with two or more filled statin

prescriptions had increased risk of prostate-specific antigen (PSA) tests (hazard ratio [HR]: 1.57 [95%CI: 1.17, 2.19]), bowel cancer screening (1.31 [95%CI: 1.12, 1.53]), breast cancer screening (1.22: [95%CI: 1.09, 1.38]), influenza vaccinations (1.21, [95%CI: 1.12, 1.31]), and pneumococcal vaccinations (1.46 [95%CI: 1.17, 1.83]) during follow-up[41]. The association with these preventative therapies potentially signifies that estimates are confounded by health-seeking behaviour.

Another area of research that has been impacted by confounding from health-seeking behaviour is influenza vaccine effectiveness research. In cohort studies of influenza vaccine effectiveness of individuals aged ≥65 years, authors have reported reductions in all-cause mortality by 40-50%42,43. These high estimates of all-cause mortality reduction have since been speculated since each winter influenza contributes to a maximum of 10% of deaths per year. Therefore, even if the influenza vaccination was 100% protective, it could be expected to directly prevent 10% of deaths per year44,45. Plausible effects beyond the direct protection of influenza-related mortality (i.e., on all-cause mortality) have previously been hypothesised. One proposed mechanism is that influenza infections can cause local and systemic inflammatory response, which in turn can lead to plaque instability, leading to plaque rupture and adverse cardiovascular events such as atherosclerosis, coronary artery disease and stroke46-48. However, it is unlikely that this indirect mechanism would contribute to the remainder of estimated all-cause deaths prevented. RCT estimates of efficacy in this age group (70 years and over) report an efficacy of only 23% for cause specific outcomes (laboratory confirmed influenza)49. Instead, it is likely that since only a subset of eligible individuals take up the influenza vaccination each year, those that do are more likely to be those with better health seeking-behaviour and fewer barriers to healthcare compared with those that do not take up the vaccine. This coincides with other health benefits as previously described and shown in the conceptual diagram in

Figure 1.

Jackson et al, 2006[50] demonstrated this issue by conducting an influenza cohort study using US administrative data (Group Health Cooperative). They estimated influenza vaccine effectiveness during each influenza season from 1995 to 2002 amongst individuals aged ≥65 years. They also estimated influenza vaccine effectiveness in vaccinated versus unvaccinated in the pre- and post-influenza season assuming that without residual bias there would be no effect when the virus is not circulating. They found that even after adjustment for age, sex, comorbidities, pneumonia hospitalisations and outpatient visits, that the relative risk of all-cause mortality was estimated to be 0.36 (95%CI: 0.30,0.44) for pre-season and 0.66 (95%CI: 0.61,0.72) post-season amongst vaccinated versus unvaccinated. These estimates show that influenza vaccine effectiveness cohort studies are likely to be highly biased by health-seeking behaviour and healthcare access even when they adjust for key potential confounders. Other authors have reported similar findings using off-season estimates[51]. A systematic literature review published in 2015, investigated the frequency and impact of confounding by indication and confounding from health-seeking behaviour in influenza vaccine research. They reported that of the twenty three studies identified, 83% showed high risk of bias, with fourteen due to confounding by indication, two for confounding by health-seeking behaviour and three for both[52]. The Centers for Disease Control and Prevention does not cite observational influenza vaccine effectiveness studies with high all-cause mortality prevention because of the likely impact of bias[53].

In terms of COVID-19 vaccine effectiveness the impact of health-seeking behaviour is less well known, due to the novel nature of this condition. At the beginning of the COVID-19 pandemic, researchers used similar methods to influenza vaccine effectiveness studies, as they were unsure of the influence of health-seeking behaviours and healthcare access. Potential evidence of this bias was identified in a study that investigated COVID-19 related mortality in those two received a BNT162b2 booster compared to those that did not receive a booster. The authors reported an estimated 90% reduction in COVID-19 related mortality in those that received the booster vaccination compared to the non-boosted[54]. More recently, a response to this article was published. Through simple calculation they estimated that the original study had 94.8% lower non-COVID-19 related mortality in the booster group compared with the booster group. This they summarised was likely due to the booster individuals possessing healthier behaviours and better healthcare access compared to their non-booster counterparts[55].

1.9 Methods to identify and quantify biases in observational research

Many different methods have been used in observational research to identify and quantify biases. The most common methods that researchers use is multiple adjustment and propensity scores[56]. Residual bias after confounding adjustment can be identified through use of negative controls. A negative control is a group or condition where no response is expected. Negative controls are required to have a) no plausible causal mechanism that it causes the outcome under study (negative control exposure) or is caused by the treatment under study (negative control outcome) and b) is also required to be affected by the same confounding structure as the treatment or outcome under study[57,58]. If condition a) and b) are met then any association between the negative control exposure and original outcome, or original exposure and negative control outcome is likely due to bias. However, negative controls are limited as they cannot be used to detect the type of bias. Moreover, the choice of negative controls relies on assumptions about the implausible association between the negative control exposure and original outcome or the original exposure and negative control outcome. If, of course in reality there is some causal association, then it is incorrectly assumed that there is residual bias in the association between the original exposure and original outcome[58]. As mentioned above a popular negative control in influenza vaccine effectiveness research is using off-season estimates[50,51,59,60]. For COVID-19, vaccine effectiveness shown in the first 14 days after vaccination (i.e., before immunogenicity is reached) has been suggested as a negative control[61-64].

Simulation studies are also very popular methods that are used to assess the potential impact of a particular bias[65,66]. For example, Smedt et al, 2018[65] assessed the potential impact of exposure and outcome misclassification in four different observational study designs (cohort, case-control, test-negative design and screening methods) on influenza vaccine effectiveness estimates. They reported that exposure and outcome misclassification led to biased estimates, with the test-negative design performing the worst. Although simulation studies are useful to examining the potential impact of biases, they use external data and therefore might not represent true associations. Other methods used include approaches such as use of directed acyclic graphs (DAG) and quantitative bias analyses[66]. However, these approaches are limited as they also rely on external data.

1.10 Methods to account for biases in observational research

Many different methodological approaches have been previously developed to control for potential biases in observational research. The use of these methods to control for potential biases in real-world evidence generation have been emphasised in recent regulatory

guidelines[67,68]. In terms of controlling for confounding in observational research using EHRs, Nørgaard et al, 2017[69] summarised different approaches. These approaches can be found summarised in Table 1 below.

Table 1 Approaches to control for potential confounding in observational research using EHRs

Type of confounding	**Example confounder**	**Example approaches to control for confounding**
Directly measurable	Age and sex	Restriction Matching Stratification Standardisation Regression analysis Propensity scores[1]
Directly unmeasurable	Disease severity; health-seeking behaviour; frailty	External adjustment Proxy measures* Imputation Test-negative-case-control design* Self-controlled design* Ratio-of-ratio method* Instrumental variable[2] Mendelian randomisation[3] Active comparator[4] Regression discontinuity design[5] Sensitivity analyses

Note: this table is adapted from Nørgaard et at, 2017[69].

*Described in the text below this table.

[1]Propensity scores: a probability of treatment score that is assigned to each individual in the data that is conditional on their covariates[56].

[2]Instrumental variables: a variable available in the data that is associated with treatment, but not associated with the health outcome except through its effect on exposure. Regression of the outcome on this variable will give the effect of the exposure on outcome in the absence of confounding. However, different assumptions have to be met, and often it is difficult to find a suitable instrumental variable[70].

[3]Mendalian randomisation: uses genetic variation as instrumental variables to investigate the effects of modifiable risk factors on health outcomes. However, there are key assumptions that need to be met, and this method only works for risk factors that are modifiable[71].

[4]Active comparator: when a drug is compared to an active comparator that is indicated for the same condition, rather than comparison to a non-user. However, often it is not possible to find a comparator drug[72].

[5]Regression discontinuation design: can be used for programmes that are introduced/discontinued at a point in time and then the outcome can be compared immediately before or after this time point. However, this only works for programmes as usually it takes longer for approved medications to be taken up in clinical practice[73].

The suggested approaches are based on whether the confounder is directly measurable i.e., can be directly identified in the data, or directly unmeasurable. The reason why confounding from

health-seeking behaviour is typically not adjusted for in observational studies using EHRs, is because it is typically not directly measurable in the data.

For confounding that is measurable, simple statistical methods such as stratification or adjustment can be used.

For confounding that is directly unmeasurable, proxy markers have previously been used to control for confounding. A proxy marker is another similar variable that is used to represent the directly unmeasurable variable[74]. For example, Farout et al, 2015[75] identified proxy markers available in US claims data to account for differences in frailty amongst treated and untreated individuals. Proxy markers included twenty different markers such as oxygen therapy, wheelchair use and arthritis, as presence of these markers might indicate evidence of frailty. Zhang et al, 2017[51] used these markers to adjust for confounding from frailty in an influenza vaccine effectiveness study in a US study of ≥65-year-olds using Medicare data. They found that adjustment using these markers reduced vaccine effectiveness estimates against all-cause mortality from 32% (95%CI: 31-33%) to 27% (95%CI: 26-28%).

Common study designs that have also been used to account for directly unmeasurable confounding are the test-negative-case-control design, the self-controlled design and the ratio-of-ratio method. The test-negative-case-control design attempts to accounts for differences in health-seeking behaviour (as well as differences in exposure to the vaccine preventable condition) by including only individuals who sought care when they experienced symptoms consistent with the vaccine-preventable disease[76]. In observational research it might not always be possible to conduct the test-negative-case-control design as test result data is required. In addition, as mentioned previously, estimates from test-negative-case-control designs are the most biased in the instance of outcome misclassification from imperfect sensitivity and specificity compared with other study designs[77]. Further details on this design can be found in **Chapter 3**, Section 3.3.3. The self-controlled design accounts for differences in health-seeking behaviour by requiring all individuals to act as their own controls[78]. The self-controlled design requires a number of assumptions to be met. For example, the outcome cannot influence subsequent exposures or the end of the outcome period[79]. The ratio-of-ratio method has previously been used to reduce confounding from health-seeking behaviour and frailty in an influenza vaccine effectiveness study[80]. This method takes advantage of a natural experiment – for some influenza seasons the circulating strain matches the vaccination, whereas in other years it does not. The method compares hazard ratios for vaccination in matched versus unmatched years with unvaccinated in matched versus unmatched years. However, this method is restricted to influenza vaccine

effectiveness only, and assumes that the influenza vaccine has no clinical benefit in unmatched seasons[79].

1.11 Rationale for research

As discussed, estimates from observational research are important as they are used in policy and clinical decision making. Particularly during pandemic contexts, they are extremely important for providing rapid vaccine effectiveness estimates to inform policy decisions during this time. Although use of these data has exploded since the late 1970s, they are also expected to further increase due to regulatory bodies providing guidance for their uses[67,68]. It is therefore necessary to ensure that the most appropriate methods that are robust to different potential biases are utilised. For example, it is necessary to understand whether the test-negative design, which is commonly used in vaccine effectiveness research, is robust to different potential biases. This is necessary to ensure the most accurate estimates of all approved interventions are generated so that informed decisions can be made. Robust methods are also required to uphold trust in use of EHRs so that their continued use can be assured. It is also important that the methods from these studies are refined, so that when the next pandemic inevitably comes, the tools that are required are available so that a rapid and efficient response can be performed. Since current methods that are used to account for potential biases in observational research (e.g., confounding from health-seeking behaviour) are limited (see Section 1.10), alternative methods are needed to identify, quantify and account for these biases. Methods that can be applied more broadly e.g., without strong assumptions, are required. This rationale informed the aim and study objectives for the thesis, which are outlined in **Chapter 2**.

2. Chapter 2: Overarching chapter

This chapter provides the overall thesis objectives and provides an overview of the layout of this thesis to guide the reader across the chapters.

2.1 Aims and research questions for thesis

The overarching aim of this thesis was to develop methods to account for biases in observational research using EHRs. The more specific aims of this thesis were:

1. To identify and quantify the size and direction of biases and alternative causal pathways in a COVID-19 vaccine effectiveness observational study using a test-negative-case-control design.
2. To systematically identify a set of markers of health-seeking behaviour available in EHRs that can potentially be used to quantify and account for this type of confounding.
3. To quantify and account for confounding from health-seeking behaviour an influenza and COVID-19 vaccine effectiveness study.

2.2 Layout for thesis

To address these objectives, three different studies were conducted. Study one, which addresses objective one, can be found in **Chapter 3**. Study two, which addresses objective two, can be found in **Chapter 6**. Study three, which addresses objective three, can be found in **Chapter 7**. **Chapter 4** includes a pragmatic literature review that identifies previous literature that has explicitly used methods to account for biases in vaccine effectiveness research using EHRs. This was conducted to understand the existing methods used before conducting study two and three. **Chapter 5** gives a general overview of the methods (datasets used, code list and variable creation for baseline variables) for study two and three, since these methods were consistent across the two studies. Lastly, discussion for the thesis overall is provided in **Chapter 8**.

3. Chapter 3: Study one: Identifying and quantifying biases in a COVID-19 observational vaccine effectiveness study

3.1 Introduction to the chapter

This chapter aimed to summarise the potential biases that were investigated in one of the first UK COVID-19 VE studies. As discussed in **Chapter 1,** COVID-19 vaccine effectiveness studies conducted during the COVID-19 pandemic were used to inform policy and clinical decisions. However, since observational studies using these data can be subject to bias, it is necessary to identify and quantify the potential impact of biases in these studies to ensure that estimates are as accurate as possible. At the early stages of the pandemic, potential biases that could be present in COVID-19 observational research studies were theorised. For example, Lewnard et al, 2021[81] discussed that outcome misclassification might be present in case-control and test-negative studies due to prolonged viral shedding and asymptomatic infections. Confounding from health-seeking behaviour and healthcare access was also potentially present as those that access and receive vaccines are likely different to those that do not (discussed in **Chapter 1** Section 1.8.1). In terms of methods that have historically been used to identify and quantify biases in observational research, these were previously discussed in **Chapter 1**. However, as previously discussed these methods are limited as they either require assumptions or external data that might not hold true.

Therefore, in the current study a different approach was used to detect and quantify the potential impact of biases in a COVID-19 vaccine effectiveness study. For this nationwide vaccination and polymerase chain reaction (PCR) COVID-19 testing data from one of the first COVID-19 vaccine effectiveness studies in the UK was utilised. This data was supplemented with data from a questionnaire that was sent to over 20,000 individuals from the original study to understand the presence of potential biases and alternative causal pathways. Biases that were investigated included exposure misclassification, outcome misclassification, confounding bias from comorbidities and deferral bias (discussed further in the paper below) as well as alternative causal pathways from vaccination to infection including riskier behaviour after vaccination and attending a vaccination visit causing infection (also discussed further below). These were investigated by comparing vaccination and testing information in the original data with the questionnaire data and by supplementing these data with additional information from the questionnaire on comorbidities and risk behaviours. The impact of the biases was assessed by updating the original vaccine effectiveness estimate that accounted for each of the biases. Alternative causal pathways from

vaccination to infection were also investigated[82]. These alternative pathways are unique to the real-world and can increase or decrease vaccine effectiveness estimates compared with clinical trial estimates. An example of an alternative causal pathway is riskier behaviour after vaccination (e.g., mixing more with individuals outside of their household), which individuals might exhibit since they have the perception of protection after they have been vaccinated; another example is contracting COVID-19 when travelling to or from, or even at, the vaccination centre. For this, the risk of SARS-CoV-2 infection amongst those with riskier behavior was reported.

This chapter first provides an overview of the original study, the original datasets and the questionnaire data. Some study context is also provided as well as further information on the original study test-negative design, study population and findings. Then the main study methods, results and discussion are found in paper one below.

3.2 Aim of chapter

To quantify the size and direction of potential biases and alternative causal pathways that may have impacted estimates from one of the first UK COVID-19 vaccine effectiveness studies.

3.3 Overview of original study

The original study was published in the *BMJ* on the 13th May 2021[63]. It was conducted by the UK Heath Security Agency (UKHSA) team during the early stages of COVID-19 vaccine deployment in the UK. They adopted a test-negative-case-control design (described below in Section 3.3.3) and included all individuals aged ≥70 years with a COVID-19 PCR test that occurred between 26th October 2020 to 21st February 2021.

3.3.1 Original study datasets

The original study used nationwide vaccination data (National Immunisation Management Service [NIMS]) linked to nationwide COVID-19 PCR data (Second Generation Surveillance System [SGSS]). These datasets are summarised below, but first an overview of the UK National Health Service (NHS) and NHS England datasets is provided.

3.3.1.1 UK National Health Service

The UK NHS, which is made up of NHS England, NHS Scotland and NHS Wales was created in 1948[83]. It provides healthcare that is free at the point of delivery to the entire population in the UK, except for some outpatient prescription charges in England which are currently £9.65 per item and some charges for dental and optician care. Certain prescriptions in England are exempt from

these charges (e.g., contraception) as well as some groups of individuals (e.g., age ≥60 years or pregnant individuals)[84].

The NHS is made up of primary, secondary, tertiary and community care services. Primary care includes general practices, community pharmacies, dental and optician services[85]. General practitioners provide primary care to individuals that are registered at their practice. Doctors and dentists have control over practice operation and are paid on a per capita basis. Primary care services provide the first point of contact to the healthcare system for any non-emergency health-related issues, acting as the 'gatekeeper' to the NHS. Individuals who visit primary care services are referred to secondary care for specialist treatment, if necessary[86]. Over 98% of the population in the UK are registered with a general practice and patients visit the same practice unless they choose to transfer out and register with a new practice[87].

Secondary care includes planned or elective care, urgent and emergency care, which includes 999 (emergency) and 111 (non-urgent helpline) services, ambulances and out-of-hours GP services and mental health care[85]. Specialists in the UK are largely based within clinics within hospitals. Patients cannot access these services without a referral, except for some small exceptions that include emergency department services and sexual health clinics[88]. Tertiary care includes highly specialist treatment such as neurosurgery, transplants, plastic surgery and secure forensic mental health services[85]. Tertiary care is usually provided in larger or teaching hospitals. Providers have access to more specialist equipment and are required to have a higher level of training than in other services. Referrals from other consultants or GPs are also required to access tertiary care. Lastly, community health includes district nurses, health visits, child health services and sexual health[85]. These services are delivered within an individuals' home and usually aim to support the independence of individuals with complex health conditions[89]. All individuals that access NHS care in the UK are assigned an NHS number which is unique to all individuals and helps to maintain a complete care records for each patient across all settings[90]. Over 12% of the population in the UK have private health insurance, which is mainly provided by an individual's employer. This mainly provides access to acute elective care[91].

3.3.1.2 NHS England datasets

The current study used the original study datasets from NHS England. For information on the NIMS dataset see Section 3.3.1.5 below and for information on the SGSS dataset see Section 3.3.1.6 below. NHS England (previously known as NHS Digital) have a statutory role in collecting data across health and social care in the UK. Overall, the collection of data by NHS England serves multiple purposes aimed at improving healthcare delivery, promoting public health,

supporting research and development, and informing healthcare policy and decision-making. These datasets can be linked at the patient level, using an individual's NHS number. NHS England have access to the patient Spine, which provides a master database, known as the Personal Demographic Service, of all the demographics of all patients in England and Wales to which all other datasets can be linked using NHS number[92].

3.3.1.3 Coding systems used in NHS England datasets

Coding classifications are used in NHS England datasets to record clinical events including diagnoses, symptoms, procedures and medications. One of the main coding systems in UK health data is Systematised Nomenclature of Medicine Clinical Terms (SNOMED-CT) code[93]. SNOMED-CT is a clinical terminology that includes more than 300,000 concepts that are organised into hierarchies. Codes are organised into 19 hierarchies. Since SNOMED-CT is an ontology codes can be organised into more than one hierarchy. Top level hierarchies include concepts based on clinical information e.g., clinical findings, observable entities, procedures and body structures. Codes can also be mapped together based on their relationships and each code can be mapped to multiple codes based on the meaning of the code[94]. International Classification of Diseases, Tenth Revision (ICD-10) is another common coding classification in EHRs in the UK. It was developed for global mortality statistics by the World Health Organisation. This is a hierarchical coding classification that organises codes into chapters according to body systems with codes organised alphabetically within each chapter[95].

3.3.1.4 Access to NHS England datasets during the COVID-19 pandemic

During the COVID-19 pandemic, research using EHRs became a priority as these data allowed the rapid assessment of risk factors for infection and the continued assessment of effectiveness and safety of the vaccinations and treatments. Data that were utilised were secondary care data, as well as new data that were collected for purposes of assessing the deployment of COVID-19 PCR testing and COVID-19 vaccinations. COVID-19 PCR testing data was collected as part of UKHSA's infectious diseases and antimicrobial resistance surveillance (see SGSS described in Section 3.3.1.6 below) and COVID-19 vaccination data was collected through the NIMS system that was previously set up for influenza vaccinations (see NIMS described in Section 3.3.1.5 below).

3.3.1.5 NIMS

In England, routine vaccinations are recorded in a patient's GP record. For children under 19 years, vaccinations are also recorded in Child Health Information Systems, which is made up of sub-registers. Vaccine monitoring through this approach relies on correct recording of

vaccinations that occur at the GP practice surgery and for vaccinations that are administered outside of the GP practice, it relies on feedback information being correctly coded in the GP practice file. NIMS was set up by the NHS to improve data flow of influenza vaccination data across different systems (e.g., pharmacies, hospitals, schools). At the beginning of the COVID-19 pandemic it was made clear that vaccinations would have to be rapidly deployed across multiple different settings. Vaccination data would also need to be made available in almost real-time so that the continued effectiveness and safety of the vaccinations could be assessed. Therefore, the NIMS system was adapted in 2020 to also include COVID-19 vaccinations. This dataset had multiple functions, but it was primarily used for influenza to identify individuals prioritised for vaccinations. Point of care applications are used at each site (e.g., GP, pharmacy) to record key information on each vaccination. Unique patient NHS number, vaccination date and batch number are mandatory items to record. When information on this patient is entered, data from the NHS Spine is used to ensure the data is complete and accurate. If data is entered on an individual not registered with a GP practice in England, a new NHS number is generated. NHS England then validate this data and link it to GP and hospital data to identify groups of individuals with clinically extremely vulnerable (CEV) status (described in Section 3.3.2.1 below) or who are pregnant. Cohort information generated by NHS England is then pushed back into NIMS. This combined data is then sent to UKHSA in a secure environment. Data is delivered to UKHSA in two separate files:

- Population denominator file: this includes an NHS number of all individuals in England with accompanying basic demographic information such as age, gender, ethnicity (as defined in the 2001 census), CEV flag and healthcare and social care flags.
- Vaccination events file: this includes information on each vaccination event including location of where the vaccination was delivered, date of vaccination administration, manufacturer information and batch number[96]. UKHSA have also compared vaccination dates and manufacturer in these data with survey responses and reported that accuracy was high (with no measures reported)[96].

UKHSA receives NIMS data daily through a structured query language server. Data cleaning processes are carried out by UKHSA before the data is made available to external researchers. The data is de-duplicated, NHS numbers are validated and any anomalies are checked for. For individuals that are not registered with a GP practice, the allocated NHS number of the unregistered individuals can still be identified. Vaccination dosing and manufacturer information is assigned a specific SNOMED-CT code (see Section 3.3.1.3). In NIMS, batch numbers are

cleaned and then provided alongside the SNOMED-CT code. They then link records to individual postcodes from GP records to assign the region based on 2011 the Office for National Statistics (ONS) rural/urban classification[97] and 2019 Index of Multiple Deprivation (IMD) decile[98]. The ONS urban/rural classification assigns individuals that are based in a built-up area with a population over 10,000 to urban and then all remaining areas are assigned rural. These are then assigned into six different settlement types. The 2019 IMD is an English index of deprivation that is based on 32,844 small geographical areas in England known as Lower-layer Super Output Areas. The indices are based on 39 separate indicators which is organised into seven domains (income, employment, education, health, crime, barriers to housing and services and living environment), which are then combined and weighted to create the index. Data for each of these indices comes from multiple different sources (e.g., census information). In all cases the most up-to-date data is used. All areas are then ranked compared to all other areas. Those in the lowest ranking are labelled as the most deprived, whereas those in the highest ranking are labelled as the least deprived. The variables available in NIMS after UKHSA variable creation can be found summarised in Table 2 below.

Table 2 Variables available in NIMS

File	Variable	Description
Population denominator file	Name	
	Date of birth	
	NHS number	
	Sex	Male, female or unknown.
	Ethnicity	Based on ethnic category code 2001[99].
	Region	Using patient postcode.
	General practice code	
	Flag for CEV*	Provided by NHS England and is a flag for individuals that were identified based on linkage to GP electronic health records, Hospital Episode Statistics (HES) and QCOVID risk assessment[100].
	Flag for front line healthcare and social workers	Provided by NHS Business Services Authority who have information on individuals that are employed by NHS organisations using an electronic staff record.
Vaccination event file	NHS number	
	Date of vaccination	
	Location code	Unique code for location where the vaccination occurred.
	Vaccination code	SNOMED-CT code.

	Vaccination procedure code	SNOMED-CT concept code.
	Route of vaccination	
	Body site	
	Batch number	Vaccination batch number.
	Manufacturer	Manufacturers name.
External sources	IMD**	2019 IMD decile[98]
	Rural/urban***	2011 ONS rural/urban classification[97]
	Care home status	Unique property reference numbers and NHS addresses are linked to the care home Care Quality Commission addresses. These are then linked to the Master patient Index provided by NHS England. This list is updated monthly.
	Age on 31 March 2021	Calculated using date of birth information.

Note: this table is adapted from Tessier et al, 2022[101].

*CEV are those individuals that were asked at the beginning of the UK COVID-19 pandemic to shield because of their high-risk status[100,102].

**IMD is a relative deprivation score that is calculated based on a patient or practice postcode.

***Urban status is assigned based on practice or patient post-code with 10,000 inhabitants or more. Rural are postcodes from all other regions.

Abbreviations: IMD: index of multiple deprivation; SNOMED-CT: Systematised Nomenclature of Medicine Clinical Terms.

3.3.1.6 SGSS

NIMS data has been linked through patient identifiers to the SGSS. SGSS is the UKHSA's system that stores and manages data on laboratory data on infectious diseases and antimicrobial resistance. Laboratories have been required to report since 2010 any positive test of listed notifiable organism to SGSS. The list of notifiable organisms includes viral infections such as Ebola, Dengue, hepatitis, influenza as well as a long list of bacterial infections. Data collection for COVID-19 SGSS began on 6 April 2020.[103] The laboratories that report into SGSS are from pillar one and pillar two. Pillar one includes any swab testing that occurs in UKHSA's laboratories or within NHS hospitals for those with a clinical need or for healthcare workers. Pillar two includes community testing for the wider population that was provided free by the UK government from July 2020 until April 2022[104,105]. Variables that are reported in SGSS can be found in Table 3 below.

Table 3 Variables available in SGSS

Information	**Variable**	**Description**
Laboratory information	Source laboratory	
	Reference laboratory	
	Reporting laboratory*	
Patient information	Name*	

	NHS number*	
	Hospital number*	
	Date of birth*	
	Sex*	
	Region	Using patient postcode.
	Ethnicity	
Testing information	Organism*	The full name of the organism and results.
	Date of onset	The date that symptoms of the illness began.
	Specimen type*	Whether it was blood, sputum, serum ect.
	Specimen date*	The date that the specimen was collected.
	Identification method	

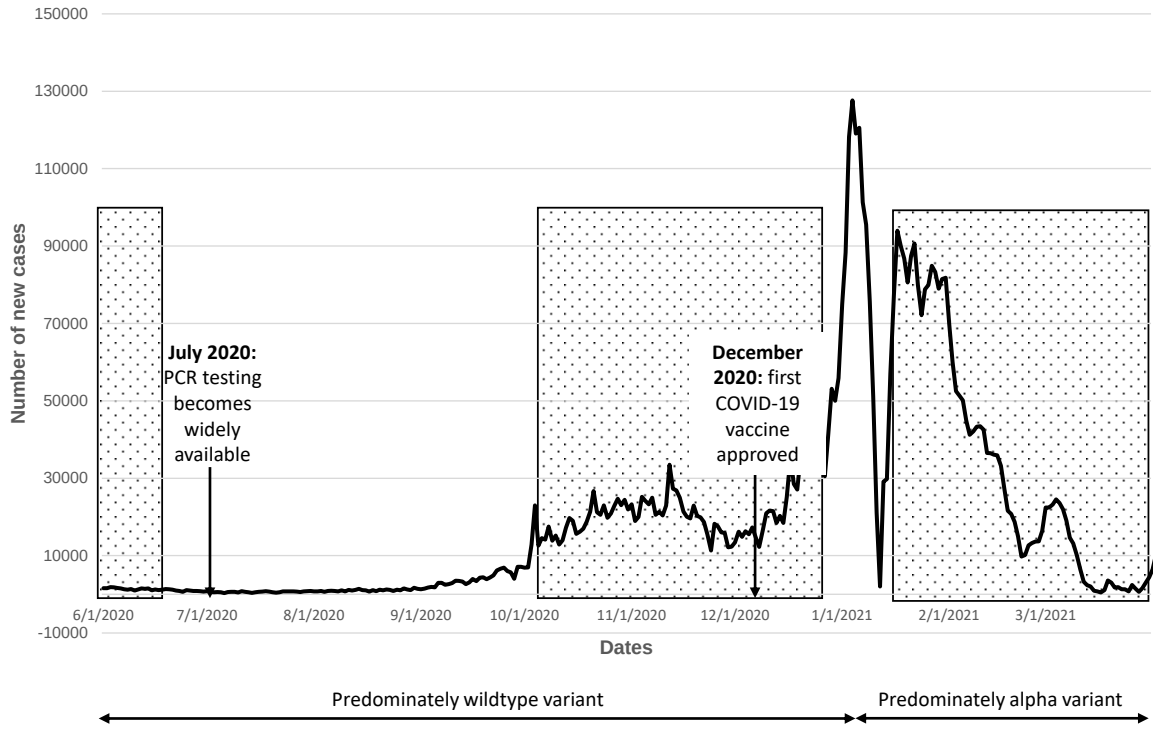
Note: this table is adapted from UKHSA's laboratory reporting guidelines[103].

*Fields that are mandatory.

3.3.2 Study context

The study period of the original study[63] spanned from October 2020 until March 2021, which was mostly during the early stages of the COVID-19 pandemic in the UK. Free governmental COVID-19 PCR community testing became available before the study period in July 2020. The first COVID-19 vaccination was approved in the UK on 8 December 2020. During this time period there were fluctuations in COVID-19 cases, evolution of a new COVID-19 variant and changes in population mixing patterns due to different lockdowns. Key dates and relevant information can be found summarised in Figure 2 below.

Figure 2 A summary of key dates in the UK COVID-19 pandemic to contextualise the original study



Note: this figure was adapted from UKHSA positive PCR COVID-19 cases data[106] and then overlayed with different key dates. The dotted boxes indicate when the different lockdowns occurred in the UK.

3.3.2.1 Clinically extremely vulnerable definition UK pandemic

CEV people were asked to shield during the UK pandemic due to their high-risk status. A shielding flag was originally added to an individual's GP record through various routes, originally through NHS England's nationally applied algorithm[102] and then later by an individuals GP, hospital doctor or later by the Q-COVID-19 algorithm[107]. The definition of CEV changed throughout the pandemic, but originally these represented influenza at risk conditions such as organ transplant, certain types of bone cancer treatment, blood or bone marrow cancer, severe lung condition, medications that increase infections and pregnant with serious heart condition[100]. Other groups of individuals were added at a later stage, for example, individuals with Down's syndrome were added in November 2020[108]. Patients could have also had their CEV flag removed through different routes, for example if their condition improved overtime. If this was the case then their flag was updated from high-risk to low or medium-risk. Therefore, to identify someone with CEV status using EHRs, it was recommended to identify someone with a high-risk flag, without a more recent medium or low risk flag. It should be noted that CEV groups combined a heterogenous set of conditions, and

so a code indicating CEV status in UK primary care records may be of unclear and mixed relevance to the risk of infection.

3.3.3 Original study design: Test-negative-case-control design

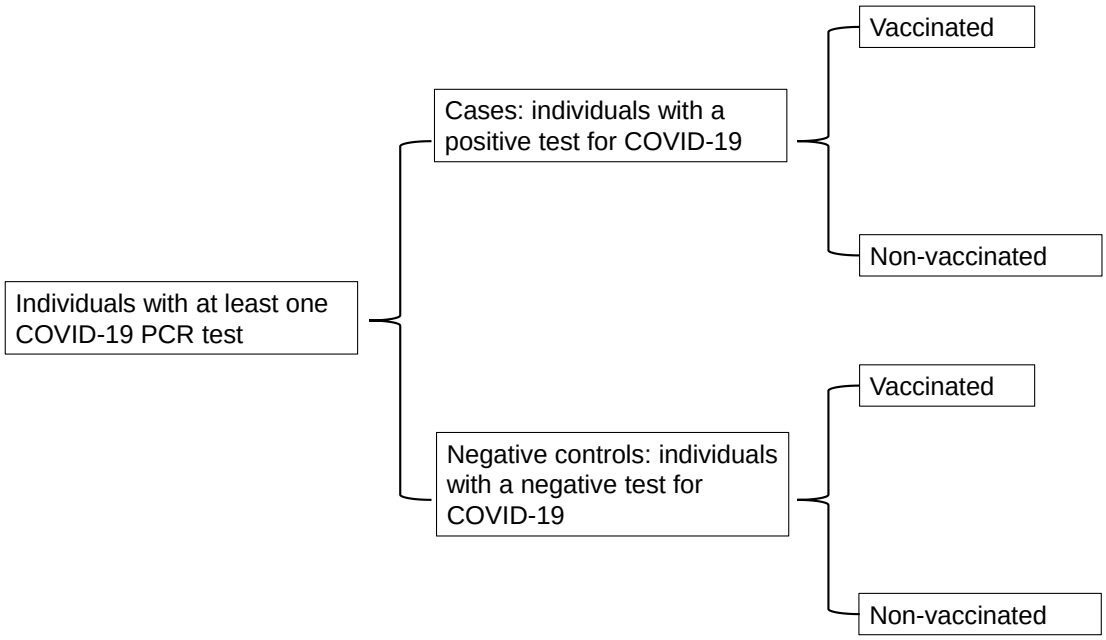
The original study[63] used a test-negative-case-control design (herein shorthanded to "test-negative design"). This design was originally developed by Jackson et al, 2013[76] for influenza vaccine effectiveness estimates to control for confounding from health-seeking behaviour, as well as control for differences in infection exposure. During the COVID-19 pandemic, it was a popular design, as the confounding structures and biases for COVID-19 were anticipated to be similar to influenza. The test-negative design is a case-control study that is conducted amongst a study population that seek care for symptoms of the vaccine preventable condition. For example, in the case of COVID-19, this could be all individuals that receive a PCR test for SARS-CoV-2. Cases are those that test positive for SARS-CoV-2 and controls are those that test negative. Odds ratios are estimated using this design by comparing the odds of vaccination amongst those that test positive compared with those that test negative. Vaccine effectiveness is derived from the odds ratios as (1-odds ratios) x 100. This study design aims to avoid confounding by health-seeking behaviour, as the study population is restricted to those who would seek care if they developed symptoms for the vaccine preventable disease[76]. This design requires many assumptions to be met. Some of the key assumptions are:

- **Similar non-COVID causes of acute respiratory infection:** the distribution of non-COVID-19 causes does not vary by COVID-19 vaccination status.
- **Similar health-seeking behaviour:** vaccine effectiveness does not vary by health-seeking behaviour status[76].

For an illustration of this design please see

Figure 3 below.

Figure 3 Illustration of the test-negative design



Abbreviations: PCR: polymerase chain reaction.

3.3.4 Original study population

The original study[63] identified all residents of England that were ≥70 years on 31 March 2021 who had a PCR COVID-19 test that occurred between 26 October 2020 to 21 February 2021. In the whole test-negative dataset, the data was restricted to the first positive test for a person since 26 October and there was only one positive test for each person. 26 October was selected as this was 6 weeks before the vaccination programme started on 8 December 2020. There were up to three randomly selected negative tests for a person in the whole test-negative dataset, which was also dated since 26 October. Negatives that occurred after a positive were removed, as were negatives within 3 weeks before a positive test result, or less than 7 days of a previous negative sample. The dataset only included individuals who were present in the NIMS denominator file

(which is all persons in England) so that they could be assigned a status of being vaccinated or unvaccinated at the time of symptom onset (89.9% of individuals).

All individuals were required to be symptomatic and test date was required to be within 0-10 days of symptom onset. They excluded individuals that had a positive PCR before 7 December 2020 (i.e., prior to COVID-19 vaccination implementation in the UK). Then they calculated the odds of vaccination amongst test negative controls compared with positive using logistic regression adjusted for age, sex, ethnicity, region, IMD, care home status and week of onset in the models (using variables from NIMS or SGSS).

3.3.5 Summary of findings and discussion from original study

The original study[63] estimated vaccine effectiveness after one dose of BNT162b2 to be 61% (95%CI: 51,69%) and after two doses of BNT162b2 in individuals aged ≥80-year-olds to be 89% (95%CI: 85,93%). After one dose of ChAdOx1 they estimated vaccine effectiveness to be to be 60% (95%CI: 41,73%) and this could not be estimated after 2 doses due to shorter follow-up in these individuals. These estimates were similar to estimates from clinical trials. For example, BNT162b2 vaccine efficacy after two doses in the trial was 93% (95%CI: 69,98)[109]. However, the original study authors were concerned about several different potential biases and alternative causal pathways (discussed in paper one below) being present in their study. For this reason, they sent out a questionnaire (see Section 0 below) to assess for potential biases that were potentially present in the original study. Data from this study, and subsequent studies using the same study design were being provided to The Joint Committee on Vaccination and Immunisation (JCVI) to inform policy decision making during the COVID-19 pandemic. Therefore, it was essential that these studies were robust to these potential biases and alternative causal pathways. My involvement in the study commenced after the questionnaire had been sent out and completed.

3.3.6 Current study: supplemented questionnaire data

The first study of my thesis also used data from NIMS and SGSS as described previously. This data was supplemented with data from a questionnaire.

3.3.6.1 Purpose of the questionnaire

Prior to the start of my thesis, UKHSA designed and sent out a questionnaire to a sub-sample of 23,963 individuals included within the original study[63]. The sub-sample of individuals included those that had a first PCR test that occurred from 1-21 February 2021. February was chosen as it covered a period where there were many infections with a range of vaccination statuses

(BNT162b2 vaccine or ChAdOx1 vaccine versus unvaccinated) to maximise statistical power. A more recent time period also ensured that any recall bias introduced through the questionnaire was minimised. The original purpose of the questionnaire was to collect additional information on comorbidities and health behaviours of individuals included in the original study. The questionnaire was sent out in March 2021, and all responses to the questionnaire were collected until August 2021.

The questionnaires were sent out as paper copies to the addresses of individuals and if email addressed could be found these were sent as electronic copies. Reminder paper copies and emails were sent to those that did not respond to the first questionnaire.

3.3.6.2 Data requested in the questionnaire

The questionnaire (see full questionnaire in Appendix B. Supplementary Materials Paper One) aimed to collect the necessary information required to assess for the presence of specific biases and alternative causal pathways between vaccination and infection. The questionnaire provided the test date of the first test (positive or negative) that occurred between 1 and 21 February 2021 and asked individuals to report specific information associated with this test date (e.g., symptomatic status). Other information that was collected in the questionnaire, was based on the date that the individual responded to the questionnaire. This included information on COVID-19 vaccination dates, COVID-19 risk factors (including comorbidities that would qualify an individual for COVID-19 'at-risk status'[110]), CEV status (see Section 3.3.2.1), care home status, household size and household type), time from vaccination invitation to vaccination (first and second dose), reasons for vaccination delay if vaccinated more than two weeks after invitation, reasons for no vaccination if not vaccinated, social mixing behaviours after vaccination (first and second dose), mode of transportation to vaccination centres (first and second dose), COVID-19 onset dates and symptomatic status. An overview of the information collected in the questionnaire and available responses are available in Table 4 below.

Table 4 Enhanced surveillance questionnaire data

Type of variable	Source	Values
Exposures		
COVID-19 vaccination brand	Questions 9, 12, 13, 16 and 17	0=BNT162b2 or 1=ChAdOx1 (if occurred)
COVID-19 vaccination date first dose	Questions 9, 12 and 16	Date of first vaccination (if occurred)
COVID-19 vaccination date second dose	Questions 9, 12 and 16	Date of second vaccination (if occurred)
Outcomes		

Symptomatic	Question 21, 22 and 24	0=asymptomatic; 1=symptomatic.
COVID-19 symptom onset date	Question 23	Date of symptom onset for identified test date (if symptoms reported)
COVID-19 severity	Question 25	1=mild, 2=moderate and 3=severe will be identified using the questionnaire (question 25). Depending on patient numbers, a scale of those that were hospitalised, admitted to accident and emergency (A&E) or were ventilated might be considered.
Covariates		
COVID-19 risk factors		
Chronic heart disease	Question 7	1=chronic heart disease
Chronic kidney disease	Question 7	1=chronic kidney disease
Chronic liver disease	Question 7	1=chronic liver disease
Chronic respiratory disease	Question 7	1=chronic respiratory disease
Asthma requiring medication	Question 7	1=asthma requiring medication
Cancer	Question 7	1=cancer
Organ or bone marrow transplant	Question 7	1=organ or bone marrow transplant
Human immunodeficiency virus (HIV)/immunodeficiency	Question 7	1=HIV/immunosuppression
Immunosuppression due to medication	Question 7	1=immunosuppression due to medication
Seizure disorder	Question 7	1=seizure disorder
Chronic neurological disease	Question 7	1=chronic neurological disease
Asplenia or dysfunction of the spleen	Question 7	1=asplenia or dysfunction of the spleen
Body Mass Index (BMI) ≥40 kg/m^2	Question 7	1=BMI => 40
CEV	Question 6	0=not CEV; 1=CEV
Household size	Question 5	1=live with no others; 2=live with one other; 2=live with 2 others; 4=live with 3 others; 5=live with 4 others; 6=live with 5 or more.
Frailty	Question 4	0=private home, other; 1=sheltered accommodation or nursing home
Behaviours		
Travelling to first vaccination	Question 14	0=walking/cycling or in a car with members of own household; 1=in a car with members from different household and public transport
Travelling to second vaccination	Question 18	As above
Mixing after first vaccine	Question 15	1=I've mixed with people outside my household of the same amount of time as I did before getting my vaccine; 2= I've mixed more with people outside my household after getting the vaccine; 3=

		I've mixed less with people outside of my household after getting the vaccine
Mixing after second vaccine	Question 19	As above
Riskier behaviours after vaccination	Question 38-40	0= not 1; 1= travel in car with someone outside household, indoors with others not from household, public transport.
First dose vaccine delay	Question 7	1=received the vaccine within 2 weeks of invitation; 2=received vaccine 2-3 weeks of invitation; 3=received the vaccine 4 or more weeks after vaccination
Reason for first dose delay	Question 11	1=I had my vaccine before I was eligible; 2I was not aware I was eligible; 3=No appointments available; 4=I prefer to wait to be vaccinated; 5=I delayed getting vaccination because I had COVID-19; 6=I was isolating and did not wish to leave home to get vaccinated; 7=I did not have time
Reason for no vaccination	Question 20	1=I have not been called for a vaccine; 2=I was not aware I was eligible; 3=there were no appointments available; 4=I would prefer not to get vaccinated at the moment; 5=I expect to get vaccinated soon but have not had a vaccine yet; 6=I am delaying getting vaccinated because I have been unwell or have had COVID-19 infection; 7=I am isolating and do not wish to leave home to get vaccinated; 8=I have not had time
Seasonal influenza flu	Question 41	0=no seasonal influenza vaccination; 1= seasonal influenza vaccination

Notes: for the full questionnaire see Appendix B. Supplementary Materials Paper One. Individuals were asked to recall information about their first test that occurred between 1 and 21 February 2021 and then all other information in the questionnaire was recalled based on latest information at the time they responded to the questionnaire.

Abbreviations: A&E: accident and emergency; BMI: body mass index; CEV: clinically extremely vulnerable; HIV: human immunodeficiency virus.

3.3.6.3 UKHSA role in the questionnaire

UKHSA developed the questionnaire themselves, they selected the study population that would be administered the questionnaire (more details in paper one below) and they sent these out over post or email. They were responsible for chasing individuals for their responses too. Once the questionnaires were returned UKHSA inputted all the data from the questionnaires into a CSV file. All patient identifiable information was removed. UKHSA cleaned all the NIMS and SGSS

data from the original study and created the questionnaire key study variables (as detailed in Table 4). After this I was responsible for all the data cleaning, data management and statistical analyses of the data according to how I wanted.

3.3.7 Open Science

For this study a transparent approach was adopted for the data study analyses. The data analyses was conducted in R and R Studio. All the R Scripts for the statistical analyses for the study one can be found on Github in this location: https://github.com/grahams99/Enhanced_surveillance_questionnaire.

3.4 Introduction to paper one

This paper was published in *Nature Communications* on 6 July 2023[111]. The aim was to identify and quantify the size and direction of potential biases that may have impacted estimates from one of the first UK COVID-19 vaccine effectiveness studies[63]. I used the original test-negative design and supplemented it with data from the questionnaire.

The potential biases that I assessed included COVID-19 vaccine exposure misclassification, outcome misclassification from symptomatic status, outcome misclassification from onset date, confounding from comorbidities and deferral bias. In addition, I investigated potential alternative causal pathways from vaccination to infection, including riskier behaviour after vaccination and attending vaccination visits being associated with COVID-19. For each of these biases I conducted a descriptive analysis, followed by logistic regression analyses to assess for the impact of each bias separately on the original estimates. Then, I conducted a final logistic regression model that accounted for all potential biases at the same time.

The paper can be found below. The supplementary materials from this paper can be found in Appendix B. Supplementary Materials Paper One.

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

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

RESEARCH PAPER COVER SHEET

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

SECTION A – Student Details

Student ID Number	2005987	**Title**	Ms
First Name(s)	Sophie		
Surname/Family Name	Graham		
Thesis Title	Advancing methods to account for biases in vaccine effectiveness research		
Primary Supervisor	Edward Parker		

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

SECTION B – Paper already published

Where was the work published?	Nature Communications		
When was the work published?	06/07/2023		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

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

SECTION C – Prepared for publication, but not yet published

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

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	The enhanced surveillance questionnaire was developed by Nick Andrews, Elise Tessier, Julia Stowe and Jamie Lopez Bernal at UKHSA. They created the questionnaire, identified the study population, sent the questionnaire out, sent reminders, collected the data from the questionnaire and inputted the data into a CSV file. This team had already previously created all the variables related to vaccination status (from the National Immunisation Management System) and infection status (from the Second Generation Surveillance System). I developed a detailed statistical analysis plan detailing my proposed analyses and this was reviewed by Helen McDonald and Jemma Walker (secondary supervisor at the time). I cleaned the data and conducted all statistical analyses outlined in the statistical analysis plan using R and R Studio. I wrote a report based on the findings from my analyses which was reviewed by Dr McDonald and Dr Walker. Then I wrote the first draft of this paper that was reviewed by Dr McDonald and Dr Walker. I updated the paper based on comments, sent it to all other authors for review and then updated the paper based on their comments. I submitted this paper and all required documents to Nature Communications. We received one round of reviewer comments, which I responded to and updated the analyses for. All authors reviewed the updated manuscript and responses to reviewer comments before re-submission when the manuscript was accepted. I uploaded all data management and data analysis files to Github.

SECTION E

Student Signature	Sophie Graham
Date	11/04/2024

Supervisor Signature	[redacted]
Date	24/04/2024

nature communications

Article

https://doi.org/10.1038/s41467-023-39674-0

Bias assessment of a test-negative design study of COVID-19 vaccine effectiveness used in national policymaking

Received: 23 December 2022

Accepted: 21 June 2023

Published online: 06 July 2023

Check for updates

Sophie Graham[1,2,3], Elise Tessier[2], Julia Stowe[2], Jamie Lopez Bernal[2], Edward P. K. Parker[1], Dorothea Nitsch[1,4,5], Elizabeth Miller[1,3], Nick Andrews[2,3], Jemma L. Walker[1,2,3,6] & Helen I. McDonald[1,3,6]

National test-negative-case-control (TNCC) studies are used to monitor COVID-19 vaccine effectiveness in the UK. A questionnaire was sent to participants from the first published TNCC COVID-19 vaccine effectiveness study conducted by the UK Health Security Agency, to assess for potential biases and changes in behaviour related to vaccination. The original study included symptomatic adults aged ≥70 years testing for COVID-19 between 08/12/2020 and 21/02/2021. A questionnaire was sent to cases and controls tested from 1–21 February 2021. In this study, 8648 individuals responded to the questionnaire (36.5% response). Using information from the questionnaire to produce a combined estimate that accounted for all potential biases decreased the original vaccine effectiveness estimate after two doses of BNT162b2 from 88% (95% CI: 79–94%) to 85% (95% CI: 68–94%). Self-reported behaviour demonstrated minimal evidence of riskier behaviour after vaccination. These findings offer reassurance to policy makers and clinicians making decisions based on COVID-19 vaccine effectiveness TNCC studies.

Test-negative-case-control (TNCC) observational studies are an important tool in the COVID-19 pandemic to monitor the continued real-world vaccine effectiveness of COVID-19 vaccinations against new variants and to assess the duration of protection[1–6]. In this design symptomatic individuals who present for testing for COVID-19 are included, categorised as cases if testing positive for COVID-19 and controls if testing negative. The design controls for confounding from health-seeking behaviour and healthcare access to some extent since both cases and controls are required to have accessed healthcare for COVID-19-like symptoms[7]. The design also controls for exposure because cases and controls have reported respiratory symptoms.

The UK Health Security Agency (UKHSA) has conducted regular COVID-19 vaccine effectiveness analyses in England using the TNCC design since vaccines were introduced in the UK in December 2020. The first published study included individuals in England aged ≥70 years who had a COVID-19 test in the community with self-reported symptoms and a symptom onset date between 8th December 2020 and 21st February 2021[4]. Patients were excluded if they had a history of a previous positive COVID-19 test from 26th October 2020 until 7th December 2020 to ensure vaccine effectiveness was assessed in those more likely to be susceptible. The study found that from 14 days from a second dose of BNT162b2 and from 14–20 days after a first dose of ChAdOx1 (i.e., the available COVID-19 vaccinations at the time), vaccine effectiveness reached 89% (95% confidence interval [CI]: 85–93%) and 60% (95% CI 41–73%), respectively, which was in line with vaccine efficacy estimates from clinical trials[8,9]. The work from this study informed governmental policy at the time[4,10] and the subsequent analyses have been used to provide regular updated estimates for national policy-makers[6].

[1]London School of Hygiene and Tropical Medicine, London, UK. [2]UK Health Security Agency, London, UK. [3]National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Vaccines and Immunisation, London, UK. [4]UK Renal Registry, Bristol, UK. [5]Renal Unit, Royal Free London NHS Foundation Trust, Hertfordshire, UK. [6]These authors contributed equally: Jemma L. Walker, Helen I. McDonald. e-mail: sophie.graham@lshtm.ac.uk

Although the TNCC design aims to control for confounding from the opportunity to be exposed, health-seeking behaviour and healthcare access, it does not implicitly control for other confounders of vaccine effectiveness and these need to be accounted for in the analysis. In the aforementioned UKHSA COVID-19 vaccine effectiveness study, it was only possible to adjust for potential confounders that were available in the national vaccination (National Immunisation Management Service (NIMS)) and COVID-19 testing (Second Generation Surveillance System (SGSS)) datasets that were utilised. Although this dataset includes some sociodemographic information such as age, gender, geographical region, index of multiple deprivation (IMD) and care home status, other key potential confounders such as detailed information on comorbidities[11] and household type[12] could not be identified in this dataset at the time (Fig. S1).

Riskier behaviour during or after vaccination may also result in real-world vaccine effectiveness estimates that are lower than the efficacy observed in randomised placebo-controlled clinical trials[13]. For example, during the national lockdowns, individuals who knew they were vaccinated may have assumed they were protected and might have therefore mixed more with individuals outside their household which would have increased their likelihood of exposure to SARS-CoV-2. They also might have had an additional risk of exposure compared to non-vaccinated individuals whilst travelling to or from, or even at, vaccination centres (Fig. S1).

The current study used a questionnaire that was sent out to a subsample of the original UKHSA TNCC COVID-19 vaccine effectiveness study in individuals aged 70 years and over. The aim of the current study was to use the questionnaire data to attempt to quantify the size and direction of potential biases that may have impacted estimates from one of the first UK COVID-19 vaccine effectiveness studies.

Results

Population description and selection bias

Among the 23713 individuals that made up the questionnaire sample, 8648 (36.5%) responded to the questionnaire ("respondents") and 15065 (63.5%) did not respond ("non-respondents"; Fig. 1). Among respondents, self-reported history of COVID-19 vaccination (one or two doses) at the time of questionnaire completion was high (Table S1).

Amongst the 8648 respondents, there were 6741 vaccinated (BNT162b2 = 3531 and ChAdOx1-S = 3210) and 1907 non-vaccinated at symptom onset date (based on SGSS onset date). Amongst the 8648 respondents there were 6541 negative controls and 2107 cases. When comparing respondents with non-respondents of the questionnaire there did appear to be some demographic and clinical differences, with respondents being younger, more likely to be of White ethnicity, less likely to live in a deprived area, and more likely to be a case when compared with non-respondents (based on a percentage absolute difference of +/−5% and p values of <0.05; Table S2). However, this selection bias did not appear to alter vaccine effectiveness estimates as after 2 doses of BNT162b2 vaccine, vaccine effectiveness in respondents (88% [95% CI: 79–94%]) was similar to non-respondents (87% [95% CI: 79–93%]) and the overall questionnaire sample (86% [95% CI: 79–91%]) (Fig. 2). There was insufficient follow-up to assess the effectiveness of two doses of ChAdOx1 vaccination, however, respondents had similar vaccine effectiveness from the first dose (14 days post-vaccination) than non-respondents (Table S3).

Respondents that were vaccinated were more likely to be a negative case, were older and were more likely to have later testing and symptoms compared with respondents that were non-vaccinated. Respondents that were cases were more likely to be non-vaccinated, more likely to be from the Northeast and Yorkshire and less likely to be from the Southwest of England, more likely to be deprived, and were more likely to have earlier testing compared with respondents that were negative controls (based on +/−5% percentage absolute difference and p values of <0.05; Table 1).

The results for the potential biases and alternative causal pathways in the original TNCC are detailed below and summarised in Fig. 3 and Table S4.

Potential biases in original TNCC study

When asking individuals in the questionnaire to report their vaccination date and comparing it to NIMS for the assessment of exposure misclassification, 9.5% (499/5276) of individuals reported a first dose vaccination date that was later in the questionnaire, whereas 7.3% (386/5276) reported a date that was earlier in the questionnaire (Fig. 3A). The remaining 83.2% (4391/5276) individuals reported the same date in NIMS and the questionnaire. The same pattern was seen for second-dose vaccinations. 89.8% of first doses self-reported in the questionnaire were within 3 days +/- of NIMS date (inclusive), whereas 3.6% were more than 3 days earlier and 6.6% were more than 3 days later in the questionnaire. For the second dose, 93.3% of self-reported vaccination dates were within 3 days +/− NIMS date, whereas 1.0% were more than 3 days earlier and 5.8% were more than 3 days later (Fig. S2A, B).

When updating vaccination dates to those self-reported in the questionnaire (or if missing using vaccination dates from NIMS), the percentage of individuals identified as non-vaccinated at symptom onset date (using SGSS) was very similar to when using NIMS (NIMS vaccination date: 22.1%; self-reported vaccination date: 23.5%). Vaccine effectiveness after two doses of BNT162b2 decreased from 88% (95% CI: 79–94%) to 84% (95% CI: 74–92%; Fig. 2). When exploring key

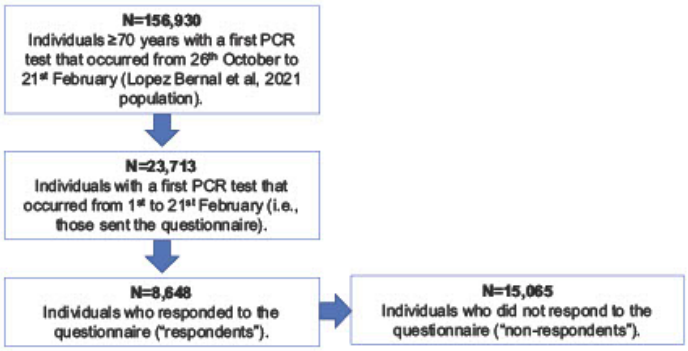


Fig. 1 | Cohort Selection. Abbreviations: PCR: polymerase chain reaction.

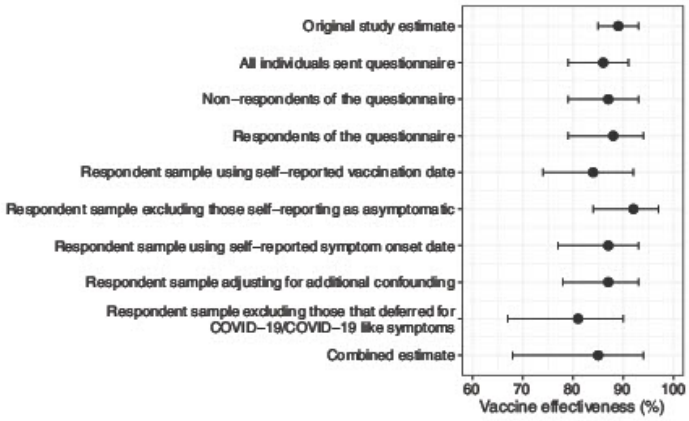


Fig. 2 | Forest Plot: Vaccine effectiveness estimates after two doses of BNT162b2. Vaccine effectiveness estimates after 2 doses of BNT162b2 in the following populations: the original TNCC study sample; all individuals who were sent the questionnaire; non-respondents of the questionnaire; respondents of the questionnaire; respondents using self-reported vaccination status; respondents using self-reported onset dates; respondents excluding those self-reporting as asymptomatic; respondents adjusting for additional confounding (for CEV, household size, and household type); respondents excluding those that delayed their vaccination because they had COVID-19/COVID-19 like symptoms; a combined estimate that accounted for all the above potential biases. All estimates adjusted for confounders that were adjusted for in the original TNCC study (age, gender, ethnicity, geography, index of multiple deprivations, care home status, and week of onset). Points represent odds ratio with corresponding 95% confidence intervals.

confounders amongst those with different self-reported vaccination status at symptom onset (from SGSS) versus unchanged status, we found that increased self-reported dose counts were associated with aged 75–79 years, most deprived IMD quintile, and COVID-19 symptom onset date in February week 1 but less associated with age 70–74 years (based on +/−5% percentage absolute difference and p values of <0.05; Table S5). On the other hand, decreased self-reported dose counts were associated with male gender, COVID-19 symptoms testing in February week 2 and were less associated with age 70–74 years, the 4th quintile of deprivation (with 5th being the lowest), and COVID-19 symptoms in February week 3 (based on +/−5% percentage absolute difference and p-values of <0.05; Table S5).

When asking individuals in the questionnaire to report their symptomatic status and comparing to SGSS for the assessment of outcome misclassification through symptomatic status, 65.5% (5539/8459) of the total population responding to this question reported that they were symptomatic despite all reporting they were symptomatic at the time of requesting their PCR test. Self-reported symptoms were 64.7% (4375/6741) and 67.4% (1285/1907) in vaccinated and non-vaccinated individuals, and 59.7% (3905/6541) and 83.5% (1759/2107) in negative controls and cases (Fig. 3B). When restricting to individuals who reported they had symptoms in the questionnaire vaccine effectiveness vaccine effectiveness for two doses of BNT162b2 increased from 88% (95% CI: 79–94%) in respondents of the questionnaire to 92% (95% CI: 84–97%; Fig. 2). Self-reported asymptomatic status was associated with older age and male sex, but no other key confounders (based on +/−5% percentage absolute difference and p-values of <0.05; Table S6).

When asking individuals in the questionnaire to report their symptom onset date and comparing to SGSS for the assessment of outcome misclassification through symptom onset date, 5.9% (514/8645) of individuals reported an earlier date, whereas 2.2% (194/8645) of individuals reported a later date in the questionnaire (Fig. 3C). 95.8% of these were within 3 days +/- of SGSS date (inclusive), whereas 3.0% were more than 3 days earlier and 1.2% were more than three days later in the questionnaire (Fig. S3).

When updating vaccination dates using self-reported onset dates, the percentage of non-vaccinated was very similar to when using SGSS (SGSS onset: 22.1%; self-reported onset: 22.6%). Vaccine effectiveness after two doses of BNT162b2 decreased marginally from 88% (95% CI: 79–94%) to 87% (95% CI: 77–93%; Fig. 2). The prevalence of confounders did not differ among individuals with differing versus unchanged self-reported symptom onset date (based on +/−5% percentage absolute difference and p-values of <0.05; Table S7).

For the assessment of confounding, when adjusting for COVID-19 risk factors self-reported in the questionnaire (household size, household type and CEV) in addition to the variables in the original study (age, gender, ethnicity, geography, index of multiple deprivations, care home status and week of onset), vaccine effectiveness estimates after two doses of BNT162b2 decreased marginally from 88% (95% CI: 79–94%) to 87% (95% CI: 78–93%; Fig. 2). Due to the later approval of ChAdOx1-S, there were insufficient individuals with two doses at symptom onset date (N = 5) for the same assessment to be made.

Individuals with COVID-19-like symptoms, recent exposure to COVID-19 or a positive SARS-CoV-2 test just before their vaccination date were recommended to defer their vaccination by 28 days according to government guidelines[14]. This deferral has the potential to increase vaccine effectiveness estimates as individuals that defer their vaccination, for this reason, might go on to test positive for SARS-CoV-2 (inflating cases among non-vaccinated individuals). In the current study, among all individuals who reported in the questionnaire that they had been vaccinated at the time of survey (8518/8613 98.9%), 9.6% (794/8251) delayed their vaccination ≥4 weeks from the invitation. Among these individuals, 25.3% (201/794) reported they delayed vaccination because of COVID-19/COVID-19 symptoms. Of all individuals who reported in the questionnaire that they had not been vaccinated (95/8613 1.2%), over a quarter (26/95 27.4%) of individuals reported that this was because they had been unwell or because they had COVID-19 (Fig. 3D and Table S2). Thus, some individuals appeared to be deferring their vaccinations because they were unwell or because they had COVID-19. When assessing for the potential impact of deferral bias, amongst those who didn't delay vaccination because of COVID-19/COVID-19 like symptoms (N = 8396), vaccine effectiveness after two doses of BNT162b2 decreased from 88% (95% CI: 79–94%) to 81% (95%

Table 1 | Baseline characteristics of respondents, by vaccination and case status, using variables from the original study data (NIMS and SGSS)

Characteristics according to NIMS and SGSS	Respondents not vaccinated at symptom onset, *N* = 1907	Respondents vaccinated at symptom onset, *N* = 6741	Difference in absolute percentage (vaccinated – non-vaccinated)	p-value	Respondents who were negative controls, *N* = 6541	Respondents who were cases, *N* = 2107	Difference in absolute percentage (cases – negative controls)	p-value
Vaccine status at symptom onset, *n* (%)								<0.001
Not vaccinated					1351 (20.7%)	556 (26.4%)	5.70%	
Vaccinated					5190 (79.3%)	1551 (73.6%)	−5.70%	
Test result				<0.001				
Negative	1351 (70.8%)	5190 (77.0%)	6.2%					
Positive	556 (29.2%)	1551 (23.0%)	−6.2%					
Age group in years, *n* (%)				<0.001				0.626
70–74	1492 (78.2%)	2931 (43.5%)	−34.7%		3348 (51.2%)	1075 (51.0%)	−0.20%	
75–79	279 (14.6%)	2056 (30.5%)	15.9%		1778 (27.2%)	557 (26.4%)	−0.80%	
80–84	57 (3.0%)	1031 (15.3%)	12.3%		826 (12.6%)	262 (12.4%)	−0.20%	
85–89	43 (2.3%)	473 (7.0%)	4.7%		381 (5.8%)	135 (6.4%)	0.60%	
=>90	36 (1.9%)	250 (3.7%)	1.8%		208 (3.2%)	78 (3.7%)	0.50%	
Gender, *n* (%)				0.421				0.115
Female	1081 (56.7%)	3749 (55.6%)	−1.1%		3685 (56.3%)	1145 (54.3%)	−2.00%	
Male	826 (43.3%)	2992 (44.4%)	1.1%		2856 (43.7%)	962 (45.7%)	2.00%	
Ethnicity, *n* (%)				0.075				<0.001
White	1756 (92.1%)	6266 (93.0%)	0.9%		6114 (93.5%)	1908 (90.6%)	−2.90%	
Non-white	84 (4.4%)	224 (3.3%)	−1.1%		185 (2.8%)	123 (5.8%)	3.00%	
Prefer not to say	67 (3.5%)	251 (3.7%)	0.2%		242 (3.7%)	76 (3.6%)	−0.10%	
Geographical region, *n* (%)				<0.001				<0.001
East of England	246 (12.9%)	814 (12.1%)	−0.8%		834 (12.8%)	226 (10.7%)	−2.10%	
London	104 (5.5%)	614 (9.1%)	3.6%		530 (8.1%)	188 (8.9%)	0.80%	
Midlands	360 (18.9%)	1415 (21.0%)	2.1%		1265 (19.3%)	510 (24.2%)	4.90%	
Northeast and Yorkshire	317 (16.6%)	1043 (15.5%)	−1.1%		948 (14.5%)	412 (19.6%)	5.10%	
Northwest	232 (12.2%)	994 (14.7%)	2.5%		922 (14.1%)	304 (14.4%)	0.30%	
Southeast	394 (20.7%)	1116 (16.6%)	−4.1%		1202 (18.4%)	308 (14.6%)	−3.80%	
Southwest	254 (13.3%)	745 (11.1%)	−2.2%		840 (12.8%)	159 (7.5%)	−5.30%	
IMD quintile, *n* (%)				0.085				<0.001
1 (most deprived)	238 (12.5%)	800 / 6736 (11.9%)	−0.6%		683 / 6536 (10.4%)	355 (16.8%)	6.40%	
2	319 (16.7%)	1018 / 6736 (15.1%)	−1.6%		963 / 6536 (14.7%)	374 (17.8%)	3.10%	
3	399 (20.9%)	1425 / 6736 (21.2%)	0.3%		1379 / 6536 (21.1%)	445 (21.1%)	0.00%	
4	477 (25.0%)	1622 / 6736 (24.1%)	−0.9%		1645 / 6536 (25.2%)	454 (21.5%)	−3.70%	
5 (least deprived)	474 (24.9%)	1871 / 6736 (27.8%)	2.9%		1866 / 6536 (28.5%)	479 (22.7%)	−5.80%	
Missing	0	5			5	0		
Week of symptom onset, *n* (%)				<0.001				<0.001
January week 1	10 (0.5%)	<5			10 (0.2%)	<5		
January week 2	32 (1.7%)	<5			35 (0.5%)	<5		
January week 3	86 (4.5%)	61 (0.9%)	−3.6%		110 (1.7%)	37 (1.8%)	0.10%	
January week 4	737 (38.6%)	987 (14.6%)	−24.0%		1238 (18.9%)	486 (23.1%)	4.20%	
February week 1	802 (42.1%)	2202 (32.73%)	−9.4%		2203 (33.7%)	801 (38.0%)	4.30%	
February week 2	178 (9.3%)	2202 (32.7%)	23.4%		1839 (28.1%)	541 (25.7%)	−2.40%	
February week 3	62 (3.3%)	1280 (19.0%)	15.7%		1106 (16.9%)	236 (11.2%)	−5.70%	
Week of COVID-19 test, *n* (%)				<0.001				<0.001
February week 1	1385 (72.6%)	2166 (32.1%)	−40.5%		2572 (39.3%)	979 (46.5%)	7.20%	
February week 2	362 (19.0%)	2038 (30.2%)	11.2%		1772 (27.1%)	628 (29.8%)	2.70%	

Table 1 (continued) | Baseline characteristics of respondents, by vaccination and case status, using variables from the original study data (NIMS and SGSS)

Characteristics according to NIMS and SGSS	Respondents not vaccinated at symptom onset, *N* = 1907	Respondents vaccinated at symptom onset, *N* = 6741	Difference in absolute percentage (vaccinated – non-vaccinated)	p-value	Respondents who were negative controls, *N* = 6541	Respondents who were cases, *N* = 2107	Difference in absolute percentage (cases – negative controls)	p-value
February week 3	160 (8.4%)	2537 (37.6%)	29.2%		2197 (33.6%)	500 (23.7%)	-9.90%	
Care home status, *n* (%)				0.059				<0.001
Not care home	1901 (99.7%)	6691 (99.3%)	-0.4%		6514 (99.6%)	2078 (98.6%)	-1.00%	
Care home[a]	6 (0.3%)	50 (0.7%)	0.4%		27 (0.4%)	29 (1.4%)	1.00%	
CEV, *n* (%)				<0.001				0.006
Not CEV	1709 (89.6%)	5746 (85.2%)	-4.4%		5601 (85.6%)	1854 (88.0%)	2.40%	
CEV	198 (10.4%)	995 (14.8%)	4.4%		940 (14.4%)	253 (12.0%)	-2.40%	

Abbreviations: *CEV* clinically extremely vulnerable, *IQR* interquartile range, *IMD* index of multiple deprivations, *n* numerator N denominator, *NIMS* National Immunisation Management System, *SGSS* Second Generation Surveillance System.
[a]Care home status is likely low in the current study because the study only included those tested in the community (pillar 2), individuals tested in care homes or in hospital are usually tested under pillar 1.
Note: all tests were conducted using two-sided Chi squared test.
Note: cells <5 have been suppressed and secondary suppression has also been conducted in order to protect patient privacy.

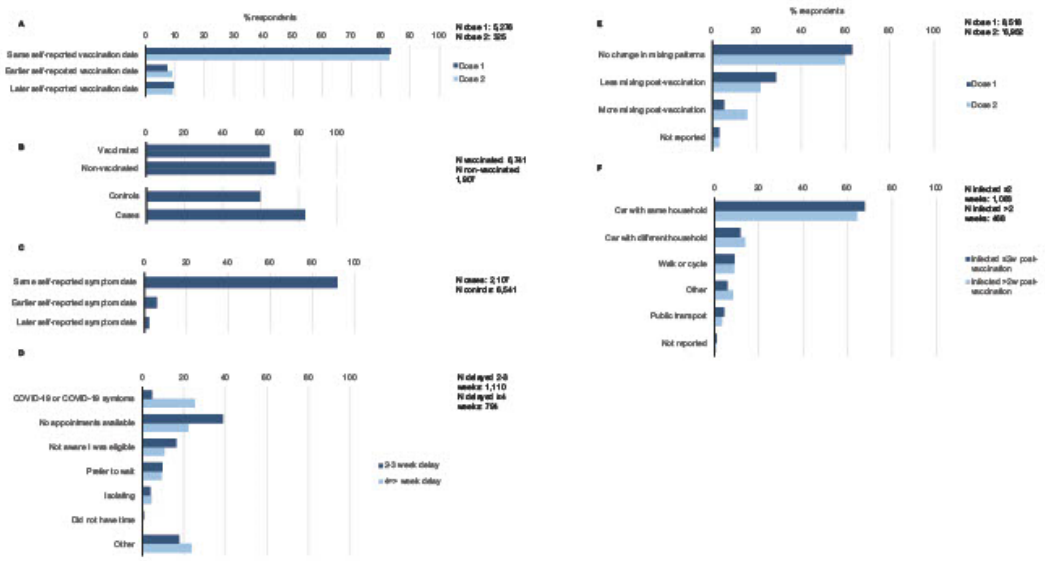


Fig. 3 | Summary of results for assessment of bias and alternative causal pathways. A Exposure (vaccination status) misclassification: comparison of vaccination dates in NIMS versus the questionnaire, by vaccination dose. **B** Outcome misclassification by symptomatic status: individuals reporting in the questionnaire they were symptomatic, by case and vaccination status. **C** Outcome misclassification by symptom onset date: comparison of onset dates in SGSS versus the questionnaire. **D** Deferral bias: individuals reporting they delayed their vaccination because they had COVID-19 or COVID-19 like symptoms, by length of vaccine delay from invitation. **E** Riskier behaviour after vaccination: individuals reporting they mixed more, the same or less after their vaccination, by vaccination dose. **F** Vaccination visit transit mode in relation to COVID-19: individuals reporting their mode of transport to the vaccination centre, amongst those with a positive COVID-19 test ≤2 weeks versus those with a positive COVID-19 test >2 weeks.

CI: 67–90%; Fig. 2). When assessing the vaccination status of those that deferred 2–3 or 4 weeks because of COVID-19/COVID-19 like symptoms we found 96% were non-vaccinated, 4.0% had one dose and 0.0% two doses.

When accounting for all of the above potential biases in the original TNCC study, vaccine effectiveness after two doses of BNT162b2 decreased slightly from 88% (95% CI: 79–94%) to 85% (95% CI: 68–94%; Fig. 2).

Potential alternative causal pathways in original TNCC study

When asking individuals in the questionnaire if they mixed more after their vaccination for assessment of alternative causal pathways via riskier behaviour after vaccination, 5.2% (445/8518) with a first dose and 15.6% (1087/6952) with a second dose, reported that they mixed more after their vaccinations, whereas the remaining 91.6% (7806/8518) for first dose and 81.4% (5658/6952) for second dose reported that they mixed the same or less (Fig. 3E and Table S2). Amongst those

that were vaccinated before symptom onset date, there was no association between mixing more and odds of COVID-19 when compared with those that reported mixing less or the same after any first dose vaccination (odds ratio [OR]: 0.92 95% CI: 0.68–1.24) after adjusting for age, gender, ethnicity, CEV, immunosuppressive conditions and month of vaccination dose. Due to a lack of individuals the same could not be assessed after second dose vaccinations.

When asking individuals in the questionnaire their mode of transport to vaccination centres, for assessment of alternative causal pathways for contracting COVID-19 individuals with a positive test within 2 weeks of vaccination did not appear to take risker types of transport compared to those that had a positive test after 2 weeks of vaccination (within 2 weeks: public transport: 4.5% [21/468], car with member outside of household: 11.8% [55/468]; after 2 weeks: 3.5% [38/1083] and 13.9% [151/1083], respectively; Fig. 3F and Table S2). Amongst those that were vaccinated before symptom onset date, there was no association between riskier transport to the vaccination centre and odds of COVID-19 (car with members outside household: OR: 1.28 95% CI: 0.98–1.67; public transport: OR: 1.26 95% CI: 0.81–2.03) when compared with those that took less risk forms of transport (drove alone or walked/cycled) after adjusting for age, gender, ethnicity, region and IMD. Therefore, there appeared to be no or minimal evidence of alternative causal pathways through riskier behaviour after vaccination or vaccination itself being associated with COVID-19 in the original TNCC study.

Discussion

Among 23713 symptomatic individuals with a positive PCR test between 1 and 21 February 2021 in England, 8648 responded to a questionnaire to assess for potential bias in an influential TNCC study of COVID-19 vaccine effectiveness. Using information from the questionnaire to produce a combined estimate that accounted for all potential biases decreased the original vaccine effectiveness estimate after two doses of BNT162b2 from 88 to 85%. Self-reported behaviour demonstrated no or minimal evidence of riskier behaviour after vaccination.

The response rate to the questionnaire was lower in those with a negative PCR test, and also among people living in areas of greater deprivation or with non-White ethnicities, both associated with increased risk of COVID-19 related death[15]. However, there was a similar response rate by vaccination status, and vaccine effectiveness point estimates were very similar in the respondent versus non-respondent and overall questionnaire samples, suggesting the selection bias did not materially affect the vaccine effectiveness estimates.

Vaccination dates and symptom onset dates were consistent between the nationwide vaccination-COVID-19 PCR testing data (NIMS-SGSS) and the questionnaire with the majority of individuals self-reporting the same date as in NIMS-SGSS. When using self-reported vaccination dates, vaccine effectiveness decreased from the original estimate of 88 to 84%. When using self-reported onset dates vaccine effectiveness decreased marginally from 88 to 87%. For vaccinations other than for COVID-19 self-reported dates have previously been shown to be unreliable[16], however in the UK, individuals were asked to carry their COVID-19 vaccination cards[17] which could explain why self-reported vaccination dates were more reliable than expected. Vaccination status using self-reported dates was more likely to be different to vaccination status when using NIMS when age increased. This likely represents the greater impact of recall bias (i.e., the questionnaire was sent in March 2021 and individuals were still responding in August 2021 and it is likely that responses to this question became more unreliable with increasing number of days between the event occurring and response to the questionnaire) in older individuals[18]. For onset date, we likely underestimated misclassification since individuals were only asked to report their onset date in the questionnaire if different from the SGSS date that was provided. It is likely that some individuals could not recall the date and left this field blank, which would have been inaccurately determined as the correct date, rather than missing.

Somewhat surprisingly, only 65.5% of individuals self-reported that they were symptomatic in the questionnaire, despite all having to be symptomatic at the time of requesting their PCR test. Cases were more likely to report being symptomatic in the questionnaire compared with negative controls, which resulted in a modest increase in vaccine effectiveness estimates (88 to 92%) when self-reportedly asymptomatic individuals were excluded. These findings may reflect a degree of outcome misclassification in the original study. They may also indicate a retrospective reassessment of symptom status by survey participants, including the downgrading of symptoms among individuals whose SARS-CoV-2 test was negative. Studies have previously found that specific comorbidities[15], household size and type[19,20] are highly associated with COVID-19. In the current study it was reassuring that when adjusting for individual CEV, household size and type, the vaccine effectiveness estimates decreased marginally from the original estimate of 88 to 87% providing limited evidence of confounding from these COVID-19 risk factors in the original TNCC study. When using these data to assess COVID-19 effectiveness early on in the pandemic, we can be more confident that missing information on these COVID-19 risk factors was less of a concern, although there could be confounding from other variables that were not collected in the questionnaire (e.g., mobility status). Factors such as occupation may also be key confounders in younger adults, but were assumed to be less relevant in the current study given our focus on individuals over 70 years of age.

Vaccine deferral because of COVID-19/COVID-19 symptoms was relatively common in the study. When we excluded individuals who deferred their vaccination because of COVID-19/COVID-19 like symptoms, vaccine effectiveness estimates decreased from the original estimate of 88 to 81%. A decrease in estimated effectiveness is expected given that this approach entails removing non-vaccinated individuals who received a positive COVID-19 test from the analysis. The effect of deferral bias appears to be modest and does not undermine conclusions from the original TNCC study regarding the high effectiveness of vaccines during the initial phases of implementation.

The combined vaccine effectiveness estimate that accounted for all potential biases saw a modest decrease in effectiveness from the original estimate of 88 to 85%. Although this small change is reassuring, 85% should not be considered a best estimate since questionnaire responses that were provided in some cases many months after the events occurred cannot be considered the gold standard.

The findings on riskier behaviour are interesting. Previously authors have suggested that information on individuals' risk behaviours and exposures should be collected when conducting vaccine effectiveness studies[13]. However, our study findings suggest that during the early stages of the pandemic in England in the elderly population, when the country was in lockdown there was low prevalence of risky behaviours following vaccination. Self-reported riskier behaviours might be susceptible to underreporting due to the impact of social desirability bias (wherein people are more likely to report behaviour in line with rules and recommendation). The lack of a significant association between mixing more after vaccination and the odds of COVID-19 may also reflect recruitment bias within the test-negative study population, whereby riskier behaviour increases exposure to both positive and negative (non-COVID-19) causes of symptoms. It would be beneficial to verify these analyses with other study designs. The widespread implementation of precautionary measures such as mask usage and physical distancing on public transport and vaccine centres may account for the lack of association between riskier transit types and COVID-19 risk.

A strength of the study was the large questionnaire sample size that meant the sample was fairly generalisable and allowed the identification of small differences in vaccine effectiveness estimates. This

study addressed an important evidence gap: previous literature[21–28] used theoretical proofs or simulated data to show the impact of substantial bias on different observational study designs. However, the current study used real-world data to detect the presence or absence of each of these biases and then quantified the true impact on vaccine effectiveness estimates. Another key strength of the study was that it assessed the robustness of an influential TNCC study that was one of the first observational studies that was used to inform governmental policy at the start of the pandemic[10].

However, despite these strengths, there were also a number of limitations. The study behavioural findings (e.g., mixing patterns) are likely only relevant to the period of time when the national population was under a strict lock down. Later on in the pandemic, when restrictions started to be lifted and individuals became "fatigued", it is likely that mixing patterns would be different to those identified in the current study[29]. Another limitation was that we were unable to assess whether collider bias[30–32] was present in the original study. Collider bias, another form of selection bias, could have potentially been introduced through the test-negative design. This type of bias is potentially introduced as health-seeking behaviour is associated with testing, vaccination uptake and infection i.e., testing is a 'collider' on the pathway between vaccination and infection[30,31,33]. We could not determine the presence of this bias because the association between health-seeking behaviour and testing could not be assessed as this information is not recorded in the data. Future studies should collect information on health-seeking behaviour so that this association can be assessed.

Based on the findings from the current study, policy makers can be more confident in their decisions made and other policy decisions that were made using the same study design in this population early on in the pandemic. Similarly, this study provides some reassurance on ongoing national vaccine effectiveness studies using TNCC with the same data sources, to support the public and healthcare workers to have continued confidence in reports of vaccine effectiveness e.g., against new strains. Future studies are required to determine whether the current findings remain applicable now that restrictions have been lifted.

Overall, there appeared to be minimal evidence of any large biases that may have affected an important TNCC COVID-19 vaccine effectiveness study that informed governmental policy early in the COVID-19 pandemic. Based on this, clinicians and policy makers can be more confident in any decisions that were made based on this study and in the TNCC studies that were conducted throughout the pandemic to assess vaccine effectiveness against new variants and to assess duration of protection of the vaccinations.

Methods

Data sources

The original study used national vaccination (NIMS) and COVID-19 testing in the community (pillar 2; SGSS) data, linked at the patient level. Details on the variables available at the time of the original analysis can be found in Table S5.

The new questionnaire data was used in combination with the NIMS and SGSS data used in the original study[4]. Data from these sources were linked on the patient level. The questionnaire was sent in March 2021 to the subset of individuals from the original TNCC study[4] who had a PCR COVID-19 test between 1 February 2021 and 21 February 2021. The most recently tested individuals were selected in order to minimise the impact of recall bias. The questionnaire (Materials S1) aimed to collect the necessary information required to assess for the presence of specific biases and behavioural changes related to vaccination. This included COVID-19 vaccination dates, COVID-19 risk factors (including comorbidities that would qualify an individual as high risk for COVID-19[11], CEV status[34], care home status, household size and household type), time from vaccination invitation to vaccination (first and second dose), reasons for vaccination delay if vaccinated more than two weeks after invitation, reasons for no vaccination if not vaccinated, social mixing behaviours after vaccination (first and second dose), mode of transportation to vaccination centres (first and second dose), COVID-19 onset dates and symptomatic status. The COVID-19 testing and symptom date of interest were specified in the survey letter (Materials S1).

Study analyses

Population description and selection bias. To assess whether the questionnaire responses were representative of all ≥70 year olds in England that had their first PCR test in February 2021 the demographics and clinical characteristics (age, gender, ethnicity, geographical region, IMD, week of COVID-19 symptom onset, week of COVID-19 test, care home status, test result, CEV status and COVID-19 vaccination status) of respondents of the questionnaire were described on 31 March 2021 and compared with non-respondents using percentage difference (with +/−5% absolute difference as threshold to define clinically meaningful differences) and Chi-squared/Fisher's exact test. Any missing data was described.

To assess whether potential selection bias had been introduced through the questionnaire sampling or response, the original vaccine effectiveness estimates (i.e., the odds of vaccination in cases and negative controls estimated using logistic regression models adjusted for potential confounders that were available in the data at the time: age, gender, ethnicity, geography, index of multiple deprivation (IMD), care home status and week of onset) were run on the entire questionnaire sample and then amongst those that responded ("respondents") and did not respond ("non-respondents"). As in the original study vaccine effectiveness was estimated as (1 − odds ratio) x 100.

Respondents that were vaccinated were compared to non-vaccinated and cases were compared to negative controls based on demographics and clinical characteristics. These were compared using percentage difference (with +/−5% absolute difference as a threshold to define clinically meaningful differences) and Chi-squared/Fisher's exact test and missing data was also described.

Sources of potential bias and behavioural changes related to vaccination were then explored among questionnaire respondents as outlined below.

Potential biases in original TNCC study. Vaccination status could be misclassified in NIMS if vaccination dates are incorrect (Fig. S1). The questionnaire, therefore, asked participants to self-report their vaccination date to identify any exposure misclassification. The number of individuals with the same, earlier or later self-reported vaccination date compared with NIMS was described as well as the distribution in difference in days using histograms for both doses. We also described the number of self-reported vaccination dates that were within 3 days +/− of NIMS (inclusive) or more and less than 3 +/− days for both doses.

We updated vaccination status using self-reported vaccination date, and if this field was missing in the questionnaire, we used the NIMS date. Amongst this population, we reported vaccination status based on self-reported vaccination dates and to assess for the potential impact of exposure misclassification on vaccine effectiveness estimates, we ran the logistic regression models from the original study (see above) using self-reported vaccine dates. To explore the potential mismeasurement of exposure misclassification within levels of confounders we described key confounders (age, gender, ethnicity, geography, index of multiple deprivation (IMD), week of onset, care home status and CEV) amongst those identified with increased or decreased number of vaccine dose counts when using self-reported vaccine dates (versus NIMS) compared to those with no change in vaccine status. These were compared using percentage difference (with +/−5% absolute difference set as threshold) and Chi-squared/Fisher's exact test.

The symptomatic status could be misclassified in SGSS if individuals incorrectly report they are symptomatic at the time of requesting their PCR test either to access free testing or because they are concerned about mild/vague symptoms (Fig. S1). This could also have affected the selection of the study population, since only symptomatic individuals were eligible for inclusion. Therefore, to assess for potential outcome misclassification through symptomatic status, the self-reported symptomatic status in the questionnaire was compared to the status reported in SGSS. Since all individuals identified in SGSS reported that they were symptomatic, the proportion of this population that reported they were asymptomatic in the questionnaire overall and by case and vaccination status was reported. The denominator in this population was all those responding to the symptomatic status question in the questionnaire. The logistic regression models from the original study (see above) were re-run amongst the population of individuals that reported they were symptomatic in the questionnaire. To explore the potential mismeasurement of outcome misclassification within levels of confounders we described key confounders (as above) amongst those self-reporting asymptomatic versus symptomatic status. These were compared using percentage difference (with +/−5% absolute difference set as threshold) and Chi-squared/Fisher's exact test.

The onset date could be misclassified in SGSS if individuals incorrectly reported their symptom onset date when booking their PCR test (Fig. S1). Individuals that reported they were symptomatic in the questionnaire were asked to report their symptom onset date (if different from the date in SGSS which was provided in the questionnaire) to assess for systematic differences. The number of individuals with the same, earlier or later self-reported onset date compared with SGSS was described as well as the distribution in difference in days using a histogram. We also described the number of self-reported onset dates that were within 3 days +/− of SGSS (inclusive) or more and less than 3 +/− days.

Vaccination status using self-reported symptom onset date from the questionnaire was updated and amongst this population, we reported vaccination status and ran the logistic regression models from the original study (see above). However, this would be interpreted with caution a priori because of the potential impact of recall bias[35]. To explore the potential mismeasurement of outcome misclassification within levels of confounders we described key confounders (as above) amongst those self-reporting different versus same onset date in the questionnaire. These were compared using percentage difference (with +/−5% absolute difference set as threshold) and Chi-squared/Fisher's exact test.

Confounding from COVID-19 risk factors was potentially present in the original study since it was not possible at the time to identify comorbidities and other risk factors for COVID-19 (e.g., household size and type) using NIMS and SGSS (composite variables including any risk group and CEV, have since been added but individual conditions remain unavailable in these datasets; Fig. S1). Therefore, to assess for potential confounding, the logistic regression models from the original study (see above) were repeated additionally adjusting for each potential COVID-19 risk factor in turn obtained from the questionnaire, including: CEV; the number of persons per household; household type; immunosuppression (separately and combined: HIV/immunodeficiency, organ or bone marrow transplant, immunosuppression due to medication and asplenia or dysfunction of the spleen); and other comorbidities that qualify an individual as high risk (separately and then combined: chronic heart disease, chronic kidney disease, chronic respiratory disease excluding asthma, cancer, seizure disorder, chronic liver disease, asthma requiring medication, chronic neurological disease and BMI ≥ 40 kg/m^2). The pre-specified analysis plan was to include all variables which changed the odds ratio of vaccination by 0.01 amongst the PCR-confirmed individuals in a multivariable model.

Deferral bias[36–38] is potentially introduced if individuals delay their vaccinations because they have a COVID-19 infection, COVID-19 like symptoms or have been recently exposed to COVID-19 (individuals in the UK are asked to delay their vaccine by 28 days if they contract COVID-19[14]; Fig. S1). Individuals that decide to defer their vaccination because of this might then go on to test positive for COVID-19 which leads to a temporary apparent protective effect of the vaccination in recently vaccinated individuals[36]. Therefore, to assess for potential deferral bias the proportion of individuals who reported they received their vaccinations ≥4 weeks from their invitation because they had COVID-19 or COVID-19 like symptoms was reported and the proportion of individuals that reported they had not yet been vaccinated because they had been unwell or had COVID-19 was also reported. The denominator population was all individuals reporting they were ever vaccinated with a first dose or second dose in the questionnaire. To assess by how much deferral bias might be expected to increase vaccine effectiveness estimates, we ran the logistic regression models from the original study (see above) removing individuals that reported they delayed either 2–3 weeks or 4 weeks because of COVID-19/COVID-19 like symptoms. We also described the vaccination status at symptom onset date of those that deferred their vaccination 2–3 or 4 week because of COVID-19/COVID-19 like symptoms.

When accounting for all biases at once, we ran the logistic regression models from the original study (see above) amongst those that did not delay their vaccination because of COVID-19/COVID-19 like symptoms, that self-reported they were symptomatic and using vaccination and symptom onset dates from the questionnaire adjusting for CEV, household size and type (as well as confounders adjusted for in the original TND study; Fig. 2).

Potential alternative causal pathways in original TNCC study. If vaccinated individuals start mixing more with individuals outside of their household after being vaccinated, then the risk of contracting COVID-19 might increase in these individuals creating an "alternative causal pathway" from vaccination to infection (Fig. S1). If increased mixing occurs at a faster rate compared to non-vaccinated individuals' then this could lower vaccine effectiveness estimates compared to true estimates. To assess for riskier behaviour after vaccination the proportion of those that reported that they mixed the same, more or less in the 3–4 weeks after the date of their first or second vaccination was reported. Amongst those that were vaccinated before the symptom onset date, the odds of COVID-19 amongst those that reported they mixed more were compared to those that mixed less or the same, using logistic regression adjusting for potential confounders (age, gender, ethnicity, CEV, immunosuppressive conditions and month of vaccination dose).

Other alternative causal pathways are potentially introduced if individuals' contract COVID-19 on the way to or back from their vaccination centres (Fig. S1). These pathways include a composite of events immediately before and after vaccination, though in practice all exposures would precede the induction of robust vaccine immunity. Exposures at the time of vaccination could have potentially lowered vaccine effectiveness estimates compared to true vaccine effectiveness estimates, especially early on in the pandemic when individuals were instructed to stay at home if they were not carrying out certain tasks (e.g., going to get vaccinated, food shopping etc.). To assess for travel to the vaccination itself being associated with COVID-19 the mode of transport to and from vaccination centres (first and second dose) was reported amongst those that had a positive COVID-19 test within 2 weeks of vaccination, compared with those that had a positive test after 2 weeks. Amongst those who were vaccinated before symptom onset date, the odds of COVID-19 amongst those who travelled to and from their vaccination centre in a car with someone outside of their household or on public transport (i.e., riskier transport modes) was compared to those that travelled either alone in a car or walked/cycled (i.e., less risky transport modes) using logistic regression

adjusted for age, gender, ethnicity, region and IMD, since these variables were likely to be associated with mode of transportation and COVID-19 risk.

All of the analyses were conducted using Stata (version 17) and R (version 4.1.3).

Ethics. This analysis was conducted as part of public health service evaluation. UKHSA has legal permission, provided by Regulation 3 of The Health Service (Control of Patient Information) Regulations 2002 to process patient confidential information for national surveillance of communicable diseases and as such, individual patient consent is not required to access records. Research ethics approval was therefore not sought.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Access to pseudonymised national datasets used in this study (National Immunisation Management Service and Second Generation Surveillance System) is managed by NHS England through the NHS COVID-19 Data Store: https://www.england.nhs.uk/contact-us/privacy-notice/how-we-use-your-information/covid-19-response/nhs-covid-19-data-store/. Questionnaire data was collected for the purposes of public health service evaluation and consent was not obtained for further sharing for research. To discuss a request for UKHSA data you would like to submit, contact DataAccess@ukhsa.gov.uk.

Code availability

The programming code for this project is available on Github: https://github.com/grahams99/Enhanced_surveillance_questionnaire.

References

1. Andrews, N. et al. Effectiveness of COVID-19 booster vaccines against COVID-19-related symptoms, hospitalization and death in England. *Nat. Med.* **28**, 831–837 (2022).
2. Andrews N., et al. Covid-19 vaccine effectiveness against the Omicron (B.1.1.529) variant. *N. Engl. J. Med.* **386**, 1532–1546 (2022).
3. Andrews, N. et al. Duration of protection against mild and severe disease by Covid-19 vaccines. *N. Engl. J. Med.* **386**, 340–350 (2022).
4. Lopez Bernal, J. et al. Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in England: test negative case-control study. *BMJ.* **373**, n1088 (2021).
5. Lopez Bernal, J. et al. Effectiveness of covid-19 vaccines against the B.1.617.2 (Delta) variant. *N. Engl. J. Med.* **385**, 585–594 (2021).
6. Kirsebom F. C. M., et al. COVID-19 vaccine effectiveness against the omicron (BA.2) variant in England. *Lancet Infect Dis.* **22**, 931–933 2022.
7. Jackson, M. L. & Nelson, J. C. The test-negative design for estimating influenza vaccine effectiveness. *Vaccine.* **31**, 2165–2168 (2013).
8. Falsey, A. R. et al. Phase 3 safety and efficacy of AZD1222 (ChAdOx1 nCoV-19) Covid-19 Vaccine. *N. Engl. J. Med.* **385**, 2348–2360 (2021).
9. Polack, F. P. et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N. Engl. J. Med.* **383**, 2603–2615 (2020).
10. UK government. Scientific evidence supporting the government response to coronavirus (COVID-19). https://www.gov.uk/government/collections/scientific-evidence-supporting-the-government-response-to-coronavirus-covid-19. Accessed 07/02/2022.
11. UK government. Greenbook Chapter 14a. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1045852/Greenbook-chapter-14a-11Jan22.pdf Accessed 07/02/22.
12. Allen, H. et al. Household transmission of COVID-19 cases associated with SARS-CoV-2 delta variant (B.1.617.2): national case-control study. *Lancet Reg. Health Eur.* **12**, 100252 (2022).
13. Lewnard, J. A. et al. Theoretical framework for retrospective studies of the effectiveness of SARS-CoV-2 vaccines. *Epidemiology.* **32**, 508–517 (2021).
14. NHS UK. COVID-19 STAFF FAQS: VACCINE INFORMATION. https://www.ouh.nhs.uk/working-for-us/staff/covid-staff-faqs-vaccine.aspx Accessed 07/02/22.
15. Williamson, E. J. et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature.* **584**, 430–436 (2020).
16. King, J. P., McLean, H. Q. & Belongia, E. A. Validation of self-reported influenza vaccination in the current and prior season. *Influenza Other Respir. Viruses* **12**, 808–813 (2018).
17. UK Health Security Agency. A guide to your COVID-19 vaccination. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1023816/UKHSA_12073_COVID-19_easy_read_guide.pdf. Published 2021. Accessed 08/08/2022.
18. Rhodes, S., Greene, N. R. & Naveh-Benjamin, M. Age-related differences in recall and recognition: a meta-analysis. *Psychon Bull Rev* **26**, 1529–1547 (2019).
19. Gillies C. L., et al. Association between household size and COVID-19: A UK Biobank observational study. *J. R. Soc. Med.* **115**, 138–144 (2022).
20. Forbes, H. et al. Association between living with children and outcomes from covid-19: OpenSAFELY cohort study of 12 million adults in England. *BMJ.* **372**, n628 (2021).
21. Endo, A., Funk, S. & Kucharski, A. J. Bias correction methods for test-negative designs in the presence of misclassification. *Epidemiol. Infect.* **148**, e216 (2020).
22. Orenstein, E. W. et al. Methodologic issues regarding the use of three observational study designs to assess influenza vaccine effectiveness. *Int. J. Epidemiol.* **36**, 623–631 (2007).
23. Jackson, M. L. & Rothman, K. J. Effects of imperfect test sensitivity and specificity on observational studies of influenza vaccine effectiveness. *Vaccine.* **33**, 1313–1316 (2015).
24. De Smedt, T. et al. Bias due to differential and non-differential disease- and exposure misclassification in studies of vaccine effectiveness. *PLoS One* **13**, e0199180 (2018).
25. Ainslie, K. E. C., Shi, M., Haber, M. & Orenstein, W. A. On the bias of estimates of influenza vaccine effectiveness from test-negative studies. *Vaccine.* **35**, 7297–7301 (2017).
26. Ciocanea-Teodorescu, I., Nason, M., Sjolander, A. & Gabriel, E. E. Adjustment for disease severity in the test-negative study design. *Am. J. Epidemiol.* **190**, 1882–1889 (2021).
27. Lewnard, J. A., Tedijanto, C., Cowling, B. J. & Lipsitch, M. Measurement of vaccine direct effects under the test-negative design. *Am. J. Epidemiol.* **187**, 2686–2697 (2018).
28. Kahn R., Schrag S. J., Verani J. R., Lipsitch M. Identifying and alleviating bias due to differential depletion of susceptible people in post-marketing evaluations of COVID-19 vaccines. *Am J Epidemiol.* **24**, 800–811 2022.
29. Smith, L. E. et al. Patterns of social mixing in England changed in line with restrictions during the COVID-19 pandemic (September 2020 to April 2022). *Sci Rep.* **12**, 10436 (2022).
30. Sullivan, S. G., Tchetgen Tchetgen, E. J. & Cowling, B. J. Theoretical basis of the test-negative study design for assessment of influenza vaccine effectiveness. *Am. J. Epidemiol.* **184**, 345–353 (2016).
31. Westreich, D. & Hudgens, M. G. Invited commentary: beware the test-negative design. *Am. J. Epidemiol.* **184**, 354–356 (2016).
32. Infante-Rivard, C. & Cusson, A. Reflection on modern methods: selection bias-a review of recent developments. *Int. J. Epidemiol.* **47**, 1714–1722 (2018).

33. Griffith, G. J. et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat. Commun.* **11**, 5749 (2020).
34. UK government. Who is at high risk from coronavirus. https://www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk/who-is-at-high-risk-from-coronavirus/. Accessed 07/02/22.
35. Miller, E. et al. Transmission of SARS-CoV-2 in the household setting: a prospective cohort study in children and adults in England. *J. Infect.* **83**, 483–489 (2021).
36. Hitchings, M. D. T. et al. Use of recently vaccinated individuals to detect bias in test-negative case-control studies of COVID-19 vaccine effectiveness. *Epidemiology.* **33**, 450–456 (2022).
37. Vasileiou, E. et al. Interim findings from first-dose mass COVID-19 vaccination roll-out and COVID-19 hospital admissions in Scotland: a national prospective cohort study. *Lancet.* **397**, 1646–1657 (2021).
38. Hall, V. J. et al. COVID-19 vaccine coverage in health-care workers in England and effectiveness of BNT162b2 mRNA vaccine against infection (SIREN): a prospective, multicentre, cohort study. *Lancet.* **397**, 1725–1735 (2021).

Acknowledgements

This work uses data provided by patients and collected by the NHS as part of their care and support (http://www.usemydata.org). In addition, we would like to thank participants of the questionnaire who provided valuable data that enabled the conduct of this study. S.G., E.M., N.A., J.L.W. and N.A. and H.I.M. are funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Vaccines and Immunisation (grant reference NIHR200929), a partnership between UK Health Security Agency and London School of Hygiene & Tropical Medicine. EPKP received funding from the UKRI COVID-19 Longitudinal Health and Wellbeing National Core Study (Phase 1 LHW-NCS, MC_PC-20059). The views expressed are those of the author(s) and not necessarily those of the NIHR, UK Health Security Agency or the Department of Health and Social Care.

Author contributions

Study concepts, E.T., J.S., J.L.B., E.M., N.A.; Study design, S.G., E.T., J.S., J.L.B., D.N., E.M., N.A., J.L.W., H.I.M.; Data acquisition, E.T., J.S., J.L.B., E.M., N.A.; Programming, S.G, E.T., J.S.; Statistical analysis, S.G.; Supervision, N.A., J.L.W., H.I.M., E.P.; Analysis and interpretation of results, All authors; Manuscript preparation, S.G.; Manuscript editing, E.P., D.N., E.T., N.A., J.L.W., H.I.M.; Manuscript review, All authors; Manuscript approval, All authors.

Competing interests

The UK Health Security Agency (UKHSA) has provided vaccine manufacturers with post-marketing surveillance reports which the companies are required to submit to the UK Licensing Authority in compliance with their Risk Management Strategy, and a cost recovery charge is made for these reports. SG is also a part-time salaried employee of Evidera, which is a business unit of Pharmaceutical Product Development (PPD), part of Thermo Fisher Scientific. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-39674-0.

Correspondence and requests for materials should be addressed to Sophie Graham.

Peer review information *Nature Communications* thanks Natalie Dean and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

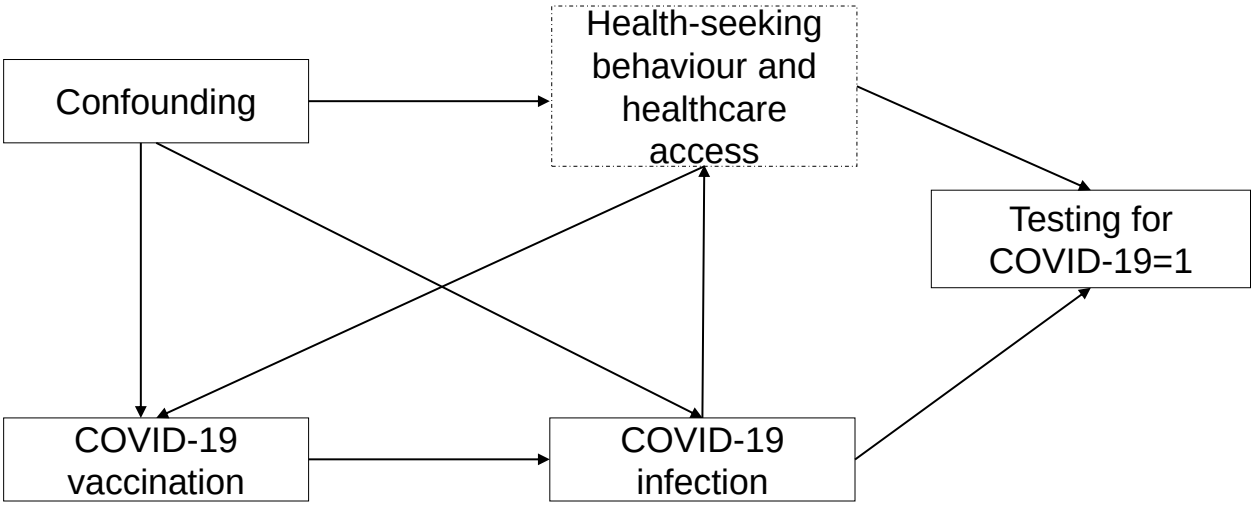
3.5 Additional Discussion

This section provides a more detailed discussion of the results from paper one above.

3.5.1 Confounding from health-seeking behaviour

A limitation that was only briefly mentioned in paper one is that potential confounding from health-seeking behaviour could not be assessed. Confounding from health-seeking behaviour was discussed in **Chapter 1** (see Section 1.8.1) as it has been shown previously to impact influenza vaccine effectiveness estimates as well as other observational study estimates. The original study used a test-negative design, which aims to account for confounding from health-seeking behaviour through its' design, however, potential collider bias threatens the validity of this claim. Collider bias is a form of selection bias. When an exposure and outcome independently cause a third variable this is termed a "collider". When the collider is inappropriately conditioned on, by study design or through statistical analysis, this results in collider bias. Controlling for a collider can introduce a distorted association between the exposure and outcome when in fact none exists[112]. Previously it has been speculated that collider bias is potentially introduced through the test-negative design. This is because both the exposure (e.g., COVID-19 vaccination; via health-seeking behaviour) and outcome (e.g., COVID-19 infection) affect the likelihood of being sampled into the study (e.g., COVID-19 testing). By conditioning on testing, this fails to block the non-causal pathway between COVID-19 vaccination and COVID-19 infection[113,114]. This is exemplified in the DAG in Figure 4 below.

Figure 4 DAG representing potential collider bias introduced through the test-negative design



Note: this figure is informed by figures from Sullivan et al[113] and Westrich et al[114]. Since health-seeking behaviour and healthcare access is not generally measurable (exemplified by the dotted box), conditioning on testing in the test-negative design fails to block the non-causal pathway from vaccination to infection. If we were able to appropriately control for confounding from health-seeking behaviour and healthcare access, then we would be able to block the non-causal pathway from vaccination to infection.

The questionnaire collected information on individuals' influenza vaccination status, which was originally thought could be used as a proxy for health-seeking behaviour. Originally it was thought that this proxy could be adjusted for to assess for potential confounding from health-seeking behaviour due to collider bias. However, this marker was not adjusted for in the final models since there is considerable positive association between COVID-19 vaccinations and influenza vaccination in the same season (e.g., in paper one 5,808 of 6,627 COVID-19 vaccinated individuals received an influenza vaccination in the 2020/21 season [88%]) and therefore adjusting for this would invertedly be adjusting for exposure status. In addition, non-vaccination for influenza in 2019/20 might also be because individuals with CEV status were asked to shield during the COVID-19 pandemic[100] (see Section 3.3.2.1) and therefore were unable to be vaccinated for influenza. Adjusting for this proxy could therefore be adjusting for underlying health conditions rather than health-seeking behaviour.

Not assessing for this type of confounding in the original test-negative study might not be problematic since the time period was when conforming to non-pharmaceutical interventions was high[115] and therefore overall the total population likely exhibited good health-seeking behaviours. Testing and vaccination capacity was also high[101,104] and therefore healthcare access was significantly improved for these services compared with pre-pandemic levels. However, although health-seeking behaviour might have improved overall for the total population during the COVID-19 pandemic, this still could be differential by vaccination status. For example, an observational

study conducted in Australia during the COVID-19 pandemic[116] reported that fully COVID-19 vaccinated individuals were more than twice as likely to report positive COVID-19 testing intentions compared to those that were unvaccinated. If this is truly the case then the association between vaccination and infection could be distorted which could lead to underestimated vaccine effectiveness estimates as vaccinated individuals are more likely to test for COVID-19. This problem would be investigated further throughout the course of the remainder of this thesis.

3.5.2 Potential misclassification of comorbidities in the questionnaire

Individuals were asked to self-report whether they had any of the COVID-19 at-risk conditions in the questionnaire, all of which were adjusted for in the analysis. However, it could be that there was potential misclassification of comorbidity status. For example, for the vast majority of individuals with chronic kidney disease, they do not know that they have their condition and in the UK there is also a wide variation on whether the GP has made a formal diagnosis in those who have laboratory evidence of chronic kidney disease[117]. Similarly for chronic heart disease in the UK – because of a national shortage in echocardiographers many people with suspected heart failure do not receive a timely diagnoses, and around 50% of heart failure cases in the UK only receive their diagnosis at hospital admission[118]. In the case of misclassification of confounders, vaccine effectiveness estimates would not be impacted if misclassification was non-differential by vaccination status. However, if this was differential, say for example, those that are vaccinated are more likely to have better health behaviours and therefore more likely to be diagnosed with a chronic condition, then adjusting for comorbidities might also in part account for confounding from health-seeking behaviour[66].

3.6 Overall chapter findings

Overall, in **Chapter 3** I used the original COVID-19 vaccination and COVID-19 PCR testing data from one of the first UK COVID-19 vaccine effectiveness studies that used a test-negative design. I supplemented data from this study with data from a questionnaire to investigate the presence of biases and alternative causal pathways in the original study. The questionnaire data revealed that there was limited evidence of bias or alternative causal pathways in the original test-negative study. However, the test-negative design is not always feasible to conduct as test result data is required. In addition, the design requires strong assumptions to be met, as discussed in **Chapter 1** (Section 1.10). Furthermore, as discussed in the current Chapter, collider bias potentially threatens the validity of the test-negative designs claim to account for confounding from health-seeking behaviour[113,114]. Therefore, alternative methods are required to identify, quantify and account for confounding from health-seeking behaviour, which is thought to highly confound

vaccine effectiveness estimates[44,45,55]. There is also value in triangulating evidence on vaccine effectiveness using different study designs[119]. Methods would also need to be developed to also confirm whether it is health-seeking behaviour that is confounding these estimates, as previous studies have only theorised this potential claim.

3.7 Unanswered questions

Some unanswered questions from this chapter are:

- To what extent can alternative methods to test-negative and other study designs be developed to account for confounding from health-seeking behaviour.
- To what extent is confounding from health-seeking behaviour impacting observational estimates and can this bias be accounted for?

3.8 How findings from paper informed rest of thesis

To answer the above questions, before developing any alternative methods, other existing methods to the test-negative and other designs needed to be investigated. It was known previously that typically health-seeking behaviour is not accounted for in observational research because it is not directly measurable in EHRs. All of the alternative methods were investigated using a pragmatic literature review that can be found described in **Chapter 4**. The focus for this pragmatic review was vaccine effectiveness research, as research into confounding from health-seeking behaviour has primarily been in this field.

4 Chapter 4: Pragmatic literature review: health-seeking behaviour in observational research

4.1 Introduction to the chapter

The aim of this chapter was to summarise methods used in observational vaccine research using EHRs to account for confounding from health-seeking behaviour. As highlighted in **Chapter 1**, this type of confounding has led to reports of 40-50% decrease in all-cause mortality in influenza vaccine effectiveness studies, which is not credible as influenza accounts for a maximum of 10% of deaths per year[44]. Although many authors have highlighted previously that these estimates are likely confounded by health-seeking behaviour[52], there were no known literature reviews at the time of this pragmatic literature review that summarised the different methods used to account for this bias.

This pragmatic literature review summarised the latest literature near to the time of thesis submission. It was conducted in January 2024 and was supplemented with studies identified throughout the course of this thesis.

4.2 Aim of chapter

To summarise the existing literature that explicitly accounts for confounding from health-seeking behaviour in vaccine effectiveness studies using EHRs.

4.3 Overall methodology

4.3.1 Approach and scope

A pragmatic review is defined as a conventional systematic literature review that takes into consideration time and resource limitations by applying limits[120]. This approach was chosen as when reading the literature for this thesis it was noticed how many observational studies aimed to account for confounding from health-seeking behaviour by adjusting for a single proxy marker, but did not explicitly mention this was the reason they were doing so. It would therefore not be possible to search for these studies in a published literature repository (e.g., Medline), without providing a search strategy that included every possible proxy marker. Potential proxy markers can also vary between datasets, and therefore it would not have been possible to come up with an exhaustive list of these. For this reason, only studies that explicitly mentioned they were intending to account for health-seeking behaviour were included. These were supplemented with papers identified throughout the course of the thesis where the intention was implied. As this was

a pragmatic literature review only Medline was used as a search engine. Grey literature (e.g., conference abstracts) was also not searched for.

4.3.2 Search strategy

The search was conducted using the Medline database, which includes literature published from 1966 to present. The database follows the Medical Subject headings (MESH) structure used by the National Library of Medicine. MESH is a clinical thesaurus that hierarchically organises terms so that related terms can be identified. For example, researchers that use the term myocardial infarction will also identify the term heart attack. It is used by PubMed for indexing articles[121]. To conduct the search in Medine, a list of relevant MESH terms for health-seeking behaviour or healthcare access, vaccine effectiveness and EHRs were compiled. No date or language restriction or any other limits were applied. Searches were set to multi-purpose which meant that terms were searched for in the title, original title, abstract, subject heading, name of substance, and registry word fields. Medline automatically provides potentially relevant synonyms from MESH, which were also used in the search. These can be seen as the capitalised terms in Table 5 below. There is also the option in Medline to explode searches which would have also identified additional searches based on all narrower terms in the MESH hierarchy. Searches were not exploded to ensure that they were as concise as possible. The final search strategy was applied in Medline on 3 January 2024 and can be found with the article numbers identified in Table 5 below.

Table 5 Search strategy applied in Medline on 3rd January 2024

#	**Search terms**	**Number of studies identified**
1	Health Behavior/ or health seeking behavio*r*.mp.	59,916
2	Health Services Accessibility/ or healthcare access.mp.	88,456
3	vaccin*.mp. or Vaccination/ or Vaccines/	497,187
4	immuni*.mp.	550,769
5	effect*.mp.	11,437,117
6	Electronic Health Records/ or electronic health record*.mp. or Medical Records Systems, Computerized/	64,641
7	electronic health data*.mp.	532
8	Primary Health Care/ or primary care record*.mp.	93,254

9	Medical Record Linkage/ or Routinely Collected Health Data/ or routinely collected health data*.mp.	5,290
10	claims data*.mp.	20,298
11	administrative data*.mp.	18,837
12	real world*.mp.	87,029
13	1 or 2	146,740
14	3 or 4	882,525
15	5 and 14	357,096
16	6 or 7 or 8 or 9 or 10 or 11 or 12	275,233
17	13 and 15 and 16	85

Abbreviations: mp: multi-purpose (searches title, original title, abstract, subject heading, name of substance, and registry word fields).
*Any character can occur where the asterisk is, including no character.
Note: word searches that are capitalised are synonyms that were identified by Medline.

Titles and abstracts were then screened to identify observational studies of vaccine effectiveness using EHRs which explicitly adjusted for confounding by health-seeking behaviour or healthcare access. EHRs were defined in this study as claims or administrative data. Studies that exclusively used prospectively collected data from surveys, medical charts, intervention study or from disease specific registries were not included as these tend to collect relevant information specific to the study research question. Literature reviews, single centre studies, case reports and systematic literature reviews were excluded. Additional relevant papers that had previously been identified during the course of my thesis were also included.

4.4 Results

Overall, the Medline search identified eighty-five potential studies, of which two were relevant[122,123] (both of which were previously identified). Six other studies were identified from my wider reading during my thesis, bringing the total number of included studies to eight[122-127]. Only two of the studies overall[122,123] (i.e., the ones identified in the search) explicitly mentioned in their abstract that they were aiming to account for confounding from health-seeking behaviour. For the other six studies, this was implied.

In summary, all of these studies either used Medicare data in the US or GP EHRs data in England linked to other datasets. Medicare is the federal insurance-based system for all individuals in the US ≥65 years, or those <65 years who have a disability, end-stage renal disease or have had an

organ transplant. In 2015, there were more than 55 million individuals covered by Medicare. It includes information on demographics, hospital and outpatient visits with corresponding diagnoses, drug and procedure codes[128]. The datasets used in England are similar to those previously described in **Chapter 3**.

All of the studies identified used proxy markers to account for confounding from health-seeking behaviour. One of these studies identified a proxy in survey data that could be linked to EHRs[123]. Theis proxy marker included self-reporting of response to the following question: "I would do almost anything to avoid going to the doctor". The other studies identified proxy markers directly in EHRs[122,124-127,129]. The number of proxy markers included in each of these studies ranged from one to thirteen. The included proxy markers were inconsistent across the datasets, even when the same group of researchers and same datasets were used. For example, Izurieta's first study using Medicare data in influenza vaccine effectiveness in the 2017/18 season [127] used five markers (pneumococcal vaccination, hospital visits, outpatient emergency, outpatient non-emergency and GP visits). For their next study of influenza vaccine effectiveness in the 2018/19 season, that also used Medicare [129], they used fourteen markers that were mostly preventative measures for example screening and vaccinations. In their next study of influenza vaccine effectiveness in the 2019/20 season, also in Medicare, they used only five markers[122], three of which were different to their prior study[129]. The markers used across other datasets (e.g., in the UK) were also different. For example, in the UK authors that assessed COVID-19 vaccine effectiveness during the early COVID-19 pandemic[126] used GP visit quartile rate, whereas, another group of researchers also assessing COVID-19 vaccine effectiveness in the same datasets used SARS-CoV-2 testing as a single proxy marker[125]. None of the studies detailed the methodology used to select their set of markers, nor did any of the researchers detail why they used different sets of markers in the same datasets.

Four of the included studies[51,122,127,129] assessed influenza vaccine effectiveness, one of them[123] assessed shingles vaccine effectiveness, whereas the other three assessed COVID-19 vaccine effectiveness[124-126]. Only two of the studies assessed the impact of adjusting for these markers on vaccine effectiveness estimates[51,123]. Both of these studies reported that adjusting for this type of confounding reduced vaccine effectiveness estimates, except for one of the three outcomes in the shingles vaccine effectiveness study[123]. They reported for ophthalmic zoster that vaccine effectiveness point estimates were similar after confounding adjustment. Both of these studies used negative controls (see Section 1.9) to assess for residual confounding after adjustment for these markers. In one of these, for influenza vaccine effectiveness they used pre-season

influenza estimates as a negative control[51]. After adjustment for age, sex, race, comorbidities, five proxies of health-seeking behaviour, and frailty proxies, their fully adjusted model showed significant evidence of potential residual confounding (pre-season estimate: 32% [95%CI: 30-33%]). Even when they added these variables into the model in different order, their final vaccine effectiveness estimates provided the same adjusted estimate (pre-season estimate: 32% [95%CI: 30-33%]). In the other study, which estimated shingles vaccine effectivness[123] they used thirteen different negative control outcomes (e.g., hip fracture, thrombosis). After adjustment for demographic factors, socio-economic conditions, healthcare utilisation characteristics, frailty characteristics, functional immunocompromising chronic conditions, immunocompromising drugs, one marker of health-seeking behaviour (survey response: “I would do almost anything to avoid going to the doctor”), a marker of mobility status (survey response: “I experience difficulty walking”) and education status (survey response: “highest level of education”), they reported vaccine effectiveness estimates that ranged from -19% to 27%. Although CIs for all the outcomes moved towards the null after adjustment for the proxy markers, none of them crossed the null, providing evidence of significant residual confounding.

The table below (Table 6) provides a summary of all eight of the included studies and the reported impact of accounting for confounding from health-seeking behaviour on the vaccine effectiveness estimates, where possible.

Table 6 Literature identified in pragmatic search and throughout course of my thesis that accounted for confounding from health-seeking behaviour in vaccine effectiveness research using EHRs

Author	Study population	Study design	Exposure (s)	Outcome(s)	Statistical methods used for confounding	Variables use din adjustment for confounding	Variables capturing confounding for health-seeking behaviour	Reported impact on vaccine effectiveness estimates
Zhang et al, 2017[51]	≥65 years in the in the 2007-2008 influenza seasons identified in the Medicare claims dataset in the US.	Cohort.	Any influenza vaccination versus unvaccinated.	All-cause mortality.	Adjusted for potential confounders in the Cox proportional hazards model.	Age, gender, race, markers of health-seeking behaviour, comorbidities and markers for frailty.	Colonoscopies, fecal occult blood tests, mammogram, PSA test and pneumococcal vaccination.	Reduced vaccine effectiveness estimates compared with minimally adjusted estimates (32% [95%CI: 31-33%] to 27% [95%CI: 26-28%]). Residual confounding potentially remained in the fully adjusted models (pre-season vaccine effectiveness estimate: 32% [95%CI: 30-33%]) and post-season estimates: 32% (95%CI: 30-33%). .
Izureita et al, 2019[123]	≥65 years from 1991 to 2013 identified in the Medicare claims dataset in the US, with a subset of the data linked to the Medicare Current Beneficiary Survey.	Cohort.	Shingles vaccination.	Community herpes zoster, antiviral-treated herpes zoster, and ophthalmic zoster.	Adjusted for potential confounders in the Cox proportional hazards model.	Demographic factors, socio-economic conditions, healthcare utilisation characteristics, frailty characteristics, chronic conditions and a marker of health-seeking behaviour.	Self-reporting doctor avoidance.	Community herpes zoster: from 41% (95%CI: 39-42) to 39% (95%CI: 36-42); antiviral-treated herpes zoster: from 34% (95%CI: 32-35) to 31% (95%CI: 28-35); ophthalmic zoster from 31% (95%CI: 27-36) to 32% (21-41).
Izurieta et al, 2019[127]	≥65 years in the 2017-2018 influenza seasons identified in the Medicare claims	Cohort.	Five different influenza vaccinations compared	Influenza related hospital visit.	Probability of treatment weighting.	Demographics, reason for entry into Medicare, region of residence, month of vaccination, markers of health-seeking	Pneumococcal vaccination, hospitalisations, outpatient emergency visits, outpatient non-emergency visits and GP visits.	Not possible to assess as markers were included in propensity weighting.

Author	Study population	Study design	Exposure (s)	Outcome(s)	Statistical methods used for confounding	Variables use din adjustment for confounding	Variables capturing confounding for health-seeking behaviour	Reported impact on vaccine effectiveness estimates
	dataset in the US.		to each other.			behaviour, chronic medical conditions.		
Izurieta et al, 2020[129]	≥65 years in the 2018-2019 influenza seasons identified in the Medicare claims dataset in the US.	Cohort.	Five different influenza vaccinations compared to each other.	Influenza related hospital visit.	Probability of treatment weighting.	Demographics, Medicaid eligibility, reason for entry into Medicare, region of residence, prior medical encounters, chronic medical conditions, markers of health-seeking behaviour, and frailty indicators.	Pneumococcal vaccine, annual wellness visit, bone density scan, cardiovascular disease screen test, bowel cancer screen, diabetes screen, initial preventative physical examination, PSA test, breast cancer screen, cervical cancer screen, pelvic examination screen, depression screen and other preventative services.	Not possible to assess as markers were included in propensity weighting.
Izureita et al, 2021[122]	≥65 years in the 2019-2020 influenza seasons identified in the Medicare claims dataset in the US.	Cohort.	Five different influenza vaccinations compared to each other.	Influenza related hospital visit.	Probability of treatment weighting.	Demographics, region, month of vaccination, chronic health conditions, frailty, prior medical encounters, and markers of health-seeking behaviour.	Annual wellness visits, counselling and health risk assessment, pneumococcal vaccine, tetanus vaccine, shingles vaccine.	Not possible to assess as markers were included in propensity weighting.
Whitaker et al, 2022[126]	≥18 years between 7 December 2021 and 16 May 2021 identified in GP EHR data of 712 practices in England.	Cohort and test-negative.	COVID-19 vaccination versus unvaccinated.	Symptoms of COVID-19 reported within 10 days before or after a positive SARS-CoV-2 test.	Adjusted for potential confounders in the poisson and logistic regression models.	Age, sex, ethnicity, IMD, recent infection, a marker of health seeking behaviour, comorbidities, shielding recommendations and smoking status.	GP consultation rate.	Not possible to assess as they did not adjust for health-seeking behaviour separately.

Author	Study population	Study design	Exposure (s)	Outcome(s)	Statistical methods used for confounding	Variables use din adjustment for confounding	Variables capturing confounding for health-seeking behaviour	Reported impact on vaccine effectiveness estimates
Horne et al, 2022[124]	≥18 years on 1 July 2021 identified in GP EHR data linked to hospital and COVID-19 PCR testing data in England.	Cohort.	COVID-19 vaccination versus unvaccinated.	COVID-19 related hospital visit, COVID-19 related death and positive SARS-CoV-2 test.	Adjusted for potential confounders in the Cox regression model.	Age, sex, IMD, ethnicity, BMI, comorbidities, pregnancy and two markers of health-seeking behaviour.	SARS tests between 18 May 2020 and first vaccination dose receipt of one or more influenza vaccination in the previous 5 years.	Not possible to assess as they did not adjust for health-seeking behaviour separately.
Hulme et al, 2022[125]	Health and social care workers vaccinated between 4 January and 28 February 2021 identified in GP EHR data linked to hospital and COVID-19 PCR testing data in England.	Cohort.	COVID-19 vaccination versus unvaccinated.	COVID-19 related hospital visit, COVID-19 related A&E visit and positive SARS-CoV-2 test.	Adjusted for potential confounders in the logistic regression model.	Age, sex, IMD, ethnicity, region, comorbidities, a marker of health-seeking behaviour, rurality and recent infection.	SARS-CoV-2 tests in the 90 days prior to study start.	Not possible to assess as they did not adjust for health-seeking behaviour separately.

Abbreviations: A&E: accident and emergency; EHR: electronic health records; GP: general practice; IMD: index of multiple deprivation; PCR: polymerase chain reaction; PSA: prostate specific antigen.

4.5 Discussion

Overall, this pragmatic literature review identified that although confounding from health-seeking behaviour is a well-recognised issue in vaccine effectiveness research, very few studies explicitly mention they are aiming account for this type of bias. The reason for this is likely because, apart from the test-negative and other designs that aim to account for this counfounding, there is no consensus for alternative methods that use proxy markers. It is also likely because previous methods that used proxy markers were not very effective as residual confounding remained.

Of the included studies, all used proxy markers to account for this type of confounding. The majority identified proxies in EHRs directly. These markers included testing for the vaccine preventable condition, GP visit quartiles and preventative measures such as routine screening and vaccinations. Testing for the vaccine preventable condition and GP consultation visits are problematic on their own as these markers are likely more strongly influenced by underlying health need than health-seeking behaviour. For SARS-CoV-2 testing, during the COVID-19 UK pandemic, individuals were only allowed to access free governmental community PCR testing if they had COVID-19 like symptoms[130]. GP visits quartiles are also likely to be highly influenced by an individual's underlying health conditions particularly during the COVID-19 pandemic and with the overburdened NHS due to austerity[131]. The other markers used in these studies (governmental screening and vaccinations) are likely to be better markers of health-seeking behaviour as the influence of underlying health need is likely to be weaker.

One of the studies used markers identified in linked survey data (response to "I would do almost anything to avoid going to the doctor"). There are potentially further issues associated with identifying markers of health-seeking behaviour from survey data. Since individuals who respond to a health survey are likely to be those that are engaged in their health[132], selection bias might distort the association between vaccination and infection.

There appeared to be no mention of how the markers were selected in any of the studies. The rationale for no mention of the methods, is likely that a systematic approach was not utilised. It is likely that decisions were made based on internal discussions, without any conceptual framework or criteria used for marker selection.

Only two of the studies adjusted for markers in a separate step in their models to allow for this type of confounding to be quantified. Both identified that vaccine effectiveness estimates declined after adjusting for these markers, except for one of the outcomes in one of the studies (zoster vaccination againnst ophthalmic zoster outcome), for which estimates were similar. In both these

studies there was evidence of residual confounding after this, as represented through their negative control analyses. Residual confounding likely remained in these studies as in one of the studies healthcare utilisation (e.g., outpatient visits) markers were used, which are highly influenced by underlying health need. In the other study, they used only one marker of health-seeking behaviour that was imputed from a survey therefore selection bias from health-seeking behaviour was likely introduced.

4.6 Conclusion

Overall, this pragmatic literature review identified that there are very few observational studies using EHRs that explicitly use alternative methods to test-negative designs and other study designs to address confounding from health-seeking behaviour. The alternative methodologies include the use of proxy markers identified in EHRs or in linkable surveys. The set of proxy markers used across these studies are very inconsistent and authors do not detail how markers are selected. Few studies have used markers to quantify this type of confounding and of the ones that did, insufficient markers sets were used and therefore residual confounding remained. Since confounding from health-seeking behaviour is a complex phenomenon, markers of health-seeking behaviour need to be selected systematically and based on a conceptual framework so that the underlying phenomenon is appropriately accounted for.

4.7 Gaps identified in the literature from pragmatic literature review

From this pragmatic literature review the following gaps were identified in the literature:

- There is a need to develop a systematic set of markers of health-seeking behaviour that are informed by a conceptual framework so that the underlying phenomenon is appropriately accounted for.
- There is a need to use these markers to quantify and account for confounding from health-seeking behaviour.

4.8 Thesis objectives informed by pragmatic review

Based on the above gaps identified in the literature, it was decided to systematically identify proxy markers of health-seeking behaviour in EHRs data. This would be informed by a conceptual framework and criteria would be developed so that additional markers could be identified by future researchers. These markers would then be adjusted for in a vaccine effectiveness study to quantify and account for this type of confounding.

Based on this, the following objectives were developed for the remainder of the thesis:

1. To systematically identify a set of markers of health-seeking behaviour available in EHRs that can potentially be used to quantify and account for this type of confounding.
2. To quantify and account for confounding from health-seeking behaviour in an influenza and COVID-19 vaccine effectiveness study.

5 Chapter 5: General theory and methods for identifying, quantifying and accounting for confounding from health-seeking behaviour

5.1 Introduction to the chapter

This chapter provides the methods used to meet the study objectives laid out in **Chapter 4** Section 4.8. These specific objectives were:

1. To systematically identify a set of markers of health-seeking behaviour available in EHRs that can be potentially used to quantify and account for this type of confounding.
2. To quantify and account for confounding from health-seeking behaviour an influenza and COVID-19 vaccine effectiveness study.

These two objectives were met in two separate publications that are provided in study two in **Chapter 6** and study three in **Chapter 7** below. The current chapter will lay out the datasets and general methods that were used in both studies since these were consistent.

5.2 Aim of chapter

To provide an overview of the datasets used in both study two and three, including information on the validity and generalisability of these data. In addition, to provide an overview of the general approach that was used to generate code lists and create variables for both studies.

5.3 Overall methodology

The EHRs that were used in both study two and three were UK primary care data (CPRD Aurum[12]) linked to secondary care (HES[133]) and death data (ONS[134]). **Chapter 3** described the UK healthcare system, NHS datasets and coding systems used in these data. The current chapter provides more information on UK primary care EHRs and then details on the CPRD Aurum, HES and ONS datasets.

5.3.1 Primary care EHRs in the UK

There are three main providers of primary care electronic healthcare software systems in use in the UK currently and these include: EMIS® health, SystmOne (which is provided by The Phoenix Partnership [TPP]) and Vision®. Data from these systems has been made available to researchers through partnerships between practices, system vendors, universities and not-for profit organisations for many years. The partnerships between these organisations led to the formation of the General Practice Research Datalink, which later became known as CPRD,

QResearch, The Health Improvement Network database and the Optimum Patient Care Research Database. The Royal College of General Practitioners (RCGP) has also separately supported surveillance research by providing data across practices. More recently the research objectives of RCGP have broadened and they have become the Research and Surveillance Centre. There are other partnerships that have arisen in more recent years.

The population coverage within each of these systems is dependent on the popularity and geographical reach of the available software systems. EMIS® health is currently the most popular amongst the systems and together with TPP cover more than 90% of practices in England. Initially the focus of EHR research using these systems was mostly in pharmacoepidemiology, but in more recent years, research objectives have broadened to include more general aspects of epidemiology[135].

5.3.2 History of CPRD and it's use in research

CPRD was first established in 1987 and was originally known as the small VAMP dataset. This dataset continued to grow until it became the General Practice Research Datalink in 1993 and then the CPRD in 2012. Data from these practices are provided to CPRD in an anonymised format and data from each patient in each of these practices is provided unless the patient has asked to opt out. Originally CPRD included data provided from the Vision® software and this formulated a dataset known as CPRD GOLD[136]. Since October 2017, CPRD released another dataset known as CPRD Aurum which provides data from practices that use EMIS® software[137].

5.3.3 Clinical Practice Research Datalink Aurum

5.3.3.1 Overview

The May 2022 CPRD Aurum release was used in the current study. This consisted of information from 1,491 current and historic patients which covered a total of 41,200,722 'acceptable' patients. CPRD deems patients acceptable when their medical records are of research quality. Of all acceptable individuals, 38,377,503 are eligible for linkage to other datasets (Set 22 linkage release). Currently there are 1,345 general practices that are contributing data, which includes 13,300,067 contributing patients (19.83% of the UK population). 99% of the practices are in England and <1% are in Northern Ireland. Mean (standard deviation) follow-up time of currently contributing patients is 7.9 (8.0) years)[138].

5.3.3.2 File contents

Data from CPRD Aurum is collected from contributing practices daily to create monthly releases that are used in observational research. Information recorded includes patient demographics

(age, gender), lifestyle factors (BMI, smoking), medical diagnoses, test results, prescriptions and interactions with secondary care (e.g., referrals). Medical diagnoses are recorded using SNOMED-CT/Read Version 2/local EMIS® codes that are each individually assigned by CPRD to a unique medcode. SNOMED-CT codes have been previously described (**Chapter 3**). Read codes contain clinical terms that are organised into chapters from 0 to 9. Within each of these chapters terms are hierarchically organised, moving from general terms at the top to very specific terms at the bottom[139]. Local EMIS® codes unique to individual EMIS practices and cannot be reliably shared across EMIS practices[140]. Where possible these codes have been mapped to SNOMED-CT codes. Information on prescription medicines are coded using the Dictionary of Medicines and Devices (dm+d), which is integrated within the SNOMED-CT system (i.e., all unique identifiers within dm+d are SNOMED-CT codes)[141]. CPRD assigns each dm+d code an individual prodcode.

CPRD data is structured on eight different file types which contains different categories of information. Patients are assigned a unique identifier which enables their records to be linked across the files, and a consultation identifier allows events from the same consultation to be identified. The key file types and contents in CPRD Aurum are outlined in Table 7 below.

Table 7 Key files and included variables in CPRD Aurum database

File	File contents	Variable	Description
Patient	Contains information on basic demographics and patient registration details.	patid	Unique encrypted patient identifier
		pracid	Unique encrypted practice identifier
		yob	Year of birth
		gender	Gender (male, female, indeterminate or unknown)
		regstartdate	Date of patient registration at their current practice.
		regenddate	Date of patient de-registration out of their current practice, which could either represent death or transfer out of the practice.
		cprd_ddate	Date of death as estimated by CPRD.
		acceptable	Flag that indicates the patient meets research quality standards. See section X below for more details.

Practice	Contains details of each practice.	pracid	Unique encrypted practice identifier
		lcd	Date of most recent data collection at the patients current practice
		region	Region (North East, North West, Yorkshire and The Humber, East Midlands, West Midlands, East of England, London, South East, South West, Wales, Scotland and Northern Ireland)
Observation	Contains the medical history data entered on the GP system including symptoms, clinical measures, laboratory test results and diagnoses as well as demographic information recorded as a medcode (e.g., ethnicity).	patid	Unique encrypted patient identifier
		obsdate	Date associated with an event
		medcodeid	CPRD unique code for the medical term associated with the event as selected by the GP in the EMIS® system
		value	Measurement of test value
Drug issue	Contains details of all prescritions on the GP system.	patid	Unique encrypted patient identifier
		issuedate	Date the prescription was issued
		prodcodeid	Unique CPRD code for the treatment as selected by the GP in the EMIS® system
		dosageid	Unique code that provides information on the dosage as provided on the preescription
		quantity	Total quantify of the prescribed treatment, as entered by the GP
		quantunitid	Unit of the treatment
Consultation	Contains information relating to the type of consultation as entered by the GP.	patid	
		consdate	Date of event
		conssourceid	Identifier that proves information on the source of the consultation as entered in the EMIS® software. For example this could “Community Clinic” or “Casulty Attendance”.
		consmedcodeid	
		consid	

		staffid	Unique encrypted staff identifier
Staff	Contains practice staff details for each staff member.	staffid	Unique encrypted staff identifier
		jobcat	Job category of the staff member

Note: this table is adapted from CPRD Data Specification[12].

5.3.3.3 COVID-19 vaccination and testing data

During the COVID-19 pandemic, COVID-19 vaccination data automatically flowed into patient GP practice records from NHS England and were recorded in the CPRD Aurum dataset as prodcodes (see Section 5.3.3.2). CPRD also reported that COVID-19 testing result data was either respectively or prospectively pushed into GP records and were recorded as medcodes (see Section 5.3.3.2) [142].

5.3.3.4 Derived death date

CPRD provide a derived death data which is based on an algorithm. This approach identifies probable deaths using three different approaches. This includes a) deaths recorded as transfer out with reason death in the patient file, b) administrative deaths recorded in the Clinical file as entity type 148 and c) deaths recorded as Read codes. Some data cleaning is conducted by CPRD. For example, deaths recorded before 1/1/1987 are removed as this was before CPRD was created. Then for individuals with more than one death date reported, the transfer out date (a) is prioritised and then the administrative death date (b), followed by the Read code event date (c)[143].

5.3.3.5 Generalisability and validity

The data with the CPRD Aurum dataset are representative of the English population in terms of age, sex, geographical spread and socioeconomic status[137]. CPRD undertakes over 900 validation checks on the data prior to the data being released. Any issues that are highlighted during these checks are then resolved beforehand. Once the data has been collected, check are conducted to assess whether the correct data has been supplied and whether the data elements are of the correct length and format. Any duplicate records are removed. Validation checks are also conducted to ensure that there are no orphan records in the data i.e., there are no records that belong to patients that have been removed from the data. Then checks are conducted on the research quality of the data. As previously mentioned CPRD provides a patient level quality metric known as the ‘acceptability flag’ which is a flag that determines whether a patient’s medical records are deemed to be of research quality. Patients are removed from the sample for any of the following reasons:

- Missing year of birth.
- Missing registration start date.
- Registration start date is after practice last collection date.
- Registration start date is before or equal to 01/01/1900.
- Registration start date is equal to or after registration end date.
- Registration start date is before year of birth.
- Gender is other than male, female or indeterminate.
- Age at the end of follow-up is greater than 115 years, based on registration end date/death date/last collection date minus year of birth.
- All recorded health care episodes have missing or invalid (before or equal to 01/01/1900, after last collection date or before year of birth) event dates.
- Patients without permanent registration.

A systematic literature review was conducted to assess the validity of diagnoses in the CPRD dataset. 212 publications were included that validated 183 different diagnoses. The review summarised that overall, the validity of diagnoses recorded in CPRD was high (median of 89% of cases were confirmed through external validation), although in some cases the reporting of validations was not of sufficient quality to permit a clear interpretation[144].

5.3.3.6 Quality and Outcomes Framework

The quality of the data recorded in CPRD Aurum is influenced by the Quality and Outcomes Framework (QOF) that was first introduced in 2004[145]. The QOF is a system that remunerates general practitioners in the United Kingdom for providing good quality care to their patients and to help improve the quality of their care[146]. Although the QOF is a voluntary system, over 99% of practices in England participate in the scheme. The 2019/2020 QOF indicators were categorised into four domains known as clinical, public health, public health services and additional services. The 2019/2020 QOF measured achievement against 68 indicators and practices could score up to a maximum of 559 points. Most of the 2019/2020 clinical indicators related to the long-term care of chronic conditions. The public health indicators relate to primary prevention of cardiovascular disease, blood pressure, obesity and smoking. The public health domain – additional services is related to cervical cancer screening. The quality improvement domain is related to prescribing safety and end-of-life care[147].

QOF has led to significant improvement in the recording of clinical events. In terms of lifestyle factors, BMI and smoking recording have improved. For example, a descriptive study that was

conducted using the CPRD GOLD database in 2019[148] summarised that QOF had led to improvements in the recording of weight in primary care records, but that this recording was selective, with 97% of individuals with diabetes having their weight recorded, whereas only 54% of individuals without diabetes having their weight recorded. In terms of smoking status, QOF incentives started in 2008 and ended in 2011. A study that assessed recording of smoking status in 28 general practices in London, identified that smoking status recording increased from 55.5% to 64.3% for men before QOF incentivization to 67.9% to 75.8% for women during QOF incentivisation[149].

In terms of ethnicity, QOF incentives started in 2004, but was then later removed in 2011. A study that used the May 2021 CPRD Aurum build found that 82.3% of currently acceptable patients had at least one ethnicity recording in CPRD. This increased to 92.9% when the researchers restricted to acceptable patients with a registration date in the QOF incentivisation period[150].

5.3.3.7 Data linkage

Linkage to additional datasets allows patient care in other settings (e.g., secondary care) to be identified, which increases the ability to identify additional medical encounters beyond primary care. CPRD Aurum can be linked to HES, and ONS (as well as other data sources) on the patient level. Consent for this linkage is provided on the practice level. For practices that consent these linkages, they submit patient identifiers to NHS England who are the assigned trusted third party for conducting the linkage. They are responsible for matching these identifiers with identifiers from external custodians using deterministic linkage. This linkage occurs through eight progressively less restrictive steps that use NHS number, date of birth, postcode and gender. Each match is ranked from 1-8, with 1 being the best quality match and 8 being the lowest quality. Only linkage that are 5 or below are provided to CPRD. It has been estimated that ~96% of patients are matched with a ranking of 1 or 2, with less than 4% of patients with a match 6-8 [151].

5.3.4 Hospital Episode Statistics

5.3.4.1 Overview

CPRD offers linkage with HES, which allows the patient to be tracked in secondary care. Linkage is only available for English practices, as HES is only available for England. As CPRD Aurum (May 2022 CPRD release) contains 13 practices in Northern Ireland, linkage to HES is available for 99.03% of practices.

Available HES datasets are HES outpatient, HES accident and emergency, HES diagnostic imaging datasets (DID) and HES admitted patient care (APC). HES APC data contains de-

identified data for all admissions to, or attendances at English NHS hospitals. This includes private and charity hospitals that are paid for by the NHS. In England it is estimated that around 98-99% of all hospital activity is funded for by the NHS[152].

5.3.4.2 File contents

HES APC includes some socio-demographic information, hospital admission and discharge dates, admission diagnoses, procedures and administrative information (e.g., information on wait times). Diagnoses are provided by hospital coders who read diagnosis lists on hospital discharge/specialty/interhospital transfer letters generated on patients on admitted in-patients. These diagnoses are coded as ICD-10 codes (see Section 3.3.1.3). Information on medications prescribed in-hospital are not reported to HES, neither is information on laboratory test results[153]. The database structure has eight distinct files. Patients are assigned a unique identifier which enables their records to be linked across the files, and a consultation identifier allows events from the same consultation to be identified. The key file types and contents in HES APC are outlined in Table 7Table 8 below.

Table 8 Key files and included variables in HES APC database

File	File contents	Variable	Description
Patient	Contains patient demographics, date of death if applicable, and date of patient's registration or deregistration from the medical practice	patid	Unique encrypted patient identifier
		pracid	Unique encrypted practice identifier
		gen_ethnicity	Patient's ethnicity*
Hospitalisations (i.e., spells)	Contains information on spell admission and discharge dates.	patid	
		spno	Spell number uniquely identifying a hospitalisation
		admidate	Date of admission
Diagnoses	Contains information on diagnoses codes recorded at admission and discharge.	patid	
		spno	
		ICD	An ICD10 diagnosis code in XXX or XXX.X format
Procedures	Contains information on procedure codes and procedure dates.	patid	
		spno	
		Operating Procedure Codes Supplement (OPCS)	An OPCS 4 procedure code
		evdate	Date of operation / procedure

Note: this table is adapted from the CPRD data minimization file[154]. Abbreviations: ICD-10: International Classification of Disease, 10th Revision; OPCS: Operating Procedure Codes Supplement.

*Ethnicity in HES APC has improved other time, it increased from 41% in 1997 to 85% in 2011[155].

5.3.4.3 Generalisability and validity

Since HES includes around 98-99% of hospital admissions in England[152] it is considered to be very generalisable.

NHS England conducts data quality checks on the data yearly. This entails extracting 'key' data items which are then assessed for validity and completeness against appropriate data standards using a set of business rules that were previously developed by the Health and Social Care Information Centre. Hospital care providers are provided with tools to verify the accuracy of reported data, with the aim of promoting high-quality data coding at the source. The submitted data undergoes an audit process to ensure completeness and identify any invalid data formats. The results of these audits are then communicated back to the hospital care providers. Additionally, the HES data obtained is thoroughly examined and validated to ensure its internal consistency and reliability[156].

There have been few studies that have assessed the validity of diagnoses in HES APC. One study compared myocardial infarction diagnoses identified in HES APC and a disease registry (Myocardial Ischaemia National Audit Project). Electrocardiographic and troponin findings from the disease registry were used as the gold standard. They found that the positive predictive value was 91.5% (90.8% to 92.1%)[157].

5.3.5 Office for National Statistics Mortality Data

5.3.5.1 Overview and file contents

CPRD also offers linkage with ONS, which is a dataset that is considered the gold standard for mortality data in the UK. It contains information on death date, cause(s) and place of death. Linkage to ONS is consented for every CPRD Aurum practice and linkage is conducted by NHS England based on a patients NHS number, date of birth and postcode. CPRD offers linkage to ONS from 2 January 1993 onwards. The underlying cause of death and then up to 15 causes of death are recorded using ICD-10 codes (see Section 3.3.1.3). The key variables in the data are 'patid', 'dod' (date of death) and 'cause' (recorded cause of death in ICD-10 format).

CPRD death date has previously been compared to ONS death date. In 69.7% of cases, CPRD death date and ONS death date were recorded as the same day. An earlier death date was identified in CPRD versus ONS for less than 3% and a later date was identified in for 27.7% of cases[143].

5.3.6 Office for Socio-economic status data

5.3.6.1 Overview and file contents

CPRD also offers linkage to IMD, which is an area-based measure of relative deprivation, which is a proxy for socio-economic status. Linkage to IMD is consented for every CPRD Aurum practice and linkage is conducted by NHS England based on a patients' or practices' registered postcode. IMD is a score that is created by the UK Ministry of Housing, Communities and Local Government, who use administrative data, such as benefits records, but also census data. It provides a weighted rank for different social domains that are based on thirty-eight different indicators across seven different domains[158]. These domains are income, employment, education and skills, health, housing, crime, access to services and living environment[159]. For the current study both linkage to patient level and practice level IMD. This is because missing data from the practice-level linkage is estimated to be around 6.7%[158], whereas this is not missing at the practice-level and therefore practice level data can be used to supplement patient level data in the cases that this is missing.

5.3.7 Rationale for why these data were used

These datasets were used since they are one of the largest datasets in the UK that can be made available to researchers with direct access to the patient level data. Although the data has to be stored in a highly secure server, the data can be visualised in its raw form. The benefit of being able to access these data in its raw form is that it is easier to come learn the structure of the data. In addition, access to these datasets via CPRD is relatively straightforward in terms of ethics and documents required. The NHS England datasets are much more difficult to acquire and require much longer time frames to access. CPRD and the linked datasets have also been heavily researched on in the past. CPRD estimates that their data has resulted in over 3,000 peer-reviewed publications[160]. The benefit of this is that the strengths and limitations of the data are well understood.

5.3.8 Ethics approvals

The ethics submission for both study two and three was submitted to both CPRD Research Data Governance (RDG) committee and the London School of Hygiene and Tropical Medicine (LSHTM) ethics committee. It was approved by CPRD's RDG committee on 1 November 2022 (#22_002202) and by LSHTM's ethnic's committee (#28169) on 9 August 2022. For the final approved ISAC see Appendix C. Approved ISAC application.

5.3.9 General methodology

5.3.9.1 Code lists

Williams et al, 2017[161] defines code list creation as *"[…] the process of assembling a set of clinical codes that represent a single clinical concept such as a diagnosis, a procedure, an observation or a medication."*

Creating code lists is generally one of the first steps in studies of EHRs as it is required to develop the operational definitions for all of the study variables. Errors in code lists can lead to selection biases that can impact study results[162]. CPRD generally recommends researchers to develop search strategies to develop code lists. These are reusable lists of key terms that can be used within their CPRD code lists browsers to identify medcodes and prodcodes. Wildcards (*) can be utilised to identify terms with multiple endings. For example, myocardial infarct* can be used to identify myocardial infarction or myocardial infarct. These lists should be exhaustive since conditions can be recorded in the EMIS® system using different medcodes or prodcodes for synonyms of the same condition. For hierachial coding systems (e.g., Read and ICD-10) the hierachies can be utilised to identify additional codes. For SNOMED-CT, since it is an ontology, similar groups of codes can be identified under the same SNOMED concept ID.

CPRD Aurum has a new data release currently every 4-6 months. Between each release practices are added that have started using the EMIS® system. Patients are added if they transfer into a practice that uses the EMIS® system. Practices are removed between each release if they stop using the EMIS® system and patients are removed if they opt out of contributing their data.

As codes change between each release, new code lists are required when using an updated release. However, since majority of comorbidities are long recognised conditions that are identified using a long lookback period in CPRD Aurum, it is unlikely that code lists for these comorbidities will change dramatically from one year to the next. Therefore, lists using an older release for comorbidities will suffice. However, for newer conditions e.g., SARS-CoV-2 infection, it is likely that new medcodes will be added between each release and therefore it is important to develop code lists using the CPRD Aurum release that will be used in the research.

An overview of the methods used to search for previously developed lists and to develop new code lists are provided below. More details on the search terms applied and inclusion/exclusion criteria for each of the variables in study two and three are found in **Chapter 6** and **7**.

5.3.9.1.1.1 Searching repositories and the literature for published code lists

The general approach to identifying code lists was conducted in the following order:

1. To identify the highest quality code lists I first checked NHS England's QOF or reference sets for published code lists.
2. If no QOF code lists were found, then I searched for existing code lists in existing repositories that have been developed by well recognised research groups (e.g., LSHTM's Data Compass[163], Health Data Research UK's Phenotype Library[164], OpenSAFELY code lists[165] and Cambridge University Primary Care Unit – Code Lists (GOLD)[166]).
3. If no code lists were found in data repositories, then I searched the literature for possible lists.
4. If no existing list was found, then I developed my own lists (see below).

5.3.9.1.1.2 Methodology for developing code lists using key search terms

The methodology for developing my own code lists using key search terms were:

1. I developed key word searches for each code lists that were based on MESH terms (previously described in Section 4.3.1), using synonyms identified in the NHS England SNOMED-CT browser[167](NHS England browser provides preferred terms and then synonyms for related medical terms) and in systematic literature review or targeted literature review searches. Each of these search terms were reviewed by a clinical epidemiologist prior to running the searches in the code lists browser.
2. Searches were run in the CPRD code list R browser (see details below). All codes that were identified in the search were exported into Excel with one tab per variable.
3. Inclusion and exclusion criteria for reviewing the code lists were developed. For example, a code was excluded if they clearly demonstrated absence of the condition in question. I firstly reviewed each code list tab line-by-line and marked each code for inclusion or exclusion. For all codes that were excluded a reason for exclusion was provided e.g., evidence of absence.
4. Once reviewed, the inclusion/exclusion decision with reason for exclusion was reviewed by a clinical epidemiologist with knowledge of UK clinical practice (Dr Helen McDonald). For those codes were there were disagreement these were discussed with a third-party reviewer (Dr Edward Parker).

5.3.9.1.1.2.1 CPRD code list R browser

This browser was developed in-house at Evidera Ltd as a more advanced code browser tool to the CPRD Aurum code browser tool. It uses the flat files from CPRD and relationship files from NHS England's SNOMED-CT flat files[168]. Either medical or product code lists are generated from this tool. For internal intellectual property reasons, I cannot further describe information on this tool.

5.3.9.2 Variable creation for sociodemographic variables at index

Variable creation for each of the sociodemographic variables in CPRD Aurum HES-APC and ONS for both paper two and three was as follows:

Age

Only year of birth (yob) is available for adults in the CPRD Aurum database and therefore date of birth was imputed as 1 July-yob for all individuals. Age at index date (described further below) was estimated as index date minus date of birth. Age was categorised into 5-year categories (65-69, 70-74, 75-79, 80-84, 85-89, 90-94, 95+).

Sex

The 'gender' variable in the CPRD Aurum patient file was used. Individuals can either be classified as female, male, indeterminate or unknown in CPRD Aurum. In the sample there were no individuals with an unknown gender as all individuals were required to have an acceptability flag (see Section 5.3.3) which excludes individuals with unknown gender. Patients in both paper two and three with indeterminate gender were also excluded as there were so few patients and therefore they were excluded for patient confidentiality reasons.

Ethnicity

Ethnicity is recorded in CPRD Aurum as specified by the patient using medcodes. In CPRD Aurum there are currently over 260 different ethnicity medcodes that can be recorded by GPs. Recording of ethnicity medcodes in EMIS® software systems was previously incentivised under QOF (see Section 5.3.3.6). Ethnicity can also be recorded in HES, as specified by the patient, as a value within one of 16 ethnicity categories based on the definitions on the UK 2001 UK Census definitions. On the 5-category level this groups individuals into White, Asian, Black, Other, Missing)[169].

For research purposes, it is necessary to categorise each of the vast number of ethnicity medcodes from CPRD into categories. Professor Rohini Mathur previously categorised each of the medcodes within CPRD GOLD into 16 ethnicity categories also based on 2001 UK Census definitions[169]. These 16 categories can be further collapsed into 5 different categories. Mathur et al, 2014[155] estimated that 11.0% of currently registered patients have more than one ethnicity medcode, which can span over different ethnicity categories. Therefore, Professor Rohini Mathur also developed an algorithm to categorise ethnicity in the case of more than one record.

Professor Mathur's algorithm that uses ethnicity medcodes within CPRD Aurum, supplemented with ethnicity categories from HES is as follows:

1. Identify all SNOMED medcodes for ethnicity in the patient record.
2. Remove duplicate records that are recorded on the same date.
3. Assign each SNOMED code to the 5 or 16 ethnicity categories.
4. If individuals are assigned to more than one ethnicity category, then the most frequently recorded ethnicity category is assigned.
5. If there is a tie between the frequency of ethnicity categories then the most recently recorded is identified.
6. If there is still a tie, then the most recently recorded ethnicity category is used.
7. If ethnicity is still missing then ethnicity is imputed from HES. Rohini et al, 2013[155] estimated in the July 2017 CPRD GOLD build that when using linked CPRD GOLD-HES data, completeness of a usable ethnicity record increased from 78.7% in CPRD alone to 97.1% for the combined database.

Region

Region was defined using the 'region' variable in the CPRD Aurum practice file. The region variable is based on ONS region from January 2022. Region in CPRD Aurum (May 2022 release) is separated in 13 categories: None, North East, North West, Yorkshire and The Humber, East Midlands, West Midlands, East of England, London, South East, South West, Wales, Scotland, Northern Ireland.

5.3.9.3 Open Science

For both study two and three all the data management and analyses were conducted using R version 4.3.1 and above. The Github pages provided for both the data management and statistical analysis can be found here: https://github.com/grahams99/Health-seeking-behaviour. The code lists developed for this study and search terms used to identify codes are also publicly available and can be found in this location: https://doi.org/10.17037/DATA.00003684.

6 Chapter 6: Study two: Identifying markers of health-seeking behaviour in UK electronic health records

6.3 Introduction to the chapter

As highlighted in the pragmatic review above (**Chapter 3**) new methods are required to identify a systematic set of markers of health-seeking behaviour that can potentially be used to account for this type of confounding in EHRs. As previously highlighted, the selection of these markers needs to be informed by a conceptual framework to ensure the underlying phenomenon is appropriately accounted for. This chapter describes the methods that were used to meet this need.

The definitions of health-seeking behaviour, healthcare access and healthcare utilisation were previously provided in **Chapter 1** Section 1.8.1. The general methods used and description of the datasets were previously described in **Chapter 5. Chapter 7** below outlines the methods used to quantify and account for this type of confounding using these markers.

The current chapter will provide information on the conceptual framework used to identify the markers and variable creation of the markers . The majority of the methods, results and discussion are provided in the paper below.

6.4 Aim of chapter

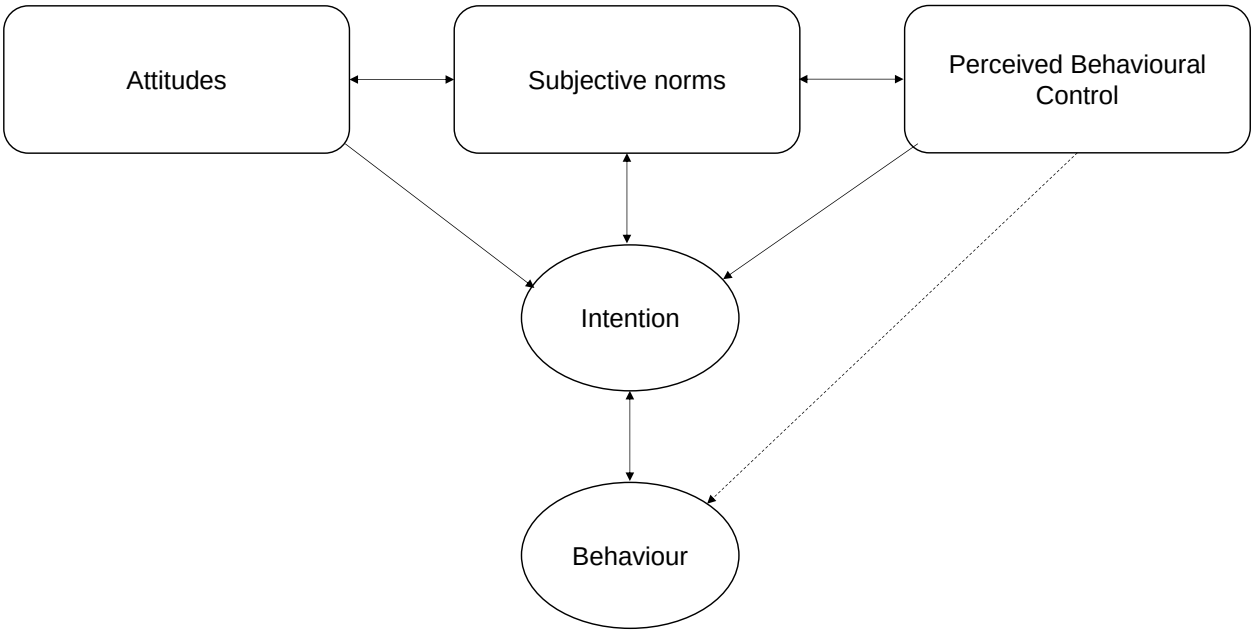
To systematically identify a set of markers of health-seeking behaviour available in EHRs that can be potentially used to quantify and account for this type of confounding.

6.5 The Theory of Planned Behaviour model

There are a range of models of determinants of healthcare uptake. The Theory of Planned Behaviour[170] was selected as a conceptual framework as it is a widely-used and accepted model. Barriers or influences of healthcare uptake can either be described on the micro-level e.g., socio-demographic influences such as age and gender, or on the macro-level e.g., geographical barriers or number of healthcare professionals. The Theory of Planned Behaviour model[170] describes the psychological barriers that influence healthcare uptake on the micro-level. The Theory of Planned Behaviour model was initially designed to predict an individual's uptake of a particular health behaviour (e.g., intention to get vaccinated). It describes how behaviours are influenced by three factors which are attitudes, subjective norms and perceived behavioral control (Figure 5 below). Some key definitions are provided below:

- Attitude towards the behaviour: this is personal attitude towards the behaviour, which is based on knowledge, attitudes and prejudices. For example, an individuals' belief about whether vaccination reduces their risk of infection.
- Subjective norms: this is personal perception of how other people view a specific behaviour i.e., social pressures. For example, how friends perceive receiving a vaccination.
- Perceived behavioral control: this is personal perception of the extent to which a behaviour is easy or difficult to conduct. For example, how easy or difficult it is to book a vaccination appointment.

Figure 5 Theory of Planned Behaviour model



Note: this figure is adapted from Azjen et al, 2005[171]. External factors can influenza behaviours regardless of intention, which is demonstrated by the dashed line.

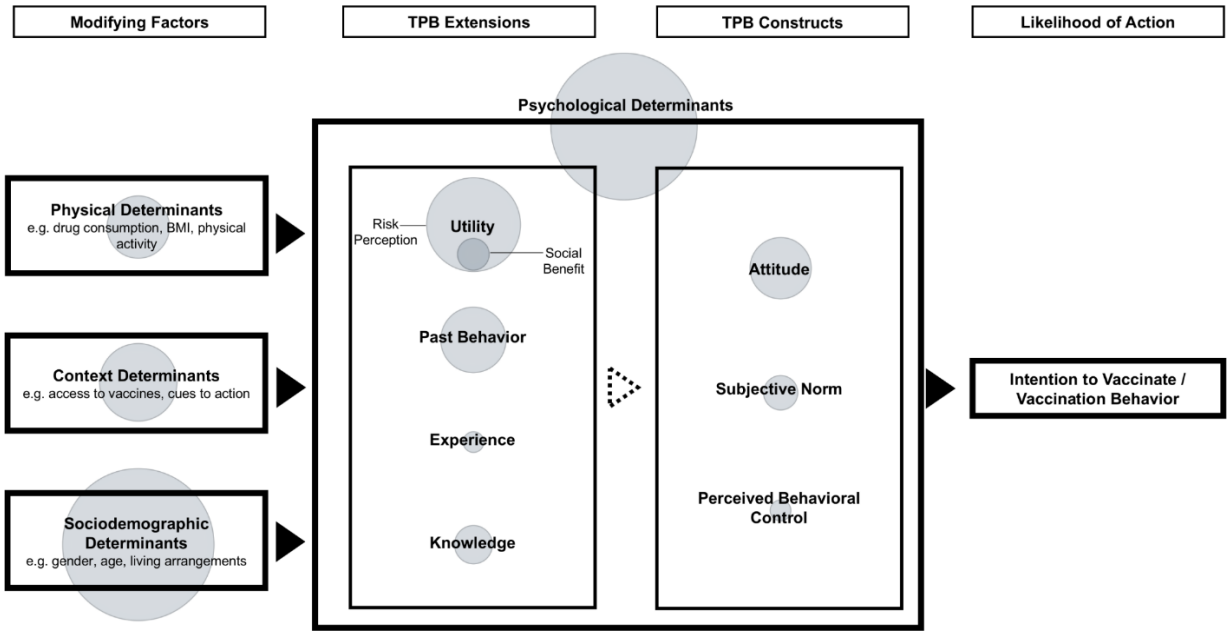
The Theory of Planned Behaviour only assesses the potentially relevant barriers/influences of healthcare uptake on the psychological level and therefore more recently authors have extended this model to include other factors such as physical, contextual, and sociodemographic aspects. Physical determinants might include underlying health conditions, lifestyle and physical activity. Contextual determinants might include GP influence and accessibility of healthcare. Sociodemographic determinants might include age, gender, ethnicity, socioeconomic status and living situation. The updated version of the Theory of Planned Behaviour model that includes these barriers/influences was first proposed by Schmid et al, 2017[172] and was developed to

understand influenza vaccination uptake (Figure 6). They adapted the original model to describe how utility, past behaviour, experience and knowledge influence attitude towards behaviour, subjective norms and perceived behavioral control. These can be defined as:

- Utility: the balance between risk perception of the vaccine preventable disease and social benefit associated with the vaccination.
- Past behaviour: whether individuals have received other vaccinations previously.
- Experience: whether individuals have been infected with the vaccine preventable infection previously.
- Knowledge: general knowledge about the vaccination and vaccine preventable condition.

This model was based on empirical and theoretical work that was based on a systematic review that included 470 research articles and the size of the cluster below is based on number of times that these determinants were reported in these articles. All of these barriers/influences are still on the micro-level.

Figure 6 Updated TPB model



Note: this figure is from Schmid et al, 2017[172]. Copyright for this figure is CC BY 4.0 Deed[173].

6.5.1 How the Theory of Planned Behaviour Model was used

Since health-seeking behaviour is a complex phenomenon, a range of markers needed to be identified that were influenced by different determinants (physical, contextual, sociodemographic and psychological) in the updated Theory of Planned Behaviour model. For example, the balance

of psychological determinants (such as social pressures and prior beliefs) and contextual determinants (such as GP encouragement and access to healthcare services) may differ for vaccination and cancer screening uptake. Markers with different underlying determinants were required to capture the influences of healthcare utilisation (i.e., the measurable outcome of health-seeking behaviour) as a whole.

The model was also used to cluster the identified markers according to how these markers were expected to behave in the data. Within clustered groups it was expected that markers would have similar strength and direction of associations with other markers. This is useful as it can help to inform expected associations with confounders, exposures and outcomes in observational research. The grouping of the markers according to this framework are detailed in paper two below.

6.6 Introduction to paper two

Paper two was submitted to *BMJ Open* on 27 September 2023 and is awaiting reviewer comments. This paper presents the methods that were used to identify markers of health-seeking behaviour in EHRs. Fifteen markers were systematically identified and the prevalence of these markers was described a population of individuals aged ≥66 years in linked UK EHR datasets (CPRD Aurum, HES APC and ONS linked - previously detailed in **Chapter 5**). The correlation between these markers was described using a data driven and theoretical approach to understand how these markers were expected to behave in relation to each other in these data.

Supplementary information for this paper is provided in Appendix D. Supplementary Materials Paper Two.

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

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

RESEARCH PAPER COVER SHEET

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

SECTION A – Student Details

Student ID Number	2005987	**Title**	Ms
First Name(s)	Sophie		
Surname/Family Name	Graham		
Thesis Title	Advancing methods to account for biases in vaccine effectiveness research		
Primary Supervisor	Edward Parker		

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

SECTION B – Paper already published

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

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

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	BMJ Open
Please list the paper's authors in the intended authorship order:	Graham, S., Walker J.L., Andrews, N., Nitsch, D., Parker, P. K. E., McDonald, H. I,
Stage of publication	**Submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I developed a detailed paper plan detailing my proposed analyses and this was reviewed by Helen McDonald and Edward Parker. I updated the plan based on their comments. I also developed the code lists for all of the markers. Firstly I created the search terms that would be run in the browser, these were then reviewed by Helen McDonald. Then I extracted all of the codes from the browser and reviewed each code list code-by-code. My inclusion/exclusion decisions and reason for exclusion were reviewed by Helen McDonald. I then updated each of the code lists based on her comments. In terms of the CPRD-HES-ONS data, I cleaned the data, conducted the data management and conducted all of the statistical analyses outlined in the detailed paper plan using R and R Studio. Then I wrote the first draft of this paper that was reviewed by Helen McDonald and Edward Parker. I updated the paper based on comments, sent it to all other authors for review and then updated the paper based on their comments. I submitted this paper and all required documents to BMJ Open.

SECTION E

Student Signature	Sophie Graham
Date	11/04/2024

Supervisor Signature	[redacted]
Date	24/04/2024

Identifying markers of health-seeking behaviour in UK electronic health records

Graham, S[1,2*]., Walker J.L[1,2,3]., Andrews, N[2,3]., Nitsch, D[1,4,5]., Parker, P. K[1,2]†. E., McDonald, H. I[1, 2,3,6]†.

1. Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK
2. National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Vaccines and Immunisation, London UK
3. UK Health Security Agency, London, UK
4. UK Renal Registry, Bristol, UK
5. Renal Unit, Royal Free London NHS Foundation Trust, Hertfordshire, UK
6. Faculty of Science, University of Bath, Bath, UK

*Corresponding author:

Sophie Graham (sophie.graham@lshtm.ac.uk)

†Equal contribution

Abstract

Objective

To assess the feasibility of identifying markers of health-seeking behaviour in UK electronic health records (EHR), for identifying populations at risk of poor health outcomes, and adjusting for confounding in epidemiological studies.

Design

Cross sectional observational study using the Clinical Practice Research Datalink (CPRD) Aurum pre-linked to Hospital Episode Statistics.

Setting

Individual-level routine clinical data from 13 million patients across general practices (GPs) and secondary data in England.

Participants

Individuals aged ≥66 years on 01/09/2019.

Main outcome measures

We used the Theory of Planned Behaviour (TPB) model and the literature to iteratively develop criteria for markers selection. Based on this we selected 15 markers: those that represented uptake of public health interventions, markers of active healthcare access/use and markers of lack of access/underuse. We calculated the prevalence of each marker using relevant lookback periods prior to index date (01/09/2019) and compared to national estimates. We assessed the correlation coefficients (phi) between markers with inferred hierarchical clustering.

Results

We included 1,991,284 individuals (mean age: 75.9 and 54.0% females). The prevalence of markers ranged from <0.1% (low-value prescriptions) to 92.6% (GP visits), and most were in line with national estimates; e.g., 73.3% for influenza vaccination in the 2018/2019 season, compared to 72.4% in national estimates. Screening markers e.g., abdominal aortic aneurysm screening were under-recorded even in age-eligible groups (54.3% in 65–69 year-olds vs 76.1% in national estimates in men). Overall, marker correlations were low (<0.5) and clustered into groups according to underlying determinants from the TPB model.

Conclusion

Overall, markers of health-seeking behaviour can be identified in UK EHRs. The generally low correlations between different markers of health-seeking behaviour suggest a range of variables are needed to capture different determinants of healthcare use.

Strengths and limitations of this study:

- This is the first known study in the UK that has identified proxies or markers of health-seeking behaviour or healthcare access.
- We utilised linked electronic health records from primary and secondary care so that a range of different health utilisation markers could be identified.
- We identified a large population of over 2 million individuals.
- For some of the markers (e.g., bone density scans), health need could not be entirely separated from health behaviour and access.
- Marker prevalences showed different patterns by age, and these findings might not be generalisable to younger age groups (<65 years).

Background:

Health-seeking behaviour can be defined as “any activity undertaken by a person believing [themselves] to be healthy, for the purpose of preventing disease or detecting it in an asymptomatic stage”[25]. Healthcare access can be defined as “the ability to obtain healthcare services such as prevention, diagnosis, treatment, and management of diseases, illness, disorders, and other health-impacting conditions”[174]. Healthcare professionals or researchers might be interested in identifying patients with a lack of health-seeking behaviour or healthcare access, since these individuals are likely to suffer from worse clinical outcomes. Health-seeking behaviour may also be a key confounder in observational studies, and failure to account for this may undermine the validity of results. This type of confounding is thought to have contributed to overestimates of the protective effect of influenza vaccinations against all-cause mortality in observational cohort studies[50]. Information on health-seeking behaviour can be collected prospectively through surveys or interviews; for example, in the English Longitudinal Study of Ageing study[175]. Typically, in routinely-recorded data such as electronic health records (EHRs) it is difficult to identify health-seeking behaviour since they are not directly recorded. Suitable markers would need to represent interactions with the healthcare system (i.e., healthcare utilisation), preferably with limited dependence on underlying health need. Behavioural scientists have a variety of models for explaining the determinants for healthcare utilisation. For example, the updated Theory of Planned Behaviour (TPB) model[172] describes the psychological , physical, contextual, and sociodemographic determinants for healthcare utilisation. Psychological determinants include influences on the micro and macro level such as societal attitudes, but also personal prior experiences. Physical determinants are on the micro-level and include lifestyle factors such as drug consumption, body mass index and physical activity. Context determinants are on the macro-level and include potential external barriers such as recommendations from healthcare professionals or geopolitical influences. Sociodemographic determinants are on the micro-level and include individual characteristics such as sex, age and living arrangements. These models demonstrate that there are a range of different determinants and therefore many different markers are likely required to capture all the underlying influences.

Three recent studies in the United States (US)[51,122,129] introduced adjusting for markers of health-seeking behaviour in observational research. However, it is not known to what extent suitable markers can be identified in UK EHR. This study aimed to identify markers of health-seeking behaviour in UK EHRs, compare their prevalence to available national estimates, and explore correlations between different markers. This study will focus on individuals aged

over 65 years as health-seeking behaviour varies by age[176] and because they have high morbidity and mortality[177].

Methods:

Data sources and population

We used the Clinical Practice Research Datalink (CPRD) Aurum pre-linked to Hospital Episode Statistics (HES) admitted patient care (APC). CPRD Aurum holds anonymised longitudinal primary care patient records collected from the EMIS® Health patient record system. At the time of data extraction (May 2022 release) this data included 1,491 currently contributing general practices for 13,300,067 currently contributing patients (19.83% of the UK population). 99% of the practices are in England and <1% are in Northern Ireland[178]. CPRD Aurum uses a combination of SNOMED, Read codes (Clinical Terms Version 3) and local EMIS codes that are each individually mapped to a unique "medcode". Prescriptions are recorded using the NHS dictionary of medicines and devices, each are mapped to a unique "prodcode". HES APC is a secondary care commissioning dataset that covers all NHS secondary care in England[133]. HES uses International Classification of Diseases 10th Revision (ICD-10) codes[179] to record diagnoses and Classification of Interventions and Procedures (OPCS) codes[180] to record procedures. Our study population included individuals in England aged 66 years or older on the 1 September 2019. We only included individuals with a GP practice registration start date before 1 September 2018 to allow for a minimum one-year pre-index period for marker identification.

Marker selection

We used the Theory of Planned Behaviour model to define our aim of identifying healthcare utilisation driven by determinants other than physical and mental health. We developed candidate markers and formal criteria for marker selection, incorporating input from two clinical epidemiologists on UK clinical practice and data recording (DN and HIM). Candidate markers from the aforementioned US studies[51,122,129] were tested against these criteria to iteratively make improvements to the criteria and identify additional potential markers. For all the markers identified from previous literature see **Supplementary Table 1**. The final criteria that were developed can be found in **Table 1** below**.** We selected fifteen markers that included abdominal aortic aneurysm (AAA) screening; breast cancer screening; bowel cancer screening; cervical cancer screening; influenza vaccination; pneumococcal vaccination; NHS health checks; prostate specific antigen (PSA) testing; bone density scans; low-value procedures; glucosamine use (low-value prescription); GP practice visits; did not attend (DNA) primary care visit; hospital visit for ambulatory care sensitive (ACS) condition; and blood pressure measurements. In general the criteria were a good fit for the markers,

but there was some tolerance for minor deviations, particularly for accepting some influence of underlying health conditions (**Supplementary Table 2)**.

Some markers represented active health-seeking behaviour, such as uptake of recommended vaccinations. Other markers represented lack of health-seeking behaviour – such as DNA for primary care visits, and hospital visits for ACS conditions. ACS conditions are conditions for which effective community care can help prevent the need for hospital admission[181]. If an individual has a visit to hospital for an ACS condition, then we can presume that they had a lack of healthcare access or health-seeking behaviour as they were unable to or did not access care when their symptoms were less severe. Low-value procedures and low value prescriptions are those that the National Institutes of Health and Care Excellence recommended to no longer provide in UK clinical practice since they were deemed to have little or no benefit, whilst still incurring an avoidable cost[182,183]. We considered both to be indicators of active health-seeking behaviour or healthcare access from a patient perspective, as patients were receiving (non-recommended) care for their perceived needs.

Table 1. Criteria used to assess inclusion of markers of health-seeking behaviour

#	Criteria	Explanation	Example of a marker that does not meet criteria
1	Should be currently or recently available in national clinical practice to all individuals (overall or by sex) at cohort entry.	Ensures that the denominator population (by sex) is eligible for each of the markers.	Shingles vaccination is currently recommended in the UK to all individuals turning 65 years (among others).[1] However, it was not historically available for all age-cohorts in this study due to the evolving age-based eligibility criteria since vaccine introduction in 2013. As a result, only selected age-cohorts would have had a period of age-based eligibility for shingles vaccination at the study index date, and this would not have been a universal marker for the study population.
2	Should be routinely recorded in the available data sources.	Ensures routine ascertainment of markers which is not dependent on other factors such as abnormal test results.	Vision and hearing tests are available through the NHS in the UK; however, most people get these tests at a private optician. Although opticians routinely send results to GPs, these may be uploaded as a PDF rather than coded in the patient's health record, particularly if no abnormality is found.
3	Should not be primarily dependent on underlying health needs.	Ensures that the determinants of healthcare utilisation are not primarily driven by underlying health conditions.	Adherence to medication could represent health-seeking behaviour; however, medication use is dependent on a diagnosed condition or health need.

Note: Shingles vaccine was first made available to immunocompetent individuals aged 70 or 79 in 2013 in the UK, with a phased catch-up programme for individuals aged 70–79 years. In 2021, the programme introduced recombinant vaccination which can be given to people with immunosuppression. At the time of the study index date, shingles vaccine was available to all individuals aged 70–79 years. Shingles vaccination is currently (1 September 2023) recommended in the UK to all individuals turning 65 years, currently aged 70–79, or aged 50 and over with immunosuppression.

Marker operational definitions

The operational definition of each marker includes code lists and lookback periods to apply in the current study datasets (**Table 2**).

For code lists, existing validated code lists were used where possible. Primarily we searched for code lists that were incentivised for national use through the Quality and Outcomes Framework[184] or those that were validated through research. If codelists were not available using these sources, then they were developed using key word searches (based on Medical Subject Headings terms with corresponding synonyms). Where possible the code lists aimed to be as specific as possible ("narrow code lists") and therefore codes were excluded if they were not clearly relevant. For example, for most screening markers we required the code to specify "screen" or "screening" but for bowel cancer we also allowed Faecal Immunochemical Tests as these are not used for symptomatic testing[185]. As a sensitivity analysis, for abdominal aortic aneurysm (AAA), breast cancer and cervical cancer screening, since the same procedure may be recorded for a screening test as for diagnostic tests investigating symptoms, we also included a broader code list that included codes that specified the relevant procedure, but did not specify "screen" or "screening". Full inclusion and exclusion list were reviewed by a clinical epidemiologist (HIM) and differences were agreed by discussion and third-party review (EP). The search terms that were used to create the code lists and the code lists that were used can be found on LSHTM data compass (https://doi.org/10.17037/DATA.00003684).

The lookback periods for each marker were developed by firstly identifying how each of these markers are recommended for use in current UK clinical practice. For markers that are available to all at any time, the lookback period reflected the expected frequency of healthcare use in UK clinical practice. For example, for markers that were expected to be frequently recorded (e.g., blood pressure measurements) we used a one-year lookback. For markers that were expected to be less frequently recorded (e.g., hospital visit for ACS conditions) a five-year lookback was used. For markers with an upper age limit of eligibility (i.e., screening and NHS health checks), we ensured the lookback period reflected timely administration of these markers (since we were interested in capturing strong evidence of health-seeking behaviour). For example, breast cancer screening is offered to women every 3 years aged 50-71 years[186] and therefore the lookback period covered the last 4-years of age-eligibility (3-years plus an additional year for uncertainty of age as only year of birth is recorded in CPRD) for breast cancer screening, until the index date. We included all follow-up time until index date to allow for delayed recording due to the transition to electronic records amongst older individuals. As a sensitivity analysis, since we were concerned that

including years after the upper age of eligibility might have meant we included more symptomatic individuals rather than healthy individuals accessing screening programmes, we also employed a restricted lookback that stopped the lookback at the upper age of eligibility (see **Supplementary Figure 1**).

Table 2. Use of markers in UK clinical practice and operational definitions

Marker	Use of this marker in current UK clinical practice	Operational definition	Sensitivity analysis	CPRD Aurum		HES APC	
				Medcode	Prodcode	ICD-10	OPCS
AAA screen	Available once to men when they turn 65 years[187].	≥1 AAA screen identified ever before index.	Alternatively using a broad code list †.	ü			
Breast cancer screen	Available every 3 years to women aged 50-71 years[186].	≥1 breast cancer screen identified from the last 4 years that they were age-eligible for screening until index date.	Alternatively using a restrictive lookback‡ and a broad code list†.	ü			
Cervical cancer screen	Available to women every 3 years between the ages of 25 and 49 years and every 5 years between the ages of 50 to 64 years[188].	≥1 cervical cancer screen identified from the last 6 years that they were age-eligible for screening until index date.	Alternatively using a restrictive lookback‡.	ü			
Bowel cancer screen	Available every 2 years to all individuals aged 60–74 years[185].	≥1 bowel cancer screen identified from the last 3 years that they were age-eligible for screening until index date.	Alternatively using a restrictive lookback‡.	ü			
NHS health checks	Available every 5 years to all individuals aged 40-74 years without pre-existing conditions*[189].	≥1 NHS health check identified from the last 6 years that they were age-eligible for NHS health checks until index date.	Alternatively using a restrictive lookback‡.	ü			
Influenza vaccination	Available annually to all individuals during the influenza season (1st September – 31st March) to all individuals aged ≥65 years[190].	≥1 influenza vaccination identified from 1 September 2018-31 March 2019. See Supplementary Table 3 for vaccination algorithm using both medcodes and prodcodes.	None	ü	ü		
Pneumococcal vaccination	Available once to all individuals when	≥1 pneumococcal vaccination identified ever before index.	None	ü	ü		

	they turn 65 years, or earlier for those with pre-existing conditions*[191].						
PSA test	Available to all men[192].	≥1 PSA test identified in the three years before index.	None	ü			
Bone density scans	Available to all individuals[193].	≥1 bone density scan identified in the three years before index.	None	ü			
GP practice visits	Available to all individuals[194].	≥1 GP practice visit(s) identified in the one year before index identified using [195]EMIS® consultation source identifiers, consultation source code identifiers and job categories to identify GP and nurse visits (excluding out-of-hours visits)[195].	None				
DNA primary care visit	Available to all individuals[194].	≥1 DNA primary care visits identified in the one year before index.	None	ü			
Low-value procedures	Available to all individuals[182].	≥1 low value procedures identified in the one year before index.	None				ü
Low value prescription (glucosamine)	Available to all individuals[183].	≥1 low value prescriptions identified in the one year before index.	None		ü		
Hospital visit for ACS condition	Available to all individuals[196].	≥1 hospital visits for an ACS condition identified in the five years before index.	None			ü (primary position only)	
Blood pressure measurements	Available to all individuals[197].	≥1 blood pressure measurement identified in the one year before index.	None	ü			

*Pre-existing conditions: chronic heart disease, chronic kidney disease, diabetes, high blood pressure, atrial fibrillation, transient ischaemic attack, inherited high cholesterol, heart failure, peripheral arterial disease, stroke, currently prescribed statins to lower cholesterol and previous checks that have found a 20% higher risk of getting cardiovascular disease over the next 10 years[189].

Abbreviations: CPRD: clinical practice research datalink; DNA: did not attend; GP: general practice; HES: hospital episode statistics; ICD-10: International Classification of Diseases 10th Revision; NHS: national health service; OPCS: Office of Population Censuses and Surveys; PSA: prostate-specific antigen.

†Broad code lists: for screening markers where the diagnostic test can be used for symptoms, broad code lists would include the diagnostic test, but did not require “screen” or “screening” to be in the medcode.

‡Restricted lookback: for markers with an upper eligible age, the lookback period would be stopped at the upper age of eligibility.

Prevalence estimates

For prevalence calculations, the denominator was all individuals aged ≥66 years on 1 September 2019 and the numerator was ≥1 occurrence of the marker in the relevant lookback period. We also calculated prevalence stratified by sex (given the inclusion of

several sex-specific markers) and age in 5-year bands (65-69, 70-74, 75-79, 80-84, 85-89, 90-95 and 95+ years).

We compared prevalence estimates to national estimates from PHE fingertips or from published literature, preferentially selecting for recent estimates from the UK in the relevant age group. The prevalence estimates from these sources can be found in **Table 4** and sources are detailed in **Supplementary Table 4**.

Correlations

The correlation of all the markers within the population sample was assessed using a phi correlation matrix. The phi coefficient is designed to measure the association between binary variables, and is equivalent to a Pearson correlation when applied to binary data. It ranges from -1 to 1, where 0 signifies no relationship between the variables, 1 is a perfect positive relationship and -1 is a perfect negative relationship[198]. Variables were ordered via complete linkage hierarchical clustering which was visualised by adorning dendrograms onto the correlation matrices (heatmaply_cor in R).

The clustering of markers was compared to a theoretical grouping using the updated TBP model[172]. The theoretical grouping was based on the underlying determinants from updated TBP model (**Table 4**). Specifically, we grouped markers into four groups: those with strong psychological influences ("psychologically determined"; e.g., vaccinations), those with strong contextual influences ("contextually determined"; e.g., screening and NHS health checks) and those fully or partially dependent on physical need. Physically determined markers were further separated into those likely to represent lack of health-seeking behaviour or healthcare access (e.g., DNA primary care visit and ACS condition hospital visit; "physically determined with lack of access") and those likely to represent active health-seeking behaviour or straightforward healthcare access ("physically determined with active access"). All programming was conducted using R (version 4.2.1-4.2.3) and the programming code can be found on Github (https://github.com/grahams99/Health-seeking-behaviour).

Results:

Overall, 1,991,284 individuals were included (54.0% females, mean (SD) age: 75.9 (7.4); **Supplementary Table 5**).

The prevalence of markers in the overall population ranged from <0.1% for low value prescriptions to 92.6% for GP visits. The proportion with at least one GP visit was so high that we conducted a *post-hoc analysis* that revealed the median (IQR) number of GP visits was 7 (4-11) with some patients having over 25 visits per year (**Supplementary Figure 2**). The prevalence of markers was similar between males and females, except for sex-specific

markers (**Table 4**). For screening and NHS health checks, broad code lists with standard lookback periods had the highest prevalence, whereas narrow code list with restrictive lookback had the lowest. For AAA screening and NHS health checks, changing the operational definition changed the prevalence <2% (**Supplementary Table 6**). The prevalence of most markers was in line with national estimates, particularly for the vaccinations, PSA testing and bone density scans. For example, 73.3% individuals in the current study had an influenza vaccination with national estimates reporting 72.4% influenza vaccination uptake among ≥65-year-olds in the 2019/20 influenza vaccination season[199]. The prevalence of screening and NHS health checks in the overall population was lower than national estimates, although this generally improved with comparison to currently eligible age-groups (**Figure 1**). Hospital visit for an ACS condition were higher than literature estimates as it was not possible to differentiate planned and unplanned hospitalisations in the current datasets (9.5% in current study vs. 0.1% in literature).

Table 4. Prevalence of markers

Variable	All Individuals	Male	Female	National estimates	Theoretical grouping from TBP model[#]
N	1,991,284	915,561	1,075,723		
AAA screen*	231,088 (11.6%)	227,844 (24.9%)	3,244 (0.3%)	76.1%	Contextual
Breast cancer screen*	346,116 (17.4%)	517 (0.1%)	345,599 (32.1%)	71.1%	Contextual
Cervical cancer screen*	397,303 (20.0%)	153 (0.0%)	397,150 (36.9%)	76.2%	Contextual
Bowel cancer screen*	1,439,412 (72.3%)	687,712 (75.1%)	751,700 (69.9%)	60.5%	Contextual
NHS health checks*†	372,244 (18.7%)	157,484 (17.2%)	214,760 (20.0%)	40%‡	Contextual
Influenza vaccine	1,460,391 (73.3%)	670,162 (73.2%)	790,229 (73.5%)	72.4%	Psychological
Pneumococcal vaccination	1,242,359 (62.4%)	568,798 (62.1%)	673,561 (62.6%)	69.0%	Psychological
PSA testing	352,272 (17.7%)	351,884 (38.4%)	388 (0.0%)	53.0%	Physical with active access
Bone density scan	100,892 (5.1%)	19,407 (2.1%)	81,485 (7.6%)	0.03-1.6%	Physical with active access
GP practice visits	1,844,823 (92.6%)	841,413 (91.9%)	1,003,410 (93.3%)	§	Physical with active access
DNA primary care visit	601,896 (30.2%)	275,449 (30.1%)	326,447 (30.3%)	§	Physical with lack of access
Low-value procedures	358,881 (18.0%)	168,746 (18.4%)	190,135 (17.7%)	0.02-0.2%	Physical with active access

Low-value prescription (glucosamine)	219 (0.0%)	75 (0.0%)	144 (0.0%)	§	Physical with active access
Hospital visit for an ACS condition	190,136 (9.5%)	82,691 (9.0%)	107,445 (10.0%)	0.1%	Physical with lack of access
Blood pressure measurement	1,470,006 (73.8%)	681,294 (74.4%)	788,712 (73.3%)	84.6%	Physical with active access

Abbreviations: AAA: abdominal aortic aneurysm; ACS: ambulatory care sensitive; DNA: did not attend; GP: general practice; PSA: prostate specific antigen; TPB: theory of planned behaviour.

*The denominator for the current study does not restrict to those that are age-eligible unlike in the national estimate. For age-eligible estimates see **Figure 1**.

†The denominator for the current study does not exclude individuals without pre-existing conditions (chronic heart disease, chronic kidney disease, diabetes, high blood pressure, atrial fibrillation, transient ischaemic attack, inherited high cholesterol, heart failure, peripheral arterial disease, stroke, currently prescribed statins to lower cholesterol and previous checks that have found a 20% higher risk of getting cardiovascular disease over the next 10 years[189]) unlike in the national estimate.

‡The national estimate for NHS health checks is the percentage of eligible individuals receiving an NHS health check in Q1 2019/2020: 2.0%. To better match the lookback period for this marker (5-years), we multiplied this estimate by 20.

§Prevalence of these markers are not knowingly presented in national estimates. NHS digital and OpenPrescribing report the total unit counts for these markers, which are reported in **Supplementary Table 4**.

[#]Theoretical grouping from the updated theory of planned behaviour model[172].

The prevalence of markers typically varied by age category, with a number of patterns evident (**Figure 1**). The recorded prevalence of markers with upper age eligibility (screening and NHS health checks) decreased with age (e.g., 28.0% in 65–69 year-olds vs 1.5% in 85–89 year-olds for NHS health checks), whereas the prevalence of ACS conditions, blood pressure measurements and vaccinations rose with age (e.g. 62.9% in 65–69 year-olds vs 80.5% in 85–89 year-olds for influenza vaccination). Although more common in younger age groups, screening marker prevalence still fell short of national estimates in currently eligible age-groups (e.g., 54.3% in 65–69 year-olds vs 76.1% in national estimates for AAA screening in men). PSA tests, bone density scans, low-value procedures, low-value prescriptions and DNA primary care visits peaked at 75-89 years, with lower prevalence in younger and older individuals. GP visits were consistent across age categories. As expected, the proportion of individuals with ≥1 GP practice visit was very high. The *post-hoc analysis* revealed the number of GP visits increased by age category until the last age strata (90+ years), when it decreased slightly (**Supplementary Figure 3**).

Using broad rather than narrow code lists, the estimated prevalences were similar for AAA screening across all age strata and for breast cancer screening in those aged 65-69 years. For all other breast cancer screening strata and for cervical cancer screening, broad code lists resulted in a higher prevalence than narrow. For standard versus restricted lookback periods, the prevalence was the same for individuals entering the cohort below the upper age of eligibility of that marker, whereas after this point there was lower prevalence in the restricted versus standard age strata.

In the overall study population, unsurprisingly, GP visits were strongly correlated with blood pressure measurements (phi φ 0.42) and influenza vaccination (0.33). Blood pressure measurements were also strongly correlated with influenza vaccination (0.23). Markers with the strongest negative correlation were blood pressure measurements and NHS health checks (-0.14) (**Figure 2**). Among males, GP visits and blood pressure measurements had the strongest positive correlation (0.45), followed by influenza and pneumococcal vaccinations (0.42). Other strong correlations included GP visits and pneumococcal vaccination (0.36) and blood pressure measurements and influenza vaccination (0.25) (**Figure 2**). Among females, GP visits were also strongly correlated with blood pressure measurements (0.40) and blood pressure measurements with influenza vaccination (0.30). There were also strong correlations between pneumococcal vaccination with influenza vaccination (0.39), bowel cancer screening and NHS health checks (0.23) (**Figure 2**).

Markers that were clustered together in the correlation matrices were: 1) blood pressure measurements, GP visits and influenza and pneumococcal vaccinations; 2) NHS health checks and bowel, cervical, breast cancer and AAA screening; and 3) ACS conditions, primary care DNA, bone density scans and low-value procedures. Markers from group 2) generally had a weak negative correlation with markers from group 3). When comparing these data-driven clusters with the theoretical grouping of markers there were some similarities. In both methods the “contextually determined” (i.e., NHS health checks and screenings) were grouped together as well as the “physically determined with a lack of healthcare access” (i.e., ACS conditions and primary care DNA). On the other hand, GP visits and blood pressure measurements were grouped with “psychologically determined” markers in the data-driven approach, but with the “physically determined with active healthcare access” in the theoretical grouping.

Discussion:

A statement of the principal findings

Overall, this study found that it is feasible to identify markers of health-seeking behaviour in UK EHRs. The prevalence of these markers ranged significantly and were generally in line with national estimates. Screening and NHS health checks were under-recorded in the EHR data, although prevalence was closer to national estimates amongst younger age groups that were currently eligible for these programmes. The prevalence and pattern of markers differed by age, with AAA screening declining with older age and hospital visits for ACS condition increasing. Correlations between markers revealed clusters that aligned well with theoretical groupings informed by the updated TBP model based on psychological, contextual and physical underlying determinants.

Strengths and weaknesses of the study

To our knowledge, this is the first study that has systematically identified proxies or markers of health-seeking behaviour or healthcare access using routinely collected data in the UK. Previous studies have adjusted for variables which may reflect confounding by health-seeking behaviour such as GP consultations, but without an explicit framework for selecting these. Our study demonstrates that a framework is beneficial since health-seeking behaviour are complex phenomena with multiple determinants, which may behave differently, and vary by age and sex. Linkage across primary and secondary care also strengthened this study as different types of healthcare utilisation with different underlying determinants could be captured. We included a large and representative cohort of over 2 million individuals aged 66 years and over in England. For some of the markers, older individuals might not have been historically eligible for services, which represents an important caveat during the interpretation of prevalence estimates. However, we accounted for this by calculating age-stratified prevalence. The study also only measured markers of health-seeking behaviour at a single point in time: these characteristics are not static, and individual behaviour and service accessibility can change over time. In addition, for some of the identified markers the influence of health need could not be entirely separated from health-seeking behaviour and therefore in some cases prevalence would be driven to some extent by health need. These findings might also not be generalisable to younger individuals where perhaps there are other contextual determinants to consider (e.g., occupation)[176].

Strengths and weaknesses in relation to other studies / discussing important differences in results

Previous studies that have used EHR to identify markers of health-seeking behaviour in the US[51,122,129] are in a considerably different context from the UK in terms of the healthcare system, claims-based recording systems and underlying determinants of health. This is likely to explain the different prevalence of markers identified in the current study. For example, the prevalence of pneumococcal vaccination was only around 11.4% in a study of ≥65 year olds identified in the Medicare database with an influenza vaccination during the 2019/2020 season[122], whereas the prevalence was 62.4% in the current study. These differences support the importance of context-specific markers of health-seeking behaviour.

Our study adds to a growing body of literature highlighting the potential to capture proxies of healthcare access and health-seeking behaviour. In prior studies in the US, these proxies were included as confounders during estimation of vaccine effectiveness, and they could play a similar role during observational studies in a UK context.[122]

The meaning of the study: possible explanations and implications for researchers, clinicians and policymakers

Based on the findings presented here, we propose several recommendations and considerations for researchers that wish to identify health-seeking behaviour in EHRs – whether to study healthcare use directly, or to quantify or adjust for confounding.

First, a range of different markers are required to fully represent both active health-seeking behaviour, or lack of these. Since health-seeking behaviour/healthcare access is such a complex phenomenon, it may be useful to include markers with different underlying determinants from the updated TBP model (psychologically, contextually and physically determined). If multiple markers are available, they can be included as separate confounders in multivariate models, or researchers may wish to consider tools such as high-dimensional propensity scores to guide study-specific confounder identification, prioritisation and adjustment[200].

Second, the optimal code lists will depend on the precise research question. Narrow code lists (e.g., using government incentivised code lists) can identify markers of health-seeking behaviour with high specificity. Broader code lists will capture more events, but may be more influenced by underlying health need. For markers with specific age-eligibility (e.g., screening or NHS health checks) look-back periods that restrict to time periods when individuals were age eligible improved specificity. However, more relaxed lookback periods might be preferred if there are expected to be artefacts in data recording such as transfer of historical information to electronic health records.

Third, prior to adjusting for health-seeking behaviour, interactions by age, sex and underlying health conditions should be considered. Markers that were recently introduced into clinical practice (e.g., AAA screening was introduced in the UK in 2013[201]) will likely decrease in prevalence with increasing age and can be supplemented with markers that increase with increasing age (e.g., ACS conditions). Otherwise, markers with relatively consistent prevalence across age strata are available (e.g., GP visits or blood pressure measurements). If markers that are restricted to specific sex (e.g., breast cancer screening) are utilised then these can be supplemented with markers of the opposite sex (e.g., AAA screening). For markers where there is some partial influence of underlying health conditions (e.g., pneumococcal vaccinations recommended to all but may be more highly prioritised among those with high-risk conditions) can be supplemented with markers that are administered to those that are healthier (e.g., NHS health checks).

Unanswered questions and future research.

Future researchers who are concerned with potential confounding from health-seeking behaviour in their study can use these markers to quantify and adjust for confounding. Where possible, a range of markers with different underlying determinants from the updated TPB model should be used and possible interactions by age, sex and underlying condition should be considered. Future research may identify key confounders within each theoretical group or cluster that are sufficient for confounding adjustment, although these are likely to be study-specific.

Common data models across datasets could increase efficiency and comparability of research investigating or adjusting for health-seeking behaviour, but future research is needed to identify suitable markers in alternative datasets and establish comparability. Additional markers may be identified in alternative datasets using the developed criteria.

Conclusion

Overall, markers of health-seeking behaviour can be identified in UK EHR, with prevalence estimates in line with national estimates. National screening programme estimates still fell short of national estimates even when restricting to currently eligible age groups. The generally low correlations between different proxy markers of health-seeking behaviour, and different age-profiles of markers, suggest a range of variables are needed to capture different determinants of healthcare use.

Contributors: Study concepts: SG, NA, JLW and HIM; study design: all authors; data acquisition: SG; Programming: SG, EPKP and HIM; Statistical analysis: SG; Supervision: EPKP, NA, JLW and HIM; Interpretation of results: all authors; Manuscript preparation: SG; Manuscript editing: all authors; manuscript review: all authors; Manuscript approval: all authors.

Funding: SG, EPKP, NA, JLW and HIM are funded by the National Institute of Health and Care Research (NIHR) Health Protection Research Unit in Vaccines and Immunisation (grant reference: NIHR200929), a partnership between UK Health Security Agency (UKHSA) and London School of Hygiene and Tropical Medicine. The views expressed are those of the author(s) and not necessarily those of the NIHR, UKHSA or the Department of Health and Social Care.

Disclaimer: The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or UKHSA.

Competing interests: SG is also a part-time salaried employee of Evidera, which is a business unit of Pharmaceutical Product Development (PPD), part of Thermo Fisher Scientific.

Patient consent for publication: Not required.

Ethics approval: The protocol for the study received scientific and ethical approval from the CPRD Research Data Governance committee (#22_002202).

Provenance and peer review: Not commissioned; externally peer reviewed.

Data availability statement: These data were obtained from the Clinical Practice Research Datalink, provided by the UK Medicines and Healthcare products Regulatory Agency. The authors' licence for using these data does not allow sharing of raw data with third parties. Information about access to Clinical Practice Research Datalink data is available here: https://www.cprd.com/research-applications. Codelists for this study are available at https://doi.org/10.17037/DATA.00003684 and code at https://github.com/grahams99/Health-seeking-behaviour.

Patient involvement statement: No patient or public involvement in the study as anonymised patient dataset was used in the analysis.

Figures titles and legends

Figure 1. Prevalence of markers, stratified by age category. Abbreviations: AAA: abdominal aortic aneurysm; DNA: do not attend; GP: general practice; NHS: National Health Service. Note: the numbers and proportions for these bar charts can be found in **Supplementary Table 7**.

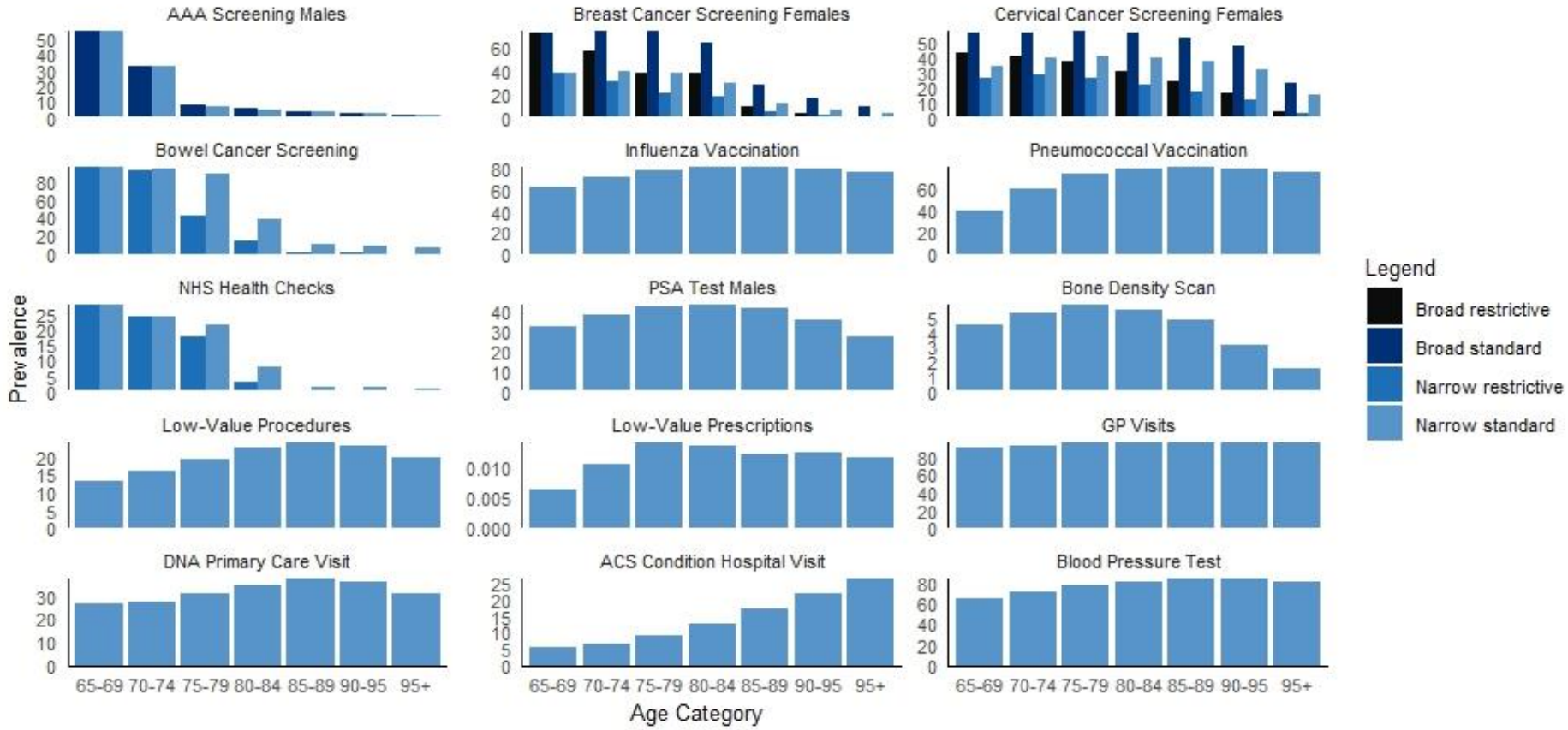
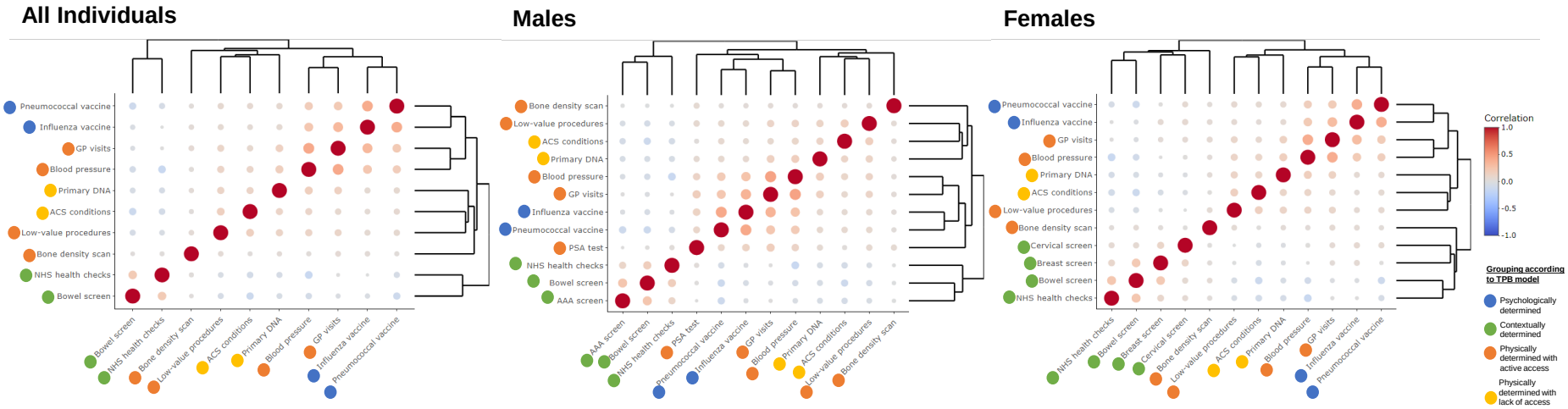


Figure 2. Correlation matrix plots. Abbreviations: ACS: ambulatory care sensitive; DNA: did not attend; NHS: National Health Service. The correlations are calculated using phi coefficient for binary variables. The clustering is visualised through the adorned dendrograms which are ordered via complete linkage hierarchical clustering. The size and the shading of the bubble represents the strength of the correlation. Note: the correlation coefficients for these plots can be found in **Supplementary Table 8-10**.



References:

1. Kasl SV, Cobb S. Health behavior, illness behavior, and sick role behavior. I. Health and illness behavior. *Arch Environ Health.* 1966;12(2):246-266.
2. University of Missouri. Health Care Access. https://medicine.missouri.edu/centers-institutes-labs/health-ethics/faq/health-care-access#:~:text=Health%20care%20access%20is%20the,and%20other%20health%2Dimpacting%20conditions. Published 2022. Accessed 05/10/2023, 2023.
3. Jackson LA, Jackson ML, Nelson JC, Neuzil KM, Weiss NS. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *Int J Epidemiol.* 2006;35(2):337-344.
4. English Longitudinal Study of Ageing. The data we collect. https://www.elsa-project.ac.uk/the-data-we-collect. Published 2002. Accessed 18/08/2023, 2023.
5. Schmid P, Rauber D, Betsch C, Lidolt G, Denker ML. Barriers of Influenza Vaccination Intention and Behavior - A Systematic Review of Influenza Vaccine Hesitancy, 2005 - 2016. *PLoS One.* 2017;12(1):e0170550.
6. Izurieta HS, Chillarige Y, Kelman J, et al. Relative Effectiveness of Influenza Vaccines Among the United States Elderly, 2018-2019. *J Infect Dis.* 2020;222(2):278-287.
7. Izurieta HS, Lu M, Kelman J, et al. Comparative Effectiveness of Influenza Vaccines Among US Medicare Beneficiaries Ages 65 Years and Older During the 2019-2020 Season. *Clin Infect Dis.* 2021;73(11):e4251-e4259.
8. Zhang HT, McGrath LJ, Wyss R, Ellis AR, Sturmer T. Controlling confounding by frailty when estimating influenza vaccine effectiveness using predictors of dependency in activities of daily living. *Pharmacoepidemiol Drug Saf.* 2017;26(12):1500-1506.
9. Cowling TE, Ramzan F, Ladbrooke T, Millington H, Majeed A, Gnani S. Referral outcomes of attendances at general practitioner led urgent care centres in London, England: retrospective analysis of hospital administrative data. *Emerg Med J.* 2016;33(3):200-207.
10. Salive ME. Multimorbidity in older adults. *Epidemiol Rev.* 2013;35:75-83.
11. Clinical Practice Research Datalink. Release Notes: CPRD Aurum May 2022. https://cprd.com/sites/default/files/2022-05/2022-05%20CPRD%20Aurum%20Release%20Notes.pdf. Published 2023. Accessed 05/09/2023, 2023.

12. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol.* 2017;46(4):1093-1093i.
13. World Health Organisation. ICD-10 Version:2019. https://icd.who.int/browse10/2019/en. Published 2019. Accessed 27/09/2023, 2023.
14. NHS England. OPCS-4 CODE. https://www.datadictionary.nhs.uk/data_elements/opcs-4_code.html. Published 2021. Accessed 27/09/2023, 2023.
15. NHS Digital. Unplanned hospitalisation for chronic ambulatory care sensitive conditions. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-outcomes-framework/may-2020/domain-2-enhancing-quality-of-life-for-people-with-long-term-conditions-nof/2-3-i-unplanned-hospitalisation-for-chronic-ambulatory-care-sensitive-conditions. Published 2020. Accessed.
16. The National Institute for Health and Care Excellence. NICE 'do not do' recommendations. https://www.nice.org.uk/media/default/sharedlearning/716_716donotdobookletfinal.pdf. Accessed 07/02/22.
17. NHS England. Items which should not be routinely prescribed in primary care. https://www.england.nhs.uk/medicines-2/items-which-should-not-be-routinely-prescribed/. Published 2023. Accessed 02/10/2023, 2023.
18. NHS Digital. Quality and Outcomes Framework (QOF). https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/quality-outcomes-framework-qof. Published 2022. Accessed 04/10/2023, 2023.
19. UK Government. Population screening programmes: bowel cancer. https://www.gov.uk/topic/population-screening-programmes/bowel Accessed.
20. UK Government. Population screening programmes: breast cancer. https://www.gov.uk/topic/population-screening-programmes/breast Accessed 07/02/22.
21. UK Government. Population screening programmes: abdominal aortic aneurysm. https://www.gov.uk/topic/population-screening-programmes/abdominal-aortic-aneurysm. Accessed 07/02/22.
22. UK Government. Cervical screening: programme overview. https://www.gov.uk/guidance/cervical-screening-programme-overview. Published 2021. Accessed 15/09/2023, 2023.
23. UK Government. NHS Health Checks: applying All Our Health. https://www.gov.uk/government/publications/nhs-health-checks-applying-all-our-

health/nhs-health-checks-applying-all-our-health. Published 2022. Accessed 15/09/2022, 2023.

24. UK Government. Greenbook Chapter 19. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/931139/Green_book_chapter_19_influenza_V7_OCT_2020.pdf Accessed 07/02/22.
25. UK Government. Greenbook Chapter 25. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/674074/GB_Chapter_25_Pneumococcal_V7_0.pdf. Accessed 07/02/22.
26. NHS England. PSA testing. https://www.nhs.uk/conditions/prostate-cancer/psa-testing/. Published 2021. Accessed 02/10/2023, 2023.
27. NHS England. Bone density (DEXA scan). https://www.nhs.uk/conditions/dexa-scan/. Published 2022. Accessed 02/10/2023, 2023.
28. NHS Digital. Appointments in General Practice report. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/appointments-in-general-practice. Published 2022. Accessed 02/10/2023, 2023.
29. Watt T, Sullivan R, Aggarwal A. Primary care and cancer: an analysis of the impact and inequalities of the COVID-19 pandemic on patient pathways. *BMJ Open.* 2022;12(3):e059374.
30. NHS England. Emergency admissions for Ambulatory Care Sensitive Conditions – characteristics and trends at national level https://www.england.nhs.uk/wp-content/uploads/2014/03/red-acsc-em-admissions-2.pdf. Published 2014. Accessed 02/10/2023, 2023.
31. NHS England. Blood pressure test. https://www.nhs.uk/conditions/blood-pressure-test/. Published 2023. Accessed 02/10/2023, 2023.
32. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J.* 2012;24(3):69-71.
33. UK Government. Fingertips, Public Health Data, Population Vaccination Coverage: Flu (aged 65 and older). https://fingertips.phe.org.uk/search/influenza#page/4/gid/1/pat/159/par/K02000001/ati/15/are/E92000001/iid/30314/age/27/sex/4/cat/-1/ctp/-1/yrr/1/cid/4/tbm/1 Published 2023. Accessed 14/02/2023, 2023.

34. Tazare J, Wyss R, Franklin JM, et al. Transparency of high-dimensional propensity score analyses: Guidance for diagnostics and reporting. *Pharmacoepidemiol Drug Saf.* 2022;31(4):411-423.
35. The Health Foundation. The Abdominal Aortic Aneurysm (AAA) screening programme. https://navigator.health.org.uk/theme/abdominal-aortic-aneurysm-aaa-screening-programme#:~:text=The%20NHS%20abdominal%20aortic%20aneurysm,previously%20could%20self%2Drefer). Published 2008. Accessed 31/07/2023, 2023.

6.7 Additional methodology

6.7.1 Detailed information about how markers are used in UK clinical practice

Table 9 below summarises the use of each of the fifteen identified markers in UK clinical practice which informed the operational definitions of these markers in the EHRs.

Table 9 Summary of the fifteen identified markers

Marker	Description	Introduced into UK clinical practice	Information on the healthcare interaction
Abdominal aortic aneurysm (AAA) screening	Offered as part of the UK government screening programme once to all men during the screening year (1 April to 31 March) that they turn 65 year[187]. Offered to all men, unless they have been treated for an AAA previously. AAA is a swelling of the aorta, which can cause life threatening bleeding if it ruptures[202]. Around 4% of men in the UK are estimated to have an AAA between the ages of 65 and 74 years[203].	The programme was fully rolled out across England in 2013[204].	AAA screening involves an ultrasound of the abdomen[202].
Bowel cancer screening	Offered as part of the UK government screening programme biannually to all individuals aged 60 to 74 years.	Originally the bowel cancer screening programme in the UK included a guaiac faecal occult blood testing that was offered biennially to those between the ages of 60 and 74 years. In 2013, flexible sigmoidoscopy ("Bowelscope") was added to this programme and was offered to all individuals once when they turned 55 years. However, from June 2019 this was replaced with immunochemical faecal occult blood testing (FOBT), otherwise known as faecal immunochemical test (FIT) tests.	FIT involves a stool sample that can be taken at home. Individuals that have an abnormal gFOBT or FIT test will be referred for a colonoscopy[185]. Colonoscopies may also be used directly for screening CEV individuals with rare conditions such as familial adenomatous polyposis[205].
Breast cancer screening	Breast cancer screening is offered as part of the UK government screening programme to all women every three years between the ages of 50 until their 71st birthday[206].	The programme began in England in 1988[207].	Breast cancer screening is provided using a mammogram[208]. High risk women with a history of breast irradiation will have an magnetic resonance

			imaging (MRI) annually between the ages of 25 and 39 years, but almost all of these women will have an MRI and mammogram by the age of 40 years[209]. Thermography is also offered privately, but not under NHS. After an abnormal mammogram an individual might be referred for a scintimammography, contract mammography or tomosynthesis for further investigation.
Cervical cancer screening	Cervical cancer screening is offered as part of the UK government screening programme to women and people with a cervix aged 25 to 64 years. Between the ages of 25 to 49 years screening is offered every three years and then from 50 to 64 years it is offered every 5 years[210].	A centralised cervical cancer screening programme was introduced in England in 1988[211].	Smears are taken using an extended tip spatula in GP surgeries by a nurse or doctor[211].
NHS Health checks	NHS health checks are provided as part of the UK government programme every 5 years to individuals aged between 40 and 74 years. They are offered to all individuals without pre-existing conditions*[212].	The programme was introduced in England 2009[213].	Checks are conducted at GP surgeries and usually include weight and height measurements, waist measurements, blood pressure test and a cholesterol test[212].
Influenza vaccination	Influenza vaccinations are provided as part of the UK government programme annually to all individuals with influenza at-risk conditions** and to any aged 65 years and older[190]. During the COVID-19 pandemic in 2021/22 the age of eligibility decreased to 50 years[214].	The programme was originally introduced in the 1960's to provide influenza vaccinations to individuals with at-risk conditions and in 2000, this was extended to over 65-year-olds[190].	Vaccinations are provided in a range of settings including pharmacies and GP surgeries[215].
Pneumococcal vaccination	Pneumococcal vaccinations are provided as part of the UK government programme once to individuals aged 65 years and over and to individuals with pneumococcal at-risk conditions***.	The programme was introduced in 1992 for individuals with at-risk conditions and this was extended to adults aged 65 years and older in 2003[191].	Vaccinations are provided at GP surgeries[216].
PSA testing	PSA tests are not currently provided as part of a government screening	PSA testing has been used as a diagnostic tool in the UK since	PSA testing is a blood test and can be conducted in a

	programme for prostate cancer in the UK[217]. Men are able to request a PSA test if they are concerned about prostate cancer and are asymptomatic or if they have symptoms. Men who have previous high PSA levels and are being monitored are also likely to receive future tests[218].	the late 1980s and early 1990s[218].	GP surgery or outpatient clinic[217].
Bone density scans	Bone density scans are not currently provided as part of a government programme for osteoporosis in the UK. However, they can be requested for individuals over 50 years with a risk of developing osteoporosis or for those with other risk factors such as smoking or a broken bone[193].	Dual-energy x ray absorptiometry (DEXA) scans were introduced into UK clinical practice in 1987[219].	DEXA scans are an x-ray and they occur within a clinic or hospital[220].
GP practice visits	GP practice visits are provided free at the point of access to anyone in the UK registered with a GP.	Since NHS creation in 1948[83].	In the UK, GP practice visits occur within primary care centres.
DNA primary care visit	DNA primary care visits is when an individual books a GP practice visit and then fails to attend that visit.	See above.	See above.
Low-value procedures	Low-value procedures are those that the National Institutes of Health and Care Excellence (NICE) no longer recommended provide in UK clinical practice since they were deemed to have little or no benefit to the patient, whilst still incurring an avoidable cost[182].	See above.	Refers to procedures conducted in the hospital.
Low-value prescriptions	As for low-value procedures, these are prescriptions that NICE no longer recommend[183].	See above.	Refers to prescriptions given in primary care.
Hospital visit for an ACS condition	ACS conditions are conditions for which effective community care can help prevent the need for hospital admission[181]. If an individual has a visit to hospital for an ACS condition, then we can presume that they had a lack of healthcare access or health-seeking behaviour as they were unable to or did not access care when their symptoms were less severe[181].	See above.	Refers to ACS conditions identified in-hospital.

Blood pressure measurements	Blood pressure tests are offered to individuals over the age of 40 to 74 years old as part of the NHS Health Checks, they are also offered to individuals that are worried about their blood pressure at anytime[197].	See above.	Offered in most pharmacies, local GPs and in some workplaces[197].

*Pre-existing conditions include heart disease, chronic kidney disease, diabetes, high blood pressure, atrial fibrillation, transient ischaemic heart attack, inherited high cholesterol (familial hypercholesterolemia), heart failure, peripheral arterial disease, stroke, the individual is currently being prescribed statins to lower cholesterol and previous checks that have found the individual to have a 20% or higher risk of getting cardiovascular disease over the next 10 years[212].

**Influenza at-risk conditions include chronic respiratory disease, chronic health disease and vascular disease, chronic kidney disease, chronic liver disease, chronic neurological disease, diabetes and adrenal insufficiency, immunosuppression, asplenia or dysfunction of the spleen, morbid obesity, pregnant women, household contacts of anyone with immunosuppression and carers[190].

***Pneumococcal at-risk conditions include asplenia or dysfunction of the spleen, chronic respiratory disease, chronic heart disease, chronic kidney disease, chronic liver disease, diabetes, immunosuppression, individuals with cochlear implants, individuals with cerebrospinal fluid leaks and occupational risk[191].

Abbreviations: AAA: abdominal aortic aneurysm; ACS: ambulatory care sensitive; DEXA: dual-energy x ray absorptiometry; FIT: faecal immunochemical test; gFOBT: guaiac faecal occult blood testing; GP: general practice; NHS: National Health Service; NICE: National Institutes of Health and Care Excellence; PSA: prostate specific antigen.

6.7.2 Variable creation for markers of health-seeking behaviour

For each of these markers that were identified, the general methodology for code list development has previously been detailed in **Chapter 5**. A more detailed methodology including the search terms used to identify each marker can be found in Table 10 below. Once the key searches had been run, codes were reviewed line-by-line and excluded if there was evidence of absence (e.g., if a screening offer was declined), where it was not clear if the event occurred (e.g., where only an invitation was sent) or where the healthcare service is not provided as part of routine UK clinical practice (e.g., sigmoidoscopy and endoscopies of the lower gastrointestinal tract are offered to individuals at high risk of bowel cancer[205]).

Table 10 Methodology used to develop code lists for markers of health-seeking behaviour

Marker	Published list used (if available)	Search terms applied in the CPRD R code browser	Published lists compared to and additional codes identified
AAA screening medcodes	None identified.	(aort|aneur|AAA) AND (scan|screen|ultra|imag|exam|abn|norm|detect|NHS|u/s)	None identified.
Breast cancer screening medcodes	None identified.	((breast|mamm) AND (scan|screen|abn|norm|lump|x-ray|detect|NHS)) OR (mammogr)	Health Data Research UK's Phenotype Library Read code list[221]. No additional codes were identified from this list.
Bowel cancer screening medcodes	None identified.	((bowel|occult|faecal|fecal|colon|rect|intest|digest|rect|colon|intest|digest|hema|FIT|FOB) AND (screen|exam|abn|norm|detect|NHS|positive|negative|test|kit|occult|immuno)) OR (gFOBt|qFIT|hemoccult|FOBt)	Read codes lists from a local general practice list[222] and SNOMED codes from Cancer Research[223]. No additional codes were identified.
Cervical cancer screening medcodes	QOF codes provided by NHS England [224].	Not relevant.	Not relevant.
NHS health checks medcodes	QOF codes provided by NHS England [224].	Not relevant.	Not relevant.
Influenza vaccination prodcodes	Prodcodes from Davidson et al, 2021[225].	Not relevant.	Not relevant.
Influenza vaccination medcodes	Medcodes from Davidson et al, 2021[225] were identified, however, since the influenza vaccination season differed to the current study new code lists were developed.	((flu|Tetra|trivalent|quadrivalent) AND (vacc|imm))OR(Fluad|Sequirus|Influvac)	Davidson's medcodes codes for influenza vaccination[225] and Cambridge university medcodes codes for influenza vaccination[166]. There were two additional codes that were added from Jennifer Davidson's list, but none from Cambridge University.
Pneumococcal vaccination prodcodes	Prodcodes from Davidson et al, 2021[225]..	Not relevant.	Not relevant.
Pneumococcal vaccination medcodes	Mecodes from Davidson et al, 2021[225].	Not relevant.	Not relevant.
PSA test medcodes	None identified.	(PSA) OR ((prostate)AND(antigen|meas|level|monitor|refer|couns|test))	None identified.
Bone density scan medcodes	None identified.	((bone|DEXA|DXA|dual|photon)AND(scan|ultra|imag|radio|dens|score|x-ray|energy|absorptiometry|result))OR(densitomet)	None identified.

Low-value procedures OPCS codes	OPCS codes identified from a local list of low-value procedures[226].	Not relevant.	Not relevant.
Low-value prescriptions prodcodes	None identified. Since there was no readily available list and since the list of low-value procedures is very long, only a code list for glucosamine was identified as an example.	glucosamine	None identified.
Primary care DNA medcodes	None identified.	((no|fail|poor|miss) AND (attend|show|encount|appoint|clinic)) OR (DNA)	None identified.
Hospital visit for ACS condition ICD-10 codes	ICD-10 codes from Carey et al, 2017[227].	Not relevant.	Not relevant.
Blood pressure test medcodes	QOF codes provided by NHS England [224]. However, since also interested in blood pressure measurements that were taken (not just the result, as in QOF) an additional search was conducted.	((blood|diastolic|systolic)AND(press)) OR (BP)	Cross compared the medcodes with Angel Wong list[*personal communication*] and no additional codes were identified.
GP visits	See approach summarised in Section 6.7.2.1 below.		

Abbreviations: AAA: abdominal aortic aneurysm; ACS: ambulatory care sensitive; DNA: did not attend; ICD: International Classification of Disease; QOF: quality outcome framework; NHS: National Health Service; SNOMED: Systematized Nomenclature of Medicine Clinical Terms.

6.7.2.1 GP practice visits

To identify GP practice visits as a marker of health-seeking behaviour, not all available visits in CPRD Aurum were included. This is because GP visits in CPRD Aurum can contain irrelevant information. For example, administrative information can be entered as a visit on the patient's file when a letter from a specialist or hospital is received, but this does not mean they necessarily had a visit.

Previously Watt et al, 2022[195] identified GP practice visits in CPRD Aurum using different variables across different CPRD Aurum files. Their study aimed to understand the impact of the COVID-19 pandemic on primary care and downstream cancer diagnoses. The algorithm used by Watt et al, 2022[195] is summarised in Appendix A. Additional Tables. All three components (variables "conssourceid", "consid" and "jobcat", which are previously defined in Section 5.3.3.2 in **Chapter 5**) were required to identify a GP visit in their algorithm. Additional values for the "courssourceid" variable were added as the current study used a more recent CPRD Aurum release than Watt et al, 2022[195] and the more recent release contained additional values for this variable. Out-of-hours visits were not included in the definition of GP visits as these are generally considered to be urgent care where the patients' GP practice is closed and therefore were regarded to reflect negative health-seeking behaviour.

6.7.3 Clustering of markers according to the Theory of Planned Behaviour

The way in which each of the markers of health-seeking behaviour were clustered according to determinants in the updated Theory of Planned Behaviour model can be found in Table 11 below. The association between these determinants and each of the markers were identified from the literature (see Table A1 Appendix A. Additional Tables). Overall, four potential groups were identified. These groups are discussed further in paper two above.

Table 11 Theoretical grouping of markers according to the updated Theory of Planned Behaviour

Marker	(1) Physical	(2) Context	(3) Sociodemographic	(4) Psychological	Grouping
AAA screen		x	x		2
Breast cancer screen		x	x		2
Cervical cancer screen		x	x		2
Bowel cancer screen		x	x		2
NHS health checks		x	x		2
Influenza vaccination		x	x	x	4
Pneumococcal vaccination		x	x	x	4
PSA testing	x				1 (active)
Bone density scans	x				1 (active)

GP practice visits	x				1 (active)
DNA primary care visit	x				1 (lack)
Low-value procedures	x	x	x		1 (active)
Low-value prescriptions	x	x	x		1 (active)
Hospital visit for an ACS condition	x		x		1 (lack)
Blood pressure measurement	x		x		1 (active)

Abbreviations: AAA: abdominal aortic aneurysm; ACS: ambulatory care sensitive; DNA: did-not attend; GP: general practice; NHS: National Health Service; PSA: prostate specific antigen.

6.8 Additional discussion of paper

This section presents a more detailed discussion of the results from the paper two above.

6.8.1 Narrow versus broad code lists and restricted versus standard lookback periods

For identification of the screening markers and NHS health checks, the main results in paper two above applied operational definitions that used narrow code lists and standard lookback periods (where relevant). Narrow code lists needed to include the word “screen” in the medcode description, whereas broad code lists additionally included codes for the relevant diagnostic tests where the word “screen” was not specified. Restricted lookbacks restricted to time periods where the individual was eligible for age-eligible markers, whereas standard allowed for these to be identified after this time point until index date.

To maximise sensitivity and specificity restricted lookbacks with narrow code lists are preferable. However, in the above paper, standard lookbacks were used instead because of electronic data transfer that was initiated in the late 1980s in the UK[228]. This is because individuals involved in the data transfer were paid to record event dates from the paper system into the electronic database. As speed was prioritised over accuracy, some event dates from the paper system were recorded as the date of electronic data transfer rather than the true event date. Therefore, for events that occurred <30 years before the index date in study two, these event dates might have been delayed and therefore recorded after the age they were age-eligible.

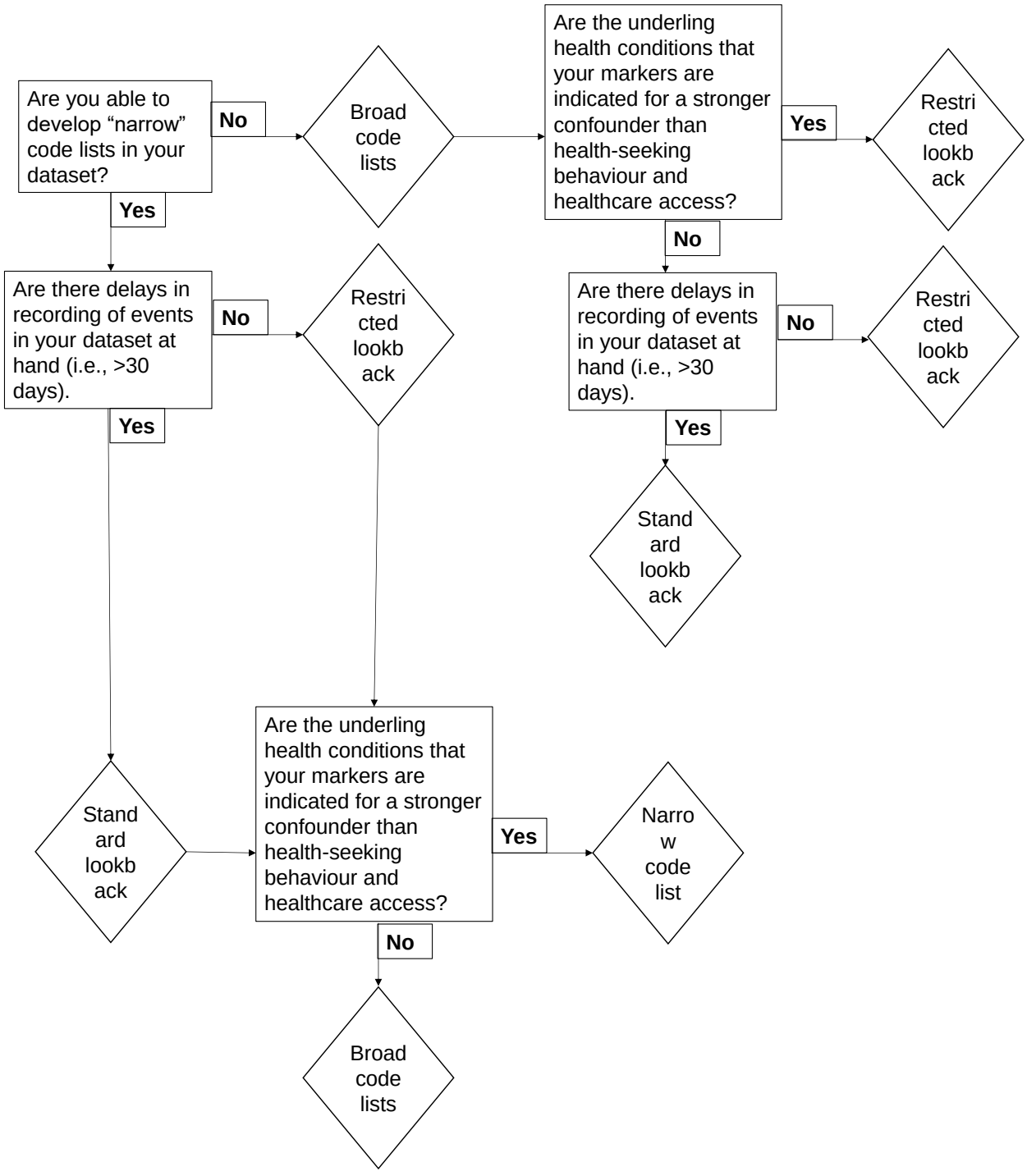
Future researchers that are deciding between narrow versus broad code lists and restricted versus narrow lookbacks should consider the following:

1) If including screening markers, whether it is possible to identify narrow code list in the dataset at hand. For some datasets that use less specific coding classifications it would not be possible to identify narrow code lists.

2) If including age-eligible markers, whether there might be potential delays in recording of these events in the data. If there are likely to be significant delays in data recordings, then standard lookbacks should be considered.
3) If including broad code lists, whether the association between the markers and health-seeking behaviour/healthcare access will be stronger than the association between the marker and the underlying conditions that these markers are indicated for. For example, if the association between NHS health checks and underlying health conditions is stronger than the association with health-seeking behaviour, then you might want to consider restricted lookback periods for age-eligible markers.

Below is a decision tree (Figure 7) to inform decision making on code list sensitivity and lookback periods.

Figure 7 Decision tree to inform broad versus narrow code lists and standard versus restricted lookback



Note: narrow code lists are those that specify that the diagnostic test was for a "screen", whereas broad code lists also include codes for the relevant diagnostic test. Restricted lookback periods only include lookback when the individual was eligible for an age-eligible marker, whereas standard lookback periods extend this time frame until index date.

6.8.2 Low prevalence of screening markers

The prevalence of most markers were in line with national estimates from the UK, however, screening and NHS health checks were under recorded in CPRD compared with national estimates, even in the age-specific strata (Figure 1 in paper two above).

For both NHS health checks and screening, the prevalence likely differs because the denominator population in the national estimate is different to the total population aged ≥66 years in England that were identified in study two. For NHS health checks, the denominator population is those eligible for an NHS health check and who were offered an NHS health check in the assigned time period[229]. For screening uptake in the national estimates the denominator population those invited or a screen. Uptake figures for a specific period will therefore only include people who are invited in that period. They will not include people who are not due at that time, or those who have been screened opportunistically if they are overdue for their test. However, it was decided to include all individuals 66 years and older as the denominator population for all the markers to ensure that the prevalence was relevant to a generalisable population, so that the prevalence could be replicated for other research questions.

In addition, the underestimation for screening could also be that only primary care was used to identify these events. For screening programmes which are delivered outside the GP surgery (bowel cancer screening, breast cancer screening and AAA screening), these events might be uploaded as attachments to the patient file without a medcode being recorded. This is likely to occur more frequently amongst normal test results compared with abnormal. These events would not be picked up in our definition of screening markers in the CPRD Aurum data. However, this explanation does not hold for cervical cancer screening which primarily occurs in GP practice surgeries and where the coding is incentivised by QOF; there is a large gap between even the broad code list and standard lookback definition in the youngest age group (65-69 years: 56.2%) and the national estimate (76.2%).

6.8.3 Correlation between markers in the data

The negative correlation between pneumococcal vaccinations and influenza vaccinations with nationwide screening programmes (AAA screening, bowel cancer screening and cervical cancer screening) and NHS health checks using the correlation matrices are likely due to presence of underlying health conditions. All individuals aged ≥65 years and those with CEV conditions[190,191] are prioritised for pneumococcal and influenza vaccinations and individuals with chronic comorbidites are not offered NHS health checks[230].

There was a strong correlation between GP visits, blood pressure measurements and the vaccinations. It is hardly surprising that these are all correlated since blood pressure measurements and vaccinations are commonly delivered in UK primary care. Schmid et al, 2017[172] identified that individuals with less frequent visits to their GP practice were less likely to get an influenza vaccination. It is also likely that in our blood pressure marker definition, individuals were captured with hypertension or hypotension who would require frequent visits to their GP and frequent blood pressure tests to be taken. Hypertension is also a risk factor for at-risk conditions for influenza vaccinations e.g., cardiovascular conditions[231].

The grouping of PSA tests, primary care DNA, ACS condition hospital visits, low-value procedures and bone density scans is also interesting. In the theoretical grouping of markers, primary care DNA and ACS condition hospital visits were grouped together as it was believed that these markers likely represented inability to easily access healthcare. This was thought to be unlike low value procedures and bone density scans which likely represented an ability to easily access healthcare. However, in the data driven approach low-value procedures and bone density scans were grouped together with PSA tests, primary care DNA and ACS hospital visits. Based on this it could be that it would be more appropriate to group the markers into three groups, rather than four. To further investigate this, in the next chapter the prevalence of these markers in COVID-19 and influenza vaccinated versus unvaccinated groups was assessed.

6.9 Overall chapter findings

Markers of health-seeking behaviour were identified in this chapter in UK EHRs. The identification of these markers was informed by the updated Theory of Planned Behaviour model and additional markers to those from the literature were identified through iteratively developed criteria. This is the first time that markers of health-seeking behaviour have been systematically identified, as previously authors have used inconsistent sets. The prevalence of these markers was mostly in line with national estimates which was reassuring to ensure the accuracy of identification of these markers in EHRs. There were some discrepancies for screening and NHS health check markers, which was likely due to differences in denominator populations. Overall, these markers could be clustered into four, and potentially three groups, based on how they were expected to behave in relation to other markers in the data. This clustering would inform how these markers are likely to behave in relation to vaccination exposures and infections in a cohort study of vaccine effectiveness (see next Chapter).

6.10 Unanswered questions

Although markers of health-seeking behaviour had been identified in UK EHRs, the following unanswered questions remained:

- To what extent are these markers associated with vaccine exposure and infections in a vaccine effectiveness study?
- To what extent can these markers be used to quantify and account for confounding from health-seeking behaviour in observational research questions?

6.11 How findings from this paper informed next chapter

These questions are addressed in the next chapter (**Chapter 7**), which used the above markers to quantify and account for confounding from health-seeking behaviour, using an influenza and COVID-19 vaccine effectiveness study as examples. The way in which the findings from the current chapter informed the next (**Chapter 7**) are:

- The narrow code lists and standard lookback periods for the markers were used.
- As the prevalence of some of the markers varied significantly by age (AAA screening, bowel cancer screening, NHS health checks and ACS conditions) interactions with age were fitted in the modelling step that included these markers.
- The data driven approach identified each of the markers into three instead of four groups. This gave some indication as to how these markers were likely to behave in association with the vaccination exposures and infections in the upcoming study. For example, it was likely that those in the same groups would all be similarly associated with influenza vaccinations.
- As the conceptual model demonstrated the different determinants of each of the marker it was known that it would be more appropriate to adjust for all of the markers at once together in the upcoming model steps, rather than combining all of these markers together into a single score.

7 Chapter 7: Study three: Quantifying and accounting for confounding from health-seeking behaviour in UK EHRs

7.1 Introduction to the chapter

As highlighted in the pragmatic review above (**Chapter 4**) there are very few vaccine effectiveness research studies that explicitly mention alternative methods to test-negative design and other study designs to account for confounding from health-seeking behaviour. Markers of health-seeking behaviour were identified in UK EHRs in **Chapter 6.** In the current chapter these markers were used to in a COVID-19 and influenza vaccine effectiveness study to identify, quantify and account for this type of confounding. The general methods, including datasets used and baseline variable creation were previously described in **Chapter 5**. The current chapter will provide additional information on the study concept and framework. The majority of the methods, results and discussion are provided in paper three below. After the paper additional information is provided on variable creation, the analytical methods and discussion.

7.2 Aim of chapter

7.3 To quantify and account for confounding from health-seeking behaviour in an influenza and COVID-19 vaccine effectiveness study. Study concept

Previous literature[44,45,55] has demonstrated that influenza vaccine effectiveness estimates are confounded, which is potentially due to health-seeking behaviour. To demonstrate whether this confounding was due to health-seeking behaviour, study three was designed as a cohort study design as this would allow for the confounding structures in the data to be most easily assessed. This analysis was also repeated for COVID-19 vaccine effectiveness, as it was unknown to what extent these estimates were also confounded by health-seeking behaviour. Since the influence of health-seeking behaviour on COVID-19 vaccine effectiveness estimates were unknown at the beginning of the pandemic, many authors were cautious and used the test-negative design. History of influenza vaccination (pre-COVID-19 pandemic) was also used as a negative control exposure (Section 1.9 in **Chapter 1**) against early COVID-19 pandemic SARS-CoV-2 infections. This negative control exposure was used to assess for any residual confounding after adjusting for the markers of health-seeking behaviour.

Therefore, there were three study populations included (that are further described in the paper below):

- Influenza cohort.
- COVID-19 cohort.
- Negative control exposure cohort.

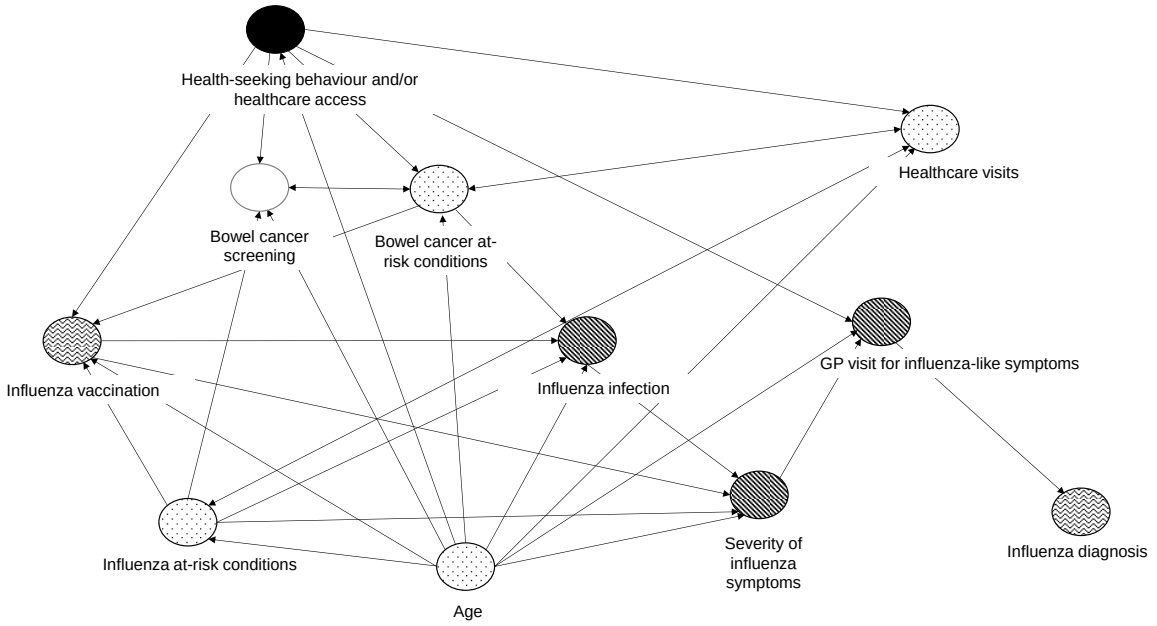
Consistent definitions and analyses were used across the three cohorts so that comparisons could be made. In each of these study populations, vaccine effectiveness was estimated from Cox regression models. Sequential adjustment of covariates was applied in each of the model steps (further described in paper three below) so that the hierarchical structures in data could be investigated. In the final step of the model all the markers of health-seeking behaviour from **Chapter 6** were adjusted for together.

7.4 Directed acyclic graph for proxy markers

Proxies markers identified in EHR data have been used in vaccine studies previously to account for confounding from other factors. For example, authors have previously used oxygen therapy, wheelchair use and arthritis and other proxies to account for differences in frailty and mobility in studies of influenza vaccine effectiveness[51].

Proxy markers that have previously been used to account for confounding from health-seeking behaviour are described in Table 6 in **Chapter 4**. For proxy markers to be sufficient to account for confounding, three factors need to occur. These factors are a) be strongly influenced by health-seeking behaviour and/or healthcare access, b) they need to not be strongly influenced by other measured confounders and c) they need to not be on the causal pathway from exposure to outcome. To understand whether these criteria are met, a DAG below (Figure 8) was used to demonstrate the potential relationships between bowel cancer screening[51,122,127,129] as an example proxy marker of health-seeking behaviour in an influenza vaccine effectiveness study.

Figure 8 DAG influenza vaccine effectiveness



Note: the arrows indicate the the causal structure of the data generating mechanism.. Exposure and outcome are indicated as zig-zag circles. Measured confounders are indicated as dotted, the directly unmeasurable confounder of health-seeking behaviour is indicated as black and the proxy marker as white. Other variables in the data are indicated as stripes. Other measured confounding e.g., region, ethnicity ect., are expected to have the same relationships in the data as age, which is why they have not been shown on the DAG.

In terms of criteria a) the association between health-seeking behaviour/healthcare access and bowel cancer screening has been described in qualitative studies. For example, one qualitative study conducted in Australia[232] reported that bowel cancer screening is more prevalent amongst non-smokers, non-drinkers, vegetable eaters and amongst those who are more physically active i.e., those with healthier behaviours.

In terms of criteria b) age would likely be a strong measured confounder. In terms of other measured confounders there might be some weak association between bowel cancer at-risk conditions and bowel cancer screening. For example, individuals that are high-risk for bowel cancer, receive a colonoscopy rather than the UK governmental screen. The conditions that qualify someone as high-risk include: familial adenomatous polyposis, Lynch syndrome, Serrated polyposis syndrome, strong family history of bowel cancer, ulcerative colitis or Crohn's disease, Polyps in the bowel or previous history of bowel cancer[205]. Since none of these conditions, except for previous history of bowel cancer, ulcerative colitis and Crohn's disease, are an influenza at-risk conditions and since these are rare conditions on the population level, the association in this direction is expected to be weak.

In terms of criteria c) it is unlikely that bowel cancer screening would be directly influenced by influenza vaccination and unlikely that bowel cancer screening would influence influenza diagnosis i.e., unlikely to be on the causal pathway from vaccination to infection.

Therefore, overall bowel cancer screening can be considered a good marker of health-seeking behaviour. For all the other markers identified in **Chapter 6**, similar associations were expected. Also as mentioned previously, since health-seeking behaviour is a complex phenomenon, multiple markers needed to be included all with varying underling determinants in the Theory of Planned Behaviour (Section 6.5). Therefore, all the markers identified in **Chapter 6** were used in the current study, with the exception of low-value prescriptions, as the prevalence of this marker was too low.

7.5 Introduction to paper three

This paper was submitted to *Journal of Infectious Diseases* on 12 April 2024 and is awaiting reviewer comments. The paper presents how the markers of health-seeking behaviour were used to quantify and account for confounding in observational research, using influenza and COVID-19 vaccine effectiveness as examples. History of influenza vaccination against early pandemic SARS-CoV-2 infections was also used as a negative control exposure to assess for residual confounding. Supplementary information for this paper is provided in Appendix E. Supplementary Materials Paper Three.

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

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

RESEARCH PAPER COVER SHEET

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

SECTION A – Student Details

Student ID Number	2005987	**Title**	Ms
First Name(s)	Sophie		
Surname/Family Name	Graham		
Thesis Title	Advancing methods to account for biases in vaccine effectiveness research		
Primary Supervisor	Edward Parker		

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

SECTION B – Paper already published

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

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

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Journal of Infectious Diseases
Please list the paper's authors in the intended authorship order:	Graham, S., Walker J.L., Andrews, N., Hulme, W.J., Nitsch, D., Parker, P. K. E., McDonald, H. I,
Stage of publication	**Submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I developed a detailed paper plan detailing my proposed analyses and this was reviewed by Helen McDonald and Edward Parker. I updated the plan based on their comments. I also developed the code lists for all of the vaccine exposures and infection outcomes. Firstly I created the search terms that would be run in the browser, these were then reviewed by Helen McDonald. Then I extracted all of the codes from the browser and reviewed each code list code-by-code. My inclusion/exclusion decisions and reason for exclusion were reviewed by Helen McDonald. I then updated each of the code lists based on her comments. In terms of the CPRD-HES-ONS data, I cleaned the data, conducted the data management and conducted all of the statistical analyses outlined in the detailed paper plan using R and R Studio. Then I wrote the first draft of this paper that was reviewed by Helen McDonald and Edward Parker. I updated the paper based on comments, sent it to all other authors for review and then updated the paper based on their comments. I submitted this paper and all required documents to Journal of Infectious Diseases.

SECTION E

Student Signature	Sophie Graham
Date	11/04/2024

Supervisor Signature	
Date	24/04/2024

Quantifying and adjusting for confounding from health-seeking behaviour in observational research

Graham, S.[1,2*], Walker J.L.[1,2,3], Andrews, N.[2,3], Hulme, W.J.[4], Nitsch, D.[1,5,6], Parker, E.P.K.[1,2†], McDonald, H.I.[1,2,3,7†]

1. Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom
2. National Institute for Health and Care Research Health Protection Research Unit in Vaccines and Immunisation, London United Kingdom
3. UK Health Security Agency, London, United Kingdom
4. The Bennett Institute for Applied Data Science, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom
5. UK Renal Registry, Bristol, United Kingdom
6. Renal Unit, Royal Free London NHS Foundation Trust, Hertfordshire, United Kingdom
7. Faculty of Science, University of Bath, Bath, United Kingdom

*Corresponding author:

Sophie Graham (sophie.graham@evidera.com)

†Equal contribution

Manuscript word count: 2,878 (excludes abstract, conclusion, tables and figures)

Abstract

Background

Health-seeking behaviour (HSB/HCA) are recognised confounders in many observational studies, but are not directly measurable in electronic health records. We used proxy markers of HSB/HCA to quantify and adjust for confounding in observational studies of influenza and COVID-19 vaccine effectiveness (VE).

Methods

This cohort study used primary care data pre-linked to secondary care and death data in England. We included individuals aged ≥66 years on 1 September 2019 and assessed influenza VE in the 2019/2020 season and early COVID-19 VE (December 2020 – March 2021). VE was estimated with sequential adjustment for demographics, comorbidities, and 14 markers of HSB/HCA. Influenza vaccination in the 2019/2020 season was also considered as a negative control exposure against COVID-19 before COVID-19 vaccine roll-out.

Results

We included 1,991,284, 1,796,667, and 1,946,943 individuals in the influenza, COVID-19 and negative control exposure populations, respectively. Markers of HSB/HCA were positively correlated with influenza and COVID-19 vaccine uptake. For influenza, adjusting for HSB/HCA markers in addition to demographics and comorbidities increased VE against infection from -1.5% (95%CI: -3.2,0.1) to 7.1% (5.4,8.7) with a less apparent trend for more severe outcomes. For COVID-19, adjusting for HSB/HCA markers did not change VE estimates against infection or severe disease (e.g., two doses of BNT162b2 against infection: 82.8% [78.4,86.3] to 83.1% [78.7,86.5]). Adjusting for HSB/HCA markers removed bias in the negative control exposure analysis (-7.5% [-10.6,-4.5] vs -2.1% [-6.0,1.7] before vs after adjusting for HSB/HCA markers).

Conclusion

Markers of HSB/HCA can be used to quantify and account for confounding in observational vaccine studies.

Background:

Health-seeking behaviour may be important confounders in observational research. Health-seeking behaviour is defined as seeking care for disease prevention, when asymptomatic or during early symptomatic stages,[233] and healthcare access as the ability to access healthcare services for these purposes.[234] Individuals with active health-seeking behaviour and the ability to easily access healthcare services generally have favourable clinical outcomes[235]. Confounding from health-seeking behaviour has previously led to overestimates of effectiveness of preventative therapies in observational research. For example, observational studies of statin use have consistently shown a reduction in hip fracture risk, even though this is not reflected in clinical trials[37]. Cohort studies of influenza vaccine effectiveness (VE), authors have reported reductions in all-cause mortality by 40-50%[42,43], despite influenza accounting for a maximum of 10% of deaths per year[44] However, systematically accounting for this type of confounding is challenging, as health-seeking behaviour are not directly measurable in routine healthcare data.

Proxy markers identified in electronic health records (EHRs) have been used to attempt to account for confounding from health-seeking behaviour[122,124-126,129]. Markers included vary considerably and optimal approaches are unclear. We recently identified a systematic set of fourteen markers of health-seeking behaviour in UK EHRs[236] that accounted for a range of determinants based on the Theory of Planned Behaviour model.[172] These markers represent healthcare system interactions that are only partially driven by an individual's underlying health need. In the current study, we aimed to assess whether these proxy markers of health-seeking behaviour can be used to quantify and adjust for confounding in observational studies, using seasonal influenza and COVID-19 VE as examples.

Methods:

Data sources

We conducted a cohort study using Clinical Practice Research Datalink (CPRD) Aurum[137] pre-linked to Hospital Episode Statistics (HES) Admitted Patient Care (APC)[133] and Office for National Statistics (ONS) data.[237] CPRD Aurum includes diagnoses (recorded using SNOMED, Read Coded Clinical Terms version 3 [CTV3], or local EMIS® codes, each mapped to an individual medcode), prescriptions (recorded using the Dictionary of Medicines and Devices [dm+d] codes, each mapped to an individual prodcode[238]), referral and testing information of patients registered to consenting GP practices in the UK. HES APC includes all admissions to NHS hospitals in England[133]. It includes inpatient hospital admission and discharge dates, diagnoses recorded using International Classification of Diseases 10th (ICD-10) Revision codes[179] and procedures recorded using Operating Procedure Codes

Supplement (OPCS) codes[180]. ONS includes date and underlying cause of death, recorded using ICD-10 codes, and socioeconomic data based on index of multiple deprivation (IMD)[239] which is based on small area geographical location. At the time of data extraction, CPRD Aurum included 13,300,067 currently contributing patients (19.8% of the UK population)[178].

Study design and population selection

We created separate cohorts to estimate influenza and COVID-19 VE. In addition, to assess potential residual confounding, we created a third "negative control exposure" cohort. Negative control exposures assume no causal mechanism between the negative control exposure and outcome, and confounding structures that reflect those of the primary exposure[58]. We used 2019/20 seasonal influenza vaccinations as a negative control exposure against COVID-19 infections before COVID-19 vaccinations were available in the UK.

We included all individuals aged ≥66 years (on 1 September 2019), who are prioritised for both vaccines and also likely to show distinct patterns of health-seeking behaviour access[176]. We required all individuals to have at least one year of registration prior to their index date, a record of 'acceptable' quality by CPRD, and linkage eligibility to HES APC and ONS. We excluded individuals with a death or registration end date before index, or with indeterminate sex (N=8). Individuals in the COVID-19 cohort were additionally excluded if their first vaccination was prior to 8 December 2020 as these likely reflected coding errors or trial participants (Figure 1 and Supplementary Table 1).

Outcomes, exposures, and follow-up

For all analyses we considered three nested outcomes of increasing severity: infections (based on primary care diagnosis, hospitalisation or death); hospitalisation/deaths; and deaths. All COVID-19 outcomes required a COVID-19 diagnosis code. For influenza we required a diagnosis of acute respiratory infection or influenza like illness (ARI/ILI)[225]. For all hospital and death outcomes the diagnosis code was required to be in the primary position.

We identified BNT162b2 and ChAdOx1 COVID-19 vaccines separately and requiring a minimum interval of 18 days between first and second doses[240]. We identified COVID-19 vaccinations using prodcodes records automatically recorded in GP records. For influenza, we identified vaccinations in the 2019/2020 season using both medcodes and prodcodes using an algorithm (Supplementary Table 2).

For influenza, the index date was 1 September 2019 and individuals were followed up until the earliest of death, transfer out of the practice or start of the COVID-19 pandemic (29

February 2020). For COVID-19, the index date was 8 December 2020, when COVID-19 vaccinations were introduced in the UK. Individuals were followed up until the earliest of death, transfer out of the practice, end of data availability (31 March 2021), first vaccination that was neither BNT162b2 or ChAdOx1, or second heterologous vaccination. For the negative control exposure analysis, the index date was 1 January 2020 when the first SARS-CoV-2 infections were identified in the UK. We included influenza vaccinations before 31 December 2019 by which time the majority of vaccinations in the UK have been delivered to reflect positive health-seeking behaviour/access, and prevent overlap with the outcome period. Individuals were followed-up until the earliest of death, transfer out of the practice, or the day before introduction of COVID-19 vaccinations in the UK (7 December 2020).

Sociodemographic variables

At index for each cohort we described: age (based on year of birth), sex, recent infection (<3 months pre-index for SARS-COV-2 or within the previous season for influenza), IMD, ethnicity[155], and influenza 'at-risk' conditions[241]. Influenza 'at-risk' groups[241] were identified from primary care records as described previously[225], grouped into immunosuppression or other conditions. We assessed missingness of ethnicity, IMD and region. For all other variables, absent codes were regarded as evidence of absence.

Markers of health-seeking behaviour

We used 14 markers of health-seeking behaviour that we identified through a framework based on the updated Theory of Planned Behaviour model[172], as described previously.[236] These included markers representing uptake of public health interventions (abdominal aortic aneurysm [AAA], breast cancer, cervical cancer and bowel cancer screening; influenza and pneumococcal vaccinations and NHS health checks), active healthcare access/use (prostate-specific antigen [PSA] testing, bone density scans, primary care visits, low value procedures[182] and blood pressure measurements) and lack of access/underuse (hospital visits for ambulatory care sensitive (ACS) conditions[181] and 'did not attend' primary care visits). Markers were identified in primary care and hospital records as described previously.[236] The lookback periods reflect use of these resources in UK clinical practice (Supplementary Table 1 with further details on all variable definitions).

Statistical analyses

We described sociodemographic variables, clinical variables, and markers of health-seeking behaviour at index, stratified by final vaccination status. To assess timeliness of vaccination, we calculated median days from index to first vaccination amongst vaccinated individuals, stratified by marker status and age categories (to reflect UK COVID-19 vaccination phased

deployment[125]). Outcome rates were represented by vaccination status as number of events divided by total person-years. We used cox regression models to estimate outcome risk in vaccinated versus unvaccinated individuals. A complete case analysis was conducted (excluding individuals with missing region, ethnicity or IMD). In the influenza and COVID-19 analyses, vaccination status was time-updated, with all individuals starting follow-up unvaccinated and vaccination status updated 14 days after a vaccination date (to provide time for immune response). For COVID-19, the analysis was brand specific. We assessed VE as [1 – hazard ratio] x 100.

We adapted a hierarchical modelling strategy[242] to understand the relationships between determinants of vaccine uptake in four steps. First, we fitted minimally-adjusted models adjusting for age (quadratic polynomial), sex, region and recent infection. Demography-adjusted models further adjusted for ethnicity and IMD. Comorbidity-adjusted models further adjusted for immunosuppressive status and other comorbidities. The fully-adjusted models further adjusted for health-seeking markers. For sex-specific markers (cervical cancer screening, breast cancer screening, AAA screening and PSA test), we included an interaction term with sex.

We conducted a sensitivity analysis fitting age interactions with AAA screening, bowel cancer screening, NHS health checks and ACS conditions (all of which vary markedly by age).[236]

Results

Study population

We included 1,946,943, 1,796,667 and 1,991,284 individuals in the influenza, COVID-19 and negative control exposure cohorts, respectively (Figure 2). Compared with individuals who remained unvaccinated, vaccinated individuals were more likely to be older, of White ethnicity, and live in less deprived areas (Table 1, and Supplementary Table 3).

Markers of health-seeking behaviour

Compared with individuals who remained unvaccinated, vaccinated individuals had a higher prevalence of all health-seeking markers (except ACS hospital visits, which should be prevented by healthcare access; Table 2). Differences in previous vaccinations were particularly marked. In the influenza analysis, 91.2% of vaccinated individuals had an influenza vaccination in the previous season, versus 22.6% of unvaccinated individuals for influenza vaccination, and a similar pattern was seen in COVID-19 analysis. Among vaccinated individuals, time-to-vaccination was not strongly associated with health-seeking marker status, except previous season influenza vaccination, which was associated with

faster uptake of both COVID-19 and influenza vaccines (Supplementary Figures 1 and 2 and Supplementary Tables 6 and 7).

Vaccine effectiveness estimates

For influenza, median (IQR) follow up time overall was 181 (0) days, which included 50 (28) days after first influenza vaccination. Unadjusted event rates ranged from 0.84 ARI/ILI-related deaths per 1,000 person-years during unvaccinated time to 117.15 influenza infections per 1,000 person-years after vaccination (Supplementary Table 4). Incremental adjustments across the models led to increased VE estimates. For influenza infections, we observed a negative VE in the minimally-adjusted baseline models of -5.5% (95%CI: -7.2,-3.9). Estimated VE increased to -1.5% (95%CI: -3.2,0.1) after adjusting for comorbidities, and to 7.1% (95%CI: 5.4,8.7) after adjusting for health-seeking markers. For severe outcomes, estimated VE increased from 42.5% (95%CI: 32.8,50.8) against ARI/ILI-related death in the baseline model to 47.5% (95%CI: 37.3,56.1) in the fully-adjusted model (Figure 3 and Supplementary Table 5).

For COVID-19 median (IQR) follow-up time was 113 (0) days overall, which included 64 (19) days after first BNT162b2 vaccination. Unadjusted event rates ranged from 0.54 COVID-19-related deaths per 1,000 person-years after two doses of BNT162b2 vaccination to 95.39 COVID-19 infections per 1,000 person-years during unvaccinated time (Supplementary Table 4). There was very minimal change in VE from the minimally-adjusted model to the fully-adjusted model that included all health-seeking markers (e.g. 2-dose VE against infection of 82.7% [95%CI: 78.3,86.2] and 83.1% [95%CI: 78.7,86.5], respectively). This was also the case for more severe outcomes (e.g. 2-dose VE against hospitalisation of 96.2% [95%CI: 93.0,98.0] and 92.3% [95%CI: 93.0,98.0] for minimally-adjusted and fully-adjusted models, respectively). For ChAdOx1, there was very limited follow-up time after two doses (Supplementary Table 5).

For the negative control exposure analysis, median follow-up time was 341 (0) days for both vaccinated and unvaccinated. Unadjusted event rates ranged from 1.39 COVID-19-related deaths per 1,000 person-years for unvaccinated individuals to 13.77 COVID-19-related infections per 1,000 person-years for influenza vaccinated individuals (Supplementary Table 4). We observed a negative VE for the effect of influenza vaccinations against COVID-19 for all minimally and demography-adjusted models (e.g. -6.4% [95%CI: -15.4,-1.9] and -12% [95%CI: -17.4,-6.9], respectively, against COVID-19-related mortality). For infections, negative VE persisted after adjusting for comorbidities (-7.5% [95%CI: -10.6,-4.5]), but not after including health-seeking markers in the fully-adjusted model (-2.1% [95%CI: -6.0,1.7]). For more severe endpoints, adjusting for comorbidities led to VE estimates consistent with a

null finding, which was also the case after additional adjustment for health-seeking markers (Figure 3 and Supplementary Table 5).

Sensitivity analyses including interaction terms between age and age-varying markers did not substantively change VE estimates (Supplementary Table 5).

Discussion

Using a range of markers of health-seeking behaviour we were able to address confounding in VE studies of influenza and COVID-19, with a negative control exposure analysis demonstrating successful control of confounding. This was assessed using a large cohort of individuals aged ≥66 years in England and confounding from health-seeking behaviour was adjusted for using proxy markers identified in UK EHRs. We found that influenza and COVID-19 vaccination uptake was higher in those with active health-seeking behaviour and better healthcare access. For VE the impact of health-seeking behaviour varied by context. Pre-COVID-19 pandemic, influenza VE against infections was underestimated when health-seeking behaviour was not adjusted for. This confounding was less apparent for more severe disease endpoints. For COVID-19 VE during a pandemic (during the early stages of COVID-19 vaccine implementation), minimally-adjusted models were very similar to fully-adjusted models that accounted for health-seeking behaviour. Residual confounding was initially present and successfully removed by adjusting for health-seeking behaviour in a negative control analysis of pre-pandemic influenza VE against early pandemic SARS-CoV-2 infections.

VE estimates from the comorbidity-adjusted models were similar to previous observational estimates. For influenza, a test-negative design study in the 2019/2020 season estimated VE against virology-confirmed disease to be 22.7% (95%CI: -38.5,56.9),[243] which is consistent with our estimate against ARI/ILI-hospital/death (24.7% [95%CI: 22.0 - 27.4]). For COVID-19, a cohort study from December 2020 to April 2021 estimated VE amongst individuals aged ≥65 years after 2 doses of BNT162b2 to be 84.7% (95%CI: 77.7%,89.5)[126] - consistent with our all infection estimate of 82.8% (95%CI: 78.4,86.3).

Our results differ from two US Medicare studies that assessed adjusting for proxy markers of health-seeking behaviour on influenza and shingles VE estimates[51,123]. Both studies saw a decrease in VE after adjusting for confounding from health-seeking behaviour, whereas we saw an increase. These discrepancies could be due to differences in healthcare settings or dataset types. The US studies[51,123] also used a smaller set of markers and therefore some residual confounding may have remained. One of the US studies[51] used pre-season influenza estimates as a negative control outcome for influenza VE and found significant residual confounding in their fully-adjusted model (32% [95%CI: 30,33%]).

In the negative control analysis we assumed that any plausible causal association between influenza vaccination and COVID-19 infection was minimal. Some studies with non-specific COVID-19 outcomes have shown there to be a minor protective effect of the influenza vaccination against COVID-19 infection[244]. A recent observational study conducted using administrative data in Canada also reported a protective effect of influenza vaccinations against COVID-19 infections, however, they also reported the same trend for previous health examination against COVID-19 infections (adjusted HR: 0.85 [95%CI: 0.78,0.91])[245]. The authors concluded that this provided evidence of residual confounding. Our study also identified and successfully removed residual confounding after adjusting for the health-seeking markers.

Future researchers will be able to use these markers to characterise health behaviours, to identify the strength and direction of confounding from health-seeking behaviour, and to account for identified confounding. We believe that particularly for seasonal influenza and COVID-19, these markers could be helpful to provide more accurate annual VE and cost-effectiveness estimates. They might also be important for chronic conditions (e.g., chronic kidney disease and diabetes), for which health-seeking behaviour have been shown to influence timeliness of seeking care and self-management.[246,247]

Usefulness of these markers is likely to vary by context. For example, they are likely to be more useful for routine rather than pandemic VE estimates. As we saw for the COVID-19 during the pandemic, sequential model adjustments had limited impact on VE estimates. This may be due to the high-risk perception of the virus and high testing and vaccination capacity during this time, but is likely to differ in a routine context. The descriptive results of this study are likely to be useful to clinicians and policymakers interested in the characteristics of individuals who are more likely to take up vaccinations and other nationwide programmes. We showed that individuals who take up UK nationwide screening programmes and NHS health checks are more likely to get vaccinated. Policy-makers could use this information to improve health equity.

Our study was strengthened by the large cohort and harmonised analyses with consistent variable definitions and modelling approaches both pre- and during the COVID-19 pandemic. Some previous VE analyses have adjusted for single variables that aim to capture health-seeking behaviour[124-126]. We included a set of proxy markers based on a theoretical model[236], providing a more systematic approach to adjusting for this complex phenomenon that can be used in other observational studies using routinely collected data. We were also able to identify and quantify residual confounding using a negative control exposure and demonstrate the impact of adjusting for health-seeking behaviour.

Despite these strengths, limitations remain. We assessed health-seeking behaviour at index date, but this might change over time, especially for the COVID-19 analysis, in which risk perception likely influenced health behaviours. There could potentially be other time-varying confounding[248] if for example, non-vaccination leads to infection and temporary ineligibility for vaccination[249]. There is scope for selection bias in the negative control exposure analysis; if individuals vaccinated against influenza in 2018/19 were less likely to die in the interim before the start of follow up for COVID-19 in January 2020, then this could overestimate VE slightly. In future, it would also be useful to understand how these markers perform in different age-groups, settings, study types and research questions, including designs that explicitly account for time-varying confounders[249].

Conclusion

We have identified markers in UK EHRs that can be used to quantify and adjust for confounding from health-seeking behaviour in observational research. Adjusting for health-seeking behaviour had a limited influence on estimates of COVID-19 VE during the pandemic early vaccine roll-out. For seasonal influenza VE, severe outcomes were robust to confounding from health-seeking behaviour, but VE against influenza infections were underestimated prior to adjustment for health-seeking behaviour. Residual confounding was also removed as demonstrated in a negative control exposure analysis of history of influenza vaccination against COVID-19 infections.

Acknowledgements

This work uses data provided by patients and collected by the NHS as part of their care and support (usemydata.org).

Ethics and CPRD requirements

We received data governance approval from CPRD (protocol #21_000737) and ethics approval from LSHTM's independent ethics committee (#28169).

Open science

All the analyses were conducted using R versions 4.2.2 to 4.2.4. All programming code from this project can be found in the Github repository: https://github.com/grahams99/Health-seeking-behaviour. The code lists and related search terms can be found on LSHTM Data Compass (https://doi.org/10.17037/DATA.00003684).

Funding: SG, EPKP, NA, JLW and HIM are funded by the National Institute of Health and Care Research (NIHR) Health Protection Research Unit in Vaccines and Immunisation (grant reference: NIHR200929), a partnership between UK Health Security Agency (UKHSA) and London School of Hygiene and Tropical Medicine. The views expressed are those of the author(s) and not necessarily those of the NIHR, UKHSA or the Department of Health and Social Care.

Disclaimer: The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or UKHSA.

Competing interests: SG is also a part-time salaried employee of Evidera, which is a business unit of Pharmaceutical Product Development (PPD), part of Thermo Fisher Scientific.

Data availability statement: These data were obtained from the Clinical Practice Research Datalink, provided by the UK Medicines and Healthcare products Regulatory Agency. The authors' licence for using these data does not allow sharing of raw data with third parties. Information about access to Clinical Practice Research Datalink data is available here: https://www.cprd.com/research-applications. Codelists for this study are available at https://doi.org/10.17037/DATA.00003684 and code at https://github.com/grahams99/Health-seeking-behaviour.

Tables

Table 1. Baseline characteristics stratified by vaccination status at the end of follow-up.

		Influenza analysis population N=1,796,667		COVID-19 analysis population N=1,796,667			Negative control exposure analysis population N=1,946,943	
Variable	**Category**	**Vaccinated N=1,473,955**	**Unvaccinated N=517,329**	**ChAdOx1-S N=845,428**	**BNT162b2 N=811,740**	**Unvaccinated N=139,499**	**Vaccinated N=1,437,356**	**Unvaccinated N=509,587**
Age category in years, N (%)	65-69	295,808 (20.1%)	152,255 (29.4%)	196,994 (23.3%)	93,761 (11.6%)	29,203 (20.9%)	288,285 (20.1%)	154,638 (30.3%)
	70-74	409,473 (27.8%)	148,126 (28.6%)	318,400 (37.7%)	185,076 (22.8%)	37,479 (26.9%)	400,663 (27.9%)	149,441 (29.3%)
	75-79	312,214 (21.2%)	89,076 (17.2%)	186,454 (22.1%)	180,327 (22.2%)	24,518 (17.6%)	305,940 (21.3%)	88,128 (17.3%)
	80-84	235,026 (15.9%)	60,466 (11.7%)	65,793 (7.8%)	191,890 (23.6%)	21,093 (15.1%)	229,632 (16.0%)	58,105 (11.4%)
	85-89	142,171 (9.6%)	39,664 (7.7%)	44,040 (5.2%)	110,191 (13.6%)	15,032 (10.8%)	137,602 (9.6%)	36,230 (7.1%)
	90-95	60,924 (4.1%)	20,019 (3.9%)	24,562 (2.9%)	41,348 (5.1%)	8,520 (6.1%)	58,241 (4.1%)	17,037 (3.3%)
	95+	18,339 (1.2%)	7,723 (1.5%)	9,185 (1.1%)	9,147 (1.1%)	3,654 (2.6%)	16,993 (1.2%)	6,008 (1.2%)
Sex, N (%)	Female	795,391 (54.0%)	280,332 (54.2%)	458,466 (54.2%)	442,085 (54.5%)	73,243 (52.5%)	776,347 (54.0%)	276,181 (54.2%)
	Male	678,564 (46.0%)	236,997 (45.8%)	386,962 (45.8%)	369,655 (45.5%)	66,256 (47.5%)	661,009 (46.0%)	233,406 (45.8%)
Ethnicity*, N (%)	Asian	49,874 (3.4%)	18,087 (3.5%)	27,362 (3.2%)	26,576 (3.3%)	9,891 (7.1%)	48,257 (3.4%)	19,013 (3.7%)
	Black	21,970 (1.5%)	14,942 (2.9%)	12,665 (1.5%)	9,685 (1.2%)	11,631 (8.3%)	21,023 (1.5%)	15,431 (3.0%)
	Mixed	6,321 (0.4%)	3,751 (0.7%)	3,744 (0.4%)	3,223 (0.4%)	2,242 (1.6%)	6,095 (0.4%)	3,853 (0.8%)
	Other	11,295 (0.8%)	6,416 (1.2%)	6,750 (0.8%)	5,942 (0.7%)	3,841 (2.8%)	10,982 (0.8%)	6,611 (1.3%)

		Influenza analysis population N=1,796,667		COVID-19 analysis population N=1,796,667			Negative control exposure analysis population N=1,946,943	
Variable	**Category**	**Vaccinated N=1,473,955**	**Unvaccinated N=517,329**	**ChAdOx1-S N=845,428**	**BNT162b2 N=811,740**	**Unvaccinated N=139,499**	**Vaccinated N=1,437,356**	**Unvaccinated N=509,587**
	White	1,331,686 (90.3%)	428,068 (82.7%)	754,194 (89.2%)	734,394 (90.5%)	94,638 (67.8%)	1,299,673 (90.4%)	418,902 (82.2%)
	Missing	52,809 (3.6%)	46,065 (8.9%)	40,713 (4.8%)	31,920 (3.9%)	17,256 (12.4%)	51,326 (3.6%)	45,777 (9.0%)
Region, N (%)	East Midlands	30,432 (2.1%)	[Redacted]	18,636 (2.2%)	14,776 (1.8%)	1,892 (1.4%)	29,726 (2.1%)	[Redacted]
	East of England	70,073 (4.8%)	25,340 (4.9%)	39,048 (4.6%)	37,023 (4.6%)	5,426 (3.9%)	67,180 (4.7%)	21,300 (4.2%)
	London	179,896 (12.2%)	83,929 (16.2%)	86,705 (10.3%)	115,643 (14.2%)	37,022 (26.5%)	174,002 (12.1%)	86,226 (16.9%)
	North East	51,140 (3.5%)	16,138 (3.1%)	30,256 (3.6%)	28,324 (3.5%)	3,873 (2.8%)	50,011 (3.5%)	16,579 (3.3%)
	North West	288,382 (19.6%)	93,210 (18.0%)	165,164 (19.5%)	158,595 (19.5%)	23,774 (17.0%)	282,152 (19.6%)	93,857 (18.4%)
	South East	330,361 (22.4%)	108,046 (20.9%)	192,498 (22.8%)	175,053 (21.6%)	27,873 (20.0%)	323,518 (22.5%)	109,375 (21.5%)
	South West	202,112 (13.7%)	68,372 (13.2%)	126,645 (15.0%)	103,675 (12.8%)	14,764 (10.6%)	196,664 (13.7%)	62,755 (12.3%)
	West Midlands	262,952 (17.8%)	94,411 (18.2%)	151,012 (17.9%)	151,300 (18.6%)	21,235 (15.2%)	256,602 (17.9%)	94,795 (18.6%)
	Yorkshire and The Humber	58,599 (4.0%)	16,206 (3.1%)	35,386 (4.2%)	27,336 (3.4%)	3,630 (2.6%)	57,467 (4.0%)	16,393 (3.2%)
	Unknown	8 (0.0%)	[Redacted]	78 (0.0%)	15 (0.0%)	10 (0.0%)	34 (0.0%)	[Redacted]
IMD, N (%)	1 (least deprived)	389,007 (26.4%)	110,866 (21.4%)	219,539 (26.0%)	213,479 (26.3%)	25,049 (18.0%)	381,652 (26.6%)	110,457 (21.7%)
	2	352,561 (23.9%)	115,220 (22.3%)	198,707 (23.5%)	195,168 (24.0%)	26,339 (18.9%)	344,778 (24.0%)	113,011 (22.2%)

		Influenza analysis population N=1,796,667		COVID-19 analysis population N=1,796,667			Negative control exposure analysis population N=1,946,943	
Variable	**Category**	**Vaccinated N=1,473,955**	**Unvaccinated N=517,329**	**ChAdOx1-S N=845,428**	**BNT162b2 N=811,740**	**Unvaccinated N=139,499**	**Vaccinated N=1,437,356**	**Unvaccinated N=509,587**
	3	291,822 (19.8%)	107,489 (20.8%)	165,923 (19.6%)	164,259 (20.2%)	27,283 (19.6%)	283,270 (19.7%)	103,275 (20.3%)
	4	245,959 (16.7%)	99,005 (19.1%)	143,497 (17.0%)	136,591 (16.8%)	30,900 (22.2%)	239,223 (16.6%)	97,624 (19.2%)
	5 (most deprived)	194,606 (13.2%)	84,749 (16.4%)	117,762 (13.9%)	102,243 (12.6%)	29,928 (21.5%)	188,433 (13.1%)	85,220 (16.7%)
Influenza 'at-risk' conditions, N (%)	Immunosuppressed status	44,445 (3.0%)	10,708 (2.1%)	20,560 (2.4%)	20,314 (2.5%)	3,181 (2.3%)	51,304 (3.6%)	11,491 (2.3%)
	Other comorbidities***	875,433 (59.4%)	227,851 (44.0%)	455,435 (53.9%)	478,945 (59.0%)	70,209 (50.3%)	859,858 (59.8%)	224,804 (44.1%)
Markers of health-seeking behaviour, N (%)	AAA screen	171,329 (11.6%)	59,759 (11.6%)	76,400 (9.4%)	127,455 (15.1%)	11,087 (7.9%)	167,605 (11.7%)	59,711 (11.7%)
	Bowel screen	1,063,252 (72.1%)	376,160 (72.7%)	556,813 (68.6%)	703,450 (83.2%)	94,562 (67.8%)	1,043,974 (72.6%)	380,264 (74.6%)
	Breast screen	268,370 (18.2%)	77,746 (15.0%)	149,098 (18.4%)	161,443 (19.1%)	16,564 (11.9%)	265,104 (18.4%)	78,616 (15.4%)
	Cervical screen	308,261 (20.9%)	89,042 (17.2%)	171,781 (21.2%)	171,059 (20.2%)	20,103 (14.4%)	301,116 (20.9%)	87,878 (17.2%)
	NHS health checks	278,539 (18.9%)	93,705 (18.1%)	134,883 (16.6%)	170,284 (20.1%)	15,862 (11.4%)	270,812 (18.8%)	92,865 (18.2%)
	Influenza vaccine†	1,343,562 (91.2%)	116,829 (22.6%)	665,364 (82.0%)	641,842 (75.9%)	56,223 (40.3%)	1,314,908 (91.5%)	112,149 (22.0%)
	Pneumococcal vaccine	1,071,867 (72.7%)	170,492 (33.0%)	575,351 (70.9%)	523,544 (61.9%)	59,781 (42.9%)	1,069,249 (74.4%)	164,248 (32.2%)
	ACS hospital care visit	146,447 (9.9%)	43,689 (8.4%)	85,734 (10.6%)	87,757 (10.4%)	17,314 (12.4%)	153,108 (10.7%)	42,755 (8.4%)

		Influenza analysis population N=1,796,667		COVID-19 analysis population N=1,796,667			Negative control exposure analysis population N=1,946,943	
Variable	**Category**	**Vaccinated N=1,473,955**	**Unvaccinated N=517,329**	**ChAdOx1-S N=845,428**	**BNT162b2 N=811,740**	**Unvaccinated N=139,499**	**Vaccinated N=1,437,356**	**Unvaccinated N=509,587**
	Blood pressure test	1,160,045 (78.7%)	309,961 (59.9%)	718,511 (88.5%)	717,133 (84.8%)	91,327 (65.5%)	1,229,509 (85.5%)	334,066 (65.6%)
	Bone density scan	81,881 (5.6%)	19,011 (3.7%)	56,679 (7.0%)	53,276 (6.3%)	5,282 (3.8%)	87,873 (6.1%)	20,667 (4.1%)
	DNA Primary care visit	459,941 (31.2%)	141,955 (27.4%)	431,120 (53.1%)	432,191 (51.1%)	72,675 (52.1%)	560,407 (39.0%)	171,514 (33.7%)
	Primary care visit	1,429,058 (97.0%)	415,765 (80.4%)	802,423 (98.9%)	828,713 (98.0%)	110,333 (79.1%)	1,420,864 (98.9%)	426,546 (83.7%)
	Low value procedures	280,546 (19.0%)	78,335 (15.1%)	259,273 (31.9%)	243,222 (28.8%)	34,634 (24.8%)	342,882 (23.9%)	93,058 (18.3%)
	PSA test	285,156 (19.3%)	67,116 (13.0%)	178,655 (22.0%)	168,574 (19.9%)	19,411 (13.9%)	294,032 (20.5%)	69,005 (13.5%)

Abbreviations: AAA: abdominal aortic aneurysm; DNA: did not attend; GP: general practice; IMD: index of multiple deprivation; N: numerator; PSA: prostate specific antigen; VE: vaccine effectiveness.

*Ethnicity was identified from primary care records as described by Mathur et al. [155]. Briefly, the algorithm uses a modal approach with ties resolved by recency. If ethnicity could not be identified in primary care, then ethnicity from HES APC was used.

**IMD was identified from the ONS at the patient level, or if missing by the primary care practice.

***Other comorbidities includes: chronic liver disease, chronic cardiac disease, chronic respiratory disease, asthma, diabetes mellitus, chronic neurological disease, chronic kidney disease, severe obesity, severe mental conditions and severe learning disability. For more information on how these were defined see **Supplementary Table 1.**

†Influenza vaccination that occurred in the influenza season prior to index date. For COVID-19 this was an influenza vaccination that occurred 1 September 2019 – 31 March 2020; for Influenza and Negative control exposure this was an influenza vaccination that occurred 1 September 2018-31 March 2019 and for Negative control exposure.

Notes: for baseline characteristics in overall analysis populations see **Supplementary Table 3**. For each analysis we are comparing individuals with ≥1 vaccination versus no vaccination throughout follow-up. Cells with <5 individuals are redacted due to CPRD's patient confidentiality requirements and secondary suppression has occurred where necessary. Age was estimated at index date for each cohort and since only year of birth is provided in CPRD, all date of birth were imputed as middle of the year (01/07).

Figure legends

Figure 1. Study design. We followed individuals until the earliest of death, transfer out of GP practice, end of data availability (COVID-19), end of influenza season (influenza), or start of COVID-19 vaccination roll out (negative control exposure). For the COVID-19 cohort, we censored individuals at first COVID-19 vaccination that was neither BNT162b2 or ChAdOx1, or on receipt of a second heterologous vaccination.

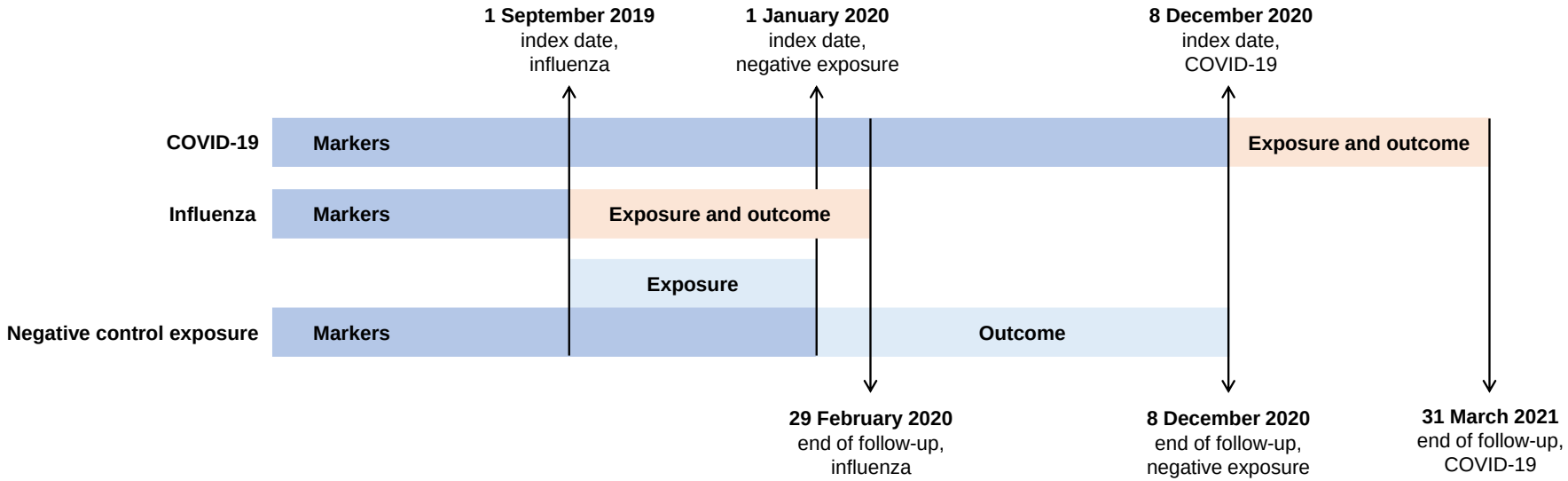


Figure 2. Study selection criteria. Additional details on population selection can be found in Supplementary Table 1. Abbreviations: CPRD: clinical practice research datalink; HES: hospital episode statistics; IMD: index of multiple deprivation; ONS: office for national statistics.

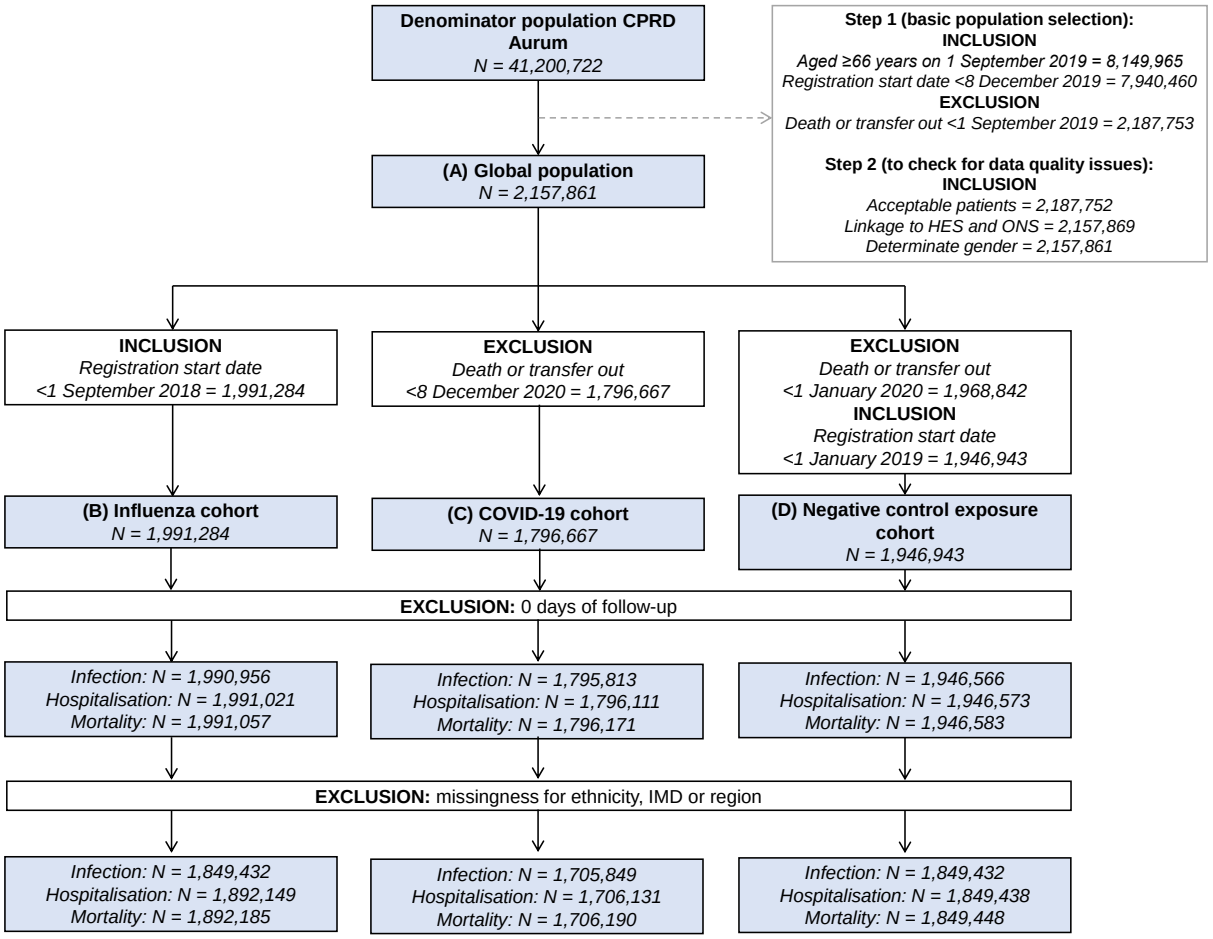
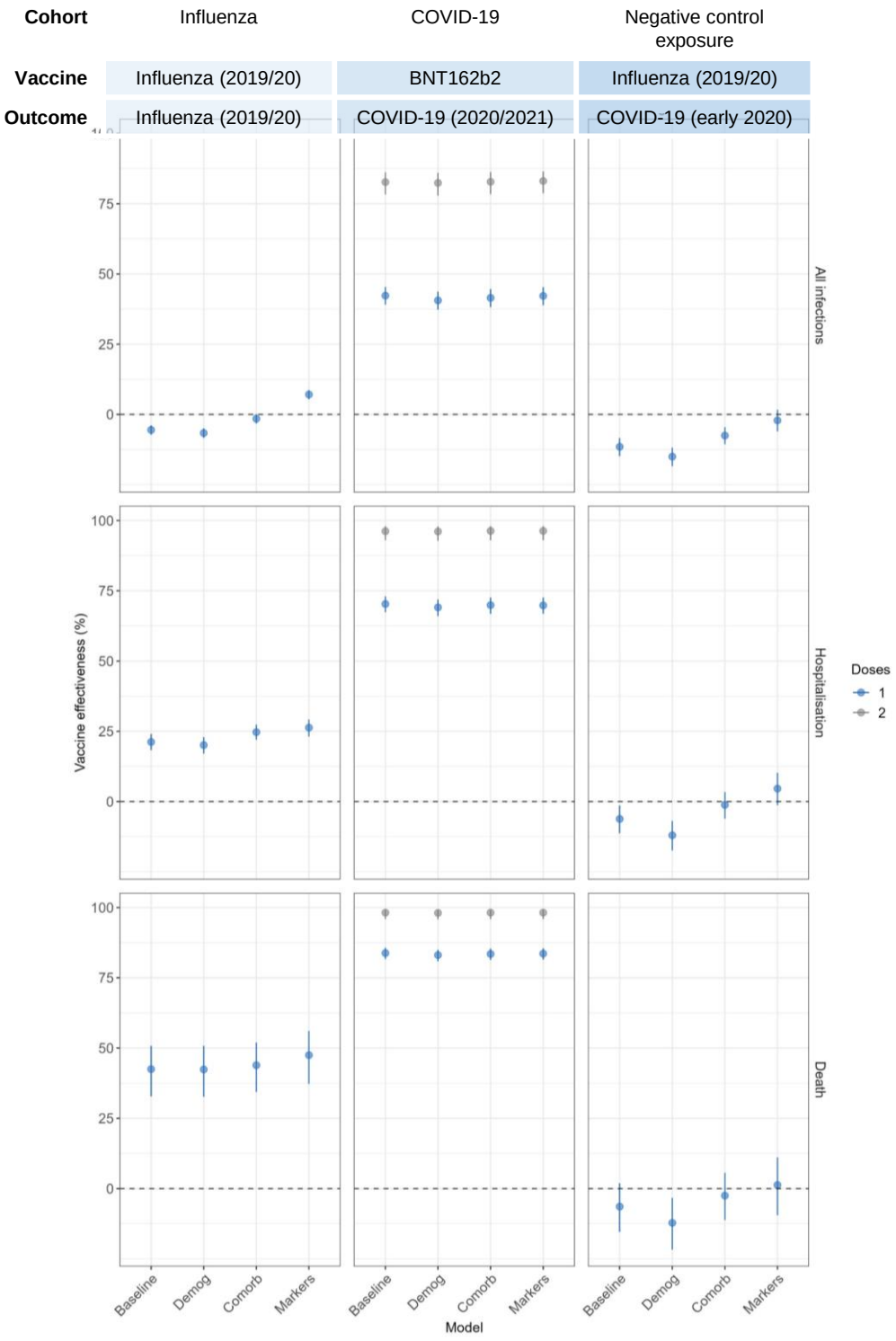


Figure 3. Estimated vaccine effectiveness following sequential confounder adjustment in each study analysis (columns) for each outcome of interest (rows). COVID-19 estimates are only for BNT162b2 versus unvaccinated as ChAdOx1 follow-up data after 2 doses was limited. Baseline models adjusted for polynomial age, sex, region and recent infection. Demography model further adjusted for ethnicity and IMD. Comorbidity models further adjusted for immunosuppressed status and other comorbidities. Marker models further adjusted for markers of health-seeking behaviour. Abbreviations: Comorb: comorbidities; demog: demography.


Cohort
Influenza
COVID-19
Negative control exposure
Vaccine
Influenza (2019/20)
BNT162b2
Influenza (2019/20)
Outcome
Influenza (2019/20)
COVID-19 (2020/2021)
COVID-19 (early 2020)
All infections
Hospitalisation
Death
Vaccine effectiveness (%)
100
75
50
25
0
Doses
1
2
Baseline
Demog
Comorb
Markers
Model

References

1. Kasl SV, Cobb S. Health behavior, illness behavior, and sick-role behavior. II. Sick-role behavior. Arch Environ Health. 1966;12(4):531-41.

2. University of Missouri. Health Care Access [Available from: https://medicine.missouri.edu/centers-institutes-labs/health-ethics/faq/health-care-access#:~:text=Health%20care%20access%20is%20the,and%20other%20health%2Dimpacting%20conditions.

3. Kinjo M, Chia-Cheng Lai E, Korhonen MJ, McGill RL, Setoguchi S. Potential contribution of lifestyle and socioeconomic factors to healthy user bias in antihypertensives and lipid-lowering drugs. Open Heart. 2017;4(1):e000417.

4. Toh S, Hernandez-Diaz S. Statins and fracture risk. A systematic review. Pharmacoepidemiol Drug Saf. 2007;16(6):627-40.

5. Jefferson T, Rivetti D, Rivetti A, Rudin M, Di Pietrantonj C, Demicheli V. Efficacy and effectiveness of influenza vaccines in elderly people: a systematic review. Lancet. 2005;366(9492):1165-74.

6. Nichol KL, Nordin JD, Nelson DB, Mullooly JP, Hak E. Effectiveness of influenza vaccine in the community-dwelling elderly. N Engl J Med. 2007;357(14):1373-81.

7. Nelson JC, Jackson ML, Weiss NS, Jackson LA. New strategies are needed to improve the accuracy of influenza vaccine effectiveness estimates among seniors. J Clin Epidemiol. 2009;62(7):687-94.

8. Horne EMF, Hulme WJ, Keogh RH, Palmer TM, Williamson EJ, Parker EPK, et al. Waning effectiveness of BNT162b2 and ChAdOx1 covid-19 vaccines over six months since second dose: OpenSAFELY cohort study using linked electronic health records. BMJ. 2022;378:e071249.

9. Hulme WJ, Williamson EJ, Green ACA, Bhaskaran K, McDonald HI, Rentsch CT, et al. Comparative effectiveness of ChAdOx1 versus BNT162b2 covid-19 vaccines in health and social care workers in England: cohort study using OpenSAFELY. BMJ. 2022;378:e068946.

10. Izurieta HS, Chillarige Y, Kelman J, Wei Y, Lu Y, Xu W, et al. Relative Effectiveness of Influenza Vaccines Among the United States Elderly, 2018-2019. J Infect Dis. 2020;222(2):278-87.

11. Izurieta HS, Lu M, Kelman J, Lu Y, Lindaas A, Loc J, et al. Comparative Effectiveness of Influenza Vaccines Among US Medicare Beneficiaries Ages 65 Years and Older During the 2019-2020 Season. Clin Infect Dis. 2021;73(11):e4251-e9.

12. Whitaker HJ, Tsang RSM, Byford R, Andrews NJ, Sherlock J, Sebastian Pillai P, et al. Pfizer-BioNTech and Oxford AstraZeneca COVID-19 vaccine effectiveness and immune response amongst individuals in clinical risk groups. J Infect. 2022;84(5):675-83.
13. Graham S, Walker JL, Andrews N, Nitsch D, Parker PKE, McDonald HI. Identifying markers of health-seeking behaviour in UK electronic health records. medRxiv. 2023:2023.11.08.23298256.
14. Schmid P, Rauber D, Betsch C, Lidolt G, Denker ML. Barriers of Influenza Vaccination Intention and Behavior - A Systematic Review of Influenza Vaccine Hesitancy, 2005 - 2016. PLoS One. 2017;12(1):e0170550.
15. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, Myles P. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol. 2019;48(6):1740-g.
16. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). Int J Epidemiol. 2017;46(4):1093-i.
17. Office for National Statistics. Main figures [Available from: https://www.ons.gov.uk/.
18. Clinical Practice Research Datalink. Defining your study population 2023 [Available from: https://www.cprd.com/defining-your-study-population#What%20coding%20systems%20are%20used%20in%20CPRD%20data?
19. World Health Organisation. ICD-10 Version:2019 2019 [Available from: https://icd.who.int/browse10/2019/en.
20. NHS England. OPCS-4 CODE 2021 [Available from: https://www.datadictionary.nhs.uk/data_elements/opcs-4_code.html.
21. Clinical Practice Research Datalink. Small area level data based on patient postcode. 2022 [Available from: https://cprd.com/sites/default/files/2022-02/Documentation_SmallAreaData_Patient_set22_v3.2.pdf.
22. Clinical Practice Research Datalink. Release Notes: CPRD Aurum May 2022 2023 [Available from: https://cprd.com/sites/default/files/2022-05/2022-05%20CPRD%20Aurum%20Release%20Notes.pdf.
23. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383-8.
24. Cowling TE, Ramzan F, Ladbrooke T, Millington H, Majeed A, Gnani S. Referral outcomes of attendances at general practitioner led urgent care centres in London, England: retrospective analysis of hospital administrative data. Emerg Med J. 2016;33(3):200-7.

25. Davidson JA, Banerjee A, Smeeth L, McDonald HI, Grint D, Herrett E, et al. Risk of acute respiratory infection and acute cardiovascular events following acute respiratory infection among adults with increased cardiovascular risk in England between 2008 and 2018: a retrospective, population-based cohort study. Lancet Digit Health. 2021;3(12):e773-e83.
26. UK Government. Greenbook Chapter 14a [Available from: https://assets.publishing.service.gov.uk/media/650c0d6afbd7bc0014e54715/Greenbook-chapter-14a-4September2023.pdf.
27. Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, vanStaa T, Grundy E, Smeeth L. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. J Public Health (Oxf). 2014;36(4):684-92.
28. UK Government. Greenbook Chapter 19 [Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/931139/Green_book_chapter_19_influenza_V7_OCT_2020.pdf
29. The National Institute for Health and Care Excellence. NICE 'do not do' recommendations [Available from: https://www.nice.org.uk/media/default/sharedlearning/716_716donotdobookletfinal.pdf.
30. NHS Digital. Unplanned hospitalisation for chronic ambulatory care sensitive conditions 2020 [Available from: https://digital.nhs.uk/data-and-information/publications/statistical/nhs-outcomes-framework/may-2020/domain-2-enhancing-quality-of-life-for-people-with-long-term-conditions-nof/2-3-i-unplanned-hospitalisation-for-chronic-ambulatory-care-sensitive-conditions.
31. Jain A, Walker JL, Mathur R, Forbes HJ, Langan SM, Smeeth L, et al. Zoster vaccination inequalities: A population based cohort study using linked data from the UK Clinical Practice Research Datalink. PLoS One. 2018;13(11):e0207183.
32. Pathirannehelage S, Kumarapeli P, Byford R, Yonova I, Ferreira F, de Lusignan S. Uptake of a Dashboard Designed to Give Realtime Feedback to a Sentinel Network About Key Data Required for Influenza Vaccine Effectiveness Studies. Stud Health Technol Inform. 2018;247:161-5.
33. Izurieta HS, Wu X, Lu Y, Chillarige Y, Wernecke M, Lindaas A, et al. Zostavax vaccine effectiveness among US elderly using real-world evidence: Addressing unmeasured confounders by using multiple imputation after linking beneficiary surveys with Medicare claims. Pharmacoepidemiol Drug Saf. 2019;28(7):993-1001.
34. Zhang HT, McGrath LJ, Wyss R, Ellis AR, Sturmer T. Controlling confounding by frailty when estimating influenza vaccine effectiveness using predictors of dependency in activities of daily living. Pharmacoepidemiol Drug Saf. 2017;26(12):1500-6.

35. Wang R, Liu M, Liu J. The Association between Influenza Vaccination and COVID-19 and Its Outcomes: A Systematic Review and Meta-Analysis of Observational Studies. Vaccines (Basel). 2021;9(5).
36. Hosseini-Moghaddam SM, He S, Calzavara A, Campitelli MA, Kwong JC. Association of Influenza Vaccination With SARS-CoV-2 Infection and Associated Hospitalization and Mortality Among Patients Aged 66 Years or Older. JAMA Netw Open. 2022;5(9):e2233730.
37. Wu S, Du S, Feng R, Liu W, Ye W. Behavioral deviations: healthcare-seeking behavior of chronic disease patients with intention to visit primary health care institutions. BMC Health Serv Res. 2023;23(1):490.
38. Prabhakar T, Goel MK, Acharya AS. Health-Seeking Behavior and its Determinants for Different Noncommunicable Diseases in Elderly. Indian J Community Med. 2023;48(1):161-6.
39. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ. 2017;359:j4587.
40. Hulme WJ, Williamson E, Horne EMF, Green A, McDonald HI, Walker AJ, et al. Challenges in Estimating the Effectiveness of COVID-19 Vaccination Using Observational Data. Ann Intern Med. 2023;176(5):685-93.

7.6 Additional methods: Variable creation

For each of the variables that were created for paper three, the general methodology for creation of the code list development has previously been detailed in **Chapter 5** and detailed methodology for code list creation and operational definition of the markers has been previously detailed in **Chapter 6**. More detailed methodology of the code list creation for the vaccination exposures and infection outcomes are described in each of the sections below and in Table 12.

7.6.1 Variables at index: Influenza at-risk conditions

Influenza at-risk conditions were described at index in each of the study cohorts and adjusted for in the comorbidity adjusted models (see paper three above). Influenza at-risk conditions are those that prioritise an individual for seasonal influenza vaccination each year as they are regarded by JCVI to increase an individuals' risk of serious illness or death should they contract an influenza infection. Influenza at-risk conditions and their definitions are provided in UKHSA's Greenbook Chapter 19[190]. In summary, at-risk conditions include nine chronic conditions that are broadly categorised into respiratory, cardiovascular and immunosuppressive conditions. Influenza at-risk conditions are similar to COVID-19 at-risk conditions[110] with differences in the condition descriptions. For example, for influenza at-risk conditions diabetes includes Type 1 diabetes, Type 2 diabetes requiring insulin or oral hypoglycaemic drugs, whereas, for COVID-19 this includes any diabetes. These conditions are different to CEV conditions, which qualified an individual for shielding status during the early COVID-19 pandemic in the UK (see Section 3.3.2.1).

For study three, for all three cohorts, influenza at-risk conditions were adjusted for. The rationale for using influenza at-risk conditions for the COVID-19 analysis was that individuals with these conditions would have known at the beginning of the pandemic that they were at higher risk of contracting a respiratory infection and therefore might be more likely to receive a timely COVID-19 vaccination. These individuals might have also been more likely to seek and receive medical attention once they contracted symptoms of COVID-19. Using the same at-risk conditions across the cohorts was also considered beneficial so that comparisons could be made across cohorts.

In addition to identifying influenza at-risk conditions, severe mental illness and learning disability were also identified. Although these were not original JCVI influenza at-risk conditions, these conditions were included as individuals with these conditions experienced barriers to accessing healthcare services during the COVID-19 pandemic and therefore had lower vaccination uptake[250]. Individuals with learning difficulties also experience a higher risk of COVID-19 severe

illness with men having a 3.1 times greater risk of COVID-19-related death and women having a 3.5 times greater risk compared with their non-disabled gendered counterparts[251].

Code lists and lookback periods for influenza at-risk conditions were previously developed for a CPRD-HES study that aimed to assess the risk of cardiovascular events following an acute respiratory infection[225]. It was decided to use these code lists and lookback periods. The definition of these conditions based on Davidson et al, 2021[225] can be found in Appendix A. Additional Tables.

Another study that compared the prevalence of these conditions using CPRD data alone versus CPRD data linked to HES found that the prevalence of the conditions was similar, except for chronic liver disease which was more prevalent when using linked data (529 per 100,000 population versus 272 per 100,000 population)[252]. Therefore, it was decided in the current study that only medcodes and prodcodes in CPRD Aurum would be used to identify the conditions. It was also decided to group these conditions into immunosuppressive conditions and other comorbidities, as previous studies have identified a much higher risk of COVID-19 related death in immunosuppressed individuals versus other conditions[253]. The decision to combine the comorbidities also builds on findings from paper one in **Chapter 3** that additional adjustment for comorbidities in COVID-19 vaccine effectiveness studies during early vaccine deployment had limited impact on vaccine effectiveness estimates.

7.6.2 Exposures: Influenza and COVID-19 vaccinations

Code list creation for influenza vaccination was previously detailed in **Chapter 6**. For COVID-19 only ChAdOx1 or BNT162b2 vaccination were included in the analysis as these were the first vaccinations to be approved in the UK and therefore maximum follow-up was accrued. The code list search terms used to identify these COVID-19 vaccinations are provided in Table 12 below.

7.6.3 Outcomes: Influenza and COVID-19 infections, hospitalisation/death and death

Code list creation for influenza and COVID-19-related outcomes can be found detailed in Table 12 below, with further information on the operational definitions in paper three above. A diagnosis of acute respiratory infection or influenza like illness (ARI/ILI) was used instead of influenza as in the UK, only a minority of influenza suspected cases undergo confirmatory testing. Influenza cases in the UK mostly go undiagnosed, but for individuals that access primary care for their symptoms, a diagnosis is based on clinical symptoms rather than diagnostic testing. Clinicians that suspect a case based on clinical symptoms will record this case as ARI/ILI, rather than influenza[254].

For COVID-19 infections, both medcodes signifying a suspected, as well as confirmed diagnosis were included. The rationale for this was because for the negative control exposure analysis, the majority of the outcomes were identified during a time period (1 January 2020 to 7 December 2020) before governmental free community PCR testing became available in July 2020 in the UK (see Figure 2 in **Chapter 3**). If a patient with COVID-19 like symptoms presented to primary care prior to the availability of nationwide testing, GPs would have likely recorded this diagnosis as suspected. Using only confirmed cases would have underestimated the number of COVID-19 events during the negative control exposure time period. For the COVID-19 analysis, the outcome period (8 December 2020 to 31 March 2021) occurred after availability of free governmental PCR testing. A consistent definition of COVID-19 infections was preferred so that comparisons could be made. Since during the COVID-19 analysis when testing capabilities were high[104], suspected cases would have accounted for the minority of coding.

In terms of influenza vaccinations, COVID-19 vaccinations and influenza outcomes, published code lists were used. For COVID-19 outcomes, since suspected cases were to be included, new code lists were generated. The search terms used for code list generations are detailed in Table 12 below. In terms of the line-by-line review of these codes the following exclusions were applied:

- Advice or education about COVID-19, but no evidence of a diagnosis.
- Antibodies for COVID-19, as these tests can be conducted to assess vaccination immunogenicity.
- Exposure to SARS-CoV-2, but no confirmation of infection.
- Patients identified as high-risk, but no confirmation of SARS-CoV-2 infection.
- Testing for COVID-19, but no results reported.
- COVID-19 vaccinations.
- Codes that are not relevant.
- Other coronaviruses.

Table 12 Influenza and COVID-19 exposures and outcomes

Code list	Published list used	Search terms	Lists compared to and additional codes identified
Exposures			
COVID-19 vaccination prodcodes	CPRD May 2022 release notes for COVID-19 vaccination counts[142]. COVID-19 vaccinations other than ChAdOx1 or	Not relevant.	Not relevant.

	BNT162b2 were included, but were flagged for censoring in the analyses.		
Influenza vaccination prodcodes	See Section 6.7.2 in **Chapter 6**		
Influenza vaccination medcodes	See Section 6.7.2 in **Chapter 6**		
Outcomes			
COVID-19 medcodes	As both suspected and confirmed cases were identified it was necessary to develop a new list.	covid|nCoV|sars|coronavirus	CPRD Aurum May 2022 COVID-19 counts[142] and Davidson et al, 2021 list[225] and no additional codes were identified.
COVID-19 ICD-10 codes	Davidson et al, 2021 list[225].	Not relevant.	Not relevant.
Influenza medcodes	Davidson et al, 2021 list[225].	Not relevant.	Not relevant.
Influenza ICD-10 codes	Davidson et al, 2021 list[225].	Not relevant.	Not relevant.

Abbreviations: CPRD: clinical practice research datalink; ICD-10: international classifications of disease, 10th revision.

7.7 Additional information on analytical methods

7.7.1 Censoring at ONS death date

Deaths can either be identified in CPRD Aurum using CPRD death date or using ONS death data. Section 5.3.3 describes how the CPRD death date algorithm is derived in CPRD and the frequency of when CPRD and ONS death dates differ. In study three, our outcome of interest was ALR/ILI or COVID-19 related death and therefore it was necessary to use the ONS data for this, as CPRD death date does not provide diagnoses codes with death date. Therefore, it was decided to censor at ONS death date only, as if censoring occurred at both ONS or CPRD death date, outcomes of interest might incorrectly be identified as censoring events.

7.7.2 Choice of analysis method

Since the number of infections was expected to fluctuate day-to-day, particularly for the negative control exposure analysis that covered the first wave of COVID-19 cases, the Cox regression model was considered appropriate. Nelson-Aalen plots[255] were used prior to conduct of the models to assess for the proportional hazards assumption and this did not appear to be violated in any of the analyses.

7.8 Additional discussion of paper

This section presents a more detailed discussion of the results from the paper three above.

7.8.1 Impact of incomplete HES or ONS linkage

The study included individuals that had complete HES or ONS linkage. Prior to applying the linkage criterion, there were 2,187,752 patients and only 29,883 (1.4%) patients were excluded during this step. Patients that were excluded were those from non-English practices, as HES linkage is not available for non-English practices. The May 2022 release of CPRD Aurum contained thirteen practices that were from Northern Ireland[138]. In addition, patients would have been excluded in this step if NHS England could not establish a linkage to either HES or ONS based on a patients NHS number and other identifiers (see Section 5.3.4.3). As the above study aimed to identify health-seeking behaviour in primary care records, it could be that those for which NHS England could not establish a linkage had poorer health-seeking behaviour compared to those with a linkage. However, since the number of individual's excluded in this step was very low, this was not deemed a concern.

7.8.2 Clustering according to Theory of Planned Behaviour

In study two, the markers of health-seeking behaviour were clustered into three or potentially four groups according to how they were expected to behave in the data. It was expected that markers in the "physically determined with lack of access" group would be negatively associated with vaccination. The reason the expected associations between the "physically determined with lack of access" markers (primary DNA and ACS conditions) and vaccinations was not identified in study three, is likely to be due to how these markers were defined in the data. For primary care DNA, the expected lower prevalence of this marker in vaccinated individual might not have occurred, as the denominator was all individuals in England aged ≥66 years. Not all these individuals would have booked GP appointments (which is required for a DNA). Potentially if the denominator was GP visits rather than the entire population, then this might have been a better marker of lack of access. ACS conditions were also similar or higher in COVID-19 unvaccinated individuals, but potentially this was because respiratory conditions are included in the list of ACS conditions, which of course will be higher amongst those unvaccinated[227]. In future, if ACS conditions are used as a marker, then the conditions for which the vaccine aim to prevent should be removed from the code list for ACS conditions.

7.8.3 Key missing variables

This study included individuals ≥66 years and therefore only risk groups 1-5 from the UK governments phased approach were relevant for this study (see Table 13 below). The first individuals to be offered a vaccination dose were those who were residents in nursing homes. These individuals could be identified in the CPRD-HES-ONS dataset using medcodes, however,

these codes would be underreported and so would be under ascertained and it is unclear with what bias. In addition, those that were 65-69 years on 31st March 2021 that were identified with CEV status (see Section 3.3.2.1), would have been offered a vaccination on the 18th January 2021, rather than the rest of their age group that were offered their vaccinations on 15th February 2021 (see Table 13 below). Other potential key missing variables were mobility status and education. The concern with not having information these key missing variables was that differences in risk between vaccinated and unvaccinated was not appropriately accounted for.

As previously discussed in Section 3.3.2.1, the CEV flag contains a very heterologous group of individuals and individual CEV status changed throughout the pandemic. Therefore, adjusting for CEV flag would not have been meaningful. It was more meaningful to adjust for specific conditions, which is what the current study did. It should also be noted that only 66-69 year olds would have been impacted by this, since all individuals aged 70 years and older were invited for vaccination at the same time or before this (see Table 13 below).

Furthermore, as calendar time was the underlying time scale in the Cox regression models it is likely that this would have invertedly accounted for differences in risk.

Table 13 UK government COVID-19 vaccination phased approach

Risk group	Description	Start date
1	Care home residents and staff	8 December 2020
2	Individuals aged 80 years and older and front line medical staff	Individuals aged 80 years: 8 December 2020; front line medical staff: 9 and 14 January 2021
3	Individuals aged 75 years and older	18 January 2021
4	Individuals aged 70 years and older or those aged 16-69 with CEV status*	18 January 2021
5	Individuals aged 65 years and older	15 February 2021
6	Individuals aged 16 to 65 years in an at-risk group**	15 February 2021
7	Individuals aged 60 years and older	1 March 2021
8	Individuals aged 55 years and older	6 March 2021
9	Individuals aged 50 years and older	17 March 2021
10	Individuals aged 40 years and older	30 April 2021
11	Individuals aged 30 years and older	26 May 2021
12	Individuals aged 18 years and older	18 June 2021

*CEV individuals were those who were asked to shield as their immune system deemed them to be at higher risk [110]. These individuals were originally identified in GP systems and then GPs were able to add additional patients based on their clinical judgement.

**At risk individuals were those that were identified with conditions that likely put them at higher risk of severe illness or death from COVID-19. Thes conditions closely reflect the influenza 'at-risk' conditions and can be found in the Greenbook Chapter 14a[110].

7.8.4 Limited impact of adjusting for confounders in COVID-19 analysis

In the COVID-19 analysis, there appeared to be limited evidence of confounding by demography variables, comorbidities or the markers of health-seeking behaviour as additional adjustments had limited impact on vaccine effectiveness estimates.

In the UK there was a phased approach for COVID-19 vaccination deployment (see Table 13). Individuals above the age of ≥65 years were prioritised for COVID-19 vaccinations if they were based in care homes or nursing homes, otherwise these vaccinations were deployed first to over 80-year-olds and then in decreasing 5-year age bands. Individuals aged 16-69 from with CEV status were asked to be vaccinated at the same time as those aged 70 years and older. Individuals aged 16-65 years were also offered the COVID-19 vaccination at the same time as those over 65 years. Although these individuals would not have been included in study three, it could be that over 65-year-olds with at-risk conditions decided to get vaccinated promptly as they knew they were at a higher risk and they also would still have been at higher risk of SARS-CoV-2 infection. However, when comorbidities were adjusted for in the COVID-19 analyses vaccine effectiveness only increased from 40.6% (95%CI: 37.3, 43.7%) in the demographic adjusted models to 41.5% (95%CI: 38.2%, 44.6%) for all infections. It is unclear why adjusting for comorbidities did not have a greater impact on the vaccine effectiveness estimates, particularly as these conditions have previously been shown to be associated with severe COVID-19 outcomes[253]. It could be that since CEV conditions could not be identified in the data (Section 3.3.2.1), the adjustment for at-risk conditions in this step is insufficient and therefore residual confounding still remained. It could be that adjustment for at-risk conditions did not impact timeliness to vaccination, as most of the population received their COVID-19 vaccination as quickly as they could. It could also be that the adjustment for calendar time in the baseline model, as mentioned above, adjusted for differences in risk, as those that received their vaccinations first were also those at highest risk. In study one (**Chapter 3**), the additional adjustment for comorbidities also did not have an impact, but differences in risk were also likely accounte for in study one as results were presented as time since vaccination. This would also explain why the findings for comorbidity adjustment differed in the influenza analysis (adjustment for comorbidities increased the vaccine effectiveness estimates from -6.6% [95%CI: -8.3, -4.9] in the demography model to -1.5% (95%CI: -3.2, 0.1) for all infections).

The limited impact of adjustment for health-seeking behaviour, could be that, as discussed in the paper above, that the risk perception of the virus and high testing and vaccination capacity during the pandemic meant that pre-pandemic markers of health-seeking behaviour were less influential.

The limited impact of comorbidities and health-seeking behaviour/healthcare access during this time is reassuring to researchers who generated vaccine effectiveness estimates for JCVI using NHS England datasets during early stages of the pandemic. This is because the NHS England datasets (e.g., COVID-19 SGSS and NIMS) lack information on key confounders e.g., comorbidities and therefore they were unable to adjust for this in their effect estimates. UK studies that also accounted for calendar time also likely accounted for CEV conditions using this approach. In future it could be that comorbidities and health-seeking behaviour/health care access have a greater influence on COVID-19 estimates in the UK as the COVID-19 vaccination programme has become seasonal[256]. As risk perception of the virus, social pressures and ease of vaccination uptake have reduced[172], it is likely that those that take up COVID-19 vaccines each year will likely be those at highest risk, or those with healthier behaviours.

7.8.5 Change in Comirnaty vaccine schedule during study period

On 31 December 2020 the UK chief medical officers announced that the second doses of COVID-19 vaccinations should be given at 12 weeks after first dose, rather than at the previously recommended 3-4 weeks[257]. In the BNT162b2 trial[109] individuals were given a second vaccination dose 21 days after their first. Some of the JCVI priority groups 1 (care home and nursing home residents) and 2 (80+ year olds) (Table 13) would have been called for first their vaccination on 12th December 2020 and therefore could have potentially received their second vaccination before the dosing schedules changed, however, for most of the current study population, they likely received their second vaccination after the dosing schedule changed. Since wider dosing schedules have previously been shown to improve duration of protection after second COVID-19 vaccine[258] it could mean that the current study estimates slightly overestimated vaccine effectiveness after two doses compared to the clinical trial. However, the effect is expected to be very minimal, especially considering the short follow-up in the current study.

7.9 Overall chapter findings

Overall, this chapter showed that identified markers of health-seeking behaviour in UK EHRs can be used to quantify and account for confounding in observational research. It was shown that for influenza vaccine effectiveness 2019/20 seasonal estimates against influenza infections (also identified in primary care) estimates were null or negative until the markers of health-seeking behaviour were adjusted for. The same impact was shown for more severe endpoints, although to a lesser degree. For COVID-19 vaccine effectiveness estimates from the early pandemic after COVID-19 vaccination deployment, there was limited evidence of confounding from health-seeking behaviour and healthcare. Adjustment for health-seeking markers had limited impact on

vaccine effectiveness estimates, with the same impact shown for other potential confounders (e.g., comorbidities). It was likely that since the COVID-19 vaccination phased approach in the UK was based on age and comorbidity risk that adjustment for calendar time in the Cox regression models accounted for differences in risk. Overall, it appeared that the markers of health-seeking behaviour, identified in study two, were very good at quantifying and removing residual confounding from health-seeking behaviour – this was demonstrated in the negative control exposure control that demonstrated null effectiveness after adjustment from these markers. This finding supports previous evidence that vaccine effectiveness estimates are speculated to be impacted by confounding from health-seeking behaviour. [44,45,55]. The negative control exposure analysis also showed that residual confounding was removed. Previous authors that have also used this approach [51,123]reported evidence that residual confounding still remained afterwards.

8 Chapter 8: Discussion

8.1 Introduction to the chapter

Overall the aim of this thesis was to develop methods to identify, quantify and account for biases in observational research using EHRs, applied in the context of vaccine effectiveness. More specifically the three aims were:

1. To identify and quantify the size and direction of biases and alternative causal pathways in a COVID-19 vaccine effectiveness observational study using a test-negative design.
2. To systematically identify a set of markers of health-seeking behaviour available in EHRs that can potentially be used to quantify and account for this type of confounding.
3. To quantify and account for confounding from health-seeking behaviour in an influenza and COVID-19 vaccine effectiveness study.

Each of these objectives were met separately through the three papers presented in **Chapter 3** (objective one), **Chapter 6** (objective two) and **Chapter 7** (objective three). The current chapter will give an overview of the entire thesis findings, discussion and interpretation.

8.2 Aim of chapter

To discuss the overall findings of this thesis, including results, strengths, limitations, interpretation, implications, learnings and recommendations for future researchers.

8.3 Overall findings of the thesis

8.3.1 Study one (objective one): Identifying and quantifying bias in COVID-19 vaccine effectiveness studies

What was known

Due to the novel nature of COVID-19 at the start of this thesis, there were limited studies that had investigated the presence of biases in COVID-19 vaccine effectiveness research. Presence of certain biases, such as outcome or exposure misclassification had been theorised by some authors previously[81]. For example, as discussed in **Chapter 1**, simulation studies have shown that in the presence of exposure and outcome misclassification influenza vaccine effectiveness are extremely biased, particularly for the test-negative design[65]. These simulation studies cannot be used to confirm the presence of or quantify biases, as external data is utilised. In the UK, the national body for public health research (UKHSA) used the test-negative design to monitor vaccine effectiveness during the COVID-19 pandemic[63]. This design at least partially accounts

for confounding from health-seeking behaviour[76], which has previously been theorised to confound vaccine effectiveness estimates[44,45,55]. Therefore, this study aimed to investigate whether certain biases were present and to quantify the impact of these biases in one of the first COVID-19 vaccine effectiveness studies that was conducted in the UK.

What this study adds

This study used questionnaire data linked to nationwide COVID-19 vaccination and PCR testing data from one of the first COVID-19 vaccine effectiveness studies in the UK. The questionnaire data was used to assess the presence of or quantify different potential biases and alternative causal pathways in the original study. It identified limited evidence of potential bias from exposure misclassification, outcome misclassification, confounding from comorbidities and deferral bias. In a combined estimate that accounted for all of these potential biases at once, estimated vaccine effectiveness decreased after two doses of BNT162b2 from 88% (95%CI: 79,94%) in the original study estimate to 85% (95%CI: 68,94%). There was also limited evidence of potential self-reported riskier behaviour after vaccination or evidence of attending a vaccination visit increasing risk of SARS-CoV-2 infections. These potential alternative causal pathways would have underestimated vaccine effectiveness estimates compared with clinical trial data, if present.

So what

There was limited evidence of potential biases that were assessed in the test-negative design. This is reassuring since this design was commonly used to assess COVID-19 vaccine effectiveness to inform government policy at the beginning of the COVID-19 pandemic. This design is also used for influenza to estimates vaccine effectiveness and cost effectiveness each year to inform recommendations for the current and next year. However, this design is not always feasible to conduct (e.g., without test result data), requires strong assumptions to be met[76] and is potentially impacted by collider bias which threatens the validity of its claim to account for confounding from health-seeking behaviour[113,114]. Alternative approaches are needed to account for confounding from health-seeking behaviour. For the rest of the thesis, alternative methods to identify, quantify and account for confounding from health-seeking behaviour were explored.

8.3.2 Pragmatic review: summarising methods used to account for confounding from health-seeking behaviour in vaccine effectiveness research

What was known

Many authors have previously reported influenza vaccine effectiveness estimates of 40-50% against all-cause mortality[42,43]. However, these estimates are implausible as influenza accounts for a maximum of 10% of deaths per year[44,45] and therefore it is speculated that this could be due to potential confounding from health-seeking behaviour. A previous systematic literature review highlighted the prevalence of this problem[52]. Previously study designs (e.g., test-negative) used to account for this type of bias (see Section 1.10), require strong assumptions to be met. Alternative methods such as proxy adjustment have been used, but it was unclear to what extent these were implemented in vaccine effectiveness research and therefore a pragmatic literature review was conducted to investigate this.

What this study adds

This pragmatic literature review identified very few (N=8)[51,122-127,129] vaccine effectiveness studies that explicitly used alternative methods test-negative and other designs to account for this potential type of confounding in EHRs. All these studies used proxy markers directly available in the EHR or from linked survey data and they either included these markers in propensity methods or adjusted for them in the model analysis. Markers from these studies included preventative measures such as screening and vaccinations, as well as healthcare utilisation and diagnostic testing for infectious diseases. Some of the markers e.g., SARS-CoV-2 testing[124,125], were problematic as they are influenced by underlying health need. The approach that was used in each of these studies to select the markers were not described, and there were inconsistencies in markers used within the same group of researchers. Two of the studies[51,123], which both used Medicare data in the US assessed the impact of adjusting for markers on vaccine effectiveness estimates. Both found that adjusting for these markers reduced estimates of vaccine effectiveness. In one of the studies that assessed influenza vaccine effectiveness[51] used healthcare utilisation markers that are influenced by underlying health need (e.g., GP visits). In the other study that assessed shingles vaccine effectiveness[123], only one marker of health-seeking behaviour (self-reported doctor avoidance) was identified and adjusted for and this was imputed from survey data. Both studies also used negative controls to assess for potential residual confounding and both found evidence of potential residual confounding after adjustment for health-seeking markers, which could be due to confounding from health-seeking behaviour.

So what

It is likely that studies have rarely accounted for confounding from health-seeking behaviour using proxy markers, as no method have been developed for authors to systemically account for this

bias using this approach. Authors are therefore unclear as to how this method can be effectively applied. Also previous studies that have used this method still reported residual confounding, so authors might think this method is ineffective.

8.3.3 Study two (objective two): Identifying markers of health-seeking behaviour in UK EHRs

What was known

A set of systematically identified markers were needed to account for potential confounding from health-seeking behaviour in observational research. It was necessary to ensure that these markers were informed by a conceptual framework and were broadly applicable to many different observational research questions.

What this study adds

A conceptual framework based on a behavioural model known as the Theory of Planned Behaviour[172] was used to systematically identify fifteen markers of health-seeking behaviour in UK EHRs. All the markers represented interactions with the healthcare system where the influence of underlying health conditions was limited. The identified markers included: AAA screening; breast cancer screening; bowel cancer screening; cervical cancer screening; influenza vaccination; pneumococcal vaccination; NHS health checks; PSA testing; bone density scans; low-value procedures; glucosamine use (low-value prescription); GP practice visits; DNA primary care visit; hospital visit for ACS condition; and blood pressure measurements. Criteria were iteratively developed that could be used to identify similar markers that future researchers could use in their data set at hand. The prevalence of the markers in a UK EHR dataset of individuals aged ≥66 years was compared to national estimates and was found to be similar. For screening markers and NHS health checks the prevalence was lower than national estimates, but this was likely due to differences in denominator populations (e.g., for national screening estimates, the denominator is all individuals sent an invite, whereas the current study included all individuals in England aged ≥66 years). These markers were grouped into three or four categories based on how they were expected to behave in relation to other markers in the data. The same groups were identified using either a theoretical approach based on the Theory of Planned Behaviour[172] or data driven approach.

So what

As these markers were selected using a conceptual framework, their selection was guided by previous research and therefore more confidence could be installed that they were appropriate markers. Furthermore, as many different markers were selected (fifteen), all which were influenced by varied determinants in the Theory of Planned Behaviour[172] model, health-seeking behaviour, as a complex phenomenon, is likely to be well represented. Now that a systematic set of markers had been identified in UK EHRs, it was necessary to investigate the performance of these markers to quantifying and adjusting for confounding from health-seeking behaviour.

8.3.4 Study three (objective three): Quantifying and accounting for confounding from health-seeking behaviour an influenza and COVID-19 vaccine effectiveness study.

What was known

Alternative methods to the test-negative design to account for confounding from health-seeking behaviour are limited. Therefore, this study aimed to quantify and account for confounding from health-seeking behaviour using the markers from study two in an influenza and COVID-19 vaccine effectiveness study.

What this study adds

Fourteen of the markers of health-seeking behaviour from study two were included (low value prescriptions dropped due to very low prevalence). A cohort study of influenza 2019/20 season and COVID-19 early pandemic, post-vaccination deployment vaccine effectiveness was conducted using UK EHRs. Markers were more prevalent amongst those vaccinated for influenza or COVID-19 compared to those who remained unvaccinated. The only exception was for ambulatory care sensitive conditions, which were less prevalent amongst COVID-19 vaccinated versus COVID-19 unvaccinated. For influenza, additionally adjusting for the markers of health-seeking behaviour increased vaccine effectiveness estimates against ARI/ILI-related infections from -1.5% (95%CI: -3.2,0.1; when adjusting for age, sex, region, recent infection ethnicity, IMD and comorbidities) to 7.1% (95%CI: 5.4,8.7). The same trend was shown for more severe outcomes (e.g., ARI/ILI-related hospitalisations and deaths), with less pronounced differences. For COVID-19, adjusting for health-seeking markers did not impact vaccine effectiveness estimates. Vaccine effectiveness was 82.7% (95%CI: 78.3,86.2) against SARS-CoV-2 infections in the minimally adjusted model that adjusted for age sex, region and recent infection and 83.1% (95%CI: 78.7,86.5) in the fully adjusted model that additionally adjusted for ethnicity, IMD, comorbidities and health-seeking markers. There was also no meaningful impact on vaccine

effectiveness estimates with additional adjustments for more severe COVID-19 outcomes. History of influenza vaccine effectiveness was used a negative control exposure against early COVID-19 pandemic SARS-CoV-2 infections to assess for residual confounding after adjusting for the markers of health-seeking behaviour. After adjusting for the health-seeking markers, estimated vaccine effectiveness was null signifying that residual confounding had been removed (before: -7.5% [95%CI: -10.6, -4.5]; after adjusting for health-seeking markers: -2.1% [95%CI: -6.0,1.7]).

Overall summary

The systematically identified set of markers of health-seeking behaviour from paper two appeared to successfully quantify and remove biases in an influenza vaccine effectiveness study. In terms of COVID-19 vaccine effectiveness, this type of confounding had minimal impact on estimates from the early pandemic when the markers were adjusted for. In future it is likely that this type of confounding has a greater influence on COVID-19 estimates with seasonal uptake of the vaccination. These findings supports previous authors that have speculated that confounding from health-seeking behaviour impacts vaccine effectiveness estimates [44,45,55]. This study also showed that residual confounding was also removed through use of the negative control exposure. Previous authors that have attempted to do this[51,127] reported residual confounding afterwards. As the markers selected in study two were broadly applicable to populations over ≥66 years, they can be used to quantify and account for this type of confounding in other observational studies with different research questions.

8.4 Overarching strengths

The detailed strengths of each of the individual studies are discussed in **Chapters 3, 4, 6** and **7**. The overall strengths of this thesis are the use of large, linked datasets, use of conceptual frameworks to define health-seeking behaviour, the use of consistent definitions, the applicability of these methods to other observational cohorts and the success of these methods in quantifying and adjusting for confounding from health-seeking behaviour.

8.4.1 Use of large, linked datasets

This thesis uses large, linked datasets of primary care data linked to national secondary care and death data and survey data in England. The use of these data allowed for precise vaccine effectiveness estimates to be generated for a nationally representative population. Even in study one that linked nationwide datasets to survey data, over 8,000 individuals responded to the survey and therefore precise estimates were generated for multiple different biases. These datasets supplemented with additional information from the survey provided a rich source of information

which allowed for confounding and other biases to be accurately assessed. In the final two studies, the primary care EHR data provides rich information on life-style factors such as BMI, which was used to identify at-risk conditions for influenza and SARS-CoV-2 infections to improve confounding adjustment.

8.4.2 Use of conceptual frameworks

Study two used a conceptual framework based on the updated Theory of Planned Behaviour model[172] to identify markers of health-seeking behaviour and to understand how each of the markers behaved in relation to other markers in the data. As discussed in **Chapter 4** in the pragmatic review, none of the previous vaccine effectiveness studies that used methods to quantify and account for confounding, described how they identified their markers. My approach used a behavioural model as a framework that was based on previously reported associations between variables. This helped to organise and guide my research and ensured greater confidence in my findings. This will also help future researchers, as the approach is transparent and reproducible and therefore can be used to guide similar research questions around health-seeking behaviour. This will also ensure improved consistency across the field and comparisons can therefore be made between studies.

In term of study three, a DAG was used to guide the appropriateness of the markers as proxies of health-seeking behaviour. This ensured that markers that were included in study three were accounting for the underlying phenomenon at hand and were not introducing additional bias e.g., by being on the causal pathway from vaccination to infection. This enabled the approach to be more robust, transparent and reproducible.

8.4.3 Use of consistent definitions

Across all three of my studies the population under investigation was individuals aged ≥66 years in England. This population was selected as these individuals are eligible for seasonal influenza vaccinations each year[190]. In addition, they were prioritised for COVID-19 vaccinations at the beginning of the UK COVID-19 pandemic (see Table 13 in Section 7.8.4), they are likely to experience similar patterns of health-seeking behaviour[176], they are eligible for many governmental preventative measures in the UK and they have high morbidity and mortality[177]. The use of consistent populations across all three studies in this thesis, meant that patterns of health-seeking behaviour were likely to be similar.

A consistent time period was also used for both the COVID-19 vaccine effectiveness studies (study one and three). Using a consistent time period across these studies meant conclusions

could be made across the studies. The use of this time period across all three studies meant that findings were likely applicable to all three. For example, the findings on self-reported risky behaviours in study one were likely also applicable to the COVID-19 vaccine effectiveness study in study three.

In both study one and three we also identified and adjusted for at-risk conditions. As the same comorbidities were identified across these studies, comparisons could be made regarding the conclusions. For example, in both study one and three, it was summarised that in a time period after COVID-19 vaccination approval, that additional adjustment for comorbidities did not impact vaccine effectiveness estimates.

In study three, a harmonised statistical approach was used across all three cohort, which aided interpretation of confounding in different contexts.

8.4.4 Applicability of these methods to other observational cohorts

In study two, the markers were selected to be broadly applicable to all individuals aged ≥66 years in England. Markers were selected that were available to the entire population, where possible and where the influence of underlying health need was as weak as possible. This meant that the markers can be used across different observational studies with different research questions, without requirement for patients to have specific conditions. The prevalence of the markers in study two were produced for a population of individuals aged ≥66 years in England. This meant that in some cases, the denominator population was not comparable to national estimates, but also meant that the expected prevalence of these markers in a generalisable population were generated. Future researchers can use these estimates to ensure that definitions have been applied appropriately in their dataset at hand.

8.4.5 Success of these methods in quantifying adjusting for confounding

The methods developed in this thesis were very successful at quantifying and accounting for confounding from health-seeking behaviour. They quantified significant confounding in cohort studies of influenza vaccine effectiveness, even when cause-specific outcomes were used and they quantified minimum confounding in early COVID-19 pandemic vaccine effectiveness estimates. Previously it was only theorised that vaccine effectiveness estimates were impacted by this type of bias[44,45,55], but now it has been confirmed. It was known that adjusting for these markers was sufficient to account for this type of confounding as the negative control exposure was null after adjusting for these markers. This is the first known study that has adjusted for markers of health-seeking behaviour and has successfully removed residual confounding. The

other two known studies that did this previously[51,123] identified significant residual confounding after adjusting for their markers, as their set of markers used were insufficient.

8.5 Overarching limitations

The detailed limitations of each of the studies are discussed in **Chapters 3, 4, 6** and **7**. The overall limitations of this thesis are the reliance on accurate clinical coding, the lack of detailed information, generalisability to younger age groups, implementation and validity in other datasets and the potential biases introduced through study design.

8.5.1 Reliance on accurate clinical coding

In study one the recording of COVID-19 PCR tests and vaccinations in the nationwide datasets used are based on the recording of these events at the time of event occurrence, however, validation of these events are limited. UKHSA report in their description of the NIMS dataset[96] that when comparing vaccination dates and manufacturer with survey data, the accuracy was high, however, measures were not reported for this comparison. For study three, COVID-19 vaccinations were automatically pushed into GP records from NHS England and then COVID-19 test results were added into the GP record also from NHS England either retrospectively or prospectively (see Section 5.3.3.3) during this time[142]. However, there are no known studies that have compared these events in the CPRD data to the original NHS England data.

As mentioned in **Chapter 5** in Section 5.3.3.5, a systematic literature review summarised that the median positive predictive value of the CPRD data to identify 189 different diagnoses was 89%[144]. This means that if an individual was identified with a condition in the CPRD dataset, 89% of the time the individual had the condition. This study was conducted using data from 1987 to 2008 and likely that improvements in coding have occurred since then, although the UK pandemic and austerity likely still contributes to underdiagnoses in these data[131]. Even with a high positive predictive value, it could be that other validity measures are low. These other measures, e.g., specificity, sensitivity and negative predictive value, require the sampling of individuals in the CPRD dataset without the condition of interest. For rare conditions, this task would be particularly cumbersome. There are a few examples of where authors have looked at sensitivity and specificity in CPRD and have reported high validity[259].

It could also be that the accuracy of clinical coding in these data varies by individual health-seeking behaviour. Individuals with strong health-seeking behaviour and very good access to healthcare might be over diagnosed in some conditions, whereas those that have poor health-seeking behaviour and a lack of access might be underdiagnosed in other conditions[260,261]. If

health-seeking behaviour does impact overdiagnosis/underdiagnosis, then the impact of adjusting for health-seeking behaviour would be underestimated in study three. This is because adjustment for the comorbidities would in part be accounting for differences in health-seeking behaviour. In this instance final estimates that adjust for comorbidities and health-seeking markers should be close to the true estimate, it is just likely that the quantification of confounding from health-seeking behaviour is underestimated.

8.5.2 Lack of detailed information

It was not possible to link to some available EHR datasets, e.g., HES Outpatient or HES DID, in the current studies due to cost restraints. These datasets include secondary care information for outpatient visits[262] and diagnostic imaging scans[263]. These datasets could have been used in the current study to identify additional markers. For example, DNA for outpatient visits could have been included as an additional marker with linkage to HES Outpatient. The capture of bone density scans might also have been improved with linkage to HES DID.

There are also likely to be other confounders that could not be identified through use of routinely collected health data. For example, in study three, as mentioned in Section 7.8.2, key variables such as nursing home resident, CEV status, mobility and education status could not be appropriately identified, so it could be that residual confounding still remained.

8.5.3 Generalisability to younger age groups

The markers that were selected in study two were designed to be broadly applicable to a population aged ≥66 years. For the markers that are only primarily available to older aged populations in the UK (AAA screening, breast cancer screening, bowel cancer screening, NHS health checks, influenza vaccination and pneumococcal vaccination), these would not be identifiable in a generalisable younger population. Researchers that were interested in assessing confounding from health-seeking behaviour in a younger age group would need to identify additional markers that are available to these individuals. In the UK, childhood vaccinations could potentially be used as markers, as well as the human papillomavirus vaccination. The criteria developed in paper two would need to be used to select different markers of interest.

8.5.4 Implementation and validity in other datasets

Although the markers were selected to be applicable to other datasets, there are some instances where some of these markers might not be identifiable. Firstly, some of the preventative markers might not be provided as part of routine healthcare service in other countries. For example, in some low-income countries, cancer prevention programmes do not exist[264] and therefore the

cancer screening markers cannot be used. Secondly, some datasets do not have primary care information included. For example, in Sweden, currently primary care datasets are only available for three regions (Stockholm, Västra Götaland, and Skåne) across Sweden, corresponding to around 52% of the Swedish population[265]. For this reason, these data are commonly not used in public health research. Researchers using these data will not be able to identify the markers specific to primary care (e.g., GP visits, DNA primary care). Thirdly, the coding system in CPRD Aurum (SNOMED-CT) is very detailed and there might be instances where these codes are not identifiable in datasets that used less detailed coding classification systems. There are also some datasets in which vaccinations are not reliably recorded. For example, The Centers for Medicare and Medicaid Services found that in Medicare claims only 17.5 million individuals aged ≥65 years had at least one dose of COVID-19 vaccination recorded, whereas, it is estimated by the Centers for Disease Control and Prevention that 44.1 million individuals had been vaccinated[15]. It could also be that recording of vaccinations in claims data is differential by health-seeking behaviour and therefore, including vaccination markers in these datasets could further bias estimates.

8.5.5 Potential biases introduced through study design

The aim of these studies was to identify and quantify potential biases in vaccine effectiveness research. However, it could be that additional biases were introduced through the study designs. For example, for study one, it could be argued that collider bias (described in Section 3.5.1) is introduced through the test-negative design[113,114]. The impact of collider bias in COVID-19 vaccine effectiveness research during the early stages of the pandemic is expected to be minimal. This is because uptake of the COVID-19 vaccination and testing, risk perception of the virus and social pressures were high[101,104], and therefore the influence of prior health-seeking behaviour was likely weak. In terms of other biases it could be that potential biases were introduced through the survey design in study one. For example, when aiming to investigate risk behaviours, it is possible that when asking individuals to report their own risk behaviours to the government (UKHSA), that they underreported these behaviours (social desirability bias), or only those with less risk behaviours responded to the survey (selection bias). The study of health-seeking behaviour using EHRs (study three) could also potentially underestimate the prevalence of minimal health-seeking behaviour or inability to access healthcare, as those that have these traits will not be registered with a GP and therefore will go undetected (selection bias). The likely impact of this though is expected to be small as over 98% of the population in the UK are registered with a general practice[87].

8.5.6 Interference in vaccine effectiveness research

An extension to consistency in the causal inference framework (see Section 1.7) is the Stable Unit Treatment Value Assignment (STUVA). STUVA assumes that the potential outcome of one group is not affected by the treatment received by the other group[266]. However, in observational vaccine research, one person's infection might be impacted by the vaccination status of another person, which is known as "interference"[21]. This was likely particularly the case for COVID-19 due to the widespread vaccination programmes, which would have reduced infectiousness amongst vaccinated individuals[101]. As interference was not accounted for in study one or study three, this might have impacted our ability to estimate the causal estimand (i.e., a study with no bias)[21].

8.5.7 Non-collapsible odds ratios or hazards ratios

Non-collapsibility occurs when the measure of association (e.g., odds ratio, risk ratio) calculated within strata of a covariate (conditional association) differs from the measure of association calculated without stratification (marginal association). Non-collapsibility occurs in odds ratios and hazard ratios due to their reliance on conditional probabilities, while risk ratios are not affected by this phenomenon because they compare absolute risks between groups. The interpretation of odds or hazards ratios depends upon the confounders that are being adjusted for[267]. The causal estimand in our study was an estimate that was not impacted by any bias. However, the difference in our estimate in study one from the original estimate (88% vs 87%) could also be because confounding is cancelled out due to non-collapsibility, rather than an indication of minimal confounding. Therefore, the final estimate that supposedly accounted for all potential biases, could have been even more biased than the original estimate as there was no way to measure this. In study three, we attempted to quantify residual bias through the use of the negative control exposure, however, we cannot rule out that a supposed removal of the residual bias was due to the non-collapsibility that was not accounted for.

8.6 Interpretation

Confounding from health-seeking behaviour has been theorised to impact influenza vaccine effectiveness estimates since the 2000s[44,45,55]. Methods study as the test-negative design have been used in vaccine effectiveness research to account for this bias. Study one of this thesis aimed to assess whether a COVID-19 test-negative design study from the early COVID-19 UK pandemic was subject to other potential biases, such as exposure and outcome misclassification. It appeared that this design was robust to these biases. However, this design cannot always be conducted due to the strong set of assumptions. Therefore, alternative methods are required to account for bias from health-seeking behaviour. Alternative methods to account for confounding

from health-seeking behaviour include adjusting for proxy markers. Previous studies that have used proxy markers are very limited and use only limited sets of markers that are inconsistent and with unclear theoretical role. Based on this identified research gap, the second study of this thesis identified a set of proxy markers that were systematically identified based on the updated Theory of Planned Behaviour model[172]. In the third study of this thesis, confounding from health-seeking behaviour was then quantified and accounted for using these markers in an influenza and COVID-19 vaccine effectiveness study. A negative control exposure of history of influenza against early pandemic SARS-CoV-2 infections identified that residual confounding was removed after adjusting for these markers.

The use of proxy markers to quantify and account for confounding from health-seeking behaviour was a novel approach, building on previous studies that used different marker sets. The markers identified in this thesis were very effective at quantifying and controlling for this type of bias and since they were developed to be broadly applicable, can be used in other observational studies.

The limited evidence of bias and alternative causal pathways in one of the first COVID-19 vaccine effectiveness studies in the UK (study one) was likely due to the accurate recording of COVID-19 vaccination and testing data during this time. The NIMS system that was set up to record COVID-19 vaccinations during the pandemic was effective at recording events that occurred across many healthcare institutions. These events were then centralised into one dataset[96]. As the majority of community COVID-19 PCR tests were confirmed in UKHSA laboratories[103], these were also centralised into one database. With accurate recording of vaccinations and testing, there was likely to be limited exposure or outcome misclassification. Another reason for the limited bias could be the high uptake of COVID-19 vaccinations and wide availability of COVID-19 PCR testing during this time. In the UK, the uptake of first and second dose COVID-19 vaccinations in over 70-year-olds was nearly 95%[101]. There were also over 500 free community test sites and the median time to a test centre was 3.7 miles[104]. With vaccination and testing rates high, this meant that previous health-seeking behaviour barriers were likely less influential during this time. It is also likely that the test-negative design, was robust to many different biases. As well as aiming to account for differences in health-seeking behaviour [76] the design also accounts for differences in infection exposure as all individuals that book a test would at least have had some possibility of infection. With similarities in infection exposure, outcome misclassification is likely to be reduced. The limited alternative causal pathways could be due to the high conforming to government non-pharmaceutical policy measures (e.g., lock-downs, mask wearing) during the UK COVID-19 pandemic. The overall level of compliance with lockdown measures in the UK was high[115], which

is likely why riskier behaviours after vaccination and during lock-down periods were at least initially infrequent.

The identified evidence of confounding from health-seeking behaviour in the influenza vaccine study (study three) algins with prior findings[44,45,50,51]. However, in terms of quantifying this type of confounding my findings did not align with previous trends. Prior studies from the US have found that adjusting for health-seeking behaviour in vaccine effectiveness research decreases vaccine effectiveness estimates[51,123]. However, for study three, influenza vaccine effectiveness estimates increased when health-seeking behaviour was accounted for. In my study when confounders, other than health-seeking behaviour were adjusted for, vaccine effectiveness estimates against infections (including those identified in primary care) was -1.5% (95%CI, -3.2, 0.1). When health-seeking behaviour was accounted for, the vaccine effectiveness estimate was 7.1% (5.4, 8.7). This likely represents that in the UK, those that are accessing care for symptoms of influenza, are also likely to be those that receive an influenza vaccination as they have favourable health-seeking behaviours and good access to health care. Once differences in health-seeking behaviour were adjusted for, a protective effect of the influenza vaccination was identified. For the previous studies, it could be that for non-specific outcomes such as all-cause mortality, the opposite trend when adjusting for these markers could also occur. This is because those that take up the influenza vaccination, their health-seeking behaviours led them to have healthier lifestyles and more preventative measures leading to better overall health outcomes. Therefore, adjusting for health-seeking behaviour for non-specific outcomes reduced vaccine effectiveness estimates.

8.7 Implications for clinical practice and policy markers

In terms of clinical practice, this research has some direct implications. The COVID-19 vaccine effectiveness findings from study one are beneficial for clinicians and policy markers as they provide further confidence in estimates generated using these methods from a similar time period. In terms of future pandemics, clinicians and policy markers can be more confident that estimates generated using this study design are robust to different potential biases. In terms of influenza vaccine estimates that do not use the test-negative design or do not adjust for confounding from health-seeking behaviour, these individuals should be cautious of potential overestimation of non-specific outcomes or underestimation of cause-specific outcomes.

The markers of health-seeking behaviour could be used by clinicians to identify groups of patients that likely have poor health-seeking behaviour or experience barriers to healthcare so that these patients can be targeted to improve health inequalities. For example, since the prevalence of all of these markers at baseline was generally higher amongst those with a subsequent influenza or

COVID-19 vaccination, these markers could be explored for their use to determine if a patient is likely to take up each of these vaccinations each year. Study three provided findings that individuals have not taken up a screening visit are also less likely to take up a vaccination appointment. Potential tools such as additional letters and resources about the benefits of vaccination can be provided to individuals who have not taken up their screening invitations each year to help improve vaccination uptake.

8.8 Implications for research

In term of implications for researchers, study one provided further evidence to support the use of a test-negative case-control design when conditions met, and in particular in early pandemic situations for emerging disease. The markers from study two can be used to adjust for health-seeking behaviour for more accurate vaccine effectiveness estimates for infections in which the conditions for test-negative case-control designs are not met, such as seasonal influenza. These will also inform more accurate cost-effectiveness estimates[15]. It is anticipated that since COVID-19 vaccinations are now to be administered seasonally, that the influence of health-seeking behaviour on future COVID-19 vaccine effectiveness studies may more closely reflect influenza. Future studies that explored this would be a logical extension to the current study. Since these markers can be used in cohort studies or other study designs, estimates can be generated across different datasets that previously could not be used in which the test-negative design could not be conducted. The markers identified can also be used across other vaccine and non-vaccine observational studies with different research questions. As mentioned in **Chapter 1**, confounding from health-seeking behaviour has also been shown to be associated with other preventative measures such as HRT[29] and statin use[37]. These markers can be used to identify, quantify and account for differences in health-seeking behaviour in these observational studies.

8.9 Unanswered questions

To understand the impact of health-seeking behaviour in younger populations and in COVID-19 non-pandemic periods, it would be beneficial to repeat the analyses for these individuals and for COVID-19 vaccinations in routine care. Future researchers can identify the markers for younger populations using the criteria developed. Furthermore, since time-varying confounding was also not accounted for in study three, these markers should be used in other study designs (e.g., target trial approaches[249]) to assess for and account for this potential type of confounding. The markers could also be used in future to assess for associations with testing to understand the influence of potential collider bias in the test-negative design [112].

8.10 Dissemination

Throughout the course of this project, I developed two abstracts, two posters and three manuscripts. Both of the posters were presented at conferences, including at the International Conference on Pharmacoepidemiology in Copenhagen (2022) and the UKHSA annual conference in Leeds (2023).

8.11 Personal learnings

I have gained a deeper understanding of the contents of UK EHRs including the NIMS, SGSS, CPRD, HES and ONS datasets and the way in which the data appears in each of these datasets. I have also gained a better understanding of regression methods such as the Cox regression model. I have also been able to develop my data management and data analyses skills through my direct access to these UK datasets. I have gained a greater understanding of how conceptual frameworks can be used to guide the design and analyses of observational study designs. I have also developed my medical writing skills through my dissemination activities.

8.12 Conclusions

This thesis identified that when using the UK EHRs and the test-negative design the impact of potential biases on early pandemic COVID-19 observational vaccine effectiveness estimates was minimal (original study: 88% [95%CI: 79-94%]; updated estimate that accounted for all potential biases: : 85% [95% CI: 68-94%]). These potential biases included exposure misclassification, outcome misclassification, confounding by comorbidities, or deferral bias (temporary apparent protective effect of the vaccination from symptomatic individuals deferring their vaccination). This thesis also identified limited evidence of riskier health behaviours associated with vaccination during this early pandemic period. The limited evidence of confounding bias is likely due to the appropriate methods (test-negative design) and the high-quality datasets (nationwide vaccination and testing data in the UK) used in the original study. The test-negative design accounts for confounding from health-seeking behaviour, however, since this design cannot be applied in all datasets, has strong assumptions that cannot always be met and is potentially impacted by collider bias (which was not accounted for in study one), alternative methods were required to account for this type of confounding.

This thesis therefore confirmed that markers of health-seeking behaviour can be systematically identified in in UK EHRs. Fourteen markers were identified that represented preventative measures where the influence of underling health need was minimal. For influenza vaccine effectiveness these markers were then used to discover that there was significant evidence of

confounding from health-seeking behaviour. This supports previous studies that speculate that this type of confounding is important. For COVID-19 vaccine effectiveness these markers also confirmed that during the early pandemic there was limited evidence of confounding from health-seeking behaviour. The minimal impact of confounding from health-seeking behaviour was identified in a COVID-19 vaccine effectiveness study from the early UK pandemic after vaccinations were deployed. These markers were also used to confirm that residual confounding had been almost entirely removed in a negative control exposure cohort study that showed almost null vaccine effectiveness after these markers were adjusted for. Future researchers can use these markers, that are broadly applicable to different study populations and datasets, to account for confounding from health-seeking behaviour and healthcare access in other observational research questions.

References

1. Senn S. Randomisation is not about balance, nor about homogeneity but about randomness. Accessed 21/10/2024, 2024. https://errorstatistics.com/2020/04/20/s-senn-randomisation-is-not-about-balance-nor-about-homogeneity-but-about-randomness-guest-post/
2. Flecha OD, Douglas de Oliveira DW, Marques LS, Goncalves PF. A commentary on randomized clinical trials: How to produce them with a good level of evidence. *Perspect Clin Res*. Apr-Jun 2016;7(2):75-80. doi:10.4103/2229-3485.179432
3. Gudmundsson LSE, O. B.; Johannsson, M. *Databases for Pharmacoepidemiological Research*. vol 205-211. Springer Series on Epidemiology and Public Health. Springer; 2021.
4. Anderson RH, Baker EI, Penny D, Redington AN, Rigby ML, Wernovsky G. *Paediatric Cardiology* 3rd ed. 2009.
5. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol*. Jan 2017;106(1):1-9. doi:10.1007/s00392-016-1025-6
6. Jick H, Watkins RN, Hunter JR, et al. Replacement estrogens and endometrial cancer. *N Engl J Med*. Feb 1 1979;300(5):218-22. doi:10.1056/NEJM197902013000502
7. NHS England. Data saves lives: reshaping health and social care with data. Accessed 25/01/2024, 2024. https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data
8. Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. *J Med Internet Res*. Mar 14 2005;7(1):e3. doi:10.2196/jmir.7.1.e3
9. McMillan B, Eastham R, Brown B, Fitton R, Dickinson D. Primary Care Patient Records in the United Kingdom: Past, Present, and Future Research Priorities. *J Med Internet Res*. Dec 19 2018;20(12):e11293. doi:10.2196/11293
10. Ehrenstein VK, H.; Lehmann H. *Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide*. 3rd ed. 2019.
11. Wood L, Coulson R. Revitalizing the General Practice Research Database: plans, challenges, and opportunities. *Pharmacoepidemiol Drug Saf*. Aug-Sep 2001;10(5):379-83. doi:10.1002/pds.608
12. Clinical Practice Research Datalink. CPRD Aurum Data Specification. Accessed 25/10/23, 2023. https://cprd.com/sites/default/files/2022-02/CPRD%20Aurum%20Data%20Specification%20v2.7%20%28002%29.pdf
13. World Health Organization. Vaccines and immunization. Accessed 03/04/2024, 2024. https://www.who.int/health-topics/vaccines-and-immunization#tab=tab_1
14. Fedson DS. Measuring protection: efficacy versus effectiveness. *Dev Biol Stand*. 1998;95:195-201.
15. Centers for Disease Control and Prevention. Why CDC Estimates Vaccine Effectiveness (VE). Accessed 14/12/2023, 2023. https://www.cdc.gov/flu/vaccines-work/estimates-vaccine-effectiveness.htm
16. Prugger C, Spelsberg A, Keil U, Erviti J, Doshi P. Evaluating covid-19 vaccine efficacy and safety in the post-authorisation phase. *BMJ*. Dec 23 2021;375:e067570. doi:10.1136/bmj-2021-067570
17. UK Government. Appendix 1: UKHSA report estimating the number needed to vaccinate to prevent COVID-19 hospitalisation for booster vaccination in autumn 2023 in England. Accessed 17/12/2023, 2023. https://www.gov.uk/government/publications/covid-19-autumn-2023-vaccination-programme-jcvi-advice-26-may-2023/appendix-1-ukhsa-report-estimating-the-number-needed-to-vaccinate-to-prevent-covid-19-hospitalisation-for-booster-vaccination-in-autumn-2023-in-engla

18. Kirsebom FCM, Harman K, Lunt RJ, et al. Vaccine effectiveness against hospitalisation estimated using a test-negative case-control study design, and comparative odds of hospital admission and severe outcomes with COVID-19 sub-lineages BQ.1, CH.1.1. and XBB.1.5 in England. *Lancet Reg Health Eur*. Dec 2023;35:100755. doi:10.1016/j.lanepe.2023.100755
19. Hernan MA RJ. *Causal Inference: What If.* . 2020.
20. World Health Organization. Vaccine efficacy, effectiveness and protection. Accessed 21/10/2024, 2024. https://www.who.int/news-room/feature-stories/detail/vaccine-efficacy-effectiveness-and-protection
21. Schnitzer ME. Estimands and Estimation of COVID-19 Vaccine Effectiveness Under the Test-Negative Design: Connections to Causal Inference. *Epidemiology*. May 1 2022;33(3):325-333. doi:10.1097/EDE.0000000000001470
22. Hernan MAR, JM. *Causal Inference: What If*. 2020.
23. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. *Nephron Clin Pract*. 2010;115(2):c94-9. doi:10.1159/000312871
24. Jager KJ, Zoccali C, Macleod A, Dekker FW. Confounding: what it is and how to deal with it. *Kidney Int*. Feb 2008;73(3):256-60. doi:10.1038/sj.ki.5002650
25. Kasl SV, Cobb S. Health behavior, illness behavior, and sick role behavior. I. Health and illness behavior. *Arch Environ Health*. Feb 1966;12(2):246-66. doi:10.1080/00039896.1966.10664365
26. University of Missouri. Health Care Access. Accessed 23/09/2023, 2023. https://medicine.missouri.edu/centers-institutes-labs/health-ethics/faq/health-care-access#:~:text=Health%20care%20access%20is%20the,and%20other%20health%2Dimpacting%20conditions
27. Eurich DT, Majumdar SR. Statins and sepsis - scientifically interesting but clinically inconsequential. *J Gen Intern Med*. Mar 2012;27(3):268-9. doi:10.1007/s11606-011-1939-7
28. van Dam RM, Li T, Spiegelman D, Franco OH, Hu FB. Combined impact of lifestyle factors on mortality: prospective cohort study in US women. *BMJ*. Sep 16 2008;337:a1440. doi:10.1136/bmj.a1440
29. Stampfer MJ, Willett WC, Colditz GA, Rosner B, Speizer FE, Hennekens CH. A prospective study of postmenopausal estrogen therapy and coronary heart disease. *N Engl J Med*. Oct 24 1985;313(17):1044-9. doi:10.1056/NEJM198510243131703
30. Varas-Lorenzo C, Garcia-Rodriguez LA, Perez-Gutthann S, Duque-Oliart A. Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case-control study. *Circulation*. Jun 6 2000;101(22):2572-8. doi:10.1161/01.cir.101.22.2572
31. Grady D, Rubin SM, Petitti DB, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. *Ann Intern Med*. Dec 15 1992;117(12):1016-37. doi:10.7326/0003-4819-117-12-1016
32. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. *JAMA*. Aug 19 1998;280(7):605-13. doi:10.1001/jama.280.7.605
33. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA*. Jul 17 2002;288(3):321-33. doi:10.1001/jama.288.3.321
34. Mosca L, Collins P, Herrington DM, et al. Hormone Replacement Therapy and Cardiovascular Disease. *Circulation*. 2001;104(4):499-503. doi:doi:10.1161/hc2901.092200
35. Hernan MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. Nov 2008;19(6):766-79. doi:10.1097/EDE.0b013e3181875e61

36. von Elm E, Egger M. The scandal of poor epidemiological research. *BMJ*. Oct 16 2004;329(7471):868-9. doi:10.1136/bmj.329.7471.868
37. Toh S, Hernandez-Diaz S. Statins and fracture risk. A systematic review. *Pharmacoepidemiol Drug Saf*. Jun 2007;16(6):627-40. doi:10.1002/pds.1363
38. Haley RW, Dietschy JM. Is there a connection between the concentration of cholesterol circulating in plasma and the rate of neuritic plaque formation in Alzheimer disease? *Arch Neurol*. Oct 2000;57(10):1410-2. doi:10.1001/archneur.57.10.1410
39. Majumdar SR, McAlister FA, Eurich DT, Padwal RS, Marrie TJ. Statins and outcomes in patients admitted to hospital with community acquired pneumonia: population based prospective cohort study. *BMJ*. Nov 11 2006;333(7576):999. doi:10.1136/bmj.38992.565972.7C
40. Setoguchi S, Glynn RJ, Avorn J, Mogun H, Schneeweiss S. Statins and the risk of lung, breast, and colorectal cancer in the elderly. *Circulation*. Jan 2 2007;115(1):27-33. doi:10.1161/CIRCULATIONAHA.106.650176
41. Brookhart MA, Patrick AR, Dormuth C, et al. Adherence to lipid-lowering therapy and the use of preventive health services: an investigation of the healthy user effect. *Am J Epidemiol*. Aug 1 2007;166(3):348-54. doi:10.1093/aje/kwm070
42. Jefferson T, Rivetti D, Rivetti A, Rudin M, Di Pietrantonj C, Demicheli V. Efficacy and effectiveness of influenza vaccines in elderly people: a systematic review. *Lancet*. Oct 1 2005;366(9492):1165-74. doi:10.1016/S0140-6736(05)67339-4
43. Nichol KL, Nordin JD, Nelson DB, Mullooly JP, Hak E. Effectiveness of influenza vaccine in the community-dwelling elderly. *N Engl J Med*. Oct 4 2007;357(14):1373-81. doi:10.1056/NEJMoa070844
44. Nelson JC, Jackson ML, Weiss NS, Jackson LA. New strategies are needed to improve the accuracy of influenza vaccine effectiveness estimates among seniors. *J Clin Epidemiol*. Jul 2009;62(7):687-94. doi:10.1016/j.jclinepi.2008.06.014
45. Simonsen L, Taylor RJ, Viboud C, Miller MA, Jackson LA. Mortality benefits of influenza vaccination in elderly people: an ongoing controversy. *Lancet Infect Dis*. Oct 2007;7(10):658-66. doi:10.1016/S1473-3099(07)70236-0
46. Chen Y, Liu S, Leng SX. Chronic Low-grade Inflammatory Phenotype (CLIP) and Senescent Immune Dysregulation. *Clin Ther*. Mar 2019;41(3):400-409. doi:10.1016/j.clinthera.2019.02.001
47. Elkind MS. Inflammatory markers and stroke. *Curr Cardiol Rep*. Jan 2009;11(1):12-20. doi:10.1007/s11886-009-0003-2
48. Ferrucci L, Fabbri E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. *Nat Rev Cardiol*. Sep 2018;15(9):505-522. doi:10.1038/s41569-018-0064-2
49. Govaert TME, Thijs CTMCN, Masurel N, Sprenger MJW, Dinant GJ, Knottnerus JA. The Efficacy of Influenza Vaccination in Elderly Individuals: A Randomized Double-blind Placebo-Controlled Trial. *JAMA*. 1994;272(21):1661-1665. doi:10.1001/jama.1994.03520210045030
50. Jackson LA, Jackson ML, Nelson JC, Neuzil KM, Weiss NS. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *Int J Epidemiol*. Apr 2006;35(2):337-44. doi:10.1093/ije/dyi274
51. Zhang HT, McGrath LJ, Wyss R, Ellis AR, Sturmer T. Controlling confounding by frailty when estimating influenza vaccine effectiveness using predictors of dependency in activities of daily living. *Pharmacoepidemiol Drug Saf*. Dec 2017;26(12):1500-1506. doi:10.1002/pds.4298
52. Remschmidt C, Wichmann O, Harder T. Frequency and impact of confounding by indication and healthy vaccinee bias in observational studies assessing influenza vaccine effectiveness: a systematic review. *BMC Infect Dis*. Oct 17 2015;15:429. doi:10.1186/s12879-015-1154-y
53. Doshi P. Influenza: marketing vaccine by marketing disease. *BMJ*. May 16 2013;346:f3037. doi:10.1136/bmj.f3037

54. Arbel R, Hammerman A, Sergienko R, et al. BNT162b2 Vaccine Booster and Mortality Due to Covid-19. *N Engl J Med*. Dec 23 2021;385(26):2413-2420. doi:10.1056/NEJMoa2115624
55. Hoeg TB, Duriseti R, Prasad V. Potential "Healthy Vaccinee Bias" in a Study of BNT162b2 Vaccine against Covid-19. *N Engl J Med*. Jul 20 2023;389(3):284-285. doi:10.1056/NEJMc2306683
56. ROSENBAUM PR, RUBIN DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. doi:10.1093/biomet/70.1.41
57. Levintow SN, Nielson CM, Hernandez RK, et al. Pragmatic considerations for negative control outcome studies to guide non-randomized comparative analyses: A narrative review. *Pharmacoepidemiol Drug Saf*. Jun 2023;32(6):599-606. doi:10.1002/pds.5623
58. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*. May 2010;21(3):383-8. doi:10.1097/EDE.0b013e3181d61eeb
59. Ray GT, Lewis N, Klein NP, Daley MF, Lipsitch M, Fireman B. Depletion-of-susceptibles Bias in Analyses of Intra-season Waning of Influenza Vaccine Effectiveness. *Clin Infect Dis*. Mar 17 2020;70(7):1484-1486. doi:10.1093/cid/ciz706
60. Ray GT, Lewis N, Klein NP, et al. Intraseason Waning of Influenza Vaccine Effectiveness. *Clin Infect Dis*. May 2 2019;68(10):1623-1630. doi:10.1093/cid/ciy770
61. Dagan N, Barda N, Kepten E, et al. BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. *N Engl J Med*. Apr 15 2021;384(15):1412-1423. doi:10.1056/NEJMoa2101765
62. Hall VJ, Foulkes S, Saei A, et al. COVID-19 vaccine coverage in health-care workers in England and effectiveness of BNT162b2 mRNA vaccine against infection (SIREN): a prospective, multicentre, cohort study. *Lancet*. May 8 2021;397(10286):1725-1735. doi:10.1016/S0140-6736(21)00790-X
63. Lopez Bernal J, Andrews N, Gower C, et al. Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in England: test negative case-control study. *BMJ*. May 13 2021;373:n1088. doi:10.1136/bmj.n1088
64. Vasileiou E, Simpson CR, Shi T, et al. Interim findings from first-dose mass COVID-19 vaccination roll-out and COVID-19 hospital admissions in Scotland: a national prospective cohort study. *Lancet*. May 1 2021;397(10285):1646-1657. doi:10.1016/S0140-6736(21)00677-2
65. De Smedt T, Merrall E, Macina D, Perez-Vilar S, Andrews N, Bollaerts K. Bias due to differential and non-differential disease- and exposure misclassification in studies of vaccine effectiveness. *PLoS One*. 2018;13(6):e0199180. doi:10.1371/journal.pone.0199180
66. Funk MJ, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. *Curr Epidemiol Rep*. Dec 2014;1(4):175-185. doi:10.1007/s40471-014-0027-z
67. Food and Drug Administration. FRAMEWORK FOR FDA'S REAL-WORLD EVIDENCE PROGRAM. Accessed 25/01/2024, 2024. https://www.fda.gov/media/120060/download?attachment
68. National Institute for Health and Care Excellence. NICE real-world evidence framework. Accessed 25/01/2024, 2024. https://www.nice.org.uk/corporate/ecd9/chapter/overview
69. Norgaard M, Ehrenstein V, Vandenbroucke JP. Confounding in observational studies based on large health care databases: problems and potential solutions - a primer for the clinician. *Clin Epidemiol*. 2017;9:185-193. doi:10.2147/CLEP.S129879
70. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*. Aug 2000;29(4):722-9. doi:10.1093/ije/29.4.722

71. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. Feb 2003;32(1):1-22. doi:10.1093/ije/dyg070
72. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nat Rev Rheumatol*. Jul 2015;11(7):437-41. doi:10.1038/nrrheum.2015.30
73. Bor J, Moscoe E, Mutevedzi P, Newell ML, Barnighausen T. Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology*. Sep 2014;25(5):729-37. doi:10.1097/EDE.0000000000000138
74. Oxford Reference. proxy variable. Accessed 05/04/2024, 2024. https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100351624
75. Faurot KR, Jonsson Funk M, Pate V, et al. Using claims data to predict dependency in activities of daily living as a proxy for frailty. *Pharmacoepidemiol Drug Saf*. Jan 2015;24(1):59-66. doi:10.1002/pds.3719
76. Jackson ML, Nelson JC. The test-negative design for estimating influenza vaccine effectiveness. *Vaccine*. Apr 19 2013;31(17):2165-8. doi:10.1016/j.vaccine.2013.02.053
77. Jackson ML, Rothman KJ. Effects of imperfect test sensitivity and specificity on observational studies of influenza vaccine effectiveness. *Vaccine*. Mar 10 2015;33(11):1313-6. doi:10.1016/j.vaccine.2015.01.069
78. Petersen I, Douglas I, Whitaker H. Self controlled case series methods: an alternative to standard epidemiological study designs. *BMJ*. Sep 12 2016;354:i4515. doi:10.1136/bmj.i4515
79. Whitaker HJ, Ghebremichael-Weldeselassie Y, Douglas IJ, Smeeth L, Farrington CP. Investigating the assumptions of the self-controlled case series method. *Stat Med*. Feb 20 2018;37(4):643-658. doi:10.1002/sim.7536
80. McGrath LJ, Kshirsagar AV, Cole SR, et al. Influenza vaccine effectiveness in patients on hemodialysis: an analysis of a natural experiment. *Arch Intern Med*. Apr 9 2012;172(7):548-54. doi:10.1001/archinternmed.2011.2238
81. Lewnard JA, Patel MM, Jewell NP, et al. Theoretical Framework for Retrospective Studies of the Effectiveness of SARS-CoV-2 Vaccines. *Epidemiology*. Jul 1 2021;32(4):508-517. doi:10.1097/EDE.0000000000001366
82. Robins J. The control of confounding by intermediate variables. *Stat Med*. Jun 1989;8(6):679-701. doi:10.1002/sim.4780080608
83. Shapiro J. The NHS: the story so far (1948-2010). *Clin Med (Lond)*. Aug 2010;10(4):336-8. doi:10.7861/clinmedicine.10-4-336
84. UK Parliament. Free NHS prescriptions: Eligibility for benefit claimants. Accessed 29/01/2024, 2024. https://lordslibrary.parliament.uk/free-nhs-prescriptions-eligibility-for-benefit-claimants/#heading-3
85. NHS Digital. The healthcare ecosystem. Accessed 29/01/2024, 2024. https://digital.nhs.uk/developer/guides-and-documentation/introduction-to-healthcare-technology/the-healthcare-ecosystem
86. NHS England. Primary care services. Accessed 08/11/2023, 2023. https://www.england.nhs.uk/get-involved/get-involved/how/primarycare/
87. NHS Digital. Patients Registered at a GP Practice. Accessed 29/01/2024, 2024. https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice
88. Fry J, Stephen WJ. Primary health care in the United Kingdom. *Int J Health Serv*. 1986;16(4):485-95. doi:10.2190/M0L4-QP4Q-50K2-8RGV
89. NHS England. What are community health services. Accessed 29/01/2024, 2024. https://www.england.nhs.uk/community-health-services/what-are-community-health-services/
90. NHS Digital. NHS number. Accessed 29/01/2024, 2024. https://digital.nhs.uk/services/nhs-number

91. Anderson M, Pitchforth E, Edwards N, Alderwick H, McGuire A, Mossialos E. United Kingdom: Health System Review. *Health Syst Transit*. May 2022;24(1):1-194.
92. NHS England. Personal Demographics Service. Accessed 26/04/2024, 2024. https://digital.nhs.uk/services/personal-demographics-service
93. Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. *J Biomed Inform*. Feb 2013;46(1):87-96. doi:10.1016/j.jbi.2012.09.006
94. Randorff Hojen A, Rosenbeck Goeg K. Snomed CT implementation. Mapping guidelines facilitating reuse of data. *Methods Inf Med*. 2012;51(6):529-38. doi:10.3414/ME11-02-0023
95. Steindel SJ. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J Am Med Inform Assoc*. May-Jun 2010;17(3):274-82. doi:10.1136/jamia.2009.001230
96. Tessier E, Edelstein M, Tsang C, et al. Monitoring the COVID-19 immunisation programme through a national immunisation Management system - England's experience. *Int J Med Inform*. Feb 2023;170:104974. doi:10.1016/j.ijmedinf.2022.104974
97. Office for National Statistics. 2011 rural/urban classification. Accessed 03/12/2023, 2023. https://www.ons.gov.uk/methodology/geography/geographicalproducts/ruralurbanclassifications/2011ruralurbanclassification#:~:text=The%202011%20rural%2Durban%20classification%20(RUC2011)%20allows%20for%20a,produced%20after%20the%202001%20Census.
98. UK Government. English indices of deprivation 2019. Accessed 11/12/2023, 2023. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019
99. NHS Digital. Ethnic Category. Accessed 12/12/2023, 2023. https://www.datadictionary.nhs.uk/data_elements/ethnic_category.html
100. NHS Digital. Shielded Patient List. https://digital.nhs.uk/coronavirus/shielded-patient-list#:~:text=A%20set%20of%20clinical%20conditions,risk%20from%20COVID%2D19%20infection.
101. Tessier E, Rai Y, Clarke E, et al. Characteristics associated with COVID-19 vaccine uptake among adults aged 50 years and above in England (8 December 2020-17 May 2021): a population-level observational study. *BMJ Open*. Mar 1 2022;12(3):e055278. doi:10.1136/bmjopen-2021-055278
102. NHS Digital. How we created the Shielded Patient List. Accessed 15/03/2024, 2024. https://digital.nhs.uk/blog/tech-talk/2020/how-we-created-the-shielded-patient-list
103. UK Health Security Agency. Laboratory reporting to UKHSA. A guide for diagnostic laboratories. Accessed 14/12/2023, 2023. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1159953/UKHSA_Laboratory_reporting_guidelines_May_2023.pdf
104. UK Government. 500 test sites now open as new lab partnerships boost capacity. Accessed 10/04/2024, 2024. https://www.gov.uk/government/news/500-test-sites-now-open-as-new-lab-partnerships-boost-capacity
105. UK Government. COVID-19 testing data: methodology note. Accessed 11/08/2023, 2023. https://www.gov.uk/government/publications/coronavirus-covid-19-testing-data-methodology/covid-19-testing-data-methodology-note
106. UK Health Security Agency. COVID-19. Accessed 03/04/2024, 2024. https://ukhsa-dashboard.data.gov.uk/topics/covid-19#cases
107. NHS England. Coronavirus (COVID-19) risk assessment. Accessed 29/04/2024, 2024. https://digital.nhs.uk/services/coronavirus-risk-assessment#:~:text=QCovid%C2%AE%20is%20a%20coronavirus,support%20the%20NHS%20coronavirus%20response.&text=QCovid%C2%AE%20is%20an%20evidence,and%20being%20admitted%20to%20hospital
108. Down's Syndrome Association. COVID19 (Coronavirus) and individuals who have Down's syndrome. Accessed 29/04/2024, 2024. https://www.downs-syndrome.org.uk/wp-

content/uploads/2023/06/COVID19_Coronavirus-and-Individuals-who-have-Downs-syndrome.pdf
109. Polack FP, Thomas SJ, Kitchin N, et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N Engl J Med*. Dec 31 2020;383(27):2603-2615. doi:10.1056/NEJMoa2034577
110. UK Government. Greenbook Chapter 14a. Accessed 07/02/22, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1045852/Greenbook-chapter-14a-11Jan22.pdf
111. Graham S, Tessier E, Stowe J, et al. Bias assessment of a test-negative design study of COVID-19 vaccine effectiveness used in national policymaking. *Nature Communications*. 2023/07/06 2023;14(1):3984. doi:10.1038/s41467-023-39674-0
112. Griffith GJ, Morris TT, Tudball MJ, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun*. Nov 12 2020;11(1):5749. doi:10.1038/s41467-020-19478-2
113. Sullivan SG, Tchetgen Tchetgen EJ, Cowling BJ. Theoretical Basis of the Test-Negative Study Design for Assessment of Influenza Vaccine Effectiveness. *Am J Epidemiol*. Sep 1 2016;184(5):345-53. doi:10.1093/aje/kww064
114. Westreich D, Hudgens MG. Invited Commentary: Beware the Test-Negative Design. *Am J Epidemiol*. Sep 1 2016;184(5):354-6. doi:10.1093/aje/kww063
115. Wright L, Steptoe A, Fancourt D. Patterns of compliance with COVID-19 preventive behaviours: a latent class analysis of 20 000 UK adults. *J Epidemiol Community Health*. Mar 2022;76(3):247-253. doi:10.1136/jech-2021-216876
116. Glasziou P, McCaffery K, Cvejic E, et al. Testing behaviour may bias observational studies of vaccine effectiveness. *J Assoc Med Microbiol Infect Dis Can*. Sep 2022;7(3):242-246. doi:10.3138/jammi-2022-0002
117. Chen TK, Knicely DH, Grams ME. Chronic Kidney Disease Diagnosis and Management: A Review. *JAMA*. Oct 1 2019;322(13):1294-1304. doi:10.1001/jama.2019.14745
118. Cowie MR. The heart failure epidemic: a UK perspective. *Echo Res Pract*. Mar 2017;4(1):R15-R20. doi:10.1530/ERP-16-0043
119. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol*. Dec 1 2016;45(6):1866-1886. doi:10.1093/ije/dyw314
120. York Health Economics Consortium. Pragmatic Review. Accessed 02/02/2024, 2024. https://yhec.co.uk/glossary/pragmatic-review/
121. National Library of Medicine. MEDLINE: Overview. Accessed 05/12/2023, 2023. https://www.nlm.nih.gov/medline/medline_overview.html
122. Izurieta HS, Lu M, Kelman J, et al. Comparative Effectiveness of Influenza Vaccines Among US Medicare Beneficiaries Ages 65 Years and Older During the 2019-2020 Season. *Clin Infect Dis*. Dec 6 2021;73(11):e4251-e4259. doi:10.1093/cid/ciaa1727
123. Izurieta HS, Wu X, Lu Y, et al. Zostavax vaccine effectiveness among US elderly using real-world evidence: Addressing unmeasured confounders by using multiple imputation after linking beneficiary surveys with Medicare claims. *Pharmacoepidemiol Drug Saf*. Jul 2019;28(7):993-1001. doi:10.1002/pds.4801
124. Horne EMF, Hulme WJ, Keogh RH, et al. Waning effectiveness of BNT162b2 and ChAdOx1 covid-19 vaccines over six months since second dose: OpenSAFELY cohort study using linked electronic health records. *BMJ*. Jul 20 2022;378:e071249. doi:10.1136/bmj-2022-071249
125. Hulme WJ, Williamson EJ, Green ACA, et al. Comparative effectiveness of ChAdOx1 versus BNT162b2 covid-19 vaccines in health and social care workers in England: cohort study using OpenSAFELY. *BMJ*. Jul 20 2022;378:e068946. doi:10.1136/bmj-2021-068946

126. Whitaker HJ, Tsang RSM, Byford R, et al. Pfizer-BioNTech and Oxford AstraZeneca COVID-19 vaccine effectiveness and immune response amongst individuals in clinical risk groups. *J Infect*. May 2022;84(5):675-683. doi:10.1016/j.jinf.2021.12.044
127. Izurieta HS, Chillarige Y, Kelman J, et al. Relative Effectiveness of Cell-Cultured and Egg-Based Influenza Vaccines Among Elderly Persons in the United States, 2017-2018. *J Infect Dis*. Sep 13 2019;220(8):1255-1264. doi:10.1093/infdis/jiy716
128. Mues KE, Liede A, Liu J, et al. Use of the Medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the US. *Clin Epidemiol*. 2017;9:267-277. doi:10.2147/CLEP.S105613
129. Izurieta HS, Chillarige Y, Kelman J, et al. Relative Effectiveness of Influenza Vaccines Among the United States Elderly, 2018-2019. *J Infect Dis*. Jun 29 2020;222(2):278-287. doi:10.1093/infdis/jiaa080
130. UK Parliament. Coronavirus: Testing for Covid-19. Accessed 11/03/2024, 2024. https://commonslibrary.parliament.uk/research-briefings/cbp-8897/
131. The Health Foundation. Nine major challenges facing health and care in England. Accessed 30/04/2024, 2024. https://www.health.org.uk/publications/long-reads/nine-major-challenges-facing-health-and-care-in-england
132. Keyes KM, Rutherford C, Popham F, Martins SS, Gray L. How Healthy Are Survey Respondents Compared with the General Population?: Using Survey-linked Death Records to Compare Mortality Outcomes. *Epidemiology*. Mar 2018;29(2):299-307. doi:10.1097/EDE.0000000000000775
133. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol*. Aug 1 2017;46(4):1093-1093i. doi:10.1093/ije/dyx015
134. Clinical Practice Research Datalink. ONS death registration data and CPRD primary care data Documentation. Accessed 26/04/2024, 2024. https://www.cprd.com/sites/default/files/2022-02/Documentation_Death_set22_v2.6.pdf
135. Edwards L, Pickett J, Ashcroft DM, et al. UK research data resources based on primary care electronic health records: review and summary for potential users. *BJGP Open*. Sep 2023;7(3)doi:10.3399/BJGPO.2023.0057
136. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. Jun 2015;44(3):827-36. doi:10.1093/ije/dyv098
137. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol*. Dec 1 2019;48(6):1740-1740g. doi:10.1093/ije/dyz034
138. Clinical Practice Research Datalink. Release Notes: CPRD Aurum May 2022. Accessed 05/10/2023, 2023. https://cprd.com/sites/default/files/2022-05/2022-05%20CPRD%20Aurum%20Release%20Notes.pdf
139. NHS Digital. Read Codes. Accessed 05/12/2023, 2023. https://digital.nhs.uk/services/terminology-and-classifications/read-codes
140. NHS Scotland. SCIMP Guide to Read Codes. Accessed 15/04/2024, 2024. https://www.scimp.scot.nhs.uk/better-information/clinical-coding/scimp-guide-to-read-codes
141. NHS Digital. Dictionary of medicines and devices. Accessed 05/12/2023, 2023. https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/dictionary-medicines-and-devices-dmd
142. Clinical Practice Research Datalink. Feasibility counts for SARS-CoV-2-related codes in CPRD primary care data. Accessed 22/11/2023, 2023. https://cprd.com/sites/default/files/2022-05/SARS-CoV-2%20counts%20May2022.pdf
143. Gallagher AM, Dedman D, Padmanabhan S, Leufkens HGM, de Vries F. The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations. *Pharmacoepidemiol Drug Saf*. May 2019;28(5):563-569. doi:10.1002/pds.4747

144. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol*. Jan 2010;69(1):4-14. doi:10.1111/j.1365-2125.2009.03537.x
145. NHS Digital. Quality and Outcomes Framework (QOF). Accessed 03/10/2023, 2023. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/quality-outcomes-framework-qof
146. Department of Health. QOF Guidance and Business Rules. Accessed 03/10/2023, 2023. https://www.health-ni.gov.uk/articles/about-quality-and-outcomes-framework-qof
147. NHS England. 2019/20 General Medical Services (GMS) contract Quality and Outcomes Framework (QOF). Accessed 03/10/2023, 2023. https://www.england.nhs.uk/wp-content/uploads/2019/05/gms-contract-qof-guidance-april-2019.pdf
148. Nicholson BD, Aveyard P, Bankhead CR, Hamilton W, Hobbs FDR, Lay-Flurrie S. Determinants and extent of weight recording in UK primary care: an analysis of 5 million adults' electronic health records from 2000 to 2017. *BMC Med*. Nov 29 2019;17(1):222. doi:10.1186/s12916-019-1446-y
149. Hamilton FL, Laverty AA, Huckvale K, Car J, Majeed A, Millett C. Financial Incentives and Inequalities in Smoking Cessation Interventions in Primary Care: Before-and-After Study. *Nicotine Tob Res*. Mar 2016;18(3):341-50. doi:10.1093/ntr/ntv107
150. Shiekh SI, Harley M, Ghosh RE, et al. Completeness, agreement, and representativeness of ethnicity recording in the United Kingdom's Clinical Practice Research Datalink (CPRD) and linked Hospital Episode Statistics (HES). *Popul Health Metr*. Mar 14 2023;21(1):3. doi:10.1186/s12963-023-00302-0
151. Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. *Eur J Epidemiol*. Jan 2019;34(1):91-99. doi:10.1007/s10654-018-0442-4
152. National Audit Office. Healthcare across the UK: A comparison of the NHS in England, Scotland, Wales and Northern Ireland. Accessed 03/10/2023, 2023. https://www.nao.org.uk/wp-content/uploads/2012/06/1213192.pdf
153. Boyd A, Cornish R, Johnson L, et al. *Understanding Hospital Episode Statistics (HES). London, UK: CLOSER*. 2017. https://www.closer.ac.uk/wp-content/uploads/CLOSER-resource-understanding-hospital-episode-statistics-2018.pdf
154. Clinical Practice Research Datalink. Data Minimisation workbook v1.8. Accessed 25/10/23, 2023. https://cprd.com/sites/default/files/2023-03/Data_Minimisation_Variable_Restriction%20v1.8_0.xlsx
155. Mathur R, Bhaskaran K, Chaturvedi N, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *J Public Health (Oxf)*. Dec 2014;36(4):684-92. doi:10.1093/pubmed/fdt116
156. NHS Digital. The HES processing cycle and data quality checks. Accessed 05/10/2023, 2023. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hes-processing-cycle-and-data-quality-checks
157. Herrett E, Shah AD, Boggon R, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ*. May 20 2013;346:f2350. doi:10.1136/bmj.f2350
158. Mahadevan P, Harley M, Fordyce S, et al. Completeness and representativeness of small area socioeconomic data linked with the UK Clinical Practice Research Datalink (CPRD). *J Epidemiol Community Health*. Jul 28 2022;76(10):880-6. doi:10.1136/jech-2022-219200
159. Clinical Practice Research Datalink. Small area level data based on patient postcode. Accessed 05/12/2023, 2023. https://cprd.com/sites/default/files/2022-02/Documentation_SmallAreaData_Patient_set22_v3.2.pdf

160. Clinical Practice Research Datalink. Clinical Practice Research Datalink. Accessed 03/10/2023, 2023. https://cprd.com/
161. Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: A review. *J Biomed Inform*. Jun 2017;70:1-13. doi:10.1016/j.jbi.2017.04.010
162. Springate DA, Kontopantelis E, Ashcroft DM, et al. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PLoS One*. 2014;9(6):e99825. doi:10.1371/journal.pone.0099825
163. London School of Hygiene and Tropical Medicine. LSHTM Data Compass. Accessed 04/10/2023, 2023. https://datacompass.lshtm.ac.uk/
164. Health Data Research. HDR UK Phenotype Library. Accessed 04/10/2023, 2023. https://phenotypes.healthdatagateway.org/
165. OpenSAFELY. OpenCodelists. Accessed 04/10/2023, 2023. https://www.opencodelists.org/
166. University of Cambridge. CPRD @ Cambridge – Codes Lists (GOLD). Accessed 04/10/2023, 2023. https://www.phpc.cam.ac.uk/pcu/research/research-groups/crmh/cprd_cam/codelists/v11/
167. NHS Digital. The NHS Digital SNOMED CT Browser. Accessed 14/10/2023, 2023. https://termbrowser.nhs.uk/?
168. NHS Digital. NHS TRUD. Accessed 12/11/2023, 2023. https://isd.digital.nhs.uk/trud/user/guest/group/0/home
169. Office for National Statistics. Census 2001 Definitions. Accessed 13/11/2023, 2023. https://ukdataservice.ac.uk/app/uploads/2001_defs_intro.pdf
170. Ajzen I. Understanding Attitudes and Predictiing Social Behavior. *Englewood cliffs*. 1980 1980;
171. Ajzen I. *Attitudes, Personality, and Behavior*. 2nd ed. Open University Press.; 2005.
172. Schmid P, Rauber D, Betsch C, Lidolt G, Denker ML. Barriers of Influenza Vaccination Intention and Behavior - A Systematic Review of Influenza Vaccine Hesitancy, 2005 - 2016. *PLoS One*. 2017;12(1):e0170550. doi:10.1371/journal.pone.0170550
173. CreativeCommons. CC BY 4.0 DEED. Accessed 16/04/2024, 2024. https://creativecommons.org/licenses/by/4.0/
174. University of Missouri. Health Care Access. Accessed 05/10/2023, 2023. https://medicine.missouri.edu/centers-institutes-labs/health-ethics/faq/health-care-access#:~:text=Health%20care%20access%20is%20the,and%20other%20health%2Dimpacting%20conditions.
175. English Longitudinal Study of Ageing. The data we collect. Accessed 18/08/2023, 2023. https://www.elsa-project.ac.uk/the-data-we-collect
176. Cowling TE, Ramzan F, Ladbrooke T, Millington H, Majeed A, Gnani S. Referral outcomes of attendances at general practitioner led urgent care centres in London, England: retrospective analysis of hospital administrative data. *Emerg Med J*. Mar 2016;33(3):200-7. doi:10.1136/emermed-2014-204603
177. Salive ME. Multimorbidity in older adults. *Epidemiol Rev*. 2013;35:75-83. doi:10.1093/epirev/mxs009
178. Clinical Practice Research Datalink. Release Notes: CPRD Aurum May 2022. Accessed 05/09/2023, 2023. https://cprd.com/sites/default/files/2022-05/2022-05%20CPRD%20Aurum%20Release%20Notes.pdf
179. World Health Organisation. ICD-10 Version:2019. Accessed 27/09/2023, 2023. https://icd.who.int/browse10/2019/en
180. NHS England. OPCS-4 CODE. Accessed 27/09/2023, 2023. https://www.datadictionary.nhs.uk/data_elements/opcs-4_code.html

181. NHS Digital. Unplanned hospitalisation for chronic ambulatory care sensitive conditions. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-outcomes-framework/may-2020/domain-2-enhancing-quality-of-life-for-people-with-long-term-conditions-nof/2-3-i-unplanned-hospitalisation-for-chronic-ambulatory-care-sensitive-conditions
182. The National Institute for Health and Care Excellence. NICE 'do not do' recommendations. Accessed 07/02/22, https://www.nice.org.uk/media/default/sharedlearning/716_716donotdobookletfinal.pdf
183. NHS England. Items which should not be routinely prescribed in primary care. Accessed 02/10/2023, 2023. https://www.england.nhs.uk/medicines-2/items-which-should-not-be-routinely-prescribed/
184. NHS Digital. Quality and Outcomes Framework (QOF). Accessed 04/10/2023, 2023. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/quality-outcomes-framework-qof
185. UK Government. Population screening programmes: bowel cancer. https://www.gov.uk/topic/population-screening-programmes/bowel
186. UK Government. Population screening programmes: breast cancer. Accessed 07/02/22, https://www.gov.uk/topic/population-screening-programmes/breast
187. UK Government. Population screening programmes: abdominal aortic aneurysm. Accessed 07/02/22, https://www.gov.uk/topic/population-screening-programmes/abdominal-aortic-aneurysm
188. UK Government. Cervical screening: programme overview. Accessed 15/09/2023, 2023. https://www.gov.uk/guidance/cervical-screening-programme-overview
189. UK Government. NHS Health Checks: applying All Our Health. Accessed 15/09/2022, 2023. https://www.gov.uk/government/publications/nhs-health-checks-applying-all-our-health/nhs-health-checks-applying-all-our-health
190. UK Government. Greenbook Chapter 19. Accessed 07/02/22, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/931139/Green_book_chapter_19_influenza_V7_OCT_2020.pdf
191. UK Government. Greenbook Chapter 25. Accessed 07/02/22, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/674074/GB_Chapter_25_Pneumococcal_V7_0.pdf
192. NHS England. PSA testing. Accessed 02/10/2023, 2023. https://www.nhs.uk/conditions/prostate-cancer/psa-testing/
193. NHS England. Bone density (DEXA scan). Accessed 02/10/2023, 2023. https://www.nhs.uk/conditions/dexa-scan/
194. NHS Digital. Appointments in General Practice report. Accessed 02/10/2023, 2023. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/appointments-in-general-practice
195. Watt T, Sullivan R, Aggarwal A. Primary care and cancer: an analysis of the impact and inequalities of the COVID-19 pandemic on patient pathways. *BMJ Open*. Mar 24 2022;12(3):e059374. doi:10.1136/bmjopen-2021-059374
196. NHS England. Emergency admissions for Ambulatory Care Sensitive Conditions – characteristics and trends at national level Accessed 02/10/2023, 2023. https://www.england.nhs.uk/wp-content/uploads/2014/03/red-acsc-em-admissions-2.pdf
197. NHS England. Blood pressure test. Accessed 02/10/2023, 2023. https://www.nhs.uk/conditions/blood-pressure-test/
198. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. Sep 2012;24(3):69-71.
199. UK Government. Fingertips, Public Health Data, Population Vaccination Coverage: Flu (aged 65 and older). Accessed 14/02/2023, 2023.

https://fingertips.phe.org.uk/search/influenza#page/4/gid/1/pat/159/par/K02000001/ati/15/are/E92000001/iid/30314/age/27/sex/4/cat/-1/ctp/-1/yrr/1/cid/4/tbm/1
200. Tazare J, Wyss R, Franklin JM, et al. Transparency of high-dimensional propensity score analyses: Guidance for diagnostics and reporting. *Pharmacoepidemiol Drug Saf*. Apr 2022;31(4):411-423. doi:10.1002/pds.5412
201. The Health Foundation. The Abdominal Aortic Aneurysm (AAA) screening programme. Accessed 31/07/2023, 2023. https://navigator.health.org.uk/theme/abdominal-aortic-aneurysm-aaa-screening-programme#:~:text=The%20NHS%20abdominal%20aortic%20aneurysm,previously%20could%20self%2Drefer).
202. NHS England. Abdominal aortic aneurysm. Accessed 06/12/2023, 2023. https://www.nhs.uk/conditions/abdominal-aortic-aneurysm/#:~:text=An%20abdominal%20aortic%20aneurysm%20(AAA,they%20could%20burst%20(rupture).
203. Kirby M. Preventing abdominal aortic aneurysm in men. *Trends in Urology and Men's Health,*. https://wchh.onlinelibrary.wiley.com/doi/epdf/10.1002/tre.811
204. Bath MF, Sidloff D, Saratzis A, Bown MJ, investigators UKAGS. Impact of abdominal aortic aneurysm screening on quality of life. *Br J Surg*. Feb 2018;105(3):203-208. doi:10.1002/bjs.10721
205. Cancer Research UK. Screening for people at high risk of bowel cancer. Accessed 14/09/2023, 2023. https://www.cancerresearchuk.org/about-cancer/bowel-cancer/getting-diagnosed/screening-for-people-high-risk
206. UK Government. NHS breast screening (BSP) programme. Accessed 17/04/2024, 2024. https://www.gov.uk/government/collections/nhs-breast-screening-bsp-programme
207. Advisory Committee on Breast Cancer S. Screening for breast cancer in England: past and future. *J Med Screen*. 2006;13(2):59-61. doi:10.1258/096914106777589678
208. NHS England. Breast cancer screening. Accessed 07/12/2023, 2023. https://www.nhs.uk/conditions/breast-screening-mammogram/further-help-and-support/
209. UK Government. Tests and frequency of testing for women at very high risk. Accessed 07/12/2023, 2023. https://www.gov.uk/government/publications/breast-screening-higher-risk-women-surveillance-protocols/tests-and-frequency-of-testing-for-women-at-very-high-risk--2
210. UK Government. Cervical screening: programme overview. Accessed 07/12/2023, 2023. https://www.gov.uk/guidance/cervical-screening-programme-overview#target-population
211. Patnick J. Cervical cancer screening in England. *Eur J Cancer*. Nov 2000;36(17):2205-8. doi:10.1016/s0959-8049(00)00310-5
212. NHS England. NHS Health Check. Accessed 07/12/2023, 2023. https://www.nhs.uk/conditions/nhs-health-check/
213. Tanner L, Kenny R, Still M, et al. NHS Health Check programme: a rapid review update. *BMJ Open*. Feb 16 2022;12(2):e052832. doi:10.1136/bmjopen-2021-052832
214. UK Health Security Agency. When can I get my flu vaccine? Accessed 07/12/2023, 2023. https://assets.publishing.service.gov.uk/media/616d473d8fa8f5297eda6725/UKHSA-12053-flu-vaccine-supplies.pdf
215. NHS England. Flu vaccine. Accessed 07/12/2023, 2023. https://www.nhs.uk/conditions/vaccinations/flu-influenza-vaccine/
216. NHS England. Pneumococcal vaccine. Accessed 07/12/2023, 2023. https://www.nhs.uk/conditions/vaccinations/pneumococcal-vaccination/#:~:text=Your%20GP%20surgery%20will%20usually,employer%20about%20getting%20the%20vaccine.
217. NHS England. PSA testing. Accessed 07/12/2023, 2023. https://www.nhs.uk/conditions/prostate-cancer/psa-testing/
218. Melia J, Moss S. PSA testing in the UK. *Wiley Online*. 2009;14(1):9-13.

219. Blake GM, Fogelman I. The role of DXA bone density scans in the diagnosis and treatment of osteoporosis. *Postgrad Med J*. Aug 2007;83(982):509-17. doi:10.1136/pgmj.2007.057505
220. Cancer Research UK. Bone density scan (DEXA, DXA). Accessed 11/12/1993, 2023. https://www.cancerresearchuk.org/about-cancer/tests-and-scans/bone-density-scan-DEXA-DXA#:~:text=You%20have%20a%20DEXA%20scan,takes%20about%2010%20%E2%80%93%2020%20minutes.
221. Carr M, Kontopantelis E, Doran T, et al. Screen Breast. Accessed 12/12/2023, 2023. https://phenotypes.healthdatagateway.org/phenotypes/PH388/version/776/detail/
222. North Central London Integrated Care System. Bowel cancer screening. Accessed 14/05/2023, 2023. https://nclhealthandcare.org.uk/
223. Cancer Research UK. Primary Care Good Practice Guide: Bowel Cancer Screening. Accessed 16/01/2024, 2024. https://www.cancerresearchuk.org/sites/default/files/bowel_good-practice-guide_feb_23.pdf
224. NHS Digital. Quality and Outcomes Framework (QOF) and primary care business rules. https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof
225. Davidson JA, Banerjee A, Smeeth L, et al. Risk of acute respiratory infection and acute cardiovascular events following acute respiratory infection among adults with increased cardiovascular risk in England between 2008 and 2018: a retrospective, population-based cohort study. *Lancet Digit Health*. Dec 2021;3(12):e773-e783. doi:10.1016/S2589-7500(21)00203-X
226. Staffordshire and Stoke on Trent Integrated Care System. Procedures of low clinical value. Accessed 06/07/2023, 2023. https://staffsstokeics.org.uk/
227. Carey IM, Hosking FJ, Harris T, DeWilde S, Beighton C, Cook DG. *An evaluation of the effectiveness of annual health checks and quality of health care for adults with intellectual disability: an observational study using a primary care database.* vol 5. Health Services and Delivery Research. 2017.
228. Benson T. Why general practitioners use computers and hospital doctors do not--Part 1: incentives. *BMJ*. Nov 9 2002;325(7372):1086-9. doi:10.1136/bmj.325.7372.1086
229. UK Government. People recieving an NHS Health Check per year. Accessed 01/08/2023, 2023. https://fingertips.phe.org.uk/profile/nhs-health-check-detailed/data#page/4/gid/1938132726/pat/159/par/K02000001/ati/15/are/E92000001/iid/91734/age/219/sex/4/cat/-1/ctp/-1/yrr/1/cid/4/tbm/1
230. NHS UK. NHS Health Check best practice guidance. Accessed 07/02/22, https://www.healthcheck.nhs.uk › seecmsfile
231. Fuchs FD, Whelton PK. High Blood Pressure and Cardiovascular Disease. *Hypertension*. Feb 2020;75(2):285-292. doi:10.1161/HYPERTENSIONAHA.119.14240
232. Carey RN, El-Zaemey S. Lifestyle and occupational factors associated with participation in colorectal cancer screening among men and women in Australia. *Prev Med*. Sep 2019;126:105777. doi:10.1016/j.ypmed.2019.105777
233. Kasl SV, Cobb S. Health behavior, illness behavior, and sick-role behavior. II. Sick-role behavior. *Arch Environ Health*. Apr 1966;12(4):531-41. doi:10.1080/00039896.1966.10664421
234. University of Missouri. Health Care Access. Accessed December 15, 2023. https://medicine.missouri.edu/centers-institutes-labs/health-ethics/faq/health-care-access#:~:text=Health%20care%20access%20is%20the,and%20other%20health%2Dimpacting%20conditions
235. Kinjo M, Chia-Cheng Lai E, Korhonen MJ, McGill RL, Setoguchi S. Potential contribution of lifestyle and socioeconomic factors to healthy user bias in antihypertensives and lipid-lowering drugs. *Open Heart*. 2017;4(1):e000417. doi:10.1136/openhrt-2016-000417

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. doi:10.1101/2023.11.08.23298256
237. Office for National Statistics. Main figures. Accessed January 2, 2024. https://www.ons.gov.uk/
238. Clinical Practice Research Datalink. Defining your study population. Accessed 18/03/2024, 2024. https://www.cprd.com/defining-your-study-population#What%20coding%20systems%20are%20used%20in%20CPRD%20data?
239. Clinical Practice Research Datalink. Small area level data based on patient postcode. Accessed 10/12/2023, 2023. https://cprd.com/sites/default/files/2022-02/Documentation_SmallAreaData_Patient_set22_v3.2.pdf
240. UK Government. Greenbook Chapter 14a. Accessed January 2, 2024. https://assets.publishing.service.gov.uk/media/650c0d6afbd7bc0014e54715/Greenbook-chapter-14a-4September2023.pdf
241. UK Government. Greenbook Chapter 19. Accessed July 2, 2022. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/931139/Green_book_chapter_19_influenza_V7_OCT_2020.pdf
242. Jain A, Walker JL, Mathur R, et al. Zoster vaccination inequalities: A population based cohort study using linked data from the UK Clinical Practice Research Datalink. *PLoS One*. 2018;13(11):e0207183. doi:10.1371/journal.pone.0207183
243. Pathirannehelage S, Kumarapeli P, Byford R, Yonova I, Ferreira F, de Lusignan S. Uptake of a Dashboard Designed to Give Realtime Feedback to a Sentinel Network About Key Data Required for Influenza Vaccine Effectiveness Studies. *Stud Health Technol Inform*. 2018;247:161-165.
244. Wang R, Liu M, Liu J. The Association between Influenza Vaccination and COVID-19 and Its Outcomes: A Systematic Review and Meta-Analysis of Observational Studies. *Vaccines (Basel)*. May 20 2021;9(5)doi:10.3390/vaccines9050529
245. Hosseini-Moghaddam SM, He S, Calzavara A, Campitelli MA, Kwong JC. Association of Influenza Vaccination With SARS-CoV-2 Infection and Associated Hospitalization and Mortality Among Patients Aged 66 Years or Older. *JAMA Netw Open*. Sep 1 2022;5(9):e2233730. doi:10.1001/jamanetworkopen.2022.33730
246. Wu S, Du S, Feng R, Liu W, Ye W. Behavioral deviations: healthcare-seeking behavior of chronic disease patients with intention to visit primary health care institutions. *BMC Health Serv Res*. May 16 2023;23(1):490. doi:10.1186/s12913-023-09528-y
247. Prabhakar T, Goel MK, Acharya AS. Health-Seeking Behavior and its Determinants for Different Noncommunicable Diseases in Elderly. *Indian J Community Med*. Jan-Feb 2023;48(1):161-166. doi:10.4103/ijcm.ijcm_106_22
248. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. *BMJ*. Oct 16 2017;359:j4587. doi:10.1136/bmj.j4587
249. Hulme WJ, Williamson E, Horne EMF, et al. Challenges in Estimating the Effectiveness of COVID-19 Vaccination Using Observational Data. *Ann Intern Med*. May 2023;176(5):685-693. doi:10.7326/M21-4269
250. Mahase E. Covid-19: All adults on learning disability register should be prioritised for vaccination, says advisory committee. *BMJ*. Feb 24 2021;372:n547. doi:10.1136/bmj.n547
251. Office for National Statistics. Updated estimates of coronavirus (COVID-19) related deaths by disability status, England: 24 January to 20 November 2020. Accessed 13/11/2023, 2023. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/coronaviruscovid19relateddeathsbydisabilitystatusenglandandwales/24januaryto20november2020

252. Walker JL, Grint DJ, Strongman H, et al. UK prevalence of underlying conditions which increase the risk of severe COVID-19 disease: a point prevalence study using electronic health records. *BMC Public Health*. Mar 11 2021;21(1):484. doi:10.1186/s12889-021-10427-2
253. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. Aug 2020;584(7821):430-436. doi:10.1038/s41586-020-2521-4
254. Hardelid P, Rait G, Gilbert R, Petersen I. Recording of Influenza-Like Illness in UK Primary Care 1995-2013: Cohort Study. *PLoS One*. 2015;10(9):e0138659. doi:10.1371/journal.pone.0138659
255. Hobbs BP. On nonparametric hazard estimation. *J Biom Biostat*. 2015;6doi:10.4172/2155-6180.1000232
256. NHS England. Autumn/Winter (AW) 2023-24 Flu and COVID-19 Seasonal Campaign. https://www.england.nhs.uk/long-read/autumn-winter-aw-2023-24-flu-and-covid-19-seasonal-campaign/#:~:text=To%20maximise%20and%20extend%20protection,cohorts%20from%20the%207%20October.
257. Iacobucci G, Mahase E. Covid-19 vaccination: What's the evidence for extending the dosing interval? *BMJ*. Jan 6 2021;372:n18. doi:10.1136/bmj.n18
258. Andrews N, Tessier E, Stowe J, et al. Duration of Protection against Mild and Severe Disease by Covid-19 Vaccines. *N Engl J Med*. Jan 27 2022;386(4):340-350. doi:10.1056/NEJMoa2115481
259. Nazareth I, King M, Haines A, Rangel L, Myers S. Accuracy of diagnosis of psychosis on general practice computer system. *BMJ*. Jul 3 1993;307(6895):32-4. doi:10.1136/bmj.307.6895.32
260. Eskin M, Simpson SH, Eurich DT. Evaluation of Healthy User Effects With Metformin and Other Oral Antihyperglycemia Medication Users in Adult Patients With Type 2 Diabetes. *Can J Diabetes*. Jul 2019;43(5):322-328. doi:10.1016/j.jcjd.2018.12.001
261. Schrauben SJ, Hsu JY, Wright Nunes J, et al. Health Behaviors in Younger and Older Adults With CKD: Results From the CRIC Study. *Kidney Int Rep*. Jan 2019;4(1):80-93. doi:10.1016/j.ekir.2018.09.003
262. NHS England. Hospital Outpatient Activity. Accessed 24/04/2024, 2024. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity
263. NHS England. Diagnostic Imaging Dataset. Accessed 29/01/2024, 2024. https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/
264. Kamaraju S, Drope J, Sankaranarayanan R, Shastri S. Cancer Prevention in Low-Resource Countries: An Overview of the Opportunity. *Am Soc Clin Oncol Educ Book*. Mar 2020;40:1-12. doi:10.1200/EDBK_280625
265. Government Offices of Sweden. Municipalities and regions. Accessed 24/04/2024, 2024. https://www.government.se/government-policy/municipalities-and-regions/
266. DB. R. *Statistics and causal inference: which ifs have causal answers.* vol 81:961–962. . 1986.
267. Daniel R, Zhang J, Farewell D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biom J*. Mar 2021;63(3):528-557. doi:10.1002/bimj.201900297
268. Jepson R, Clegg A, Forbes C, Lewis R, Sowden A, Kleijnen J. The determinants of screening uptake and interventions for increasing uptake: a systematic review. *Health Technol Assess*. 2000;4(14):i-vii, 1-133.
269. Bunten A, Porter L, Gold N, Bogle V. A systematic review of factors influencing NHS health check uptake: invitation methods, patient characteristics, and the impact of interventions. *BMC Public Health*. Jan 21 2020;20(1):93. doi:10.1186/s12889-019-7889-4
270. Ellis DA, McQueenie R, McConnachie A, Wilson P, Williamson AE. Demographic and practice factors predicting repeated non-attendance in primary care: a national retrospective

cohort analysis. *Lancet Public Health*. Dec 2017;2(12):e551-e559. doi:10.1016/S2468-2667(17)30217-7
271. Wallar LE, De Prophetis E, Rosella LC. Socioeconomic inequalities in hospitalizations for chronic ambulatory care sensitive conditions: a systematic review of peer-reviewed literature, 1990–2018. *International Journal for Equity in Health*. 2020/05/04 2020;19(1):60. doi:10.1186/s12939-020-01160-0
272. Augustsson H, Ingvarsson S, Nilsen P, et al. Determinants for the use and de-implementation of low-value care in health care: a scoping review. *Implement Sci Commun*. Feb 4 2021;2(1):13. doi:10.1186/s43058-021-00110-3
273. Wood ME, Chrysanthopoulou S, Nordeng HME, Lapane KL. The Impact of Nondifferential Exposure Misclassification on the Performance of Propensity Scores for Continuous and Binary Outcomes: A Simulation Study. *Med Care*. Aug 2018;56(8):e46-e53. doi:10.1097/MLR.0000000000000800

Appendix A. Additional Tables

A 1 Defining GP visits in CPRD Aurum

Variable in CPRD Aurum	Value included
Conssourceid	Acute visit, Casualty attendance, Clinic, Emergency appointment, Emergency consultation, Enterprise consultation, Face to face consultation, Follow-up/routine visit, Gp surgery, Home visit, Home visit note, Main surgery, Nursing home, Nursing home visit note, Online services message, Other, Residential home, Residential home visit note, Same day appointment, Surgery consultation, Telephone encounter, Urgent consultation, Walk-in centre, Walk-in clinic. Additional values added in study two and three*: Extra Appointment, Practice Nurse, visit, Visit-Home, Booked Appointment, Normal Home Visit (08:00 - 11:00), Consultation, G P Consultation, Home Visit - In Surgery Hours, Attendance, Branch Surgery, G.P Surgery (Pm), G.P. Morning Surgery, GP Practice, Daytime Visits patients home, Diabetic Clinic, Home of Patient, Seen by Practice Nurse, Seen in Nurses Surgery, Surgery or Clinic, Seen in GP's surgery, Surgery, Surgery Clinic, P.Nurse Clinic, Practice Nurse Clinic, Seen in GPs surgery, Seen in Health Centre, Seen in own home, Nurse Practitioner, Nurse Practitioner Surgery, Nurse Surgery, Nurse Visit, Surgery Attendance, Telephone Advice, Telephone Appt, Telephone Consultation, Telephone Surgery, Treatment Room, Weekday Surgery, Nurse's Treatment Room Clinic, Treatment Room (Nurse), Clinic Premises, Nurse Assessment Clinic, Nurse Minor Illness Clinic, Nurse Practitioner Telephone Advice, G.P. Evening Surgery, Nurse Surgery Triage, Nurse Triage, Nurse Triage Clinic, Nurse Triage Consultation, Telephone (Triage), Telephone Triage Triage By Phone, Nurse telephone triage, Telephone triage encounter, Telephone Triage By Doctor, Triage Telephone, Phone, Telephone Call, Telephone call from a patient, Telephone call to a patient, Open Access Surgery, Open Surgery, Duty Doctor Telephone, Emergency Gp Surgery, G.P.Surgery Urgent Consultation, Emergency Doctor, Emergency Surgery, Surgery Emergency, Urgent Surgery, Urgent Appointment, Saturday Morning Surgery, Same Day Clinic, Urgent Slots, Duty Doctor Urgent Appointment, Duty Telephone Appt, Emergency Nurse Clinic, OPEN DOOR SURGERY, Unbooked Clinic, Walk-In Surgery, Primary Care Centre, Seen in other clinic, Clinic NHS, Clinic note, Community Clinic, Community health clinic, Health Centre, Minor Operations Clinic, Minor Ops Clinic, Three Minute Surgery, Primary care organisation
Consid (only used when conssourceid is “awaiting review”)	Consultation, visit, seen in gp unit, seen in private clinic, seen in rapid access clinic at gp surgery, seen in urgent care centre, online communication.

jobcat	GP – 4, 5, 15, 24, 31, 181, 183 Dr – 1, 41, 91, 116, 119, 121, 126, 173, 177, 197 Nurse – 8, 9, 27, 33, 47, 48, 50, 55, 59, 60, 61, 111 Other healthcare professional - 2, 3, 6, 7, 10:14, 16, 17, 34:37, 42, 43, 52, 54, 58, 62:65, 68, 72, 73, 77, 80, 82, 83, 86:89, 94, 95, 97, 100:102, 105, 106, 112:114, 118, 122, 125, 127, 131, 135, 136, 138, 141, 142, 145, 148, 149, 154, 156, 158, 168, 185, 186, 188, 189, 204, 208

*Since study two and three used a later release of CPRD Aurum, additional values for conssourceid were used.

Note: table adapted from Watt et al, 2022[195].

A 2 Literature that influenced determinants of healthcare utilisation for each of the markers of health-seeking behaviour

Marker	Barriers and influences of uptake
Breast cancer screening	Jepson et al, 2000 conducted a systematic literature review to assess the determinants of breast cancer screening uptake[268]. 34 studies global were included and they assessed the proportion of studies that reported determinants that were significantly associated with breast cancer screening uptake in each of the studies: **Sociodemographic:** • Having insurance (58%). • Being black (20%). • Being African American (7%). • Being white (7%). **Knowledge/behaviour/attitudes/beliefs:** • Having a previous mammogram (65%). • Expressing an intention to attend screening (54%). • Having a previous Pap smear (33%). • Perceiving own health to be poor (25%). • Knowing about mammograms (20%). **Health:** • Perceiving self to be susceptible or vulnerable to cancer (12%). • Visited GP ≤7 times in preceding year (40%). • Having a family history of breast cancer (33%). • Being at moderate risk of breast cancer development (33%). • Having a history of ≥2 major illnesses (25%). • Having a history of breast cancer (25%). **Barriers and facilitating conditions:** • Visiting a GP 4-6 times in previous year (20%). • Receiving a recommendation from doctor (50%). • Being worried about breast cancer (20%). Individuals with the following determinants were less likely to attend screening: **Sociodemographic:** • Being native American (7%). **Knowledge/behaviour/attitudes/beliefs:** • Being a smoker (33%). **Barriers and facilitating conditions:**

	• Having concerns about radiation and mammography (20%). Determinants were association with screening uptake is unclear: **Sociodemographic:** • Age (39%). • Being single/divorced/widowed (27%). • Having a higher level of education (17%).
Bowel cancer screening	Jepson et al, 2000 conducted a systematic literature review to assess the determinants of breast cancer screening uptake[268]. 12 studies global were included and they assessed the proportion of studies that reported determinants that were significantly associated with bowel cancer screening uptake in each of the studies: **Sociodemographic:** • Being older than 65 (50%). • Having a higher level of education (14%). **Knowledge/behaviour/attitudes/beliefs:** • Having had a previous FOBT (80%). • Perceived self-susceptible to cancer (33%). **Health** • Being capable of performing activities of daily living (67%). Individuals with the following determinants were less likely to attend screening: **Barriers and facilitating conditions:** • Being affected by barriers ('Barriers' refers to combined barriers, as in the Health Belief Model; 33%).
Influenza vaccination	Schmid et al, 2017[172] conducted a systematic review to assess barriers to influenza intention and behaviour across the globe. They included 470 articles and clustered according to a conceptual framework according to an extended version of the Theory of Planned behaviour[170], that also included physical, contextual and sociodemographic aspects to the conceptual framework. They found the following barriers were significantly associated with influenza vaccination uptake (with a cut-off of at least 6 studies identifying significance): **Psychological barriers** • Utility/risk perception: higher perceived risk of disease results in higher vaccination uptake, whereas higher perceived risk of adverse events from the vaccination results in lower uptake. • Social benefit: individuals that do not acknowledge the social benefit of the vaccination or perceive low risk of influenza results in lower vaccination uptake. • Subjective norm: low pressure from significant others results in low vaccination uptake. • Perceived behavioural control: lacking perceived behavioural control (e.g., self-efficacy) results in low vaccination uptake. • Attitude: having a negative attitude to the vaccination results in lower uptake. • Past behaviour: individuals who had been vaccinated in previous years showed higher vaccination uptake. • Experience: individuals who had not suffered from influenza previously were less likely to get vaccinated. • Knowledge: lacking general knowledge about influenza and the vaccination was identified as a barrier. **Physical barriers**

	• Unhealthy lifestyles: some articles reported that increased smoking, alcohol consumption and decreased physical activity and higher BMI were associated with lower vaccination uptake. However, the results for these lifestyle factors were mixed and were likely to be confounded by other factors e.g., health status. **Contextual barriers** On the meso-level, the SAGE model acknowledges the influence of external contextual factors on vaccine uptake. • Access: general access due to political, geographical or economic issues influencing production and reliability of supply was not identified as a barrier to vaccination uptake. • Interaction with healthcare system: individuals who interacted less frequently with the health-care system were less likely to get vaccinated. • Cues to action: individuals who do not receive a direct recommendation from medical personnel were less likely to get vaccinated. • System factors: those from deprived areas were less likely to be vaccinated. **Sociodemographic factors:** • Higher age, being female, and being white was associated with higher and lower uptake, with being white being more likely to be reported as a promoter. • Living alone and being unmarried was associated with lower uptake.
NHS health checks	Bunten et al, 2020 conducted a systematic literature review to assess factors influencing uptake of NHS health checks[269]. The review included 9 studies and factors influencing uptake were described: **Sociodemographic** • Age: all studies found that older individuals were more likely to have a health check than younger individuals. • Gender: the majority of studies found that uptake was higher for females vs. males. • Deprivation: lower deprivation was associated with lower uptake. • Ethnicity: results for ethnicity were mixed. One study found that Asian, Black and mixed ethnicity groups had the highest uptake, whereas another found that females from Black African ethnicity had the lowest uptake, with higher uptake among Black Caribbean ethnicity of both genders • Medical and lifestyle risk: one study found that there was higher uptake for individuals with a family history of coronary heart disease and presence of non-CVD comorbidities was associated with higher uptake. **Physical barriers:** • Practice list size: two studies found some evidence that practice list size impacted health check attendance, however the direction of these effects was different, and Cochrane and colleagues found that practice size was not significantly related to uptake.
Primary care DNA	Ellis et al, 2017 conducted a study using routine primary care data from Scotland from 2013 to 2016[270]. They assessed determinants associated missing primary care appointments (zero, low, medium, high). Determinants of missing primary care appointments were: **Sociodemographic** • Males: were more likely to miss primary care appointments than females, but only when this is offset by the number of appointments made. • Deprivation: the most deprived were more likely to miss primary care appointments than less deprived. • Age: older patients were more likely to miss appointments than younger patients. **Physical barriers:**

	• Practices with appointment delays of 2-3 days were more likely to have missed appointments than those that had on the day appointments. • Urban appointments were more likely to have missed appointments than rural.
ACS conditions	Wallar et al, 2020[271] conducted a systematic literature review to assess the socioeconomic determinants of chronic ambulatory care sensitive hospitalisations. They found the following socioeconomic factors were reported to be significantly associated with ACS hospitalisations: • Income: of the 12 studies that fully adjusted for confounding variables, 11 found that lower income was associated with a higher risk or rate of ACS hospitalisation, with 10 studies reporting significant effects. • Education: of the 6 studies that fully adjusted for confounding variables, all found that lower education is associated with higher risk of ACS hospitalisation, with 3 reporting significant effects. • Occupation: one study found that lower occupational class was weakly associated with higher risk of ACS hospitalisations. • Deprivation: all 5 fully adjusted studies observed that higher deprivation is associated with higher risk or rate of ACSC hospitalisations.
Low-value procedures	Augustsson et al, 2021[272] conducted a scoping review to assess the determinants of low-value care, defined as "care that is unlikely to benefit the patient given the harms, cost, available alternatives, or preferences of the patient". 101 studies were included and the most common determinants of low value care was described: **Patient determinants:** • There was no consistent pattern. Some studies reported that younger age was associated with higher low value care, whereas others reported that higher age was. • Severity of illness and characteristics of disease led to higher low-value care use in 17 studies. • Patients who requested non-indicated prescriptions were more likely to receive low-value care in 6 studies. • Expectations from relatives contributed in 2 studies.

Note: publications could not be found for the other markers of health-seeking behaviour. Similar markers were expected to have the same determinants.

Abbreviations: ACS: ambulatory care sensitive; GP: general practice; NHS: National Health Service.

A 3 Operational definitions used to define influenza at-risk and other conditions

Condition	Definition	Lookback period
Chronic respiratory disease	A previous diagnosis of a chronic (long-term) respiratory disease, such as chronic obstructive pulmonary disease (COPD), emphysema or bronchitis, including cystic fibrosis and fibrosing interstitial lung diseases.	Ever before index.
	A current diagnosis of asthma.	Three months prior to index.
Chronic heart disease and vascular disease	A previous diagnosis of chronic heart disease likely to cause long-term increased risk of severe respiratory infection, including angina or myocardial infarction, heart disease, major	Ever before index.

	congenital anomalies requiring long-term follow up such as Fallot's tetralogy.	
Chronic kidney disease	CKD stage 3–5 or estimated glomerular filtration rate (eGFR) ≤60 mL/min/1.73m2 using serum creatinine test results.	Most recent prior to index.
	A previous diagnosis of end stage renal disease.	Ever before index.
Chronic liver disease	A previous diagnosis of a chronic liver disease including cirrhosis, oesophageal varices, biliary atresia and chronic hepatitis.	Ever before index.
Chronic neurological disease	A previous diagnosis of stroke, transient ischaemic attack, or conditions in which respiratory function may be compromised due to neurological disease such as Parkinson's disease, motor neurone disease, multiple sclerosis (MS).	Ever before index.
Diabetes and adrenal insufficiency	A previous diagnosis of diabetes.	Ever before index.
Morbid obesity	Latest body mass index on the index date ≥40 kg/m2, based on latest adult records of height and weight (18 years and above), and reported for age groups ≥20 years.	Most recent prior to index.
Immunosuppression (including asplenia or dysfunction of the spleen)	A previous diagnosis of any solid organ transplant (unless as donor) or any history in the previous year of: aplastic anaemia, leukaemia, lymphoma, receiving a bone marrow transplant, or receiving chemotherapy or radiotherapy or any previous history of asplenia or dysfunction of the spleen (including sickle cell disease but not sickle cell trait). Any history of HIV or other permanent immunosuppression (such as genetic conditions compromising immune function).	Ever before index.
	A previous diagnosis for immunosuppression without further details.	Three months prior to index.
	A previous prescription for high dosage for immunosuppressants. High dosage was identified as follows: • Steroids: >40mg per day for more than a week or >20mg per day for more than 14 days. • Azathioprine: >225mg per day. • Mercaptopurine: >112.5mg per day. • Methotrexate: >25mg per week. If "dosageid" (identifier that allows dosage information on the event to be retrieved)	Three months prior to index.

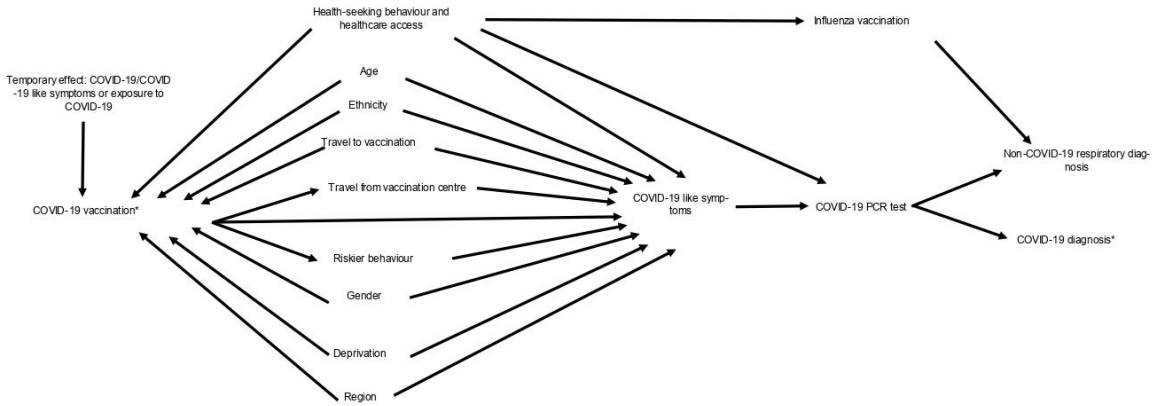
	information was not missing then this would be used to populate tablets per day. Strength information was taken from the product name using a word grab. Product name was taken by linking the "prodcodeid" to the product name in CPRD Aurum product file. Product name was sometimes missing. If information on duration was missing then we imputed using quantity (i.e., assuming one tablet per day). Then we would calculate tablets per day as quantity / duration and then dose per day as the tablets per day * strength. Then scripts were identified as high dose if they were higher than those doses listed above. The limitation of this approach was that we likely underestimated high dose, since assuming one tablet per day amongst those missing is likely an underestimate.	
	A previous prescription of biologic therapies.	One year prior to index.
Severe mental illness	A previous diagnosis of schizophrenia or bipolar disorder, or any mental illness that causes severe functional impairment.	Ever before index.
Severe learning difficulties	Mixed approach that uses register or diagnostic code.	Ever before index.

Note: operational definitions were adapted from Davison et al, 2021[225].

Appendix B. Supplementary Materials Paper One

Supplementary figure titles and legends

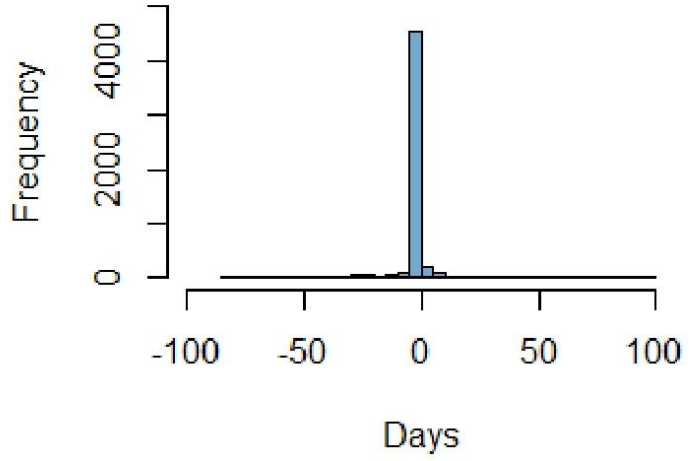
Supplementary Figure 1. Key pathways under investigation in the current study. It should be noted that not all possible pathways are represented in the below figure, however, the key pathways are represented for exposure misclassification, outcome misclassification, confounding, deferral bias, riskier behaviour after vaccination and vaccination itself associated with COVID-19.



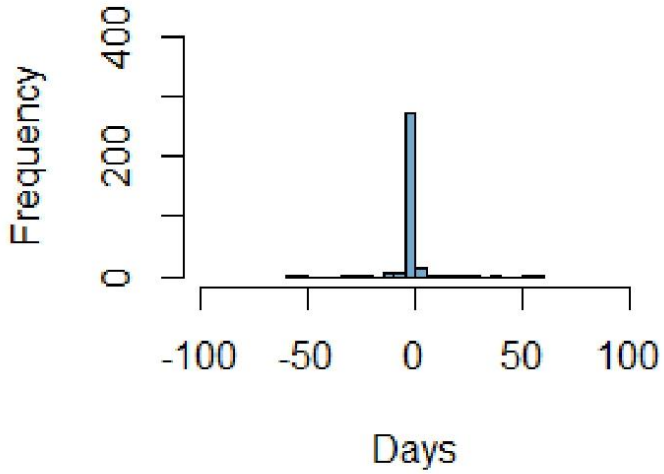
Abbreviations: PCR: polymerase chain reaction. Note: * represents classified as exposed or diagnosed.

Supplementary Figure 2. Histograms representing difference in days between NIMS and questionnaire vaccination date for both A) dose 1 and B) dose 2. ***Negative values indicate that the self-reported vaccination date is earlier than NIMS.***

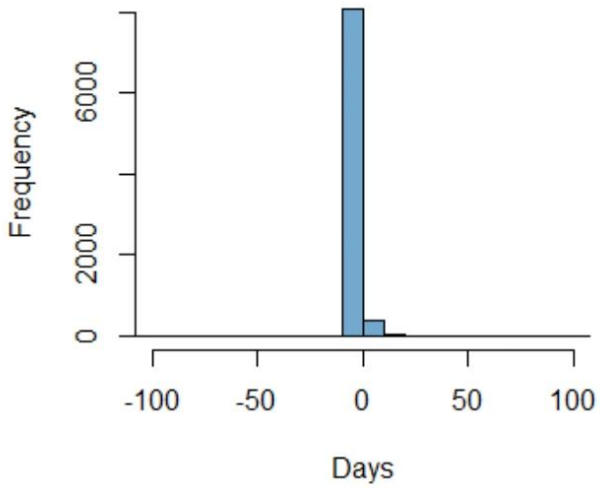
A)



B)



Supplementary Figure 3. Histogram representing difference in days between SGSS and questionnaire onset date (assuming questionnaire is the earlier date). ***Negative values indicate that the self-reported onset date is earlier than SGSS.***



Supplementary table titles and legends

Supplementary Table 1. Summary of Responses in the Questionnaire

Responses to the questionnaire summarised				
		n	**N**	**%**
Respondents:		8,648	23,713	36.5%
Self-reported vaccination status at date of questionnaire response:	Vaccinated:	8,518	8,613	98.9%
	Non-vaccinated:	95	8,613	1.2%
Self-reported CEV status*:		2,337	8,648	27.1%
Self-reported comorbidities:	Chronic heart disease:	663	8,648	7.7%
	Chronic kidney disease:	158	8,648	1.8%
	Chronic liver disease:	29	8,648	0.3%
	Chronic respiratory disease:	881	8,648	10.2%
	Asthma requiring medication:	1,032	8,648	11.9%
	Cancer:	486	8,648	5.6%
	Organ or bone transplant:	18	8,648	0.2%
	HIV/immunodeficiency:	12	8,648	0.14%
	Immunosuppression due to medication:	181	8,648	2.1%
	Seizure disorder:	63	8,648	0.7%
	Chronic neurological disease:	112	8,648	1.3%
	Asplenia or dysfunction of the spleen:	22	8,648	0.3%
	BMI ≥40 kg/m^2:	101	8,648	1.2%
Self-reported symptomatic status when requesting PCR COVID-19 test:		5,539	8,459	65.5%
Amongst those with symptoms (N=5,539), health services accessed during illness:	GP:	1,922	5,539	34.7%
	NHS 111:	659	5,539	11.9%
	Hospital:	503	5,539	9.1%
	Emergency department:	216	5,539	3.9%
	Other healthcare:	121	5,539	2.2%
Amongst vaccinated (N=8,518) length of time from invitation to first dose vaccination:	Less than 2 weeks:	6,131	8,518	72.0%
	2-3 weeks:	1,110	8,518	13.0%
	4 or more weeks:	794	8,518	9.3%
	I had my vaccine before I was eligible:	216	8,518	2.5%
	Missing:	267	8,518	3.1%
Amongst vaccinated (N=8,518), mixing patterns after first dose:	Mixed same amount	5,371	8,518	63.1%
	Mixed more:	445	8,518	5.2%
	Mixed less:	2,435	8,518	28.6%
	Missing	267	8,518	3.1%
Amongst those that delayed their vaccination 2-3 weeks (N=1,110), reason for delay	Not aware I was eligible:	182	1,110	16.4%
	No appointments available:	430	1,110	38.7%
	Prefer to wait to be vaccinated:	107	1,110	9.6%
	Delayed because I had COVID-19 or symptoms:	51	1,110	4.6%

	I was isolating:	41	1,110	3.7%
	I did not have time:	8	1,110	0.7%
	Other:	195	1,110	17.6%
	Missing:	96	1,110	8.6%
Amongst those that delayed their vaccination ≥4 weeks (N=794), reason for delay:	Not aware I was eligible:	84	794	10.6%
	No appointments available:	176	794	22.2%
	Prefer to wait to be vaccinated:	73	794	9.2%
	Delayed because I had COVID-19 or symptoms:	201	794	25.3%
	I was isolating:	35	794	4.4%
	I did not have time:	<5	794	-
	Other:	189	794	23.8%
	Missing:	<5	794	-
Amongst vaccinated with two doses (N=6,952), mixing patterns after second dose	Mixed same amount	4,153	6,952	59.7%
	Mixed more:	1,087	6,952	15.6%
	Mixed less:	1,505	6,952	21.6%
	Missing	207	6,952	3.0%
Amongst non-vaccinated (N=95), reason for no vaccination:	Not called for a vaccine:	<5	95	-
	Not aware eligible:	0	95	0.0%
	No appointments available:	<5	95	-
	Prefer to wait to be vaccinated:	32	95	33.7%
	Expect to get vaccinated soon:	5	95	5.3%
	Have been unwell or have had COVID-19:	26	95	27.4%
	I have been isolating:	<5	95	-
	I did not have time:	0	95	0.0%
	Other:	17	95	17.9%
	Missing:	7	95	7.4%

Abbreviations: CEV: clinically extremely vulnerable; n: numerator; N: denominator; PCR: polymerase chain reaction.

*Phrased in the questionnaire as: "Have you been advised you are part of the clinically extremely vulnerable group?".

Note: since surveys were sent out in March 2021 and individuals were responding to the questionnaire until August 2021. Numbers above reflect self-reported numbers at the time of survey response, rather than at the time of symptom onset in the TNCC study.

Note: Cells <5 have been suppressed and secondary suppression has also been conducted in order to protect patient privacy.

Supplementary Table 2. Baseline characteristics of respondents versus non-respondents using variables from the original study data (NIMS and SGSS)

Characteristic	Respondents, N = 8,648	Non-respondents, N = 15,062	Percentage absolute difference (respondents – non-respondents)	p-value
Vaccine status at symptom onset, n (%)				<0.001
Not vaccinated	1,907 (22.1%)	3,826 (25.4%)	-3.30%	
Vaccinated	6,741 (77.9%)	11,236 (74.6%)	3.30%	
Test result, n (%)				<0.001
Negative	6,541 (75.6%)	12,756 (84.7%)	-9.10%	
Positive	2,107 (24.4%)	2,306 (15.3%)	9.10%	
Age group in years, n (%)				<0.001
70-74	4,423 (51.1%)	6,561 (43.6%)	7.50%	
75-79	2,335 (27.0%)	3,896 (25.9%)	1.10%	
80-84	1,088 (12.6%)	2,260 (15.0%)	-2.40%	
85-89	516 (6.0%)	1,427 (9.5%)	-3.50%	
=>90	286 (3.3%)	918 (6.1%)	-2.80%	
Gender, n (%)				<0.001
Female	4,830 (55.9%)	8,884 (59.0%)	-3.10%	
Male	3,818 (44.1%)	6,178 (41.0%)	3.10%	
Ethnicity, n (%)				<0.001
White	8,022 (92.8%)	12,773 (84.8%)	8.00%	
Non-white	308 (3.6%)	1,572 (10.4%)	-6.80%	
Prefer not to say	318 (3.7%)	717 (4.8%)	-1.10%	
Geographical region, n (%)				<0.001
East of England	1,060 (12.3%)	1,665 (11.1%)	1.20%	
London	718 (8.3%)	1,738 (11.5%)	-3.20%	
Midlands	1,775 (20.5%)	3,299 (21.9%)	-1.40%	
Northeast and Yorkshire	1,360 (15.7%)	2,278 (15.1%)	0.60%	
Northwest	1,226 (14.2%)	2,221 (14.7%)	-0.50%	
Southeast	1,510 (17.5%)	2,352 (15.6%)	1.90%	
Southwest	999 (12.3%)	1,509 (10.0%)	2.30%	
IMD quintile, n (%)				<0.001
1 (most deprived)	1,038 (12.0%)	2,879 (19.1%)	-7.10%	
2	1,337 (15.5%)	2,918 (19.4%)	-3.90%	

3	1,824 (21.1%)	3,091 (20.5%)	0.60%	
4	2,099 (24.3%)	3,196 (21.2%)	3.10%	
5 (least deprived)	2,345 (27.1%)	2,966 (19.7%)	7.40%	
Missing	5	12		
Week of symptom onset, n (%)				0.099
January week 1	12 (0.1%)	21 (0.1%)	0.00%	
January week 2	39 (0.5%)	97 (0.6%)	-0.10%	
January week 3	147 (1.7%)	284 (1.9%)	-0.20%	
January week 4	1,724 (19.9%)	2,797 (18.6%)	1.30%	
February week 1	3,004 (34.7%)	5,294 (35.1%)	-0.40%	
February week 2	2,380 (27.5%)	4,207 (27.9%)	-0.40%	
February week 3	1,342 (15.5%)	2,362 (15.7%)	-0.20%	
Week of COVID-19 test, n (%)				0.821
February week 1	3,551 (41.1%)	6,211 (41.2%)	-0.10%	
February week 2	2,400 (27.8%)	4,212 (28.0%)	-0.20%	
February week 3	2,697 (31.2%)	4,639 (30.8%)	0.40%	
Care home status, n (%)				<0.001
Not care home	8,592 (99.4%)	14,504 (96.3%)	3.10%	
Care home†	56 (0.6%)	558 (3.7%)	-3.10%	
CEV, n (%)				<0.001
Not CEV	7,455 (86.2%)	12,311 (81.7%)	4.50%	
CEV	1,193 (13.8%)	2,751 (18.3%)	-4.50%	

Abbreviations: CEV: clinically extremely vulnerable; IQR: interquartile range; IMD: index of multiple deprivation; n: numerator; N = denominator; NIMS: National Immunisation Management System; SGSS: Second Generation Surveillance System.

†Care home status is likely low in the current study because the study only included those tested in the community (pillar 2), individuals tested in care homes or in hospital are usually tested under pillar 1. In addition, care home status was identified in the current study using an algorithm based on address and the list of official care home residencies in the UK, however, some individuals might have been missed through this.

Note: all tests were conducted using Chi squared test.

Supplementary Table 3. Adjusted odds of COVID-19 after two doses of BNT162b2 or one dose of ChAdOx1 amongst questionnaire sample, respondents and non-respondents, by days since vaccination

	Questionnaire sample	Respondents	Non-respondents
	aOR* (95% CI)	aOR* (95% CI)	aOR* (95% CI)
ChAdOx1 1 dose 0-13	0.87 (0.79-0.96)	0.84 (0.72-0.97)	0.81 (0.71-0.93)
ChAdOx1 1 dose 14+	0.74 (0.65-0.85)	0.73 (0.59-0.90)	0.66 (0.55-0.78)
BNT162b2 1 dose 0-13	0.97 (0.87-1.09)	0.90 (0.76-1.07)	0.91 (0.78-1.06)
BNT162b2 1 dose 14+	0.53 (0.47-0.60)	0.47 (0.39-0.58)	0.51 (0.43-0.59)
BNT162b2 2 dose	0.14 (0.09-0.21)	0.12 (0.06-0.21)	0.13 (0.07-0.21)

Abbreviations: aOR: adjusted odds ratio; CI: confidence interval.

*Adjusted for age, gender, ethnicity, geography, index of multiple deprivation, care home status and week of onset.

Supplementary Table 4. Bias or alternative causal pathways, description, analysis, results, limitations and conclusions

Bias name	Definition	Analysis	Results	Limitation	Conclusions
Exposure (vaccination status) misclassification	Occurs when vaccination status is misclassified. In the context of the current study it was thought that exposure misclassification could be introduced through inaccurate vaccination dates in NIMS.	Vaccination dates (first and second) were compared in NIMS and the questionnaire by reporting the number and proportion of individuals that had an earlier vaccination date in NIMS, earlier vaccination dates in the questionnaire and the same date in both. The percentage of self-reported vaccine dates that were within 3 days +/- of NIMS date was reported. Vaccine effectiveness estimates were re-run using self-reported vaccination dates.	There was no evidence of inaccurate vaccination dates in NIMS (first dose: 9.5% of individuals reported a date that was later, and 7.3% reported a date that was earlier in the questionnaire when compared with NIMS). 89.8% of first dose and 93.3% of second dose self-reported vaccination dates were within 3 days +/- of NIMS date. Vaccine effectiveness after two doses of BNT162b2 decreased from 88% (95% CI: 79-94%) to 84% (95% CI: 74-92%).	A number of individuals did not provide their vaccinations dates in the questionnaire. For example, 38.4% of individuals that reported they received their second vaccination in the questionnaire did not provide a vaccination date. Therefore, the comparison of vaccination dates had to be made amongst those with non-missing data and it had to be assumed that those with missing and non-missing dates did not differ in reporting a vaccination date that differed to the date in NIMS. The use of vaccination cards during the COVID-19 pandemic could have reduced the impact of recall bias on self-reported vaccination dates.	Limited evidence of exposure misclassification.
Outcome misclassification	Occurs when outcome status is misclassified. In the context of the current study it was thought that outcome	Identify the proportion of individuals in the questionnaire reporting they were symptomatic. Note: all individuals included in the	65.5% of individuals reported they were symptomatic in the questionnaire, which was lower in vaccinated (64.7%; versus non-vaccinated: 67.4%) and negative	It is not possible to determine whether either or both of these biases are influencing these results. Individuals were only asked to report symptom onset date if different from the date reported in SGSS	Unclear for outcome misclassification from symptomatic status as

	misclassification could be introduced if individuals were incorrectly reporting their symptomatic status or symptom onset date when requesting their PCR test.	original TNCC study were identified as symptomatic in SGSS. This was reported overall and by vaccination and case status. Compared symptom onset dates in SGSS and questionnaire. Vaccine effectiveness estimates were re-run separately excluding those reporting they were asymptomatic and using self-reported symptom onset dates.	controls (59.7%; versus cases: 83.5%). Vaccine effectiveness for two doses of BNT162b2 increased from 88% (95% CI: 79-94%) in respondents of the questionnaire to 92% (95% CI: 84-97%). Symptom onset dates were not too dissimilar in the questionnaire (5.9% of individuals reported an earlier date, whereas, 2.2% of individuals reported a later date in the questionnaire when compared with SGSS). Vaccine effectiveness after two doses of BNT162b2 decreased from 88% (95% CI: 79-94%) to 87% (95% CI: 77-93%).	and therefore individuals that could not remember their onset date potentially could have left this question blank, which would have been incorrectly interpreted as the same date, rather than missing data.	difference by case status could be subject to recall bias (i.e., those that received a positive test were more likely to recall symptoms) or outcome misclassification (i.e., individuals incorrectly reporting they had symptoms in order to access free testing). No or limited evidence of outcome misclassification from COVID-19 symptom onset date.

Vaccinee bias from confounders	Occurs when vaccinated individuals differ systematically from non-vaccinated due to factors such as underlying health, health-seeking behaviour and access to healthcare which are risk factors for protection against the vaccine preventable disease. In the context of the current study, it was thought that confounding was introduced since comorbidities and other COVID-19 risk factors could not be identified in NIMS or SGSS.	Adjusted for risk factors from questionnaire separately in logistic regression models (also adjusting for each of the variables adjusted for in the original TNCC study). Since there were no variables that changed the vaccine effectiveness estimates when adjusted for separately in the model, it was decided that household size, household type and CEV would be adjusted for all together in a *post hoc analysis* since these variables were considered to be of clinical importance.	Adjusting for household size, household type and CEV (i.e., variables thought to potentially be confounders) as well as other variables adjusted for in the original TNCC study decreased the vaccine effectiveness from 88% (95% CI: 79-94%) to 87% (95% CI: 78-93%) after a second dose of BNT162b2.	Relied on accurate reporting of COVID-19 'at-risk' conditions and other COVID-19 risk factors that was not differential by exposure or case status. There are also likely to be other COVID-19 risk factors such as mobility or frailty that could not be measured in the context of the current study.	No or limited evidence of confounding.

Healthy vaccinee bias from vaccine delay when unwell	Occurs when those that receive a vaccination are more likely to be healthy in the short time period around their vaccination, since individuals are asked to defer vaccination if they are unwell or have the vaccine preventable disease. The impact of this can persist if there are inaccuracies with the vaccine preventable disease onset date.	Identified the proportion of individuals that delayed their first vaccination 2+ weeks from their invitation that reported they delayed their vaccines due to COVID-19/COVID-19 symptoms. Identified the proportion of individuals that have not been vaccinated because they were unwell or had COVID-19 infection. Vaccine effectiveness estimates were re-run excluding those reporting they delayed their vaccination because of COVID-19/COVID-19 like symptoms.	Several individuals who delayed their vaccination 4+ weeks (9.3 %) reported they did so because of COVID-19/COVID-19 symptoms (25.3%). Over a quarter (27.4%) of individuals that had not been vaccinated reported they had done so because they had been unwell or because they had COVID-19. Vaccine effectiveness after two doses of BNT162b2 decreased from 88% (95% CI: 79-94%) to 81% (95% CI: 67-90%).		There was some evidence of individuals delaying their vaccinations because they were unwell or because they had COVID-19. However, this had limited effect on vaccine effectiveness estimates.
Riskier behaviour after vaccination	Occurs when individuals might adopt riskier behaviours after they have received a vaccination, which	Identified the proportion of individuals reporting they mixed more after their COVID-19 vaccination (first and second). Then, the odds of COVID-19 amongst those	Individuals did not report mixing more after vaccinations (first dose: 5.2%; second dose: 15.6%). Those that reported they mixed more after their first vaccination dose did not have an increased odds of COVID-19	There was insufficient data to assess the odds of COVID-19 in those that had riskier behaviour after a second vaccination dose. The responses on the questionnaire might have been subject to desirability bias. Since the study	No or limited evidence of riskier behaviour after vaccination.

	increases their risk of infection compared to non-vaccine recipients (**Figure S1**).	that mixed more versus same/less was compared using logistic regression adjusting for age, gender, ethnicity, CEV, immunosuppressive conditions and month of vaccination dose.	(OR: 0.92, 95% CI: 0.68-1.24) compared to those that mixed the same.	population selected for those that only had their first ever COVID-19 test in February 2021, and the population were also those that responded to a governmental survey, it could be that the study population were those with less risky behaviours that the overall English population. The questions were also answered when there were COVID-19 restrictions in the UK and when the prevalence of COVID-19 was high and therefore individuals might have had less risky behaviours for reasons other than their vaccination.	
Vaccination itself associated with higher risk of COVID-19	Occurs when individuals contract the vaccine preventable disease when they are travelling to, from, or even at, their vaccination centre (**Figure S1**).	Identified the mode of transport taken to vaccination centres (first and second dose) stratified by those that had a positive PCR test within 2 weeks (inclusive) of vaccination and then those that had after 2 weeks. Then the odds of COVID-19 within 2 weeks since first vaccination amongst those	Individuals with a positive test within 2 weeks of first dose were more likely to have taken public transport (4.5% vs. 3.5%) but were less likely to have taken car with individuals outside of their household (11.8% vs. 13.9%) compared with those with a positive test more than 2 weeks after vaccination. There was no association with riskier transport to vaccination centre and odds of COVID-19 (car	This was only assessed by the mode of transport that was taken to the vaccination center. There are other potential factors that could have increased an individual's risk, such as the number of individuals queuing at the vaccination center and the mode of transport taken from the vaccination center (if different from the mode taken there).	No or limited evidence of vaccination itself being associated with COVID-19.

		that took riskier modes of transport (car with those outside of household or public transport) to their vaccination centre would be compared to those that took less risky forms of transport (car alone or with members within household or walked/cycled) using a logistic regression adjusting for age, gender, ethnicity, region and IMD.	with members outside household: OR: 1.28, 95% CI: 0.98-1.67; public transport: OR: 1.26, 95% CI: 0.81-2.03) compared to those that walked/cycled/car alone or with members from own household.		

Abbreviations: CEV: clinically extremely vulnerable; CI: confidence interval; IMD: Index of Multiple Deprivation; NIMS: National Immunisation Management System; PCR: polymerase chain reaction; SGSS: Second Generation Surveillance System; TNCC: test-negative-case-control study; VE: vaccine effectiveness.

Supplementary Table 5. Description of key confounders in those with increased or decreased number of doses using self-reported vaccination date onset date (SGSS) versus unchanged vaccination status.

Characteristic	Questionnaire increases number of doses, N = 81	Unchanged, N = 8,377	Percentage point difference (increased doses – unchanged)	p-value (increased doses vs unchanged)	Questionnaire decreases number of doses, N = 189	Percentage point difference (decreased doses – unchanged)	p-value (decreased doses vs unchanged)
Age				0.309			0.079
70-74	35 (43.2%)	4,307 (51.4%)	-8.2%		81 (42.9%)	-8.5%	
75-79	29 (35.8%)	2,246 (26.8%)	9.0%		60 (31.7%)	4.9%	
80-84	10 (12.3%)	1,054 (12.6%)	-0.3%		24 (12.7%)	0.1%	
85-89	<5	495 (5.9%)			18 (9.5%)	3.6%	
=>90	<5	275 (3.3%)			6 (3.2%)	-0.1%	
Gender				0.805			0.012
Female	47 (58.0%)	4,694 (56.0%)	2.0%		88 (46.6%)	-9.4%	
Male	34 (42.0%)	3,683 (44.0%)	-2.0%		101 (53.4%)	9.4%	
Ethnicity				0.528			0.811
White	77 (95.1%)	7,771 (92.8%)	2.3%		173 (91.5%)	-1.3%	
Non-White	<5	299 (3.6%)	-		8 (4.2%)	0.6%	
Prefer not to say	<5	307 (3.7%)	-		8 (4.2%)	0.5%	
Geographical region				0.643			0.569
London	6 (7.4%)	695 (8.3%)	-0.9%		17 (9.0%)	0.7%	
South England ex-London	30 (37.0%)	3,468 (41.4%)	-4.4%		71 (37.6%)	-3.8%	
North England	45 (55.6%)	4,214 (50.3%)	5.3%		101 (53.4%)	3.1%	
IMD				0.299			0.469
1 (least deprived)	6 (7.4%)	1,007 (12.0%)	-4.6%		25 (13.2%)	1.2%	
2	14 (17.3%)	1,295 (15.5%)	1.8%		28 (14.8%)	-0.7%	
3	13 (16.0%)	1,765 (21.1%)	-5.1%		46 (24.3%)	3.2%	
4	19 (23.5%)	2,044 (24.4%)	-0.9%		36 (19.0%)	-5.4%	
5 (most deprived)	29 (35.8%)	2,261 (27.0%)	8.8%		54 (28.6%)	1.6%	

Missing	0	5			0		
Week COVID-19 symptom onset				0.566			0.032
Jan week 1	<5	12 (0.1%)			<5		
Jan week 2	0 (0.0%)	38 (0.5%)	-0.5%		<5		
Jan week 3	<5	146 (1.7%)			0 (0.0%)	-1.7%	
Jan week 4	13 (16.0%)	1,678 (20.0%)	-4.0%		33 (17.5%)	-2.5%	
Feb week 1	36 (44.4%)	2,898 (34.6%)	9.8%		69 (36.5%)	1.9%	
Feb week 2	17 (21.0%)	2,295 (27.4%)	-6.4%		68 (36.0%)	8.6%	
Feb week 3	14 (17.3%)	1,310 (15.6%)	1.7%		18 (9.5%)	-6.1%	
Week COVID-19 test				0.630			0.123
Feb week 1	37 (45.7%)	3,442 (41.1%)	4.6%		71 (37.6%)	-3.5%	
Feb week 2	19 (23.5%)	2,316 (27.6%)	-4.1%		65 (34.4%)	6.8%	
Feb week 3	25 (30.9%)	2,619 (31.3%)	-0.4%		53 (28.0%)	-3.3%	
Care home status				1			1
Not care home	-	8,323 (99.4%)			-		
Care home	<5	54 (0.6%)			<5		
CEV NIMS				0.834			0.609
Not CEV	71 (87.7%)	7,223 (86.2%)	1.5%		160 (84.7%)	-1.5%	
CEV	10 (12.3%)	1,154 (13.8%)	-1.5%		29 (15.3%)	1.5%	

Abbreviations: CEV: clinically extremely vulnerable; IQR: interquartile range; IMD: index of multiple deprivation; n: numerator; N: denominator; NIMS: National Immunisation Management System; SGSS: Second Generation Surveillance System.

Note: all tests were conducted using two-sided Chi squared test.

Note: cells <5 have been suppressed and secondary suppression has also been conducted in order to protect patient privacy.

Supplementary Table 6. Description of key confounders in those self-reporting they were symptomatic versus asymptomatic in the questionnaire.

Characteristic	Symptomatic, N = 5,539	Asymptomatic, N = 2,920	Percentage point difference (symptomatic – asymptomatic)	p-value
Age				<0.001
70-74	2,976 (53.7%)	1,376 (47.1%)	6.60%	
75-79	1,492 (26.9%)	783 (26.8%)	0.10%	
80-84	605 (10.9%)	446 (15.3%)	-4.40%	
85-89	301 (5.4%)	201 (6.9%)	-1.50%	
=>90	165 (3.0%)	114 (3.9%)	-0.90%	
Gender				<0.001
Female	3,250 (58.7%)	1,468 (50.3%)	8.40%	
Male	2,289 (41.3%)	1,452 (49.7%)	-8.40%	
Ethnicity				0.7944
White	5,145 (92.9%)	2,706 (92.7%)	0.20%	
Non-White	193 (3.5%)	110 (3.8%)	-0.30%	
Prefer not to say	201 (3.6%)	104 (3.6%)	0.00%	
Geographical region				0.385
London	455 (8.2%)	250 (8.6%)	-0.40%	
South England ex-London	2,273 (41.0%)	1,234 (42.3%)	-1.30%	
North England	2,811 (50.7%)	1,436 (49.2%)	1.50%	
IMD				0.875
1 (least deprived)	-	-	-0.10%	
2	849 (15.3%)	467 (16.0%)	-0.70%	
3	1,158 (20.9%)	623 (21.4%)	-0.50%	
4	1,372 (24.8%)	690 (23.6%)	1.20%	
5 (most deprived)	1,501 (27.1%)	790 (27.1%)	0.00%	
Missing	<5	<5		
Week COVID-19 symptom onset				0.543
Jan week 1	6 (0.1%)	6 (0.2%)	-0.10%	
Jan week 2	25 (0.5%)	14 (0.5%)	0.00%	

Jan week 3	91 (1.6%)	51 (1.7%)	-0.10%	
Jan week 4	1,092 (19.7%)	596 (20.4%)	-0.70%	
Feb week 1	1,944 (35.1%)	989 (33.9%)	1.20%	
Feb week 2	1,504 (27.2%)	830 (28.4%)	-1.20%	
Feb week 3	877 (15.8%)	434 (14.9%)	0.90%	
Week COVID-19 test				0.473
Feb week 1	2,247 (40.6%)	1,221 (41.8%)	-1.20%	
Feb week 2	1,544 (27.9%)	810 (27.7%)	0.20%	
Feb week 3	1,748 (31.6%)	889 (30.4%)	1.20%	
Care home status				0.268
Not care home	5,508 (99.4%)	2,897 (99.2%)	0.20%	
Care home	31 (0.6%)	23 (0.8%)	-0.20%	
CEV NIMS				<0.001
Not CEV	4,845 (87.5%)	2,449 (83.9%)	3.60%	
CEV	694 (12.5%)	471 (16.1%)	-3.60%	

Abbreviations: CEV: clinically extremely vulnerable; IQR: interquartile range; IMD: index of multiple deprivation; n: numerator; N: denominator; NIMS: National Immunisation Management System; SGSS: Second Generation Surveillance System.

Note: all tests were conducted using two-sided Chi squared test.

Note: cells <5 have been suppressed and secondary suppression has also been conducted in order to protect patient privacy.

Supplementary Table 7. Description of key confounders by those in those with different versus same symptomatic status using self-reported symptomatic date from the questionnaire.

Characteristic	Different onset date, N = 708	Same onset date, N = 7,937	Percentage point difference (different – same)	p-value
Age				<0.001
70-74	390 (55.1%)	4,032 (50.8%)	4.3%	
75-79	211 (29.8%)	2,122 (26.7%)	3.1%	
80-84	63 (8.9%)	1,025 (12.9%)	-4.0%	
85-89	32 (4.5%)	484 (6.1%)	-1.6%	
=>90	12 (1.7%)	274 (3.5%)	-1.8%	
Gender				0.931
Female	397 (56.1%)	4,431 (55.8%)	0.3%	
Male	311 (43.9%)	3,506 (44.2%)	-0.3%	
Ethnicity				0.761
White	653 (92.2%)	7,367 (92.8%)	-0.6%	
Non-White	27 (3.8%)	280 (3.5%)	0.3%	
Prefer not to say	28 (4.0%)	290 (3.7%)	0.3%	
Geographical region				0.644
London	65 (9.2%)	652 (8.2%)	1.0%	
South England ex-London	286 (40.4%)	3,283 (41.4%)	-1.0%	
North England	357 (50.4%)	4,002 (50.4%)	0.0%	
IMD				0.498
1 (least deprived)	100 (14.1%)	937 (11.8%)	2.3%	
2	100 (14.1%)	1,235 (15.6%)	-1.5%	
3	142 (20.1%)	1,682 (21.2%)	-1.1%	
4	173 (24.4%)	1,926 (24.3%)	0.1%	
5 (most deprived)	193 (27.3%)	2,152 (27.1%)	0.2%	
Missing	0	5		
Week COVID-19 symptom onset				0.525

Jan week 1	0 (0.0%)	12 (0.2%)	-0.2%	
Jan week 2	5 (0.7%)	34 (0.4%)	0.3%	
Jan week 3	16 (2.3%)	131 (1.7%)	0.6%	
Jan week 4	147 (20.8%)	1,577 (19.9%)	0.9%	
Feb week 1	235 (33.2%)	2,767 (34.9%)	-1.7%	
Feb week 2	188 (26.6%)	2,191 (27.6%)	-1.0%	
Feb week 3	117 (16.5%)	1,225 (15.4%)	1.1%	
Week COVID-19 test				0.761
Feb week 1	300 (42.4%)	3,250 (40.9%)	1.5%	
Feb week 2	192 (27.1%)	2,206 (27.8%)	-0.7%	
Feb week 3	216 (30.5%)	2,481 (31.3%)	-0.8%	
Care home status				0.595
Not care home	-	7,884 (99.3%)	-	
Care home	<5	53 (0.7%)	-	
CEV NIMS				0.019
Not CEV	631 (89.1%)	6,822 (86.0%)	3.1%	
CEV	77 (10.9%)	1,115 (14.0%)	-3.1%	

Abbreviations: CEV: clinically extremely vulnerable; IQR: interquartile range; IMD: index of multiple deprivation; n: numerator; N: denominator; NIMS: National Immunisation Management System; SGSS: Second Generation Surveillance System.

Note: all tests were conducted using two-sided Chi squared test.

Note: cells <5 have been suppressed and secondary suppression has also been conducted in order to protect patient privacy.

Supplementary materials

Supplementary Materials 1. Questionnaire sent out to individuals aged ≥70 years with a PCR test from February 1 to 21 2021

COVID-19 Vaccine Effectiveness Survey

This is a request for your help from Public Health England about COVID-19. A few weeks ago you had a test for COVID-19 on <<testdate>>. To understand better important questions about how different activities may affect the chance of catching COVID-19 we are asking you if you can help by filling in this short questionnaire. We need you to do this whether your result was positive or negative.

This one-off survey has 4 parts and should take 15-20 minutes to complete.

Part 1: Details about you at the time you had your COVID-19 test

Part 2: COVID-19 Vaccination Details

Part 3: Details about your illness before getting tested for COVID-19

Part 4: Details in the week before you had symptoms (or a test taken if you did not have symptoms)

Please complete this survey as full as possible by Friday 21 May 2021

If the person this was addressed to can't complete the form themselves it would be really helpful if someone could complete this on their behalf if they are happy for you to do so. Please then answer the questions as if you are that person.

Please enter today's date D D / M M / Y Y

Part 1: Details about you at the time you had your COVID-19 test

1. Please enter your forename _____

2. Please enter your surname _____

3. Please enter your date of birth D D / M M / Y Y

4. What type of accommodation did you live in?

☐ Private home ☐ Care home / Nursing home

☐ Sheltered accommodation ☐ Other

If other please describe

_____

5. You were tested for COVID-19 on <<testdate>>. How many people were you living with on <<testdate>>?

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 or more

6. Have you been advised you are part of the clinically extremely vulnerable group?

For example, have you received a letter or telephone call from your GP to say you are at risk and eligible for a vaccine?

☐ Yes ☐ No

7. Do you have any of the following conditions? *Please tick all that apply*

☐ Chronic Heart Disease ☐ Diarrhoea

☐ Chronic Kidney Disease ☐ Chronic Liver Disease

☐ Chronic Respiratory Disease (excluding asthma) ☐ Asthma requiring medication

☐ Cancer ☐ Organ or Bone Marrow Transplant

☐ HIV/Immunodeficiency ☐ Immunosuppression due to medication*

☐ Seizure Disorder ☐ Chronic Neurological Disease

☐ Asplenia or dysfunction of the spleen ☐ BMI ≥40 kg/m²

☐ None of the above

☐ *If you have immunosuppression, please give further details:

_____

—

Part 2: COVID-19 Vaccination Details

8. Did you receive an invitation for a COVID-19 vaccine (e.g. from your GP or the NHS)?

☐ Yes

☐ No

9. As of today, have you received one or both doses of the COVID-19 vaccine?

☐ Yes, I received 1 dose

☐ Yes, I received 2 doses

☐ No, I have not had a COVID-19 vaccine *(please go to question 20)*

10. How long after you received your invite (or if you were not invited how long after you became eligible) did you receive your first dose of vaccine?

☐ Less than 2 weeks *(Please go to question 12)*

☐ 2-3 weeks *(Please go to question 11)*

☐ 4 or more weeks *(Please go to question 11)*

☐ I had my vaccine before I was eligible *(Please go to question 12)*

11. Why were you not vaccinated sooner?

☐ I was not aware I was eligible

☐ No appointments available

☐ I preferred to wait to get vaccinated

☐ I delayed getting vaccinated because I had COVID-19

☐ I was isolating and did not wish to leave home to get vaccinated

☐ I did not have time

☐ Other _____

12. Please give the date of your first COVID-19 vaccine *It may be difficult to remember exactly; approximate dates are fine.*

D D / M M / Y Y

13. Please specify the brand/type of COVID-19 vaccine you had for your first dose?

☐ Pfizer

☐ AstraZeneca

☐ Unsure

14. How did you travel to the vaccination site?

☐ Walking/ cycling

☐ In a car alone or with members of own household

☐ In a car with member(s) from a different household

☐ Public transport

☐ Other _____

15. In the 3-4 weeks after receiving your first dose, how often have you met/mixed with others outside of your household (e.g. to go to shops, see friends and family)?

☐ I've mixed with people outside of my household for the same amount of time as I did before getting my vaccine

☐ I've mixed more with people outside of my household after getting the vaccine

☐ I've mixed less with people outside of my household after getting the vaccine

16. If you received a second dose, please give the date of your second COVID-19 vaccine *It may be difficult to remember exactly; approximate dates are fine.* D D / M M / Y Y

17. Please specify the brand/type of COVID-19 vaccine you had for your second dose?

☐ Pfizer

☐ AstraZeneca

☐ Unsure

18. How did you travel to the vaccination site for your second dose?

☐ Walking/ cyclingl

☐ In a car alone or with member(s) of own household

☐ In a car with member(s) from a different household

☐ Public transport

☐ Other _____

19. In the 3-4 weeks after receiving your second dose, how often have you met/mixed with others outside of your household (e.g. to go to shops, see friends and family)?

☐ I've mixed with people outside of my household for the same amount of time as I did before getting my vaccine

☐ I've mixed more with people outside of my household after getting the vaccine

☐ I've mixed less with people outside of my household after getting the vaccine

20. If you have not received a COVID-19 vaccine, please can you give us a reason from the options below

Please select all that apply

☐ I have not been called for a vaccine

☐ I was not aware I was eligible

☐ There were no appointments available

☐ I would prefer not to get vaccinated at the moment

☐ I expect to get vaccinated soon but have not had a vaccine yet

☐ I am delaying getting vaccinated because I have been unwell or have had COVID-19 infection……

……..

☐ I am isolating and do not wish to leave home to get vaccinated

☐ I have not had time

☐ Other

_____

Part 3: Details about your illness before getting tested for COVID-19

21. Why were your tested for COVID-19?

☐ I had COVID-19 symptoms

☐ In contact with a case

☐ I was tested in a care home

☐ I was tested in hospital

☐ I was tested as part of surge testing for a variant in my area

☐ I had another illness

☐ Other_____

When you were tested on <<testdate>>, our records showed that you had COVID-19 symptoms starting on <<symptomdate>>. Can you please provide further details about your symptoms.

22. Please confirm if you had symptoms

☐ Yes, I had symptoms *(Please go to question 23)*

☐ No, I did not have symptoms *(Please go to Part 4)*

23. If the date that your symptoms started (<<symptomdate>>) is incorrect, please update when you had your first symptoms.

D	D	/	M	M	/	Y	Y

24. Which of the following symptoms did you have?

Please tick all that apply

- ☐ Fever or chills
- ☐ Runny nose
- ☐ Cough *(Please go to question 21)*
- ☐ Shortness of breath
- ☐ Sore throat
- ☐ Loss of taste and/or smell
- ☐ Nausea
- ☐ Diarrhoea
- ☐ Headache
- ☐ Muscle/ body pain
- ☐ Fatigue
- ☐ Other, please describe: _____

25. How severe would you describe your symptoms?

- ☐ Mild
- ☐ Moderate
- ☐ Severe

26. Have you accessed any healthcare services during your illness (either in person or over the phone)?

Please tick all that apply

- ☐ GP *(Please complete Part 4)*
- ☐ NHS 111 *(Please complete Part 4)*
- ☐ A&E Department *(Please complete Part 4)*
- ☐ Hospital *(Please go to question 27)*
- ☐ None of the above *(Please complete Part 4)*

27. Did you get admitted to hospital due to your illness?

- ☐ Yes *(Please go to question 28)*
- ☐ No *(Please complete Part 4)*

28. If hospitalised, what was the reason for your hospital admission?

☐ COVID-19 related

☐ Unrelated to COVID-19

29. What was your date of admission to the hospital? *It may be difficult to remember exactly; approximate dates are fine.*

D	D	/	M	M	/	Y	Y

30. How many days were you in the hospital?

_____ days

31. Did you receive oxygen in hospital?

☐ Yes ☐ No ☐ Unsure

32. Were you admitted to ICU/ITU (intensive care)?

☐ Yes ☐ No ☐ Unsure

33. Did you receive care involving a ventilator?

☐ Yes ☐ No ☐ Unsure

Part 4: Details in the week before you had symptoms (or a test taken if you did not have symptoms)

34. In the week before you had symptoms (or your COVID-19 test if you did not have symptoms) were you in contact with someone who was unwell with COVID-19 symptoms?

☐ Yes ☐ No ☐ Unsure

35. In the week before you had symptoms (or your COVID-19 test if you did not have symptoms) were you in contact with someone who tested positive for COVID-19?

☐ Yes ☐ No ☐ Unsure

36. In the week before you had symptoms (or your COVID-19 test if you did not have symptoms), did anyone visit your home?

Please tick all that apply

☐ No

☐ Yes, a friend or relative

☐ Yes, a carer

☐ Yes, a doctor or nurse

☐ Another person please describe

_____

37. In the week before you had symptoms (or your COVID-19 test if you did not have symptoms), did you go to a shop or supermarket?

☐ Yes ☐ No ☐ Unsure

38. In the week before you had symptoms (or your COVID-19 test if you did not have symptoms), did you travel in a car with someone outside your home?

☐ Yes ☐ No ☐ Unsure

39. In the week before you had symptoms (or your COVID-19 test if you did not have symptoms), did you go indoors somewhere not in your home where other people go as well (e.g. place of worship, workplace)?

☐ Yes ☐ No ☐ Unsure

40. In the week before you had symptoms (or your COVID-19 test if you did not have symptoms), did you seek medical care outside your home (e.g. dentist, GP, hospital)?

☐ Yes ☐ No ☐ Unsure

41. In the week before you had symptoms (or your COVID-19 test if you did not have symptoms), did you use public transport (e.g. bus, tube)?

☐ Yes ☐ No ☐ Unsure

42. Have you been vaccinated with this season's flu vaccine (since September 2020)?

☐ Yes ☐ No ☐ Unsure

43. Please add any additional information you think would be useful for us to know (e.g. I was part of a vaccine clinical trial)

_____

_____

_____

Appendix C. Approved ISAC application

Medicines & Healthcare products Regulatory Agency

CPRD

1 General information

Protocol reference Id
22_002202

Study title
Exploring the feasibility and effect of adjusting for confounding using markers of health-seeking behaviour and healthcare access in observational cohort studies of influenza and COVID-19 vaccine effectiveness.

Research Area

Drug Effectiveness

Methodological

Does this protocol describe an observational study using purely CPRD data?
No

Does this protocol involve requesting any additional information from GPs, or contact with patients?
No

2 Research team

Role	Chief Investigator
Title	Clinical Assistant Professor
Full name	Helen McDonald
Affiliation/organisation	London School of Hygiene & Tropical Medicine (LSHTM)
Email	helen.mcdonald@lshtm.ac.uk
Will this person be analysing the data?	Yes
Status	pending_confirmation

Role	Corresponding Applicant
Title	Research Associate
Full name	Sophie Graham
Affiliation/organisation	Evidera, Inc
Email	sophie.graham@evidera.com
Will this person be analysing the data?	No
Status	Confirmed

Role	Collaborator
Title	Statistician
Full name	Nick Andrews
Affiliation/organisation	UK Health Security Agency (UKHSA)
Email	nick.andrews@ukhsa.gov.uk
Will this person be analysing the data?	No
Status	Confirmed

Role	Collaborator
Title	PhD student
Full name	Sophie Graham
Affiliation/organisation	London School of Hygiene & Tropical Medicine (LSHTM)
Email	sophie.graham@lshtm.ac.uk
Will this person be analysing the data?	Yes
Status	Confirmed

Role	Collaborator
Title	Professor of clinical epidemiology
Full name	Dorothea Nitsch
Affiliation/organisation	London School of Hygiene & Tropical Medicine (LSHTM)
Email	dorothea.nitsch@lshtm.ac.uk
Will this person be analysing the data?	No
Status	Confirmed

Role	Collaborator
Title	Assistant Professor in Statistics
Full name	Jemma Walker
Affiliation/organisation	London School of Hygiene & Tropical Medicine (LSHTM)
Email	jemma.walker@lshtm.ac.uk
Will this person be analysing the data?	Yes
Status	Confirmed

3

Access to data

Sponsor

London School of Hygiene & Tropical Medicine (LSHTM)

Funding source for the study

Is the funding source for the study the same as Chief Investigator's affiliation?

No

Funding source for the study

National Institute for Health Research - NIHR London Office

Institution conducting the research

Is the institution conducting the research the same as Chief Investigator's affiliation?

Yes

Institution conducting the research

London School of Hygiene & Tropical Medicine (LSHTM)

Method to access the data

Indicate the method that will be used to access the data

Institutional multi-study licence

Is the institution the same as Chief Investigator's affiliation?

Yes

Institution name

London School of Hygiene & Tropical Medicine (LSHTM)

Extraction by CPRD

Will the dataset be extracted by CPRD

No

Multiple data delivery

This study requires multiple data extractions over its lifespan

No

Data processors

Data processor is	Same as the chief investigator's affiliation
Processing	Yes
Accessing	Yes
Storing	Yes
Processing area	UK

4 Information on data

Primary care data

CPRD Aurum

Do you require data linkages

Yes

Patient level data

HES Admitted Patient Care

ONS Death Registration Data

NCRAS data

Covid 19 linkages

Area level data

Do you require area level data?

Yes

Practice level (UK)

Practice Level Index of Multiple Deprivation

Patient level (England only)

Patient Level Index of Multiple Deprivation

Withheld concepts

Are withheld concepts required?
No

Linkage to a dataset not listed

Are you requesting a linkage to a dataset not listed?
No

Patient data privacy

Does any person named in this application already have access to any of these data in a patient identifiable form, or associated with an identifiable patient index?
No

5

Protocol information

Lay Summary

It is useful to study vaccine effectiveness using patient medical records data (e.g., using Clinical Practice Research Datalink) to assess the effect of the vaccinations against new virus mutations and to assess the continued success of these vaccines.

One issue with using patient medical records to study vaccination success is that individuals who regularly access healthcare services for routine health check-ups or vaccinations are likely to have better health outcomes compared to those who never or irregularly visit their healthcare system. This is therefore not a fair comparison when you want to assess the effect of a vaccination by comparing people who are vaccinated to people who are not vaccinated, unless differences in health-seeking behaviour and healthcare access can be accounted for.

Therefore, this study will aim to assess the association between markers or proxies of health-seeking behaviour and healthcare access with vaccinations and COVID-19 and influenza infections. Then the study will assess the impact of accounting for these markers in a COVID-19 and influenza vaccine effectiveness study. This study will be conducted amongst people aged 66 years or older in England.

This work is part of the Health Protection Research Unit in Vaccines and Immunisation, a partnership between the London School of Hygiene & Tropical Medicine and the UK Health Security Agency.

Technical Summary

Observational studies of vaccine effectiveness (VE) are important for assessing VE particularly of new strains of influenza and COVID-19 and assessing duration of protection. However, cohort studies of VE are susceptible to confounding, including by health-seeking behaviour/healthcare access, which can result in over or underestimated VE. This study will explore whether identified markers of health-seeking behaviour/healthcare access (e.g., uptake of screening in nationwide programmes) are associated with influenza/COVID-19 vaccination and infections, and the effect of adjusting for these markers in cohort studies of influenza/COVID-19 VE.

The primary exposures of interest will be COVID-19 and influenza vaccinations, whilst the primary outcomes of interest will be COVID-19 infections, COVID-19-related hospitalisation, COVID-19-related death, acute respiratory infection or influenza/influenza-like-illness (ARI/ILI) infections, ARI or ILI-related hospitalisation and ARI or ILI-related death. This will be conducted in a population aged 66 years identified in CPRD Aurum. The CPRD Aurum dataset will be linked to Hospital Episode Statistics (HES) to identify COVID-19 and ARI/ILI related infections, as well as markers of health-seeking behaviour/healthcare access. The data will also be linked to Office for National Statistics (ONS) data to identify ARI/ILI or COVID-19 related deaths and to the Index of Multiple Deprivation (IMD) to adjust for differences in socioeconomic status. The study analyses will include multivariable logistic regression to assess associations between markers of health-seeking behaviour/healthcare access and each of the exposures and outcomes and Poisson regression to assess COVID-19 and influenza VE with and without adjusting for a combination of markers of health-seeking behaviour/healthcare access.

This study will ensure a better understanding of confounding in VE studies, which will allow for better interpretation of current and future estimates of VE, allowing for better policy-decisions on vaccination strategy. This work is part of the HPRU in Vaccines and Immunisation, a partnership with the UKHSA.

Outcomes to be measured

Primary outcomes:

• COVID-19 infections, which will include the following:
o COVID-19-related primary care record; or
o COVID-19-related hospitalisations; or
o COVID-19-related death.
• COVID-19 related hospitalisations or COVID-19-related death.
• COVID-19 related death.
• ARI or ILI infection, which will include the following:
o ARI or ILI-related primary care record; or
o ARI or ILI-related hospitalisations; or
o ARI or ILI-related death.
• ARI or ILI-related hospitalisations or ARI/ILI-related death.
• ARI or ILI-related death.

Secondary outcome:

• COVID-19 PCR test, regardless of the result (positive, negative or void).

Objectives, specific aims & rationale

The overall research objective is to assess whether identified markers of health-seeking behaviour and healthcare access are associated with COVID-19 and influenza vaccinations and infections, and then quantifying the impact of adjusting for these markers on confounding from health-seeking behaviour and healthcare access in both a COVID-19 and influenza vaccination effectiveness study. The primary hypothesis to be tested is that these markers will appropriately control for confounding from health-seeking behaviour and healthcare access in vaccine effectiveness research.

As a secondary objective, the associations between markers of health-seeking behaviour and healthcare access and COVID-19 PCR testing will also be assessed, so that the potential impact of collider bias in test-negative-case-control studies can be estimated in future studies.

The specific primary objective aims are, among adults aged 66 and over:
1. a) To assess whether markers of health-seeking behaviour and healthcare access are associated with COVID-19 vaccination and COVID-19 outcomes from the introduction of the vaccination programme (8 December 2020 to 31 March 2021); and b) to quantify the impact of confounding from heath-seeking behaviour and healthcare access in a COVID-19 vaccine effectiveness study.
2. a) To assess whether identified markers of health-seeking behaviour and healthcare access are associated with influenza vaccination and acute respiratory infection or influenza/influenza-like-illness (ARI or ILI) outcomes in the 2019/2020 influenza season; and b) to quantify the impact of confounding from health-seeking behaviour and healthcare access in an influenza vaccine effectiveness study.

In addition, COVID-19 infections will be used as a negative control outcome in an influenza vaccine effectiveness study (see objective 3 below). COVID-19 infections are presumed to be a viable negative control outcome, since the confounding structures between influenza and COVID-19 exposures and outcomes are likely to be similar, however, there is expected to be no or minimal biological effect of influenza vaccinations against COVID-19 infections(1). Any impact of the influenza vaccination against COVID-19 infections will be assumed to be due to residual confounding, which will be used to interpret the results of objective 1 and 2. COVID-19 infections will be identified in a time period before COVID-19 vaccinations were approved to avoid positive association between influenza and COVID-19 vaccinations(2). These association between markers and COVID-19 infections in this earlier time period is required in part a) since the confounding structures between COVID-19 infections earlier on in the pandemic might be different to those later on. Therefore, the specific aim for objective 3 is:

3. a) To assess whether identified markers of health-seeking behaviour and healthcare access are associated with COVID-19 outcomes of interest before vaccination (1 July 2020 to 7 December 2020); and b) to quantify the impact of confounding from health-seeking behaviour and healthcare access in an influenza vaccine effectiveness study using COVID-19 infections as a negative control outcome.

The specific secondary objective aim is:
4. To assess the association between identified markers of health-seeking behaviour and healthcare access with COVID-19 PCR testing.

Rationale:
This study aims to increase understanding of potential bias from health-seeking behaviour and healthcare access in observational vaccine effectiveness studies and to improve methods to account for this potential bias. In addition, the study will provide parameters which can be used to quantify the potential impact of collider bias in test-negative-case-control studies that are commonly used in vaccine effectiveness research and have throughout the pandemic been used to inform governmental policy(3-7). A better understanding of both of these biases will allow for better interpretation of current and future estimates of vaccine effectiveness. The current study will also either provide an improved approach to accounting for confounding from health-seeking behaviour and healthcare access or will warrant new methods to control for this bias. Patients would benefit since they would benefit from policy-decisions on vaccination strategy informed by better quality estimates of vaccine effectiveness.

This study is a component of the programme of work of the Health Protection Research Unit

(HPRU) in Vaccines and Immunisation, a research partnership between UK Health Security Agency (UK HSA, formerly Public Health England) and the London School of Hygiene and Tropical Medicine.

Study background

Observational vaccine effectiveness studies are useful for assessing effectiveness of the vaccine against new viral strains and to assess duration of protection. However, there is extensive evidence of confounding in cohort studies of influenza vaccine effectiveness. For example, influenza vaccine effectiveness cohort studies have reported reductions in all-cause mortality by 50%, however, these estimates are implausible since influenza accounts for a maximum of 10% of deaths per year(8). This is because healthcare access and health-seeking behaviour are found in qualitative research to be predictors of vaccine uptake as well as seeking care once infectious disease symptoms present(10).

One option for addressing confounding from health-seeking behaviour / health care access is to use alternative study designs such as the test-negative case-control study design, which is a case-control study conducted only amongst individuals who are tested for the vaccine preventable disease. This design has frequently been used throughout the pandemic to inform UK government policy(3-7). However, this design can only be conducted using databases that include testing result data (negative and positive results required) for the vaccine preventable disease and preferentially also includes information on whether the individual was symptomatic or not. In addition, some authors have theorised that this design could be susceptible to collider bias since selection of the study population is dependent on attendance for testing(11). Therefore, for some studies of vaccine effectiveness, traditional study designs such as cohort study designs are required.

Alternative approaches are therefore required to quantify and account for confounding from health-seeking behaviour / health care access in traditional study designs such as the cohort study. Research is also required to assess the association between health-seeking behaviour / healthcare access and accessing healthcare for symptoms of the vaccine preventable disease so that the extent of potential collider bias can be estimated in test-negative-case-control studies.

It is not known to what extent proxies of health-seeking behaviour and healthcare access can be identified in routinely-collected health records, or which would be the most appropriate proxies to use. Both health-seeking behaviour and healthcare access are drivers of healthcare utilisation, however, healthcare utilisation is a complex phenomenon that is also influenced by a number of micro and macro level factors(12). However, different aspects of healthcare utilisation may be measurable in routine-collected health records, including uptake of preventative care and (non)-attendance to routine care and factors that may be barriers or enablers of accessing care (e.g., ethnicity) are also measurable to some extent. Adjusting for proxy markers of health-seeking behaviour and healthcare access in vaccine effectiveness cohort studies may reduce confounding from health-seeking behaviour and healthcare access in these studies. In addition, assessing the association between proxy markers of health-seeking behaviour and healthcare access and COVID-19 polymerase chain reaction (PCR) testing may confirm whether the test-negative-case-control design introduces collider bias, which is key to interpreting results from vaccine effectiveness studies using these commonly used designs.

Potential markers of health-seeking behaviour and healthcare access were selected based on literature searches for markers that had previously been used in vaccine effectiveness cohort studies. They were also identified based on discussing with clinicians and epidemiologists how available data in the EHR record could represent different aspects of health behaviour. Underlying barriers and influences of each of these different aspects of healthcare behaviour were interpreted using the updated Theory of Planned Behaviour model(12). More information on this model and the barriers and influences to different health behaviours can be found in the supplementary materials. All of the potential markers of health-seeking behaviour and healthcare access will be identified in a pre-pandemic period (i.e., pre-December 2019), since during the COVID-19 pandemic healthcare resource utilisation was highly impacted(13, 14).

The study will include all adults in England 66 years or over, since a) this population is at highest risk of COVID-19 or influenza infections, b) they are all eligible for seasonal influenza vaccination each year and have been prioritised for COVID-19 vaccination c) healthcare resource utilisation is more uniform in this group (since in younger populations, occupation has a huge influence on healthcare resource use(13)).

The study will explore the association between markers of health-seeking behaviour and healthcare access and each of the influenza and COVID-19 exposures and outcomes. In addition, the effect of adjusting for markers of health-seeking behaviour and healthcare access on estimates of vaccine effectiveness for COVID-19 and seasonal influenza. These infections have been selected because both have varying strains and so observational vaccine effectiveness studies are important. COVID-19 vaccine effectiveness exposures and outcomes will be identified in an early pandemic period, after vaccinations were approved in the UK (i.e., December 2020 until March 2021). This time period will be an area of focus since it is when there were high numbers of COVID-19 cases (and therefore more accurate COVID-19 PCR testing), the alpha variant was the dominant variant at the time, human behaviour was relatively similar since the UK was in the second and third lock-down and because vaccine effectiveness estimates are likely to be more similar to clinical trial estimates, which can be used as benchmark estimates for the study (Figure 1). It is also expected that HES linked data will only be available until 31 March 2021 at the time of data extraction, however, if more recent data is available at the time of data extraction, an extended time period may be considered (in which case an amendment would be submitted). Influenza vaccine effectiveness will be assessed pre-pandemic, since influenza infections were highly influenced by changes in mixing patterns caused by the COVID-19 pandemic(15). Looking at influenza vaccine effectiveness in a pre-pandemic period might also be helpful to predict how confounding structures might behave for COVID-19 in a post-pandemic period (if COVID-19 vaccination uptake starts to resemble that of influenza vaccine uptake i.e., seasonal and given to those at highest risk).

Negative controls can be used to assess for residual confounding that remains after adjustment for potential confounding from health-seeking behaviour and healthcare access. Negative controls are either exposure or outcome variables that have the same confounding structure as the exposure and outcome of interest, however, the negative control must not have any association with either the exposure or outcome of interest(15, 16). Therefore, COVID-19 infections will be used as a negative control outcome in an influenza vaccine effectiveness study. This is considered a viable negative control since the confounding structures between influenza and COVID-19 vaccine effectiveness studies are expected to be similar, however, there is expected to be no biological effect of influenza vaccination on COVID-19 infections(1). This analysis will be conducted in a time period before COVID-19 vaccinations were available in order to avoid positive associations between influenza and COVID-19 vaccinations(2).

This study will aim to improve estimates of vaccine effectiveness used for public health planning for new strains of influenza and COVID-19. Linked data from HES and ONS will be required to

improve the robustness of the identification of study outcomes. This study will be conducted in partnership between UK Health Security Agency (UK HSA, formerly Public Health England) and the London School of Hygiene and Tropical Medicine.

Study type

This is a hypothesis-testing study:

Null hypotheses:
• There is no relationship between any of the markers of health-seeking behaviour and healthcare access and COVID-19 vaccine coverage or COVID-19 infections.
• There is no relationship between any of the markers of health-seeking behaviour and healthcare access and influenza vaccine coverage or ARI or ILI infections.
• Adjusting for identified markers of health-seeking behaviour and health-care access does not change COVID-19 vaccine effectiveness estimates in a cohort study design.
• Adjusting for identified markers of health-seeking behaviour and health-care access does not change influenza vaccine effectiveness estimates in a cohort study design.

Alternative hypotheses:
• Some of the markers of health-seeking behaviour and health care access are collectively associated with increased or decreased COVID-19 coverage and COVID-19 infections.
• Some of the markers of health-seeking behaviour and health care access are collectively associated with increased or decreased influenza coverage and ARI or ILI infections.
• Adjusting for the identified markers of health-seeking behaviour and healthcare access in a COVID-19 vaccine effectiveness study decreases vaccine effectiveness estimates in a cohort study design.
• Adjusting for the identified markers of health-seeking behaviour and healthcare access in an influenza vaccine effectiveness study decreases vaccine effectiveness estimates in a cohort study design.

Secondary objective (description of association of markers of health-seeking behaviour and healthcare access with COVID-19 PCR testing)
Null hypothesis
• There is no association between any of the markers of health-seeking behaviour and healthcare access and COVID-19 PCR testing.
Alternative hypothesis
• There is a positive association between markers of health-seeking behaviour and healthcare access and COVID-19 PCR testing.

Study design

This will be an observational cohort study. This design has been selected since it is the most appropriate design for identifying and quantifying confounding structures. Figure 2 that can be found in the supplement represents the study design.

For all three objectives the following definitions will be applied:
• Index date: 1 September 2019 for all individuals. This is the start of the influenza season for objective two (see below), and hence is selected as the baseline date at which pre-pandemic health-seeking behaviour and health care access will be assessed for all analyses.
• Pre-index period: all available time in a patient's record in CPRD Aurum before 1 September 2019.
• Pre-pandemic period: 1 September 2014 to 1 September 2019. This time period will be utilised to identify markers of health-seeking behaviour / health care access. For each marker these will be identified in different look back periods according to how these programmes are administered in UK clinical practice (see Table 1 in the supplement). The different look-back period for each potential marker will be as follows:
o Breast cancer screening: three years prior to index.
o Bowel cancer screening: two years prior to index.
o Influenza vaccination: one year prior to index.

o NHS health checks: five years prior to index.
o Primary care did not attend (DNA): one year prior to index.
o Attendance for ambulatory care sensitive conditions: one year prior to index.
o Low-value procedures: one year prior to index.
o GP practice visits: one year prior to index.

For objective 1 (association of markers of health-seeking behaviour and healthcare access with COVID-19 vaccination and infections, and COVID-19 vaccine effectiveness study), the following time period will be identified:
• COVID-19 period: 8 December 2020 until 31 March 2021. This covers the period from introduction of COVID-19 vaccinations on 8 December 2020 to 31 March 2021 as the expected latest date of data collection in HES at the time of data extraction. If at the time of data extraction there is more recent HES data, then this time period could be extended to cover the latest HES data (in which case an amendment will be submitted). In part 1b) follow-up will start on 8 December 2020 and end at the earliest of death, transfer out of the practice, unclear COVID-19 vaccination status (e.g., the brand is not specified), date of first COVID-19 infection, or the date of receival of any other COVID-19 vaccination other than Comirnaty, or end of the study period (31 March 2021). Vaccination status will be time updated at first and second dose.

For objective 2 (association of markers of health-seeking behaviour and healthcare access with influenza vaccination and infections, and influenza vaccine effectiveness study), the following time period will be identified:
• Influenza period: 1 September 2019 to 29 February 2020 (i.e., the 2019/2020 influenza season until end February). Typically, the influenza season would continue until the end of March, however, this will end a month earlier to prevent overlap with the COVID-19 pandemic, which might have influenced influenza vaccination uptake and its determinants. Ending the study in February rather than March also has the advantage of restricting the study to influenza vaccinations received in time to be effective during the influenza season. In part 2b) follow-up will start on 1 September 2019 and end at the earliest of death, transfer out of the practice, end of the study period (29 February 2020) or the date of first influenza infection. Vaccination status will be time updated at first vaccination dose.

For objective 3 (association of markers of health-seeking behaviour and healthcare access with COVID-19 outcomes [during an earlier time period than objective 1] and an influenza vaccine effectiveness study with a negative control outcome), the following time periods will be identified:
• Negative control outcome period: 1 July 2020 until 7 December 2020. This time period starts once PCR testing was made widely available in the UK and ends at the introduction of the COVID-19 vaccination programme, to avoid confounding by COVID-19 vaccination(2). In part 3b) follow-up will start on 1 July 2020 and end at the earliest of death, transfer out of the practice, end of the study period (7 December 2020) or date of first COVID-19 infection. A history of influenza vaccination during the previous influenza season will be defined as a binary variable.

For the secondary objective (association of markers of health-seeking behaviour and healthcare access and COVID-19 PCR testing), the following time period will be identified:

• COVID-19 testing period: 1 July 2020 until 31 March 2021. This time period was selected since this is after PCR testing was made widely available in the UK. Follow up will end at the earliest of death, transfer out of the practice or end of the study period.

Feasibility counts

In 2020, the population of England over the age of 65 years was 10,464,019 (mid-year estimate: ONS(17). Since Aurum currently covers 19.83% of the UK population, that is majority in England (99.03%) it is estimated that there will be around 2,075,015 individuals that could potentially be eligible for inclusion in the study sample. However, it is assumed that the drop in sample size will be minimal since the inclusion/exclusion are not very restrictive. Therefore, as a conservative estimate it is anticipated that the study sample size will lie between 1,000,000 and 1,850,000. Since the expected number of patients is over 600,000 patients a data minimisation file will be completed.

Planned use of linked data and benefit to patients in England and Wales

Data from CPRD will be linked with data from Hospital Episode Statistics (HES) admitted patient care, the practice and patient level index of multiple deprivation (IMD), and the Office for National Statistics (ONS) mortality records.

Justification on planned use of linked data:
HES-linked data is required to identify the following:
• Outcomes: COVID-19 and ARI/ILI infections and COVID-19 and ARI/ILI-related hospitalisations.
• Potential markers of health-seeking behaviour: nationwide screening programmes, low-value procedures and attendance for ambulatory care sensitive conditions.
• Covariates: COVID-19 'at-risk' groups.

CPRD data linked to patient level IMD is required to assess socioeconomic status. Patient level IMD will be adjusted for in the models as a confounder of vaccine effectiveness. Practice-level IMD will be used to supplement IMD for individuals with missing patient-level IMD.

Linked ONS data is also required to obtain death registration data, as well as the cause of death using ICD-10 codes, which is required to identify COVID-19-related and ARI/ILI-related death. Patients will be censored during follow-up at the time of death.

Through the use of linked data, the study findings will benefit patients in England since it will enable more robust studies that assess COVID-19 and influenza vaccine effectiveness to be conducted so that we can obtain a better understanding of the effect of these vaccinations, particularly since observational data is continuously being used to assess vaccine effectiveness against new variants, to assess the duration of protection and to inform mathematical modelling of the ongoing pandemic.

Definition of the study population

For all objectives the following general population criteria will be applied to identify the overall study population of interest. As previously mentioned, the study population will comprise older adults (66 years) since this population is at highest risk for COVID-19 and ARI or ILI infections, since they are prioritised each year for influenza vaccinations and were prioritised for COVID-19 vaccinations early on in the pandemic and since healthcare resource utilisation is expected to be more uniform in this group compared to a younger population, whose healthcare utilisation is highly influenced by occupation. Those 66 years or older have been selected rather than those 65 years or older, to ensure that those that had just become eligible for influenza vaccinations in the one year before 1 September 2019 could be identified.

Inclusion criteria:
• Individuals with a registration start date (variable: 'regstartdate') on or before 1 September 2018 and did not have a registration end date (variable 'regenddate') before 1 September 2019. This will allow for at least one year registration before the index date for all individuals to assess pre-pandemic health-seeking behaviour and healthcare access.
• Individuals 66 years on 1 September 2019.
• Individuals registered at a general practice with linkage available to HES on 1 September 2019.
• Individuals with an acceptable patient record (variable: acceptable=1).

For each of the following objectives the following exclusion criteria will also be applied to ensure that only susceptible individuals will be included in each of the vaccine effectiveness estimates:

Objective 1 a) and b) exclusion criteria:
• Individuals with a COVID-19 infection (see definition in exposures, outcomes and covariates section below) that occurred any time before 8 December 2020.
• Any death or transfer out of the practice that occurred between 1 September 2019 and 8 December 2020, as follow-up will start on 8 December 2020.

Objective 2a) and c) exclusion criterion:
• Individuals with an ARI/ILI infection (see definition in exposures, outcomes and covariates section below) that occurred between 1 April 2018 and 31 August 2019.

Objective 3 a) and b) exclusion criteria:
• Individuals with a COVID-19 infection (see definition in exposures, outcomes and covariates section below) that occurred any time before 1 July 2020.
• Any death or transfer out of the practice that occurred between 1 September 2019 and 1 July 2020, as follow-up will start on 1 July 2020.

Secondary objective exclusion criterion:
• Any death or transfer out of the practice that occurred between 1 September 2019 and 1 July 2020, as follow-up will start on 1 July 2020.

Selection of comparison groups/controls

Vaccination status will be time-updated and therefore all individuals will be included as non-vaccinated until the time point at which they received their first vaccination, separately for influenza and COVID-19 vaccines. For COVID-19 (objective 1), vaccination status will also be further updated at the time point that individuals receive their second vaccination. Since it is expected that data will only be available until March 2021, there will be no individuals who have received a third or even fourth COVID-19 vaccination during follow up, however, if there is more recent HES data at the time of extraction, then vaccination status will also be updated when individuals receive a third and fourth vaccination (and an amendment submitted). This approach was selected since it avoids the introduction of immortal time bias (as individuals that receive vaccinations are limited to those that survive long enough to do so(25).

Exposures, outcomes and covariates

Below includes the operational definitions that will be used to identify the exposures, outcomes and covariates.

A current study applicant (Helen McDonald) collaborated on a previous study in CPRD Aurum led by Jennifer Davidson (approved CPRD protocol application number: 20_000135 and 21_000380), which developed SNOMED and ICD-10 codes for COVID-19 vaccinations, influenza vaccinations, ARI/ILI infections and influenza 'at-risk' groups, which the current study proposes to re-use. References to these published code lists are provided below and code lists for exposures and outcomes are provided in the supplementary materials.

Exposures:
Objective 1:
• COVID-19 vaccination: any dose of Comirnaty (medcodes(26)) identified in CPRD Aurum during the COVID-19 period (8 December 2020 to 31 March 2021). For the descriptive statistics and objective 1a) vaccination status will be polytomous and defined as only having one dose, only having 2 doses and non-vaccinated using all follow-up data. The rationale for only including Comirnaty is that these patients are expected to have the longest length of follow-up in the data (since this vaccine was approved the earliest). For objective 1b) vaccination status will be time updated 14 days after first dose second dose since this is the time period required for an antibody immune response to the vaccine to develop(27).
Objective 2 and 3:
• Influenza vaccination: an influenza vaccination (medcodes(28, 29)) identified in CPRD Aurum during the influenza period (1 September 2019 to 29 February 2020). For the descriptive statistics and objectives 2a) and 3b) vaccination status will be binary (vaccinated at any point during the influenza period, or non-vaccinated). For objective 2b) vaccination status will be time updated 14 days after first dose since this is the time period required for an antibody immune response to the vaccine to develop(30).

Outcomes:
For all of the primary outcomes for objective 1, 2 and 3, these will be identified as binary outcomes using all follow-up data for part a) of each objective and then as time updated variables for part b) of each objective:

Primary objective:
Objective 1 and 3:
• COVID-19 infection: a primary care record of COVID-19 infection (SNOMED codes(31)) in CPRD Aurum, a hospitalisation (ICD-10 codes: U07.1 and U07.2) in HES or a death (ICD-10 codes: U07.1 and U07.2) in ONS for COVID-19. In HES both a primary and secondary diagnoses will be identified and in ONS both main and underlying diagnoses (any position) will be identified.
• COVID-19-related hospitalisation or COVID-19-related death: a hospitalisation in HES or death in ONS for COVID-19 (ICD-10 codes: U07.1 and U07.2). In HES, codes only in the primary position and in ONS only codes in the main position will be considered.
• COVID-19-related death: a death in ONS for COVID-19 (ICD-10 codes: U07.1 and U07.3). Only codes in the main position will be considered.
Objective 2:
• ARI/ILI infection: a primary care record (SNOMED codes(32)) in CPRD Aurum, a hospitalisation (ICD-10 codes:(33)) or a death (ICD-10 codes(33)) in ONS for ARI/ILI. In HES, codes in both the primary and secondary position will be identified and in ONS both main and underlying diagnoses (any position) will be identified.
• ARI/ILI-related hospitalisation or ARI/ILI-related death: a hospitalisation in HES for ARI/ILI (ICD-10 codes(33)). In HES, codes only the primary position will be considered.
• ARI/ILI-related death: a death (ICD-10 codes(33)) in ONS for ARI/ILI. Only codes in the main position will be considered.

Secondary objective:
• COVID-19 PCR testing: a COVID-19 PCR test with any result (positive, negative or void) recorded as a SNOMED code in CPRD Aurum.

The below potential markers of health-seeking behaviour and healthcare access have been selected since they reflect individuals' interaction with routine care and access to preventative care services for which all individuals aged 66 years and above in England are eligible. Multiple markers need to be looked at, since each of these markers reflect different aspects of healthcare utilisation since they have different determinants (barriers and influences). For each potential marker these will be identified in each of the relevant look back periods, which each reflect the recommended or typical frequency of these events in routine clinical practice (Table 1).

Potential markers of health-seeking behaviour/health care access:
• Breast cancer screening: SNOMED codes in CPRD Aurum.
• Bowel cancer screening: SNOMED codes in CPRD Aurum.
• Prior influenza vaccination: Medcodes in CPRD Aurum.
• NHS health care checks: SNOMED codes in CPRD Aurum. It should be noted that NHS healthcare checks are only eligible to those that do not have an underlying health condition(34). Therefore, it is likely that this marker will be assessed in combination with did not attend primary care visits (see below), since these represent opposite determinants (i.e., NHS health checks are amongst those without chronic conditions that face limited barriers to accessing healthcare, whereas primary care DNAs are mostly amongst those with chronic conditions that face barriers to accessing healthcare).
• Low-value procedures: OPCS codes in HES APC that are reported in the National Institute for health and Care Excellence "do not do" book of procedures(35).
• GP practice visits: any 'consdate' variable identified in CPRD Aurum.

Potential markers of lack of health-seeking behaviour/health care access:
• Did not attend primary care visit: SNOMED codes in CPRD Aurum.
• Ambulatory care sensitive: method of admission is accident and emergency (ADMIMETH value 21=accident and emergency or dental casualty department of the Health Care Provider) and ICD-10 codes in HES APC for ACS conditions(36). Only the primary cause of admission will be utilised. This definition and the codes used is the same as Carey et al, 2017(37).

Covariates:
The following potential confounders of COVID-19 and influenza vaccine effectiveness will be included as covariates:
• Age: will be estimated using the 'yob' variable in CPRD Aurum and will be defined on individuals index date. Since only year of birth will be available, all individuals will be assumed to be born at the middle of each year. Age will be described as a continuous variable, but may be modelled as a categorical variable in 5-year bands depending on model fit.
• Gender: will be identified using the 'gender' variable in CPRD Aurum.
• IMD: individual level deprivation score will be identified in the relevant IMD dataset, categorised in quintiles from most to least deprived, or missing. If patient level IMD, is missing then practice level IMD score will be used from the practice level IMD dataset.
• Ethnicity: will be identified using clinical codes in the observational file of CPRD Aurum, categorised in the 5+1 Census categories as South Asian, Black, Other, Mixed, White or not recorded. If missing in CPRD Aurum the data will be supplemented with data from HES. The method by Mathur et al, 2014(35) used to select ethnicity in the instance of multiple records of ethnicity will be utilised.
• Region: will be identified using the 'region' variable in CPRD Aurum.
• Influenza 'at-risk' groups: will be identified based on clinical risk groups that the government used to prioritise vaccination for influenza. Influenza risk groups have been used instead of COVID-19, since these individuals would have known from the start of the pandemic that they were at higher risk of COVID-19 and therefore would have been more likely to get vaccinated or would have received their vaccination in a timely manner. The comorbidities that will be identified using SNOMED or ICD-10 codes in CPRD Aurum in HES in the pre-index period (unless otherwise specified) using codes lists from Davison et al, 2021(36), unless otherwise specified. The categories for these comorbidites are:
o Chronic respiratory disease.
o Chronic heart disease (using SNOMED and ICD-10 codes from OpenCodelists(37).
o Chronic kidney disease.

o Chronic liver disease.
o Chronic neurological disease.
o Diabetes mellitus.
o Immunosuppression: a relevant SNOMED or ICD-10 diagnoses or OPCS code (to identify transplants and chemotherapy) in CPRD Aurum or HES APC in the pre-index period.
o Asplenia or dysfunction of the spleen.
o Morbid obesity: BMI score 40 kg/m2 or higher recorded in CPRD Aurum in the one year prior to 1 September 2019. The height and weight values that are closest to and before index will be used.
• Additional categories for COVID-19 'at-risk' groups: the below comorbidities will be treated separately to the above comorbidities, since these individuals likely experience barriers to accessing health services and therefore these groups are likely to have lower vaccine uptake. Again, these comorbidities will be identified using SNOMED or ICD-10 codes in CPRD Aurum in HES in the pre-index period. The categories that will be identified are:
o Severe mental illness (using SNOMED and ICD-10 codes from(38) Davison et al, 2022(39).
o Learning difficulties (using SNOMED and ICD-10 codes from Davidson et al, 2022(38).
• Underlying health conditions: influenza and COVID-19 at risk groups will be combined to increase sample size for the modelling steps in the analyses into those with an

immunosuppressive condition (immunosuppression and asplenia and dysfunction of the spleen), those with other comorbidities and those without any comorbidities identified in the pre-index period.

Data/statistical analysis

Descriptive analysis

Firstly, the coverage of each of COVID-19 and influenza vaccinations within the relevant study periods will be described overall, and then stratified by age and calendar month. All covariates will be described at index date (1 September 2019), comparing first amongst those that are vaccinated with Comirnaty versus those that are non-vaccinated against COVID-19 during the COVID vaccination period, and then separately comparing amongst those that are vaccinated with influenza vaccination versus those that are not vaccinated against influenza within the 2019/2020 influenza vaccination season (1 September 2019 to 29 February 2020). Missing data for each variable will also be described during this step. Covariates will also be described at 8 December 2020, so that those eligible for a COVID-19 vaccination at the start of the COVID-19 period can also be described. The prevalence of each of the markers of health-seeking behaviour and healthcare access will also be described at index date using each of the relevant look-back periods. All categorial covariates will be described using number and proportion for each category and compared using the Chi-squared test. Age will be described using mean, standard deviation (SD), median and interquartile ranges and compared using the t- or Mann Whitney U-test.

A sensitivity analysis will be conducted to identify the presence of potential misclassification of markers of health-seeking behaviour and healthcare access that could be introduced through high rates of patient attrition (i.e., patients changing practices regularly). To assess this, the distribution of lookback period amongst all patients will also be described using mean, median, SD and interquartile ranges. If majority of patients have less than a 5-year lookback then, the prevalence of each of the markers will be described using only a one-year lookback only. The prevalence of these markers will be compared to the original prevalence estimates to assess for the potential impact of misclassification.

Objective 1: a) To assess whether markers of health-seeking behaviour and healthcare access are associated with COVID-19 vaccination and COVID-19 outcomes in the COVID-19 period and b) to quantify the impact of confounding from health-seeking behaviour and healthcare access in a COVID-19 vaccine effectiveness study.

For part a) of objective 1 multivariable logistic regression will be used to assess the association between each marker with both COVID-19 vaccination and COVID-19 infections, COVID-19-related hospitalisations with and without COVID-19 related deaths during the COVID-19 period. The binary outcomes of interest will be vaccination status, COVID-19 infection, hospitalisation and death. The association between each marker of health-seeking behaviour and healthcare access will be assessed separately for its association with each outcome. Once these have been looked at separately, different combination of markers will be included as exposures in the logistic regression models. Different combinations of markers will be considered using a priori knowledge (e.g., NHS health checks and primary care DNAs will be combined into one marker) and based on associations observed for individual markers. Before combining variables a correlation matrix of all the markers will be plotted to check for potential multi-collinearity. A priori confounders will include age (continuous, categorical or quadratic, depending model fit) and gender. Additional potential confounders (region, underlying health condition, IMD and ethnicity) will also be explored. Each will be added into the models in a stepwise process and will only be included in the models if the odds ratio changes, otherwise they will be removed and the next covariate will be assessed. Ethnicity will be adjusted for as the final covariate each using a complete case analysis – the main analysis will be restricted to those with non-missing ethnicity. There are not expected to be any other variables with missing data, but if there are then the same process will be repeated. As with the other potential confounders, if adjusting for ethnicity does not change the results, then this variable will not be included in the final model since it could be that although these are risk factors for COVID-19 infection, they have already been adjusted for

through other variables (e.g., region).

For part b), the incidence rate ratio (IRR) of COVID-19 infections, COVID-19-related hospitalisations with and without COVID-19-related death will be estimated using Poisson regression models comparing vaccinated individuals versus non-vaccinated, using time updated vaccination status. The approach for adjusting for confounders will be the same as part a). The follow-up periods will be split on calendar time (interval to be assessed at data delivery stage) so that differences in the incidence rate of COVID-19 infections can be assessed.

Then, using all the markers that had the strongest association (either positively or negatively) with each of the COVID-19 outcomes in part a), these will be additionally adjusted for in the model one at a time. These will be added in combination, until the IRR no longer changes. If the IRR doesn't change, then the last marker will be removed from the model, and the next marker will be added into the model. This process will be repeated until all markers have been assessed for inclusion to see if they impact the IRR. This model will be compared to the model that does not adjust for markers of health-seeking behaviour or healthcare access.

Objective 2: a) To assess whether identified markers of health-seeking behaviour and healthcare access are associated with influenza vaccination and influenza outcomes and b) to quantify the impact of confounding from health-seeking behaviour and healthcare access in an influenza vaccine effectiveness study.

The analyses of 1a and 1b) will be repeated for influenza vaccination (binary for 2a and time updated for 2b) during the 2019/2020 influenza vaccination season (1 September 2019 to 29 February 2020). Outcomes of interest (influenza vaccination status, ARI/ILI infection, ARI/ILI-related hospitalisation and ARI/ILI-related death) will be identified in the influenza period.

Objective 3: a) To assess whether identified markers of health-seeking behaviour and healthcare access are associated with COVID-19 outcomes of interest; and b) to quantify the impact of confounding from health-seeking behaviour and healthcare access in an influenza vaccine effectiveness study using COVID-19 infections as a negative control outcome.

The analysis in 1a) will be repeated to assess the association between markers of health-seeking behaviour and healthcare access and COVID-19 infections, COVID-19-related hospitalisations and COVID-19-related death in the negative control outcome period. The associations between the markers and COVID-19 outcomes during this period will be compared to the associations with COVID-19 outcomes in objective 1a). If there is a large difference in these associations, then this will be considered in the interpretation when considering the extent to which residual confounding in objective 3 can be extrapolated to objectives 1 and 2. For objective 3b) the exposure of interest will be history of influenza vaccination during the previous influenza season, which will be defined as a binary variable.

Secondary Objective: To assess the association between identified markers of health-seeking behaviour and healthcare access with COVID-19 PCR testing.

For the secondary objective, Poisson regression models will be used to assess the association between each marker of health-seeking behaviour/healthcare access and COVID-19 PCR testing. Poisson regression models are to be utilised since it is necessary to identify the rate of COVID-19 PCR testing over the course of the study period (1 July 2020 to 31 March 2021). Repeat COVID-19 PCR tests will be considered since it is likely that an elderly population will have more than one COVID-19 PCR test within the study period (pandemic period). The exposures of interest will be each of the markers of health-seeking behaviour / healthcare access and the outcome of interest will be COVID-19 PCR testing. Each model will be adjusted for potential covariates as in part 1a). Then based on the markers that are most strongly positively associated with COVID-19 PCR testing will be added into the models in combination. Then the markers that are most negatively associated with COVID-19 PCR testing will be added into the models in combination. The most positive and the most negative estimates will then be used to get a range of estimates that can be used a parameters in simulation studies to assess the impact of potential

collider bias in different observational study designs. The follow-up time in the poisson regression models will be split by calendar time, since it is expected that associations will change over the course of the study period. Robust standard errors will be used to account for clustering.

Plan for addressing confounding

During each analysis step all covariates that are identified as confounders will be adjusted for in both steps of the analysis for all objectives. The negative control outcome influenza vaccine effectiveness study (objective 3) aims to quantify any remaining residual confounding from health-seeking behaviour and healthcare access and will aid in the interpretation of any residual confounding that likely remains in objective 1 and 2.

Plans for addressing missing data

Missing data for individual-level deprivation will be supplemented with practice-level deprivation status.

Missing data on ethnicity in primary care records will be supplemented with ethnicity recorded in Hospital Episode Statistics. To handle remaining missing ethnicity data, we will not attempt multiple imputation since the assumption that data are missing at random is unlikely to be met. We will explore potential confounding in a complete case analysis, since the assumption required (that missing missingness is unrelated to the study outcomes, given the covariates included in the model) is more likely to be met.

Missing data on BMI will not be supplemented with any additional information, however, it will be noted that missing data for BMI could overestimate vaccine effectiveness estimates when adjusted for since individuals with a recorded BMI might be those that have better healthcare access/are interested in bettering their health outcomes than those with missing BMI.

Patient or user group involvement

Findings on the role of markers of health-seeking behaviour / healthcare access will be discussed for their interpretation with a Public Involvement Panel which will be organised through the NIHR Health Protection Research Unit (HPRU) in Vaccines and Immunisation. The HPRU runs regular PPIE events (minimum of 3x/year) and attendance is targeted each time to target audiences appropriate to the study, with regular attendance varying around 15-20 attendees. Recruitment channels for this event will include the HPRU mailing list and local Healthwatch organisations.

Plans for disseminating & communicating

The results of this study will be disseminated in peer-reviewed journals and at scientific conferences. The study is being conducted in partnership with the UK Health Security Agency and results will be used to inform interpretation and design of observational studies of influenza and COVID-19 vaccine effectiveness for public health planning.

Conflict of interest statement

SG, NA, JLW and HIM are funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Vaccines and Immunisation (grant reference NIHR200929), a partnership between UK Health Security Agency and London School of Hygiene & Tropical Medicine. UKHSA has provided vaccine manufacturers with post-marketing surveillance reports which the companies are required to submit to the UK Licensing Authority in compliance with their Risk Management Strategy, and a cost recovery charge is made for these reports.

SG is also a part-time salaried employee of Evidera, which a business unit within Pharmaceutical Product Development, LLC, (PPD), which is owned by Thermo Fisher Scientific. None of these companies are involved in the direct development of vaccinations, however, Evidera consults pharmaceutical or biotechnology companies that develop vaccinations. PPD are a clinical research organisation who are contracted by pharmaceutical or biotechnology companies to conduct vaccination trials on their behalf. Thermo Fisher Scientific produces several devices and laboratory products that aid in vaccination research.

DN confirms that she has no relevant COIs.

Limitations of study design

This study has several limitations.

Firstly, the identification of potential markers of health-seeking behaviour and healthcare access relies on accurate recording of these variables in primary and secondary care data. It also relies on almost consistent recording of these variables between practices and hospitals. If there is inaccurate recording of these variables and variability between practices then adjusting for these markers might instead be accounting for other sociodemographic factors. The study attempts to account for geographical variation in recording by adjusting for NHS region. Even if these proxies are well recorded and consistently recorded across practices, these markers are just proxies and therefore relies on the extent to which they can quantify and adjust for the underlying phenomenon of confounding by health-seeking behaviour and health care access. It is likely that these markers will not be able to capture all confounding from health-seeking behaviour and healthcare access and therefore adjusting for these in combination might still underestimate confounding from this. However, it will give an indication of the extent to which these can be used to adjust for confounding and the negative control outcome will indicate the likely extent of residual confounding to estimate how much confounding remains unaddressed.

Secondly, the study uses markers of health-seeking behaviour and healthcare access from a pre-pandemic period because it was assumed that behaviour and access would change during the pandemic. The limitation here is that it is assumed that individuals underlying characteristics driving the behaviour and access remained unchanged despite the pandemic. Future research may be useful to update these analyses with markers of health-seeking behaviour and healthcare access post-pandemic, once these are stable.

Another limitation is that COVID-19 outcomes used in the negative control outcome analysis are identified in a different study period to the COVID-19 vaccine effectiveness outcomes. The association between health-seeking behaviour / healthcare access and COVID-19 outcomes might change between these periods, which would mean that the quantified residual confounding in objective 3 might not be directly applicable to objective 1 and 2. However, these associations will be compared between these two periods, and will aid in the interpretation of these results.

References

1. Su W, Wang H, Sun C, Li N, Guo X, Song Q, et al. The Association Between Previous Influenza Vaccination and COVID-19 Infection Risk and Severity: A Systematic Review and Meta-analysis. Am J Prev Med. 2022;63(1):121-30.
2. Doll MK, Pettigrew SM, Ma J, Verma A. Effects of Confounding Bias in COVID-19 and Influenza Vaccine Effectiveness Test-Negative Designs Due to Correlated Influenza and

COVID-19 Vaccination Behaviors. Clin Infect Dis. 2022.
3. Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. Covid-19 Vaccine Effectiveness against the Omicron (B.1.1.529) Variant. N Engl J Med. 2022.
4. Andrews N, Stowe J, Kirsebom F, Toffa S, Sachdeva R, Gower C, et al. Effectiveness of COVID-19 booster vaccines against COVID-19-related symptoms, hospitalization and death in England. Nat Med. 2022;28(4):831-7.
5. Andrews N, Tessier E, Stowe J, Gower C, Kirsebom F, Simmons R, et al. Duration of Protection against Mild and Severe Disease by Covid-19 Vaccines. N Engl J Med. 2022;386(4):340-50.
6. Lopez Bernal J, Andrews N, Gower C, Gallagher E, Simmons R, Thelwall S, et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. N Engl J Med. 2021;385(7):585-94.
7. Lopez Bernal J, Andrews N, Gower C, Robertson C, Stowe J, Tessier E, et al. Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in England: test negative case-control study. BMJ. 2021;373:n1088.
8. Nelson JC, Jackson ML, Weiss NS, Jackson LA. New strategies are needed to improve the accuracy of influenza vaccine effectiveness estimates among seniors. J Clin Epidemiol. 2009;62(7):687-94.
9. Robertson E, Reeve KS, Niedzwiedz CL, Moore J, Blake M, Green M, et al. Predictors of COVID-19 vaccine hesitancy in the UK household longitudinal study. Brain Behav Immun. 2021;94:41-50.
10. Sullivan SG, Tchetgen Tchetgen EJ, Cowling BJ. Theoretical Basis of the Test-Negative Study Design for Assessment of Influenza Vaccine Effectiveness. Am J Epidemiol. 2016;184(5):345-53.
11. Schmid P, Rauber D, Betsch C, Lidolt G, Denker ML. Barriers of Influenza Vaccination Intention and Behavior - A Systematic Review of Influenza Vaccine Hesitancy, 2005 - 2016. PLoS One. 2017;12(1):e0170550.
12. Cowling TE, Ramzan F, Ladbrooke T, Millington H, Majeed A, Gnani S. Referral outcomes of attendances at general practitioner led urgent care centres in London, England: retrospective analysis of hospital administrative data. Emerg Med J. 2016;33(3):200-7.
13. Moynihan R, Sanders S, Michaleff ZA, Scott AM, Clark J, To EJ, et al. Impact of COVID-19 pandemic on utilisation of healthcare services: a systematic review. BMJ Open. 2021;11(3):e045343.
14. Lu Y, Wang Y, Shen C, Luo J, Yu W. Decreased Incidence of Influenza During the COVID-19 Pandemic. Int J Gen Med. 2022;15:2957-62.
15. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383-8.
16. Statistics OfN. Estimates of the population for the UK, Englan and Wales, Scotland and Northern Ireland 2021 [Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimate
Rothman KJ, Greenland S. Planning Study Size Based on Precision Rather Than Power. Epidemiology. 2018;29(5):599-603.
18. Whitaker HJ, Tsang RSM, Byford R, Andrews NJ, Sherlock J, Sebastian Pillai P, et al. Pfizer-BioNTech and Oxford AstraZeneca COVID-19 vaccine effectiveness and immune response amongst individuals in clinical risk groups. J Infect. 2022;84(5):675-83.
19. Robinson L, and Jewell, Nicholas P. Some Surprising Results about Covariate Adjustment in Logistic Regression Models. International Statistical Review. 1991;Vol. 59, No. 2 (Aug., 1991), pp. 227-240.
20. Millett ER, Quint JK, Smeeth L, Daniel RM, Thomas SL. Incidence of community-acquired lower respiratory tract infections and pneumonia among older adults in the United Kingdom: a population-based study. PLoS One. 2013;8(9):e75131.
21. Lewnard JA, Patel MM, Jewell NP, Verani JR, Kobayashi M, Tenforde MW, et al. Theoretical Framework for Retrospective Studies of the Effectiveness of SARS-CoV-2 Vaccines. Epidemiology. 2021;32(4):508-17.
22. J. D. Clinical code list - CPRD Aurum - COVID-19. London School of Hygiene and Tropical Medicine. 2022.
23. Chan RWY, Liu S, Cheung JY, Tsun JGS, Chan KC, Chan KYY, et al. The Mucosal and

Serological Immune Responses to the Novel Coronavirus (SARS-CoV-2) Vaccines. Front Immunol. 2021;12:744887.
24. Davidson J W-GC, Mcdonald H, Smeeth L and Banerjee A. . Clinical code list - CPRD Aurum - influenza vaccine. London School of Hygiene and Tropical Medicine.
25. Davidson J W-GC, Mcdonald H, Smeeth L and Banerjee A. Therapy codelist - CPRD Aurum - Influenza vaccine. London School of Hygiene and Tropical Medicine.
26. Krammer F. The human antibody response to influenza A virus infection and vaccination. Nat Rev Immunol. 2019;19(6):383-97.
27. J. D. Clinical codelist - CPRD Aurum - COVID-19. London School of Hygiene and Tropical Medicine.
28. C. DJaW-G. Clinical codelist - CPRD Aurum - acute respiratory infection and influenza/influenza-like-illness codes. London School of Hygiene and Tropical Medicine.
29. C. DJaW-G. Clinical codelist - HES - acute respiratry infection and influenza/influenza-like-illness codes. London School of Hygiene and Tropical Medicine.
30. Service NH. NHS Health Check 2019 [Available from: https://www.nhs.uk/conditions/nhs-health-check/
31. (NICE) TNIfHaCE. NICE 'do not do' recommendations [Available from: https://www.nice.org.uk/media/default/sharedlearning/716_716donotdobookletfinal.pdf.
32. QualityWatch. Focus on prevantable admissions: Trends in emergency admissions for ambulatory care sensitive conditions, 2001 to 2013 [Available from: https://www.health.org.uk/sites/default/files/QualityWatch_FocusOnPreventableAdmissions.pdf.
33. Carey IM, Hosking FJ, Harris T, DeWilde S, Beighton C, Cook DG. An evaluation of the effectiveness of annual health checks and quality of health care for adults with intellectual disability: an observational study using a primary care database. Health Services and Delivery Research. Southampton (UK)2017.
34. Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, vanStaa T, Grundy E, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. J Public Health (Oxf). 2014;36(4):684-92.
35. Davidson JA, Banerjee A, Smeeth L, McDonald HI, Grint D, Herrett E, et al. Risk of acute respiratory infection and acute cardiovascular events following acute respiratory infection among adults with increased cardiovascular risk in England between 2008 and 2018: a retrospective, population-based cohort study. Lancet Digit Health. 2021;3(12):e773-e83.
36. OpenCodelists. Chronic heart disease codes 2021 [Available from: https://www.opencodelists.org/codelist/primis-covid19-vacc-uptake/chd_cov/v1.2.1/.
37. Davidson JaS, H. Clinical codelist - CPRD Aurum - severe mental illness. London School of

Hygiene & Tropical Medicine, London, United Kingdom. 2022.
38. Davidson JaW-G, C Clinical codelist - CPRD Aurum - learning and intellectual disability. London School of Hygiene & Tropical Medicine, London, United Kingdom. 2022.

Appendices

figure-1-covid19-cases-lockdowns-variants-and-key-dates.pdf

figure-2-time-periods.pdf

supplementary-material-code-lists_0.pdf

supplementary-materials-theory-of-planned-behaviour-model.pdf

table-1-how-markers-of-healthseeking-behaviour-and-healthcare-access-are-administered-in-uk-clir

table-2-vaccine-effectiveness-sample-size-calculation.pdf

table-3-vaccine-effectiveness-estimates-percentage-decrease_0.pdf

Grant ID
The study is funded by the National Institute for Health Research (NIHR) Health Protection Research Unit (HPRU) in Vaccines and Immunisation (grant reference NIHR200929).

Appendix D. Supplementary Materials Paper Two

Supplementary information

Tables

Supplementary Table 1. Current available literature and markers derived from each

Author and year	Study type	Markers identified	Included/excluded	Reason for exclusion
Izurieta et al, 2020[1]	Observational cohort study of relative influenza vaccine effectiveness using US claims data	Pneumococcal vaccination	Included	
		Annual wellness visit	Included as NHS health checks	
		Bone mass measurements	Included as bone density scan	
		Cardiovascular disease screen tests	Excluded	Failed criteria 3
		Colorectal cancer screen	Included as bowel cancer screen	
		Diabetes screen	Included as NHS health checks	
		Initial Preventive Physical Examination	Included as NHS health checks	
		Prostate Cancer Screen	Included as PSA test	
		Screen Mammography	Included as breast cancer screen	
		Screen Pap test	Included as cervical cancer screen	
		Screen pelvic examination	Excluded	Failed criteria 3
		Depression screen	Excluded	Failed criteria 3
		Other preventative services	Excluded	Unclear what this entails
Izurieta et al, 2021[2]	Observational cohort study of influenza vaccine effectiveness using US claims data	Annual Wellness Visits	Included as NHS health checks	
		Counseling & Health Risk Assessment	Excluded	Failed criteria 3
		Pneumococcal Vaccination	Included	
		Tetanus-containing vaccination	Included	Failed criteria 3
		Shingrix vaccination	Excluded	Failed criteria 1
Zhang et al, 2017[3]	Observational cohort study of influenza vaccine effectiveness using US claims data	Outpatient visits	Included as GP visits	
		Hospital visits	Excluded	Failed criteria 3
		Colonoscopies	Included as bowel cancer screen	
		Fecal occult blood tests	Included as bowel cancer screen	

Abbreviations: DNA: did not attend; GP: general practice; NHS: national health service; PSA: prostate-specific antigen.

Supplementary Table 2. How each of the markers meet each of the criteria

Marker	Criteria 1 Should have been currently or recently available in national clinical practice to all individuals (overall or by sex) at cohort entry.	Criteria 2 Should be routinely recorded in the available data sources.	Criteria 3 Should not be primarily dependent on underlying health needs.
AAA screen	~ AAA screening was introduced in 2013 in the UK[4] so for men entering the cohort at age 71 years and above they might not have ever been offered a screen (since this is offered to all men when they turn 65 years).	✓	✓ Universally available except for men that have been treated for an AAA previously. The incidence of AAA in those under 65 years is very low[5].
Breast cancer screen	✓	✓	✓

Marker	Criteria 1 Should have been currently or recently available in national clinical practice to all individuals (overall or by sex) at cohort entry.	Criteria 2 Should be routinely recorded in the available data sources.	Criteria 3 Should not be primarily dependent on underlying health needs.
Cervical cancer screen	✓	✓	✓
Bowel cancer screen	✓	✓	✓
NHS health checks	✓	✓	~ Universally available to those without pre-existing high-risk conditions[6].
Influenza vaccination	✓	✓	✓
Pneumococcal vaccination	✓	✓	✓
PSA testing	✓	✓	~ Men can request a PSA test with no prior diagnosis of prostate cancer who are asymptomatic, men presenting with symptoms, and men who have had a previous high PSA level and are being monitored[7].
Bone density scans	✓	✓	~ Can be requested for individuals over 50 years with a risk of developing osteoporosis or for those with other risk factors such as smoking or broken bone[8].
GP practice visits	✓	✓	~ Occurs for those with symptoms.
Low-value procedures	✓	✓	~ Requested for those with symptoms.
Low-value prescriptions	✓	~ Although these can be prescribed in primary care, individuals can also buy some low values prescriptions over the counter and therefore we might see under recording of this. However, those that access primary care for something that can be bought over the counter likely have very active access to healthcare.	✓
Hospital visit for an ACS condition	✓	✓	~ Occurs for those with symptoms.
DNA primary care visit	✓	✓	~ Occurs for those with symptoms.
Blood pressure measurements	✓	~	~ Universally available, but disproportionately requested for those with underlying health conditions.

Abbreviations: DNA: did not attend; GP: general practice; NHS: national health service; PSA: prostate-specific antigen.

Supplementary Table 3. Influenza Vaccination Algorithm

Combination of codes on same day	Total number of vaccination events	Decision	Rationale
Given and neutral	725432	Record as valid vaccination.	
Given and absent	1149 (of these only 924 are the first vaccination dose)	Do not record as valid vaccination.	The prevalence of this marker may be underestimated very slightly, however, in this instance better to be more specific than sensitive when it comes to confounders[9].
Given and adverse	14	Record as valid vaccination.	Likely that this patient received the vaccination, but then had an adverse event on the same day.
Given and product	395881	Record as valid vaccination.	
Given and given with lag	4700	Record as valid vaccination.	
Neutral and absent	1178	Do not record as valid vaccination.	
Neutral and adverse	9	Record as valid vaccination.	
Neutral and product	295746	Record as valid vaccination.	
Neutral and given lag	7748	Record as valid vaccination.	
Absent and adverse	27	Do not record as valid vaccination.	Likely that these patients are reporting a previous adverse event as the reason for not wanting to get vaccinated.
Absent and product	218 (of these only 194 are first vaccination dose)	Do not record as valid vaccination.	The prevalence of this marker will be underestimated very slightly, however, in this instance better to be more specific than sensitive when it comes to confounders[9].
Absent and given with lag	570	Record as valid vaccination.	This most likely reflects that the reason a vaccine wasn't given on the event date is that the patient had already had it elsewhere. We can be reasonably confident that the patient was vaccinated, but we don't know the exact date.
Adverse and product	1	Record as valid vaccination.	Likely that this patient received the vaccination, but then had an adverse event on the same day.
Adverse and given with lag	0	Ignore – no events.	
Given with lag and product	784	Record as valid vaccination.	

Note: since influenza vaccinations can be identified using both medcodes and prodcodes and since medcodes do not always insinuate presence of a vaccination, an algorithm was developed for combinations of codes that occurred on the same day. Medcodes were separated into those that were clearly given ("given" or "administered"), given with delay ("given" or "administered" but evidence this occurred in another setting previously), neutral (vaccination mentioned but no "given" or "administered") and absent (vaccination "refused" or "not consented). Then we looked at vaccination events (using both medcodes and prodcodes) that were recorded on the same date and categorised these according to the above framework.

Supplementary Table 4. National estimates

Marker	Estimate (%, unless otherwise specified)	Year	Numerator	Denominator
AAA screen[10]	76.1	2019/20	Number of men eligible for the initial screen who have had a conclusive scan result within the screening year plus an additional 3 months (in the event of non attendance and cancellations at the end of the year this allows men to be reinvited and screened).	Number of eligible men in their 65th year to whom the screening programme propose that a screening encounter during the reporting period should be offered.
Breast cancer screen[11]	71.1	2019/20	The number of persons registered to the practice who were screened adequately in the previous 36 months	The number of eligible persons on last day of the review period
Cervical cancer screen[12]	76.2	2019	The number of women in the resident population eligible for cervical screening aged 50 to 64 years at end of period reported who were screened adequately within the previous 5.5 years.	The number of women in the resident population eligible for cervical screening aged 50 to 64 years
Bowel cancer screen[13]	60.5	2019	Adequately screened (numerator) is the number of eligible men and women who have had an adequate gFOBT screening result recorded in the past 30 months.	Eligible population (denominator) is the number of men and women aged 60 to 74 years resident in the area (determined by postcode of residence) who are eligible for bowel cancer screening at a given point in time, excluding those whose recall has been ceased for clinical reasons (e.g. no functioning colon) or if they opt out of the programme.

NHS health checks[14]	2.0	2019/20 Q1	Number of people aged 40-74 eligible for an NHS Health Check who were recorded as receiving an NHS Health Check in the current quarter	Number of people aged 40-74 eligible for an NHS Health Check in the financial year.
Influenza vaccination[15]	72.4	2019/20	The number of adults aged 65 and over, who received the flu vaccination between 1st September to the end of February as recorded in the GP record.	Number of adults aged 65 years and older
Pneumococcal vaccination[16]	69	2019/20	These data describe pneumococcal polysaccharide vaccine (PPV) uptake for the survey year, for those aged 65 years and over.	Those aged 65 years and over
PSA test[17]	52.95	2002-2011	Number of men with at least one PSA test identified over a 10 year period	Number of men
Bone density scan[18]	0.3 to 16.2 per 1000 weighted population	2013/14	Number of bone scans.	Total population.
GP practice visits[19]	39127 per 100 000 patient-months	2019	Number of primary care consultations in CPRD Aurum	Number of patient months
GP practice visits[19]	26919 consultations per 100 000 patient-months	2020	Number of primary care consultations in CPRD Aurum	Number of patient months
DNA primary care visit[19]	Number: 23578484	Dec-19	Number of primary care appointments that were DNA	
Low value procedures[20]	0.02-0.2	2017/18	Number of category 1 interventions.	Total English population, age and sex-standardised.
Low-value prescriptions (glucosamine)	21,961 items	2019/20	Total quantity prescribed in primary care	
Hospital visit for an ACS condition[21]	0.1	2018/19	Number of unplanned hospitalisations for chronic ambulatory care sensitive conditions	Total unplanned hospitalisations
Blood pressure test[22]	84.6	2012-13	Number of individuals aged ≥50 years with a blood pressure check in the last year	Number of individuals aged ≥50 years

Abbreviations: DNA: did not attend; GP: general practice; NHS: national health service; PSA: prostate-specific antigen.

Supplementary Table 5. Study population.

Group	Category	All N= 1,991,284
Age continuous, mean (SD)		75.9 (7.4)
Age category, N (%)	65-69	448,063 (22.5%)
	70-74	557,599 (28.0%)
	75-79	401,290 (20.2%)
	80-84	295,492 (14.8%)
	85-89	181,835 (9.1%)
	90-95	80,943 (4.1%)
	95+	26,062 (1.3%)
Sex, N (%)	Female	1,075,723 (54.0%)
Ethnicity, N (%)	Asian	67,961 (3.4%)
	Black	36,912 (1.9%)
	Mixed	10,072 (0.5%)
	Other	17,711 (0.9%)
	White	1,759,754 (88.4%)
	Missing	98,874 (5.0%)
Region, N (%)	East Midlands	42,106 (2.1%)
	East of England	95,413 (4.8%)
	London	263,825 (13.2%)
	North East	67,278 (3.4%)

	North West	381,592 (19.2%)
	South East	438,407 (22.0%)
	South West	270,484 (13.6%)
	West Midlands	357,363 (17.9%)
	Yorkshire and The Humber	74,805 (3.8%)
	Unknown	11 (0.0%)

Abbreviations: N: number; SD: standard deviation.

Supplementary Table 6. Prevalence of markers using different definitions

Variable	All	Male	Female
N	1,991,284	915,561	1,075,723
AAA screen	231,088 (11.6%)	227,844 (24.9%)	3,244 (0.3%)
AAA screen broad*	238,186 (12.0%)	233,574 (25.5%)	4,612 (0.4%)
Breast cancer screen restrictive†	253,610 (12.7%)	227 (0.0%)	253,383 (23.6%)
Breast cancer screen	346,116 (17.4%)	517 (0.1%)	345,599 (32.1%)
Breast cancer screen broad* restrictive†	484,402 (24.3%)	573 (0.1%)	483,829 (45.0%)
Breast cancer screen broad*	686,521 (34.5%)	1,517 (0.2%)	685,004 (63.7%)
Cervical cancer screen restrictive†	256,565 (12.9%)	40 (0.0%)	256,525 (23.8%)
Cervical cancer screen	397,303 (20.0%)	153 (0.0%)	397,150 (36.9%)
Cervical cancer screen broad* restrictive†	373,369 (18.8%)	93 (0.0%)	373,276 (34.7%)
Cervical cancer screen broad*	588,288 (29.5%)	361 (0.0%)	587,927 (54.7%)
NHS health checks restrictive†	338,441 (17.0%)	143,325 (15.7%)	195,116 (18.1%)
NHS health checks	372,244 (18.7%)	157,484 (17.2%)	214,760 (20.0%)
Bowel screen restrictive†	1,151,943 (57.8%)	553,109 (60.4%)	598,834 (55.7%)
Bowel screen	1,439,412 (72.3%)	687,712 (75.1%)	751,700 (69.9%)

Abbreviations: AAA: abdominal aortic aneurysm; DNA: did not attend; GP: general practice; PSA: prostate-specific antigen.
*Broad code list: code lists were less specific e.g., screening markers could mention the relevant test, but without requiring "screen" in the code.
†Restrictive lookback: for markers with an upper age of eligibility (e.g., cancer screening and NHS health checks) the lookback period stopped at the age of upper eligibility.

Supplementary Table 7. Prevalence of markers by age.

Marker	Age category	Count	Prevalence
AAA screen broad*	65-69	121,345	55.0
AAA screen broad*	70-74	87,776	32.7
AAA screen broad*	75-79	13,957	7.4
AAA screen broad*	80-84	6,930	5.3
AAA screen broad*	85-89	2,748	3.7
AAA screen broad*	90-95	709	2.5
AAA screen broad*	95+	109	1.6
AAA screen	65-69	119,707	54.3
AAA screen	70-74	85,529	31.8
AAA screen	75-79	13,075	7.0
AAA screen	80-84	6,360	4.9
AAA screen	85-89	2,445	3.3
AAA screen	90-95	634	2.3
AAA screen	95+	94	1.4
Breast cancer screen broad*	65-69	165,587	72.8
Breast cancer screen broad*	70-74	214,147	74.1
Breast cancer screen broad*	75-79	160,055	74.9
Breast cancer screen broad*	80-84	104,683	63.5
Breast cancer screen broad*	85-89	30,543	28.2
Breast cancer screen broad*	90-95	8,351	15.8
Breast cancer screen broad*	95+	1,638	8.5
Breast cancer screen broad* restrictive†	65-69	165,575	72.8
Breast cancer screen broad* restrictive†	70-74	163,148	56.4
Breast cancer screen broad* restrictive†	75-79	81,247	38
Breast cancer screen broad* restrictive†	80-84	62,039	37.6
Breast cancer screen broad* restrictive†	85-89	10,162	9.4
Breast cancer screen broad* restrictive†	90-95	1,500	2.8
Breast cancer screen broad* restrictive†	95+	158	0.8
Breast cancer screen	65-69	86,335	37.9
Breast cancer screen	70-74	112,561	38.9
Breast cancer screen	75-79	81,241	38
Breast cancer screen	80-84	48,422	29.4
Breast cancer screen	85-89	13,182	12.2
Breast cancer screen	90-95	3,300	6.2
Breast cancer screen	95+	558	2.9
Breast cancer screen restrictive†	65-69	86,327	37.9
Breast cancer screen restrictive†	70-74	87,674	30.3
Breast cancer screen restrictive†	75-79	44,831	21
Breast cancer screen restrictive†	80-84	29,008	17.6
Breast cancer screen restrictive†	85-89	4,817	4.4
Breast cancer screen restrictive†	90-95	679	1.3
Breast cancer screen restrictive†	95+	47	0.2
Cervical cancer screen broad*	65-69	127,792	56.2

Cervical cancer screen broad*	70-74	160,051	55.4
Cervical cancer screen broad*	75-79	122,449	57.3
Cervical cancer screen broad*	80-84	91,696	55.6
Cervical cancer screen broad*	85-89	57,058	52.7
Cervical cancer screen broad*	90-95	24,576	46.5
Cervical cancer screen broad*	95+	4,305	22.3
Cervical cancer screen broad* restrictive†	65-69	95,951	42.2
Cervical cancer screen broad* restrictive†	70-74	114,760	39.7
Cervical cancer screen broad* restrictive†	75-79	78,703	36.8
Cervical cancer screen broad* restrictive†	80-84	49,255	29.9
Cervical cancer screen broad* restrictive†	85-89	25,412	23.5
Cervical cancer screen broad* restrictive†	90-95	8,537	16.1
Cervical cancer screen broad* restrictive†	95+	658	3.4
Cervical cancer screen	65-69	76,059	33.4
Cervical cancer screen	70-74	112,676	39
Cervical cancer screen	75-79	85,484	40
Cervical cancer screen	80-84	63,964	38.8
Cervical cancer screen	85-89	39,581	36.6
Cervical cancer screen	90-95	16,649	31.5
Cervical cancer screen	95+	2,737	14.2
Cervical cancer screen restrictive†	65-69	59,641	26.2
Cervical cancer screen restrictive†	70-74	81,978	28.4
Cervical cancer screen restrictive†	75-79	55,434	25.9
Cervical cancer screen restrictive†	80-84	35,276	21.4
Cervical cancer screen restrictive†	85-89	17,966	16.6
Cervical cancer screen restrictive†	90-95	5,846	11.1
Cervical cancer screen restrictive†	95+	384	2.0
Bowel cancer screen	65-69	423,248	94.5
Bowel cancer screen	70-74	521,588	93.5
Bowel cancer screen	75-79	354,378	88.3
Bowel cancer screen	80-84	114,222	38.7
Bowel cancer screen	85-89	17,514	9.6
Bowel cancer screen	90-95	6,743	8.3
Bowel cancer screen	95+	1,719	6.6
Bowel cancer screen restrictive†	65-69	423,248	94.5
Bowel cancer screen restrictive†	70-74	513,796	92.1
Bowel cancer screen restrictive†	75-79	170,462	42.5
Bowel cancer screen restrictive†	80-84	42,108	14.3
Bowel cancer screen restrictive†	85-89	1,903	1.0
Bowel cancer screen restrictive†	90-95	391	0.5
Bowel cancer screen restrictive†	95+	35	0.1
NHS health checks	65-69	125,225	27.9
NHS health checks	70-74	135,076	24.2
NHS health checks	75-79	85,472	21.3
NHS health checks	80-84	22,626	7.7
NHS health checks	85-89	2,781	1.5

NHS health checks	90-95	844	1.0
NHS health checks	95+	220	0.8
NHS health checks restrictive†	65-69	125,225	27.9
NHS health checks restrictive†	70-74	134,622	24.1
NHS health checks restrictive†	75-79	69,649	17.4
NHS health checks restrictive†	80-84	8,933	3.0
NHS health checks restrictive†	85-89	6	<0.1
NHS health checks restrictive†	90-95	6	<0.1
NHS health checks restrictive†	95+	0	0
Influenza vaccination	65-69	281,993	62.9
Influenza vaccination	70-74	400,143	71.8
Influenza vaccination	75-79	310,417	77.4
Influenza vaccination	80-84	237,347	80.3
Influenza vaccination	85-89	146,427	80.5
Influenza vaccination	90-95	64,238	79.4
Influenza vaccination	95+	19,826	76.1
Pneumococcal vaccination	65-69	171,814	38.3
Pneumococcal vaccination	70-74	329,895	59.2
Pneumococcal vaccination	75-79	287,843	71.7
Pneumococcal vaccination	80-84	229,129	77.5
Pneumococcal vaccination	85-89	142,058	78.1
Pneumococcal vaccination	90-95	62,385	77.1
Pneumococcal vaccination	95+	19,235	73.8
PSA testing	65-69	71,613	32.5
PSA testing	70-74	101,779	37.9
PSA testing	75-79	79,356	42.3
PSA testing	80-84	56,678	43.4
PSA testing	85-89	30,517	41.5
PSA testing	90-95	10,108	36
PSA testing	95+	1,833	27.2
Bone density scans	65-69	19,797	4.4
Bone density scans	70-74	29,459	5.3
Bone density scans	75-79	23,462	5.8
Bone density scans	80-84	16,393	5.5
Bone density scans	85-89	8,844	4.9
Bone density scans	90-95	2,544	3.1
Bone density scans	95+	393	1.5
GP visits	65-69	396,941	88.6
GP visits	70-74	511,875	91.8
GP visits	75-79	377,556	94.1
GP visits	80-84	281,987	95.4
GP visits	85-89	174,447	95.9
GP visits	90-95	77,495	95.7
GP visits	95+	24,522	94.1
DNA primary care	65-69	118,146	26.4
DNA primary care	70-74	151,958	27.3

DNA primary care	75-79	124,316	31
DNA primary care	80-84	102,601	34.7
DNA primary care	85-89	67,750	37.3
DNA primary care	90-95	29,068	35.9
DNA primary care	95+	8,057	30.9
Low value procedures	65-69	59,141	13.2
Low value procedures	70-74	87,714	15.7
Low value procedures	75-79	77,838	19.4
Low value procedures	80-84	66,686	22.6
Low value procedures	85-89	43,871	24.1
Low value procedures	90-95	18,545	22.9
Low value procedures	95+	5,086	19.5
Low value prescriptions	65-69	29	<0.1
Low value prescriptions	70-74	58	<0.1
Low value prescriptions	75-79	57	<0.1
Low value prescriptions	80-84	40	<0.1
Low value prescriptions	85-89	22	<0.1
Low value prescriptions	90-95	<5	<0.1
Low value prescriptions	95+	<5	<0.1
Hospital visit ACS condition	65-69	23,948	5.3
Hospital visit ACS condition	70-74	36,985	6.6
Hospital visit ACS condition	75-79	36,772	9.2
Hospital visit ACS condition	80-84	37,255	12.6
Hospital visit ACS condition	85-89	30,822	17
Hospital visit ACS condition	90-95	17,519	21.6
Hospital visit ACS condition	95+	6,835	26.2
Blood pressure measurement	65-69	288,444	64.4
Blood pressure measurement	70-74	394,102	70.7
Blood pressure measurement	75-79	307,207	76.6
Blood pressure measurement	80-84	240,202	81.3
Blood pressure measurement	85-89	151,973	83.6
Blood pressure measurement	90-95	67,185	83
Blood pressure measurement	95+	20,893	80.2

Abbreviations: AAA: abdominal aortic aneurysm; ACS: ambulatory care sensitive; DNA: did not attend; GP: general practice; NHS: National Health Service; PSA: prostate-specific antigen.

*Broad code list: code lists were less specific e.g., screening markers could mention the relevant test, but without requiring "screen" in the code.

†Restrictive lookback: for markers with an upper age of eligibility (e.g., cancer screening and NHS health checks) the lookback period stopped at the age of upper eligibility.

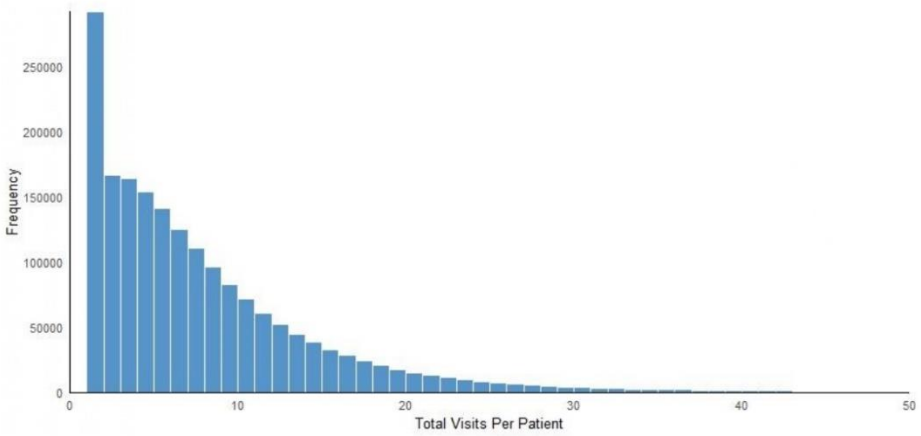
Supplementary Table 8. Phi coefficients

	NHS health checks	Bone density scan	Bowel screen	Primary DNA	Blood pressure	Pneumococcal vaccine	Influenza vaccine	ACS conditions	Low-value procedures	GP visits
NHS health checks	1									
Bone density scan	0.02	1								
Bowel screen	0.21	0.02	1							
Primary DNA	-0.06	0.03	-0.05	1						
Blood pressure	-0.14	0.03	-0.08	0.16	1					
Pneumococcal vaccine	-0.06	0.03	-0.12	0.07	0.18	1				
Influenza vaccine	-0.01	0.04	-0.03	0.06	0.23	0.41	1			
ACS conditions	-0.08	0.02	-0.12	0.11	0.13	0.07	0.05	1		
Low-value procedures	-0.05	0.04	-0.06	0.12	0.11	0.07	0.06	0.15	1	
GP visits	0.01	0.05	-0.02	0.14	0.42	0.21	0.33	0.07	0.11	1

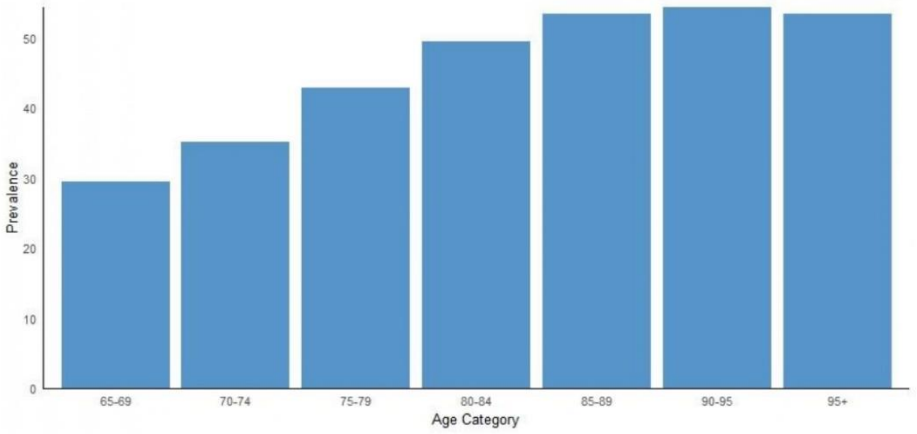
Abbreviations: ACS: ambulatory care sensitive; DNA: did not attend; GP: general practice; NHS: National Health Service.
Note: the correlations are calculated using phi coefficient for binary variables.

Supplementary Table 9. Phi coefficients males

	AAA screen	NHS health checks	PSA test	Bone density scan	Bowel screen	Primary DNA	Blood pressure	Pneumococcal vaccine	Influenza vaccine	ACS conditions	Low-value procedures	GP visits
AAA screen	1											
NHS health checks	0.11	1										
PSA test	0.00	0.03	1									
Bone density scan	-0.01	0.00	0.06	1								
Bowel screen	0.24	0.17	0.01	-0.01	1							
Primary DNA	-0.05	-0.05	0.06	0.03	-0.05	1						
Blood pressure	-0.04	-0.14	0.12	0.03	-0.06	0.17	1					
Pneumococcal vaccine	-0.12	-0.07	0.1	0.03	-0.12	0.07	0.22	1				



Supplementary Figure 3: Prevalence of over the median number of GP visits, stratified by age category. As majority of individuals in the study had at least one GP visit, we conducted a *post-hoc analysis* that identified the prevalence of over the median number of visits per year (7), stratified by age category.



References

1. Izurieta HS, Chillarige Y, Kelman J, et al. Relative Effectiveness of Influenza Vaccines Among the United States Elderly, 2018-2019. *J Infect Dis.* 2020;222(2):278-287.
2. Izurieta HS, Lu M, Kelman J, et al. Comparative Effectiveness of Influenza Vaccines Among US Medicare Beneficiaries Ages 65 Years and Older During the 2019-2020 Season. *Clin Infect Dis.* 2021;73(11):e4251-e4259.
3. Zhang HT, McGrath LJ, Wyss R, Ellis AR, Sturmer T. Controlling confounding by frailty when estimating influenza vaccine effectiveness using predictors of dependency in activities of daily living. *Pharmacoepidemiol Drug Saf.* 2017;26(12):1500-1506.
4. Bath MF, Sidloff D, Saratzis A, Bown MJ, investigators UKAGS. Impact of abdominal aortic aneurysm screening on quality of life. *Br J Surg.* 2018;105(3):203-208.
5. Howard DP, Banerjee A, Fairhead JF, Handa A, Silver LE, Rothwell PM. Age-specific incidence, risk factors and outcome of acute abdominal aortic aneurysms in a defined population. *Br J Surg.* 2015;102(8):907-915.
6. UK Government. NHS Health Checks: applying All Our Health. https://www.gov.uk/government/publications/nhs-health-checks-applying-all-our-health/nhs-health-checks-applying-all-our-health. Published 2022. Accessed 15/09/2022, 2023.
7. Melia J, Moss S. PSA testing in the UK. *Wiley Online.* 2009;14(1):9-13.
8. NHS England. Bone density (DEXA scan). https://www.nhs.uk/conditions/dexa-scan/. Published 2022. Accessed 02/10/2023, 2023.
9. Wood ME, Chrysanthopoulou S, Nordeng HME, Lapane KL. The Impact of Nondifferential Exposure Misclassification on the Performance of Propensity Scores for Continuous and Binary Outcomes: A Simulation Study. *Med Care.* 2018;56(8):e46-e53.
10. UK Government. Fingertips, Public Health Data, Abdominal Aortic Aneurysm Screening Coverage. https://fingertips.phe.org.uk/search/92317#page/4/gid/1/pat/159/par/K02000001/ati/15/are/E92000001/iid/92317/age/94/sex/1/cat/-1/ctp/-1/yrr/1/cid/4/tbm/1 Published 2023. Accessed 14/02/2023, 2023.
11. UK Government. Breast screening coverage: aged 50 to 70 years old. https://fingertips.phe.org.uk/search/breast%20screening#page/4/gid/1/pat/159/par/K02000001/ati/15/are/E92000001/iid/91339/age/265/sex/4/cat/-1/ctp/-1/yrr/1/cid/4/tbm/1. Published 2022. Accessed 01/08/2023, 2023.
12. UK Government. Cancer screening coverage: cervical cancer (aged 50 to 64 years old). https://fingertips.phe.org.uk/search/cervical#page/4/gid/1/pat/159/par/K02000001/ati/15/are/E92000001/iid/93561/age/273/sex/2/cat/-1/ctp/-1/yrr/1/cid/4/tbm/1. Published 2023. Accessed 01/08/2023, 2023.
13. UK Government. Cancer screening coverage: bowel cancer. https://fingertips.phe.org.uk/search/screening#page/4/gid/1/pat/159/par/K02000001/ati/15/are/E92000001/iid/91720/age/280/sex/4/cat/-1/ctp/-1/yrr/1/cid/4/tbm/1. Published 2023. Accessed 01/08/2023, 2023.
14. UK Government. Percentage of NHS Health Checks received by the total eligible population in the quarter. https://fingertips.phe.org.uk/profile/nhs-health-check-detailed/data#page/6/gid/1938132726/pat/159/par/K02000001/ati/15/are/E92000001/iid/91041/age/219/sex/4/cat/-1/ctp/-1/yrr/1/cid/4/tbm/1. Published 2023. Accessed 01/08/2023, 2023.
15. UK Government. Population vaccination coverage: Flu (aged 65 and over). https://fingertips.phe.org.uk/search/92317#page/4/gid/1/pat/159/par/K02000001/ati/15/are/E92000001/iid/92317/age/94/sex/1/cat/-1/ctp/-1/yrr/1/cid/4/tbm/1. Published 2023. Accessed 31/07/2023, 2023.
16. UK Government. Fingertips, Public Health Data, Population Vaccination Coverage, PPV. https://fingertips.phe.org.uk/search/ppv#page/4/gid/1/pat/159/par/K02000001/ati/15/are/E920000

01/iid/30313/age/27/sex/4/cat/-1/ctp/-1/yrr/1/cid/4/tbm/1. Published 2023. Accessed 14/02/2023, 2023.
17. Young GJ, Harrison S, Turner EL, et al. Prostate-specific antigen (PSA) testing of men in UK general practice: a 10-year longitudinal cohort study. *BMJ Open.* 2017;7(10):e017729.
18. UK Government. Rate of dual-energy X-ray absorptiometry (DEXA) activity per weighted population by CCG. https://fingertips.phe.org.uk/documents/Atlas_2015_MuscSkel.pdf. Published 2015. Accessed 01/08/2023, 2023.
19. NHS Digital. Appointments in General Practice report. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/appointments-in-general-practice. Published 2022. Accessed 02/10/2023, 2023.
20. NHS England. Evidence-Based Interventions: Consultation Document. https://www.england.nhs.uk/wp-content/uploads/2018/06/04-b-pb-04-07-2018-ebi-consultation-document.pdf. Published 2018. Accessed 21/08/2023, 2023.
21. NHS Digital. Unplanned hospitalisation for chronic ambulatory care sensitive conditions. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-outcomes-framework/may-2020/domain-2-enhancing-quality-of-life-for-people-with-long-term-conditions-nof/2-3-i-unplanned-hospitalisation-for-chronic-ambulatory-care-sensitive-conditions. Published 2020. Accessed.
22. Shiue I. Cold homes are associated with poor biomarkers and less blood pressure check-up: English Longitudinal Study of Ageing, 2012-2013. *Environ Sci Pollut Res Int.* 2016;23(7):7055-7059.

Appendix E. Supplementary Materials Paper Three

Supplementary information

Tables

Supplementary Table 1. Overview of study design and population selection in all analyses

	COVID-19	Influenza	Negative control exposure
Index date/start of follow-up for all individuals	8 December 2020	1 September 2019	1 January 2020
End of follow-up	Earliest of death, transfer out of the practice, end of data availability (29 March 2021), date of first COVID-19 vaccination that was neither BNT162b2 or ChAdOx1 or date of second heterologous vaccination (Figure 1).	Earliest of death, transfer out of the practice or end of influenza season (defined as 29 February 2020; Figure 1).	Earliest of death, transfer out of the practice or start of COVID-19 vaccination availability (7 December 2020; Figure 1).
Population selection	**Inclusion criteria:** 1. Aged ≥66 years on 1 September 2019. 2. Registration start date one year before index date. 3. Acceptable flag. 4. Eligible for HES APC and ONS linkage. **Exclusion criteria:** 1. Registration end date or death before index date. 2. Indeterminate sex.		
Additional criteria	**Exclusion criteria:** 1. Individuals with a COVID-19 vaccination that occurred before 8 December 2020 (as these were likely to be trial participants with different risk).		

Outcomes	• SARS-CoV-2 infection (primary care visit, hospital visit or death with COVID-19 specific medcode or ICD-19 code). • Hospital visit or death with COVID-19 specific ICD-10 code. • Death with COVID-19 specific ICD-10 code. Both suspected and confirmed COVID-19 medcodes were used since we wanted a consistent definition with Negative control exposure (see Negative control exposure column).	• Acute respiratory infection or influenza/influenza-like-illness infection (ARI/ILI; primary care visit, hospital visit or death with ARI/ILI specific medcode or ICD-19 code). • Hospital visit or death with ARI/ILI specific ICD-10 code. • Death with ARI/ILI specific ICD-10 code.	• SARS-CoV-2 infection (primary care visit, hospital visit or death with COVID-19 medcode or ICD-10 specific code). • Hospital visit or death with COVID-19 specific ICD-10 code. • Death with COVID-19 specific ICD-10 code. Both suspected and confirmed COVID-19 medcodes were used since majority of the outcome period occurred before the availability of widespread free polymerase chain reaction testing in the UK.
Exposures	• 1 or 2 doses of BNT162b2 or ChAdOx1 from 8 December 2020 onwards. We only used prodcodes to identify these, since it was not possible to identify the brand using medcodes.	• 1 dose of any influenza vaccination in the 2019/2020 influenza season (1 September 2019-29 February 2020). Both prodcodes and medcodes were used to identify influenza vaccinations. An algorithm was developed for code that occurred on the same day – see Supplementary Table 2.	• A history of 1 dose of any timely influenza vaccination in the 2019/2020 season (1 September 2019-31 December 2019; binary, assessed at baseline). Both prodcodes and medcodes were used to identify influenza vaccinations. An algorithm was developed for code that occurred on the same day – see Supplementary Table 2.
Variables described at index date	• Age in years calculated as index date – date of birth. Day and month imputed as 01/07 for all individuals as only year of birth is recorded in CPRD. Categorised as 65-69, 70-74, 75-79, 80-84, 85-89, 90-95, 95+. • Sex (male, female) • Recent infection (COVID-19 VE analysis: COVID-19 infection in the last 3-months; Influenza VE analysis: influenza infection in the previous influenza season [1 April 2018 to 31 August 2019]) • IMD was identified from the ONS at the patient level, or if missing by the primary care practice. Categorised as from 1 [least deprived] to 5 [most deprived].		

	• Ethnicity was identified from primary care records as described by Mathur et al.(27). Briefly, the algorithm uses a modal approach with ties resolved by recency. If ethnicity could not be identified in primary care, then ethnicity from HES APC was used. Categorised as Asian, Black, Missing, Mixed, Other and White • 'Influenza at risk' conditions (immunosuppressed status and other – see below). Comorbidities in influenza 'at-risk' groups were identified according to Greenbook chapter 19[241] in the pre-index period using medcodes (unless otherwise specified) and the below specified lookback periods: Immunosuppressed status: • Organ recipient: any time prior to index. • Immunosuppression therapies: biologic within the one year prior to index; or corticosteroids >40mg prednisolone per day for more than 1 week or corticosteroids >20mg prednisolone per day for more than 14 day or methotrexate >25mg per week; azathioprine >3.0mg/kg/day; 6-mercaptopurine >1.5mg/kg/day or corticosteroid injections; other disease-modifying antirheumatic drugs or other immunosuppressant medications in the 3-months prior to index. • Other immunosuppression: any time prior to index. Other conditions: • Chronic liver disease: any time prior to index. • Chronic cardiac disease: any time prior to index. • Chronic respiratory disease: any time prior to index. • Asthma: 3-months prior to index. • Diabetes mellitus: any time prior to index. • Chronic kidney disease: dialysis or transplant any time prior to index; or latest chronic kidney disease code is stage 3-5 (and not 1-2); or latest serum creatinine test result value ≤60 mL/min/1.73m2. • Chronic neurological disease: any time prior to index. • Severe obesity: latest body mass index recording prior to index ≥40 kg/m^2. • Severe mental conditions: any time prior to index. • Severe learning disability: any time prior to index. Code lists from Davidson et al, 2021[225] were utilised. The code lists from can be found listed on London School of Hygiene and Tropical Medicine (LSHTM) data compass: https://datacompass.lshtm.ac.uk/id/eprint/2240/. Previous estimates have shown that use of medcodes alone give plausible prevalence estimates.[252]

Markers of health-seeking behaviour and look back period	Markers were previously identified in Graham et al.[236] All of these conditions were identified using medcodes, ICD-10, OPCS or prodcodes identified in CPRD Aurum or HES APC. The same operational definitions from Graham et al.[236] were utilised: : • Abdominal aortic aneurysm screening (sex specific): any time prior to index date. • Breast cancer screening (sex specific): the last 4 years that they were age-eligible for screening prior to index date. • Cervical cancer screening (sex specific): the last 6 years that they were age-eligible for screening prior to index date. • Bowel cancer screening: from the last 3 years that they were age-eligible for screening prior to index date. • NHS healthcare checks: the last 6 years that they were age-eligible for NHS health checks prior to index date. • Influenza vaccination: from 1 September 2018-31 March 2019 (influenza and negative control exposure analysis); from 1 September 2019-31 March 2020 (COVID-19 analysis). • Pneumococcal vaccination: any time prior to index. • Prostate specific antigen testing: the last three years prior to index. • Bone density scans: the last three years prior to index. • GP practice visits: the last year prior to index. • Did not attend primary care visit: the last year prior to index. • Low value procedures: the last year prior to index. • Hospital visit for ambulatory care sensitive conditions: the last five years prior to index. • Blood pressure measurements: the last year prior to index. Code lists from Graham et al.[236] were utilised. The code lists from this project can be found listed on LSHTM data compass: https://doi.org/10.17037/DATA.00003684.

Abbreviations: HES: Hospital Episode Statistics; ICD-10: International Classification of Diseases 10th revision; IMD: index of multiple deprivation; LSHTM: London School of Hygiene and Tropical Medicine; OPCS: Operating Procedure Codes Supplement.

Supplementary Table 2. Influenza vaccination algorithm

Combination of codes on same day	Total number of vaccination events	Decision	Rationale
Given and neutral	725432	Record as valid vaccination.	

Given and absent	1149 (of these only 924 are the first vaccination dose)	Do not record as valid vaccination.	The prevalence of this marker may be underestimated very slightly, however, in this instance better to be more specific than sensitive when it comes to confounders[273].
Given and adverse	14	Record as valid vaccination.	Likely that this patient received the vaccination, but then had an adverse event on the same day.
Given and product	395881	Record as valid vaccination.	
Given and given with lag	4700	Record as valid vaccination.	
Neutral and absent	1178	Do not record as valid vaccination.	
Neutral and adverse	9	Record as valid vaccination.	
Neutral and product	295746	Record as valid vaccination.	
Neutral and given lag	7748	Record as valid vaccination.	
Absent and adverse	27	Do not record as valid vaccination.	Likely that these patients are reporting a previous adverse event as the reason for not wanting to get vaccinated.
Absent and product	218 (of these only 194 are first vaccination dose)	Do not record as valid vaccination.	The prevalence of this marker will be underestimated very slightly, however, in this instance better to be more specific than sensitive when it comes to confounders[273].
Absent and given with lag	570	Record as valid vaccination.	This most likely reflects that the reason a vaccine wasn't given on the event date is that the patient had already had it elsewhere. We can be reasonably confident that the patient was vaccinated, but we don't know the exact date.
Adverse and product	<5	Record as valid vaccination.	Likely that these patients received the vaccination, but then had an adverse event on the same day.
Adverse and given with lag	0	Ignore – no events.	
Given with lag and product	784	Record as valid vaccination.	

Note: this table is attributed from Graham et al.[236] Since influenza vaccinations can be identified using both medcodes and prodcodes and since medcodes do not always insinuate presence of a vaccination, an algorithm was developed for combinations of codes that occurred on the same day. Medcodes were separated into those that were clearly given ("given" or "administered"), given with delay ("given" or "administered" but evidence this occurred in another setting previously), neutral (vaccination mentioned but no "given" or "administered") and absent (vaccination "refused" or "not consented). Then we looked at vaccination events (using both medcodes and prodcodes) that were recorded on the same date and categorised these according to the above framework. We found 33.5% of individuals with >1 influenza prodcode or medcode during the 2019/2020 season. However, since individuals can have both a prodcode and medcode recorded for each vaccination event these were unlikely to be true vaccination events and therefore were ignored. Cells with <5 individuals are redacted due to CPRD's patient confidentiality requirements and secondary suppression has occurred where necessary.

Supplementary Table 3. Baseline characteristics stratified in overall analysis populations.

Variable	Category	COVID-19 analysis population N=1,796,667	Influenza analysis population N=1,991,284	Negative control exposure analysis population N=1,946,943
Age category in years, N (%)	65-69	448,063 (22.5%)	319,958 (17.8%)	442,923 (22.7%)
	70-74	557,599 (28.0%)	540,955 (30.1%)	550,104 (28.3%)
	75-79	401,290 (20.2%)	391,299 (21.8%)	394,068 (20.2%)
	80-84	295,492 (14.8%)	278,776 (15.5%)	287,737 (14.8%)
	85-89	181,835 (9.1%)	169,263 (9.4%)	173,832 (8.9%)
	90-95	80,943 (4.1%)	74,430 (4.1%)	75,278 (3.9%)
	95+	26,062 (1.3%)	21,986 (1.2%)	23,001 (1.2%)
Sex, N (%)	Female	1,075,723 (54.0%)	973,794 (54.2%)	1,052,528 (54.1%)
	Male	915,561 (46.0%)	822,873 (45.8%)	894,415 (45.9%)
Ethnicity, N (%)	Asian	67,961 (3.4%)	63,829 (3.6%)	67,270 (3.5%)
	Black	36,912 (1.9%)	33,981 (1.9%)	36,454 (1.9%)
	Missing	98,874 (5.0%)	89,889 (5.0%)	97,103 (5.0%)
	Mixed	10,072 (0.5%)	9,209 (0.5%)	9,948 (0.5%)
	Other	17,711 (0.9%)	16,533 (0.9%)	17,593 (0.9%)
	White	1,759,754 (88.4%)	1,583,226 (88.1%)	1,718,575 (88.3%)
Region, N (%)	East Midlands	95,413 (4.8%)	35,304 (2.0%)	38,029 (2.0%)
	East of England	263,825 (13.2%)	81,497 (4.5%)	88,480 (4.5%)
	London	67,278 (3.4%)	239,370 (13.3%)	260,228 (13.4%)
	North East	381,592 (19.2%)	62,453 (3.5%)	66,590 (3.4%)
	North West	438,407 (22.0%)	347,533 (19.3%)	376,009 (19.3%)
	South East	270,484 (13.6%)	395,424 (22.0%)	432,893 (22.2%)
	South West	357,363 (17.9%)	245,084 (13.6%)	259,419 (13.3%)
	West Midlands	74,805 (3.8%)	323,547 (18.0%)	351,397 (18.0%)

	Yorkshire and The Humber	11 (0.0%)	66,352 (3.7%)	73,860 (3.8%)
	Unknown	499,873 (25.1%)	103 (0.0%)	38 (0.0%)
IMD, N (%)	1 (least deprived)	467,781 (23.5%)	458,067 (25.5%)	492,109 (25.3%)
	2	399,311 (20.1%)	420,214 (23.4%)	457,789 (23.5%)
	3	344,964 (17.3%)	357,465 (19.9%)	386,545 (19.9%)
	4	279,355 (14.0%)	310,988 (17.3%)	336,847 (17.3%)
	5 (most deprived)	95,413 (4.8%)	249,933 (13.9%)	273,653 (14.1%)
Influenza 'at-risk' conditions, N (%)	Immunosuppressed status	55,153 (2.8%)	44,055 (2.5%)	62,795 (3.2%)
	Other comorbidities*	1,103,284 (55.4%)	1,004,589 (55.9%)	1,084,662 (55.7%)
Markers of health-seeking behaviour, N (%)	AAA screen	231,088 (11.6%)	214,942 (12.0%)	227,316 (11.7%)
	Breast screen	346,116 (17.4%)	327,105 (18.2%)	343,720 (17.7%)
	Cervical screen	397,303 (20.0%)	362,943 (20.2%)	388,994 (20.0%)
	Bowel screen	1,439,412 (72.3%)	1,354,825 (75.4%)	1,424,238 (73.2%)
	NHS health checks	372,244 (18.7%)	321,029 (17.9%)	363,677 (18.7%)
	Influenza vaccine†	1,460,391 (73.3%)	1,363,429 (75.9%)	1,427,057 (73.3%)
	Pneumococcal vaccine	1,242,359 (62.4%)	1,158,676 (64.5%)	1,233,497 (63.4%)
	PSA test	352,272 (17.7%)	366,640 (20.4%)	363,037 (18.6%)
	Bone density scan	100,892 (5.1%)	115,237 (6.4%)	108,540 (5.6%)
	Low value procedures	358,881 (18.0%)	537,129 (29.9%)	435,940 (22.4%)
	Primary care DNA	601,896 (30.2%)	935,986 (52.1%)	731,921 (37.6%)
	Hospital visit for ambulatory care sensitive conditions	190,136 (9.5%)	190,805 (10.6%)	195,863 (10.1%)
	Blood pressure test	1,470,006 (73.8%)	1,526,971 (85.0%)	1,563,575 (80.3%)
	Primary care visit	1,844,823 (92.6%)	1,741,469 (96.9%)	1,847,410 (94.9%)

Abbreviations: AAA: abdominal aortic aneurysm; DNA: did not attend; GP: general practice; IMD: index of multiple deprivation; N: numerator; NHS: National Health Service; PSA: prostate specific antigen; VE: vaccine effectiveness.

*Other comorbidities includes: chronic liver disease, chronic cardiac disease, chronic respiratory disease, asthma, diabetes mellitus, chronic neurological disease, chronic kidney disease, severe obesity, severe mental conditions and severe learning disability. For more information on how these were defined see **Supplementary Table 1.**

†Influenza vaccination that occurred in the influenza season prior to index date. For COVID-19 this was an influenza vaccination that occurred 1 September 2019 – 31 March 2020; for Influenza and Negative control exposure this was an influenza vaccination that occurred 1 September 2018-31 March 2019.

Notes: we are comparing individuals with ≥1 vaccination versus no vaccination throughout whole of follow-up. Cells with <5 individuals are redacted due to CPRD's patient confidentiality requirements and secondary suppression has occurred where necessary.

Supplementary Table 4. Unadjusted rates for each analysis and outcome

	Infections			Hospital or death			Death		
Exposure	**Events**	**Person years**	**Rate per 1,000 person years**	**Events**	**Person years**	**Rate per 1,000 person years**	**Events**	**Person years**	**Rate per 1,000 person years**
COVID-19 cohort analysis									
BNT162b2									
Unvaccinated	14,516	152,174.8	95.39	5525	153,293.7	36.04	3076	153,698.3	20.01
One dose	2,381	107,497.5	22.15	622	10,8495	5.73	366	10,8655	3.37
Two doses	77	11,062.77	6.96	10	11,121.57	0.9	6	11,128.44	0.54
ChAdOx1									
Unvaccinated	22559	186384.6	121.03	7206	188467.5	38.23	3097	189131.7	16.37
One dose	1626	90524.3	17.96	459	91875.61	5	311	92119.05	3.38
Two doses	[Redacted]	76.75	[Redacted]	0	79.27	0	0	79.5	0
Influenza cohort analysis									
Unvaccinated	40420	427719.1	94.5	7451	432900.5	17.21	364	434087.6	0.84

One dose	54210	462753.6	117.15	9263	476088.3	19.46	428	478248.4	0.89
Negative control exposure cohort analysis									
Unvaccinated	6192	528783.4	11.71	2257	528788.2	4.27	736	528789.8	1.39
One dose	22095	1604967	13.77	8176	1604968	5.09	2865	1604973	1.79

Notes: events represents the total number of events identified in the study follow up period. Population represents the total number of individuals included in each group. Person years is the total time in years until end of follow-up. It should be noted that for the COVID-19 and influenza analyses person years is time whilst unexposed/exposed, whereas for the negative control exposure analysis, since we used a binary exposure at baseline, person-years is all available follow-up from index until end of follow-up. Rate is calculated as the total number of events divided by the total person years multiplied by 1,000. Cells with <5 events are redacted due to CPRD's patient confidentiality requirements and secondary suppression has occurred where necessary.

Supplementary Table 5. Vaccine effectiveness estimates

Model	**All infections VE (95%CI)**	**Hospitalisation or death VE (95%CI)**	**Death VE (95%CI)**
Influenza			
Baseline	-5.5 (-7.2, -3.9)	21.2 (18.3, 24.0)	42.5 (32.8, 50.8)
Demography	-6.6 (-8.3, -4.9)	20.1 (17.1, 22.9)	42.4 (32.7, 50.8)
Comorbidities	-1.5 (-3.2, 0.1)	24.7 (22.0, 27.4)	43.9 (34.4, 52.0)
Markers	7.1 (5.4, 8.7)	26.3 (23.1, 29.2)	47.5 (37.3, 56.1)
Sensitivity	7.2 (5.5, 8.9)	26.4 (23.3, 29.4)	47.4 (37.1, 55.9)
COVID-19 (BNT162b2) dose one			
Baseline	42.3 (39.1, 45.4)	70.3 (67.4, 73.0)	83.8 (81.7, 85.7)
Demography	40.6 (37.3, 43.7)	69.1 (66, 71.9)	83.1 (80.9, 85.0)
Comorbidities	41.5 (38.2, 44.6)	69.9 (66.8 - 72.6)	83.5 (81.4, 85.4)
Markers	42.2 (38.9, 45.3)	69.8 (66.8, 72.6)	83.6 (81.5, 85.5)
Sensitivity	42.1 (38.9, 45.2)	69.8 (66.8, 72.5)	83.6 (81.5, 85.5)
COVID-19 (BNT162b2) dose two			
Baseline	82.7 (78.3, 86.2)	96.2 (93.0, 98.0)	98.2 (95.9, 99.2)

Model	All infections VE (95%CI)	Hospitalisation or death VE (95%CI)	Death VE (95%CI)
Demography	82.4 (77.9, 86.0)	96.1 (92.8, 97.9)	98.1 (95.8, 99.2)
Comorbidities	82.8 (78.4, 86.3)	96.3 (93.0, 98.0)	98.2 (95.9, 99.2)
Markers	83.1 (78.7, 86.5)	96.3 (93.0, 98.0)	98.2 (95.9, 99.2)
Sensitivity	83.0 (78.7, 86.5)	96.2 (93.0, 98.0)	98.2 (95.9, 99.2)
COVID-19 (ChAdOx1)			
Baseline	7.6 (0.4, 14.2)	24.9 (14.7 - 33.9)	51.0 (43.0 - 57.8)
Demography	5.0 (-2.3 - 11.9)	21.9 (11.3 - 31.2)	49.3 (41.1 - 56.4)
Comorbidities	6.5 (-0.8 - 13.2)	23.5 (13.1 - 32.7)	50.4 (42.3 - 57.3)
Markers	9.6 (2.6 - 16.2)	25.7 (15.6 - 34.6)	52.5 (44.8 - 59.2)
Sensitivity	8.8 (1.7 - 15.4)	25.3 (15.2 - 34.3)	52.4 (44.6 - 59.1)
Negative control exposure			
Baseline	-11.5 (-14.8 - -8.4)	-6.2 (-11.3 - -1.4)	-6.4 (-15.4 - 1.9)
Demography	-15 (-18.4 - -11.8)	-12 (-17.4 - -6.9)	-12.2 (-21.7 - -3.3)
Comorbidities	-7.5 (-10.6 - -4.5)	-1.2 (-6.1 - 3.4)	-2.5 (-11.2 - 5.6)
Markers	-2.1 (-6.0 - 1.7)	4.6 (-1.3 - 10.2)	1.3 (-9.5 - 11.1)
Sensitivity	-2.2 (-6.1 - 1.5)	4.5 (-1.5 - 10.1)	1.2 (-9.6 – 11.0)

Notes: baseline: adjusted for polynomial age, sex, region and recent infection. Demography model: baseline model + adjusted for ethnicity and IMD. Comorbidities: demography model + adjusted for immunosuppressed status and other comorbidities. Markers: comorbidities model + adjusted for each marker of health-seeking behaviour separately with sex interactions for sex-specific markers. Sensitivity: markers model + age interactions for AAA screening, bowel cancer screening, NHS health checks and ACS conditions. Vaccine effectiveness is estimated as (1-HR)*100.
Abbreviations: SES: socioeconomic status.

Supplementary Table 8. Amongst vaccinated individuals only, median days from index to influenza vaccination stratified by marker status and age category

	Presence of marker		Absence of marker	
Marker	**Age category**	**Median (q1 – q3)**	**Age category**	**Median (q1 – q3)**
AAA Screen Males	Overall	50 (39 - 67)	Overall	50 (38 - 66)

	65-69	52 (39 - 69)	65-69	53 (40 - 72)
	70-74	50 (39 - 67)	70-74	51 (39 - 67)
	75-79	48 (38 - 62)	75-79	49 (38 - 65)
	80-84	48 (37 - 62)	80-84	48 (38 - 64)
	85+	47 (37 - 62)	85+	50 (39 - 67)
Bowel Cancer Screen	Overall	50 (39 - 67)	Overall	51 (39 - 68)
	65-69	52 (39 - 69)	65-69	53 (41 - 72)
	70-74	50 (39 - 67)	70-74	51 (40 - 68)
	75-79	48 (38 - 65)	75-79	51 (39 - 66)
	80-84	48 (38 - 64)	80-84	50 (38 - 66)
	85+	50 (38 - 65)	85+	52 (39 - 69)
Breast Cancer Screen Females	Overall	51 (39 - 67)	Overall	51 (39 - 68)
	65-69	51 (39 - 68)	65-69	52 (39 - 69)
	70-74	50 (38 - 65)	70-74	50 (39 - 67)
	75-79	48 (38 - 64)	75-79	50 (39 - 66)
	80-84	50 (38 - 65)	80-84	50 (39 - 66)
	85+	50 (39 - 66)	85+	52 (40 - 71)
Cervical Cancer Screen Females	Overall	51 (39 - 67)	Overall	51 (39 - 67)
	65-69	52 (40 - 69)	65-69	51 (39 - 69)
	70-74	50 (39 - 66)	70-74	50 (39 - 66)
	75-79	50 (38 - 65)	75-79	50 (38 - 65)
	80-84	50 (38 - 65)	80-84	50 (39 - 66)
	85+	51 (39 - 68)	85+	53 (40 - 72)
NHS Health Checks	Overall	50 (39 - 67)	Overall	51 (39 - 67)
	65-69	52 (39 - 69)	65-69	52 (39 - 69)
	70-74	50 (39 - 66)	70-74	50 (39 - 67)
	75-79	48 (38 - 64)	75-79	50 (38 - 65)
	80-84	48 (38 - 64)	80-84	50 (38 - 65)
	85+	52 (40 - 71)	85+	51 (39 - 68)
Influenza Vaccination	Overall	50 (39 - 67)	Overall	59 (41 - 86)

	65-69	51 (39 - 67)	65-69	61 (43 - 89)
	70-74	50 (38 - 65)	70-74	59 (41 - 87)
	75-79	48 (38 - 64)	75-79	57 (40 - 83)
	80-84	48 (38 - 64)	80-84	57 (40 - 81)
	85+	51 (39 - 67)	85+	59 (41 - 86)
Pneumococcal Vaccination	Overall	50 (39 - 67)	Overall	53 (40 - 72)
	65-69	50 (38 - 66)	65-69	54 (41 - 74)
	70-74	48 (38 - 65)	70-74	53 (40 - 72)
	75-79	48 (38 - 64)	75-79	52 (39 - 69)
	80-84	48 (38 - 64)	80-84	52 (39 - 68)
	85+	51 (39 - 68)	85+	53 (40 - 73)
ACS Hospital Visit	Overall	50 (39 - 67)	Overall	50 (39 - 67)
	65-69	52 (39 - 71)	65-69	52 (39 - 69)
	70-74	51 (39 - 68)	70-74	50 (39 - 67)
	75-79	50 (38 - 67)	75-79	48 (38 - 65)
	80-84	51 (39 - 68)	80-84	48 (38 - 65)
	85+	53 (40 - 73)	85+	51 (39 - 68)
Blood Pressure Test	Overall	50 (39 - 67)	Overall	52 (39 - 69)
	65-69	51 (39 - 69)	65-69	53 (40 - 72)
	70-74	50 (39 - 66)	70-74	51 (39 - 68)
	75-79	48 (38 - 65)	75-79	51 (39 - 66)
	80-84	48 (38 - 65)	80-84	51 (39 - 67)
	85+	51 (39 - 68)	85+	52 (40 - 71)
Bone Density Scan	Overall	50 (39 - 67)	Overall	51 (39 - 67)
	65-69	51 (39 - 68)	65-69	52 (39 - 69)
	70-74	50 (38 - 65)	70-74	50 (39 - 67)
	75-79	48 (38 - 64)	75-79	50 (38 - 65)
	80-84	48 (38 - 64)	80-84	50 (38 - 65)
	85+	50 (38 - 66.75)	85+	52 (39 - 69)
DNA Primary Care Visit	Overall	50 (39 - 67)	Overall	50 (39 - 66)

	65-69	52 (39 - 71)	65-69	52 (39 - 69)
	70-74	51 (38 - 68)	70-74	50 (39 - 66)
	75-79	50 (38 - 66)	75-79	48 (38 - 65)
	80-84	50 (38 - 66)	80-84	48 (38 - 64)
	85+	52 (39 - 70)	85+	51 (39 - 68)
Primary care visit	Overall	50 (39 - 67)	Overall	53 (39 - 72)
	65-69	52 (39 - 69)	65-69	54 (39 - 75)
	70-74	50 (39 - 66)	70-74	53 (39 - 72)
	75-79	48 (38 - 65)	75-79	52 (39 - 69)
	80-84	50 (38 - 65)	80-84	52 (39 - 67)
	85+	51 (39 - 68)	85+	54 (40 - 73)
Low-Value Procedure	Overall	50 (39 - 67)	Overall	51 (39 - 67)
	65-69	51 (39 - 69)	65-69	52 (39 - 69)
	70-74	50 (38 - 67)	70-74	50 (39 - 67)
	75-79	49 (38 - 65)	75-79	50 (38 - 65)
	80-84	50 (38 - 65)	80-84	50 (38 - 65)
	85+	52 (39 - 69)	85+	51 (39 - 68)
PSA Test	Overall	50 (39 - 67)	Overall	51 (39 - 68)
	65-69	52 (40 - 69)	65-69	52 (39 - 71)
	70-74	50 (39 - 66)	70-74	51 (39 - 68)
	75-79	48 (38 - 64)	75-79	50 (38 - 66)
	80-84	48 (38 - 62)	80-84	50 (38 - 65)
	85+	48 (38 - 65)	85+	51 (39 - 68)

Abbreviations: AAA; abdominal aortic aneurysm; ACS: ambulatory care sensitive; DNA: did not attend; GP: general practice; PSA: prostate specific antigen; q1: first quartile; q3: third quartile.

Note: this table does not include unvaccinated individuals. We combined 85-89, 90-94 and 95+ age categories due to low counts.

Supplementary Table 7. Amongst vaccinated individuals only, median days from index to first COVID-19 vaccination by marker status and age category

	Presence of marker		Absence of marker	
Marker	**Age category**	**Median (q1 – q3)**	**Age category**	**Median (q1 – q3)**
AAA Screen Males	Overall	67 (58 - 75)	Overall	64 (55 - 72)
	65-69	80 (75 - 84)	65-69	80 (74 - 84)
	70-74	71 (66 - 74)	70-74	70 (66 - 74)
	75-79	60 (57 - 66)	75-79	60 (57 - 66)
	80-84	53 (48 - 58)	80-84	53 (47 - 59)
	85+	52 (46 - 58)	85+	53 (47 - 60)
Bowel Cancer Screen	Overall	66 (57 - 74)	Overall	56 (50 - 65)
	65-69	80 (74 - 84)	65-69	80 (75 - 84)
	70-74	71 (66 - 74)	70-74	71 (66 - 75)
	75-79	60 (57 - 67)	75-79	60 (57 - 67)
	80-84	53 (47 - 59)	80-84	53 (49 - 59)
	85+	52 (46 - 59)	85+	53 (48 - 60)
Breast Cancer Screen Females	Overall	66 (57 - 74)	Overall	66 (56 - 74)
	65-69	80 (74 - 84)	65-69	80 (74 - 84)
	70-74	71 (66 - 74)	70-74	71 (66 - 74)
	75-79	60 (56 - 66)	75-79	60 (57 - 67)
	80-84	53 (48 - 59)	80-84	53 (48 - 59)
	85+	53 (48 - 59)	85+	53 (48 - 61)
Cervical Cancer Screen Females	Overall	66 (57 - 74)	Overall	66 (57 - 75)
	65-69	80 (74 - 84)	65-69	80 (74 - 84)
	70-74	71 (66 - 74)	70-74	71 (66 - 74)
	75-79	60 (57 - 66)	75-79	60 (57 - 67)
	80-84	53 (48 - 59)	80-84	53 (48 - 59)
	85+	53 (48 - 60)	85+	53 (48 - 61)
NHS Health Checks	Overall	66 (57 - 74)	Overall	66 (56 - 74)
	65-69	80 (75 - 84)	65-69	80 (74 - 84)

	70-74	71 (66 - 74)	70-74	71 (66 - 74)
	75-79	60 (57 - 66)	75-79	60 (57 - 67)
	80-84	53 (47 - 59)	80-84	53 (48 - 59)
	85+	53 (49 - 60)	85+	53 (47 - 60)
Influenza Vaccination	Overall	66 (57 - 74)	Overall	70 (60 - 78)
	65-69	80 (74 - 84)	65-69	80 (76 - 86)
	70-74	70 (66 - 74)	70-74	72 (67 - 76)
	75-79	60 (57 - 66)	75-79	62 (58 - 67)
	80-84	53 (47 - 59)	80-84	55 (50 - 62)
	85+	53 (47 - 60)	85+	56 (50 - 65)
Pneumococcal Vaccination	Overall	66 (57 - 74)	Overall	70 (60 - 78)
	65-69	79 (74 - 83)	65-69	80 (75 - 85)
	70-74	71 (66 - 74)	70-74	71 (66 - 74)
	75-79	60 (57 - 66)	75-79	61 (57 - 67)
	80-84	53 (47 - 59)	80-84	53 (49 - 60)
	85+	53 (47 - 60)	85+	53 (49 - 62)
ACS Hospital Visit	Overall	66 (57 - 74)	Overall	67 (58 - 75)
	65-69	79 (72 - 83)	65-69	80 (75 - 84)
	70-74	70 (65 - 74)	70-74	71 (66 - 74)
	75-79	60 (56 - 67)	75-79	60 (57 - 67)
	80-84	53 (49 - 60)	80-84	53 (48 - 59)
	85+	54 (49 - 63)	85+	53 (47 - 60)
Blood Pressure Test	Overall	66 (57 - 74)	Overall	71 (60 - 78)
	65-69	80 (74 - 84)	65-69	80 (77 - 85)
	70-74	71 (66 - 74)	70-74	72 (67 - 75)
	75-79	60 (57 - 66)	75-79	61 (58 - 67)
	80-84	53 (48 - 59)	80-84	53 (49 - 59)
	85+	53 (47 - 60)	85+	54 (49 - 61)
Bone Density Scan	Overall	66 (57 - 74)	Overall	66 (57 - 74)
	65-69	79 (73 - 83)	65-69	80 (74 - 84)

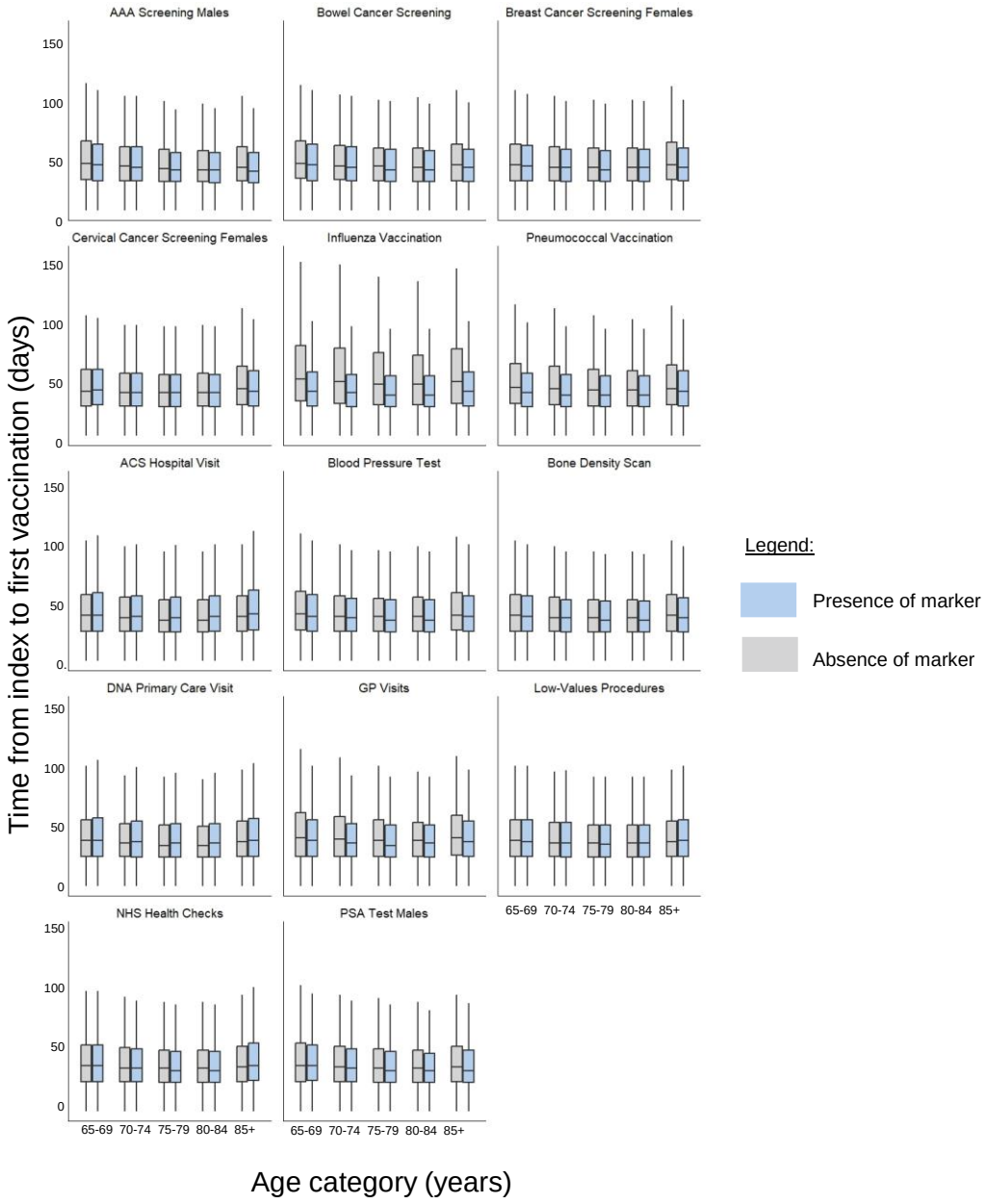
	70-74	70 (65 - 74)	70-74	71 (66 - 74)
	75-79	60 (57 - 66)	75-79	60 (57 - 67)
	80-84	53 (48 - 59)	80-84	53 (48 - 59)
	85+	53 (47 - 60)	85+	53 (47 - 60)
DNA Primary Care Visit	Overall	66 (57 - 74)	Overall	67 (58 - 75)
	65-69	80 (74 - 84)	65-69	80 (75 - 84)
	70-74	71 (66 - 74)	70-74	71 (66 - 74)
	75-79	60 (57 - 67)	75-79	60 (57 - 66)
	80-84	53 (48 - 59)	80-84	53 (47 - 59)
	85+	53 (48 - 61)	85+	53 (47 - 60)
Primary care visit	Overall	66 (57 - 74)	Overall	74 (66 - 81)
	65-69	80 (74 - 84)	65-69	82 (78 - 88)
	70-74	71 (66 - 74)	70-74	73 (68 - 77)
	75-79	60 (57 - 67)	75-79	65 (59 - 70)
	80-84	53 (48 - 59)	80-84	57 (51 - 64)
	85+	53 (47 - 60)	85+	56 (51 - 66)
Low-Value Procedure	Overall	66 (57 - 74)	Overall	67 (58 - 75)
	65-69	79 (73 - 83)	65-69	80 (75 - 84)
	70-74	70 (65 - 74)	70-74	71 (66 - 74)
	75-79	60 (57 - 66)	75-79	60 (57 - 67)
	80-84	53 (48 - 59)	80-84	53 (48 - 59)
	85+	53 (47 - 61)	85+	53 (47 - 60)
PSA Test	Overall	67 (58 - 75)	Overall	68 (58 - 76)
	65-69	80 (74 - 83)	65-69	80 (75 - 85)
	70-74	70 (66 - 74)	70-74	71 (66 - 74)
	75-79	60 (57 - 66)	75-79	60 (57 - 67)
	80-84	53 (47 - 58)	80-84	53 (48 - 59)
	85+	53 (46 - 59)	85+	53 (47 - 60)

Abbreviations: AAA; abdominal aortic aneurysm; ACS: ambulatory care sensitive; DNA: did not attend; GP: general practice; PSA: prostate specific antigen; q1: first quartile; q3: third quartile.

Note: this table does not include unvaccinated individuals and this only includes days until first COVID-19 vaccination. We combined 85-89, 90-94 and 95+ age categories due to low counts.

Figures

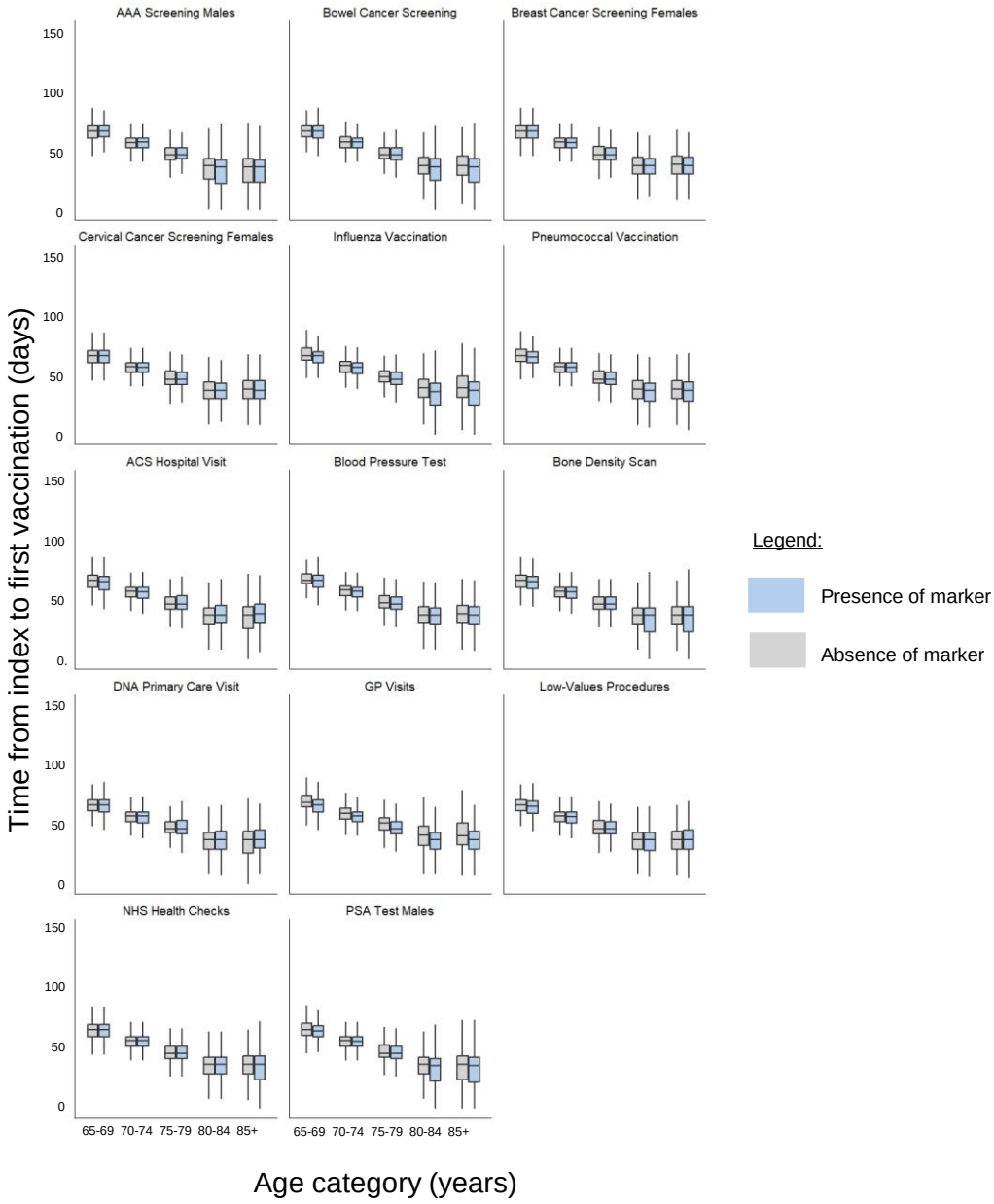
Supplementary Figure 1. Amongst vaccinated individuals only, box plots for median days from index date to influenza vaccination stratified by marker status and age category



Abbreviations: AAA: abdominal aortic aneurysm; ACS: ambulatory care sensitive; DNA: did not attend; GP: general practice; NHS: national health service; PSA: prostate specific antigen.

Note: the raw data from the figure can be found in **Supplementary Table 6.** This figure does not include unvaccinated individuals. We combined 85-89, 90-94 and 95+ age categories due to low counts.

Supplementary Figure 2. Amongst vaccinated individuals only, box plots for median days from index to first COVID-19 vaccination stratified by marker status and age category



Abbreviations: AAA: abdominal aortic aneurysm; ACS: ambulatory care sensitive; DNA: did not attend; GP: general practice; NHS: national health service; PSA: prostate specific antigen.

Note: the raw data from the figure can be found in **Supplementary Table 7**. We combined 85-89, 90-94 and 95+ age categories due to low counts.