OpenSAFELY Protocol: Assessing Covid-19 Vaccine Effectiveness

This is a collaboration between the following institutions as part of OpenSAFELY.org:

The DataLab, Nuffield Department of Primary Care Health Sciences, University of Oxford

Electronic Health Records Research Group, Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine

Population Health Science Institute, University of Bristol

Version: v1.3

Date: 10/05/2021

Table of Contents

Background

Vaccines are a critical tool in the fight against SARS-CoV-2 infection. Results from RCTs of vaccines are encouraging, and The UK Medicines and Healthcare products Regulatory Agency (MHRA) has so far approved the Pfizer-BioNTech mRNA vaccine (2 Dec 2020, first administered 8 Dec 2020), the Oxford-AstraZeneca adenovirus vector vaccine (30 Dec 2020, first administered 4 Jan 2021), and the Moderna vaccine (8 Jan 2021) for use in the United Kingdom. More approvals are expected.

On 8 December 2020, the Pfizer-BioNTech mRNA vaccine was administered to patients in the UK for the first time in a non-trial setting. This was followed by the Oxford-AstraZeneca adenovirus vector vaccine on 4 January 2021. Vaccines were offered to patients based on prioritisation groups defined with the intention to first protect those at highest risk of infection (for example health care workers) and those at highest risk of experiencing severe post-infection outcomes (for example the elderly and sick).

Assessing vaccines in non-trial settings is crucial to ensure the vaccination programmes are effective public health interventions, to understand vaccine effectiveness in different patient groups, and to understand possible differential effectiveness against new SARS-CoV-2 variants.

Objectives

Primary Objectives

The primary objective is to assess the out-of-trial effectiveness of (at least) one dose of a COVID-19 vaccine administered as part of the COVID-19 vaccination programme in England to patients aged 70 and over. The outcomes are recorded SARS-CoV-2 infection, COVID-19-related hospitalisation and death.

Secondary Objectives

The secondary objectives are to assess vaccine effectiveness in the effectiveness of the Pfizer and AstraZeneca brands separately, and specific clinical subgroups.

Methods

Database Description

We use data from general practice (GP) records, obtained from the GP software provider TPP, linked to:

- Second Generation Surveillance System (SGSS) data on SARS-CoV-2 test results;
- Secondary Uses Service (SUS) data on hospital admissions (Admitted Patient Care Statistics (APCS) dataset);
- Death registration data held by the Office for National Statistics (ONS), including date and ICD-10 coded cause of death for all deaths occurring in England and Wales.

The data was accessed, linked and analysed through openSAFELY.org - a data analytics platform created by our team on behalf of NHS England to address urgent questions relating to the epidemiology and treatment of COVID-19(REF). OpenSAFELY provides a secure software interface allowing the analysis of pseudonymised primary care patient records from England in near real-time within the EHR vendor's highly secure data centre, avoiding the need for large volumes of potentially disclosive pseudonymised patient data to be transferred off-site. This, in addition to other technical and organisational controls, minimises any risk of re-identification. Similarly pseudonymised datasets from other data providers are securely provided to the EHR vendor and linked to the primary care data. Descriptions of OpenSAFELY have been previously published (REF), and more information can be found on https://opensafely.org/.

Primary care records retrieved from the TPP SystmOne electronic health record system include diagnoses (Read 3 CTV3), prescriptions (dm+d), basic sociodemographics and laboratory results for 22 million individuals – approximately 40% of the English population. Data extracted by SystmOne have previously been used in medical research, as part of the ResearchOne dataset (REFS).

All data is held in a secure research environment hosted by TPP, which is a Tier 3 data centre, accredited to NHS Digital standards for centrally hosted clinical systems (ISO 27001 standard and IG Toolkit version 2). We received ethics approval to conduct the data linkage and analyses by the London - City & East Research Ethics Committee on the 2nd of April 2020 (REC reference: 20/LO/0651) and LSHTM Ethics Board (ref 21863). No further ethical or research governance approval was required by the University of Oxford but copies of the approval documents were reviewed and held on record.

Information Governance

NHS England is the data controller; TPP is the data processor; and the key researchers on OpenSAFELY are acting on behalf of NHS England. This implementation of OpenSAFELY is hosted within the TPP environment, which is accredited to the ISO 27001 information security standard and is NHS IG Toolkit compliant 52,53; patient data have been pseudonymized for analysis and linkage using industry standard cryptographic hashing techniques; all pseudonymized datasets transmitted for linkage onto OpenSAFELY are encrypted; access to the platform is through a virtual private network (VPN) connection; the researchers hold contracts with NHS England and only access the platform to initiate database queries and statistical models; all database activity is logged; and only aggregate statistical outputs leave the platform environment following best practice for anonymization of results such as statistical disclosure control for low cell counts54. The OpenSAFELY research platform adheres to the data protection principles of the UK Data Protection Act 2018 and the EU General Data Protection Regulation (GDPR) 2016. In March 2020, the Secretary of State for Health and Social Care used powers under the UK Health Service (Control of Patient Information) Regulations 2002 (COPI) to require organizations to process confidential patient information for the purposes of protecting public health, providing healthcare services to the public and monitoring and managing the COVID-19 outbreak and incidents of exposure55. Together, these provide the legal bases to link patient datasets on the OpenSAFELY platform. GP practices, from which the primary care data are obtained, are required to share relevant health information to support the public health response to the pandemic, and have been informed of the OpenSAFELY analytics platform.

Design

Overview

A full cohort design following all patients aged 80 and over within the OpenSAFELY-TPP database, under follow-up from the start of 8 December 2020 (the start of the national vaccine roll-out in England) until [the most recent date with adequate outcome coverage].

Study Population

The study will be undertaken independently in patients aged 80 years and over, and patients aged 70-79, as at 31 March 2020. These correspond to age-based criteria defining JCVI priority groups 2 and groups 3 and 4 respectively.

Adults aged 70 and over are chosen as:

- They are easily-identifiable in the primary care record.
- They were amongst the first people to be vaccinated in England, so there is sufficient exposed-person-time for analysis.
- The vast majority are non-working, and so effectiveness is not significantly confounded by occupation, unlike the under 65s which includes frontline health and social care workers who were high-priority for vaccination (JCVI priority group 2) and are at higher risk of exposure to SARS-CoV-2. Occupation is not easily-identifiable in the primary care record.

All patients who are alive and registered at a GP practice using TPP's SystmOne Clinical Information System on 7 December 2020 are considered.

Exclusion criteria

- Missing age, sex, IMD, ethnicity, geography.
- Care home residents. Timing of vaccination and risk of infection is highly correlated for patients residing in the same home. Once a care home starts vaccinating residents, they are likely to reach close to 100% coverage quickly. This group may be considered in a separate analysis.
- Patients with evidence of prior Covid-19 infection, either with a positive test or probable/suspected cases identified in primary care. These people may be less likely to accept vaccine and may have a different immuno-response to the vaccine.

- ~~Participants in UK-based vaccine trials, who after unblinding were found to be vaccinated. Only a very small proportion of patients but nevertheless contributes to measurement error for the exposure.~~ [Not currently available]
- People who have been vaccinated but the brand is unknown, to ensure consistency between any vaccine analysis and brand-specific analyses.There are only a very small number of such cases (<1%).
- People who are considered to be near death, for example on palliative care pathways, since vaccination rates may be lower in this group.

Follow-up

Patients are followed-up until the first occurrence of:

- The outcome of interest
- 22 February 2021 (to be extended when more follow up time is available)
- Death
- De-registration

The end date is chosen as the latest date before which coverage of all primary outcome variables was deemed complete. This is around 4-6 weeks prior to the latest reported hospital admission in APCS, as this data-source has the highest lag from events occurring to appearing in the OpenSAFELY-TPP database.

Study Measures

Exposure

The exposure of interest is vaccination for COVID-19 in a non-trial setting. We will consider first-dose only. The exposure will be modelled as the time since first vaccination dose to allow for time-varying effects.

Patient vaccination status will be ascertained from the primary care record, as previously described (https://www.medrxiv.org/content/10.1101/2021.01.25.21250356v2). The record provides details on the vaccine date including first and subsequent doses, and vaccine brand.

Brand-specific exposures (currently either the P-B vaccine or the Ox-AZ vaccine though other brands will appear at a later date depending on approval and roll-out) will also be considered, with vaccination by one brand considered a censoring event for the risk of vaccination by any other brand.

Outcomes

The primary outcomes are:

- SARS-CoV-2 test positivity, from SGSS
- COVID-19-related hospital admission, from SUS-APCS
- ~~COVID-19-related ICU admission, from SUS-APCS~~
- ~~COVID-19-related death, from ONS death registry data~~ [not included as positive test is both a time-varying predictor of vaccination and a *structural component* of this outcome, so cannot be used as a confounder]
- All-cause death, from ONS registry data

Additionally, non-COVID-19 deaths will be modelled as a negative control outcome.

These outcomes will be modelled independently. Recurrent events for infection and admission are not considered.

The following outcomes are of interest but will not be considered here because of current difficulties in ascertaining them quickly or reliably with currently available data.

- Symptomatic infection
- Severe disease (hospitalised patients may include those with mild disease but who may be considered at high risk of deterioration)
- Transmission

Covariates

Covariates for the substantive vaccine-outcome models are chosen based on availability with the primary care record and linked data and anticipated role as confounders / mediators for the substantive vaccine-outcome relationship. In particular, characteristics anticipated to be risk factors for COVID-19 infection, poorer post-infection outcomes, and for being offered/accepting the vaccine.

Covariates for the IPW vaccine / censor models are chosen based on availability and their expected information value in predicting vaccination or censoring events.

Some covariates may be dropped from the IPW models if the case-count is especially low (leading to unstable effect estimates and over-fitting).

Time-independent

These covariates will be calculated at study start date (with the exception of age) and are assumed fixed for the duration of the study period:

- Age. This is defined as at 31 March 2020 as per JCVI priority groups
- Sex
- Ethnicity (in 5 categories: Black, Mixed, South Asian, White, Other)
- Deprivation: defines using quintiles of the English Index of Multiple Deprivation
- ~~Rurality (potentially influencing vaccine access)~~
- ~~Frailty, which may lower chances of vaccination due to mobility/access issues, may lower chances of infection due to shielding.~~
- Evidence of COVID-19 infection prior to study start date. This could reduce risk or severity of subsequent infection due to the induced immune response and may influence vaccine uptake (in either direction)
- Other comorbidities / clinical characteristics: chronic cardiac disease; chronic obstructive pulmonary disease (COPD); obesity (most recent adult body mass index (BMI) ≥30); dialysis; immunosuppressive diagnoses or medications; severe mental illness (psychosis, schizophrenia and bipolar disorder); learning disabilities including Down syndrome; dementia; lung cancer; haematological cancer; other cancers
- Previous flu vaccination
- ~~Medically homebound~~
- ~~Inferred occupation: this is a challenging issue, relevant to co-variates and stratification but not necessarily easily implementable in the same way as others in this list. At this point in the vaccination campaign younger vaccine recipients are especially likely to be healthcare workers. They will likely have higher exposure, and more ready access to tests. Looking at the positive test results broken down by age band for vaccinated people, rates are indeed substantially higher among younger people. This could be a challenge for some types of control, we can't readily identify unvaccinated care workers in the same age bands, and non-care workers in the same age band are likely to have lower exposure and lower likelihood of getting tested. It's a strong argument for stratifying by age, and it may be worth considering in extremis e.g. dichotomising cohort into over 70 (assumed non-care worker) and below.~~

Time-dependent

These covariates will be calculated on each day of follow-up for all individuals:

- Unplanned hospitalisation status, as a proxy for recent acute illness which may affect the short-term likelihood of vaccination
    - Separate by infectious / non-infectious disease related admissions
- ~~Recent interaction with health care services. Occurrence of both in-patient and out-patient episodes may increase vaccination propensity as these were used to target vaccinations.~~
- ~~Self-isolation: coded in primary care, probably specific and probably not sensitive~~

- ~~New diagnosis of a condition in the JCVI at-risk list~~
- Possible or probable SARS-CoV-2 infection as indicated in primary care records
- ~~[not easily available] estimates of local community infection rates~~
- Calendar time (by region) as a proxy for underlying changes in infection rates
- Evidence of COVID-19 infection up to 21 days after vaccination.

Geographical / administrative

Geographical region is an important component of infection risk due to the varying prevalence of infectious individuals in each region. The highest-resolution geographic region available is MSOA. There are currently no reliable, timely, structured sources of localised infection rates over time.

~~Similarly, administrative health bodies (Practice, PCN, CCG, STP) play a role in vaccine administration. These organisations are arranged geographically and in the absence of good local infection data, these may be used to adjust for both local infection risk (outcome) and vaccination (exposure).~~ [just using region / time interaction, due to computational constraints]

Statistical Analysis

Descriptive statistics

Key demographic and clinical characteristics will be reported as the study start date using appropriate summary statistics (frequencies and percentages; means and quantiles). These will also be reported at 28-day snapshots, stratified by vaccination status. Person-time, event counts, and event rates per person-time will be reported by vaccination status.

For context, vaccine uptake and case-status over the duration of the study period will be reported in a cumulative incidence plot, stratified by key demographic characteristics.

Statistical Modelling

The primary analysis will comprise a full cohort study, with all eligible patients contributing to unvaccinated and vaccinated person-time. We will use Marginal structural models (MSM) to account for time-varying confounding, by reweighting each day of person-time according to the inverse probability of remaining unvaccinated and uncensored at time t. The primary effect of interest is the hazard ratio for each outcome for vaccinated versus unvaccinated person-time.

Marginal Structural Models

Fitting MSMs is a multi-step process: first the data must be reshaped as one-row-per-patient-per-day of follow-up time, to encode the time-varying variables and estimate the exposure and censoring propensity for each day of follow-up; second, the exposure and censoring probabilities are estimated using logistic regression; third, the substantive model is fit using cluster-robust logistic regression, with person-time weighted by the inverse of the probabilities estimated in the previous step. The estimated exposure-outcome odds ratio from this approach is a good approximation of the hazard ratio assuming the outcome rate is low on each day of follow-up. The steps are outlined in more detail below.

First, a one-row-per-patient-per-day dataset is created, from the study start date (7 December 2021) until the study end date or an earlier censoring date. Time-varying variables (vaccination status, covariates, outcomes) are updated each day, with value changes (for instance vaccination status) assumed to occur at the end of each day. For example if a patient receives their first vaccine dose on day 3 they are considered to be at risk of vaccination on days 1, 2, and 3, and no longer at risk from day 4 onwards.

Second,models to estimate the time-dependent probability of vaccination are fit (henceforth the IPW models). Two models are necessary for stabilised weights: a full model including both time-dependent and time-independent covariates; and a reduced model excluding time-dependent covariates *except* variables related to time itself (for instance calendar time). The weight at time t for patient i derived from each IPW model is the probability of the patient's actual vaccination history, conditional on their time-dependent covariates, i.e.,

$$w_{it} = \prod_{t=1}^{t} P(vaccination\ status\ at\ time\ t \mid vaccination\ status\ at\ t-1,\ patient\ characteristics\ at\ time\ t)$$

The stabilised weight for each day of follow-up for each person is the ratio of the reduced model against the full model $\frac{w_{it}^{fxd}}{w_{it}}$.

Note that
$P(patient\ i\ vaccinated\ at\ time\ t \mid vaccinated\ at\ t-1,\ patient\ characteristics\ at\ time\ t) = 1.$

This estimation procedure is repeated for censoring events (in this case non-covid-related death). The final weights are the product of the vaccination weights and the censoring weights.

Finally, a logistic regression model is fit on the one-row-per-patient-per-day, reweighted by w_{it}, and adjusting for time-independent covariates. The vaccinated/unvaccinated odds-ratio is a good approximation for the hazard ratio if the event rate at each time t is small.

Variable specification

Time since start date will be modelled non-linearly with splines separately by region, to account for anticipated differences in infection risk over time and space. Age will be modelled using both a linear and a quadratic term to account for highly non-linear age-outcome relationships, in particular with respect to death.

Time-varying vaccine effects (i.e. non-proportional hazards) will be modelled using a) piecewise-constant hazards or b) splines. This is necessary to capture the (assumed) divergence of risk between the vaccinated and unvaccinated given the vaccine does not have an immediate biological effect in reducing outcomes, and may even have intermediate adverse risks due to behavioural changes (i.e., undertaking activities with higher infection risk soon after being vaccinated).

Sub-sampling

Due to the computation time involved in fitting models against person-time datasets on millions of patients, we will take only a sample of patients for the substantive analyses. For each outcome, sampling will proceed as follows: all patients experiencing the outcome will be included: a 10% sample of patients who do not experience the outcome will be included; all models will be reweighted to account for this sampling; patients will be sampled based on a ranked hash of the patient identifier, to ensure sufficient sampling overlap across outcomes.

Proposed stratified analyses (subgroups)

The proposed analysis will be undertaken overall, and independently across the following subgroups to allow for the possibility of vaccine-outcome effect modification:

- Sex
- Immunosuppressed patients
- Prior infection
- Oral steroids

Vaccine effect estimates between subgroups will be compared using Welch's t-test.

Sensitivity analyses

- Symptomatic positive tests as an outcome, not all positive tests. This is useful to determine whether vaccines may be effective for reducing symptomatic infections (and possibly transmission).
- Different ethnicity sources -- primary care only (~75% complete) versus primary care + SUS (~90% complete)
- Do not exclude those with evidence of SARS-CoV-2 infection prior to study start date, and assess effect modification of prior infection
- Include prior positive infection as a confounder
- Censor at occurrence of second dose

Priority groups for vaccination advised by the Joint Committee on Vaccination and Immunisation
Reproduced from the Green Book provisional guidance, 5 December 2020 update; Public Health England. Immunisation against Infectious Disease (the Green Book). Chapter 14a: COVID-19. (Provisional guidance subject to MHRA approval of vaccine supply). Available from: https://web.archive.org/web/20201127125222/https://www.gov.uk/government/publications/covid-19-the-green-book-chapter-14a.

Priority group	Risk group
1	Residents in a care home for older adults Staff working in care homes for older adults
2	All those 80 years of age and over Frontline Health and social care workers
3	All those 75 years of age and over
4	All those 70 years of age and over Clinically extremely vulnerable individuals (not including pregnant women and those under 16 years of age)
5	All those 65 years of age and over
6	Adults aged 16 to 65 years in an at-risk group*
7	All those 60 years of age and over
8	All those 55 years of age and over
9	All those 50 years of age and over

* 'at-risk' groups comprise individuals with any of: chronic respiratory disease including severe asthma, chronic heart disease including atrial fibrillation, peripheral vascular disease, a history of venous thromboembolism, chronic kidney disease, chronic liver disease, stroke, transient ischaemic attack, cerebral palsy, Down's syndrome, or other chronic neurological conditions which may compromise respiratory function, diabetes mellitus, immunosuppression due to disease or treatment including any history of haematological malignancy, morbid obesity, severe mental illness, adult carers, and younger adults in residential care settings.