AUTHORS

CONVALESCENCE Study Team

TITLE

Understanding the risk of adverse health events following COVID-19 diagnosis prior to vaccines becoming available and in the era of delta among the fully vaccinated and the electively unvaccinated

RESEARCH QUESTIONS

1. Among individuals in the time prior to vaccines becoming available, are there higher rates (expressed as hazard ratios with time since COVID-19 diagnosis) of an incident outcome in those with a COVID-19 diagnosis compared to those before or without a COVID-19 diagnosis, before and after adjustment for potential confounders?

2. Among vaccinated individuals in the era of the delta variant of SARS-CoV-2, are there higher rates (expressed as hazard ratios with time since COVID-19 diagnosis) of an incident outcome in those with a COVID-19 diagnosis compared to those before or without a COVID-19 diagnosis, before and after adjustment for potential confounders?

3. Among unvaccinated individuals (i.e., individuals eligible for vaccination that have chosen not to receive it) in the era of the delta variant of SARS-CoV-2, are there higher rates (expressed as hazard ratios with time since COVID-19 diagnosis) of an incident outcome in those with a COVID-19 diagnosis compared to those before or without a COVID-19 diagnosis, before and after adjustment for potential confounders?

DATA SOURCES

This research will be conducted using OpenSafely and requires the following data sources:

- Primary care data (TPP)
- Second Generation Surveillance System (SGSS) for Pillar 1 and Pillar 2 SARS-COV-2 infection laboratory testing data
- Secondary Uses Service (SUS)
- Office of National Statistics (ONS) death registry
- Index of Multiple Deprivation (IMD)

STUDY POPULATION

Patients will be included if they meet ALL the following criteria:

- Alive on the cohort start date
- Known age between 18 and 110 inclusive on the cohort start date
- Known sex
- Known deprivation
- Known region
- Registered in an English GP with TPP software for at least 6 months prior to the study start date

Additional criteria may be applied for certain outcomes and are summarized in the outcome specific documents (see: ‘outcomes and potential confounders’ section for links).

QUALITY ASSURANCE

We will ensure data quality by applying the following quality assurance rules:

1. Remove individuals who are missing year of birth
2. Remove individuals whose year of birth is after their year of death
3. Remove individuals whose year of birth is after today
4. Remove individuals whose date of death is after today
5. Remove men whose records contain pregnancy and/or birth codes
6. Remove men whose records contain HRT or COCP medication codes
7. Remove women whose records contain prostate cancer codes

COHORTS

	Cohort 1: Pre-vaccination	***Cohort 2: Vaccinated***	***Cohort 3: Unvaccinated***
Start date	01/01/2020, which is the approximate start date of the pandemic in the UK.	01/06/2021, which is the date that the delta variant was thought to be ubiquitous in England.	01/06/2021, which is the date that the delta variant was thought to be ubiquitous in England.
End date - exposure	18/06/2021, which is the date when the Joint Committee for Vaccination and Immunisation (JCVI) phase 2, group 12 (all adults aged 18 years and older) become eligible for a COVID-19 vaccination.	14/12/2021, which is the day that the UK Health Security Agency stated that over half of English cases they sampled have S Gene Target Failure, meaning they were likely Omicron, in this report.	14/12/2021, which is the day that the UK Health Security Agency stated that over half of English cases they sampled have S Gene Target Failure, meaning they were likely Omicron, in this report.
End date - outcome	14/12/2021, which is the day that the UK Health Security Agency stated that over half of English cases they sampled have S Gene Target Failure, meaning they were likely Omicron, in this report.	14/12/2021, which is the day that the UK Health Security Agency stated that over half of English cases they sampled have S Gene Target Failure, meaning they were likely Omicron, in this report.	14/12/2021, which is the day that the UK Health Security Agency stated that over half of English cases they sampled have S Gene Target Failure, meaning they were likely Omicron, in this report.
Exclusion criteria	Patients will be excluded if they meet any of the following criteria: • COVID-19 diagnosis recorded prior to their index date	Patients will be excluded if they meet any of the following criteria: • COVID-19 diagnosis recorded prior to their index date [Note: these individuals are required for a sensitivity analysis and so should not be removed at the data extraction stage]	Patients will be excluded if they meet any of the following criteria: • COVID-19 diagnosis recorded prior to their index date [Note: these individuals are required for a sensitivity analysis and so should not be removed at the data extraction stage]

		• They do not have a record of two vaccination doses prior to the study end date • They received a vaccination prior to 08-12-2020 (i.e., the start of the vaccination program) • They received a second dose vaccination before their first dose vaccination • They received a second dose vaccination less than three weeks after their first dose • They received mixed vaccine products before 07-05-2021	• They have a record of one or more vaccination doses prior to their index date • They could not be assigned to a vaccination group as defined by the Joint Committee on Vaccination and Immunisation (JCVI)
Follow-up start (i.e., an individual's index date)	Study start date.	Follow-up will start at the latest of the following dates: • Two weeks after their second vaccination • Study start date	Follow-up will start at the latest of the following dates: • 12 weeks after they became eligible for vaccination • Study start date
Follow-up end for exposure	Follow-up will end at the earliest of the following dates: • Death • Outcome event • Study end date exposure • Deregistration date • Vaccination • Date when eligible for vaccination according to	Follow-up will end at the earliest of the following dates: • Death • Outcome event • Study end date exposure • Deregistration date	Follow-up will end at the earliest of the following dates: • Death • Outcome event • Study end date exposure • Deregistration date • Vaccination

	JCVIs priority groupings		
Follow-up end for outcomes	Follow-up will end at the earliest of the following dates: • Death • Outcome event • Study end date outcome • Deregistration date	Follow-up will end at the earliest of the following dates: • Death • Outcome event • Study end date outcome • Deregistration date	Follow-up will end at the earliest of the following dates: • Death • Outcome event • Study end date outcome • Deregistration date
Cox regression time periods, full	[0], [1,7), [7,14), [14,28), [28,56), [56,84), [84,197), [197, 365), [365,714)	[0], [1,7), [7,14), [14,28), [28,56), [56,84), [84,197)	[0], [1,7), [7,14), [14,28), [28,56), [56,84), [84,197)
Cox regression time periods, collapsed	[0], [1,28), [28,197), [197, 365), [365,714)	[0], [1,28), [28,197)	[0], [1,28), [28,197)

EXPOSURES

COVID-19 diagnosis

Exposure will be defined as the first date of a COVID-19 diagnosis post index date. Exposures can be recorded in any of the following data sources:

Data source	Definition
SGSS	Date of positive SARS-COV-2 PCR or antigen test
Primary care	Date of confirmed diagnosis code
SUS	Start date of episode with COVID-19 diagnosis in any position
ONS death registry	Date of death with COVID-19 listed as primary or underlying cause

COVID-19 severity

Individuals with a hospital admission record that includes a COVID-19 diagnosis in the primary position within 28 days of first COVID-19 diagnosis will be defined as 'COVID-19 diagnosis with hospitalisation'. All other individuals will be defined as 'COVID-19 diagnosis without hospitalisation'.

OUTCOMES

Outcomes can be recorded in any of the following data sources:

Data source	Definition
Primary care	Date of diagnosis or prescription code
SUS	Start date of episode with confirmed diagnosis in any position
ONS death registry	Date of death with diagnosis listed as primary or underlying cause

Details of outcomes are provided in the following outcome specific documents:

- Cardiovascular
- Mental health
- Diabetes
- Neurodegenerative disease
- Autoimmune disorders
- Gastrointestinal disorders
- Respiratory

POTENTIAL CONFOUNDERS

The common potential confounders are as follows:

Covariate	Type	Definition
Age	Continuous	Modelled as age in years using a restricted cubic spline with 3 knots at the 10th, 50th and 90th percentiles
Sex	Categorical	Male, Female
Ethnicity	Categorical	1: White 2: Mixed 3: South Asian 4: Black 5: Other
Deprivation	Categorical	10 categories from Index of Multiple Deprivation 2019
Region	Categorical	East of England London Midlands North East and Yorkshire North West South East South West Scotland Wales
Smoking status	Categorical	E: Ever smoker M: Missing N: Never smoker S: Current smoker
Care home status	Binary	1 if care home resident; 0 otherwise
Consultation rate	Continuous	Number of GP consultations 12 months prior to the start of the study
Health care worker	Binary	1 if healthcare worker; 0 otherwise
Dementia	Binary	1 if diagnosis present; 0 otherwise.
Liver disease	Binary	1 if diagnosis present; 0 otherwise.
Chronic kidney disease	Binary	1 if diagnosis present; 0 otherwise.
Cancer	Categorical	1 if diagnosis present; 0 otherwise.
Hypertension	Binary	1 if diagnosis present; 0 otherwise.
Diabetes	Binary	1 if diagnosis present; 0 otherwise.
Obesity	Binary	1 if BMI>=30 or coded diagnosis for obesity;

		0 otherwise.
Chronic obstructive pulmonary disease (COPD)	Binary	1 if diagnosis present; 0 otherwise.
Acute myocardial infarction	Binary	1 if diagnosis present; 0 otherwise.
Ischaemic stroke	Binary	1 if diagnosis present; 0 otherwise.

Details of additional potential confounders are provided in the following outcome specific documents.

Covariates will be checked prior to the analysis and the following rules applied to ensure the models run:

- Remove binary or categorical variables if any level contains <=2 individuals with both the exposure and the outcome
- If the covariate ‘smoking status’ is required for the analysis but would be removed due to low numbers, merge ‘Ever smoker’ and ‘Current Smoker’ into a single ‘Ever smoker’ category so that the variable is ever/never rather than ever/never/current.
- If the covariate ‘deprivation’ is required for the analysis but would be removed due to low numbers, merge the deciles in quintiles – i.e., 1-2, 3-4, 5-6, 7-8, 9-10.

MAIN ANALYSES

Descriptive statistics

Initial descriptive statistics will be used to describe the demographic and clinical characteristics of the baseline cohort, overall and for the subgroups hospitalised and non-hospitalised with COVID-19 diagnosis.

Cox regression

We will split follow up time for each person into periods before and after COVID-19 diagnosis, and into time periods since diagnosis defined in days using the time periods specified for each cohort above. We will tabulate numbers of outcome events, person-years of follow-up and rates of events before and with time since exposure. If any of these time periods contain no events, we will collapse the time periods after COVID-19 diagnosis into the collapsed time periods specified for each cohort above prior to analysis.

We will fit Cox regression models with calendar time scale using the start of study date as the origin. This will ensure that all analyses account for changes with calendar time in rates of the outcome event. Using this approach, we will estimate hazard ratios for events of different types before and after exposure, and by time since exposure.

For computational efficiency, we will use a sampling procedure for datasets containing more than 4,000,000 individuals. For these datasets, we will include all people with the outcome event (i.e., the cases), all people with the exposure, and a random subset of non-case-non-exposed individuals as per the table below. Analyses will incorporate inverse probability weights for data from the non-case-non-exposed individuals. For example, consider a sample of N people, X of whom are cases. We will choose the number of non-case-non-exposed individuals per case, Y, based on the number of cases. We will then sample Y*X people non-case-non-exposed individuals and assign a weight of (N-X)/(Y*X) to each of them and 1 to each case and each exposed individual. Confidence intervals will be derived using robust standard errors when sampling has occurred. [Agreed 24/08/2022]

Number of cases, X	Number of non-case-non-exposed individuals per case, Y
X < 100,000	20
100,000 <= X < 500,000	10
X >= 500,000	5

Potential confounders (see: **Error! Reference source not found.**) will be based on data recorded on or before the start of follow-up in each analysis. We will exclude potential confounders from any analysis when there are ≤2 disease events at any level. All models will be stratified by region so that risk sets are constructed within region, hence accounting for between-region variation in the baseline hazard.

We will estimate: (i) age and sex adjusted and (ii) maximally adjusted HRs. We will examine the fit of the restricted cubic splines used for age.

We will analyse outcomes for which there are at least 50 events after exposure. This is an arbitrary threshold chosen on the basis that outcomes which are this rare in such a large sample are unlikely to have population level impact. We will apply the same criterion to subgroup analyses.

Absolute excess risk

The absolute excess risk analysis is performed for each outcome in each cohort. To compare the outcomes across the cohorts, each of which have different lengths of follow-up, we will calculate the absolute excess risk at 196 days.

This analysis requires the following summary statistics for eight age/sex groups (female_18_39; female_40_59; female_60_79; female_80_110; male_18_39; male_40_59; male_60_79; male_80_110):

- Number of unexposed person days,
- Number of unexposed outcome events
- Total number of people exposed
- Sample size

Plus, the maximally adjusted HR from the main model.

Create an empty life table with one row per day from 0 to 196. The life table is then constructed as follows:

1. Calculate the average daily incidence of the outcome in the unexposed by age/sex group:
   incidence_unexp = unexposed_events/unexposed_person_days
2. Calculate cumulative risk over time in the unexposed by age/sex group:
   cumulative_survival_unexp = cumprod(1 - incidence_unexp)
3. Label each day with the relevant HR (e.g., days 0 to 27 will have the coefficient for the term ‘days0_28’, while days 28 to 196 will have the coefficient for the term ‘days28_197’).
4. Predict the expected cumulative survival in the exposed by age/sex group by multiplying the daily incidence in the unexposed (from step 1) by the HR (from step 4):
   cumulative_survival_exp = cumprod(1 - (hr * incidence_unexp))
5. Calculate the daily excess risk as the difference in cumulative survival for the unexposed (from step 2) and the expected cumulative survival in the exposed (from step 4):
   cumulative_difference_absolute_excess_risk = cumulative_survival_unexp - cumulative_survival_exp

An example life table can be found here: AER example calculation.xlsx

The overall absolute excess risk will be estimated using a weighted sum of the age- and sex-specific excess risks, weighted by the proportions of individuals in age and sex strata in the pre-vaccination cohort. Ultimately, total excess events, total post exposure follow-up (years) and excess events per 100 000 Covid-19 diagnosis will be reported for all cohorts at 196 days to allow comparison between the cohorts.

SUBGROUP ANALYSES

We will repeat the main analysis to estimate post-exposure hazard ratios for the following subgroups unless specified otherwise in the outcome specific documents:

- Subgroups according to severity (hospitalised / non-hospitalised)
- Subgroups according to age group (18-39 / 40-59 / 60-79 / 80-110)
- Subgroups according to sex (male / female)
- Subgroups according to ethnicity (White / Asian or Asian British / Black or Black British / Mixed / Other Ethnic Groups)
- Subgroups according to prior history of outcome subcategory (prior history of outcome subcategory / no prior history of outcome subcategory)

For the subgroups according to age group (18-39 / 40-59 / 60-79 / 80-110), we will include age and age squared as covariates in place of the cubic restricted spline for age.

For the subgroups according to severity (hospitalised / non-hospitalised), include all individuals who were exposed (hospitalised) plus 20 unexposed (non-hospitalised - i.e., those who either never had a COVID-19 diagnosis or had a COVID-19 diagnosis but were not hospitalised) controls per every 1 case. Also, merge the regions 'London' and 'South East' into a single region called 'South East, including London' and the regions 'East Midlands' and 'West Midlands' into a single region called 'Midlands'. [Agreed 31/05/2022; JS, VW]

Outcome specific subgroup analyses are detailed in the outcome specific documents, as needed.

SENSITIVITY ANALYSES

Prior infection analysis

We will repeat the main analyses in individuals who had a COVID-19 diagnosis prior to the start of the study.

Outcome specific sensitivity analyses

Outcome specific sensitivity analyses are detailed in the outcome specific documents, as needed.

Day zero analysis

We will repeat the main analyses splitting the time period [0,7) into [0,1) and [1,7) if the full time periods were used and, if possible, splitting the time period [0,28) into [0,1) and [1,28) if the reduced time periods were used.

MISSING DATA

Individuals with missing age, sex, or deprivation are excluded from the analysis by the study definition. We will include a missing category for ethnicity. All other covariates are defined using the presence versus absence of specific codes, so have no identifiable missing values. We will not use multiple imputation.