UK Biobank

COVID-19 re-imaging study: Selection and case-control matching

---

Version 3.0

www.ukbiobank.ac.uk
September 2023

1. Study background

The COVID-19 re-imaging study was set up to investigate the potential effects of SARS-CoV-2 infection on internal organs by comparing imaging scans taken from participants both before and after infection. Between 2014 and 2021, approximately 49,000 UK Biobank participants had attended an imaging assessment in one of four UK Biobank centres, based in Cheadle (Stockport), Newcastle, Reading and Bristol. About 3,200 participants had also attended a repeat imaging visit.

The pandemic forced a temporary closure of all imaging clinics, but in February 2021 they re-opened to perform repeat scans for the purposes of COVID-19 research. The study aimed to invite about 2,000 of the 49,000 participants who had previously attended an imaging assessment for a repeat assessment. Half of the participants had been previously infected with SARS-COV-2 (designated as 'cases') and half had not ('controls'). Identification of previous infection was based on data from linked PCR antigen (swab) tests, clinical codes identified in the electronic health record data or home-based lateral flow SARS-CoV-2 antibody tests sent to participants.

The eligibility criteria for this study, including the details of the case-control matching, are described below.

2. Sample selection

UK Biobank participants were eligible for selection if they met all of the following criteria:

- attended an imaging assessment at either Cheadle, Newcastle or Reading before the pandemic and had not already attended a repeat assessment (the Bristol site had just opened prior to the pandemic and there were too few numbers to invite back for a repeat assessment);
- still lived within the catchment area of the clinic they attended for their first imaging assessment, although this criteria was removed in May 2021 with the easing of travel restrictions;
- no incidental findings were identified from their scans taken at the first imaging assessment;
- had not withdrawn or died;
- had a valid email and postal address;
- high-quality scans were obtained from the first imaging assessment (using data field 25733[1] as a proxy);

[1] There are a small number of participants in the study who will not have data in this field, as this eligibility criterion was introduced shortly after this study started.

- Lived within 60km of the clinic (extended to 75km in Feb 2021), although this criteria was removed in May 2021 with the easing of the travel restrictions.

Due to national restrictions caused by the pandemic in early 2021, selection was additionally restricted to those not living with another UK Biobank participant who was also eligible for the study, in order to maintain social distancing measures between all participants at the clinic. However, this restriction was removed in May 2021, in accordance with the easing of the lockdown, and those who lived with another study member were invited to participate.

3. Case identification

There are two routes that participants were identified as having been infected with SARS-CoV-2.

1. A record of a positive SARS-CoV-2 PCR antigen test result, as identified through record linkage to these datasets in England, Wales and Scotland, or a clinical code of confirmed infection with SARS-CoV-2 among hospital inpatient data (ICD10 code U07.1) or primary care data (CTV3 codes in TPP: XaLTE, Y20d1, Y213a, Y228d, Y22b8, and Y23f7; SNOMED-CT codes in EMIS: 1240581000000104, 1300721000000109, 1321541000000108, 1321551000000106, and 1321661000000108; EMIS local code EMISNQCO303);

2. A record of a positive SARS-CoV-2 antibody test result obtained from a home-based lateral flow test (LFT) kit sent to all participants who agreed to receive one. For those who had reported they had not yet been vaccinated, a second kit was sent to all participants who recorded an initial positive result in order to reduce the number of false-positives. From July 2021 onwards, case identification was relaxed to having a positive result from the first kit only, given emerging evidence of the low rate of false-positives. For those who had had at least one dose of the vaccine and initially tested positive, a further sample was collected and sent to the Thriva lab to determine whether the antibodies were specifically a result of previous infection (or vaccination).

The source of the positive COVID-19 test result is given in data field 41001, using encoding 1400.

NB. In 2023, when this field was updated for the last time, some identified cases with multiple sources of positive COVID-19 test results had LFT or a PCR source value removed. The original assignment of their values was made before the LFT and PCR data were fully processed and made available for general release via Showcase. Following this data cleaning exercise, it was noted that the test dates for a small number of cases occurred after a self-reported vaccination and/or imaging assessment date. Importantly, this does not change their case status given the existence of other sources of a positive COVID-19 result.

4. Control matching

Participants were eligible to enter the study as control participants if they met the main selection criteria (as above), and they had no record of confirmed or *suspected* COVID-19 in their linked health records. One control was randomly assigned to each case, matching on the following criteria:

- gender
- ethnicity (white; non-white)
- date of birth (+/- 6 months)
- location of baseline imaging assessment clinic
- date of baseline imaging assessment (+/- 6 months)

5. Case-control status

Any participant who has attended their repeat imaging assessment as part of the COVID-19 study will have a case-control status in data field 41000. Date of attendance and imaging site is recorded under Instance 3 of data field 53 and data field 54, respectively. The numeric value given to each case-control participant in data field 41000 comprises the unique number assigned to a case-control set followed by the last digit denoting the case-control status of that individual, where '0' is a control and '1' is a case. For example, in a case-control set with a unique identifier of 324, the control participant would have a value in data field 41000 of '3240' and the corresponding matched case would have the value '3241'.

Field 41000 was updated regularly while the study was progressing and there were incomplete case-control sets appearing in the data at any one time, i.e. where only one case or control in any given set had attended their appointment. If a missing case or control chose to decline their imaging invitation, or if a control became a case in the time interval between being assigned to a matched case-control set and attending their imaging appointment, then the member of a set who had already attended was re-matched to a different case or control whenever possible. In this circumstance, a new set ID in data field 41000 was provided for that particular case/control participant. Now that the study is completed, the cases and controls without a corresponding matched pair have been re-coded in Data field 41000 as -998 and -999 for controls and cases, respectively. If researchers wish to maximise the number of participants in their analyses, it might be worth considering breaking the case-control matches, and instead adjusting for the matching variables in their analytical model.

There are also a small number of cases and controls who had their status changed in data field 41000 as the study progressed, following the emergence of new linkage data or SARS-CoV-2 antibody test result data, e.g. they tested positive for COVID-19 prior to their appointment at the imaging clinic (or SARS-CoV-2 antibody test results showed they tested positive up to two weeks after their appointment).