LongCOVID Phenotype Definitions for Analyses

**COVID-19 Host Genetics Initiative (HGI)**

25 October 2021, updated 9 February 2022, v1.1

[**Working group**](#_heading=h.gjdgxs)

[**Study Rationale**](#_heading=h.30j0zll)

[**Phenotype Inclusivity**](#_heading=h.1fob9te)

[**Phenotype Notation**](#_heading=h.3znysh7)

[**Meta-analyses**](#_heading=h.2et92p0)

[**Phenotype definitions**](#_heading=h.tyjcwt)

[**Wide Questionnaire-based Phenotypes (WQ)**](#_heading=h.3dy6vkm)

[Analysis WQ1.3: Long COVID vs. No Long COVID, within COVID+](#_heading=h.1t3h5sf)

[Analysis WQ2.3: Long COVID vs. Population Control](#_heading=h.4d34og8)

[**Narrow Questionnaire-based Phenotypes (NQ)**](#_heading=h.2s8eyo1)

[Analysis NQ1.3: Long COVID vs. No Long COVID, within COVID+](#_heading=h.17dp8vu)

[Analysis NQ2.3: Long COVID vs. Population Control](#_heading=h.3rdcrjn)

[**Wide Electronic Health Record (EHR)-based Phenotypes (WE)**](#_heading=h.26in1rg)

[**Analysis WE1.3: Long COVID vs. No Long COVID, within COVID+**](#_heading=h.lnxbz9)

[**Analysis WE2.3: Long COVID vs. Population Control**](#_heading=h.35nkun2)

[**Narrow Electronic Health Record (EHR)-based Phenotypes (NE)**](#_heading=h.1ksv4uv)

[**Analysis NE1.3: Long COVID vs. No Long COVID, within COVID+**](#_heading=h.44sinio)

[**Analysis NE2.3: Long COVID vs. Population Control**](#_heading=h.2jxsxqh)

[**COVID-19 diagnoses**](#_heading=h.z337ya)

[Questionnaire](#_heading=h.3j2qqm3)

[EHR](#_heading=h.1y810tw)

[**Long COVID diagnoses**](#_heading=h.4i7ojhp)

[**Specific diagnosis codes for long COVID**](#_heading=h.2xcytpi)

[**Long COVID-19 symptoms**](#_heading=h.1ci93xb)

[Questionnaire](#_heading=h.3whwml4)

[EHR](#_heading=h.2bn6wsx)

[**Pre-existing symptoms**](#_heading=h.qsh70q)

[**Data from various timepoints**](#_heading=h.xqp0jiah8qt)

[Concerning patients who died during follow-up](#_heading=h.awx3ztdlqz5u)

[**New phenotype definitions for symptom clusters**](#_heading=h.omov32ms0g1o)

# Working group

(in alphabetical order)

**Beatriz Cortés, MSc**: population genetics, bioinformatics

**Brent Richards, MD, MSc**: endocrinology, clinical experience of COVID, genetic epidemiology

**Chiara Gabbi, MD, PhD:** internal medicine physician, clinical/translational scientist

**Dylan Williams**, **PhD**; neuro-epidemiology, gerontology, Mendelian randomisation & gen epi

**Emma Duncan,**

**Guan Wang**, **PhD**: population genetics, bioinformatics

**Hanna Ollila, PhD**: immune genetics, statistical genetics, sleep disorders

**Louise Wain, PhD**, genetic epidemiology, respiratory disease, post-hospitalised COVID study co-lead

**Michael Hultström, MD, PhD,** critical care, genomics, clinical experience of COVID

**Minttu Marttila, PhD:** muscle disorders, genetics, pathogenetic mechanisms

**Nasa Sinnott-Armstrong, ScM**: statistical genetics, cardiometabolic & pulmonary disease, gene-environment interactions

**Alessandra Renieri, MD PhD**

Clinical and Molecular Geneticist

**Samuel Jones, PhD**: statistical genetics, genetic epidemiology, sleep and circadian genetics

**Tomoko Nakanishi, MD:** respiratory disorders, clinical expertise, human genetics

**Vilma Lammi, PhD**: sleep disorders, human genetics, coordinating the Long COVID-19 Host Genetics Initiative collaboration

# Study Rationale

COVID-19 continues to place a huge burden on healthcare systems worldwide. It is becoming apparent that a significant proportion of COVID-19 patients report ongoing symptoms long after the SARS-CoV-2 virus is no longer detectable in the naso-pharingeal district. It is currently unknown why these individuals develop post-viral symptoms, although this phenomenon is not unique to SARS-CoV-2; and data on these individuals will help elucidate the mechanisms of so-called “Long COVID”.

Through large-scale meta-analyses involving over a hundred studies, the COVID-19 Human Genetics Initiative (HGI) identified genetic variants associated with both COVID-19 susceptibility and severity. We propose to extend this effort in an attempt to understand the biology of Long COVID and the factors influencing risk of developing this condition. As with the original HGI effort, analyses will be simple and scalable. Initially, we will focus on simple definitions and aim to maximise statistical power. As time goes on and more data are collected, additional and more specific Long COVID phenotypes will be defined.

# Phenotype Inclusivity

Long COVID is very difficult to define (and thus diagnose), partly because of the subjective nature of many symptoms. Many affected individuals are also missing the “official” COVID-19 diagnosis records required to be included as a long COVID patient. Given that these individuals are a key source of data and that we rely on ongoing support of Long COVID support groups, we should aim to include such individuals in the broader definitions of questionnaire-based phenotypes.

# Phenotype Notation

Each phenotype will be assigned a letter and numbers so that they can be distinguished from one another. To avoid ambiguity, we propose that phenotypes are assigned a letter based on whether they are based on questionnaires (Q) or Electronic Health Records, EHR (E). Additional letters may be assigned later, for example in the case of hybrid questionnaire-EHR phenotypes (e.g. H).   
Phenotypes with more narrow (strict) inclusion criteria will be marked with N, and wider ones allowing e.g. for self-report COVID-19 infection with W.

The analysis will then be given a number (1 or 2) which corresponds to the study design (practically the definition of control individuals used, see Figure 1).

1. Cases: Long COVID; Controls: Had COVID-19 but not long COVID (=recovered from COVID).
2. Cases: Long COVID; Controls: No long COVID = Population control = Everyone not a case.

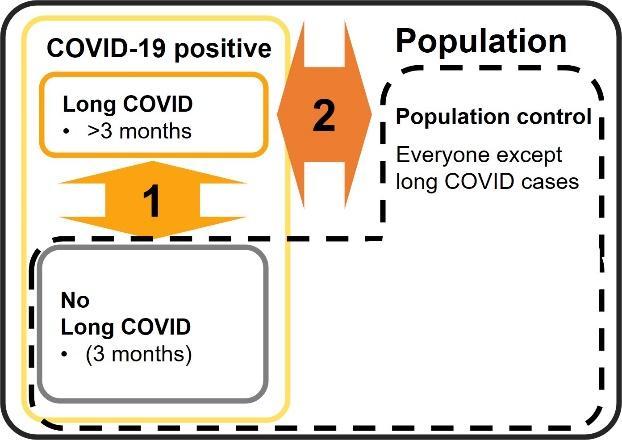


Figure 1. Analyses comparing Long COVID cases to either COVID those who have recovered from COVID within 3 months (1), or the population (2).

Finally, as the phenotypes are defined using length of symptoms after infection (default 3 months/12 weeks) and there are plans to investigate over shorter and longer periods, the phenotype label is also given a number indicating this duration in months. The default phenotype will be defined using a three month cut-off, starting from the onset of COVID (positive SARS-CoV-2 test result / date of COVID diagnosis / onset of COVID symptoms).

The phenotype label will take the (regexp) format [NW][QE][12].[1-9][0-9]\*. For example, “WE1.3” would be a “wide” EHR-based definition with Long COVID cases and controls that have had COVID-19 but not Long COVID, with Long COVID being defined as ongoing symptoms 3 months after COVID infection.

# Meta-analyses

Meta-analyses will be run using summary statistics from GWAS analyses from the individual studies. Not all studies have data for all of the phenotypes but each meta-analysis will comprise of those studies that have run that particular GWAS and gone through the quality control.

The GWAS results using questionnaire (Q) and EHR (E) based phenotypes will be combined in the meta-analysing phase (see figure 2). Thus, no same individuals should be included in the Q and E analyses (concerning the cohorts/studies that have both questionnaire and EHR data on their participants).

The narrow (N) phenotypes from Q and E will be meta-analysed together. Similarly, the wide (W) phenotypes will be meta-analysed together, separately from the narrow phenotypes. Thus, individuals meeting the narrow criteria should also be included in the wide analysis.

Meta-analyses for control definition 1, within COVID cases, will be run together, separately from analyses 2 that use a population control.

Phenotype definitions using different cutoffs for time since COVID-19 diagnosis will be introduced as the project progresses, and these will then be meta-analysed separately. The primary phenotypes use a 3-month cutoff.

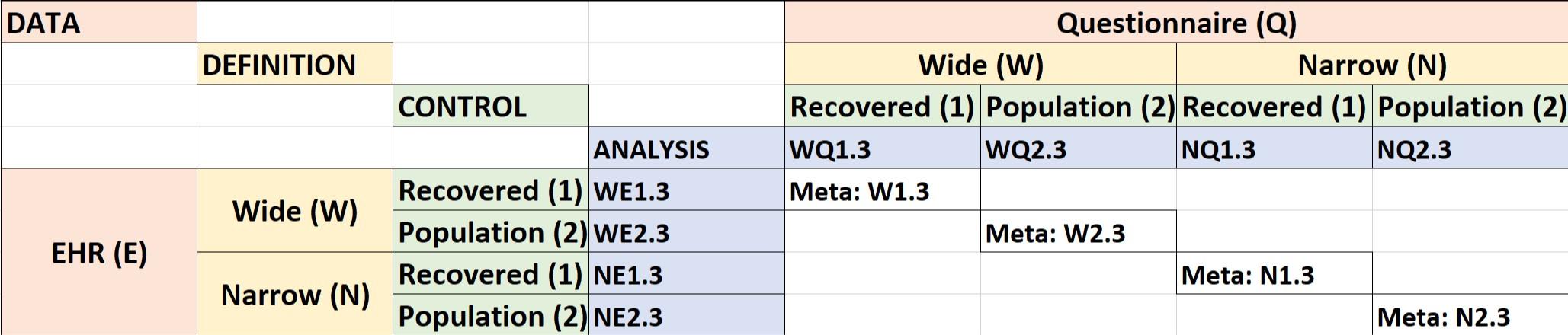


Fig 2. Meta-analyses combining questionnaire and EHR-based data.

# Phenotype definitions

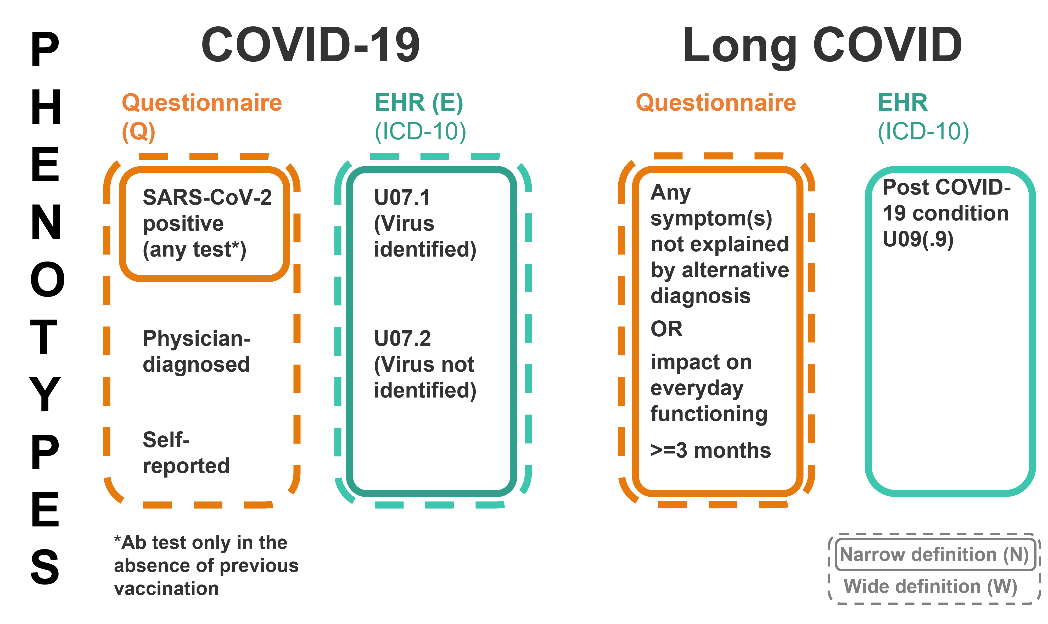


Figure 3. An overview for the definitions of COVID and Long COVID in our narrow and wide phenotypes.

## Wide Questionnaire-based Phenotypes (WQ)

### Analysis WQ1.3: Long COVID vs. No Long COVID, within COVID+

* **Cases**: {Positive SARS-CoV-2 test[[1]](#footnote-1) OR COVID-19 diagnosis OR self-reported suspected COVID-19 with no confirmation} AND {self-reported symptom(s)[[2]](#footnote-2) that cannot be explained by an alternative diagnosis at least 3 months after COVID onset OR report of significant impact in day-to-day life at least 3 months after COVID onset}.
* **Controls**: {Positive SARS-CoV-2 test1 OR self-reported COVID-19 diagnosis OR self-report suspected COVID-19 with no confirmation} AND not defined as a Long COVID case (with this WQ definition).

### Analysis WQ2.3: Long COVID vs. Population Control

* **Cases:** {Positive SARS-CoV-2 test1 OR COVID-19 diagnosis OR self-reported suspected COVID-19 with no confirmation} AND {self-reported symptom(s)2 that cannot be explained by an alternative diagnosis at least 3 months after COVID onset OR report of significant impact on day-to-day life at least 3 months after COVID onset}.
* **Controls:** Everyone not defined as a Long COVID case (with this WQ definition).

## Narrow Questionnaire-based Phenotypes (NQ)

### Analysis NQ1.3: Long COVID vs. No Long COVID, within COVID+

* **Cases**: {Positive SARS-CoV-2 test1} AND {self-reported symptom(s)2 that cannot be explained by an alternative diagnosis at least 3 months after COVID onset OR report of significant ongoing impact in day-to-day life at least 3 months after COVID onset}.
* **Controls**: {Positive SARS-CoV-2 test1} AND not defined as a Long COVID case (with this NQ definition).

### Analysis NQ2.3: Long COVID vs. Population Control

* **Cases:** {Positive SARS-CoV-2 test1} AND {self-reported symptom(s)2 that cannot be explained by an alternative diagnosis at least 3 months after COVID onset OR report of significant ongoing impact in day-to-day life at least 3 months after COVID onset}.
* **Controls:** Everyone not defined as a Long COVID case (with this NQ definition).

## Wide Electronic Health Record (EHR)-based Phenotypes (WE)

### Analysis WE1.3: Long COVID vs. No Long COVID, within COVID+

* **Cases:** COVID-19 positive diagnosis with or without identification of SARS-CoV-2[[3]](#footnote-3) AND non pre-existing[[4]](#footnote-4) symptom(s)[[5]](#footnote-5) that cannot be explained by an alternative diagnosis, with a diagnosis date between three months to 1 year after COVID-19 diagnosis date[[6]](#footnote-6).
* **Controls**: COVID-19 positive diagnosis with or without identification of SARS-CoV-23 AND not defined as a Long COVID case (with this WE definition).

### Analysis WE2.3: Long COVID vs. Population Control

* **Cases**: COVID-19 positive diagnosis with or without identification of SARS-CoV-23 AND non pre-existing4 symptom(s)5 that cannot be explained by an alternative diagnosis, with a diagnosis date between three months to 1 year after COVID-19 diagnosis date6.
* **Controls**: Everyone not defined as a Long COVID case (with this WE definition).

## Narrow Electronic Health Record (EHR)-based Phenotypes (NE)

### Analysis NE1.3: Long COVID vs. No Long COVID, within COVID+

* **Cases:** a specific long COVID diagnosis[[7]](#footnote-7).
* **Controls**: COVID-19 positive diagnosis AND not defined as a Long COVID case (with this NE definition and 3 months of follow-up time in the EHR data after the COVID diagnosis).

### Analysis NE2.3: Long COVID vs. Population Control

* **Cases**: a specific long COVID diagnosis6.
* **Controls**: Everyone not defined as a Long COVID case (with this NE definition).

## COVID-19 diagnoses

### Questionnaire

For the wide questionnaire definition, participants can be considered to have had COVID-19 if they

1. **have had a positive SARS-CoV-2 test result (see below)**,
2. report they have had COVID diagnosed by a clinician but it has not been verified by any positive test result, or
3. strongly suspect that they have had COVID-19 but did not get tested or did not receive a positive test result

For the narrow questionnaire definition, only the first of the above criteria (in bold) is considered.

Positive test result refers to any one of the following testing methods, whether done in clinical laboratory settings or by the patient:

* PCR test
* Antigen test
* Lateral flow test (LFAT = lateral flow antigen test)
* Antibody test - only to be considered in the absence of previous vaccination
* Next generation sequencing (NGS) test
* Nucleic acid amplification test (NAAT)

### EHR

For the wide EHR definition, COVID-19 diagnosis can be any one of the following diagnosis codes:

* ICD-10: **U07.1**, U07.2
* ICD-11: **RA01.0**, RA01.1
* SNOMED: [**840533007**](http://snomed.info/id/840533007), [**840536004**](http://snomed.info/id/840536004), [**840539006**](http://snomed.info/id/840539006),[**871553007**](http://snomed.info/id/871553007), [**871555000**](http://snomed.info/id/871555000), [**871556004**](http://snomed.info/id/871556004), [**871557008**](http://snomed.info/id/871557008), [**871558003**](http://snomed.info/id/871558003), [**871559006**](http://snomed.info/id/871559006), [**871560001**](http://snomed.info/id/871560001)*,*[**871562009**](http://snomed.info/id/871562009)*,* [**1240411000000100**](http://snomed.info/id/1240411000000107)*,*[**1240581000000100**](http://snomed.info/id/1240581000000104)*,* [840544004](http://snomed.info/id/840544004),[1017214008](http://snomed.info/id/1017214008), [1119302008](http://snomed.info/id/1119302008)

## Long COVID diagnoses

### Specific diagnosis codes for long COVID

The following codes should be used to identify cases of long COVID for the narrow EHR-based phenotype(s):

* ICD-10: U09(.9) (‘Post COVID-19 condition’)ICD-11: RA02
* SNOMED: 1325161000000100, 1325181000000100, 1325021000000110, 1325031000000110, 1325041000000100, 1325051000000100, 1325061000000100, 1325071000000100, 1325081000000110, 1325091000000110, 1325101000000100, 1325121000000100, 1325131000000110, 1325141000000100, 1325151000000100
* TPP local codes: Y2b87, Y2b88, Y2b89, Y2b8a (see <https://biobank.ndph.ox.ac.uk/showcase/coding.cgi?id=8708&nl=1>)
* (If you wish other coding systems to be included, please contact us.)

## Long COVID-19 symptoms

### Questionnaire

The following is an incomplete list of COVID-19 symptoms that can be used to determine Long COVID. Any symptom or combination of symptoms captured by the questionnaires will be considered as a potential Long COVID symptom, given that it cannot be explained by an alternative diagnosis in the individual in question.

* Excessive fatigue or tiredness
* Fever or raised temperature
* Chills (feeling cold)
* Loss of or change in sense of smell
* Loss of or change in sense of taste
* Confusion, disorientation, dizziness or drowsiness
* Problems with memory and concentration ("brain fog")
* Sore (or painful) throat
* Hoarse voice
* Shortness of breath
* Breathlessness affecting normal activities
* Persistent cough
* Runny or blocked nose
* Reduced appetite
* Nausea and/or vomiting
* Abdominal pain
* Joint pain
* Diarrhoea
* Weight loss (otherwise unexplained)
* Abnormal muscle pains or aches
* Limbs feeling heavy
* Numbness or tingling in the body
* Headache (more often than normal)
* Chest pain
* Difficulty sleeping or insomnia
* Depression or anxiety

### EHR

A definitive list of symptoms and their codes does not yet exist. However, in addition to the [specific Long COVID diagnosis codes](#_heading=h.2xcytpi) listed above, we recommend initially considering the codes used in the algorithm developed by the 4CE

[<https://www.medrxiv.org/content/10.1101/2021.04.25.21255923v3.full-text>] and kindly provided by Shefali Verma of Penn Medicine. These codes are (ICD-10, *ICD-9*):

* Bluish lips/face: R23.0, *782.5*
* Breathing problems: R06.0, R06.00, R06.01, R06.02, R06.03, R06.09, *518.82*, *770.89*, *786.00*, *786.02*, *786.05*, *786.09*
* Chest pressure/pain: R07.1, R07.8, R07.81, R07.89, R07.9, *786.50*, *786.52*, *786.59*
* Chills: R50.9, R68.83, *780.60*, *780.64*
* Complete loss of smell (anosmia): R43.0, *781.1*
* Complete loss of taste (ageusia): R43.2, *781.1*
* Confusion or inability to rouse: R41.0, R46.4, *780.97*, *799.89*
* Coronavirus as the cause of other diseases: B97.2
* Cough: R04.2, R05, *786.2*, *786.30*
* Diarrhea (>= 3 loose/looser than normal stools in 24 hr period): R19.7, *787.91*
* Fatigue/lethargy: R53.81, R53.82, R53.83, *780.71*, *780.79*
* Fever > 100.4: R50.81, R50.9, *780.60*, *780.61*
* Headache: R51.9
* Mild conjunctivitis/red eye: H10\*, H11\*, H57.9, analogous ICD-9 (*372\**), *V41.1*
* Muscle pain: M79.1, *729.1, ​​M60.9, M79.6, M79.7*
* Nausea or vomiting: R11.0, R11.10, R11.11, R11.12, R11.2, *787.01*, *787.02*, *787.03*
* Other: R00.2, R09.89, R41.840, R42, M25.51, M35.81, F32.9, F41.9, *296.20*, *300.00*, *719.41*, *780.4*, *785.1*, *785.9*, *799.51*
* Partial loss of smell (partial anosmia): R43.1, R43.9, *781.1*
* Partial loss of taste (partial ageusia): R43.8, R43.9, *781.1*
* Runny nose: R09.81, R09.89, *478.19*, *785.9*
* Sore throat: J02.9, R07.0, R09.89, *462*, *784.1*, *785.9*
* Weight loss: R63.4, *783.21*
* Sequelae of other specified infectious and parasitic diseases: B94.8?

Short list of symptoms:

* Fatigue/lethargy: R53.81, R53.82, R53.83, 780.71, 780.79
* Brain fog (memory loss, inability to concentrate, sleep problems) 780.93, R41.840
* Muscle pain/myalgia and weakness: M60.9, M79.1, M79.6, M79.7, (729.1 ICD 9)
* Headache: R51.9
* Fever: R50.81, R50.9, 780.60, 780.61
* Cough: R04.2, R05, 786.2, 786.30
* Breathing problems/dyspnea: R06.00, R06.01, R06.02, R06.03, R06.09, 518.82, 770.89, 786.00, 786.02, 786.05, 786.09
* Chest pain R07
* Arrhythmia I49.9
* Loss of smell (anosmia): R43.0, 781.1 and loss of taste (ageusia): R43.2, 781.1

## Pre-existing symptoms

In the wide EHR-based phenotype, to minimise the misclassification of COVID symptoms that are unlikely to be caused by COVID-19 infection, we should, for each patient, rule out symptoms that precede the COVID infection or that the patient has previously been diagnosed with.

* We therefore propose that COVID-19 symptoms will only be considered if the symptom is not present for the specific individual in the year (365 days) preceding their first COVID-19 diagnosis date.

There will be no ideal way of characterising a symptom as pre-existing, but we aim to minimise the number of non-COVID related symptoms that are used to determine Long COVID status. This definition is currently open and we support discussion on improving this. Therefore, please either leave your comments here or raise this as a discussion point in the regular meetings.

## Data from various timepoints

If there is data available from several questionnaires in different timepoints, an individual defined as a Long COVID case using the criteria above, will remain a Long COVID case in the GWAS even if they have subsequently recovered based on a later questionnaire.

## Concerning patients who died during follow-up

1: Questionnaire based studies should include all patients who were alive to finish the questionnaire(s) during the follow-up after SARS-CoV-2 infection.

2: Electronic health record based studies should include all patients who received a Long-COVID diagnosis, irrespective of whether they died.

3: Each study should report the number of deaths among cases and controls for each analysis.

4: A sensitivity analysis should be performed with patients who died excluded before the final analysis before publication. The time for this will be communicated separately to reduce the total burden of analysis work.

## New phenotype definitions for symptom clusters

We will adopt the N2.3 definition for the duration of symptoms and COVID-19 diagnosis.

The number of both cases and controls need to be each ≥50.

For each symptom cluster, those with the following symptoms are defined as cases.

Controls could include those with any other symptoms which last ≥3 months after SARS-CoV-2 infection.

ANALYSIS: RESP

* shortness of breath
* cough

ANALYSIS: FATIGUE

* fatigue / malaise

ANALYSIS: ANOSMIA

* loss of taste / smell

ANALYSIS: NEURO

* brain fog
* problems with memory / concentration
* confusion / disorientation

ANALYSIS: GI

* nausea
* vomiting
* diarrhea

1. A list of acceptable COVID-19 diagnosis methods is provided in the [COVID-19 diagnoses](#_heading=h.z337ya) section. [↑](#footnote-ref-1)
2. Any symptom or symptoms, please see the [Long COVID-19 symptoms](#_heading=h.1ci93xb), [Questionnaire](#_heading=h.3whwml4) section. [↑](#footnote-ref-2)
3. A list of acceptable COVID-19 diagnosis methods is provided in the [COVID-19 diagnoses](#_heading=h.z337ya), [EHR](#_heading=h.1y810tw) section. [↑](#footnote-ref-3)
4. To determine whether a symptom is pre-existing, please see the [Pre-existing symptoms](#_heading=h.qsh70q) section. [↑](#footnote-ref-4)
5. For a list of symptoms/diagnoses, please see the [Long COVID-19 symptoms](#_heading=h.1ci93xb), [EHR](#_heading=h.2bn6wsx) section. [↑](#footnote-ref-5)
6. In case of re-infections, the period starts 3 months after the first COVID-19 diagnosis date and ends 1 year after the last COVID-19 diagnosis date. [↑](#footnote-ref-6)
7. A list of long COVID-19 diagnosis codes is provided in the [Specific diagnosis codes for long COVID](#_heading=h.2xcytpi) section. [↑](#footnote-ref-7)