Protocol & Statistical Analysis Plan

CCU079_01 - Risk of new diagnoses in secondary care following SARS-CoV-2 and other respiratory infections among school-aged individuals in England.

Authors:	Ms Eleanor Walsh (University of Bristol) Dr Katharine Looker (University of Bristol) Dr Alexia Sampri (University of Cambridge) Mr John Nolan (BHF Data Science Centre) Dr Rachel Denholm (University of Bristol)
Research team:	Professor Alastair Hay (University of Bristol) Professor Hannah Christensen (University of Bristol) Dr Patrick Nguipdop-Djomo (London School of Hygiene & Tropical Medicine [LSHTM]) Professor Punam Mangtani (LSHTM) Mr Elliot McClenaghan (LSHTM) Ms Alison Judd (Office for National Statistics [ONS]) Dr Ana Torralbo (University College London [UCL]) Professor Spiros Denaxas (UCL)
Advisory panel:	Dr Stefania Vergnano (University Hospitals Bristol & Weston NHS Foundation Trust) Professor Adam Finn (University of Bristol) Professor Caroline Relton (University of Bristol) Professor Sir Terence Stephenson (Great Ormond Street Institute of Child Health) Dr Michael Absoud (King's College London and Evelina London Children's Hospital) Dr Sunil Bhopal (Northumbria Healthcare NHS Foundation Trust) Professor Shamez Ladhani (St George's University Hospitals NHS Foundation Trust and UK Health Security Agency [UKHSA]) Dr Chris Bonell (LSHTM) Professor Charlotte Warren-Gash (LSHTM)

Version History	Date	Description
v0.1	08/04/2024	First draft – based on approval BHF DSC proposal v1
v0.2	17/07/2024	Refined title, background and aims relevant to analyses
v0.3	03/09/2024	Background, cohort specification and design
v1.0	01/10/2024	Reviewed by COVID-IMPACT consortium, ELUCIDate team and advisory panel
v1.1	10/12/2024	Responded to reviewer comments

Background

Children and young people are an often-overlooked group in research on SARS-CoV-2 infection, despite a substantial number experiencing long-COVID. Estimates of long-COVID incidence following SARS-CoV-2 infection in children and young people currently range from 1.8% (ZOE app study[1]) to 14% (Children & young people with Long Covid [CLoCk] study[2]). Generating more precise estimates is made difficult by lack of a clear clinical definition. The National Institute for Health and Care Excellence (NICE) defines long-COVID as, "signs and symptoms that continue or develop after acute COVID-19", encompassing both 'ongoing symptomatic COVID-19' (signs and symptoms 4-12 weeks after infection) and 'post-COVID-19 syndrome' (signs and symptoms >12 weeks after infection[3]). Health data science approaches, such as high-throughput phenotyping (an automated process useful for examining a large number of potential associations – in this instance, with a wide variety of possible diagnoses – within databases for millions of individuals), have identified a range of symptoms associated with long-COVID[4]. These are largely based on adult populations, raising questions about their applicability to children and young people. The range of symptoms linked to long-COVID is broad, with many non-specific to long-COVID (such as headache, nausea, and fatigue)[1]. Expanding analyses to consider other outcomes, such as any new diagnoses, and prescriptions, may be helpful in defining long-COVID, and understanding the prognosis for children with lingering health effects from SARS-CoV-2 infection.

The trajectory of long-COVID in children and young people, including subsequent diagnoses, and frequency and types of health service attendance, alongside associated risk factors (e.g., living in deprived areas and/or with pre-existing health conditions), has not been extensively studied to date. As a result, the information available to families, healthcare providers, schools, and the public is limited. This lack of data complicates service planning for NHS Integrated Care Boards (ICBs).

Most studies of long-COVID have focused on SARS-CoV-2 infections acquired during the early phases of the COVID-19 pandemic[5]. Yet evidence from France suggests that timing of SARS-CoV-2 infection is associated with incidence of long-COVID symptoms[6 7]. Therefore time-dependent factors, such as dominant circulating SARS-CoV-2 variant and population level of immunity (whether natural or vaccine-derived), could be expected to be important determinants of risk[8]. The contributions of SARS-CoV-2 variant as well as pre-existing immunity to long-COVID risk and its trajectory have important clinical implications. For instance, a large number of individuals who reported a SARS-CoV-2 infection early in the pandemic, continue to experience substantial long-term effects on health 12 months after infection[9].

In the post-COVID-19 testing era, acute cases of SARS-CoV-2 infection reported in healthcare records will be grouped with the wide array of other non-SARS-CoV-2 respiratory tract infections (RTIs) based on common presenting symptoms without identification of the causative pathogen. Thus, differences in trajectories within the group of "RTIs" according to their different infectious aetiologies are obscured. This is exacerbated by poor understanding of the long-term health outcomes of non-SARS-CoV-2 RTIs generally[10]. Evidence is therefore needed to enable clinicians to advise on the long-term prognosis for a child presenting either with (1) an acute RTI with unknown pathogen; (2) an acute RTI with identified pathogen; (3) long-COVID; or (4) ongoing health concerns following an RTI of unknown aetiology.

During the time when COVID-19 testing was routine, healthcare records of SARS-CoV-2 infection will include those with a positive SARS-CoV-2 test (whether from an LFT antigen test or PCR test) and clinical diagnoses made in healthcare settings without testing. This may lead to disproportionate inclusion of symptomatic cases tested and those reporting to care, over representing more severe

cases in healthcare records. However, asymptomatic or mildly-symptomatic infections are included in national COVID-19 surveillance records during the period of twice-weekly testing of secondary school students for SARS-CoV-2 infection via LFTs distributed through schools.

The first planned analysis is outlined here and will investigate the incidence of a range of diagnoses subsequent to SARS-CoV-2 infection and subsequent to non-SARS-CoV-2 RTIs in school-aged individuals. It will be important to describe patterns in reported SARS-CoV-2 infection and non-SARS-CoV-2 RTIs, and in diagnoses, prior to investigating associations between infection and subsequent diagnoses. Since many individuals are still experiencing long-term effects of their first SARS-CoV-2 infection, the analyses will focus on the first infection (separately for SARS-CoV-2 and non-SARS-CoV-2 RTI). Three different cohorts will be considered, corresponding to three different exposure periods aligned to the dominating SARS-CoV-2 variants at that time: pre-Delta, Delta, and post-Delta. This will be done for both SARS-CoV-2 exposures and for non-SARS-CoV-2 RTI exposures, since patterns of both SARS-CoV-2 infection and RTIs are likely aligned with SARS-CoV-2 variant waves and associated COVID-19 mitigation and testing strategies. A range of diagnoses will be explored utilising a high-throughput phenotyping data-driven approach, in order not to limit analyses to pre-conceived ideas of where associations lie. Outcomes will be investigated for three time periods: 4 weeks to 12 weeks, 12 weeks to 6 months, and 6 months to 4 years after infection, to reflect NICE's definitions for ongoing symptomatic COVID-19 and post-COVID-19 syndrome. Subsequent proposals will focus on the number and type of health service attendances, and prescriptions.

Aims

The aim of this project is to investigate the association of a range of diagnoses and previous SARS-CoV-2 infection, as well as previous non-SARS-CoV-2 RTIs in school-aged individuals, and understand differences by demographic and clinical characteristics.

Research Questions

1. Among school-aged individuals in England, is the first SARS-CoV-2 infection associated with higher rates of subsequent diagnoses, before and after adjusting for covariates, compared to the (collective) period before or in the absence of SARS-CoV-2 infection?
2. Among school-aged individuals in England, is the first non-SARS-CoV-2 RTI associated with higher rates of subsequent diagnoses, before and after adjusting for covariates, compared to the period before or in the absence of non-SARS-CoV-2 RTI?
3. What is the absolute excess risk of subsequent diagnoses after first SARS-CoV-2 infection compared to the period before or in the absence of SARS-CoV-2 infection?
4. What is the absolute excess risk of subsequent diagnoses after first non-SARS-CoV-2 RTI compared to the period before or in the absence of non-SARS-CoV-2 RTI?

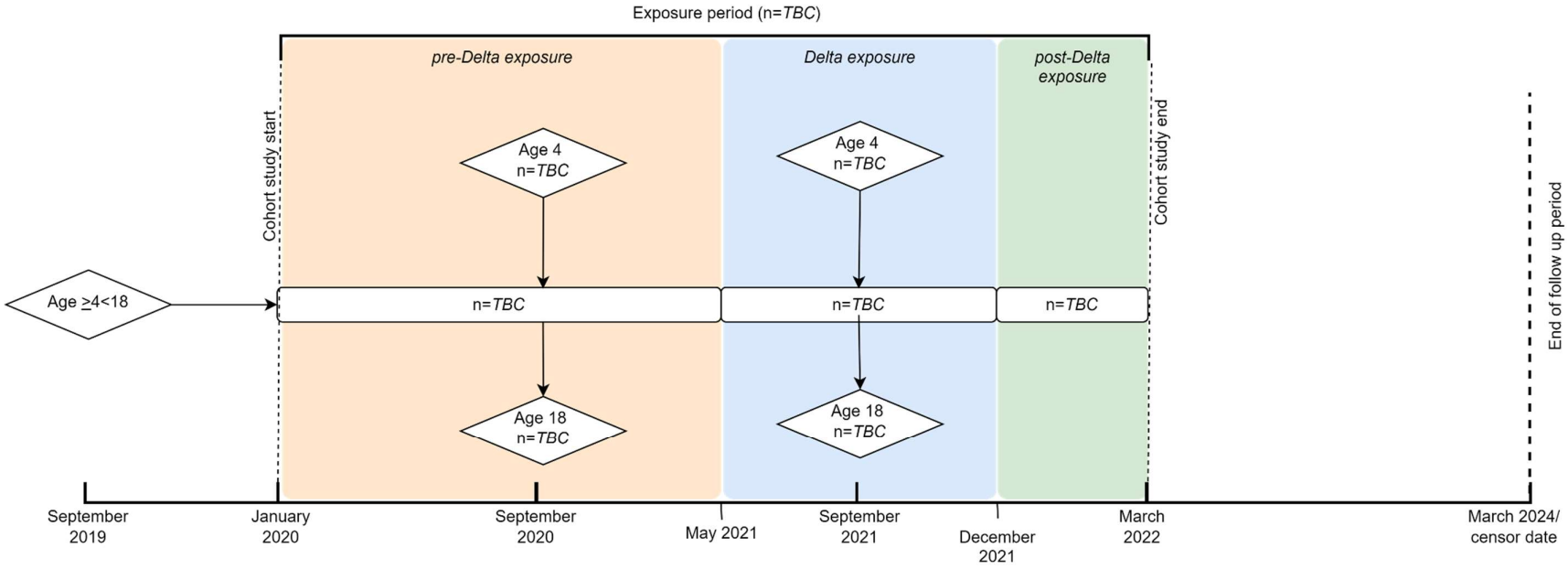
Objectives

1. Describe the trends in SARS-CoV-2 infections amongst school-aged individuals in England between January 2020 to March 2022.
    a. Compare trends in hospitalised versus non-hospitalised SARS-CoV-2 infections

2. Describe the trends in non-SARS-CoV-2 RTIs amongst school-aged individuals in England pre-pandemic (January 2019-December 2020) and during the pandemic (January 2020-March 2022).
    a. Compare trends in hospitalised versus non-hospitalised non-SARS-CoV-2 RTIs
3. Describe the trends in categories of diagnoses made in secondary care amongst school-aged individuals in England between January 2020 to March 2024 generated using a data-driven high throughput phenotyping approach.
4. Compare the rate of, and risk factors for, a range of diagnoses following SARS-CoV-2 infection (exposed group) versus before or in the absence of SARS-CoV-2 infection (unexposed group).
    a. Compare the rate of a range of diagnoses following hospitalised versus non-hospitalised SARS-CoV-2 infection
5. Compare the rate of, and risk factors for, a range of diagnoses following a non-SARS-CoV-2 respiratory infection (exposed group) versus before or in the absence of non-SARS-CoV-2 respiratory infection (unexposed group).
    a. Compare the rate of a range of diagnoses following hospitalised versus non-hospitalised non-SARS-CoV-2 respiratory infection
6. Investigate trends, rates and risks of range of diagnoses following infection within specific clinically-relevant subgroups of both SARS-CoV-2 infection and non-SARS-CoV-2 RTI.
7. Estimate the absolute excess risk of diagnoses following SARS-CoV-2 infection and following non-SARS-CoV-2 respiratory infection regionally and nationally.

Methods

Study population



SAP Figure 1. Eligible study population – selection of eligible school-aged population between January 2020 and March 2022

SAP Table 1. Cohort specification

		Cohort 1: pre-Delta exposure	Cohort 2: Delta exposure	Cohort 3: post-Delta exposure
Inclusion		At start date for each cohort:- • 4-18 years old at start of academic school year during cohort study period • Known sex • Alive • Primary care record • Resident in England		
Exclusion		At start date for each cohort:- • Confirmed SARS-CoV-2 and non-SARS-CoV-2 RTI exposure before cohort start date (but retained for sensitivity analyses of prior infection for Delta and post-Delta cohorts) • Confirmed non-SARS-CoV-2 RTI within 12 months of cohort start date (to exclude those with recent RTI before January 2020) • Fails quality assurance checks (e.g., date of death relative to birth, sex-specific diagnoses)		
Exposure	**1) SARS-CoV-2 infection**	1st confirmed +ve LFT (antigen)/PCR test or SARS-CoV-2 diagnosis from • National surveillance/testing (SGSS & Pillar 2*) • Primary care diagnosis (GDPPR: SNOMED*) • Secondary care (HES: ICD-10*)		
	2) Non-SARS-CoV-2 RTI	Confirmed diagnosis of respiratory infection (URTI, influenza, LRTI, pneumonia) that is not SARS-CoV-2 from • Primary care (SNOMED*) • Secondary care (ICD-10)		
Exposure period	**Start date**	• 1st January 2020[2]	• 22nd May 2021[2]	• 18th December 2021[2]
	End date	• 21st May 2021	• 17th December 2021	• 31st March 2022
Follow-up period for outcomes	**Start date**	• Exposure start date		
	End date	Earliest of: • Death (includes exposure-related deaths) • Outcome event • Latest data release (e.g., 31st March 2024) Sensitivity analysis of censoring at subsequent exposure		

Outcomes	• Diagnoses (ICD-10 from the phecode system and/or the Disease Atlas) after infection compared to before/in the absence of infection during the following time periods (4-12 weeks/12 weeks-6 months/6 months-end of follow-up in days)		
	[28, 84), [84, 183), [183, 1151)	[28, 84), [84, 183), [183,1044)	[28, 84), [84, 183), [183, 834)
Covariates at exposure start	• Age • Sex • Ethnicity • Deprivation • Location in England • Pre-existing outcome as co-morbidity(diagnosis in secondary care) • JCVI* & shielding (SNOMED CT code 1300561000000107) risk group • Vaccination**		
Subgroups* **	• Age, sex, ethnicity • Hospitalised vs non-hospitalised infection • Vulnerable at-risk groups (e.g. JCVI/shielding (including asthma); also see potential sensitivity analyses)		

[1] 90 days is the UKHSA definition of re-infection https://ukhsa.blog.gov.uk/2022/02/04/changing-the-covid-19-case-definition/
[2] Dates when over half of sequenced isolates were of a particular variant from sampled cases in England
https://covid19.sanger.ac.uk/lineages/raw?lineageView=1&lineages=B.1.1.7%2CB.1.617.2%2CB.1.1.529&colours=1%2C6%2C2

* SGSS - Second Generation Surveillance System (see Data Sources)
GDPR - GPES Data for Pandemic Planning and Research (see Data Sources)
HES – Hospital Episodes Statistics (see Data Sources)
LFT – Lateral Flow Test
PCR – Polymerase chain reaction
RTI – respiratory tract infection
URTI – upper respiratory tract infection
LRTI – lower respiratory tract infection
SNOMED - Systematized Nomenclature of Medicine Clinical Terms
ICD-10 - International Classification of Diseases 10th edition
JCVI – Joint Committee on Vaccination and Immunisation
** SARS-CoV-2 infection: COVID-19 vaccination; non-SARS-CoV-2 RTI: will be guided by availability of vaccination codes in healthcare records
*** potential – will be guided by descriptive analyses

Data Sources

COVID-19 testing & vaccinations:

- COVID-19 SGSS (Second Generation Surveillance System)
- Pillar 2 Antigen testing
- COVID-19 Vaccination Status

Secondary care:

- Admitted Patient Care (Hospital Episode Statistics; HES)
- Critical care (HES) (if available)
- Outpatients (HES)
- Accident & Emergency (HES)
- COVID Hospitalisations Surveillance Service (CHESS)

Primary care:

- GDPPR: GPES Data for Pandemic Planning and Research

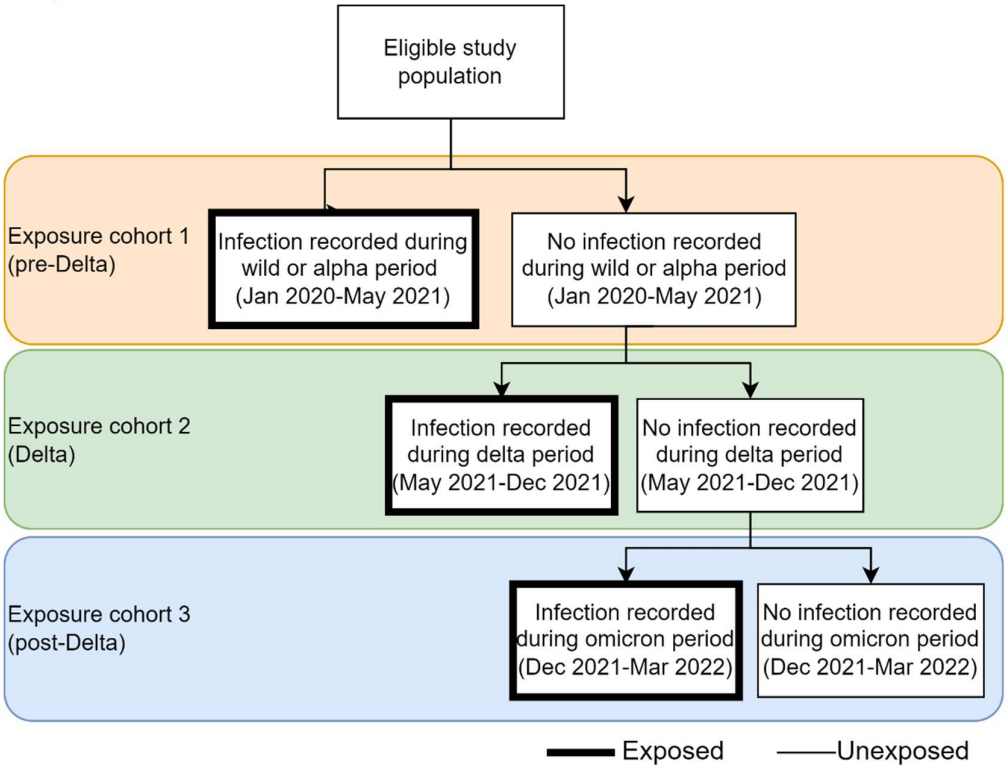
Civil Registrations:

- Deaths (ONS guidance/NHSD mortality data review)

For a breakdown of where data sources are being used, see Appendices

Exposures

Individuals will move from the unexposed group to the exposed group once they become infected. Outcomes of those exposed will be compared to those before or in the absence of exposure collectively.



SAP Figure 2. Exposure cohorts based on dominant circulating SARS-CoV-2 variant. Individuals will be included in the exposed cohort where there is a record of their first infection during the period and not included in subsequent cohorts. Individuals with no infection recorded will be treated as unexposed until an exposure is recorded, therefore some individuals will be included as unexposed in more than one cohort.

SARS-CoV-2 infection

A positive antigen test or SARS-CoV-2 diagnosis will be used to define SARS-CoV-2 infection, as defined in the following primary care, secondary care and national surveillance sources between January 2020-March 2022 when routine testing ended:

SAP Box 1. Definition of SARS-CoV-2 infection exposure

Data source	Definition
SGSS, Pillar 2 antigen testing	Positive SARS-CoV-2 PCR or antigen test
Primary care	Confirmed SNOMED diagnosis code
Secondary care	Episode with confirmed admission in any position (ICD-10 code U071/U072)

Non-SARS-CoV-2 RTI

A non-SARS-CoV-2 RTI will be defined based on the following primary and secondary care sources and codelists. Different types of non-SARS-CoV-2 RTI will also investigated both separately, (e.g. influenza, pneumonia, respiratory syncytial viral (RSV) infections) and in broad categories (i.e., upper and lower respiratory tract infections). SARS-CoV-2-specific codes will be excluded from these definitions. These codelists are a starting point to develop for this project, and will be reviewed by a clinician to assess and refine for relevance to school-aged individuals. Finalised codelists will be validated for use in this project and added to an open-access repository (e.g. HDRUK Phenotype Library).

SAP Box 2. Definition of non-SARS-CoV-2 RTI exposure

Infection type	Data source	Definition	Codelists
Upper respiratory tract infections (URTI)	Primary care	Confirmed SNOMED diagnosis code of URTI defined in codelists	https://www.opencodelists.org/codelist/bristol/upper-respiratory-tract-infections/2b52ba22/ https://www.opencodelists.org/codelist/nhsd-primary-care-domain-refsets/c19flurti_cod/20200812/
	Secondary care	Episode with confirmed admission in any position	https://phenotypes.healthdatagateway.org/phenotypes/PH158/version/316/detail/ https://phenotypes.healthdatagateway.org/phenotypes/PH488/version/1521/detail/
Lower respiratory tract infections (LRTI)	Primary care	Confirmed SNOMED diagnosis code of LRTI defined in codelists	https://www.opencodelists.org/codelist/nhsd-primary-care-domain-refsets/c19flurti_cod/20200812/
	Secondary care	Episode with confirmed admission in any position	https://phenotypes.healthdatagateway.org/phenotypes/PH204/version/408/detail/ https://phenotypes.healthdatagateway.org/phenotypes/PH488/version/1521/detail/

Influenza & 'flu-like symptoms	Primary care	Confirmed SNOMED diagnosis code	https://www.opencodelists.org/codelist/nhsd-primary-care-domain-refsets/c19flurti_cod/20200812/ https://phenotypes.healthdatagateway.org/phenotypes/PH800/version/2224/detail/ https://phenotypes.healthdatagateway.org/phenotypes/PH806/version/2230/detail/
	Secondary care	Episode with confirmed admission in any position	https://phenotypes.healthdatagateway.org/phenotypes/PH801/version/2225/detail/ https://phenotypes.healthdatagateway.org/phenotypes/PH807/version/2231/detail/ https://www.opencodelists.org/codelist/bristol/influenza_icd10/71a06879/
Pneumonia	Primary care	Confirmed SNOMED diagnosis code	https://phenotypes.healthdatagateway.org/phenotypes/PH15/version/30/detail/ https://www.opencodelists.org/codelist/bristol/pneumonia_v2/0f871dfa/
	Secondary care	Episode with confirmed admission in any position	https://phenotypes.healthdatagateway.org/phenotypes/PH15/version/30/detail/ https://www.opencodelists.org/codelist/bristol/pneumonia_icd10/7453286f/ https://www.opencodelists.org/codelist/opensafely/pneumonia-secondary-care/2020-10-05/
Respiratory syncytial viral (RSV) infections	Primary care	Confirmed SNOMED diagnosis code	https://www.opencodelists.org/codelist/nhsd-primary-care-domain-refsets/c19flurti_cod/20200812/ https://www.opencodelists.org/codelist/bristol/upper-respiratory-tract-infections/2b52ba22/
	Secondary care	Episode with confirmed admission in any position	https://phenotypes.healthdatagateway.org/phenotypes/PH488/version/1521/detail/

			https://www.opencodelists.org/codelist/bristol/rsv_icd10/5890f544/

Categories for infection type not mutually exclusive

Covariates

SAP Table 2. Covariates recorded at cohort exposure start

Covariate	Type	Definition
Age	Continuous	Within included range ≥4<18 years old at start of school year (1st September 2019/2020/2021)
Sex	Categorical	Male, Female
Ethnicity	Categorical	1: White 2: Mixed 3: Asian or Asian British 4: Black or Black British 5: Other
Deprivation	Categorical	10 categories from Index of Multiple Deprivation 2019 (Using LSOA 2011 at/closest record to start of exposure period)
Geographical location	Categorical	East of England London Midlands North East and Yorkshire North West South East South West Scotland Wales (Using LSOA 2011 at/closest record to start of exposure period)
Pre-existing medical diagnosis	Binary	Presence or absence of Level 1 phecode (Disease Atlas from ICD-10 in hospital records; see **Outcomes**)
JCVI*/shielding risk group	Binary	Criteria for clinical risk group met or unmet
Vaccination history	Binary	Presence or absence of at least 1 dose of vaccination for SARS-CoV-2 or other RTIs (e.g. seasonal influenza)

* JCVI – Joint Committee on Vaccination and Immunisation

Outcomes

A validated reference catalogue of diseases ("Disease Atlas"; https://www.ucl.ac.uk/health-informatics/research/disease-atlas) across clinical specialities will be used. The catalogue contains phenotyping algorithms generating phecodes for all common, uncommon and rare diseases recorded in electronic health record data. The idea of this data-driven approach to create a comprehensive phenotype for each disease is that these phenotypes enable a systematic comparison across all diseases.

Four levels of phecodes have been developed, with the level of detail and specificity of phenotype increasing as the levels increase. For this analysis, level 1 phecodes (n=1857), the most broad and comparable to category of diagnosis in the ICD-10 code structure, will be used. Phecodes relevant to school-aged individuals will be selected with clinical guidance, that have most clinical relevance for children and young people, and reflect the broad range of possible symptoms experienced by, and of concern to, the PPI contributors.

Data Analysis

All data analyses will be done using the "R" programming language. Reproducible code and statistical analysis plans will be published open-source.

Descriptive analyses

See SAP Table 1 for details of study design.

Descriptive analyses will be conducted in the first instance to investigate demographic and clinical characteristics of the cohorts.

We will describe the incidence rates (number of events and person-years) of exposures and outcomes during the follow up-period. This will inform subgroup associations and prioritising outcomes (subsequent diagnoses) of interest for further analysis.

Respiratory infection (including SARS-CoV-2) is an acute exposure, meaning children and young people may have multiple reports. Incidence of different infection types (e.g., SARS-CoV-2, pneumonia, respiratory syncytial viral (RSV), upper RTIs and influenza) will be described to investigate where infections co-exist within the same exposure period, such as the proportion of the cohort with a pneumonia diagnosis reported within 28 days of SARS-CoV-2 infection. This will inform further subgroup analyses to investigate the risk of specific infection types and combination of co-existing infections.

Cox regression

Mixed-effects Cox regression models will be used to calculate hazard ratios (HRs) for the association of diagnoses outcomes following both i) SARS-CoV-2 infection and ii) non-SARS-CoV-2 RTI. Clinical diagnoses occurring over (1) 4-12 weeks, (2) 12 weeks-6 months and (3) >6 months since exposure will be examined. See Appendices for splitting follow-up time and blank example tables.

We will fit Cox regression models with calendar time scale using the exposure period start date as the time origin (T0). This will ensure that all analyses account for changes with calendar time in rates of the outcome event. Using this approach, we will estimate HRs for events of different types before and after exposure, and by time since exposure.

It is unlikely to be feasible for the regression models to be run when the full sample contains more than 4 million children. For computational efficiency, we will use a sampling procedure for datasets containing more than 4,000,000 individuals. Cox models will be fitted to datasets including all individuals with the outcome event (i.e., the cases), all exposed individuals, and a random subset of unexposed individuals without the outcome event (i.e., the controls) equal to Y times the number of individuals with the outcome event (X), where Y is determined by the range of values X falls within. Analyses will incorporate inverse probability weights for data from unexposed individuals without the outcome event. For example, consider a sample of N people, X of whom have the outcome. We want to sample Y * X people without the exposure or the outcome and assign a weight of (N-X)/(Y *X) to each control and 1 to each case and exposed individual. If 20X>=N-X, we will use the whole sample. Confidence intervals will be derived using robust standard errors.

Proportional hazards (PH) have not been violated in other similar analyses. However, if patterns in PH are found to deviate from previous work, we will investigate appropriateness of follow-up timeframes and consider alternative approaches.

We will estimate: (i) age, sex and region-adjusted and (ii) maximally-adjusted HRs. We will exclude potential covariates with ≤2 occurrences at any level.

Absolute excess risk

We will calculate absolute excess risk (in time intervals since exposure) as the sum of the difference in the estimated daily incidence in the unexposed population and the expected daily incidence in the exposed population. The latter is estimated using life tables by applying the time-dependent hazard ratios to the estimated daily incidence in the unexposed population.

Sensitivity analysis

Prior infection

We will repeat the main analyses not excluding individuals who had SARS-CoV-2 or non-SARS-CoV-2 infection prior to the start of the Delta and post-Delta exposure cohort date.

Censoring at subsequent infection

In the main analyses, follow-up time for outcomes following first infection (whether SARS-CoV-2 or non-SARS-CoV-2) will not be censored at any subsequent infection. This means that subsequent infections may influence the risk and rate of outcomes. The rationale for not censoring at subsequent infection is that we are interested in long-term outcomes following a first infection versus no infection from a clinical perspective. We will however perform a sensitivity analysis that does censor at subsequent infection. This will enable us to compare the association of infection with outcomes both with (main analysis) and without (sensitivity analysis) possible multiple infections.

Risk categories

Children and young people who have pre-existing conditions who were deemed to be in vulnerable at-risk groups (by JCVI or recommended to shield; https://digital.nhs.uk/coronavirus/shielded-patient-list/guidance-for-general-practice) may have had different experience of exposure and outcomes to those not considered vulnerable. If feasible, we will perform a sensitivity analysis without these at-risk subgroups.

Missing data

Individuals with missing data on age or sex will be excluded from the analysis. We will include missing categories for ethnicity, deprivation and geographical location. All other covariates will be defined using the presence versus absence of specific codes in the electronic health records, so they have no identifiable missing values. Multiple imputation will not be utilised because missing data is part of the exclusion criteria to ensure linkage reliability and therefore levels of missing data are low.

Appendices

Data Sources

SAP Appendix Table 1. Data sources for each parameter (i.e. exposures, covariates, outcomes)

Domain	Data source	Exposure				Covariates		Vaccination	Outcomes	QA
		SARS-CoV-2 infection	Respiratory infection			Age: location	Pre-existing medical diagnosis			
			URTI	Influenza	LRTI/ Pneumonia					
COVID-19 surveillance	COVID-19 SGSS (Second Generation Surveillance System)	X								
	Pillar 2 Antigen testing	X								
	COVID-19 Vaccination Status							X		
Primary care	GDPPR: GPES Data for Pandemic Planning and Research	X	X	X	X	X				
Secondary care	Admitted Patient Care (Hospital Episode Statistics; HES)	X			X		X – first diagnosis position		X – first diagnosis position	
	Outpatients (HES)	X			X					
	Accident & Emergency (HES)	X			X					
	Emergency Care Data Set (ECDS)				X					
	Critical care (HES)	X			X					
	COVID Hospitalisations Surveillance Service (CHESS)	X								
Civil Registrations	Deaths				X				X	

Draft tables

Manuscript Table 1a. Characteristics of those with SARS-CoV-2 infection between January 2020-March 2022

		Total cohort	Pre-Delta		Delta		Post-Delta	
			All (N [%])	SARS-CoV-2 infection (N [%])	All (N [%])	SARS-CoV-2 infection (N [%])	All (N [%])	SARS-CoV-2 infection (N [%])
Total (%)								
Hospitalisation within 28 days of infection								
Age (Years; mean, s.d)								
COVID-19 vaccination history								
Sex	Male							
	Female							
Ethnicity	White							
	Mixed							
	Asian or Asian British							
	Black or Black British							
	Other							
Deprivation	% most deprived area							
Location	East of England							
	London							
	Midlands							
	North East and Yorkshire							
	North West							
	South East							
	South West							
Pre-existing medical diagnosis	% Yes							
Vulnerable at-risk group	% Yes							

Manuscript Table 1b. Characteristics of those with non-SARS-CoV-2 respiratory tract infection (RTI) between January 2020-March 2022

		Total cohort	Pre-Delta		Delta		Post-Delta	
			All (N [%])	RTI (N [%])	All (N [%])	RTI (N [%])	All (N [%])	RTI (N [%])
Total (%)								
Hospitalisation within 28 days of infection								
Age (Years; mean, s.d)								
Vaccination history								
Sex	Male							
	Female							
Ethnicity	White							
	Mixed							
	Asian or Asian British							
	Black or Black British							
	Other							
Deprivation	% most deprived area							
Location	East of England							
	London							
	Midlands							
	North East and Yorkshire							
	North West							
	South East							
	South West							
Pre-existing medical diagnosis	% Yes							
Vulnerable at-risk group	% Yes							

Manuscript Table 2a. Crude incidence rates (per X,000 person-years) for diagnoses following SARS-CoV-2 infection by exposure period

Outcomes	Pre-Delta		Delta		Post-Delta	
	N of events/person-years	Incidence rate (95% CI)	N of events/person-years	Incidence rate (95% CI)	N of events/person-years	Incidence rate (95% CI)
Diagnosis 1						
No SARS-CoV-2						
Hospitalised for SARS-CoV-2						
Not hospitalised for SARS-CoV-2						
Diagnosis 2						
No SARS-CoV-2						
Hospitalised for SARS-CoV-2						
Not hospitalised for SARS-CoV-2						
Diagnosis 3						
No SARS-CoV-2						
Hospitalised for SARS-CoV-2						
Not hospitalised for SARS-CoV-2						
Diagnosis…n						
No SARS-CoV-2						
Hospitalised for SARS-CoV-2						
Not hospitalised for SARS-CoV-2						

Manuscript Table 2b. Crude incidence rates (per X,000 person-years) for diagnoses following non-SARS-CoV-2 RTI by exposure period

Outcomes	Pre-Delta		Delta		Post-Delta	
	N of events/person-years	Incidence rate (95% CI)	N of events/person-years	Incidence rate (95% CI)	N of events/person-years	Incidence rate (95% CI)
Diagnosis 1						
No non-SARS-CoV-2 RTI						
Hospitalised for non-SARS-CoV-2 RTI						
Not hospitalised for non-SARS-CoV-2 RTI						
Diagnosis 2						
No non-SARS-CoV-2 RTI						
Hospitalised for non-SARS-CoV-2 RTI						
Not hospitalised for non-SARS-CoV-2 RTI						
Diagnosis 3						
No non-SARS-CoV-2 RTI						
Hospitalised for non-SARS-CoV-2 RTI						
Not hospitalised for non-SARS-CoV-2 RTI						
Diagnosis…n						
No non-SARS-CoV-2 RTI						
Hospitalised for non-SARS-CoV-2 RTI						
Not hospitalised for non-SARS-CoV-2 RTI						

Manuscript Table 3a. Adjusted hazard ratio estimates for diagnosis following SARS-CoV-2 infection by exposure period

	Time since infection	Pre-Delta		Delta		Post-Delta	
		aHR	95% CI	aHR	95% CI	aHR	95% CI
Diagnosis 1	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis 2	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis 3	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis…n	4-12 weeks						
	13 weeks-6 months						
	7-39 months						

Manuscript Table 3b. Adjusted hazard ratio estimates for diagnosis following non-SARS-CoV-2 RTI by exposure period

	Time since infection	Pre-Delta		Delta		Post-Delta	
		aHR	95% CI	aHR	95% CI	aHR	95% CI
Diagnosis 1	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis 2	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis 3	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis…n	4-12 weeks						
	13 weeks-6 months						
	7-39 months						

Manuscript Table 4an. Adjusted hazard ratio estimates for [diagnosis following] SARS-CoV-2 infection by exposure period [subgroup *n*]

	Time since infection	Pre-Delta		Delta		Post-Delta	
		aHR	95% CI	aHR	95% CI	aHR	95% CI
Diagnosis 1	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis 2	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis 3	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis…n	4-12 weeks						
	13 weeks-6 months						
	7-39 months						

Manuscript Table 4bn. Adjusted hazard ratio estimates for [diagnosis following] non-SARS-CoV-2 RTI by exposure period [subgroup *n*]

	Time since infection	Pre-Delta		Delta		Post-Delta	
		aHR	95% CI	aHR	95% CI	aHR	95% CI
Diagnosis 1	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis 2	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis 3	4-12 weeks						
	13 weeks-6 months						
	7-39 months						
Diagnosis…n	4-12 weeks						
	13 weeks-6 months						
	7-39 months						

Table 5a. Excess events per 100,000 people at [x months] post SARS-CoV-2 infection in the pre-Delta, Delta and post-Delta cohorts

Outcome	Pre-Delta			Delta			Post-Delta		
	n	Difference	% difference	n	Difference	% difference	n	Difference	% difference
Diagnosis 1									
Diagnosis 2									
Diagnosis 3									
Diagnosis…n									

Table 5b. Excess events per 100,000 people at [x months] post non-SARS-CoV-2 RTI in the pre-Delta, Delta and post-Delta cohorts

Outcome	Pre-Delta			Delta			Post-Delta		
	n	Difference	% difference	n	Difference	% difference	n	Difference	% difference
Diagnosis 1									
Diagnosis 2									
Diagnosis 3									
Diagnosis…n									

Draft figures

Manuscript Figure 1. Data pipelines

Manuscript Figure 2. Upset plot of incidence of infections across data sources

Manuscript Figure 3. Adjusted hazard ratios for [diagnosis following] infections by severity (hospitalised vs non-hospitalised) for pre-Delta, Delta and post-Delta cohorts

Manuscript Figure 4. Adjusted hazard ratios for [diagnosis following] infections by [subgroups] for pre-Delta, Delta and post-Delta cohorts

Manuscript Figure 5. Absolute excess risk up to [x months] for [diagnosis following] infections [by age-group] for pre-Delta, Delta and post-Delta cohorts

Splitting follow-up time

Consider the following definitions:

- Time scale: days since the start of the study.
- Outcome of interest: time to event D measured at T_D with indicator I_D in days.
- Exposure of interest: binary exposure E measured at T_E with indicator I_E, parameterised as days since T_E. This will be categorised, for example, into: E1 = [28, 84); E2 = [84, 183); E3 = [84, 1551/1044/834);
- Administrative censoring time: set as day T_C.

For individuals without exposure and without an event, then T_D=T_C, I_D = 0, T_E=end of exposure period (506/209/103 days), T1=T_C (end of follow-up period (T0+1551/1044/834)), I_E=0, T0=0 (e.g., individual 1 in tables).

For individuals without exposure and with an event at time *t*, then T_D=*t*, T1=T_D, I_D = 1, T_E=*t*, I_E=0, T0=0 (e.g., individual 2 in tables).

For individuals with exposure at T_E and without an event, then: (1) split follow-up time at T_E, and (2) split follow-up time>T_E at T_E+84; T_E+183; (T_E+1551, T_E+1044, T_E+834) and then censor at earliest of T_E+1551/+1044/+834 or T_C (e.g., individual 3 in table below).

For individuals with exposure at T_E and an event at T_D, then first (1) split follow-up time at T_E, and then (2) split follow-up time>T_E at T_E+84; T_E+183; (T_E+1551, T_E+1044, T_E+834) and then censor at earliest of T_E+1551/+1044/+834 or T_D (e.g., individual 4 in table below).

SAP Appendix Table 2a. Splitting follow-up time for Cohort 1 (pre-Delta period: 1/1/2020 – 21/05/2021 (506 days))

id	**T_E**	**T_D**	**T_C**	**T0**	**T1**	**I_E**	**I_D**	**E1**	**E2**	**E3**
Individual	*Exposure time*	*Follow-up time*	*Administrative censor time*	*Index time*	*Time to outcome*	*Exposure indicator*	*Outcome indicator*	[28, 84) days	[84, 183) days	[183, 1551) days
1	506	1552	1552	0	1552	0	0	0	0	0
2	47	47	1552	0	47	0	1	0	0	0
3	35	1552	1552	0	35	0	0	0	0	0
3	35	1552	1552	35	119	1	0	1	0	0
3	35	1552	1552	119	218	1	0	0	1	0
3	35	1552	1552	218	1552	1	0	0	0	1
4	105	136	1552	0	105	0	0	0	0	0
4	105	136	1552	105	136	1	0	1	0	0

SAP Appendix Table 2b. Splitting follow-up time for Cohort 2 (Delta period: 22/05/2021 – 17/12/2021 (209 days))

id	**T_E**	**T_D**	**T_C**	**T0**	**T1**	**I_E**	**I_D**	**E1**	**E2**	**E3**
Individual	*Exposure time*	*Follow-up time*	*Administrative censor time*	*Time to exposure*	*Time to outcome*	*Exposure indicator*	*Outcome indicator*	[28, 84) days	[84, 183) days	[183, 1044) days
1	209	1045	1045	0	1045	0	0	0	0	0
2	47	47	1045	0	47	0	1	0	0	0
3	35	1045	1045	0	35	0	0	0	0	0
3	35	1045	1045	35	119	1	0	1	0	0
3	35	1045	1045	119	218	1	0	0	1	0
3	35	1045	1045	218	1045	1	0	0	0	1
4	105	136	1045	0	105	0	0	0	0	0
4	105	136	1045	105	136	1	0	1	0	0

SAP Appendix Table 2c. Splitting follow-up time for Cohort 3 (post-Delta period: 18/12/2021 – 28/02/2022 (72 days))

id	**T_E**	**T_D**	**T_C**	**T0**	**T1**	**I_E**	**I_D**	**E1**	**E2**	**E3**
Individual	*Exposure time*	*Follow-up time*	*Administrative censor time*	*Time to exposure*	*Time to outcome*	*Exposure indicator*	*Outcome indicator*	[28, 84) days	[84, 183) days	[183, 834) days
1	72	835	835	0	835	0	0	0	0	0
2	47	47	835	0	47	0	1	0	0	0
3	35	835	835	0	35	0	0	0	0	0
3	35	835	835	35	119	1	0	1	0	0
3	35	835	835	119	218	1	0	0	1	0
3	35	835	835	218	835	1	0	0	0	1
4	105	136	835	0	105	0	0	0	0	0
4	105	136	835	105	136	1	0	1	0	0

References

1. Molteni E, Sudre CH, Canas LS, et al. Illness duration and symptom profile in symptomatic UK school-aged children tested for SARS-CoV-2. *Lancet Child Adolesc Health* 2021;5(10):708-18. doi: 10.1016/s2352-4642(21)00198-x [published Online First: 2021/08/07]
2. Pinto Pereira SM, Nugawela MD, McOwat K, et al. Symptom Profiles of Children and Young People 12 Months after SARS-CoV-2 Testing: A National Matched Cohort Study (The CLoCk Study). *Children (Basel)* 2023;10(7) doi: 10.3390/children10071227 [published Online First: 2023/07/29]
3. NICE. COVID-19 rapid guideline: managing the long-term effects of COVID-19: National Institute for Heath and Care Excellence, 2024.
4. Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. *Nature* 2021;594(7862):259-64. doi: 10.1038/s41586-021-03553-9
5. Zheng YB, Zeng N, Yuan K, et al. Prevalence and risk factor for long COVID in children and adolescents: A meta-analysis and systematic review. *J Infect Public Health* 2023;16(5):660-72. doi: 10.1016/j.jiph.2023.03.005 [published Online First: 2023/03/18]
6. Scherlinger M, Lemogne C, Felten R, et al. Excess of Post-Acute Sequelae of COVID-19 After the First Wave of the Pandemic. *Infectious Diseases and Therapy* 2022;11(6):2279-85. doi: 10.1007/s40121-022-00698-6
7. Matta J, Wiernik E, Robineau O, et al. Association of Self-reported COVID-19 Infection and SARS-CoV-2 Serology Test Results With Persistent Physical Symptoms Among French Adults During the COVID-19 Pandemic. *JAMA Internal Medicine* 2022;182(1):19-25. doi: 10.1001/jamainternmed.2021.6454
8. Mariette X. Long COVID: a new word for naming fibromyalgia? *Ann Rheum Dis* 2024;83(1):12-14. doi: 10.1136/ard-2023-224848 [published Online First: 20240102]
9. Sk Abd Razak R, Ismail A, Abdul Aziz AF, et al. Post-COVID syndrome prevalence: a systematic review and meta-analysis. *BMC Public Health* 2024;24(1):1785. doi: 10.1186/s12889-024-19264-5 [published Online First: 20240704]
10. Løkke FB, Hansen KS, Dalgaard LS, et al. Long-term complications after infection with SARS-CoV-1, influenza and MERS-CoV - Lessons to learn in long COVID? *Infect Dis Now* 2023;53(8):104779. doi: 10.1016/j.idnow.2023.104779 [published Online First: 2023/09/08]
11. Kuan V, Denaxas S, Gonzalez-Izquierdo A, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *The Lancet Digital Health* 2019;1(2):e63-e77. doi: 10.1016/S2589-7500(19)30012-3