Medicines & Healthcare products Regulatory Agency

CPRD

Second Generation Surveillance System (SGSS) COVID-19 positive virology data results

2021-03 CPRD SGSS Documentation

Version 1.2

Date: 12 March 2021

NIHR | National Institute for Health Research

Documentation Control Sheet

Over time, it may be necessary to issue amendments or clarifications to parts of this document. This form must be updated whenever changes are made.

Version	Affected Areas Summary of Change	Prepared By	Reviewed By
1.0	Initial	Eleanor Yelland	Rachael Williams/ Susan Hodgson
1.1	Modified	Susan Hodgson	
1.2	Modified	Eleanor Yelland	Susan Hodgson

Summary of Changes

Version 1.1

- Updated footer with new NIHR logo

Version 1.2

- Updated for March 2021 release; added DOI

Second Generation Surveillance system COVID-19 positive virology data linked to CPRD primary care data

This document provides an overview of the Second Generation Surveillance system (SGSS) data, and the available subset that is linked to CPRD GOLD and CPRD Aurum.

What is the Second Generation Surveillance system?

SGSS is the national laboratory reporting system used in England to capture routine laboratory data on infectious diseases and antimicrobial resistance. Diagnostic laboratories are required to notify Public Health England when specified causative agents are found in a human sample. The Health Protection (Notification) Regulations 2010 made provision for this and the requirement came into effect in October 2010.

The Second Generation Surveillance system is the application through which the data on laboratory notifications and isolates is stored and managed. The data is stored centrally within Public Health England. Laboratories in England are required to notify PHE of a positive result relating to a range of organisms, including viral such as: Ebola, dengue fever, hepatitis A-E, measles, and bacterial infections such as: Campylobacter, Legionella, or Salmonella. The full list of notifiable organisms and details of their urgency can be found here in guide for laboratory reporting to Public Health England(1):

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/739854/PHE_Laboratory_Reporting_Guidelines.pdf

What subset of these data has CPRD primary care data been linked to?

Currently, the subset of SGSS data that CPRD primary care data has been linked to includes only notifications of positive SARS-CoV-2 (COVID-19 [coronavirus]) specimen tests. All testing being reported currently is swab testing using the polymerise chain reaction (PCR) test method; antibody testing is not included. The laboratories reporting to PHE on SARS-CoV-2 are NHS hospitals and PHE laboratories carrying out pillar 1 testing; diagnostic swab testing for those with a clinical need, and health and care workers (2). Some SARS-CoV-2 tests are carried out by private laboratories and these may not be represented in the data. Only those that have been forwarded by the NHS to private laboratories are captured.

Pillar 2 testing for the wider population is not currently included in the available data. Pillar 2 testing includes tests conducted in community settings, such as regional test sites, mobile testing units, satellite test centres and via home tests.

Linkage method

The linkage method applied for the SGSS data differs from the 8-step linkage algorithm that has previously been used for other linkages undertaken by the CPRD trusted third-party, NHS Digital. In this instance, the linkage has been carried out using NHS number only.

Unlike the other linkages there is no specific linkage eligibility flag for these SGSS data. The HES eligibility flag (hes_e) in the source file may be used to identify a denominator population potentially eligible for linkage to SGSS (i.e. comprising patients registered at a practice participating in the linkage scheme, with a valid NHS-number). The linkage was carried out by NHS Digital using Jan 2021 EMIS and Vision identifiers, which is also the basis for the set 21 source file; as such, the set 21 source file is the most appropriate to use for understanding the subset of the population potentially eligible for linkage.

Please note that no personal identifiers are held by CPRD, or included in the CPRD GOLD, CPRD Aurum, or linked SGSS data available.

Linkage coverage period

This March 2021 release of SGSS COVID-19 positive virology data linked to CPRD primary care data covers the period 01/03/2020 – 29/09/2020 inclusive.

Data structure and formatting
As far as possible, the SGSS data is supplied "as is" without any modification or cleaning during processing by CPRD. Where CPRD has modified the data, these are detailed below.

- **Duplicate variables:**

The raw SGSS data provided to CPRD contained two versions of each variable, one marked 'cleaned'. However, upon review, the content of these variables was found to be identical to each of the raw versions. The excess 'cleaned' version has been dropped. This will be reviewed if data in each set of two variables are found to differ in later updates of the SGSS data.

DOI
Please cite in any publications using these data:

CPRD GOLD SGSS March 2021 - https://doi.org/10.48329/g0dh-h917

CPRD Aurum SGSS March 2021 - https://doi.org/10.48329/3jj1-4q52

Data Quality concerns – Known issues

These data are being made available for research purposes on accelerated timelines. Minimal changes have been made to the raw data and there is likely to be some requirement for data cleaning. See below some known potential concerns:

- **Duplicates:** Initial exploration by CPRD has found the presence of complete duplicates as well as records that appear to be duplicates for all but one or two variables. This is generally due to the value of the 'care_home' variable differing across the two records. These records are likely to be due the submission of 'updated' or 'refreshed' records to SGSS that may contain updated information. Unfortunately, it is not possible to determine which is the most recent record based on the currently available data.

- **Potential bias:** Initially, the majority of tests were carried out in London and laboratories in different locations started testing and reporting at different times. This may result in some geographical bias, particularly in early data.

- **Age recording:** A very small number of records contain the value of '-1' for the patient's age. Year of birth is recorded in the primary care data and can be used where age recording in these data look erroneous.

- **Multiple CPRD GOLD or CPRD Aurum patient records linked to a single specimen record (n_patid_spec):** In CPRD GOLD and CPRD Aurum, individuals are assigned a new patient identifier (patid) each time they move from one contributing practice to another. A single individual can therefore be represented by two or more different patient identifiers, and each can be linked to the same specimen record (cdr_specimen_request as unique specimen ID). This information is captured by the variable n_patid_spec which indicates how many different patient identifiers within the same database (i.e. CPRD GOLD or CPRD Aurum) are linked to the same specimen number.

  In a very small number of instances (affecting <0.02% of all specimen IDs in the linked SGSS data), the number of primary care patient records linked to the same specimen record is large

(n_patid_spec>20). Although this is a very rare occurrence, the linked SGSS data for these patients may not be reliable. A simple approach when using linked SGSS data would therefore be to exclude from any analyses all patients where n_patid_spec is deemed to be large.

SGSS – SARS-CoV-2 data: Data dictionary

1. Specimen file (CPRD_GOLD_SGSS_March_2021.txt & CPRD_Aurum_SGSS_March_2021.txt)

Column name	*Description*	*Type*	*Format*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum [primary key]	INTEGER	20
pracid	Encrypted unique key given to a practice in CPRD GOLD or CPRD Aurum	INTEGER	5
n_patid_spec	Number of unique CPRD patid within a database linked to this crd_specimen_request_sk	INTEGER	3
pseudo_specimen_id	Pseudonymised specimen identifier – unique to each specimen within the dataset	NUMBER	12
organism_species_name	The name of the organism tested for (all SARS-CoV-2 by definition here)	TEXT	33
lab_report_date	Date of laboratory report	TEXT	23
age_in_years	Age in years when specimen was taken	INTEGER	8
reporting_lab_id	ID for lab reporting test result – pseudonymised	INTEGER	8
specimen_date	Date specimen was taken	TEXT	23
care_home	Indicator for care home residence (FALSE | TRUE | missing)	TEXT	5

References

1. Public Health England. Laboratory reporting to Public Health England. PHE publications; 2016.
2. COVID-19 testing data: methodology note [Internet]. GOV.UK. [cited 2020 Nov 10]. Available from: https://www.gov.uk/government/publications/coronavirus-covid-19-testing-data-methodology/covid-19-testing-data-methodology-note