In-pandemic development of an application ontology for COVID-19 surveillance in a primary care sentinel network

Simon de Lusignan, Harshana Liyanage, Dylan McGagh, Bhautesh Dinesh Jani, Jorgen Bauwens, Rachel Byford, Dai Evans, Tom Fahey, Trisha Greenhalgh, Nicholas Jones, Frances S Mair, Cecilia Okusi, Vaishnavi Parimalanathan, Jill P Pell, Julian Sherlock, Oscar Tamburis, Manasa Tripathy, Filipa Ferreira, John Williams, FD Richard Hobbs

Submitted to: JMIR Public Health and Surveillance
on: July 01, 2020

Table of Contents

In-pandemic development of an application ontology for COVID-19 surveillance in a primary care sentinel network

Simon de LusignanMD, ; Harshana LiyanagePhD, ; Dylan McGaghBSc, ; Bhautesh Dinesh JaniMBChB, ; Jorgen BauwensMSc, MPH, ; Rachel ByfordBSc, ; Dai EvansBSc, MBBS, ; Tom FaheyMSc, MD, ; Trisha GreenhalghFMedSci, ; Nicholas JonesMBBS, MSc, ; Frances S Mair; Cecilia Okusi; Vaishnavi Parimalanathan; Jill P PellMSc, MD, ; Julian SherlockBSc, ; Oscar TamburisMEng, PhD, ; Manasa TripathyBSc, MSc, ; Filipa FerreiraBEng, MSc, PhD, ; John WilliamsMSc, FRCGP, FFCI, ; FD Richard HobbsFMedSci, FRCGP, MA, FRCP,

Corresponding Author:
Simon de LusignanMD,
Phone: +441865617283
Email: simon.delusignan@phc.ox.ac.uk

Abstract

Background: Creating an ontology for coronavirus disease 2019 (COVID-19) surveillance should help ensure transparency and consistency. Ontologies formalise conceptualisations at either domain or application level. Application ontologies cross domains and are specified through testable use cases. Our use case was extension of the role of the Oxford Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) to monitor the current pandemic and become an in-pandemic research platform.

Objective: To develop an application ontology for COVID-19 which can be deployed across the various use case domains of the Oxford- RCGP RSC research and surveillance activities.

Methods: We described our domain-specific use case. The actor was the RCGP RSC sentinel network; the system the course of the COVID-19 pandemic; the outcomes the spread and effect of mitigation measures. We used our established three-step method to develop the ontology, separating ontological concept development from code mapping and data extract validation. We developed a coding system–independent COVID-19 case identification algorithm. As there were no gold standard pandemic surveillance ontologies, we conducted a rapid Delphi consensus exercise through the International Medical Informatics Association (IMIA) Primary Health Care Informatics working group and extended networks.

Results: Our use case domains included primary care, public health, virology, clinical research and clinical informatics. Our ontology supported: (1) Case identification, microbiological sampling and health outcomes at both an individual practice and national level; (2) Feedback through a dashboard; (3) A national observatory, (4) Regular updates for Public Health England, and (5) Transformation of the sentinel network to be a trial platform. We have identified a total of 8,627 people with a definite COVID-19 status, 4,240 with probable, and 59,147 people with possible COVID-19, within the RCGP RSC network (N=5,056,075).

Conclusions: The underpinning structure of our ontological approach has coped with multiple clinical coding challenges. At a time when there is uncertainty about international comparisons, clarity about the basis on which case definitions and outcomes are made from routine data is essential.

(JMIR Preprints 01/07/2020:21434)
DOI: https://doi.org/10.2196/preprints.21434

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✔

Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

In-pandemic development of an application ontology for COVID-19 surveillance in a primary care sentinel network

Authors:
Simon de Lusignan 1,2*
Professor of Primary Care and Clinical Informatics
Director Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC)
SdeL simon.delusignan@phc.ox.ac.uk

Harshana Liyanage
Research Fellow
HL harshana.liyanage@phc.ox.ac.uk

Dylan McGagh
Medical Student
DMcG dylan.mcgagh@magd.ox.ac.uk

Bhautesh Dinesh Jani
Clinical Senior Lecturer
Bhautesh.jani@glasgow.ac.uk

Jorgen Bauwens
jbauwens@ssphplus.ch

Rachel Byford
rachel.byford@phc.ox.ac.uk

Dai Evans
Lead Clinical Advisor
Dai.Evans@nottingham.ac.uk

Tom Fahey
Professor of General Practice
tomfahey@rcsi.ie

Trisha Greenhalgh
Professor of Primary Care Health Sciences
trish.greenhalgh@phc.ox.ac.uk

Nicholas Jones
Research Fellow
nicholas.jones2@hmc.ox.ac.uk

Frances S Mair
Norie Miller Professor of General Practice
FSM frances.mair@glasgow.ac.uk

Cecilia Okusi
cecilia.okusi@phc.ox.ac.uk

Vaishnavi Parimalanathan
vaishnavi.parimalanathan@phc.ox.ac.uk

Jill P Pell
Henry Mechan Professor of Public Health
University of Glasgow
JPP jill.pell@glasgow.ac.uk

Julian Sherlock
julian.sherlock@phc.ox.ac.uk

Oscar Tamburis
oscar.tamburis@unina.it

Manasa Tripathy
manasa.tripathy@phc.ox.ac.uk

Filipa Ferreira
filipa.ferreira@phc.ox.ac.uk

John Williams
Senior Clinical Research Fellow
JS john.williams@phc.ox.ac.uk

FD Richard Hobbs
Nuffield Professor of Primary Care Health Sciences

FDRH richard.hobbs@phc.ox.ac.uk

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, New Radcliffe House, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG
[2]Royal College of General Practitioners, London, United Kingdom

[*] Corresponding Author:

Professor Simon de Lusignan

E: simon.delusignan@phc.ox.ac.uk

In-pandemic development of an application ontology for COVID-19 surveillance in a primary care sentinel network

Abstract 292 2020/06/02

Background:
Creating an ontology for coronavirus disease 2019 (COVID-19) surveillance should help ensure transparency and consistency. Ontologies formalise conceptualisations at either domain or application level. Application ontologies cross domains and are specified through testable use cases. Our use case was extension of the role of the Oxford Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) to monitor the current pandemic and become an in-pandemic research platform.

Objective:

Methods:
We described our domain-specific use case. The actor was the RCGP RSC sentinel network; the system the course of the COVID-19 pandemic; the outcomes the spread and effect of mitigation measures. We used our established three-step method to develop the ontology, separating ontological concept development from code mapping and data extract validation. We developed a coding system–independent COVID-19 case identification algorithm. As there were no gold standard pandemic surveillance ontologies, we conducted a rapid Delphi consensus exercise through the International Medical Informatics Association (IMIA) Primary Health Care Informatics working group and extended networks.

Results:
Our use case domains included primary care, public health, virology, clinical research and clinical informatics. Our ontology supported: (1) Case identification, microbiological sampling and health outcomes at both an individual practice and national level; (2) Feedback through a dashboard; (3) A national observatory, (4) Regular updates for Public Health England, and (5) Transformation of the sentinel network to be a trial platform. We have identified a total of 8,627 people with a definite COVID-19 status, 4,240 with probable, and 59,147 people with possible COVID-19, within the RCGP RSC network (N=5,056,075).

Conclusions:
The underpinning structure of our ontological approach has coped with multiple clinical coding challenges. At a time when there is uncertainty about international comparisons, clarity about the basis on which case definitions and outcomes are made from routine data is essential.

Funding: Public Health England, *Wellcome.*

In-pandemic development of an application ontology for COVID-19 surveillance in a primary care sentinel network

Introduction

The coronavirus-2019 (COVID-19) pandemic has many features of a complex system.[1,2] Complexities include repeated name changes of both the causative organism and associated disease,[3–6] evolving understanding of core clinical features at presentation,[7,8] and differing rates of testing and approaches to outcome reporting between countries.[9,10] This complexity presents a significant challenge for consistent clinical coding within computerised medical records (CMR) systems.[11]

Creating an ontology for COVID-19 surveillance should help to facilitate reproducibility and interoperability between various key stakeholders; from clinicians and epidemiologists to data scientists and software developers. Ontologies are formalisations of conceptualisations and exist in reference or application formats.[12] Reference ontologies are at a domain level and describe a group of related concepts. Application ontologies are more specific and are used when modelling across multiple domains.[13] Application ontologies should be evaluated against a testable use case, which represents the scope and requirements of the specific application.[14,15] The emergence of a new disease means that corresponding original ontologies need to be developed.

We report the development of a COVID-19 application ontology using the Oxford Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) network's adaptations to COVID-19 as its use case. The RCGP RSC, is an established primary care sentinel network, which extracts pseudonymised data from a nationally representative sample of over 500 general practices twice weekly (N=4,857,084).[16] RCGP RSC has collaborated with Public Health England (PHE) for over 50 years, conducting influenza and respiratory disease surveillance and vaccine effectiveness studies.[17,18] The RCGP RSC has extended these routine surveillance activities to include monitoring the spread of COVID-19, assessing the effectiveness of containment measures, and becoming a platform for an in-pandemic COVID-19 trial.[19,20]

Building on our previous experience of developing ontologies,[12,21,22] we created an application ontology for COVID-19 extended surveillance.

Methods

Our application ontology was developed in three stages: Stage 1: creating and testing our use case; Stage 2: Developing the COVID-19 surveillance ontology; and Stage 3: External validation using a rapid Delphi consensus exercise. We classified this as an application ontology because the system crosses a range of domains to deliver specific goals. The data source was routine primary care data from the RCGP RSC sentinel network, combined with virology and serology sampling data from PHE. Additionally, a practice dashboard and a practice liaison team helped ensure data quality.[23] Episode type, whether a case is a first or incident case or follow-up is important for surveillance, whilst we can infer this from records it is better it is collected as primary information.[24]

Stage 1: Creating and testing the use case

We created a testable narrative use case for COVID-19 surveillance, using previously described methods.[25,26] The primary actor was the RCGP RSC, the system it interacts with was the national response to the COVID-19 pandemic, and its outcomes monitoring spread and effect of mitigation measures.

The use case has been progressively implemented in-pandemic. We report our implementation across the domains identified. As our ontology developed and formalised, there was post hoc checking to ensure extracts were ontology compliant.

Stage 2: Developing the COVID-19 surveillance ontology

We developed an application ontology to support extended surveillance using routine CMR data. The terminology and clinical understanding and response to COVID-19 were rapidly changing during the period of development. The ontology has inbuilt flexibility to accommodate these and further changes.

We used our three-step ontological process to identify codes to meet our requirements:[12,21,22]

- Step 1: the ontology layer, defines relevant COVID-19 surveillance concepts (COVID-19) and may include exposure, investigations, diagnoses, or other ‘processes of care’. Part of our ontological process is to iterate whether cases identified are definite, probable, or possible; based on the specificity of the codes used, an approach developed in diabetes research.[27]
- Step 2: the coding layer, applies concepts of the ontology layer to the specific coding system used in the CMR. Individual codes are classified as having direct, partial or no clear mapping to the criteria considered.[28] In this case, we extended this to exclude suspected cases where there was a subsequent negative test. Post hoc data validation was largely via our practice liaison team. One member of the team is entirely dedicated to ensuring data quality providing anticipatory training and coding aids, and a responsive service.
- Step 3: the logical data extract model, systematically tests the codes identified to ensure that data outputs are consistent with study requirements.

We wanted our resource to be findable, accessible, interoperable and reusable (FAIR),[29] so used standard tools in its development: namely the Protégé ontology development environment[30] and Web Ontology language (OWL).[31]

The scope of the ontology included:

- Demographic details, including age, gender, ethnicity, deprivation, rurality and linking key identifiers.
- Recording of monitored conditions and key clinical features – symptoms and signs.
- Relevant comorbidities and risk factors.
- Tests and test results – COVID-19-specific and test results that might imply susceptibility or resilience.
- Key outcome measures including hospitalization, oxygen therapy, intensive care admission and mortality.

Stage 3: External evaluation of the ontology

We carried out a rapid Delphi consensus exercise by inviting a panel (n=9) of international primary care clinicians and informaticians through the International Medical Informatics Association (IMIA) Primary Health Care Informatics work group and extended networks.[32,33] The consensus exercise consisted of 3 rounds:

1. Round 1: We shared our initial ontology and requested panel members to inform us about additional concepts that were not present in the ontology but present in their clinical workflows. In order to facilitate rapid consensus, we used email correspondence for this stage.
2. Round 2: We shared the revised ontology with panel members who were asked to indicate their level of agreement, on a 5-point Likert scale, to statements related to the coverage of concepts and applicability of the ontology to their primary care system. This was delivered through on an online survey (panel members and questions in supplementary material). Consensus was defined as ≥ 80% agreement. Statements not meeting 80% agreement were modified according to feedback provided by the expert panel and redistributed to panelists for round 3.
3. Round 3: We conducted an online discussion to review and approve the final ontology.

Ethical considerations

COVID-19 surveillance is carried out by RCGP RSC in collaboration with PHE, and approved under Regulation 3 of The Health Service (Control of Patient Information) Regulations 2002 by PHE's Caldicott Guardian.[34] No specific permissions were needed for our ontology development as no additional processing of data was required.

Results

Stage 1: Creating and testing the use case

We developed a summary narrative use case (Table 1). The success scenarios listed are goals we want to achieve.

Table 1: Summary narrative use case

Actor	RCGP RSC	
Scope	Delivery of COVID-19 surveillance and research	
Level	Healthcare system wide	
Stakeholders and interests	Patients and public	Safe and timely guidance through pandemic
	General Practices	Professional interest, payment, providing high-quality, evidence-based care
	PHE	Need data to predict transmission, monitor effectiveness of interventions
	RCGP	Care for/protect members, contribute to pandemic response
	Primary Care Clinical Trials Unit	Data governance policies control which data can be viewed at Recruit to trial to mitigate COVID-19
Precondition	Legal basis, Permissions for data extracts, Data extraction and analytics capability within the network	
Minimal guarantee	Delivery of data and analytics at pre-pandemic scale	
Success guarantee	1. Larger network with high quality data 2. Outputs to meet changed requirements during the pandemic 3. Authoritative source of primary care data, evidenced by academic publication	
Main success scenario	1. High quality primary care data, feedback to practices via customized dashboards 2. Representative sampling of virology and serology with the collection of the specified number of samples (900 virology, 1000 serology per week). 3. Twice weekly data feeds to PHE to meet their data requirements. 4. National observatories and weekly return that represent the impact of COVID-19 5. Ensure that we fully recruit to the PRINCIPLE and other trials through the Oxford-RCGP RSC system	

	6. High quality publication of the learning from surveillance
Extensions	1. Trebling the number of virology practices (we have gone from 100 to 300 virology sampling practices, from 10 to 200 serology sampling practices) 2. Adjusting to the effect of lockdown on: a) Extending the network to over 1,000 practices to support large-scale clinical trials, embedded in clinical practice, e.g. recruitment into PRINCIPLE trial b) Sampling all eligible patients – due reduced number seen on surgery premises 3. Post convalescent serology – we will collect convalescent serology at 28 days from a wide range of practices 4. Managing unforeseen problems: a) Refusal of some post offices to allow sample postage b) Postage delays c) Swab supply problems 5. Piloting new methods of swab delivery to patients 6. Add resilience to the surveillance system a) Human resilience – extending data team and support b) System resilience – direct feeds from major CMR suppliers 7. Other studies: large numbers of study requests that need managing

The success scenarios and extensions reflect the cross-domain activities within the use case. We list the outcomes across five domains: primary care, public health, virology, clinical research and clinical informatics (Table 2). We implemented our ontology through practical activities across these domains.

Table 2: Application use case outcomes by domain

Domain	**Description**	**Outcomes**
Primary care	• COVID 19- Observatory – temporal and geographic surveillance • COVID-19 dashboard – practice level data quality	• Data quality feedback to practices • Feedback from practices
Public health	• COVID – supplementary report • Public health policy - containment measures	• Trends of community transmission after social distancing ends • Estimates of COVID-19-related community morbidity and mortality
Virology	Swabbing – Investigation • Virology • Serology	• Virologically confirmed incidence • Representative collection of serology for sero-epidemiology • Ordering stock control and swab and virology container supply
Clinical Research	Recruitment to clinical trials	• Health outcomes: chest infections, hospitalisation, ICU, mechanical ventilation, oxygen therapy and death

Clinical informatics	• IG – legal basis, data sharing agreements, contracts • Hardware and its resilience. • Semantic interoperability across domains	• Data quality, usability, FAQs – continuous improvement of our interface. • Adaptability with changing clinical knowledge • Ontology with annotations to clinical terms/codes

Primary care domain: data quality and feedback to general practice (COVID-19 dashboard)

Our COVID-19 dashboard presented weekly data on respiratory conditions to practices within the sentinel network. Data were presented on incidence of COVID-19 for the individual practice, and at the regional and national levels for reference, along with rates of other respiratory infections (Figure 1). Post implementation feedback had to keep pace with multiple data changes and different timetables of code releases between CMR system providers. It included constant updating of coding prompt cards.[35]. It also required liaison with computer template developers to change design to incorporate episode type.

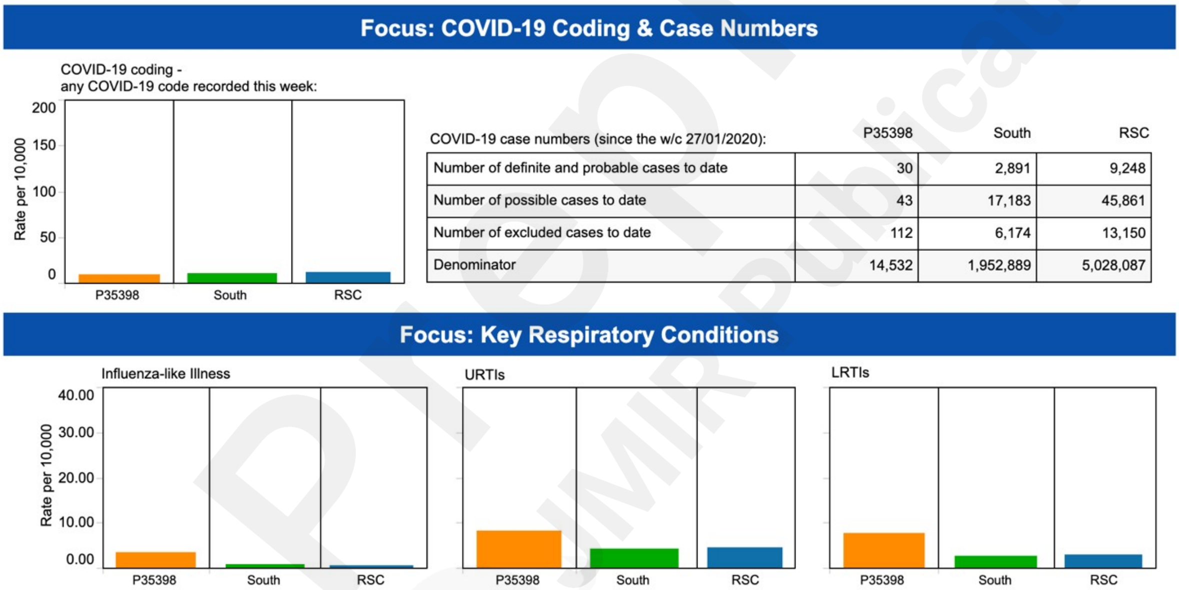


Figure 1: COVID-19 dashboard for each RCGP RSC network practice
The column starting P35398 is that practices data; "South" is their region; RSC is the rate across the whole network. URTI & LRTI are Upper and Lower Respiratory Infection, respectively.[36]

Public health domain: data visualisation with COVID-19 Observatory

Our ontology ensured consistency between our classic weekly return, which now includes COVID-19 surveillance. In addition, we developed customised outputs for epidemiologists at PHE and an observatory to present data on the incidence of COVID-19 across the network. This is based on coding described in the ontological layer (Figure 4) and presents incidence rate per

10,000 of COVID-19. Up to the week commencing 09/06/2020, we have identified a total of 8,627 people with definite COVID-19, 4,240 with probable, and 59,147 people with possible COVID-19, within the RCGP RSC network (N=5,056,075) (Figure 2).

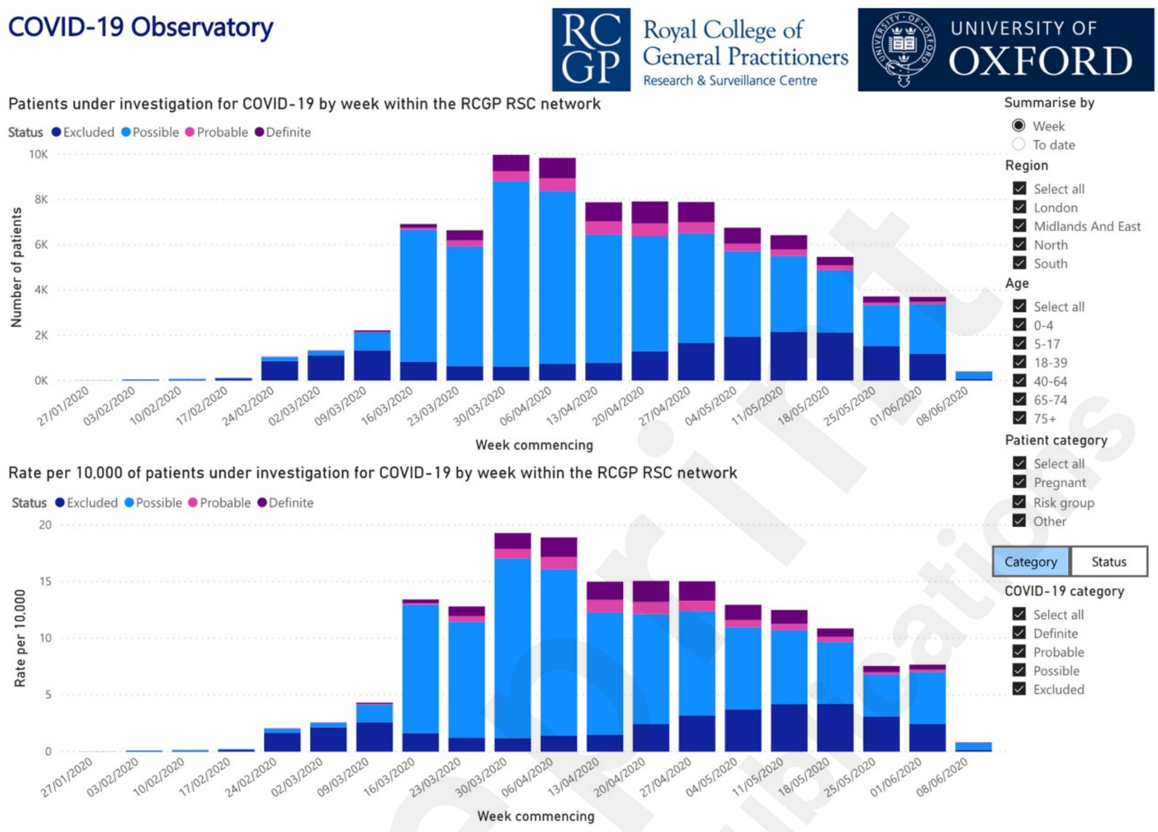


Figure 2: Oxford RCGP RSC interactive COVID-19 observatory
Users can select the cumulative or week-by-week view of data, and visualise data by age-band, region, risk group, and COVID-19 status – Definite, Probable, Possible and Excluded.[37]

The biggest area of challenge was attribution of codes to certainty of diagnosis. We have had to evolve this with coding system changes. (See supplementary material for our final SNOMED CT concept list.)

Virology domain: weekly virologic surveillance reports

Similarly, our ontology drove the consistent extension of our virology reporting. Sound data structures have also been important because the number of participating virology sampling practices trebled from 100 to 300 to provide more data. The weekly virology report provides a visualisation of the absolute number and rate per 10,000 by week of the swabs taken, combined with the matched week from previous year's figures for background context. There is a similar observatory for serology (included in supplementary material).

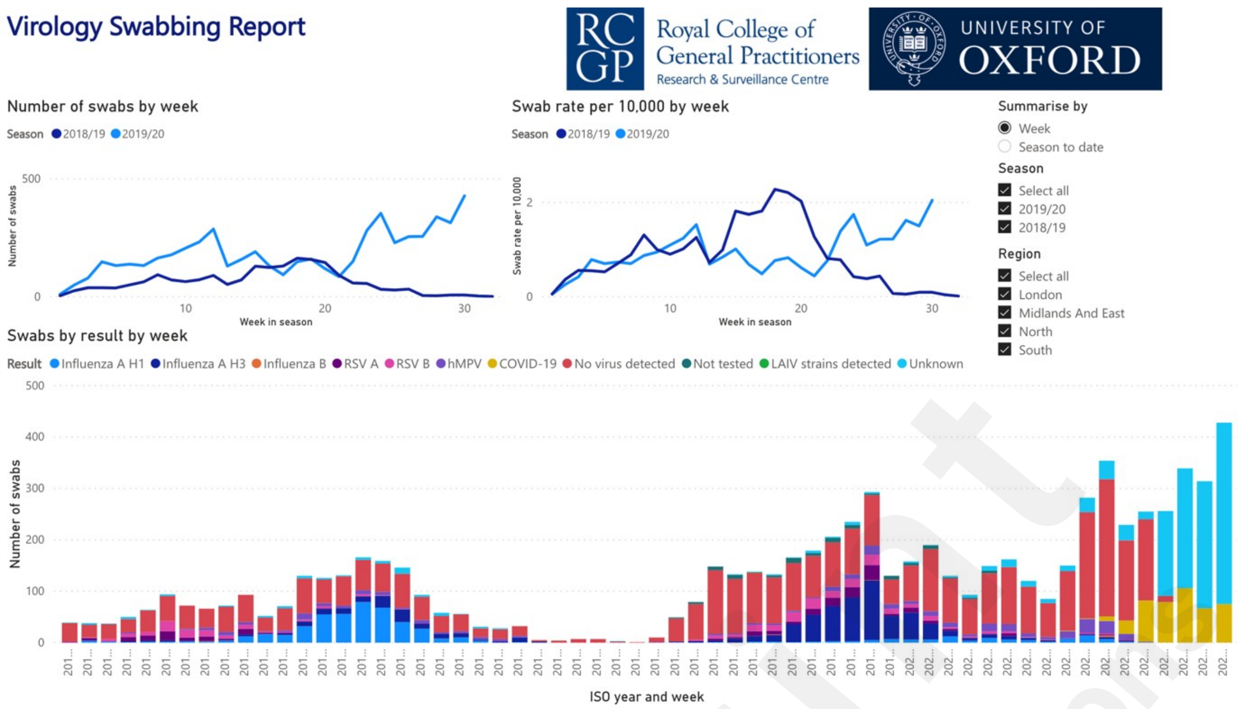


Figure 3: **Oxford RCGP RSC interactive virology swabbing report** (link)
Users can look at the cumulative, or by week, or compare with last year, and look by infected organism or region. "Unknown" is used where no testing is done, currently samples are only tested for COVID-19.

Clinical research domain: participation in observational and interventional studies

The COVID-19 surveillance application ontology supported consistent reporting of findings in observational and interventional clinical research. We have a series of ongoing observational studies, the first of which has reported results.[38] The network is also supporting the PRINCIPLE trial, a UK platform randomised control trial (RCT) of interventions for COVID-19 in Primary Care. The study is assessing the effectiveness of trial treatments in reducing the need for hospital admission and death for patients with suspected COVID-19 infection aged ≥50 years with serious comorbidity, and aged ≥65 with or without comorbidity.[20] To date, 830 practices have signed up, with 415 patients randomised; 468 (56.4%) of these are RCGP RSC practices, and they have recruited 342 (82.4%) of the included patients so far.

Clinical informatics domain: creating the COVID-19 ontology

The annotated application ontology was published on the BioPortal Ontology Repository:[39] and will continue to be developed as our understanding of COVID-19 advances and new interventions, for example vaccination, are introduced. The detail of the ontological development is set out in Stage 2, reporting our three-step process.

Stage 2: Developing the COVID-19 surveillance ontology

Step 1: Ontological layer

We reviewed emerging case definitions of COVID-19 to identify key concepts used for case ascertainment and their relationships. Concepts included in the ontology were consistent with the WHO data dictionary for COVID-19 case-based reporting.[40]

We have limited our presentation of results to the case definition of COVID-19. This has involved grouping concepts into: (1) Definite, which include definitive codes for a laboratory confirmed case of COVID-19: (2) Probable, which included clinical diagnosis of COVID-19 and use of out-of-date codes created in the previous SARS outbreak; (3) Possible, contain a range of coding alternatives related to suspected COVID-19, investigation but no result and exposure codes; and (4) Excluded, where a test requested is reported as negative. At the individual level the tests work hierarchically, with the most specific driving the categorisation.

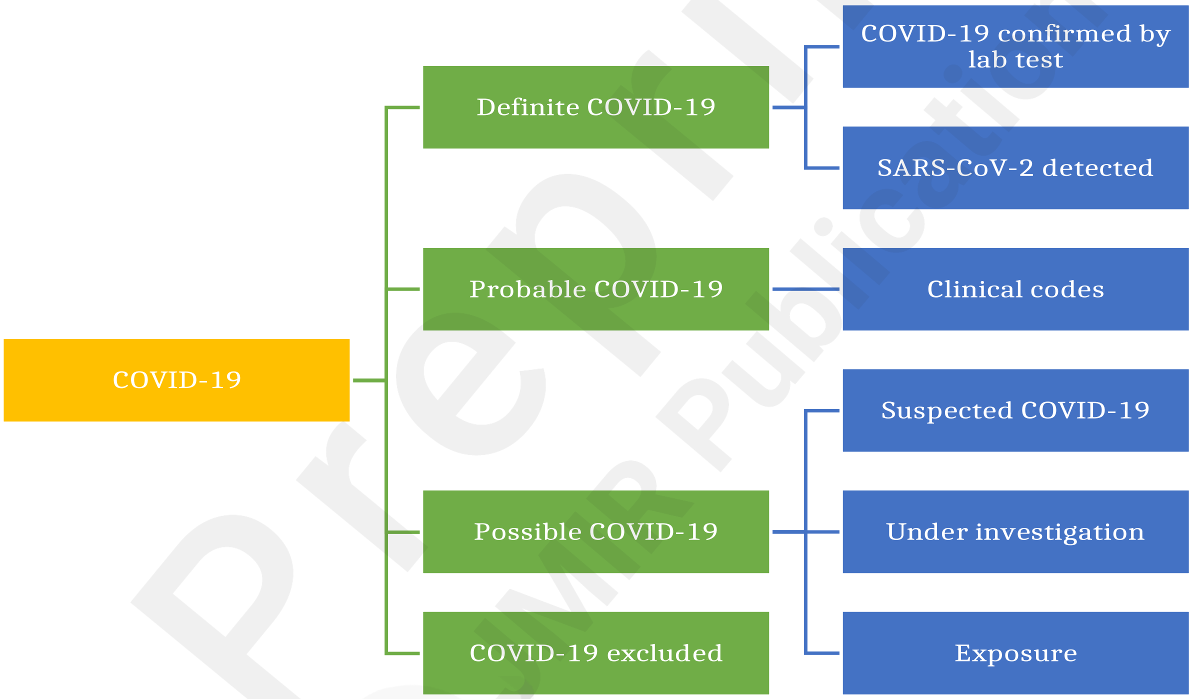


Figure 4: Foundational ontological concepts used for COVID-19 surveillance

Step 2: Coding layer

We completed a dynamic process of mapping clinical terminology codes to concepts which emerged from our ontological layer.

The National Health Service (NHS) uses the UK SNOMED CT system of coding, which is normally only updated twice yearly. In early February 2020, there were no clinical codes specific to COVID-19. Initially, CMR suppliers created five new system wide local codes to support essential COVID-19-related recording within a week of being requested.[11,19] Subsequently, two emergency releases of novel COVID-19-related UK SNOMED CT codes were developed through a rapid consultation process conducted by the NHS Digital (NHSD) Information Representation Service (IReS),[41] as greater clinical insight into COVID-19 and stability around nomenclature emerged. These UK SNOMED CT concepts were developed independently of international SNOMED CT terminology development, however, this open-source ontology can be mapped to international terms with ease. We iteratively annotated the ontological concepts with these stepwise-released COVID-19 SNOMED CT clinical concepts.

Table 3: Migration across SNOMED CT concepts released from February to May 2020.

Clinical concepts that should be coded in CMR – from ontological layer	**Temporary codes** ***–used until replacement with SARS-CoV-2/COVID-19 concepts***	**Final SNOMED CT Description**
Covid-19 definite	Confirmed 2019 nCoV (Wuhan) infection *or* Confirmed 2019 nCoV (novel coronavirus) infection	COVID-19 confirmed by laboratory test SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) detected
Covid-19 probable	*No specific codes*	COVID-19 COVID-19 confirmed by clinical diagnostic criteria
Covid-19 possible		
- Exposure to infectious agent	Exposure to 2019 nCoV (Wuhan) infection *or* Exposure to 2019 nCoV (novel coronavirus) infection	Exposure to SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) infection
- Suspected infection	Suspected 2019 nCoV (Wuhan) infection *or* Suspected 2019 nCoV (novel coronavirus) infection	Suspected COVID-19
- Test for infectious agent offered or taken	*No specific codes*	Swab for SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) taken by healthcare professional Self-taken swab for SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) offered Self-taken swab for SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) completed
	Tested for 2019 nCoV (Wuhan) infection *or* Tested for 2019 nCoV (novel coronavirus) infection	
Covid-19 excluded	Excluded 2019 nCoV (Wuhan) infection *or* Excluded 2019 nCoV (novel coronavirus) infection	COVID-19 excluded COVID-19 excluded by laboratory test SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) not detected

Step 3: Logical data extract layer

We incorporated the annotated ontology into the routine surveillance platform of the RCGP RSC data. The ontology identified various states of COVID-19 diagnosis in the incoming data feeds used for surveillance. We conducted week-by-week analysis of incoming data modifying our outputs to take account of supplier-specific changes in reporting. We are planning for cloud-based extracts and customised extracts from individual CMR vendors; to do this we are creating an Oxford RCGP Clinical Informatics Digital Hub (ORCHID).[42]

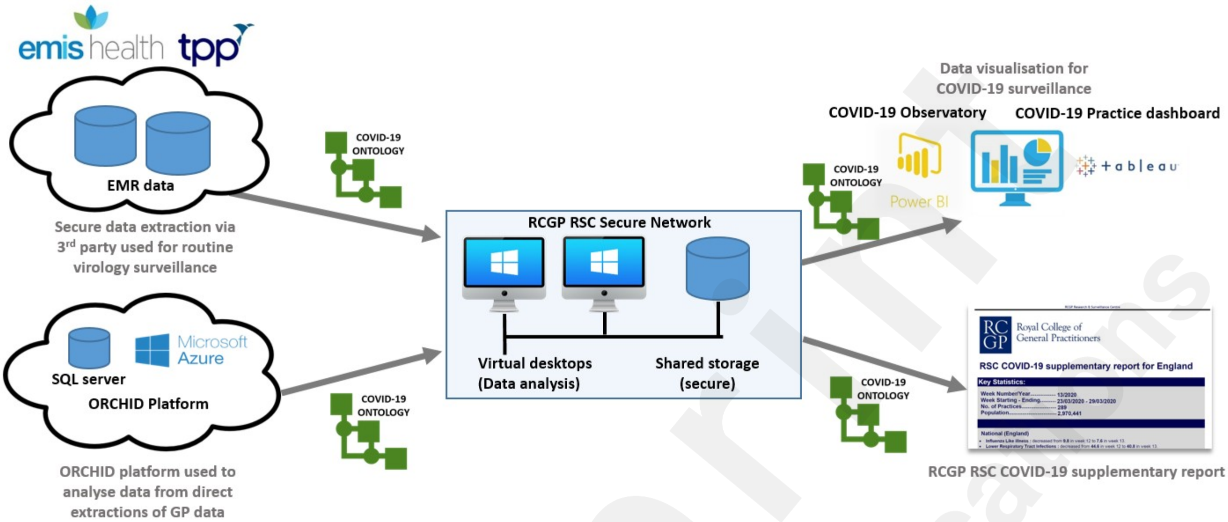


Figure 5: Use of the COVID-19 surveillance ontology across the RCGP RSC processes to achieve semantic consistency in data extraction, visualisations and surveillance reports.

External evaluation of the ontology

Whilst we obtained a good consensus in our Delphi exercise, there was important learning and priorities flagged for development. Consensus was obtained for 7/8 (87.5%) of the statements related to coverage of concepts under the upper level headings of the COVID-19 ontology. All panel members bar one agreed with statements relating to the applicability of the ontology for case finding activities in their local primary care setting (Table 3). Input from panel members guided expansion of the concepts related to statements not reaching consensus and this was reviewed by panel members in round 3 of the Delphi exercise.

Table 3: Number of responses and % agreement (strongly agree/agree) to the following:

Please indicate your level of agreement with the coverage of concepts given under each upper level heading of the COVID-19 surveillance ontology.

Upper level heading	SD	D	N	A	SA	%A
Symptoms and signs	0	1	0	3	5	88·9
Past medical history/At-risk conditions	0	0	3	2	4	66·66
Exposure	0	0	0	2	7	100
Investigations	0	0	0	5	4	100
COVID-19 case status	0	0	0	1	8	100
Interventions	0	1	0	3	5	88·9
Process of care	0	0	1	1	7	88·9
Outcomes	0	0	0	5	4	100

Statement	SD	D	N	A	SA	%A
The COVID-19 ontology in its current format is suitable for COVID-19 case ascertainment in my local primary care setting.	0	0	1	5	3	88·9

(SD = strongly disagree, D = disagree, N = neither agree or disagree, A = agree, SA = strongly agree, %A = % agreement)

Softer important discussion points emerged, for example, our symptom collection is relatively poorly developed and that there remains uncertainty about risk and protective factors. There was strong feeling among one expert that vaccination and exposures should be part of the ontology; these were subsequently added.

Discussion

Principal findings

We rapidly developed an application ontology in-pandemic to support extended surveillance and research activities across the five clinical and informatics domains described in our use case. This application ontology has provided a framework which we have used to help ensure the reliability and consistency of our outputs at a time of change. This iterative ontological approach is flexible and robust enough to match the pace and direction of the evolving clinical landscape of COVID-19.

The focus of our work has been on case identification and associated test results, as these are the foundations on which epidemiological and interventional studies are based. We felt it appropriate to give a flag of the certainty with which a diagnosis is made. We have already used this ontology in observational and interventional studies.[20,38]

The separation of the coding layer from the ontological (conceptual) layer allows surveillance to be resilient while new case definitions and clinical codes are added to general practice CMR systems. This approach ensures transparency in case definitions used for reporting and facilitates clear communication by allowing clinicians, database developers (involved in extracting data from practices' data sources) and practice liaison officers (who advise practices about data recording best practices) to maintain consistency within an organisation.

This application ontology could easily and rapidly be adapted for COVID-19 surveillance and clinical research in various other countries and healthcare networks. As the COVID-19 pandemic continues, there is enormous global pressure on healthcare systems to understand trends in incidence rates and conduct high-quality research, this ontology is open-source and can be mapped onto local clinical coding systems to permit consistency in analyses.

Comparison with previous literature

To our knowledge, this is the first time that a systematic ontological approach has been developed in-pandemic for extended disease surveillance, using structured routine clinical data. This application ontology aligns with previous clinical informatics literature on application ontology engineering and validation through the testable use case approach.[43,44]

There are other pandemic surveillance systems which look at open-source, unstructured data, such as media reports and clusters of symptom-related internet searches, extracting information of epidemiological relevance.[45] Examples of such systems include BioCaster,[46] the Global Public Health Intelligence Network (GPHIN),[47] ProMed,[48] and HealthMap.[49] The latter three systems are working under the WHO collaborative, the Epidemic Intelligence from Open Sources (EIOS), which played a role in the identification of the COVID-19 outbreak from early media reports from China in December 2019.[50] Some of the event-based pandemic surveillance systems have published ontological foundations in the public health and surveillance domains.[46,51,52] While useful for providing supplementary information to epidemiologists on the emergence of an outbreak in real-time, these knowledge representations do not specifically address the types of information described in clinical data, such as presenting complaint, comorbidities, virology or health outcomes.

There are very limited studies of data platforms' performance within integrated clinical surveillance systems.[45] The lack of accurate and available data to underpin epidemic forecasting in emerging outbreaks has been highlighted.[53]

We found no literature using an ontological approach for COVID-19 surveillance. There are domain ontologies related to coronavirus published on BioPortal. The first focuses on the wider *coronaviridae* family and their biochemical and microbiological properties,[54] while the second was developed to provide semantic assistance for clinical research form completion.[55] None were designed to integrate the various clinical data streams necessary to carry out COVID-19 surveillance.

Strengths and limitations

The three-step iterative ontological process which we have implemented has proven to be suitably flexible to cope with the changes in COVID-19 terminology and CMR system codes. A further strength was the implementation and deployment of this ontology, considering the FAIR guiding principles.[29] The ontology is discoverable and accessible on the BioPortal ontology repository. This application ontology, built using best practices around defining and testing a use case, is inherently interoperable and reusable.[29] In the absence of a gold standard infectious disease surveillance ontology, we believe our attempts at achieving a degree of consensus and external validity from a range of international experts in the field of clinical informatics and primary care as a major strength of the current study. While the Delphi panel size was relatively small and a limitation, we purposefully selected panel members from a range of countries with varied clinical coding systems

We focussed on case finding and results, we now need to turn our attention to presenting symptoms, particularly looking to focus on those that may be of prognostic value and emerging treatments including vaccination. Additional limitations were its development in a single sentinel system and it was not developed ready to integrate into a common data model.[56]

Call for further research

We welcome any requests for information on applying our COVID-19 surveillance application ontology to other healthcare settings, both domestically and internationally.

Conclusions

We have created a COVID-19 application ontology, its strengths are its speed of development, being openly shared via BioPortal, and its adaptability. The limitations are its development in a single sentinel network and its current limited focus. The ontology should make conclusions based on primary care sentinel data more transparent and facilitate pooled analyses in COVID-19 surveillance and research.

Contributions

SdeL conceived the need for this ontology with important input from HS, JW, DE and DMcG. HS and DMcG produced an initial draft and SdeL completed the first complete manuscript. All authors contributed to the scope of the ontology, contributed comments, and read and approved the final version.

Declaration of interests

Acknowledgements

The practices and patients of the Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC), who allowed their pseudonymised clinical medical records to be used for this work. We would also like to acknowledge the members of the Primary Health Care Informatics Working Group of the International Medical Informatics Association for their contribution to validate the ontology.

References

1 Lipsitz LA. Understanding Health Care as a Complex System. *JAMA* 2012. DOI:10.1001/jama.2012.7551.

2 Sturmberg J, Lanham HJ. Understanding health care delivery as a complex system: Achieving best possible health outcomes for individuals and communities by focusing on interdependencies. *J Eval Clin Pract* 2014. DOI:10.1111/jep.12142.

3 Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet. 2020. DOI:10.1016/S0140-6736(20)30185-9.

4 WHO. Novel Coronavirus(2019-nCoV) Situation Report - 10. .

5 Gorbalenya AE. Severe acute respiratory syndrome-related coronavirus – The species and its viruses, a statement of the Coronavirus Study Group. *bioRxiv* 2020. DOI:10.1101/2020.02.07.937862.

6 WHO. Novel Coronavirus (2019-nCoV) Situation Report – 22. 2020 https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf.

7 Huang C, Wang Y, Li X, *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020. DOI:10.1016/S0140-6736(20)30183-5.

8 Giacomelli A, Pezzati L, Conti F, *et al.* Self-reported olfactory and taste disorders in SARS-CoV-2 patients: a cross-sectional study. *Clin Infect Dis* 2020. DOI:10.1093/cid/ciaa330.

9 Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. Real estimates of mortality following COVID-19 infection. Lancet Infect. Dis. 2020. DOI:10.1016/S1473-3099(20)30195-X.

10 Verity R, Okell LC, Dorigatti I, *et al.* Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* 2020. DOI:10.1016/s1473-3099(20)30243-7.

11 de Lusignan S, Williams J. To monitor the COVID-19 pandemic we need better quality primary care data. *BJGP Open* 2020.

12 De Lusignan S. In this issue: Ontologies a key concept in informatics and key for open definition of cases, exposures and outcome measures. J. Innov. Heal. Informatics. 2015. DOI:10.14236/jhi.v22i2.170.

13 Musen MA. Domain ontologies in software engineering: use of protégé with the EON architecture. *Methods Inf Med* 1998. DOI:10.1055/s-0038-1634543.

14 Malone J, Parkinson H. Reference and Application Ontologies. *Ontogenesis* 2010.

15 Kumarapeli P, De Lusignan S, Ellis T, Jones B. Using Unified Modelling Language (UML) as a process-modelling technique for clinical-research process improvement. *Informatics Heal Soc Care* 2007. DOI:10.1080/14639230601097705.

16 Correa A, Hinton W, Govern AM, *et al.* Royal college of general practitioners research and surveillance centre (RCGP RSC) sentinel network: A cohort profile. *BMJ Open* 2016. DOI:10.1136/bmjopen-2016-011092.
17 De Lusignan S, Correa A, Smith GE, *et al.* RCGP Research and Surveillance Centre: 50 years' surveillance of influenza, infections, and respiratory conditions. Br. J. Gen. Pract. 2017. DOI:10.3399/bjgp17X692645.
18 Pebody R, Warburton F, Ellis J, *et al.* Effectiveness of seasonal influenza vaccine for adults and children in preventing laboratory-confirmed influenza in primary care in the United Kingdom: 2015/16 end-ofseason results. *Eurosurveillance* 2016. DOI:10.2807/1560-7917.ES.2016.21.38.30348.
19 de Lusignan S, Lopez Bernal J, Zambon M, *et al.* Emergence of a Novel Coronavirus (COVID-19): Protocol for Extending Surveillance Used by the Royal College of General Practitioners Research and Surveillance Centre and Public Health England. *JMIR Public Heal Surveill* 2020; **6**: e18606.
20 Butler C. ISRCTN86534580: A trial evaluating treatments for suspected coronavirus infection in people aged 50 years and above with pre-existing conditions and those aged 65 years and above. 2020. DOI:https://doi.org/10.1186/ISRCTN86534580.
21 Cole NI, Liyanage H, Suckling RJ, *et al.* An ontological approach to identifying cases of chronic kidney disease from routine primary care data: A cross-sectional study. *BMC Nephrol* 2018. DOI:10.1186/s12882-018-0882-9.
22 Liyanage H, Williams J, Byford R, De Lusignan S. Ontology to identify pregnant women in electronic health records: Primary care sentinel network database study. *BMJ Heal Care Informatics* 2019. DOI:10.1136/bmjhci-2019-100013.
23 Smith S, Morbey R, de Lusignan S, Pebody RG, Smith GE, Elliot AJ. Investigating regional variation of respiratory infections in a general practice syndromic surveillance system. *J Public Health (Bangkok)* 2020. DOI:10.1093/pubmed/fdaa014.
24 Smith N, Livina V, Byford R, Ferreira F, Yonova I, De Lusignan S. Automated differentiation of incident and prevalent cases in primary care computerised medical records (CMR). In: Studies in Health Technology and Informatics. 2018. DOI:10.3233/978-1-61499-852-5-151.
25 Cockburn A. Writing effective use cases. Addison-Wesley Professional, 2000.
26 Liyanage H, de Lusignan S, Liaw ST, *et al.* Big Data Usage Patterns in the Health Care Domain: A Use Case Driven Approach Applied to the Assessment of Vaccination Benefits and Risks. Contribution of the IMIA Primary Healthcare Working Group. Yearb. Med. Inform. 2014. DOI:10.15265/IY-2014-0016.
27 De Lusignan S, Khunti K, Belsey J, *et al.* A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: A pilot and validation study of routinely collected data. *Diabet Med* 2010. DOI:10.1111/j.1464-5491.2009.02917.x.
28 Rollason W, Khunti K, De Lusignan S. Variation in the recording of diabetes diagnostic data in primary care computer systems: Implications for the quality of care. *Inform Prim Care* 2009. DOI:10.14236/jhi.v17i2.723.
29 Wilkinson MD, Dumontier M, Aalbersberg IjJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016. DOI:10.1038/sdata.2016.18.
30 Musen MA. The Protégé Project: A Look Back and a Look Forward. *AI Matters* 2015. DOI:10.1145/2757001.2757003.
31 McGuinness DL, van Harmelen F. OWL Web Ontology Language Overview. W3C Recomm. 2004.
32 Liaw S-T, Liyanage H, Kuziemsky C, *et al.* Ethical Use of Electronic Health Record Data and Artificial Intelligence: Recommendations of the Primary Care Informatics Working Group

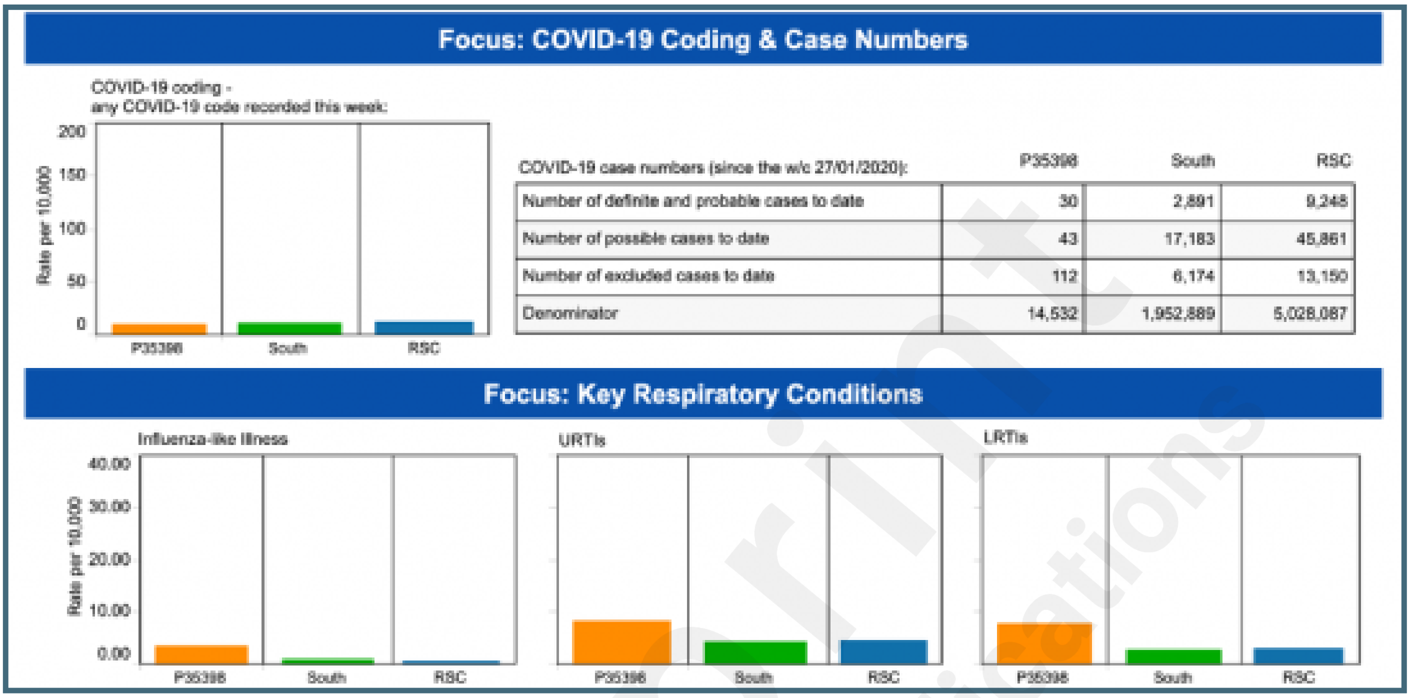
of the International Medical Informatics Association. *Yearb Med Inform* 2020; published online April 17. DOI:10.1055/s-0040-1701980.
33 Liyanage H, Liaw ST, Jonnagaddala J, *et al.* Artificial Intelligence in Primary Health Care: Perceptions, Issues, and Challenges. In: Yearbook of medical informatics. 2019. DOI:10.1055/s-0039-1677901.
34 Taylor MJ. Legal bases for disclosing confidential patient information for public health: Distinguishing between health protection and health improvement. *Med Law Rev* 2015. DOI:10.1093/medlaw/fwv018.
35 RCGP RSC. RCGP RSC COVID-19 Surveillance. 2020. https://clininf.eu/index.php/cov-19/ (accessed June 13, 2020).
36 RCGP RSC. RCGP RSC COVID-19 Practice-Level Dashboard. 2020. https://clininf.eu/index.php/covid/ (accessed June 13, 2020).
37 RCGP RSC. COVID-19 Observatory. Publ. 2020. https://app.powerbi.com/view?r=eyJrIjoiZTU5ZDE5MGYtMzUzMy00ZjRmLTg4MGEtMTM3ZGJiZDNhODFkIiwidCI6IjZiOTAyNjkzLTEwN zQtNDBhYS05ZTIxLWQ4OTQ0NmEyZWJiNSIsImMiOjh9 (accessed March 31, 2020).
38 De Lusignan S, Dorward J, Correa A, Joy M. Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study. *Lancet Infect Dis* 2020; **May 15 ePu**. DOI:doi.org/10.1016/S1473-3099(20)30371-6.
39 Liyanage H, de Lusignan S, Williams J. COVID-19 Surveillance Ontology. 2020. https://bioportal.bioontology.org/ontologies/COVID19 (accessed May 4, 2020).
40 WHO. Global surveillance for COVID-19 caused by human infection with COVID-19 virus: Interim guidance 20 March 2020. 2020. https://apps.who.int/iris/bitstream/handle/10665/331506/WHO-2019-nCoV-SurveillanceGuidance-2020.6-eng.pdf (accessed March 30, 2020).
41 NHSDigital. Clinical guidance on COVID-19 SNOMED codes. 2020. https://theprsb.org/standards/covid-19snomedcodes/ (accessed April 10, 2020).
42 de Lusignan S, Jones N, Dorward J, *et al.* Oxford Royal College Of General Practitioners Clinical Informatics Digital Hub: Rapid Innovation To Deliver Extended COVID-19 Surveillance And Trial Platforms. *JMIR Public Heal Surveill* 2020. DOI:10.2196/preprints.19773.
43 Malone J, Holloway E, Adamusiak T, *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 2010. DOI:10.1093/bioinformatics/btq099.
44 Doing-Harris K, Bray BE, Thackeray A, *et al.* Development of a cardiac-centered frailty ontology. *J Biomed Semantics* 2019. DOI:10.1186/s13326-019-0195-3.
45 Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: A systematic review. Milbank Q. 2014. DOI:10.1111/1468-0009.12038.
46 Collier N, Doan S, Kawazoe A, *et al.* BioCaster: Detecting public health rumors with a Web-based text mining system. *Bioinformatics* 2008. DOI:10.1093/bioinformatics/btn534.
47 Dion M, AbdelMalik P, Mawudeku A. Big Data and the Global Public Health Intelligence Network (GPHIN). *Canada Commun Dis Rep* 2015. DOI:10.14745/ccdr.v41i09a02.
48 Yu VL, Madoff LC. ProMED-mail: An Early Warning System for Emerging Diseases. *Clin Infect Dis* 2004. DOI:10.1086/422003.
49 Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media

Reports. *J Am Med Informatics Assoc* 2008. DOI:10.1197/jamia.M2544.
50 WHO. Epidemic Intelligence from Open Sources (EIOS). 2020. https://www.who.int/eios (accessed April 3, 2020).
51 Okhmatovskaia A, Chapman WW, Collier N, Espino J BD. SSO: The Syndromic Surveillance Ontology. 2009. https://ncbo.bioontology.org/sites/default/files/SSO.pdf (accessed April 3, 2020).
52 Crubézy M, O'Connor M, Pincus Z, Musen MA, Buckeridge DL. Ontology-centered syndromic surveillance for bioterrorism. IEEE Intell. Syst. 2005. DOI:10.1109/MIS.2005.91.
53 Buckee C. Improving epidemic surveillance and response: big data is dead, long live big data. Lancet Digit. Heal. 2020. DOI:10.1016/S2589-7500(20)30059-5.
54 He Y. Coronavirus Infectious Disease Ontology. 2020. http://bioportal.bioontology.org/ontologies/CIDO (accessed April 14, 2020).
55 Bonino L. WHO COVID-19 Rapid Version CRF semantic data model. 2020. https://bioportal.bioontology.org/ontologies/COVIDCRFRAPID?p=summary (accessed April 23, 2020).
56 Liyanage H, Liaw ST, Jonnagaddala J, Hinton W, De Lusignan S. Common Data Models (CDMs) to Enhance International Big Data Analytics: A Diabetes Use Case to Compare Three CDMs. In: Studies in Health Technology and Informatics. 2018. DOI:10.3233/978-1-61499-921-8-60.

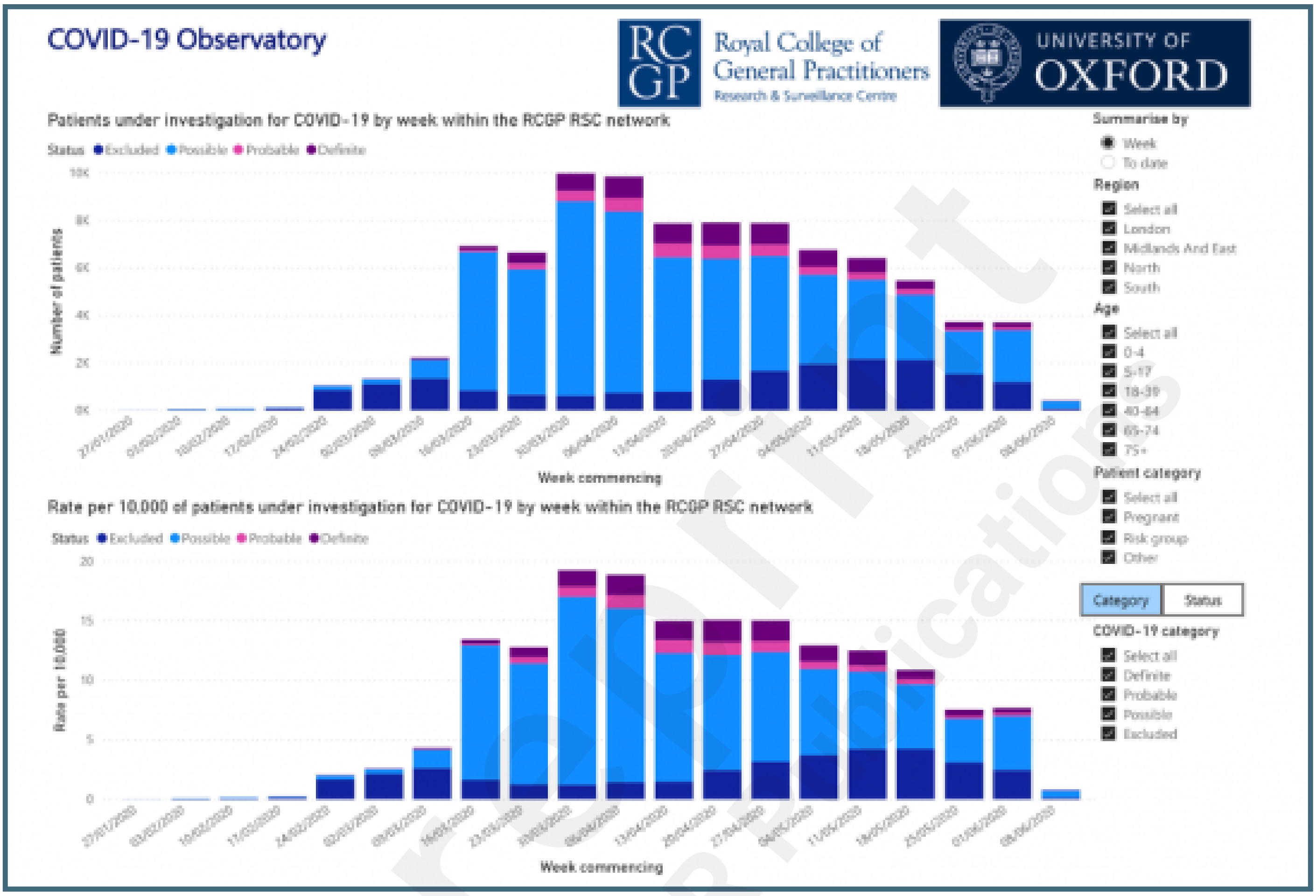
Supplementary Files

Figures

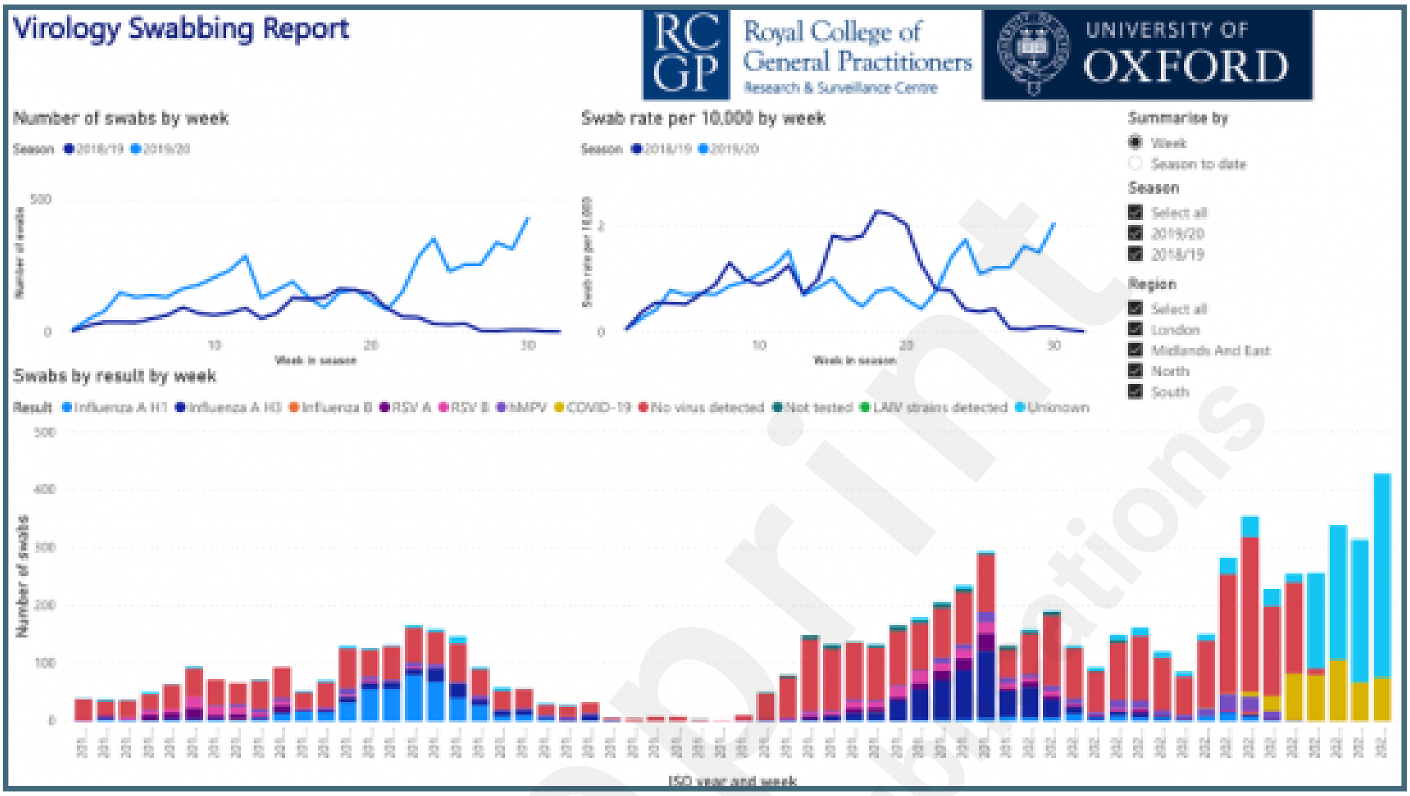
Covid-19 dashboard for each RCGP RSC network practice. The column starting P35398 is that practices data; “South” is their region; RSC is the rate across the whole network. URTI & LRTI are upper and lower respiratory infection, respectively.



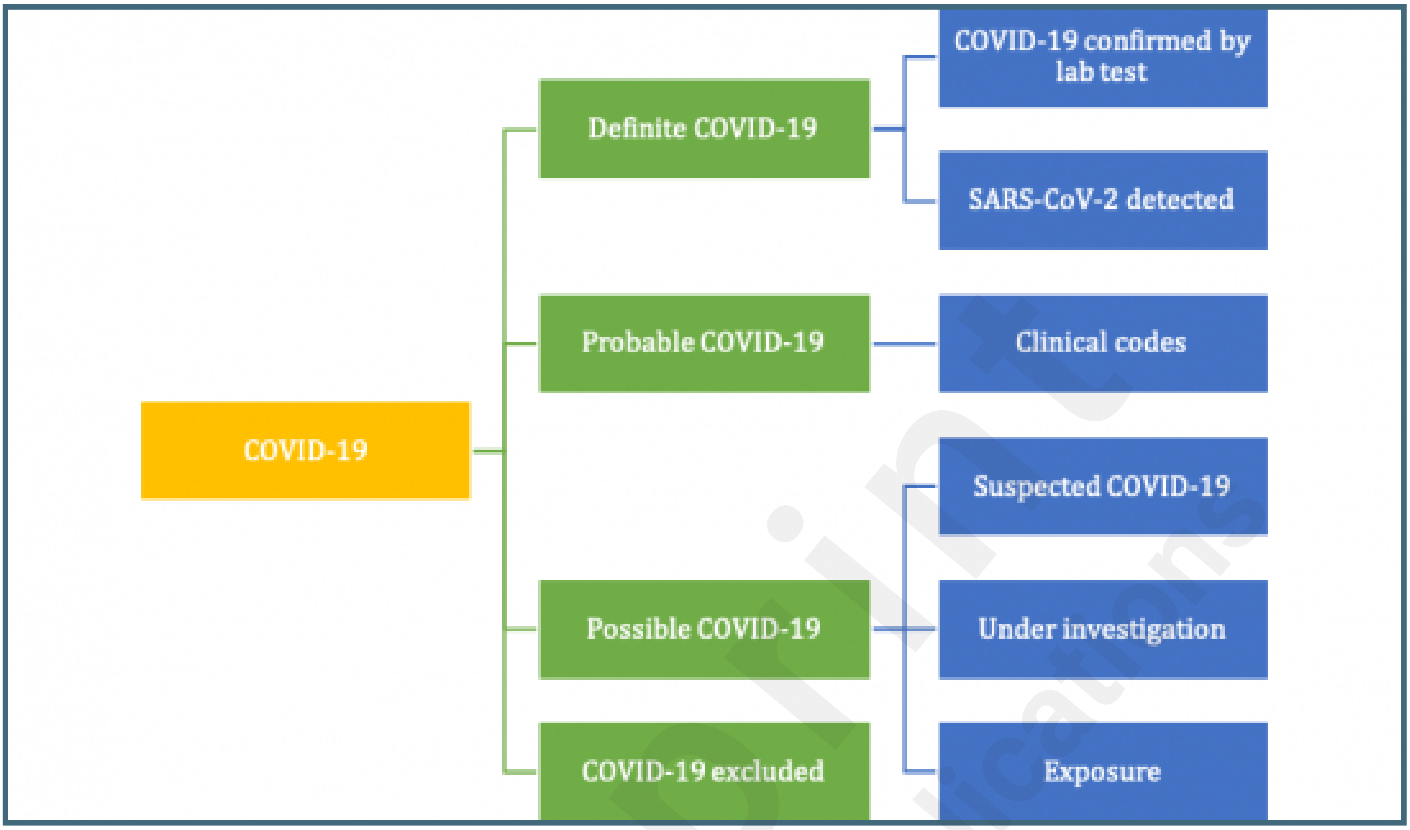
Oxford RCGP RSC interactive COVID-19 observatory Users can select the cumulative or week-by-week view of data, and visualise data by age-band, region, risk group, and Covid-19 status – Definite, Probable, Possible and Excluded.37.



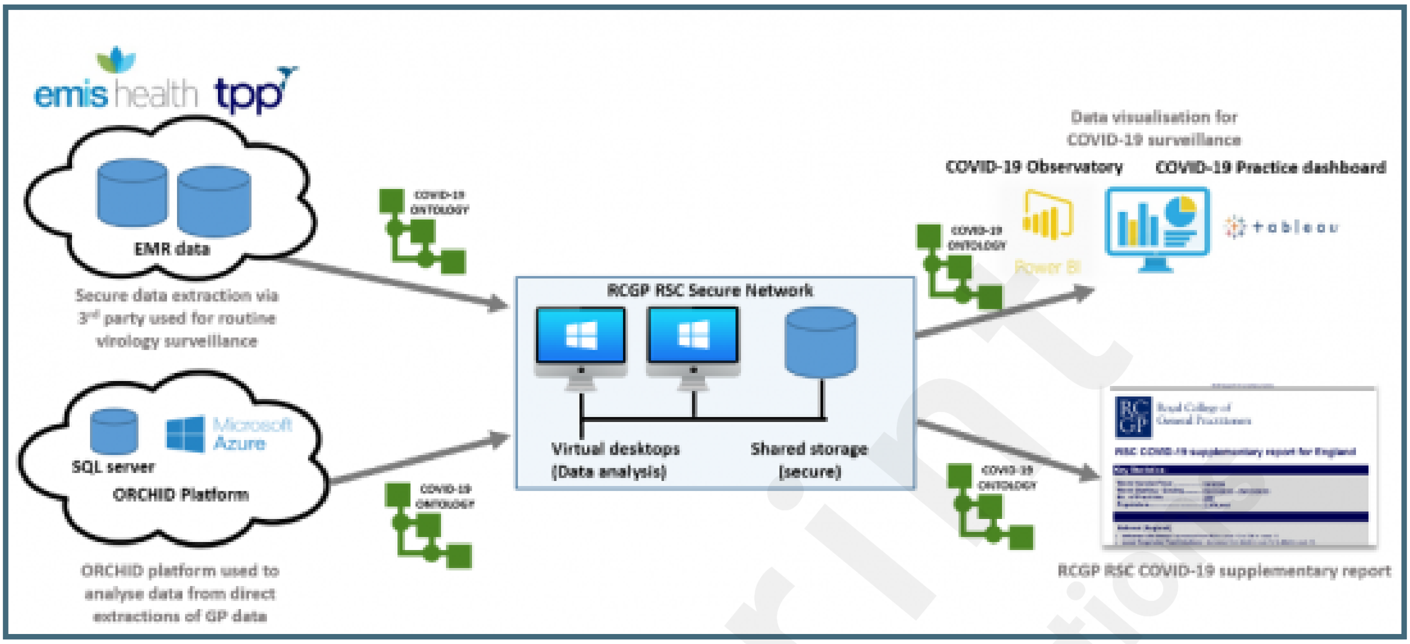
Oxford RCGP RSC interactive virology swabbing report Users can look at the cumulative, or by week, or compare with last year, and look by infected organism or region. “Unknown” is used where no testing is done, currently samples are only tested for Covid-19.



Foundational ontological concepts used for Covid-19 surveillance.



Use of the Covid-19 surveillance ontology across the RCGP RSC processes to achieve semantic consistency in data extraction, visualisations and surveillance reports.



Multimedia Appendixes

S1: Oxford RCGP RSC interactive serology sampling dashboard Users can select the cumulative or week-by-week view of data on samples collected for serological investigation, and visualise data by age-band and region.
URL: https://asset.jmir.pub/assets/f37a41a89dcf615e547802a426acad0b.png

Supplementary Material In-pandemic development of an application ontology for Covid-19 surveillance in a primary care sentinel network.
URL: https://asset.jmir.pub/assets/c3205dc9b98e2bde9d261cf9780cf46a.docx