Deliverable D7.3 COVID-19 metadata mapping model across COVID-19 studies in federated EGA

Project Title (Grant agreement no.):	ELIXIR-CONVERGE: Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services (871075)		
Project Acronym (EC Call):	ELIXIR-CONVERGE (H2020-INFRADEV-2018-2020)		
WP No & Title:	WP7 Federated European Genome-phenome Archives for transnational access of COVID-19 host data		
WP leader(s):	Jordi Rambla (CRG) & Thomas Keane (EMBL-EBI)		
Deliverable Lead Beneficiary:	11 - CSC		
Contractual delivery date:	31/05/2021	**Actual delivery date:**	01/06/2021
Delayed:	No		
Partner(s) contributing to this deliverable:	CSC, CRG		
Authors: Francesca Morello (CSC), Heikki Lehväslaiho (CSC), Ilkka Lappalainen (CSC), Jordi Rambla (CRG) **Contributors:** **Acknowledgments (not grant participants):**			
Reviewers:	ELIXIR-CONVERGE Management Board (MB) members.		

Log of changes

DATE	Mvm	Who	Description
05/05/2021	0v1	Heikki Lehväslaiho (CSC)	Initial version
17/05/2021	0v2	Heikki Lehväslaiho (CSC)	Sent to PMU after incorporating internal WP feedback
17/05/2021	0v3	Nikki Coutts (ELIXIR Hub)	Circulated to the MB for final review before submission
25/5/2021	0v4	Ilkka Lappalainen (CSC)	MB comments addressed

02/06/2021	1v0	Nikki Coutts (ELIXIR Hub)	Final version to be uploaded into EC Portal

Table of contents

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

1. Executive Summary

We summarise efforts from the past year of the COVID-19 pandemic to categorize and guide the collection of phenotype information from patients infected with the SARS-CoV-2. The COVID-19 Host Genetics Initiative data dictionary brings together observations from many independent sources and has been revised multiple times to match availability of data and research needs. In this document, we describe how these data, deposited in controlled access repositories such as EGA and dbGaP, were mapped to the data dictionary and each other, to form the basis of the proposed CONVERGE minimal COVID-19 phenotype metadata model.

2. Contribution toward project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives/key results:

Objective no. / Key Result no. Description	**Contributed to:**
Objective 1: Develop a sustainable and scalable operating model for transnational life-science data management support by leveraging national capabilities **(WP1, WP5**)	
Key Result 1.1: Established European expert network of data stewards that connect national data centres and similar infrastructures and drive the development of interoperable solutions following international best practice, including national interpretations of the General Data Protection Regulation (GDPR)	**No**
Key Result 1. 2: Development of joint guidelines and common toolkit that are adopted into funder recommendations, with support available nationally and in local languages	**Yes**
Key Result 1.3: The catalogue of successful national business models incorporated into national strategies	**No**
Key Result 1.4: The developed “sustainable and scalable operating model for transnational life-science data management support” is adopted into national ELIXIR Node	**No**
Objective 2: Strengthen Europe’s data management capacity through a comprehensive training programme delivered throughout the European Research Area **(WP2, WP6)**	
Key Result 2.1: A comprehensive ELIXIR Training and Capacity building programme in Data Management, directed at both data managers and ELIXIR	**No**

This project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

users, and connected to the national training programmes in Data Management in the ELIXIR Nodes and prospective ELIXIR Member countries.	
Key Result 2.2: Development of a collective group of trainers that support scalable deployment of Data Management training across ELIXIR Nodes.	**No**
Key Result 2.3: A substantial cohort of data managers, Node coordinators and researchers with specific data management skills, business planning and knowledge of transnational operations across the ELIXIR Nodes	**No**
Objective 3: Align national data management standards and services through a sustainable, scalable and cost-effective data management toolkit **(WP2, WP3, WP5**)	
Key Result 3.1: Assemble a full-stack harmonised common toolkit comprising all aspects of data management: from data capture, annotation, and sharing; to integration with analysis platforms and making the data publicly available according to international standards.	**Yes**
Key Result 3.2: Provide exemplar toolkit configurations for prioritised demonstrators to serve as templates for future use.	**No**
Key Result 3.3: Establish national capacity in using as well as updating, extending and sustaining the toolkit across the ERA.	**No**
Key Result 3.4: Enable 'FAIR at source' practice for data generation, and analytical process pipeline implementation by flexible deployment of the toolkit in national operations	**No**
Objective 4: Align national investments to drive local impact and global influence of ELIXIR **(WP4,WP6**)	
Key Result 4.1: Development of a Node Impact Assessment Toolkit based on RI-PATHS methodology.	**No**
Key Result 4.2: Adoption of Impact assessment in ELIXIR Nodes, supported by Node coordinators network and feedback on applicability from dialogues with national funders.	**No**
Key Result 4.3: Creation of national public-private partnerships and industry outreach where open life-science data and services stimulate local bioeconomy	**No**
Key Result 4.4: Growth in reach, impact and engagement of stakeholder communication assessed by established ELIXIR Communications metrics	**No**
Key Result 4.5: Initiating and advancing discussions on Membership (EU and international) or strategic partnerships (international countries) following ELIXIR-CONVERGE workshops.	**No**
Objectives - WP7 - Federated European Genome-phenome Archives for transnational access of	

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

COVID-19 host data.	
O7.1: Federation architecture, interfaces, and compliance tests for the Federated EGA network (Task 7.1).	**No**
O7.2: Coordination of the development of a reference implementation sufficient to create a functional Federated EGA node (Task 7.1).	**No**
O7.3: Development of phenotype metadata model to enable mapping and linkage of COVID-19 host-clinical measures across European national nodes, and robust linkage between host and viral datasets (Task 7.3).	**Yes**
O7.4: The development of documentation and guidelines for the operational practices of federated EGA nodes (Task 7.4).	**No**

3. Introduction

The pandemic has caused an abundance of research efforts into Covid-19 and hence a corresponding need to rapidly disseminate research results that traditional publishing methods have struggled to manage. Similarly, the audience interested in obscure details of virus replication, development stages of immunological response or epidemiological terminology has expanded to laypeople who passionately and obsessively try to understand what are the implications for their daily life. This situation poses unique challenges for data collection and data dissemination to ensure a consistent and unified interpretation of the data coming from heterogeneous sources.

The International Severe Acute Respiratory and emerging Infection Consortium (ISARIC) and the World Health Organization (WHO) were first to develop case report forms (CRFs) to collate anonymized clinical data of hospitalized patients with suspected or confirmed infection with COVID-19. These CRFs are organized in modules to allow data collection on the first day of admission, during hospitalization, and at discharge or death[1][2]. See Appendix 1 for module information.

The forms are designed to be completed by medical professionals, and data are meant to be filled during patient examinations or interviews, or retroactively transcribed from hospital records. These CRFs quickly became the standard reference for national organizations starting to collect clinical information of COVID19 patients. For example, Kurth et al. (2020) established a basic platform for harmonized, scalable data collection, pathophysiological analysis, and deep phenotyping of COVID-19 in Germany based on these forms (the Pa-COVID-19 study). Their protocol parameters were designed by an interdisciplinary panel of experts and were based on the German translation of ISARIC-CRF. Their data elements are organized into three categories: core data, recommended data, and supplemental data elements. Similar to other studies, the Pa-COVID protocol aimed to collect

[1] https://isaric.org/research/covid-19-clinical-research-resources/covid-19-crf/
[2] https://www.who.int/publications/i/item/WHO-2019-nCoV-Clinical_CRF-2020.4

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

blood sampling that would also allow deep molecular and immunological phenotyping, transcriptomic profiling, and comprehensive biobanking (Kurth et al., 2020).

Private companies (e.g. Ancestry®[3] and 23andMe[4]) are collecting datasets of self-reported COVID-19 surveys via web or mobile application. In particular, Ancestry.com has collected over 500,000 COVID-19 survey responses between April and May 2020 that accompany genetic data from the AncestryDNA database (Roberst et al., 2020).

The COVID-19 Host Genetics Initiative (HGI[5]) is an international collaboration of academic researchers harmonizing phenotype data collection from more than 150 independent international studies covering over 50 countries (The COVID-19 Host Genetics Initiative, 2020). Sequence data and associated phenotypic information are collected from biobanks for retrospective studies or from hospitalized patients for prospective studies.

4. Description of work accomplished

4.1 The COVID-19 Host Genetics Initiative data model

We propose to develop the European COVID-19 phenotypic data model based on the COVID-19 Host Genetics Initiative (HGI) project model with minor changes. The proposed alterations include information about vaccine and clinical outcome of the affected patient. Both data types are relevant for the research use cases and have also been highlighted as the pandemia has progressed globally.

We have analysed the original HGI project phenotype model (Bernasconi et al. 2020) and observed how the model has been applied to real data during the project lifetime. Based on our work the HGI phenotype model also complies well with the WHO and ISARIC CRFs. The HGI phenotype model is centered around a patient with separate tables for observations, events, and genomic or viral sequences. It creates a flexible structure to harmonize data from various national cohorts or EHRs. For example, the HGI phenotype model has been applied to more than 3500 patients collated from Spain, Belgium, Brazil and Italy. The host genomes together with harmonised phenotypes will be submitted to the EGA service later this year. The current HGI model (Figure 1) is simpler with roughly half the attributes compared to the initial version (Table 1).

The minimal European COVID-19 phenotype data model together with the associated metadata will form the foundation for future Federated EGA submissions (CONVERGE WP7 deliverables X and Y). The relevant service documentation and training materials should be prepared jointly with CONVERGE WP1 and WP2.

[3] Ancestry® https://www.ancestry.com
[4] 23andMe https://www.23andme.com/
[5] The COVID-19 Host Genetics Initiative https://www.covid19hg.org/

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Table 1. Evolution of the HGI models from the initial (v1) version to the latest one (v4) showing the reduction of both attribute categories and the number of attributes over time.

Attribute group	HGI initial model number of attributes	HGI latest model number of attributes
Identification	2	1
Demographics	9	6
Exposure	3	0
Risk factors	4	0
Comorbidities	33	24
Hospitalisation	27	17
Symptoms at admission	16	4
Medical/Treatment	16	5
Lab tests	34	26
Total	**144**	**83**

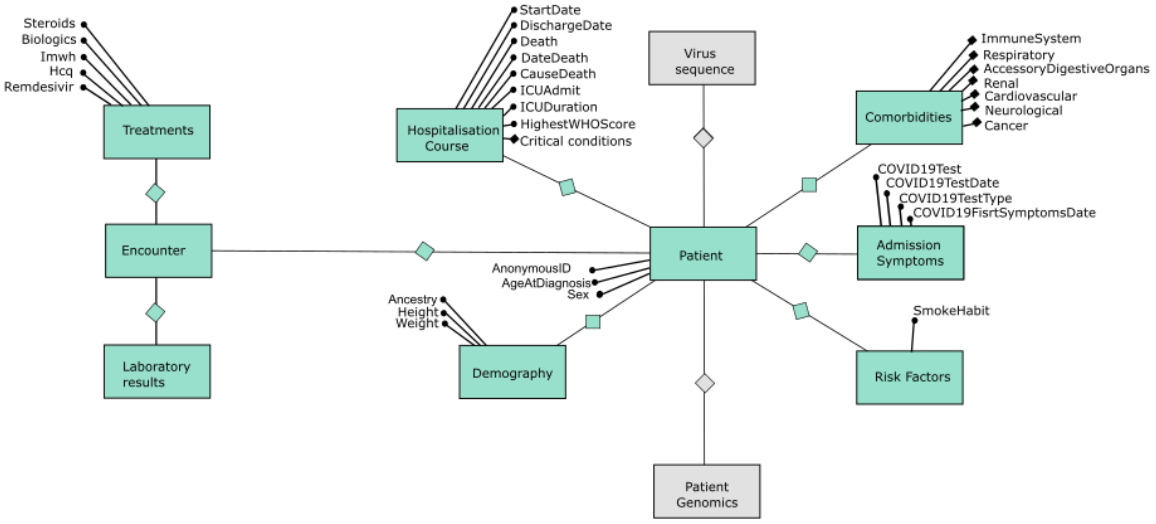

Figure 1. The latest, version 4 HGI phenotype data model is shown in the same format as the published v1 in Bernasconi et al. (2020). The grey boxes represent the virus and host genome sequences that are the data these metadata categories annotate. They are linked to using permanent identifiers given by databases holding them.

4.2 COVID-19 data structure in studies in EGA and dbGAP

At this time, in May 2021, there are 11 COVID-19 studies in EGA. Appendix 2 summarises these studies. While their technologies range from single-cell RNA-sequencing to GWAS, they usually lack

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

included phenotype data. However, most of them have separately made available supplementary aggregate phenotype information.

4.2.1 Studies with externally shared phenotype information

Two of these studies analysed data from CONTAGIOUS (COvid-19 Advanced Genetic and Immunologic Sampling), an observational clinical trial conducted on hospitalized patients in Belgium. These studies analysed respectively 16S amplicon data of nasopharyngeal swabs (EGAS00001004951[6], 125 samples) and Single-cell RNA sequencing of bronchoalveolar lavages (EGAS00001004717[7], 35 samples), published in Wauters et al. (2021). No individual-level phenotype data was deposited in EGA. However, the summary of demographic and clinical data of the prospectively recruited patient cohort, including comorbidities, were reported in the supplementary information.

Three studies combined clinical data (based on two prospective cohorts, Pa-Covid19 Berlin cohort and Bonn cohort) and single-cell sequencing to analyse COVID-19 patient's airway cellular composition, hypertensions effects, and whole-blood and peripheral-blood mononuclear cells profiling (EGAS00001004772[8], 33 samples, EGAS00001004481[9] 36 samples, EGAS00001004571[10], 141 individuals).

The study EGAS00001004571[11] uses data from two different cohorts (the Berlin cohort, the Bonn cohorts and samples from the study Reyes at al., 2020);(Schulte-Schrepping et al., 2020). The cohort details (donors, samples and methods) used in the manuscript are available in an Excel file obtainable both in the publication supplementary information[12] or from the EGA API in JSON format.

The attributes reported are:

- Patient ID: (format varies depending on the cohort)
- Condition: control / COVID19 mild / COVID19 severe
- Outcome: n/a / discharged / ongoing / deceased
- WHO classification: 1-10; sampling day (after onset of symptoms): digit
- Sex: m / f / n/a
- Age: digit (5 y)
- Comorbidities: n/a, not cardiovascular diseases, obesity, chronic lung disease, covid-related medications

Patient ID, WHO classification, Sex, Age, comorbidities and COVID-19 related medication and anti-microbials used map to the attribute groups (or elements) of the common minimal metadata model.

[6] https://ega-archive.org/studies/EGAS00001004951
[7] https://ega-archive.org/studies/EGAS00001004717
[8] https://ega-archive.org/studies/EGAS00001004772
[9] https://ega-archive.org/studies/EGAS00001004481
[10] https://ega-archive.org/studies/EGAS00001004571
[11] https://ega-archive.org/studies/EGAS00001004571/
[12] https://www.cell.com/cms/10.1016/j.cell.2020.08.001/attachment/5428d9f5-15

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

The attribute "condition" has possible values: control, COVID-19 mild, COVID-19 severe, FLI, corresponding to mild (WHO 2-4) or severe (5-7) disease according to the WHO clinical ordinal scale (WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection, 2020). The attribute "outcome" can have these values: n/a, discharged, ongoing, deceased. The age is reported as a stratified value (in steps of 5 years).

Most of the comorbidities map to the minimal metadata model, however some of the possible values are more specific.

The "COVID-19 related medication and anti-microbials" attribute has several values that are not present in the minimal metadata model: antibiotics and type of antibiotics, convalescent plasma, emtricitabine/tenofovir alafenamide fumarate; opinavir/ritonavir, raltegravir; steroids; ampicillin/sulbactam, clarithromycin.

According to the HGI data dictionary, "biological" attribute refers to Medications called "biologics" or "monoclonal antibodies" such as abatacept (Orencia), adalimumab (Humira), anakinra (Kineret), certolizumab pegol (Cimzia), etanercept (Enbrel), golimumab (Simponi), infliximab (Inflectra, Remicade), rituximab (Rituxan), tocilizumab (Actemra), tofacitinib (Xeljanz), and upadacitinib (Rinvoq).

The dbGAP study phs002245.v1.p1[13] reported on 585 patients with life-threatening COVID-19 pneumonia (Zhang et al., 2020; Bastard et al., 2020). Patients who developed Kawasaki-like syndrome were excluded. The attributes are:

- SUBJECT_ID: de-identified Subject ID, string
- SEX (gender of the participant): F /M / unknown
- AGE (subject age at enrollment): numerical value, >89, no data, unknown
- COVID19 clinical severity: A:Critical, B:Asymptomatic, C:Severe, D:Mild, E:Moderate, F:Moderate Severe, G:Unknown
- Clinical outcome: Alive / Deceased / Outcome unknown

These attributes correspond well with values in the HGI model but they are presented as strings and not in the numerical norm.

Currently, nine studies involved in the HGI (from Spain, Belgium, Brazil and Italy) have submitted information from 3500 samples with genotype data and harmonized clinical information to EGA using the HGI model but their metadata are not yet available.

4.3 Studies with individual-level phenotype data

Just one study deposited genotype and phenotype information to EGA: The Ancestry study (EGAS00001004716) on 15000 individuals. As mentioned previously, these data are obtained from surveys and questionnaires, and hence this Covid-19 study is unique because the phenotype

[13] https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002245.v1.p1

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

information is self-reported. The Ancestry data dictionary in its entirety is in Appendix 3. Mapping from it to the minimal data model is shown in Appendix 4.

The Ancestry data dictionary has the survey questions as attributes and possible answers as string values. The data dictionary is not organized in attribute groups and does not use any specific terminology or ontology.

The Ancestry metadata model can be summarized in the following attribute groups:

- Demographics (4),
- Risk factors (2),
- Comorbidities (1) attribute, 16 possible values),
- Exposure (5),
- Hospitalization (5),
- Symptoms at admission (3, 15x6).

All the Ancestry attributes map to the minimal HGI model except the exposure attributes. Severity of the disease is defined based on the yes/no answer to the questions: "have you been hospitalized?". Sex of the person was deduced from sequences. It is worth noting that the dataset contains only individuals who had contracted the disease (Roberts et al., 2020).

4.2.3. Laboratory results in Covid-19 data

COVID-19 patients are subjected to laboratory tests at hospitals or during longitudinal studies. The attribute group "lab tests" is used by ISARIC and WHO CRFs to collect basic laboratory test values. In addition to these, the HGI and the Pa-Covid19 (extended version of lab tests) studies include more specific measurements, for example the count of cd4 and cd8 lymphocytes and the analysis of specific coagulation factors (fibrinogens). The WHO post-COVID-19 CRF also considers the analysis of coronavirus antibodies (IgA, IgM and IgG) that require specific analysis and as such are excluded from our minimal model. The detailed mapping between ISARIC and HGI data dictionaries is given in **Appendix 5**.

5. Results

5.1 The COVID-19 minimal phenotype metadata model

The proposed minimal COVID-19 phenotype metadata model is available in a separate, normative data dictionary document: CONVERGE WP7 D7.3: Minimal COVID-19 phenotype metadata data dictionary[14].

[14] https://docs.google.com/spreadsheets/d/1xdAhYuZHMliyxE0fwHUjEiquIVcngTmdAKuBBgIaoi0/edit#gid=1658857689

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

In summary, the model is based on the latest minimal HGI model with the following elements and modifications (Figure 2):

Demographics: age at diagnosis, sex, ancestry, height, weight.

Note that no exposure information is included in the model.

Comorbidities: Immune System (HIV, Immunocompromised status, Organ transplant, Autoimmune or rheumatologic disease (e.g., rheumatoid arthritis, systemic lupus erythematosus, multiple sclerosis, inflammatory bowel disease), Type I diabetes, Type II diabetes; Respiratory system (Asthma, Chronic obstructive pulmonary disease (COPD), Sleep Apnea); Accessory digestive organs (Liver disease, Gallbladder, Pancreas); Renal disease (Chronic kidney disease, Dialysis, Congestive heart failure, Hypertension, Myocardial infarction, Peripheral vascular disease, Stroke, Atrial fibrillation or flutter); Neurological disorders: (Dementia, Neurological or neuropsychiatric disease); Cancer: (Leukemia, Lymphoma Malignant solid tumor);

Risk factors: smoking

Attributes related to death have been moved from hospitalisation to a new group "Clinical outcome" and a boolean attribute "readmission" have been added to it.

Hospitalisation: Date at the start of hospitalization, Date at the end of hospitalization, readmission, ICU admit, Number of days in ICU, Highest WHO clinical progression scale, Highest level of respiratory support, Days of ventilation, Deep vein thrombosis (DVT), Thromboembolism (PE), Stroke, Myocardial infarction, Renal failure requiring dialysis, Major bleeding (i.e. gastrointestinal bleeding, intracranial haemorrhage), Hepatic failure.

Clinical outcome: Death, Date of death, Cause of death, Recovered from COVID 19, COVID19 complications (cardiovascular, dermatological, endocrine, gastrointestinal, fatigue, Musculoskeletal, Mentha health, Neurological, pulmonary, renal).

The HGI attribute group "**Symptoms at admission / longitudinal**" was renamed to "COVID-19 test":

COVID-19 test: COVID19 test result, COVID19 test date, COVID19 type of test used. No symptoms were included in this model.

The HGI data dictionary attribute group "COVID-19 therapy during hospitalisation" is now called "Treatments":

Treatments: Steroids, Medications called "biologics" or "monoclonal antibodies" such as abatacept (Orencia), adalimumab (Humira), anakinra (Kineret), certolizumab pegol (Cimzia), etanercept (Enbrel), golimumab (Simponi), infliximab (Inflectra, Remicade), rituximab (Rituxan), tocilizumab (Actemra), tofacitinib (Xeljanz), and upadacitinib (Rinvoq), Low-molecular-weight heparin, Hydroxychloroquine, Remdesivir.

Laboratory results include a long list basic tests:

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Laboratory results: See the list of attributes in the data dictionary. Next to the laboratory results, every study needs to report the reference values and units for lab tests, necessary to be able to interpret the results and harmonize the data across studies.

A new attribute group was added to cover vaccination events. Each administered dose should receive its own entry:

Vaccines: COVID19 Vaccine administration, Vaccine Date of administration, Vaccine Type (Moderna, Pfizer, AstraZeneca, Johnson and Johnson, Other).

We propose all these attributes as required fields, without any optional fields. Most fields contain a value "unknown" as an option. As specified earlier, most of the COVID-19 studies are based on observational clinical trials which collect complex information about each hospitalized patient. Moreover, even the information collected with the use of self reported questionnaires designed in the first months of the pandemic can map to most of the attributes proposed in the common minimal model.

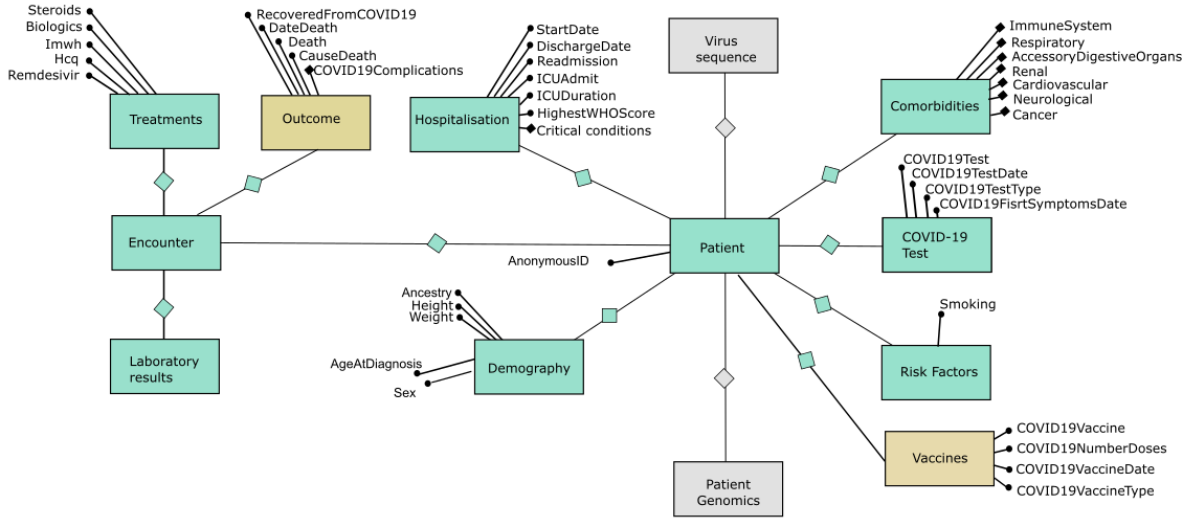

Figure 2. The proposed CONVERGE COVID-19 minimal metadata model. The tan coloured attribute groups are new additions on top of the HGI model in the previous figure.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

6. Conclusions

Phenotype data are collected from a wide range of formats, from electronic forms, questionnaires or even data transcribed by hand to spreadsheets. Researchers from the HGI study developed a metadata model and a simple data dictionary that allowed the collection and harmonization of COVID19 phenotypic data worldwide. For this reason, we propose the same model for future data submission across Federated EGA and we propose minimal changes to this model.

It is worth noting that none of the limited number of studies submitted to EGA and dbGAP use named ontologies to describe data elements or values. The uptake and use of these precise, comprehensive and complex ways of describing phenotypic values seem to face a very high barrier in clinical studies. The trend towards simplicity can also be seen in the diminishing number of attributes included in the successive versions of the HGI data dictionary. Although it is tempting to include ever more details to collect from this still relatively unknown disease, it is clearly better to try to focus on as few and important phenotypic attributes as possible to maximise the amount of data. Our minimal COVID-19 data model hopes to help in gaining that goal.

Our analysis also emphasises the importance of study-specific data dictionaries when trying to understand the values used. Every study submitted to EGA and future Federated EGA should ideally include a public data dictionary file.

The approach used here to use number coded controlled vocabularies, should not deter from expanding the model to wider and more comprehensive controlled vocabularies in the future. For example, the expansion of the ancestry CV terms to Human Ancestry Ontology HANCESTRO is a good candidate.

7. Impact

This is the first practical guide to collecting European COVID-19 host information linked to sequences.

8. Next Steps

WP1, share with other nodes

WP2, training material for a broad public and researchers from different fields

9. Deviation from Description of Action

The initial aim of the task was to analyse the submitted COVID-19 data in Federated EGA to understand and unify the semantics of data collection. However, as the founding of the Federated EGA organisation has been delayed, we analysed the data submitted to both EGA and dbGaP.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

10. Bibliography

Anna Bernasconi, Arif Canakoglu, Marco Masseroli, Pietro Pinoli, Stefano Ceri, A review on viral data sources and search systems for perspective mitigation of COVID-19, Briefings in Bioinformatics **2021**, 22(2):664–675, https://doi.org/10.1093/bib/bbaa359

Bastard P, Rosen LB, Zhang Q, Michailidis E, Hoffmann HH, Zhang Y, Dorgham K, Philippot Q, Rosain J, Béziat V, Manry J, Shaw E, Haljasmägi L, Peterson P, Lorenzo L, Bizien L, Trouillet-Assant S, Dobbs K, de Jesus AA, Belot A, Kallaste A, Catherinot E, Tandjaoui-Lambiotte Y, Le Pen J, Kerner G, Bigio B, Seeleuthner Y, Yang R, Bolze A, Spaan AN, Delmonte OM, Abers MS, Aiuti A, Casari G, Lampasona V, Piemonti L, Ciceri F, Bilguvar K, Lifton RP, Vasse M, Smadja DM, Migaud M, Hadjadj J, Terrier B, Duffy D, Quintana-Murci L, van de Beek D, Roussel L, Vinh DC, Tangye SG, Haerynck F, Dalmau D, Martinez-Picado J, Brodin P, Nussenzweig MC, Boisson-Dupuis S, Rodríguez-Gallego C, Vogt G, Mogensen TH, Oler AJ, Gu J, Burbelo PD, Cohen JI, Biondi A, Bettini LR, D'Angio M, Bonfanti P, Rossignol P, Mayaux J, Rieux-Laucat F, Husebye ES, Fusco F, Ursini MV, Imberti L, Sottini A, Paghera S, Quiros-Roldan E, Rossi C, Castagnoli R, Montagna D, Licari A, Marseglia GL, Duval X, Ghosn J; HGID Lab; NIAID-USUHS Immune Response to COVID Group; COVID Clinicians; COVID-STORM Clinicians; Imagine COVID Group; French COVID Cohort Study Group; Milieu Intérieur Consortium; CoV-Contact Cohort; Amsterdam UMC Covid-19 Biobank; COVID Human Genetic Effort, Tsang JS, Goldbach-Mansky R, Kisand K, Lionakis MS, Puel A, Zhang SY, Holland SM, Gorochov G, Jouanguy E, Rice CM, Cobat A, Notarangelo LD, Abel L, Su HC, Casanova JL. Autoantibodies against type I IFNs in patients with life-threatening COVID-19. Science. 2020 Oct 23;370(6515):eabd4585. doi: 10.1126/science.abd4585. Epub 2020 Sep 2 https://doi.org/10.1126/science.abd4585

Chua, R. L.; Lukassen, S.; Trump, S.; Hennig, B. P.; Wendisch, D.; Pott, F.; Debnath, O.; Thürmann, L.; Kurth, F.; Völker, M. T.; Kazmierski, J.; Timmermann, B.; Twardziok, S.; Schneider, S.; Machleidt, F.; Müller-Redetzky, H.; Maier, M.; Krannich, A.; Schmidt, S.; Balzer, F.; Liebig, J.; Loske, J.; Suttorp, N.; Eils, J.; Ishaque, N.; Liebert, U. G.; von Kalle, C.; Hocke, A.; Witzenrath, M.; Goffinet, C.; Drosten, C.; Laudi, S.; Lehmann, I.; Conrad, C.; Sander, L.-E.; Eils, R. COVID-19 Severity Correlates with Airway Epithelium–Immune Cell Interactions Identified by Single-Cell Analysis. *Nat Biotechnol* **2020**, *38* (8), 970–979. https://doi.org/10.1038/s41587-020-0602-4.

Wauters, E.; Van Mol, P.; Garg, A. D.; Jansen, S.; Van Herck, Y.; Vanderbeke, L.; Bassez, A.; Boeckx, B.; Malengier-Devlies, B.; Timmerman, A.; Van Brussel, T.; Van Buyten, T.; Schepers, R.; Heylen, E.; Dauwe, D.; Dooms, C.; Gunst, J.; Hermans, G.; Meersseman, P.; Testelmans, D.; Yserbyt, J.; Tejpar, S.; De Wever, W.; Matthys, P.; Neyts, J.; Wauters, J.; Qian, J.; Lambrechts, D. Discriminating Mild from Critical COVID-19 by Innate and Adaptive Immune Single-Cell Profiling of Bronchoalveolar Lavages. *Cell Res* **2021**, *31* (3), 272–290. https://doi.org/10.1038/s41422-020-00455-9.

Kurth, F.; Roennefarth, M.; Thibeault, C.; Corman, V. M.; Müller-Redetzky, H.; Mittermaier, M.; Ruwwe-Glösenkamp, C.; Heim, K. M.; Krannich, A.; Zvorc, S.; Schmidt, S.; Kretzler, L.; Dang-Heine, C.; Rose, M.; Hummel, M.; Hocke, A.; Hübner, R. H.; Opitz, B.; Mall, M. A.; Röhmel, J.; Landmesser, U.; Pieske, B.; Knauss, S.; Endres, M.; Spranger, J.; Mockenhaupt, F. P.; Tacke, F.; Treskatsch, S.; Angermair, S.; Siegmund, B.; Spies, C.; Weber-Carstens, S.; Eckardt, K.-U.; Schürmann, D.; Uhrig, A.; Stegemann, M. S.; Zoller, T.; Drosten, C.; Suttorp, N.; Witzenrath, M.; Hippenstiel, S.; von Kalle, C.; Sander, L. E. Studying the Pathophysiology of Coronavirus Disease 2019: A Protocol for the Berlin Prospective COVID-19 Patient Cohort (Pa-COVID-19). *Infection* **2020**, *48* (4), 619–626. https://doi.org/10.1007/s15010-020-01464-x.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Murray, M. F.; Kenny, E. E.; Ritchie, M. D.; Rader, D. J.; Bale, A. E.; Giovanni, M. A.; Abul-Husn, N. S. COVID-19 Outcomes and the Human Genome. *Genet Med* **2020**, *22* (7), 1175–1177. https://doi.org/10.1038/s41436-020-0832-3.

Roberts, G. H. L.; Park, D. S.; Coignet, M. V.; McCurdy, S. R.; Knight, S. C.; Partha, R.; Rhead, B.; Zhang, M.; Berkowitz, N.; AncestryDNA Science Team; Haug Baltzell, A. K.; Guturu, H.; Girshick, A. R.; Rand, K. A.; Hong, E. L.; Ball, C. A. *AncestryDNA COVID-19 Host Genetic Study Identifies Three Novel Loci*; preprint; Epidemiology, 2020. https://doi.org/10.1101/2020.10.06.20205864.

Schulte-Schrepping, J.; Reusch, N.; Paclik, D.; Baßler, K.; Schlickeiser, S.; Zhang, B.; Krämer, B.; Krammer, T.; Brumhard, S.; Bonaguro, L.; De Domenico, E.; Wendisch, D.; Grasshoff, M.; Kapellos, T. S.; Beckstette, M.; Pecht, T.; Saglam, A.; Dietrich, O.; Mei, H. E.; Schulz, A. R.; Conrad, C.; Kunkel, D.; Vafadarnejad, E.; Xu, C.-J.; Horne, A.; Herbert, M.; Drews, A.; Thibeault, C.; Pfeiffer, M.; Hippenstiel, S.; Hocke, A.; Müller-Redetzky, H.; Heim, K.-M.; Machleidt, F.; Uhrig, A.; Bosquillon de Jarcy, L.; Jürgens, L.; Stegemann, M.; Glösenkamp, C. R.; Volk, H.-D.; Goffinet, C.; Landthaler, M.; Wyler, E.; Georg, P.; Schneider, M.; Dang-Heine, C.; Neuwinger, N.; Kappert, K.; Tauber, R.; Corman, V.; Raabe, J.; Kaiser, K. M.; Vinh, M. T.; Rieke, G.; Meisel, C.; Ulas, T.; Becker, M.; Geffers, R.; Witzenrath, M.; Drosten, C.; Suttorp, N.; von Kalle, C.; Kurth, F.; Händler, K.; Schultze, J. L.; Aschenbrenner, A. C.; Li, Y.; Nattermann, J.; Sawitzki, B.; Saliba, A.-E.; Sander, L. E.; Angelov, A.; Bals, R.; Bartholomäus, A.; Becker, A.; Bezdan, D.; Bonifacio, E.; Bork, P.; Clavel, T.; Colome-Tatche, M.; Diefenbach, A.; Dilthey, A.; Fischer, N.; Förstner, K.; Frick, J.-S.; Gagneur, J.; Goesmann, A.; Hain, T.; Hummel, M.; Janssen, S.; Kalinowski, J.; Kallies, R.; Kehr, B.; Keller, A.; Kim-Hellmuth, S.; Klein, C.; Kohlbacher, O.; Korbel, J. O.; Kurth, I.; Landthaler, M.; Li, Y.; Ludwig, K.; Makarewicz, O.; Marz, M.; McHardy, A.; Mertes, C.; Nöthen, M.; Nürnberg, P.; Ohler, U.; Ossowski, S.; Overmann, J.; Peter, S.; Pfeffer, K.; Poetsch, A. R.; Pühler, A.; Rajewsky, N.; Ralser, M.; Rieß, O.; Ripke, S.; Nunes da Rocha, U.; Rosenstiel, P.; Saliba, A.-E.; Sander, L. E.; Sawitzki, B.; Schiffer, P.; Schulte, E.-C.; Schultze, J. L.; Sczyrba, A.; Stegle, O.; Stoye, J.; Theis, F.; Vehreschild, J.; Vogel, J.; von Kleist, M.; Walker, A.; Walter, J.; Wieczorek, D.; Ziebuhr, J. Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell* **2020**, *182* (6), 1419-1440.e23. https://doi.org/10.1016/j.cell.2020.08.001.

The COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum Genet* **28,** 715–718 (2020). https://doi.org/10.1038/s41431-020-0636-6

Trump, S.; Lukassen, S.; Anker, M. S.; Chua, R. L.; Liebig, J.; Thürmann, L.; Corman, V. M.; Binder, M.; Loske, J.; Klasa, C.; Krieger, T.; Hennig, B. P.; Messingschlager, M.; Pott, F.; Kazmierski, J.; Twardziok, S.; Albrecht, J. P.; Eils, J.; Hadzibegovic, S.; Lena, A.; Heidecker, B.; Bürgel, T.; Steinfeldt, J.; Goffinet, C.; Kurth, F.; Witzenrath, M.; Völker, M. T.; Müller, S. D.; Liebert, U. G.; Ishaque, N.; Kaderali, L.; Sander, L.-E.; Drosten, C.; Laudi, S.; Eils, R.; Conrad, C.; Landmesser, U.; Lehmann, I. Hypertension Delays Viral Clearance and Exacerbates Airway Hyperinflammation in Patients with COVID-19. *Nat Biotechnol* **2020**. https://doi.org/10.1038/s41587-020-00796-1.

Zhang Q, Bastard P, Liu Z, Le Pen J, Moncada-Velez M, Chen J, Ogishi M, Sabli IKD, Hodeib S, Korol C, Rosain J, Bilguvar K, Ye J, Bolze A, Bigio B, Yang R, Arias AA, Zhou Q, Zhang Y, Onodi F, Korniotis S, Karpf L, Philippot Q, Chbihi M, Bonnet-Madin L, Dorgham K, Smith N, Schneider WM, Razooky BS, Hoffmann HH, Michailidis E, Moens L, Han JE, Lorenzo L, Bizien L, Meade P, Neehus AL, Ugurbil AC, Corneau A, Kerner G, Zhang P, Rapaport F, Seeleuthner Y, Manry J, Masson C, Schmitt Y, Schlüter A, Le Voyer T, Khan T, Li J, Fellay J, Roussel L, Shahrooei M, Alosaimi MF, Mansouri D, Al-Saud H, Al-Mulla F, Almourfi F, Al-Muhsen SZ, Alsohime F, Al Turki S, Hasanato R, van de Beek D, Biondi A, Bettini LR, D'Angio' M, Bonfanti P, Imberti L, Sottini A,

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Paghera S, Quiros-Roldan E, Rossi C, Oler AJ, Tompkins MF, Alba C, Vandernoot I, Goffard JC, Smits G, Migeotte I, Haerynck F, Soler-Palacin P, Martin-Nalda A, Colobran R, Morange PE, Keles S, Çölkesen F, Ozcelik T, Yasar KK, Senoglu S, Karabela ŞN, Rodríguez-Gallego C, Novelli G, Hraiech S, Tandjaoui-Lambiotte Y, Duval X, Laouénan C; COVID-STORM Clinicians; COVID Clinicians; Imagine COVID Group; French COVID Cohort Study Group; CoV-Contact Cohort; Amsterdam UMC Covid-19 Biobank; COVID Human Genetic Effort; NIAID-USUHS/TAGC COVID Immunity Group, Snow AL, Dalgard CL, Milner JD, Vinh DC, Mogensen TH, Marr N, Spaan AN, Boisson B, Boisson-Dupuis S, Bustamante J, Puel A, Ciancanelli MJ, Meyts I, Maniatis T, Soumelis V, Amara A, Nussenzweig M, García-Sastre A, Krammer F, Pujol A, Duffy D, Lifton RP, Zhang SY, Gorochov G, Béziat V, Jouanguy E, Sancho-Shimizu V, Rice CM, Abel L, Notarangelo LD, Cobat A, Su HC, Casanova JL. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science. 2020 Oct 23;370(6515):eabd4570. doi: 10.1126/science.abd4570. Epub 2020 Sep 24. PMID: 32972995; PMCID: PMC7857407. https://doi.org/10.1126/science.abd4585.

WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection. A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis. 2020 Aug;20(8):e192-e197. https://doi.org/10.1016/S1473-3099(20)30483-7. Epub 2020 Jun 12. Erratum in: Lancet Infect Dis. 2020 Oct;20(10):e250. PMID: 32539990; PMCID: PMC7292605.https://doi.org/10.1016/s1473-3099(20)30483-7.

11. Appendices

Appendix 1. Summary of the WHO and ISARIC CRFs

The modules and CRFs of the WHO and ISARIC form the ISARIC Data Dictionary[15] with 321 unique variables.

Module	WHO-CRF	ISARIC CRF (rapid)
Module 1	Clinical Inclusion criteria	Clinical Inclusion Criteria
	Demographics	Demographics
	Data of onset and vital signs	Data of onset and vital signs
	Comorbidities	Comorbidities
	Pre-admission and chronic medications	Preadmission and Chronic medications
	Signs and symptoms on admission	Signs and symptoms on admission
		Vaccinations
	Medication on the day of admission	Medications

[15] http://clinbuild.com/blog/downloads/ISARICnCoV_DataDictionary_2020-03-27.csv

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

	Supportive care	Supportive care
	Laboratory results on admission	Laboratory results on admission
Module 2	Vital signs	Vital signs
	Daily clinical features	Daily clinical features
	Laboratory results	Laboratory results
	Medications	Medications
	Supportive care	Supportive care
Module 3	Diagnostic pathogen testing	Diagnostic pathogen testing
	Complications	Complications
	Medications	Medications
	Supportive care	Supportive care
	Outcome	Outcome

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Appendix 2. List of COVID19 studies in EGA

EGA Study ID	Description	Source Phenotype info	Samples	Individual level phenotype dataset
EGAS00001004717	Single-cell RNA seq	Contagious, observational clinical trial Hospitalized patients (Belgium)	35	no
EGAS00001004951	16S amplicon data of nasopharyngeal swabs (upper respiratory microbiome)	Contagious, observational clinical trial Hospitalized patients (Belgium)	125	no
EGAS00001004412	single-cell RNA and VDJ sequencing of antigen-enriched B cells	Discharged patients (China)	1 (pulled single cell, patient 1 to 12 plus 6 batches of 10 patients)	no
EGAS00001004503	two datasets of single cells from blood Immunoprofiling neutrofiles	Hospitalized patients Rhineland Study used as control (Germany)	79 + 41	no
EGAS00001004772	Hypertention single-cell sequencing data of airway samples	Based on two prospective cohorts (Germany)	33 single cells (144 phenotype not published) plus reanalysed	no
EGAS00001004481	COVID-19 severity correlates with airway epithelium–immune	Dual-center cohort from Charité - Universitätsmedizin Berlin and University Hospital Leipzig (Germany)	36	no
EGAS00001004571	ScRNA-seq of PBMC	Berlin (pa-covid) cohort Bonn cohort (cohort 2)	141	no (but as public info)
EGAS00001004716	**Phenotype** and genotype GWAS (Ancestry study)	questionnaire /survey (US)	15000	Yes

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

phs002245	Genetic Determinants of Susceptibility to Severe COVID-19 Infection **Phenotype**	hospitalized Phenotype and sequences	585	Yes (limited)
phs002258	metatranscriptomics (total-RNA-seq) for host, viral, and microbial profiling.	Hospitalized Patients (US)	669	no
EGAS00001005060	FACS-purified CD8 T lymphocytes from two Austrian patients.	2 Austrian Patients	2	no
EGAS00001004419	ALI culture bronchial cells and alveolar lung surgical resection scRNA-Seq	Lung Biobank Heidelberg (Germany)	16	no (sex, age and smoking hystory reported in the publication)
EGAS00001004502	Swarm Learning	93Intensive Care Unit of the Radboud University Medical Centre in Nijmegen, the Netherlands, 41 samples, Sotiria Athens General Hospital or the ATTIKON University General Hospital in Athens, Greece	650	no

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Appendix 3. Ancestry DNA COVID-19 Survey Data Dictionary

Ancestry DNA COVID-19 Survey Data Dictionary

EGA Release

Direct any questions to covid19research@ancestry.com

Ancestry COVID-19 Survey Question	**Ancestry COVID-19 Survey Response Values**
Age	Integer (# Years)
Genetic Sex (*Inferred from Genetic Data*)	Single Value from Options: "Male", "Female"
Have you been swab tested for COVID-19, commonly referred to as coronavirus?	Single Value from Options: "Yes, and was positive", "Yes, and was negative"
Did you experience symptoms as a result of your condition?	Single Value from Options: "Yes"', "No", "Not sure"
Between the beginning of February and now, have you had any of the following symptoms?	Matrix Selection: Rows: "Fever", "Shortness of breath", "Dry cough", "Nasal congestion", "Runny nose", "Sore throat", "Feeling tired or fatigue", "Chills", "Body aches", "Headache", "Cough producing phlegm", "Abdominal pain", "Nausea/vomiting", "Diarrhea", "Change in taste or smell" Columns: "None", "Very Mild", "Mild", "Moderate", "Severe", "Very Severe"
Were you hospitalized due to these symptoms?	Single Value from Options: "Yes", "No", "Not sure"
How long were you hospitalized?	Integer (# Days)
Were you hospitalized in the Intensive Care Unit (ICU) with a ventilator?	Single Value from Options: "Yes", "No", "Not sure"
Were you hospitalized in the Intensive Care Unit (ICU) with oxygen?	Single Value from Options: "Yes", "No", "Not sure"

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Have you had any of the following complications due to your illness? Select all that apply.	Select All That Apply from Options: "None", "Pneumonia", "Severe pneumonia leading to hospitalization", "Respiratory failure", "Septic shock", "Multiple organ dysfunction or failure", "Blood clots", "Stroke", "Other"
Have any of your biological relatives tested positive for COVID-19? Select all that apply.	Select All That Apply from Options: "Grandparent(s)", "Parent(s)", "Child(ren)", "Sibling(s)", "Other", "None"
Are you a healthcare professional working directly with or in close physical proximity to patients who have tested positive for COVID-19?	Single Value from Options: "Yes", "No", "Not sure"
Has someone in your household tested positive for COVID-19?	Single Value from Options: "Yes", "No", "Not sure"
How many people in your household have tested positive for COVID-19?	Integer (# People)
Have you been directly exposed to someone who has tested positive for COVID-19?	Single Value from Options: "Yes", "No", "Not sure"
Do you currently have any of the following health conditions? Select all that apply.	"Asthma", "COPD (Chronic Obstructive Pulmonary Disease)", "Other lung condition", "Cancer (treated in the past year)", "Cardiovascular disease", "Chronic kidney disease", "Diabetes", "Hypertension", "Organ failure requiring a transplant (in the last year)", "Blood disorder requiring hematopoietic stem cell/bone marrow transplant", "Sickle cell disease", "Other autoimmune disease", "Other immunodeficiency disorder", "Other", "None", "Not sure"
How many packs of cigarettes do you usually smoke a day?	Integer (# Packs of Cigarettes per Day)
During your entire life, have you smoked at least 100 cigarettes?	Single Value from Options: "Yes", "No", "Not sure"
How tall are you?	Float (cm)
How much do you weigh?	Float (kg)

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Appendix 4. Minimal data model mapping of attributes to other data sources

Metadata mapping between the common minimal model (attributes in black are from the HGI data dictionary v4 and attributes in red are proposed additions), the Ancestry questionnaire, dbGAP study phs002245, and the study EGAS00001004571.

Abbreviations: GAP: no possible mapping term

Common minimal model Attribute Group	Attribute Name	Ancestry questionnaire	dbGAP (phs002245)	EGAS00001004571
Identification	anonymized_patient_id		SUBJECT_ID	Patient ID
Demographics	age_at_diagnosis	age	AGE	
	sex	Genetic sex	SEX	Sex
	ancestry			
	height	How tall are you?		
	weight	How much do you weigh?		
Vaccine	covid19_vaccine			
	covid19_numb_doses			
	covid19_vaccine_date			
	covid19_vaccine_type			
Comorbidities: Immune system	com_hiv			
	com_immunocomp			
	com_transplant	Organ failure requiring a transplant (in the last year)		
	com_autoimm_rheum			Hypothyroidism Hashimoto's thyroiditis
	com_type_i_diabetes	Diabetes		Type 1 diabetes mellitus
	com_type_ii_diabetes	Diabetes		type 2 diabetes
Comorbidities: Respiratory	com_asthma	Asthma		
	com_chronic_pulm	COPD Chronic Obstructive Pulmonary Disease		chronic lung disease
	com_sleep_apnea	Other lung condition		
Comorbidities: Accessory Digestive organs	com_liver			
	com_gallbl			
	com_pancreas			
Comorbidities: Renal	com_chronic_kidney	Chronic kidney disease		chronic kidney disease
	com_dialysis			Kidney transplant chronic kidney disease Diabetic nephropathy
Comorbidities: cardiovascular	GAP	Sickle cell disease		Brutonagammaglo-bulinemia (XLA)
	com_heart_failure	Cardiovascular disease		Cardiovascular diseases
	com_hypertension	Hypertension		
	com_infarction			
	com_vascular			
	com_stroke			
	com_afib			
Neurological	com_dementia			
	com_neurological			
Comorbidities: Cancer	com_leukemia			

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

	com_lymphoma	Blood disorder requiring hematopoietic stem cell/bone marrow transplantCancer (treated in the past year)		
	com_malignant_solid			Metastasizing lung canc er Prostate cancer
Risk factors	smoking	How many packs of cigarettes do you usually smoke a day?During your entire life, have you smoked at least 100 cigarettes?		
	GAP			obesity

Common minimal model Attribute Group	Attribute name	Ancestry Questionnaire	phs002245 (dbGAP)	EGA
Hospitalization information	hospitalization	Were you hospitalized due to these symptoms?		
	hospitalization_end	How long were you hospitalized?		
	readmission			
	icu_admit			
	icu_duration			
	highest_who_score		COVID19 clinical severity	WHO Classification
	highest_respiratory_support	Were you hospitalized in the Intensive Care Unit (ICU) with oxygen?		
	days_ventilator	Were you hospitalized in the Intensive Care Unit (ICU) with a ventilator?		
	hosp_dvt			
	hosp_thrombo			
	hosp_stroke			
	hosp_infarction			
	hosp_renal			
	hosp_bleeding			
	hosp_hepatic			
Symptoms At admission Pathogen testing	covid19_test	Have you been swab tested for COVID-19, commonly referred to as coronavirus?		
	covid19_test_date			
	covid19_test_type			
	covid19_first_symptoms_date			
	GAP	Did you experience symptoms as a result of your condition? Between the beginning of February and now, have you had any of the following symptoms?		

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

COVID-19 therapy during hospitalization	steroids			steroids
	biologics			Plasma Tocilizumab
	lmwh			
	hcq			
	remdesivir			remdesivir
	GAP			antibiotics
Outcome and Post COVID complications	death		Outcome	Outcome deceased
	cause_of_death			
	date_of_death			
	covi19_complic_cardiovascular			
	covi19_complic_dermatological			
	covi19_complic_endocrine			
	covi19_complic_gastrointestinal			
	covi19_complic_fatigue			
	covi19_complic_Musculoskeletal			
	covi19_complic_mentalhealth			
	covi19_complic_neurological			
	covi19_complic_pulmonary			
	covi19_complic_renal			

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Appendix 5. Attribute mapping between laboratory results

Table illustrating metadata mapping between the laboratory results attributes reported in the ISARIC COVID-19 data dictionary, the WHO CRF and the HGI data dictionary (HGI Clinical Values V4). The attributes colored in blue are present only in the HGI data dictionary.

ISARIC Data dictionary		WHO CRF	HGI Data Dictionary	
Attribute name	CRF	CRF	Attribute name	CRF
daily_lbdat	Date of assessment		lab_result_date	Lab Result Date
daily_hb_lbyn	2.2 Haemoglobin available	Heamoglobin		
daily_hb_lborres	Haemoglobin			
daily_hb_lborresu	Haemoglobin Unit1, g/L|2, g/dL			
daily_wbc_lbyn	2.3 WBC count available			
daily_wbc_lborres	WBC count	WBC Count	lab_wbc	lab_wbc
daily_wbc_lborresu	WBC Unit1, x 10^9 / L|2, x 10^3/µL			
daily_lymp_lbyn	2.4 Lymphocyte count available			
daily_lymp_lborres	Lymphocyte count cells//µL		lab_lymphocytes	Lymphocytes
			lab_cd4	CD4 count
			lab_cd4	CD8 count
daily_neutro_lbyn	2.5 Neutrophil count available			
daily_neutro_lborres	Neutrophil count cells//µL		lab_neutrophils	Neutrophils
daily_haematocrit_lbyn	2.6 Haematocrit available			
daily_haematocrit_lborres	Haematocrit %			
daily_plt_lbyn	2.7 Platelets available			
daily_plt_lborres	Platelet Count	Platelets	lab_platelets	Platelets
daily_plt_lborresu	Platelets Unit 1, x 10^9 / L|2, x 10^3/µL			
			lab_eosinophils	Eosinophils
			lab_basophils	Basophils
daily_aptt_lbyn	2.8 APTT/APTR available			
daily_aptt_lborres	APTT/APTR	APTT/APTR	lab_appt	Activate partial thromboplastin time (APTT)
daily_pt_inr_lbyn	2.9 PT or INR available 1, PT done|2, INR done|3, PT or INR not done			
daily_pt_lborres	PT seconds	PT		
daily_inr_lborres	INR	INR	lab_inr	Prothrombin time test international normalized ratio (PT-INR)
ALT/SGPT	2.10 ALT / SGPT available	ALT/SGPT	lab_alt	ALT
daily_alt_lborres				
daily_bil_lborres	Total Bilirubin	Total bilirubin	lab_bilirubin	Total Serum Bilirubin
daily_bil_lborresu	Total Bilirubin Unit 1, µmol/L|2, mg/dL			

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

daily_ast_lbyn	2.12 AST/SGOT available	AST/SGO	lab_ast	AST (Aspartate amino transferase)
daily_ast_lborres	U/L AST/SGOT			
daily_glucose_lbyn	2.13 Glucose available			
daily_glucose_lborres	Glucose			
daily_glucose_lborresu	Glucose Unit 1, mmol/L|2, mg/dL			
daily_bun_lbyn	2.14 Blood Urea Nitrogen (urea) available	BUN		
daily_bun_lborresu	Blood Urea Nitrogen (urea)1, mmol/L|2, mg/dL			
daily_lactate_lbyn	Lactate	Lactate	lab_ldh	Lactate dehydrogenase
daily_lactate_lborresu	Lactate Unit 1, mmol/L|2, mg/dL			
daily_creat_lbyn	2.16 Creatinine available	lab_creatinine	lab_creatinine	Serum Creatinine
daily_creat_lborres	Creatinine			
daily_creat_lborresu	Creatinine Unit 1, µmol|2, mg/dL|3, umol/L			
		Creatine kinase	lab_ck	Creatine Kinase
daily_sodium_lbyn	2.17 Sodium available			
daily_sodium_lborres	SodiummEq/L	Sodium		
daily_potassium_lbyn	2.18 Potassium available	Potassium		
daily_potassium_lborres	mEq/L			
daily_procal_lbyn	2.19 Procalcitonin available	Procalcitonin	lab_procalcitonin	Procalcitonin
daily_crp_lbyn	2.20 C-reactive protein (CRP) available	CRP	lab_crp	CRP (C-reactive protein)
daily_crp_lborres	C-reactive protein (CRP) mg/L			
		Troponin	lab_trop_t	TROP T (Troponin T)
			lab_ggt	Gamma-Glutamyl Transferase (GGT)
			lab_alp	Alkaline phosphatase
		D-Dimer	lab_d_dimer	D-dimer
		IL-6	lab_il_6	IL-6 (Interleukin 6)
		Ferritin	lab_serum_ferritin	Serum Ferritin
			lab_fibrinogen	Fibrinogen
			lab_monocytes	Monocytes

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.