



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



**EXCELENCIA
SEVERO
OCHOA**

Alfonso Valencia. Ph.D.

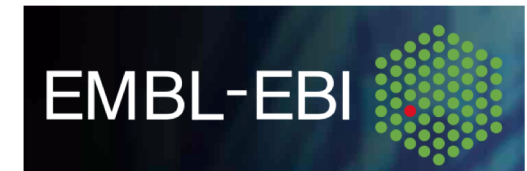
ICREA Prof.

Director Life Sciences Dept

Director INB-ISCIH / ELIXIR-ES

Barcelona Supercomputing Center - BSC

European COVID-19 Data



You are viewing cochrane's screen View Options Speaker View

COVID-19 Data Layers

COVID-19 Portal

SARS-Cov-2 Virus Biology

Human COVID19 Biology

Population Dynamics of Humans + Virus

Service (enabling research)

Research (doing)

Data Hubs + ENA

Other DBs

Federated EGA

National PHI & Infrastructure

Clinical Research

Protein and RNA Structures

Chemobiology

Phylogeny

Protein/RNA Interactomes

Human Genetics

Analysing human responses

6 EMBL-EBI

Unmute Stop Video Participants 51 Chat 3 Share Screen Record Reactions Leave Meeting

Rolf Apweiler

avalenci

Christine Durinx

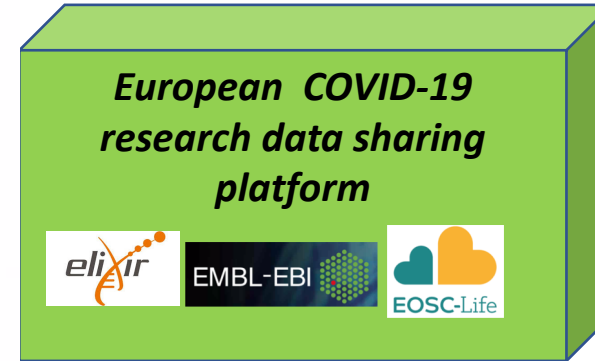
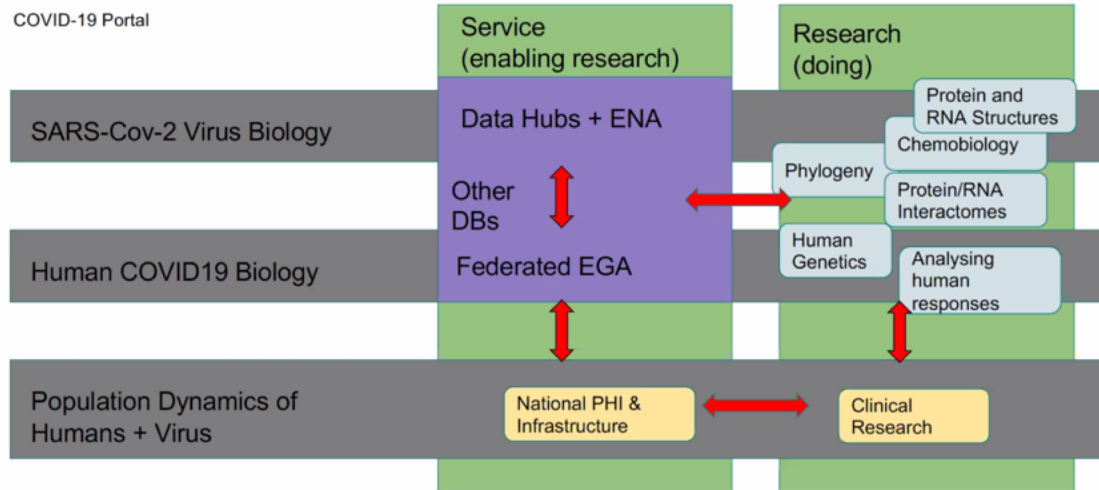
blomberg

marionkoopmans

Building the bridges between molecular and medical information

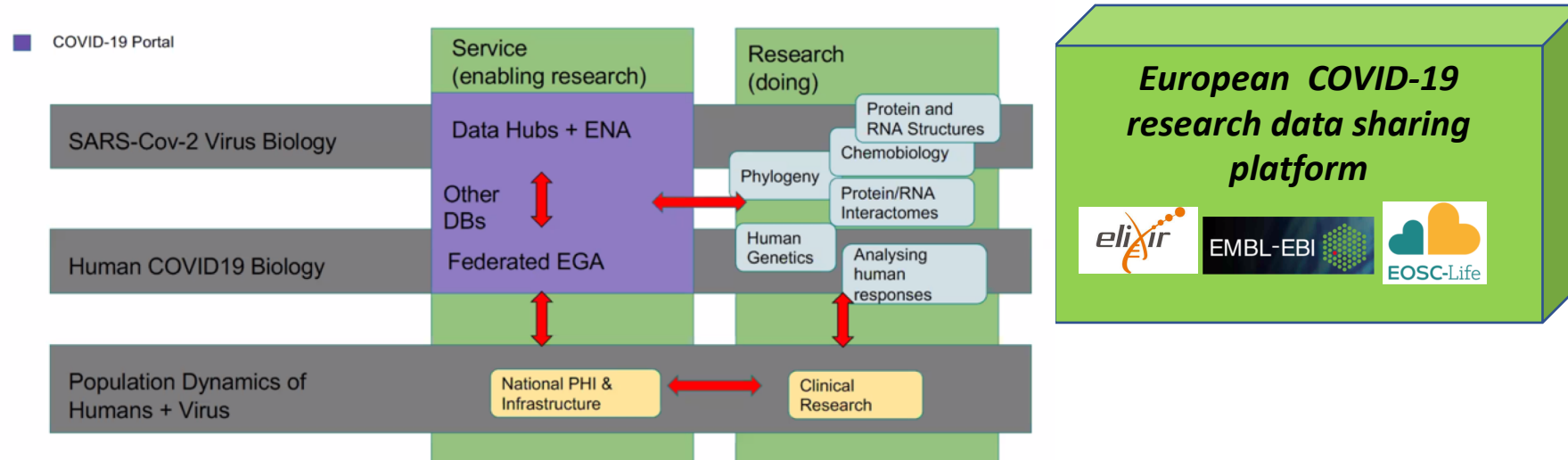
COVID-19 Data Layers

COVID-19 Portal



Building the bridges between molecular and medical information

COVID-19 Data Layers

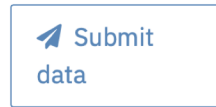


- About
- Data Hubs
- Federated EGA
- Related resources
- Our partners
- Bulk downloads
- Submit data

- Viral Sequences
- Host Sequences
- Expression
- Proteins
- Biochemistry
- Literature



- About
- European Portal
- Support Services



- en
- sv

Genomics & transcriptomics

Protein data

Imaging data

Compound and target data

Health data

Spanish National Repository for COVID-19

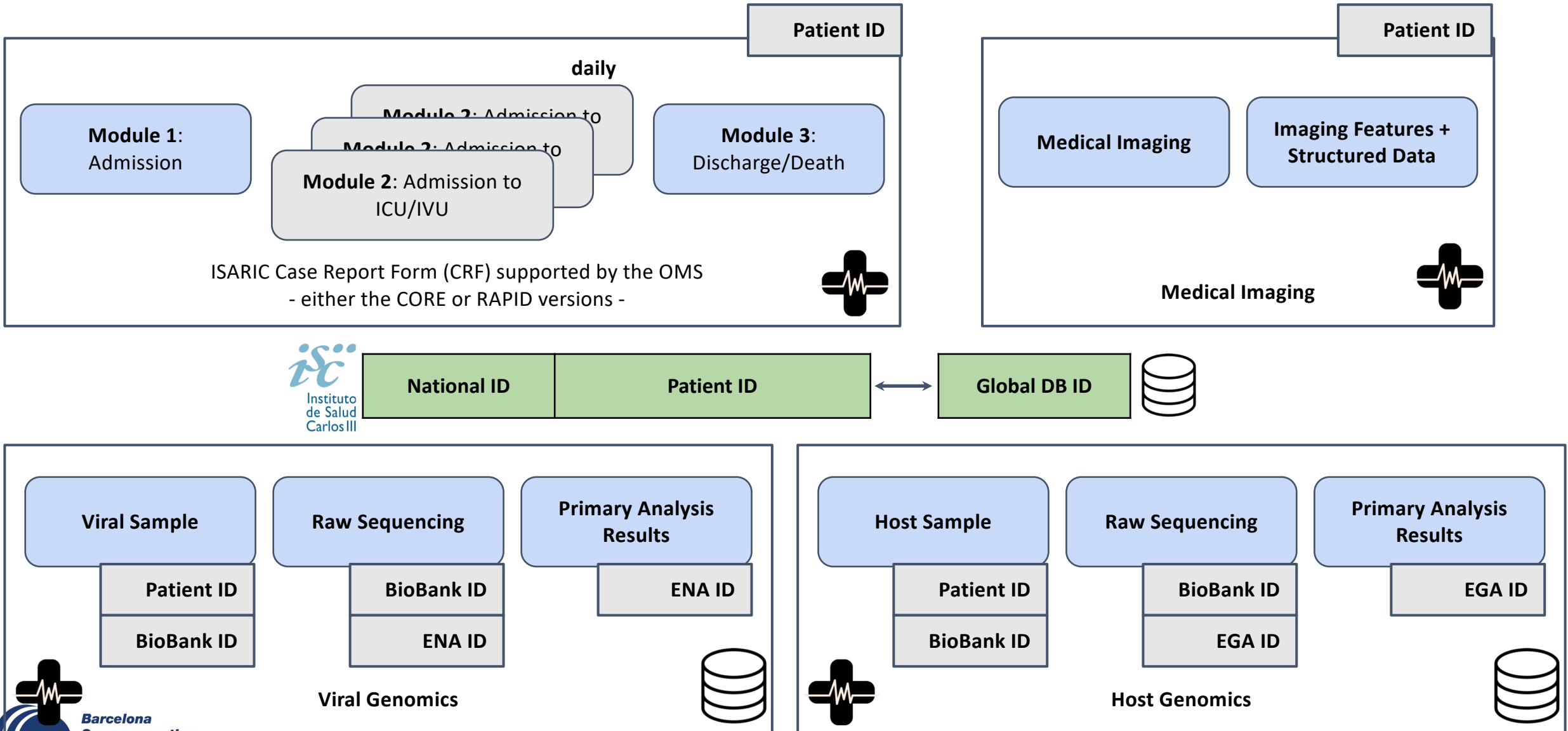
Viral and Microbiota Data: Aligned with **ENA** protocols and technology. Coordinated with **ENA** repository.

Host Data: Based on local **EGA** technology. Coordinated with EGA-CRG

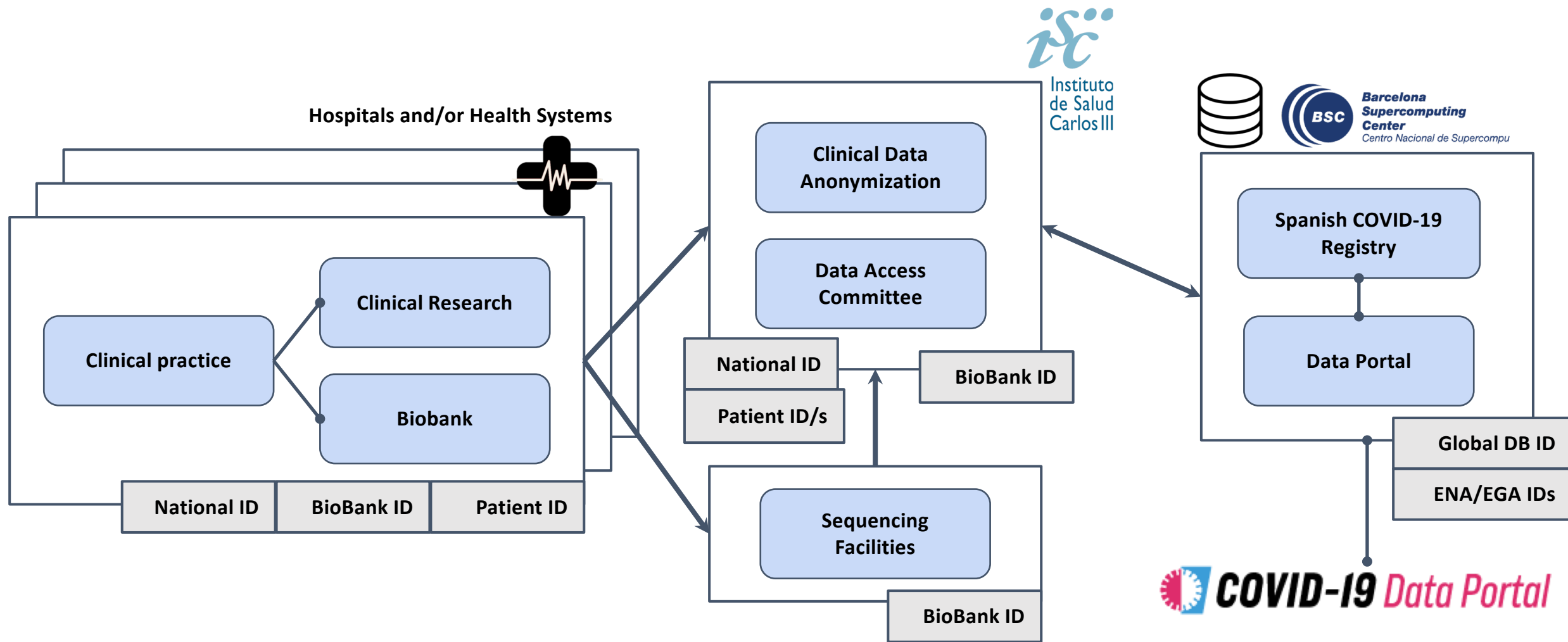
Imaging Data: Based on **XNAT** technology.

Clinical Data Collection: Based on the ISARIC/WHO COVID-19 **Case Report Forms**. Use of **RedCap** to facilitate data capture and submission

Spanish COVID-19 Registry



Expected data flow across the Spanish system



1

2

3

Spanish COVID-19 Registry: IDs, privacy and provenance.

- ❑ Each patient receives a **Patient ID** at admission. This patient ID can be specific to the COVID-19 or it can be an existing patient ID in their hospitals. Depending on how the health system is organized, patient IDs can be unique to each hospital or across the entire system.
- ❑ As a mechanism to guarantee that all information for a specific patient is managed together, the ISCIII will keep a registry with the correspondence between patient ID/s and their personal **National ID**. In this way, a patient that has been treated at different places (hospitals, regions, etc) can have all their information together.
- ❑ A **Global DB ID** (e.g. ESC19XXXXXXX), which is *permanent* and *unique*, will be issued as part of the submission of the anonymized data for a given patient. The ISCIII will keep the correspondence between that Global DB ID and the IDs associated to the patient. It will contribute towards **data provenance** and also will contribute towards **data anonymization** where all data leading to the re-identification of patients will be removed.

Running Projects

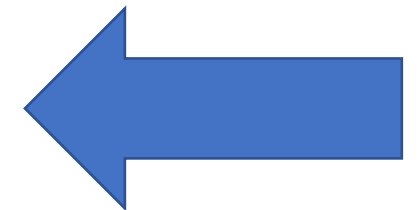


COV20_00140	Addressing unknowns of COVID-19 transmission and infection combining pathogen genomics and epidemiology to inform public health interventions	AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS, M.P.	Iñaki Comas Espadas	There is a growing pressure to understand the transmission patterns of COVID-19. As a new, emergent virus we have important gaps in the epidemiology of its infection. Those gaps translate also to limited understanding on who is infectious and for how long and how is connected to clinical outcome. Here we propose several objectives that involve combining viral genome data to epidemiological, clinical information to address those gaps. We will obtain viral complete genome sequences and epidemiological data and combine to (1) inform public health measures by revealing highways of transmission; (2) identify when an individual is infectious; (3) identify early warnings of local spread; (4) viral diversity connected to immune, drug and diagnostic; (5) complement current diagnostic approaches, and (6) identify the connection to adverse clinical outcomes. To this end, we have put together a nation-wide group of infectious disease, genomics, bioinformaticians and clinical researchers from more than 30 institutions.	01/05/2020
COV20_00622	Determinantes Genéticos y Biomarcadores genómicos de riesgo en pacientes con infección por coronavirus SARS-COV-2	Consortio Centro de Investigación Biomédica en Red M.P. (CIBER)	Angel Carracedo Alvarez (PI) Pablo Lapunzina (CO-PI)	Objetivo: Búsqueda de biomarcadores genómicos de riesgo que puedan predisponer a infección más grave estratificación pronóstica y/o biomarcadores genómicos de respuesta en pacientes que hayan recibido tratamiento durante la infección. Se realizará un estudio de asociación de genoma completo (GWAS) en una muestra de 7000 pacientes españoles y un estudio de secuenciación de genoma completo en un grupo de pacientes seleccionados (300), recogidos por un consorcio nacional e internacional que incluye más de 15 hospitales, complementado con grupo de investigación de alto nivel. Los datos clínicos (signos, síntomas y hallazgos) y los antecedentes personales y epidemiológicos se recogerán en una base de datos común diseñada ad hoc para este proyecto. Encontraremos así determinantes genético/genómicos de riesgo y mala evolución que permitan una mejor estratificación de los pacientes y la adecuación y optimización de protocolos terapéuticos.	01/05/2020
COV20_00110	Factores de riesgo, pronóstico personalizados y seguimiento a un año de los enfermos ingresados en las unidades de Ciudadanos Intensivos Españolas infectados por el virus COVID:CIBERSESUCICOVID	CONSORCIO CENTRO DE INVESTIGACION BIOMEDICA EN RED (CIBER)	Antoni Torre Martí	El proyecto que se presenta tiene el objetivo de conocer los factores de riesgo y pronóstico y desenlaces a corto y largo plazo (6 meses y un año) de los pacientes con COVID19 ingresados en las UCIs Españolas. Otros dos objetivos adicionales son el estudio epigenético y biomarcadores esenciales de enriquecimiento predictivo para ayudar a individualizar el tratamiento en base a las vías biológicas alteradas en cada paciente. CIBERES cuenta con el aval de SEMICYUC y SEPAR. La hipótesis al final de la pandemia que habrán ingresado 10000 pacientes en las UCIs españolas seleccionadas a través de SEPAR Y SEMICYUC. Los datos clínicos se analizarán con inteligencia artificial. Se hará un análisis interino con los 1000 primeros pacientes lo que puede ser de utilidad para el manejo clínico de los pacientes con COVID19 graves.	30/04/2020

Viral genomics

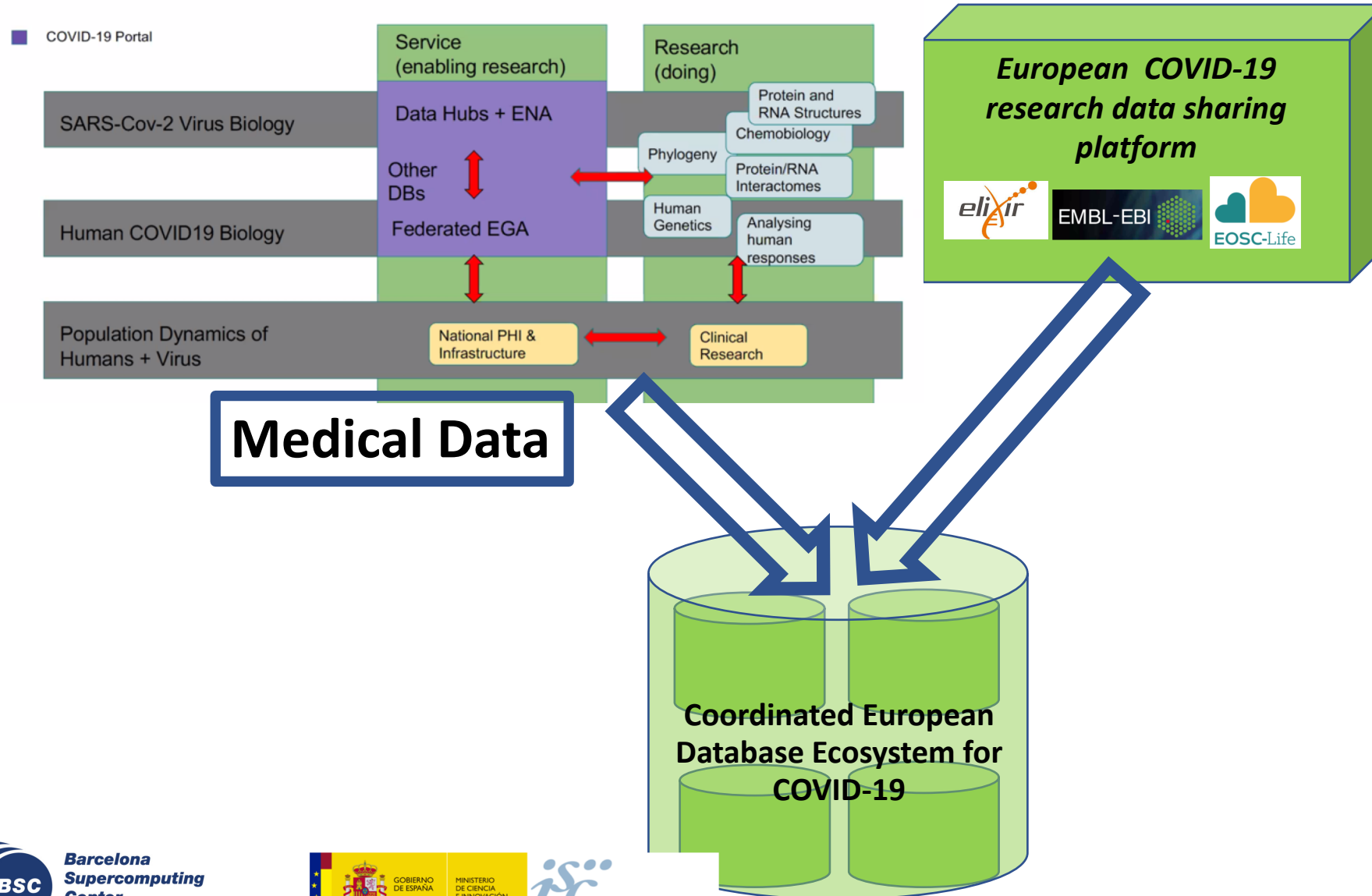
Human genomics

Medical data (i.e. ICU)



Building the bridges between molecular and medical information

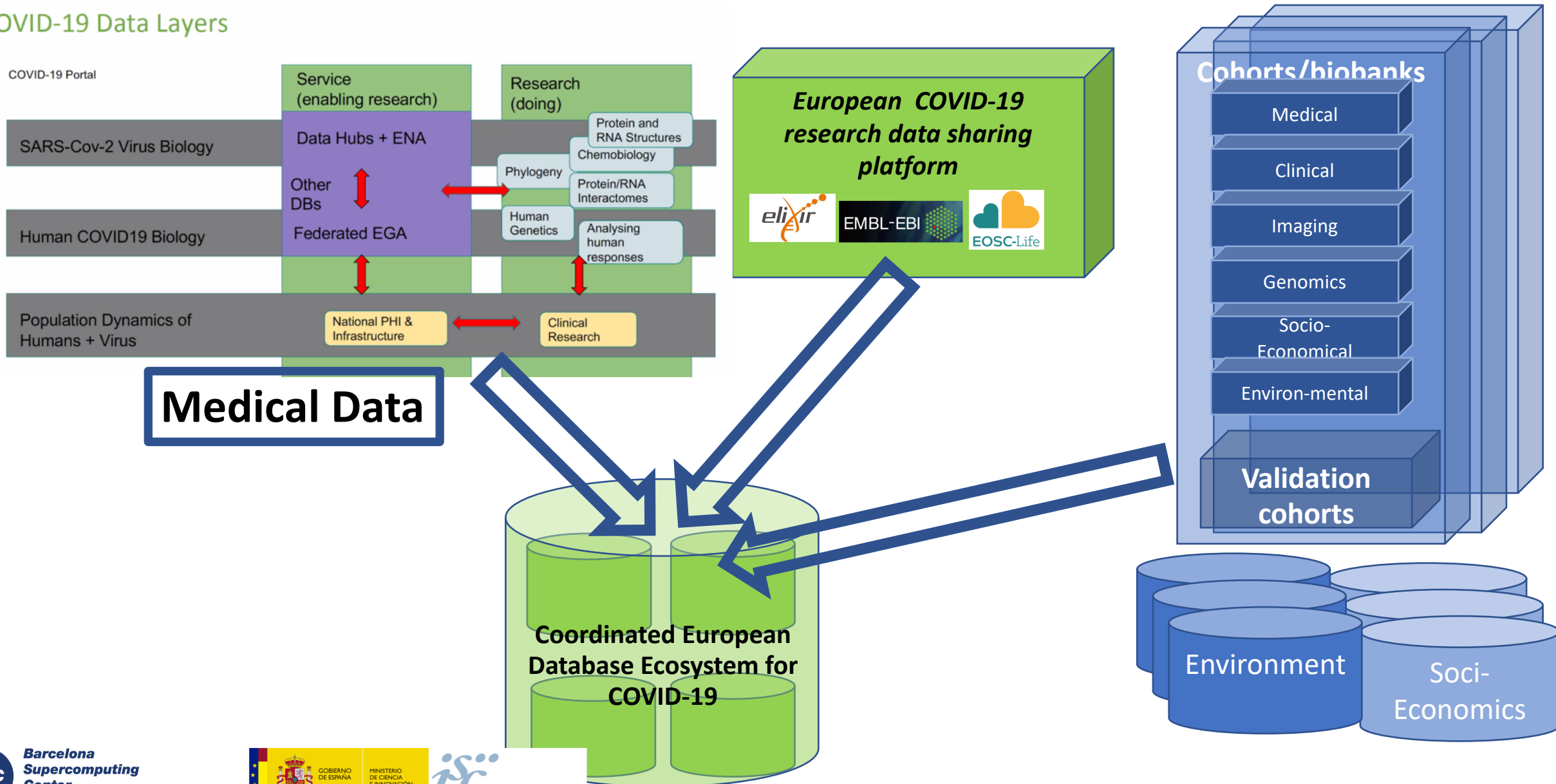
COVID-19 Data Layers



Building the bridges between molecular and medical information

COVID-19 Data Layers

■ COVID-19 Portal



Data Harmonization: Integration



National COVID Cohort Collaborative



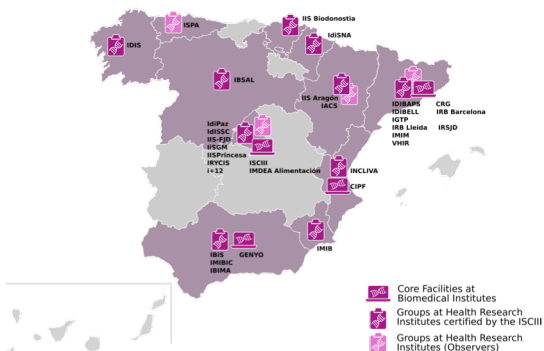
Final Merge

- OMOP versioned data from all sources will be combined into analytic database
- Analytic database will migrate to Palantir Analytic Platform

EHR DREAM Challenge COVID-19

The banner features a virus icon on the left and various medical and data icons on the right. Below the banner are logos for the following organizations: University of Washington, ITHS, NIH National Center for Advancing Translational Sciences, National Center for Data to Health, and SageBionetworks.

Spanish National Bioinformatics Institute (INB)
Spanish Node of ELIXIR (ELIXIR-ES)



TransBioNet
Network of Translational Bioinformatics Groups in Research Institutes of Spanish Hospitals.
(INB hosted)



TransBioNet (mainly Institutes in Research Hospitals)

Acronym	TransBioNet member
GENYO	Centre for Genomics and Oncological Research
IBIMA	Instituto de Investigación Biomédica de Málaga
IBIS	Instituto de Biomedicina de Sevilla
IMIBIC	Instituto Maimónides de Investigación Biomédica de Córdoba
IACS	Instituto Aragonés de Ciencias de la Salud
IIS Aragón	Instituto de Investigación Sanitaria Aragón
ISPA	Instituto de Investigación Sanitaria del Principado de Asturias
IBSAL	Instituto de Investigación Biomédica de Salamanca
CRG	Centre for Genomic Regulation
IDIBAPS	Institut d'Investigacions Biomèdiques August Pi i Sunyer
IDIBELL	Instituto de Investigaciones Biomédicas de Bellvitge
IGTP	Germans Trias i Pujol Research Institute
IRB Barcelona	Institute for Research in Biomedicine Barcelona
IRBLEida	Institut de Recerca Biomèdica de Lleida
IRSJD	Institut de Recerca Sant Joan de Déu
PSMAR-IMIM	Parque de Salud MAR (PSMAR) / Instituto Hospital del Mar de Investigaciones Médicas (IMIM)
VHIR	Vall d'Hebron Institut de Recerca
CIPF	Centro de Investigación Príncipe Felipe
INCLIVA	Institut d'Investigació Sanitària INCLIVA
IDIS / FPGMX	Instituto de Investigación Sanitaria de Santiago de Compostela / Fundación Pública Galega de Medicina Xenómica
IdISBa	Instituto de Investigación Sanitaria Islas Baleares
i+12	Instituto de Investigación Hospital 12 de Octubre
IdiPaz	Instituto de Investigación Hospital Universitario La Paz
IDIPHIM	IIS Instituto de Investigación Sanitaria Puerta de Hierro
IdISSC	Instituto de Investigación Sanitaria Hospital Clínico San Carlos
IIS-FJD	IIS-Fundación Jiménez Díaz
IISGM	Gregorio Marañón Health Research Institute
IISPrincesa	Instituto de Investigación Sanitaria Hospital de la Princesa
IMDEA	Instituto Madrileño de Estudios Avanzados
IRYCIS	Instituto Ramón y Cajal de Investigación Sanitaria
ISCIII	Instituto de Salud Carlos III
IMIB	Instituto Murciano de Investigación Biosanitaria
IDISNA	Instituto de Investigación Sanitaria de Navarra
IISBiodonostia	Instituto de Investigación Sanitaria BioDonostia
ibs.GRANADA	Instituto de Investigación Biosanitaria de Granada
IDIVAL	Instituto de Investigación Sanitaria Marqués de Valdecillas
CNIC	Centro Nacional de Investigaciones Cardiovasculares
IIB Sant Pau	Institut d'Investigació Biomèdica de Sant Pau

Hospitals



**H V del Rocio
Sevilla**



**H Clínic
Barcelona**

**Hospital Universitario
12 de Octubre**

**H 12 Octubre
Madrid**

Regional Health Systems

SAS

CatSalut
(AQUAS)

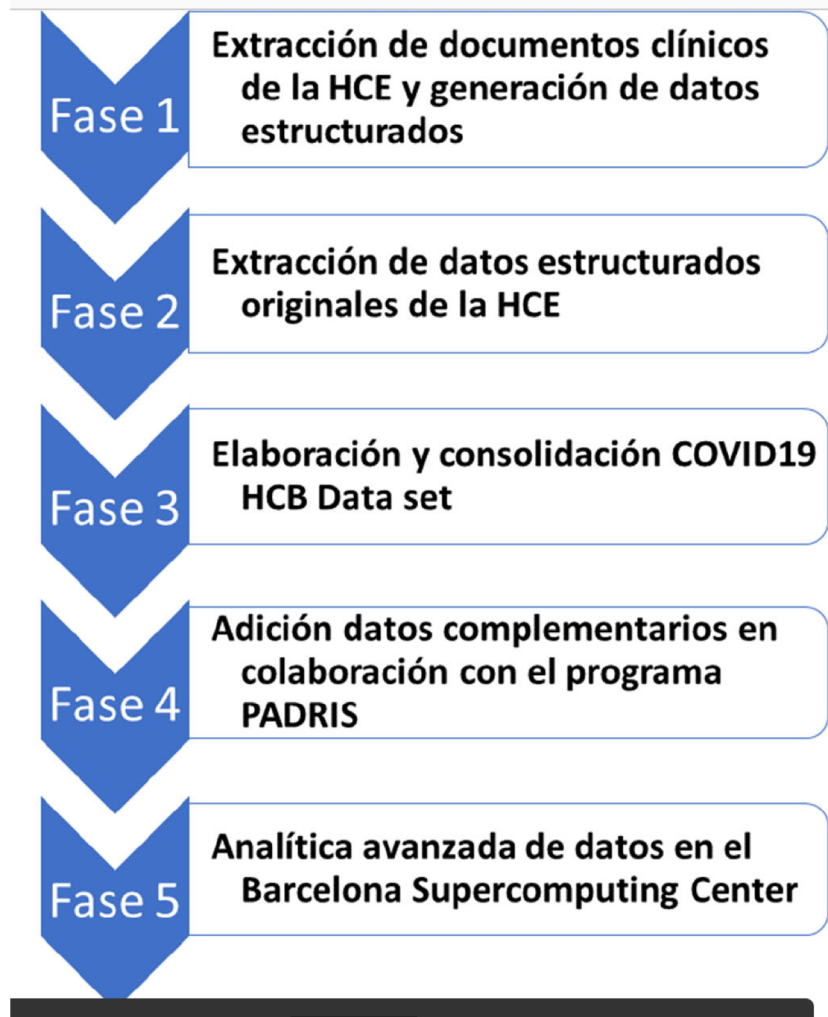
SERMAS

*BSC
Research contracts as part of the
National NLP Plan*



<https://inb-elixir.es/transbionet>

Access to COVID-19 Related Medical Information



COVID-19 clinical questions:

1. Prediction of the **risk of serious respiratory involvement** at the time of hospital admission or in its early stages.
2. Prediction of the **global evolution of the admitted** patient based on his personal characteristics, clinical evolution, biological parameters, findings in diagnostic imaging tests, and therapeutic and life support measures.
3. **Stratification of patients** according to the evolutionary pattern of the disease.

variables:

- 1.- **Structured variables** at origin (EHR)
- 2.- Variables encoded using standard terminologies (ICD10 - MC, SNOMEDCT), **obtained applying NLP** to the clinical texts of the EHRs (context of the infection, clinical signs and symptoms, pathophysiological states, diagnoses and procedures)

Matching variables and codes

Medical complications

Concept

Relevance

Priority

Synonyms

SNOMED

ICD10

GRUPO	COMPLICACIONES			
SUBGRUPO				
orden	1	2	3	4
Concepto	convulsiones	encefalitis	encefalopatía	ictus
Relevancia	Primera manifestación posible si encefalitis. Recogido por cualquier especialidad, aunque no siempre son crisis reales (sino movimientos anormales de causa indeterminada)	Casos aislados y descrito incluso ante ausencia de clínica respiratoria como manifestación principal de la enfermedad. Puede ser por diversas causas con diferentes implicaciones de tratamiento (por el efecto directo del virus en el sistema nervioso central, -presencia de virus en LCR demostrada en algunos- o puramente inflamatorio)	Principal síntoma recogido en las historias puede ser "confusión", con prevalencia descrita en COVID+ entre 9-25% de pacientes hospitalizados y hasta 65% en pacientes en UCI, relacionado con gravedad respiratoria también y asociado a mayor mortalidad	Principal complicación neurológica del COVID. Afectación de grandes vasos, aterotrombosis arterial y venosa. Mayor en aquellos casos respiratoriamente graves
Prioridad	alta	alta	alta	alta
Sinónimos	convulsión CTCG crisis ausencia crisis de desconexión del medio estado epiléptico estatus epilepticus estatus	rombencefalitis meningoencefalitis lesiones cerebrales desmielinizantes inflamación del sistema nervioso central (SNC)	alteración del estado mental alteración cognitiva	ictus isquémico ictus hemorrágico hematoma cerebral embolia cerebral trombosis cerebral infarto cerebral accidente vascular cerebral ACV
SNOMED CT	91175000	45170000	81308009	432504007
Descript (PT)	convulsión	encefalitis	trastorno del encéfalo	infarto cerebral
Descript (FSN)	convulsión (hallazgo)	encefalitis (trastorno)	trastorno del encéfalo (trastorno)	infarto cerebral (trastorno)
ICD10	R56.9	G04.90	G93.40	I63.9
Descript	Convulsiones no especificadas	Encefalitis y encefalomielititis, no especificadas	Encefalopatía, no especificada	Infarto cerebral, no especificado
Obs codificación ICD10	Hay otros códigos más específicos para identificar el tipo de epilepsia y el estado epiléptico			

Data Model: Tests, treatments, and observations

	A	B	C	D	E	F	G	H	I	J
1			Visor COVID-19 H120							
2			Observaciones							
3										
4		Fecha-hora:	3/29/2020 12:36							
5		+ Enlaces:	Recursos de referencia [+]							
14										
15										
16				Estandarización [+]						
17				Enlace terminológico						
18	Árbol	Componente	Descripción	Tipo de enlace	Cardinalidad	Estándar	Contexto	Código	Término	Referencias
19	c1	c4	c5	c6	c7	c8	c9	c11	c14	c20
20	-	COMPOSITION		Componentes		SNOMED CT	artefacto de registro		-1	
21	-	SECTION		Componentes		SNOMED CT	artefacto de registro		-1	
22	-	ENTRY		Componentes		SNOMED CT	artefacto de registro		-1	
23	+	Cluster	Observaciones epidemiológicas en relación con COVID-19	Componentes		SNOMED CT	artefacto de registro	63261000122108	Observaciones epidemiológicas en relación con COVID-19	
49	+	Cluster	Observaciones sobre síntomas en relación con COVID-19	Componentes		SNOMED CT	artefacto de registro	63271000122103	Observaciones sobre síntomas en relación con COVID-19	
74	+	Cluster	Observaciones clínicas en relación con COVID-19	Componentes		SNOMED CT	artefacto de registro	63281000122100	Observaciones clínicas en relación con COVID-19	
127	+	Cluster	Observaciones de laboratorio en relación con COVID-19	Componentes		SNOMED CT	artefacto de registro	63291000122102	Observaciones de laboratorio en relación con COVID-19	

Observaciones epidemiológicas en relación con COVID-19
Observaciones sobre síntomas en relación con COVID-19
Observaciones clínicas en relación con COVID-19
Observaciones de laboratorio en relación con COVID-19

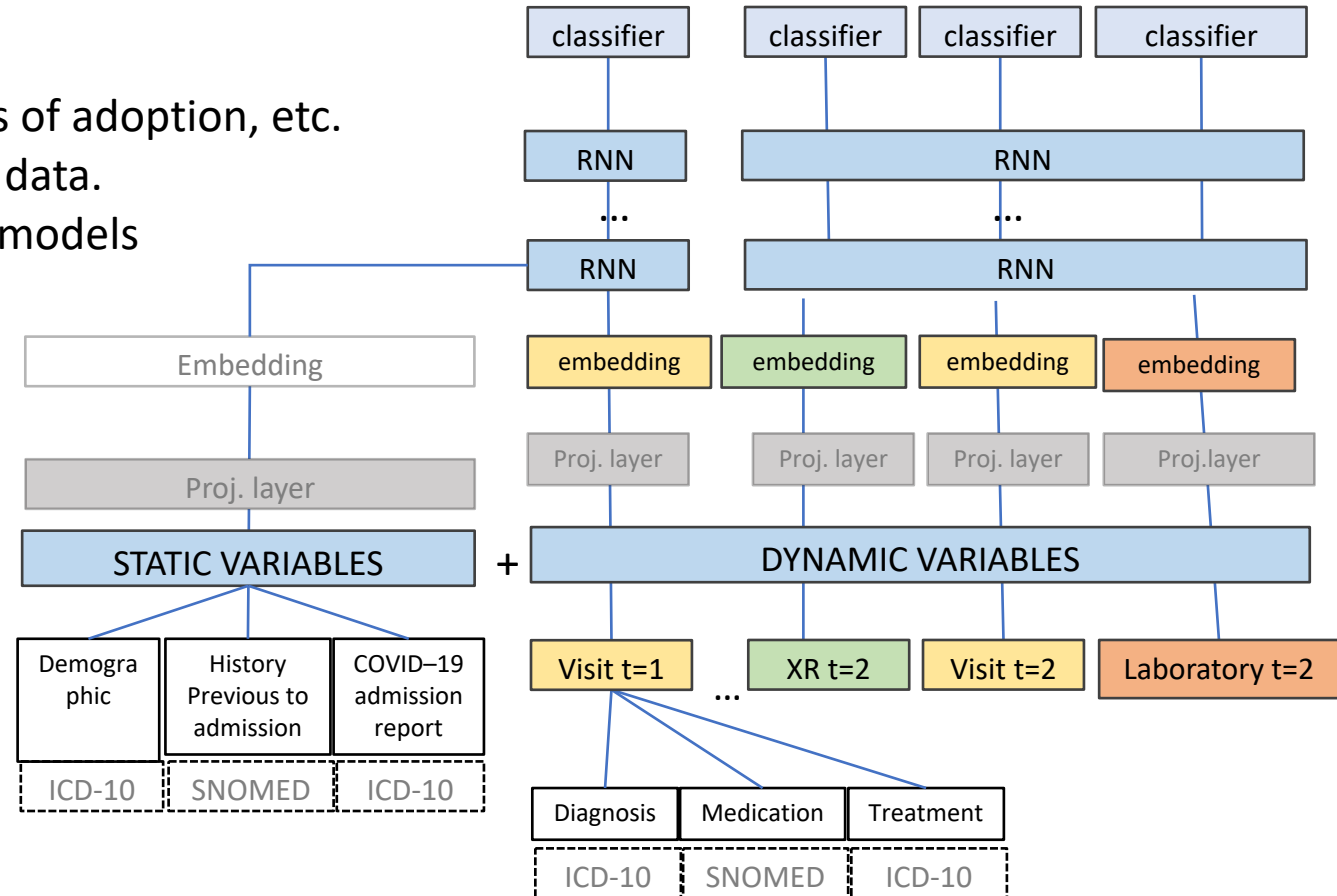
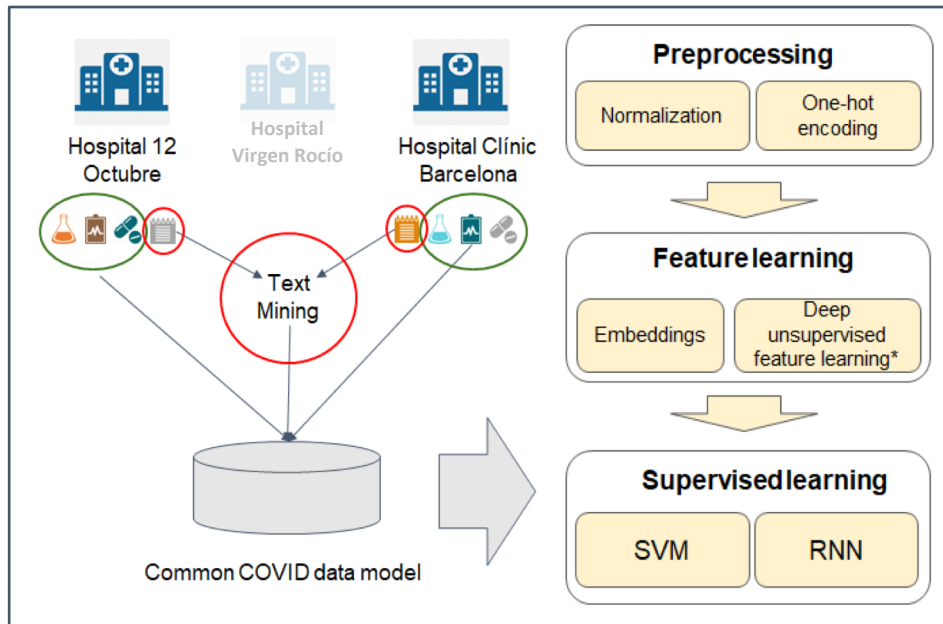
Prognosis of patients with COVID-19

OBJECTIVES

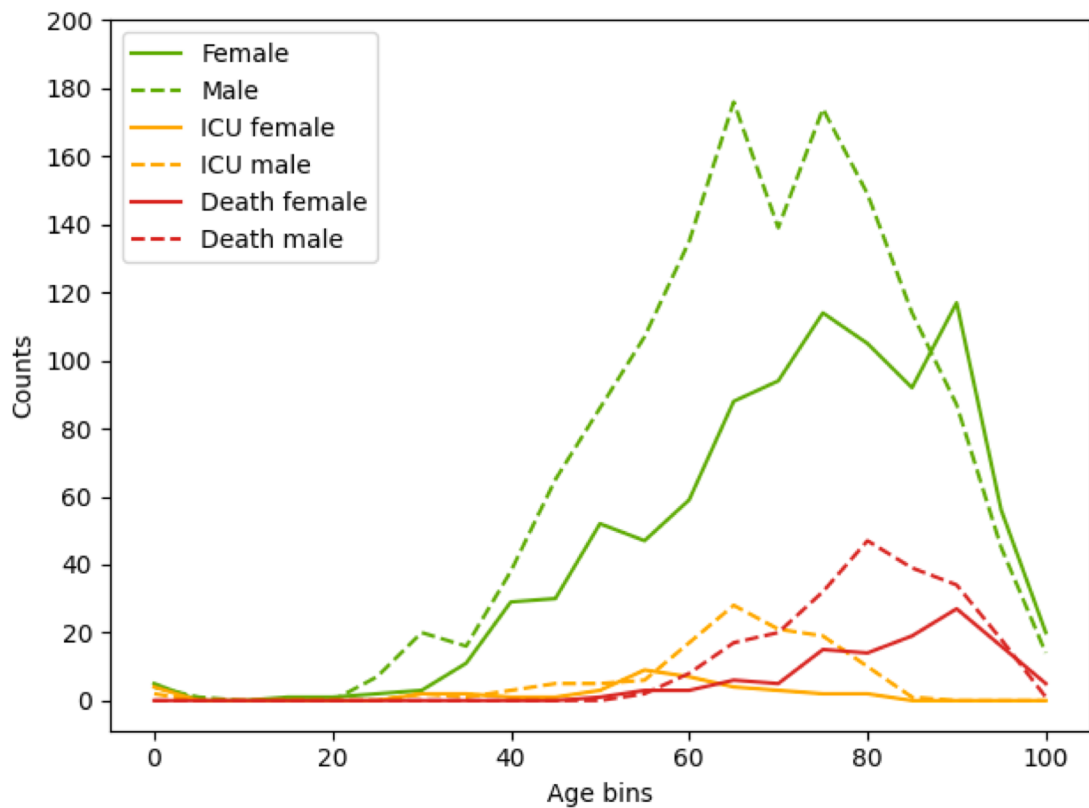
AI to predict the course of the COVID-19 disease based on the patient's EHR and clinical-care event records

MAIN TASKS:

1. Common data model & protocol: scalable with degrees of adoption, etc.
2. Development of COVID predictive models using clinica data.
3. Use patient's entire medical history for the predictive models

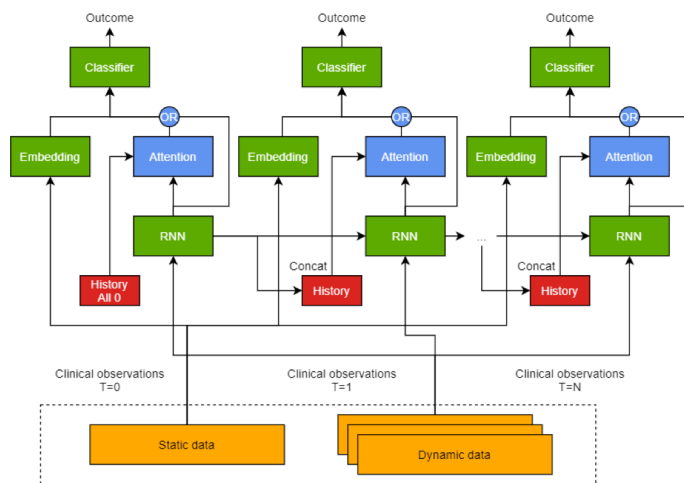


2,307 COVID-19 patients



3	EDAD/AGE	si
4	SEXO/SEX	si
5	DIAG_ING/INPAT	si COVID/NO_COVID
6	F_INGRESO/ADMISSION_D_ING/INPAT	FECHA REFERENCIA para todas las fechas
7	F_ENTRADA_UC/ICU_DATE_IN	si
8	F_SALIDA_UCI/ICU_DATE_OUT	no (ya tenemos los dias que ha estado en uci)
9	UCI_DIAS/ICU_DAYS	si
10	F_ALTA/DISCHARGE_DATE_ING	
11	MOTIVO_ALTA/DESTINY_DISCHARGE_ING	si
12	F_INGRESO/ADMISSION_DATE_URG/EMERG	
13	HORA/TIME_ADMISION/ADMISSION_URG/EMERG	
14	ESPECIALIDAD/DEPARTMENT_URG/EMERG	
15	DIAG_URG/EMERG	NO Parece que l diagnóstico es más preciso en tabla ?
16	DESTINO/DESTINY_URG/EMERG	NO. Todos son ingresos
17	HORA/TIME_CONSTANT_PRIMERA/FIRST_URG/EMERG	si? hay algunas medidas que no salen en la tabla de laboratotirio, aunque hay que comprobar los counts (hay muchos 0) t ver si vale la pena
18	TEMP_PRIMERA/FIRST_URG/EMERG	si? (hay 548 = 0)
19	FC/HR_PRIMERA/FIRST_URG/EMERG	si? (hay 529 = 0)
20	GLU_PRIMERA/FIRST_URG/EMERG	si? (hay 2283 = 0)
21	SAT_O2_PRIMERA/FIRST_URG/EMERG	si? (sólo hay 499 = 0)
22	TA_MAX_PRIMERA/FIRST/EMERG_URG	si? (842 = 0)
23	TA_MIN_PRIMERA/FIRST_URG/EMERG	si? (843 = 0)
24	HORA/TIME_CONSTANT_ULTIMA/LAST_URG/EMERG	si? hay algunas medidas que no salen en la tabla de laboratotirio, aunque hay que comprobar los counts (hay muchos 0) t ver si vale la pena
25	FC/HR_ULTIMA/LAST_URG/EMERG	
26	TEMP_ULTIMA/LAST_URG/EMERG	
27	GLU_ULTIMA/LAST_URG/EMERG	2278 = 0
28	SAT_O2_ULTIMA/LAST_URG/EMERG	381 = 0
29	TA_MAX_ULTIMA/LAST_URGEMERG	
30	TA_MIN_ULTIMA/LAST_URG/EMERG	
**		

Predicting Disease Outcome with RNNs

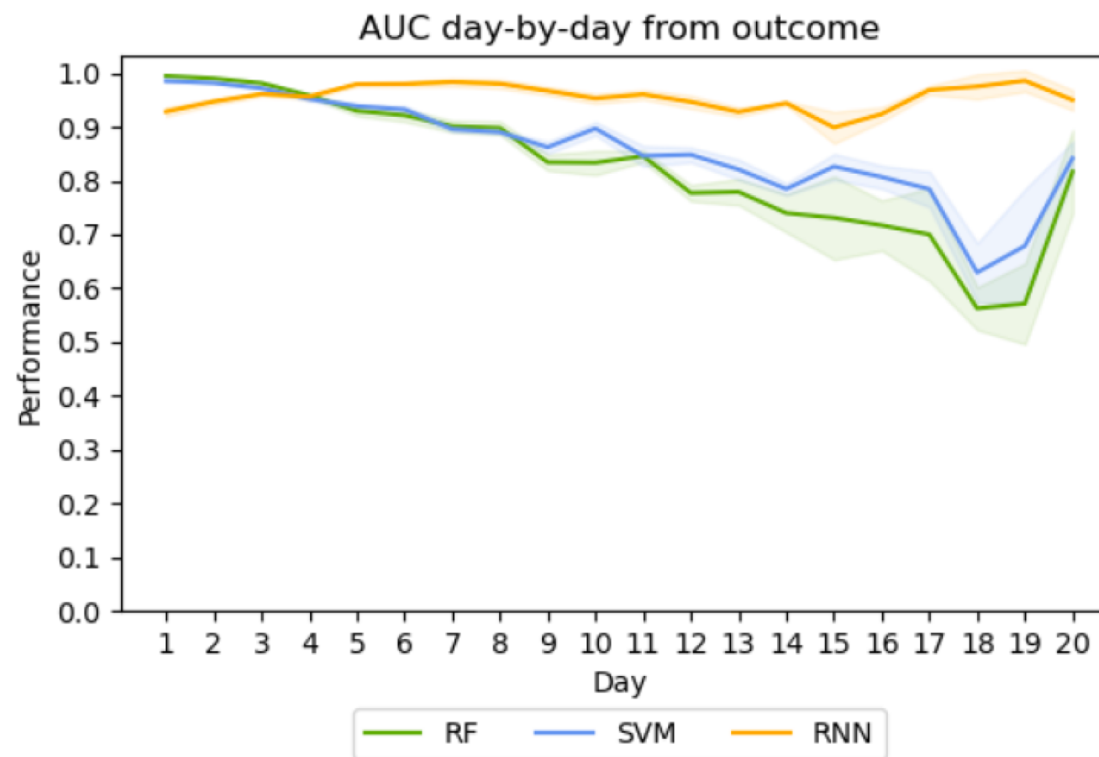


RNN's params

- Batch size
- Dropout
- Static embedding size
- Dynamic embedding size
- RNN hidden state size
- RNN architecture (either LSTM or GRU)
- Loss function (and γ value for the Focal Loss)
- Attention mechanism
- Learning rate
- Number of features (top-k)

Method	Accuracy	Sensitivity	Specificity	AUC
SVM	0.9026±0.0125	0.5077±0.0360	0.9753±0.0093	0.9166±0.0140
Random Forest	0.9111±0.0084	0.5552±0.0132	0.9766±0.0103	0.9372±0.0110
Our best model	0.9197±0.0094	0.7043±0.0437	0.9689±0.0096	0.9538±0.0047
Our top 5 models	0.9166±0.0085	0.7135±0.0492	0.9631±0.0155	0.9552±0.0042
Our top 10 models	0.9179±0.0089	0.7170±0.0465	0.9638±0.0128	0.9548±0.0052

Table 3: Final evaluation results on cross-validations.



TOWARDS PREDICTING THE EVOLUTION OF COVID-19 MORTALITY RISK:
A RECURRENT NEURAL NETWORK APPROACH

CP Carrino, A Gonzalez-Agirre, J Armengol-Estapé, A Gutiérrez-Fandiño, D Pérez Fernández, M Villegas. A Valencia

NLP applications to COVID-19 EHRs

Pre-
processing

Term
extraction

Doc
similarity

Annotation &
guidelines

Model
training

Automatic
labelling

Output



0

1

2

3

4

5



Input: emergency
discharge reports,
radiology reports (Chest
X/CT-scans (H12O,
Clínic)

Document
selection for
annotation

Expert knowledge
(human in the loop)
adaptation/fine tuning

Deep learning : procedures,
symptoms, drugs, diseases,
negation, pathogens, socio-
economic...

Structured
clinical data

SPECIES **SINTOMA**
Al ingreso, la paciente reportó cuatro defecaciones al día con sangre

ENFERMEDAD **ENTIDAD-OBSERVABLE**
Las pruebas analíticas indicaron anemia grave (hemoglobina 7

PROCEDIMIENTO **ENFERMEDAD**
Una sigmoidoscopia mostró úlceras grandes y profundas en el colon

- Detection of COVID-19 cohorts
- Extraction of: other diagnoses, symptoms/ severity, and outcomes
- Early warning: predict the clinical course of the disease (increase in FIO2, admission to intensive care)

Building the bridges between molecular and medical information

COVID-19 Data Layers

COVID-19 Portal

