

# Big data y Medicina Personalizada en Andalucía

---

**MEDICINA DE PRECISIÓN: CIENCIA Y TECNOLOGÍA AL SERVICIO DE LA TRANSFORMACIÓN  
DEL SISTEMA SANITARIO**  
UIMP (Virtual) 1 de julio de 2021

## Joaquín Dopazo

Clinical Bioinformatics Area, Fundación Progreso y Salud,  
Instituto de Biomedicina de Sevilla (IBIS)  
FPS/ELIXIR-es, RD-Bioinformatics CIBERER)  
Hospital Virgen del Rocío, Sevilla, Spain

<http://www.clinbioinfospa.es>

<http://www.babelomics.org>

 @xdopazo, @ClinicalBioinfo



**Junta de Andalucía**

Consejería de Salud y Familias

FUNDACIÓN PROGRESO Y SALUD



# Clinical data is becoming Big Data

Medical knowledge doubles every 2 months (2020)

Journal List > Trans Am Clin Climatol Assoc > v.122; 2011 > PMC3116346

## Transactions of the American Clinical and Climatological Association

Trans Am Clin Climatol Assoc. 2011; 122: 48-58.

PMCID: PMC3116346

PMID: 21686208

### Challenges and Opportunities Facing Medical Education

Peter Densen, MD

Author information Copyright and License information Disclaimer

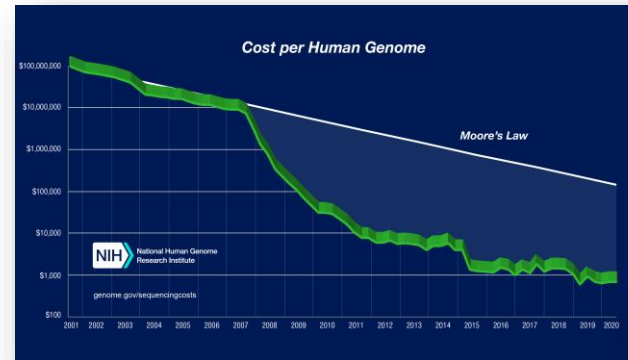
This article has been cited by other articles in PMC.

#### Abstract

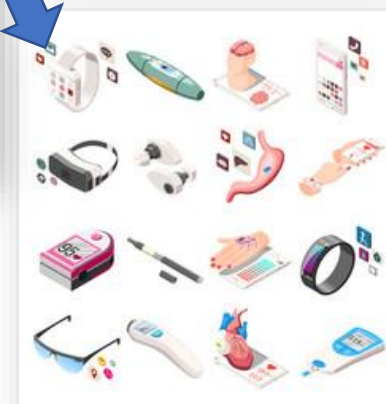
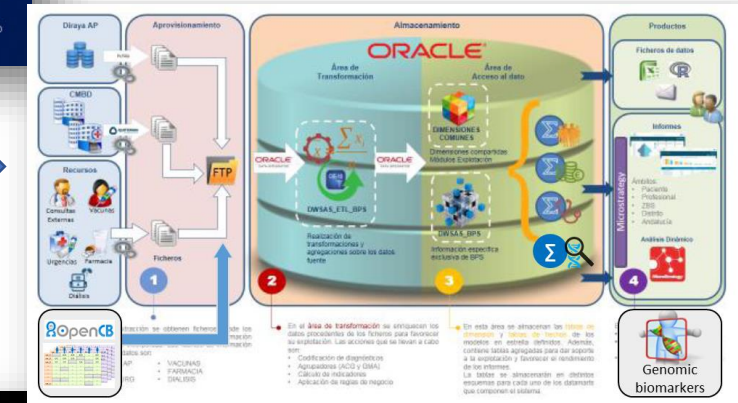
Go to:

Medical education is at a crossroads. Although unique features exist at the undergraduate, graduate, and continuing education levels, shared aspects of all three levels are especially revealing, and form the basis for informed decision-making about the future of medical education.

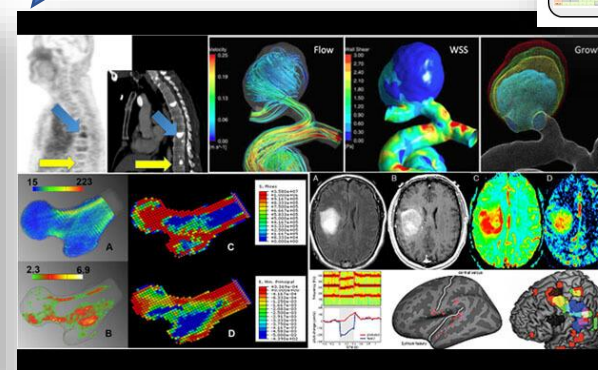
This paper describes some of the internal and external challenges confronting education. Key internal challenges include the focus on disease to the relative inpatient versus outpatient education, and implications of a faculty whose research is at the molecular or submolecular level. External factors include the exponential technological ("disruptive") innovations, and societal changes. Addressing this institutional leadership with an eye to 2020 and beyond—the period in which they begin their careers. This paper presents a spiral-model format for a curriculum on disease mechanisms, that addresses many of these challenges and incorporates principles.



Whole genomes by < 600€



Even more data will come from portable medical devices

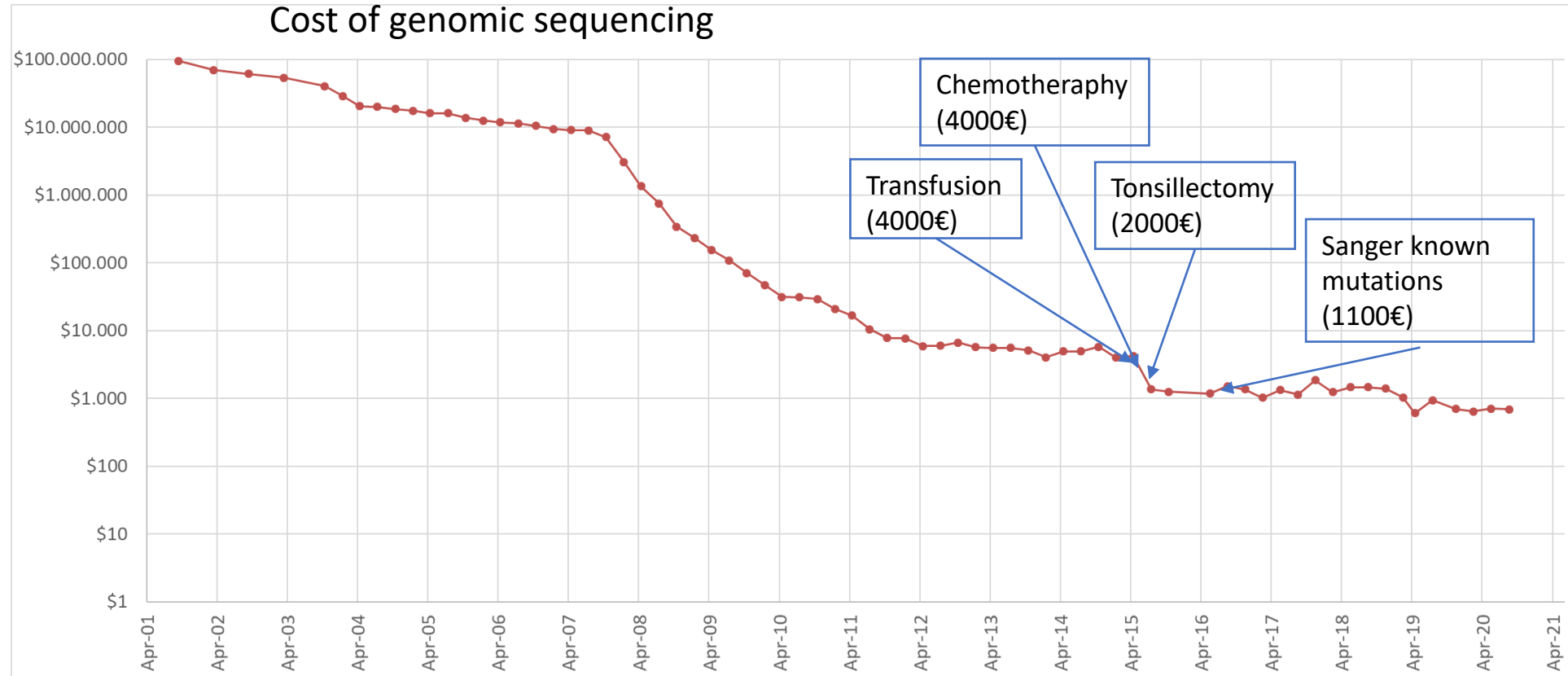


TBs of medical image are produced on daily basis

Clinical data:  
The Population Health Database

Medical image

# The future of genomic data generation



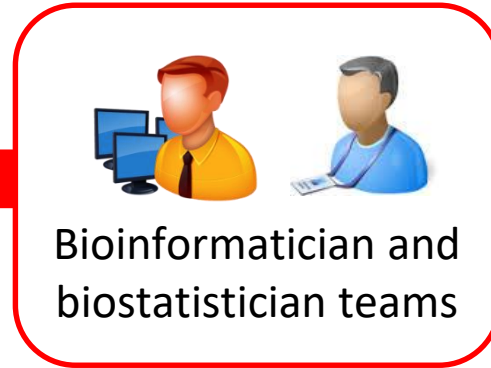
DNA sequencing prices are in the range on many current conventional tests.



**Global Alliance**  
for Genomics & Health  
Collaborate. Innovate. Accelerate.

By 2023 about 80% of omics data will be generated in the context of healthcare (not research)

# Genomic data in health care and solutions for its management



- No scalability
- Expensive
- Long response times
- Lack of experts
- Inequity



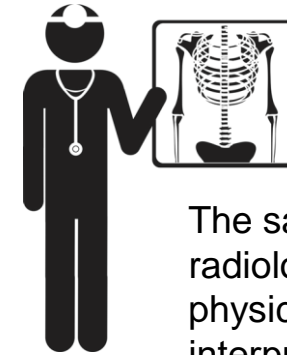
- Expensive (pay per use)
- **GDPR non compliant**
- Inequity

## Our solution: corporative software

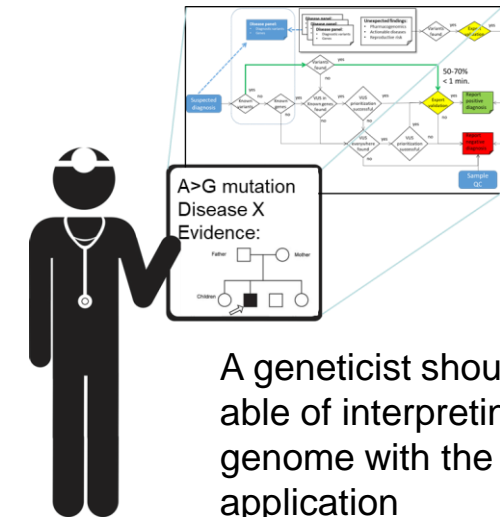
- **Equity**
- **Scalability**
- **Affordable**
- **End user: geneticist (clinician)**
- **GDPR compliant**

# Use of genomic data in the public health system requires sustainability

- User: **clinician** (not bioinformatician)
  - **Routine** genomic analysis (diagnosis, treatment recommendation) must be done with tools for end users, which involves **hiding the complexity** of the analysis to the clinician (geneticist)
  - A solution for the management of genomic data must be **integrated** the same way other analyses of the health system are.
- GDPR compliance and data reusability
  - **Genomic** data must be **stored** in the system, **linked** to **clinical** data the same way that other data are for further potential **prospective clinical studies**



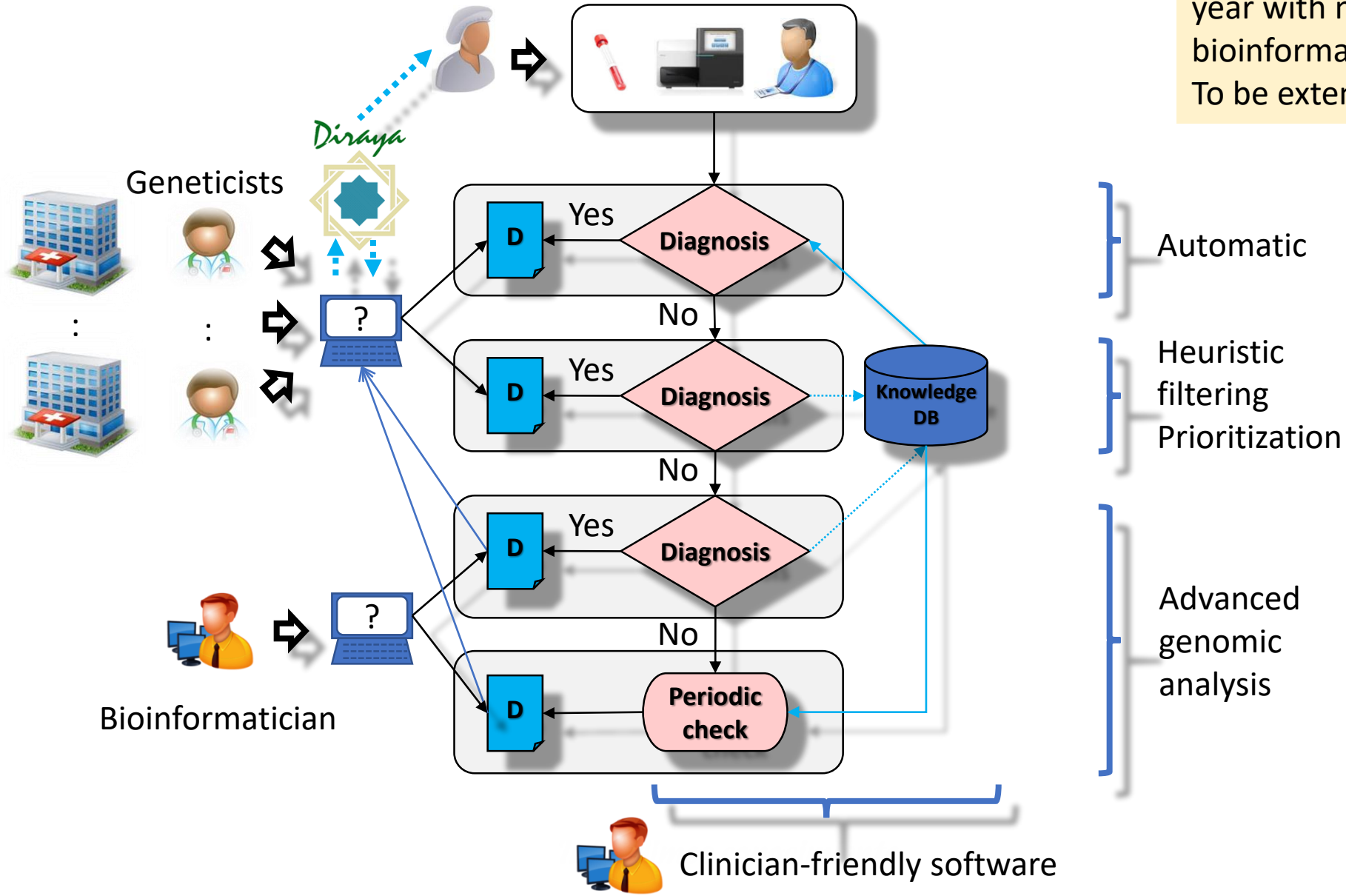
The same way that a radiologist do not need a physicist or an engineer to interpret a radiography...



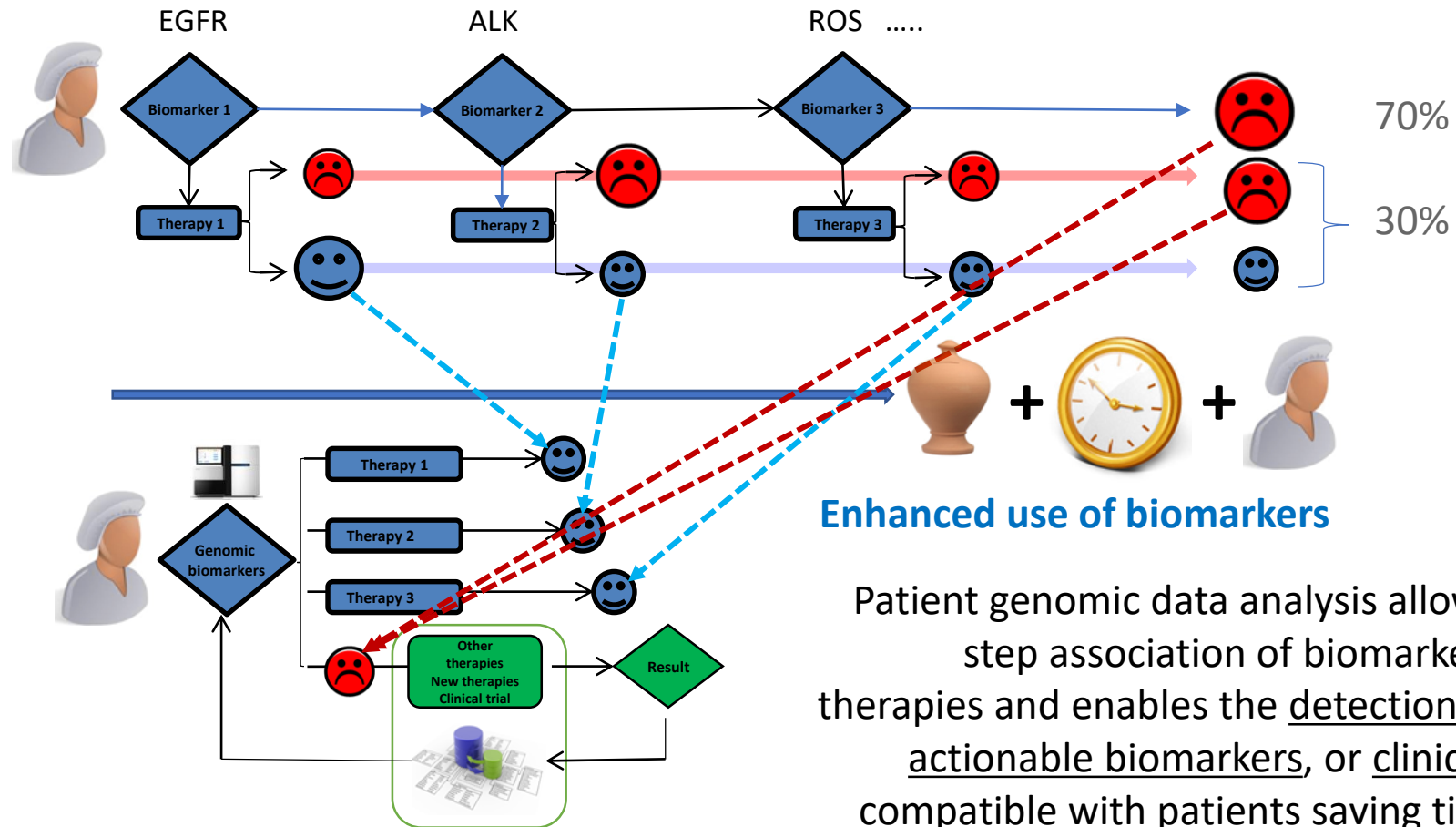
A geneticist should be able of interpreting a genome with the proper application

# The four levels of diagnosis

Hospital Virgen del Rocío:  
Diagnosis circuit for Rare  
Diseases. > 1000 diagnosis per  
year with no intervention of  
bioinformatics.  
To be extended to the SAS



# Beyond rare diseases diagnosis: Personalized Medicine in cancer



Cheaper than the conventional clinical circuit!!

# First wave: SARS-CoV-2 genomic sequencing initiatives in Spain and Andalusia.

The pandemics as an opportunity to build and test a clinical circuit of genome sequencing.

SARS-CoV-2 sequences are:

- Small and manageable
- Not sensitive data

Previous efforts in Spain: *SeqCOVID*  
<http://seqcovid.csic.es/es/>



Previous efforts in Andalusia: *covseq* project  
<http://clinbioinfospa.es/projects/covseq>

The screenshot shows the covseq project website with the following content:

**Sequencing of the SARS-CoV-2 virus genome for the monitoring and management of the Covid-19 epidemic in Andalusia and the rapid generation of prognostic and response to treatment biomarkers**

The differences in the spread, severity of symptoms, mortality and other characteristics of **COVID-19** are due to a combination of epidemiological factors, although it is expected that the genetics of the virus will play a very important role, which is still completely unknown.

A consortium, made up of 14 hospitals and 5 Health Research Institutes from the region, the **Technical Office for Information Management**, the **Regional Management of Public Health (Epidemiological Surveillance and Occupational Health Service)** and the **Fundacion Progreso y Salud**, intends to use the network of diagnostic centers of the Andalusian community and the sequencing platforms of the Health Research Institutes and associated centers, to sequence about 1,000 samples of the **SARS-CoV-2** virus covering the whole of Andalusia obtained from different type of patients (age, sex, previous complications, previous treatments), as well as different clinical profiles observed and responses to treatments.

# Beyond human genomes: A clinical circuit of SARS-CoV-2 WGS

Implementation: **WHO recommendation**, **Ponencia de Alertas y Planes de Preparación y Respuesta** y por la Comisión de Salud Pública del Consejo Interterritorial, **Instrucción** Consejería de Salud y Familias.

## SARS-CoV-2 genomic sequencing for public health goals

Interim guidance  
8 January 2021



### Key messages:

- Global surveillance of SARS-CoV-2 genetic sequences and related metadata contributes to the COVID-19 outbreak response. This contribution includes tracking the spread of SARS-CoV-2 geographically over time and ensuring that mutations that could potentially influence pathogenicity, transmission or countermeasures (such as vaccines, therapeutics and diagnostics) are detected and assessed in a timely manner.
- While the cost and complexity of genetic sequencing have dropped significantly over time, effective sequencing programmes still require substantial investment in terms of staff, equipment, reagents and bioinformatic infrastructure. Additionally, effective collaboration is needed to ensure that generated data are of good quality and are used in a meaningful way.
- Countries are encouraged to rapidly deposit SARS-CoV-2 sequences in a public database in order to share them with the scientific community for public health purposes. Investments in a tiered global sequencing network for SARS-CoV-2 will contribute to the development of resilient, high-quality global sequencing programmes for the detection and management of other outbreak pathogens in the future.

Rapid implementation  
because of the  
experience of the  
CovSeq project

GOBIERNO DE ESPAÑA MINISTERIO DE SANIDAD  
SECRETARÍA DE ESTADO DE SANIDAD, DIRECCIÓN GENERAL DE SALUD PÚBLICA  
Centro de Coordinación de Alertas y Emergencias Sanitarias Instituto de Salud Carlos III

### INTEGRACIÓN DE LA SECUENCIACIÓN GENÓMICA EN LA VIGILANCIA DEL SARS-CoV-2

22 de enero de 2021  
Actualización 12 de febrero de 2021

Junta de Andalucía Consejería de Salud y Familias  
CONSEJERÍA DE SALUD Y FAMILIAS  
Dirección General de Salud Pública y Ordenación Farmacéutica  
Dirección General de Asistencia Sanitaria y Resultados en Salud

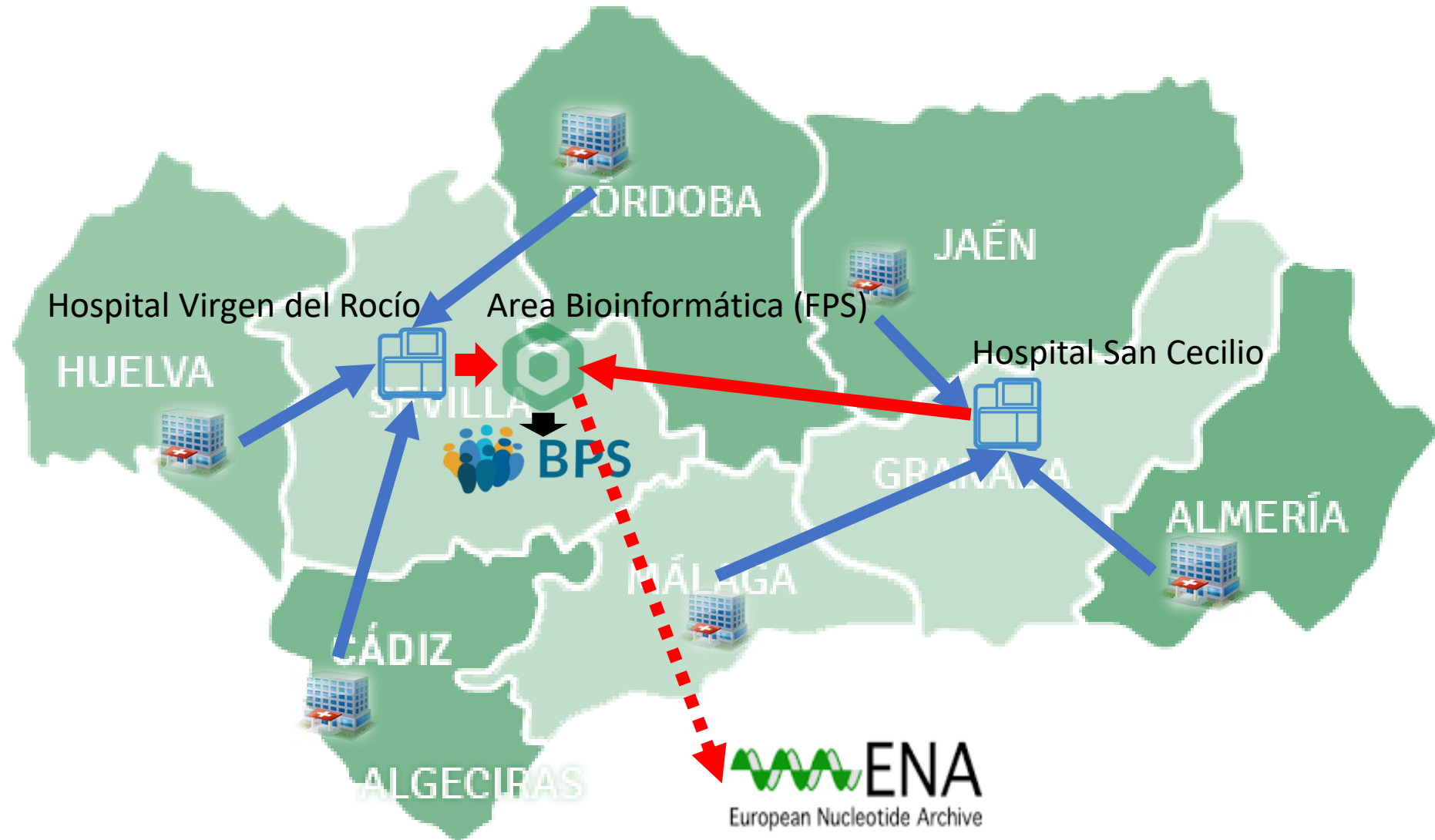
### Instrucción para la integración de la secuenciación genómica en la Vigilancia del SARS-CoV-2 en Andalucía

#### 1. INTRODUCCIÓN

La secuenciación de genomas completos o parciales de SARS-CoV-2 constituye el procedimiento más adecuado para complementar al actual sistema de vigilancia del SARS-CoV-2 de forma que permita detectar la aparición de nuevas variantes del virus. La aparición de variantes que aumenten la transmisibilidad de este virus, su virulencia o que escapen a la acción de los anticuerpos neutralizantes generados tras la infección natural o la vacuna, constituyen un problema de salud pública de primer orden que puede repercutir de forma importante en el control de la pandemia.

Es por ello, que en el marco de la Ponencia de Alertas y Planes de Preparación y Respuesta y la Comisión de Salud Pública del Consejo Interterritorial se ha aprobado el documento de Integración de la secuenciación genómica en la Vigilancia del SARS-CoV-2 que introduce las indicaciones y objetivos para tal fin (disponible en: [https://www.msbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Integracion\\_de\\_la\\_secuenciacion\\_genomica-en\\_la\\_vigilancia\\_del\\_SARS-CoV-2.pdf](https://www.msbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Integracion_de_la_secuenciacion_genomica-en_la_vigilancia_del_SARS-CoV-2.pdf)).

# Beyond human genomes: A clinical circuit of SARS-CoV-2 WGS

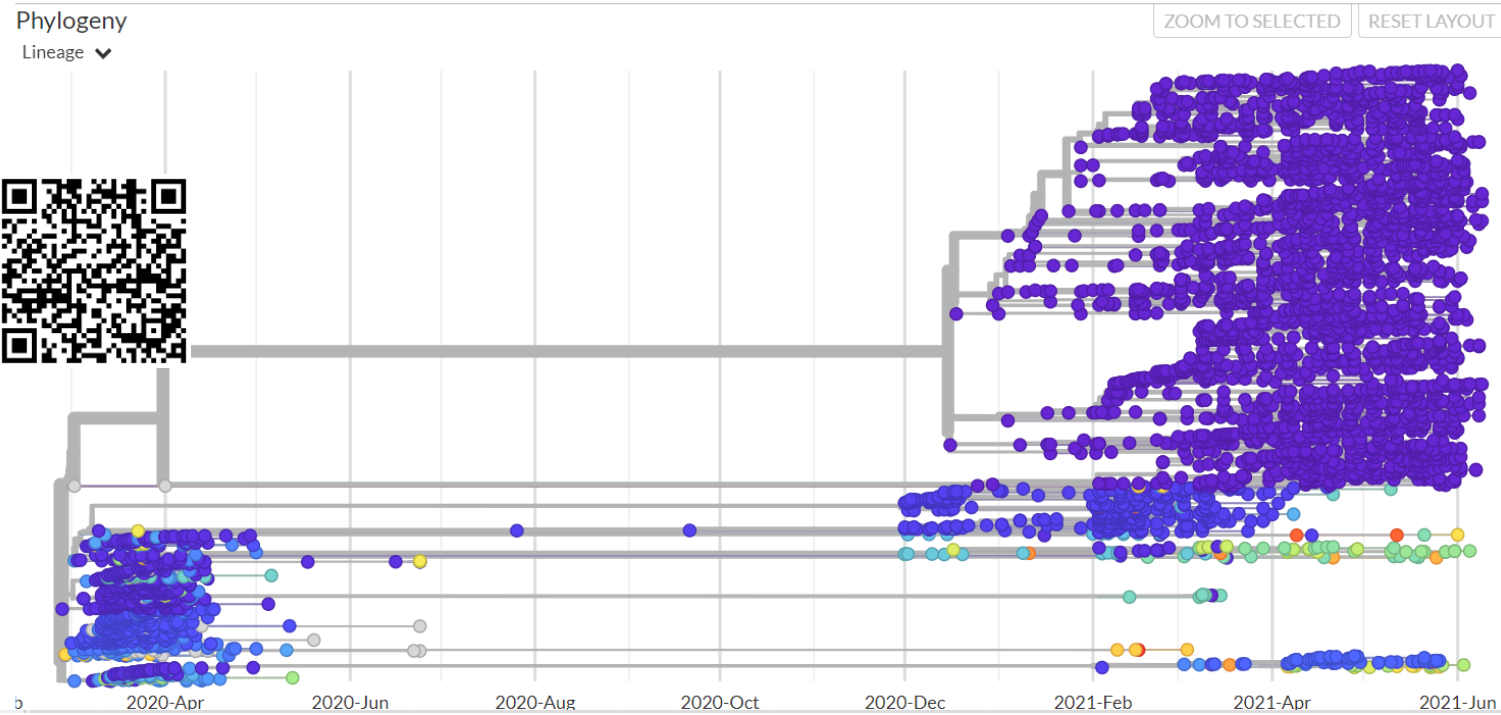


# Results so far

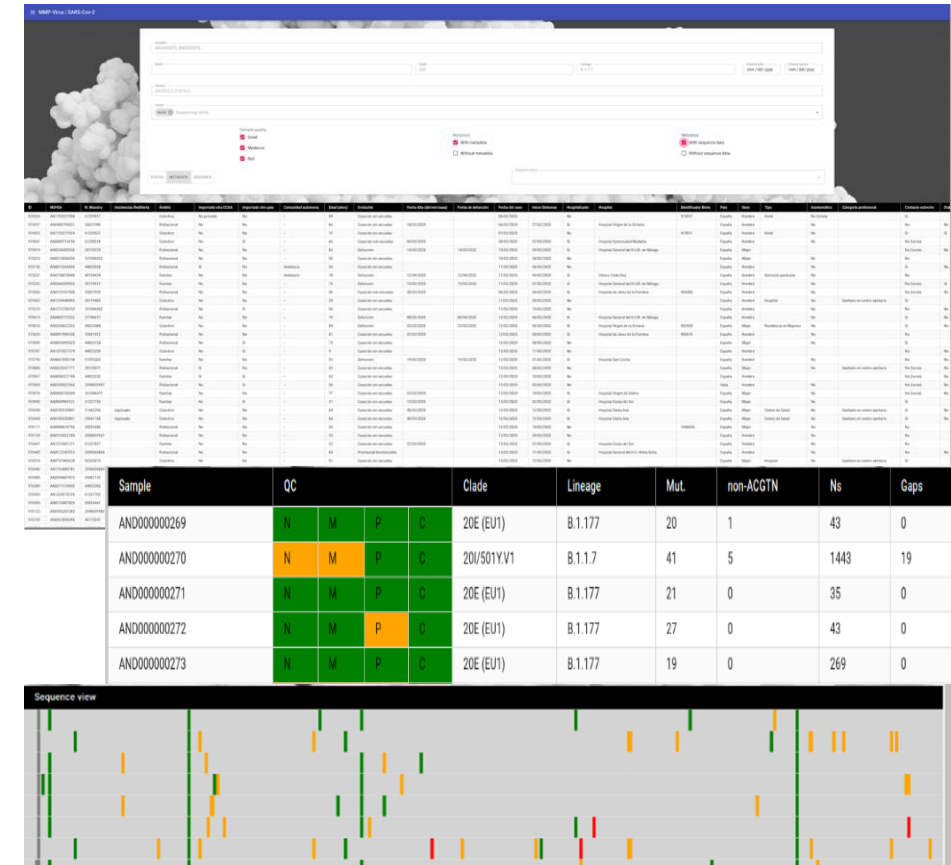
## SARS-COV-2 outbreak in Andalucia

Maintained by [Clinical Bioinformatics Area](#).

Showing 5051 of 5051 genomes sampled between Feb 2020 and Jun 2021.



More than 5000 SARS-CoV-2 whole genomes sequenced (28/6/2021)



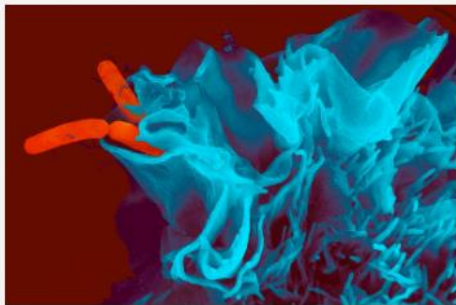
New tools: MMP-virus.  
Keeps track of all the variants in Andalucia, accessible from all the hospitals

# SIEGA: Sistema Integrado de Epidemiología Genómica de Andalucía

## MAPAS EPIDEMIOLÓGICOS

Mapa epidemiológico de los aislados de *Salmonella enterica*, *Listeria monocytogenes* y *Campylobacter jejuni* de la comunidad andaluza.

*Salmonella enterica*



*Listeria monocytogenes*



*Campylobacter jejuni*



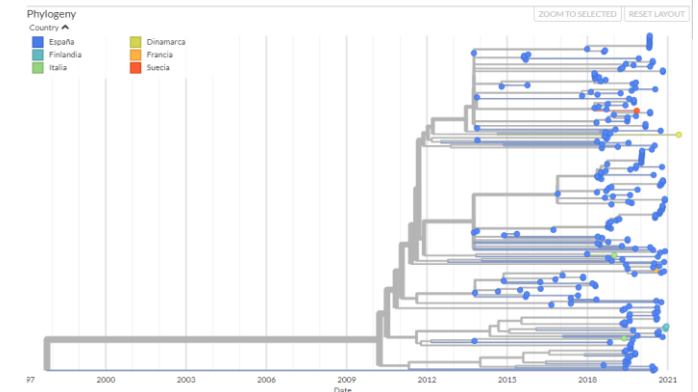
Integral One-Health control of environmental pathogens that can be transmitted to humans.

Currently includes Salmonella, Listeria and Campylobacter. Recently West Nile Virus has been added



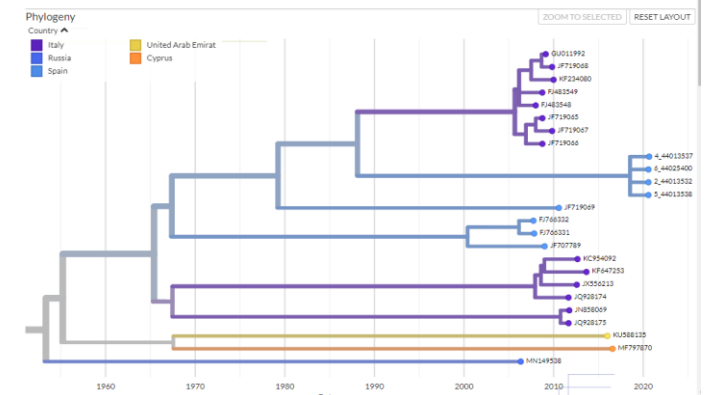
### Salmonella enterica in Andalucía

Maintained by Clinical Bioinformatics Area.  
Showing 251 of 251 genomes sampled between Oct 2013 and Jun 2021.



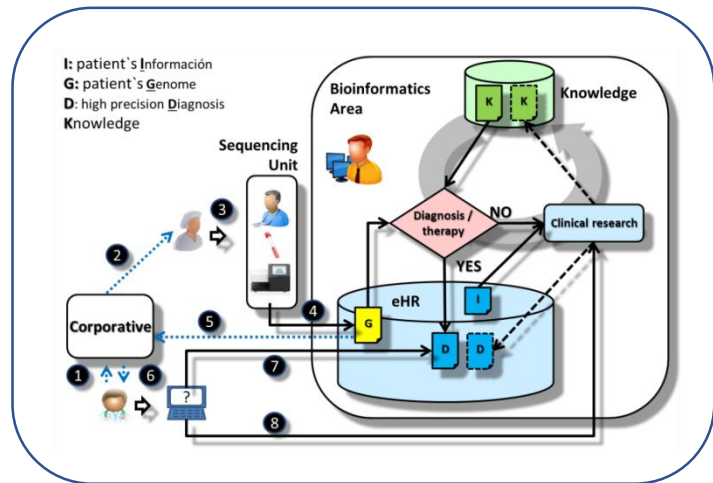
### WNV outbreak

Maintained by Clinical Bioinformatics Area.  
Showing 25 of 153 genomes sampled between May 2006 and Aug 2020.



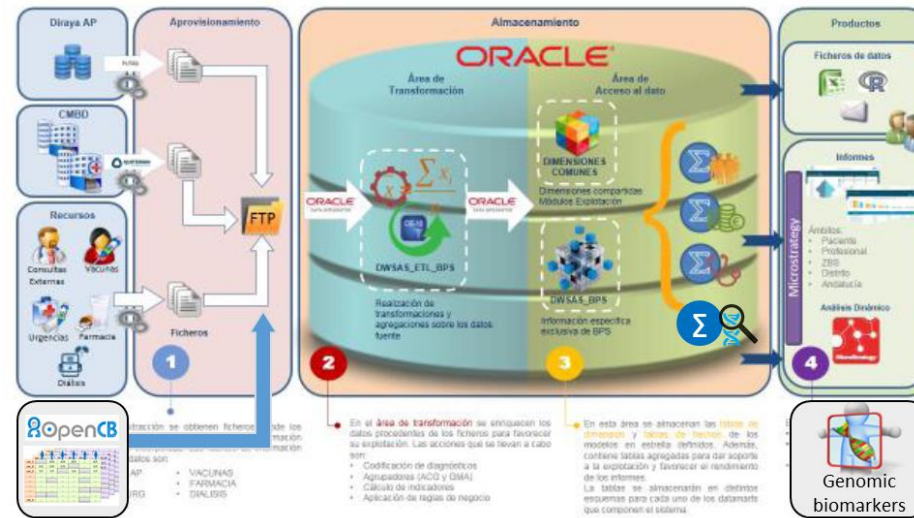
# The two tiers of personalized medicine:

## Primary and secondary use of genomic data to generate new knowledge



**Primary use:** use of patient genomic data for **precision diagnosis** (typically RDs) and **treatment recommendation** (typically cancer).

Widely implemented



**Secondary use:** use **clinical data (eHR)** along with **structured genomic data** for **preventive medicine** and **knowledge generation** (e.g. biomarker discovery, etc.)  
Andalusian **Population Health Database**, with over 13M people since 2001.

**Aim:** facilitating RWD / RWE studies within the Andalusian Public Health System (SSPA)



Poorly implemented

# GDPR and conventional clinical data management for secondary data analysis

“Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales” GDPR compliant, replaces the old LOPD since 25 May 2018.  
(<https://www.boe.es/eli/es/lo/2018/12/05/3>)



Processing of “data concerning health,” “genetic data,” and “biometric data” is prohibited unless one of several conditions applies:

- data subject gives “explicit consent” (conventional for **prospective** studies)
- or processing is necessary for reasons of public interest in the area of public health (covers **retrospective** studies)

# The Population Health Database (BPS)

A comprehensive repository with all the clinical information (EHRs) of more than 13 million patients of the Andalusian Public Health System.

BPS is used for secondary data analysis



The screenshot shows the website for the Base Poblacional de Salud (BPS) within the Servicio Andaluz de Salud. The header includes the organization's logo and name, and the 'e\_atención al profesional' logo. Navigation tabs are present for 'El SAS', 'Ciudadanía', 'Profesionales', and 'Proveedores'. A search bar is located below the navigation. The breadcrumb trail indicates the current page: 'Inicio > Profesionales > Sistemas de información > Base poblacional de salud'. The main heading is 'Base poblacional de salud'. A descriptive paragraph states: 'La Base Poblacional de Salud (BPS) es un sistema de información sanitaria que recoge datos clínicos y del uso de recursos sanitarios de cada una de las personas que reciben asistencia sanitaria en el Servicio Andaluz de Salud.' To the right of this text is an icon of a group of people. At the bottom, a paragraph explains that data from the BPS is used to estimate health status, user behavior, and stratify the population. A 'Normativa' link is also visible in the bottom right corner.

# The paradox: Data + knowledge and computing + expertise are separated

Relevant question



**Health system**



**Firewall**

Relevant analysis

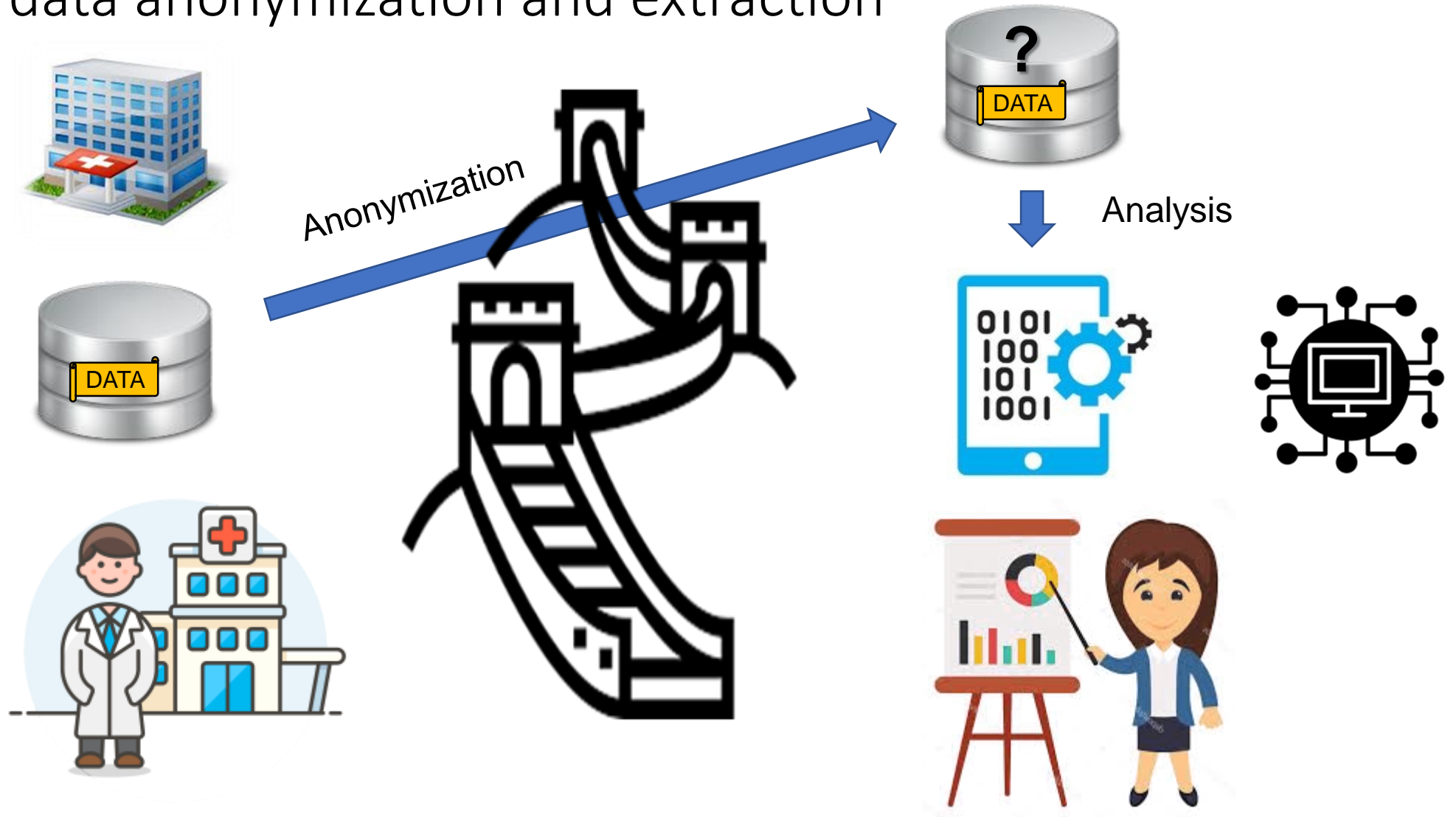


**World**



# Problems with secondary analysis of clinical Big Data

Solution: data anonymization and extraction




# Risk: patient re-identification.

## Many examples show that anonymized data can be re-identified

### 'Anonymous' browsing data can be easily exposed, researchers reveal

A journalist and a data scientist secured data from three million users easily by creating a fake marketing company, and were able to de-anonymise many users



▲ 'We wrote and called nearly a hundred companies, and asked if we could have the raw data, the clickstream from people's lives.' Photograph: Steve Marcus/Reuters

A judge's porn preferences and the medication used by a German MP were among the personal data uncovered by two German researchers who acquired the "anonymous" browsing habits of more than three million German citizens.

<https://www.theguardian.com/technology/2017/aug/01/data-browsing-habits-brokers>

### The disclosure of diagnosis codes can breach research participants' privacy

Grigorios Loukides, Joshua C Denny, Bradley Malin

Journal of the American Medical Informatics Association, Volume 17, Issue 3, May 2010, Pages 322-327, <https://doi.org/10.1136/jamia.2009.002725>  
Published: 01 May 2010 Article history

PDF Split View Cite Permissions Share

#### Abstract

**Objective** De-identified clinical data in standardized form (eg, diagnosis codes), derived from electronic medical records, are increasingly combined with research data (eg, DNA sequences) and disseminated to enable scientific investigations. This study examines whether released data can be linked with identified clinical records that are accessible via various resources to jeopardize patients' anonymity, and the ability of popular privacy protection methodologies to prevent such an attack.

**Design** The study experimentally evaluates the re-identification risk of a de-identified sample of Vanderbilt's patient records involved in a genome-wide association study. It also measures the level of protection from re-identification, and data utility, provided by suppression and generalization.

**Measurement** Privacy protection is quantified using the probability of re-identifying a patient in a larger population through diagnosis codes. Data utility is measured at a dataset level, using the percentage of retained information, as well as its description, and at a patient level, using two metrics based on the difference between the distribution of Internal Classification of Disease (ICD) version 9 codes before and after applying privacy protection.

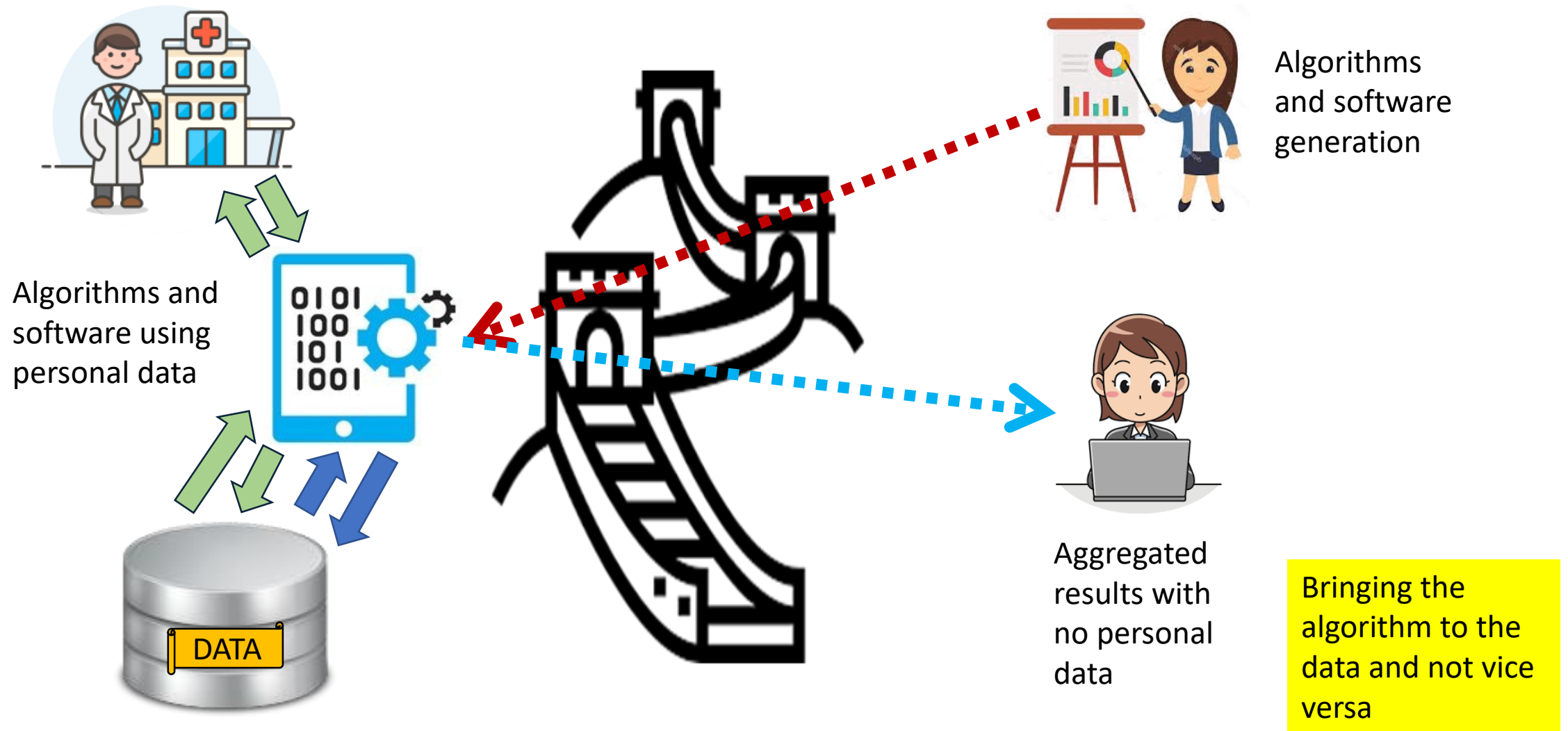
**Results** More than 96% of 2800 patients' records are shown to be uniquely identified by their diagnosis codes with respect to a population of 1.2 million patients. Generalization is shown to reduce further the percentage of de-identified records by less than 2%, and over 99% of the three-digit ICD-9 codes need to be suppressed to prevent re-identification.

**Conclusions** Popular privacy protection methods are inadequate to deliver a sufficiently protected and useful result when sharing data derived from complex clinical systems. The development of alternative privacy protection models is thus required.

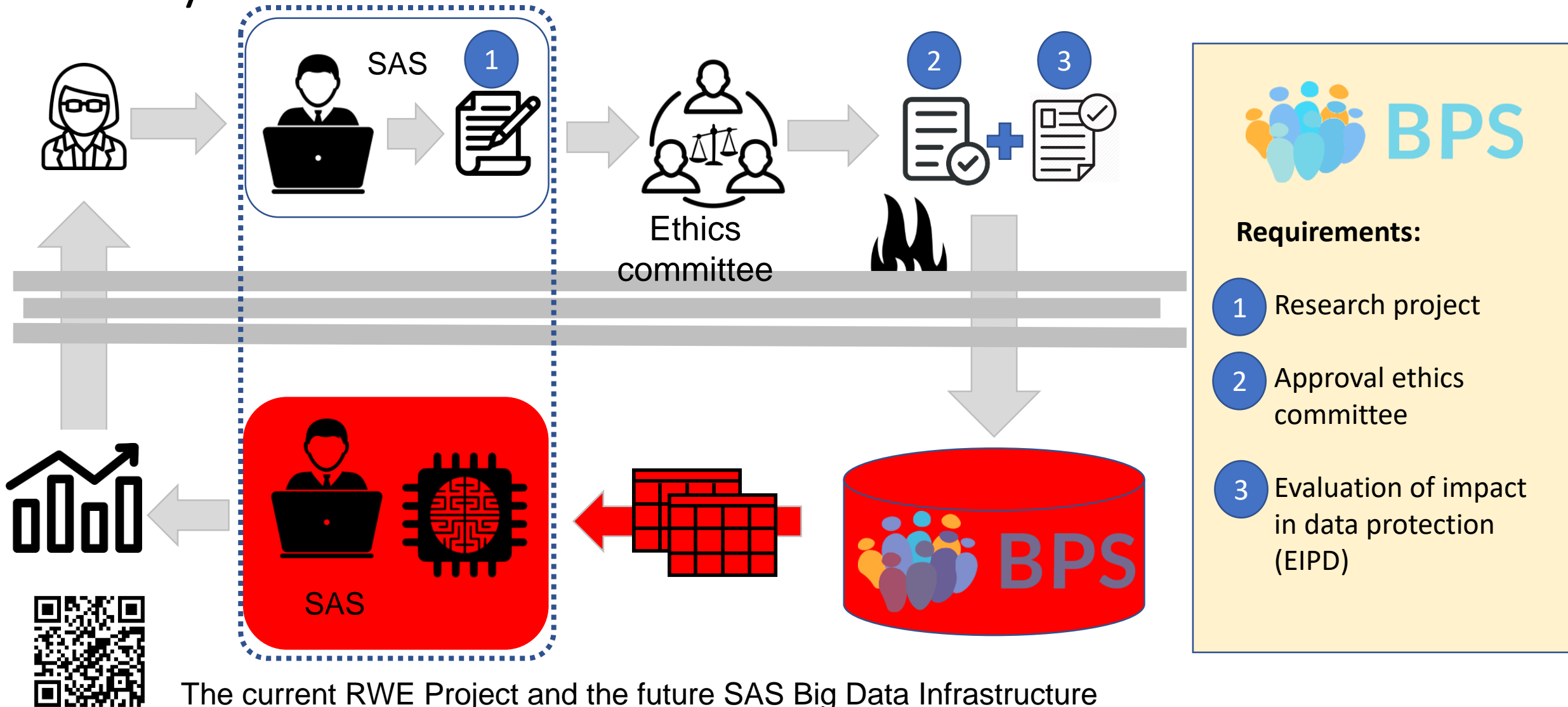
Studies have shown that de-identified hospital discharge data could be re-identified using basic demographic attributes.

# Problems with secondary analysis of clinical Big Data

Solution: Bring the algorithm to the data

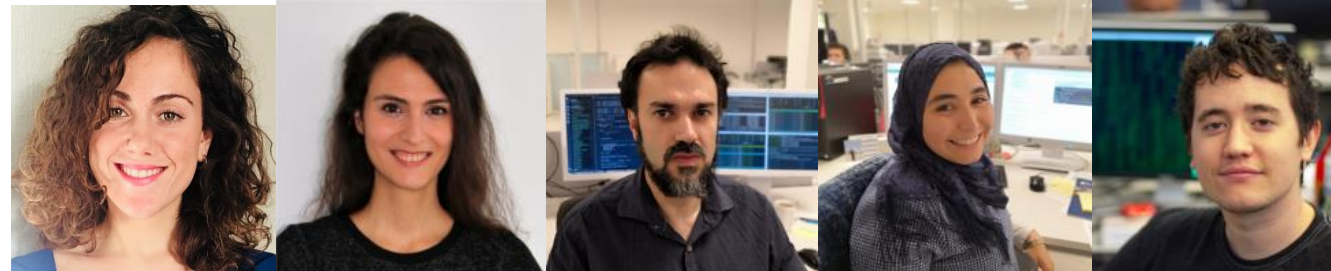


# Infrastructures for secure secondary data analysis



The current RWE Project and the future SAS Big Data Infrastructure

# Systematic drug repurposing in SARS-CoV-2 with machine learning and mechanistic models

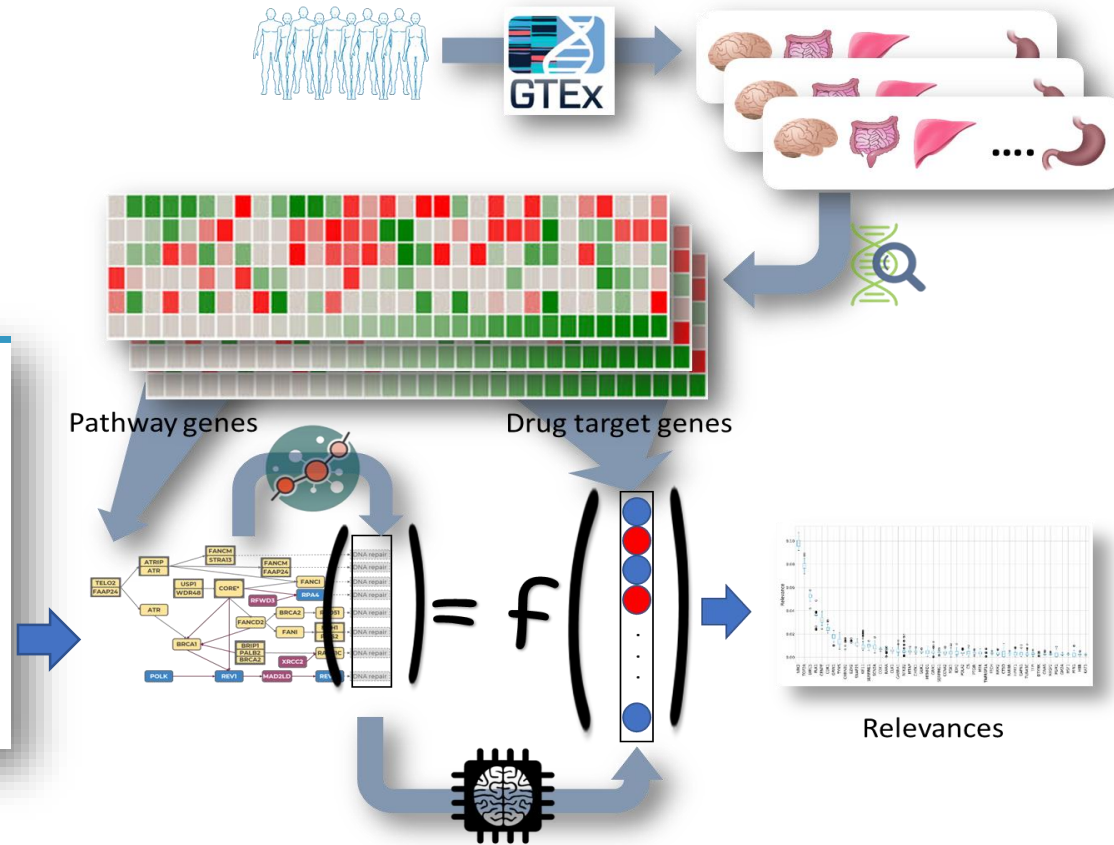


Funded by the ISCIII

Marina Esteban-Medina  
 María Peña-Chilet  
 Carlos Loucera  
 Kinza Rian  
 Matias M Falco  
 Joaquin Dopazo

REFERENCIA	TITULO	ENTIDAD SOLICITANTE	INVESTIGADOR/RES PRINCIPAL/LES	DESCRIPCION/RESUMEN	FECHA RESOLUCIÓN CONCESION
COV20_0070	Ensayo clínico piloto, abierto, randomizado de uso combinado de Hidroxicloroquina, Azitromicina y Tocilizumab para el tratamiento de la infección por SARS-CoV-2 (COVID-19)	Fundación Privada Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau	Pere Domingo Pedrol	El objetivo principal del estudio es evaluar la mortalidad intrahospitalaria, necesidad de ventilación mecánica, o necesidad de dosis de rescate de tocilizumab en los pacientes con infección confirmada por COVID-19 en tratamiento con hidroxicloroquina y azitromicina combinado con tocilizumab. El estudio comparará 2 ramas, terapia habitual en la práctica clínica, rama control; vs. tratamiento con tocilizumab, rama experimental.	30/03/2020
COV20_0072	Ensayo clínico aleatorizado multicéntrico de terapia con plasma de convalecientes añadido al mejor tratamiento disponible para COVID-19 en pacientes hospitalizados	FUNDACION PARA LA INVESTIGACION BIOMEDICA DEL HOSPITAL UNIVERSITARIO PUERTA DE HIERRO	Cristina Avendaño Solá y Rafael Duarte Palomino	El plasma hiperinmune de convalecientes (PC) se usa habitualmente en infecciones respiratorias graves de causa vírica, en situación de urgencia y en base a estudios no comparativos de baja calidad. Ante la pandemia por SARS-CoV-2, proponemos realizar un ensayo randomizado en sujetos hospitalizados con formas no críticas de COVID-19 en más de 20 hospitales. ISCIII, que nos permitirá seguridad del tratamiento anticuerpos y respuesta de carga viral y la seroconversión.	
COV20_00788	Modelo mecanístico basado en inteligencia artificial para la reutilización de fármacos contra la infección por SARS-CoV-2	Fundación Pública Andalucía Progreso y Salud	Joaquín Dopazo Blázquez	La rápida expansión del el SARS-CoV-2, un virus nuevo, ha cogido a los sistemas sanitarios sin terapias adecuadas. Aunque el desarrollo de un nuevo fármaco es inviable a corto plazo, la solución de reutilizar fármacos con otras indicaciones acercaría la disponibilidad de nuevas terapias. Las aproximaciones computacionales han demostrado ser más rápidas y baratas que las experimentales. Dentro de estas, las basadas en modelos mecanísticos, que permiten inferir relaciones causa-efecto, han demostrado mayor efectividad que soluciones tradicionales como cribados virtuales u otras. Proponemos el uso de un modelo mecanístico basado en inteligencia artificial que ya ha demostrado su efectividad en enfermedades raras para la reutilización de fármacos para Covid-19.	08/05/2020

# Machine learning and potential causal relationships



Signal Transduction and Targeted Therapy

Explore our content | Journal information | Publish with us

nature > signal transduction and targeted therapy > letters > article


Letter | Open Access | Published: 11 December 2020

### Drug repurposing for COVID-19 using machine learning and mechanistic models of signal transduction circuits related to SARS-CoV-2 infection

Carlos Loucera, Marina Esteban-Medina, Kinza Rian, Matías M. Falco, Joaquín Dopazo & María Peña-Chilet

Signal Transduction and Targeted Therapy 5, Article number: 290 (2020) | Cite this article

921 Accesses | 17 Altmetric | Metrics



www.nature.com/scientificdata

## SCIENTIFIC DATA

OPEN COMMENT

### COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms

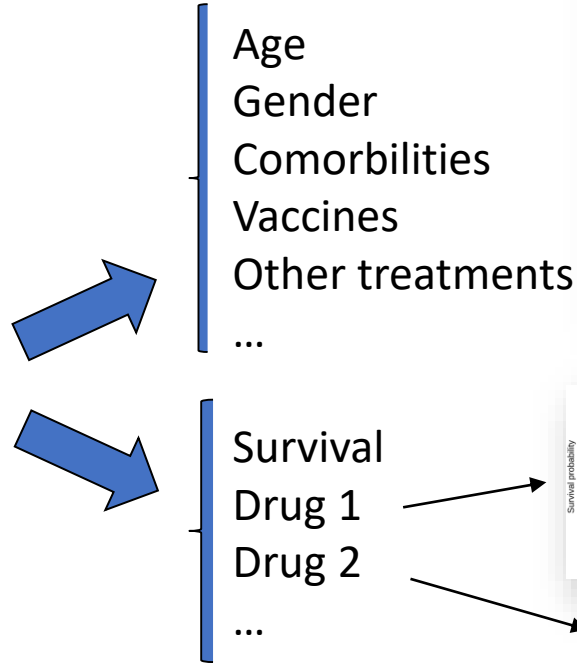
Marek Otsaszewski<sup>1</sup>, Alexander Mazein<sup>1,2</sup>, Marc E. Gillespie<sup>1,4</sup>, Inna Kuperstein<sup>5</sup>, Anna Niarakis<sup>6</sup>, Henning Hermjakob<sup>7</sup>, Alexander R. Pico<sup>8</sup>, Egon L. Willighagen<sup>9</sup>, Chris T. Evelo<sup>10</sup>, Jan Hasenauer<sup>11,12,13</sup>, Falk Schreiber<sup>14,15</sup>, Andreas Dräger<sup>16,17,18</sup>, Emek Demir<sup>19</sup>, Olaf Wolkenhauer<sup>20</sup>, Laura I. Furlong<sup>21</sup>, Emmanuel Barillot<sup>22</sup>, Joaquín Dopazo<sup>23,24,25</sup>, Aurelio Ortega-Rosendo<sup>26,27</sup>, Francesco Messina<sup>28,29</sup>, Alfonso Valencia<sup>30,31</sup>, Akira Funahashi<sup>32</sup>, Hiroaki Kitano<sup>33,35,36</sup>, Charles Auffray<sup>37</sup>, Rudi Balling<sup>38</sup> & Reinhard Schneider<sup>1,35</sup>

Researchers around the world join forces to reconstruct the molecular processes of the virus-host interactions aiming to combat the cause of the ongoing pandemic.

Machine learning is used to detect potential causal relationships between known drug targets and the activity of COVID-19 hallmarks

# Systematic RWE validation of the protective potential of drugs in COVID-19

Real world evidence of the use of drugs prescribed for other reasons and mortality rate of COVID-19 hospitalized in a large retrospective cohort of 15,968 Andalusian patients hospitalized between February and November 2020



medRxiv THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

HOME | ABOUT | Search

Real world evidence of calcifediol use and mortality rate of COVID-19 hospitalized in a large cohort of 16,401 Andalusian patients

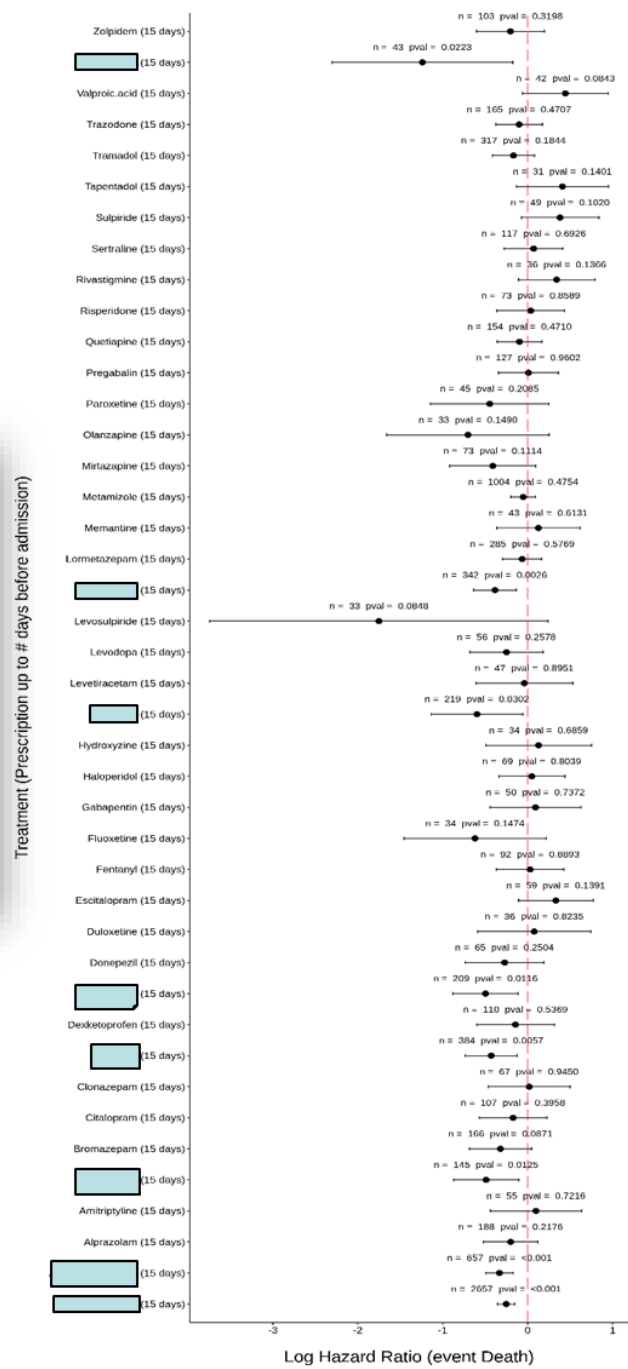
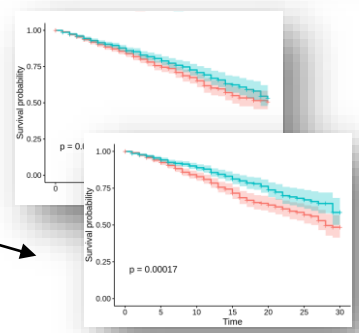
Comments (2)

Carlos Loucera, María Peña-Chilet, Marina Esteban-Medina, Dolores Muñozerro-Muñiz, Román Villegas, Jose Lopez-Miranda, Jesus Rodriguez-Baño, Isaac Túniz, Roger Bouillon, Joaquin Dopazo, Jose Manuel Quesada Gomez

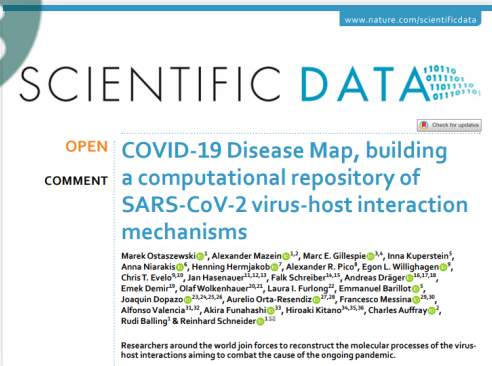
doi: <https://doi.org/10.1101/2021.04.27.21255937>

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.

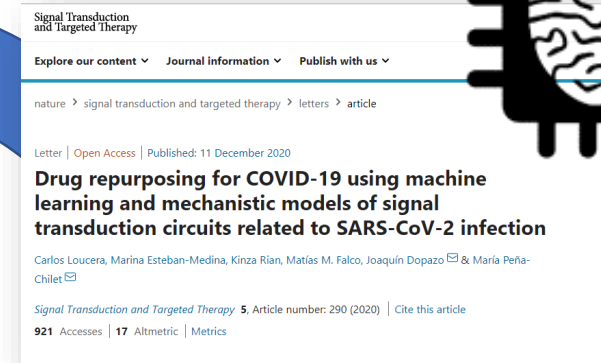
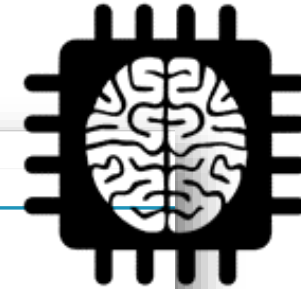
Abstract Full Text Info/History Metrics Preview PDF



# Whole cycle from the hypothesis to the validation: data science (big data + AI) with no experiments



1. Modeling Biological knowledge



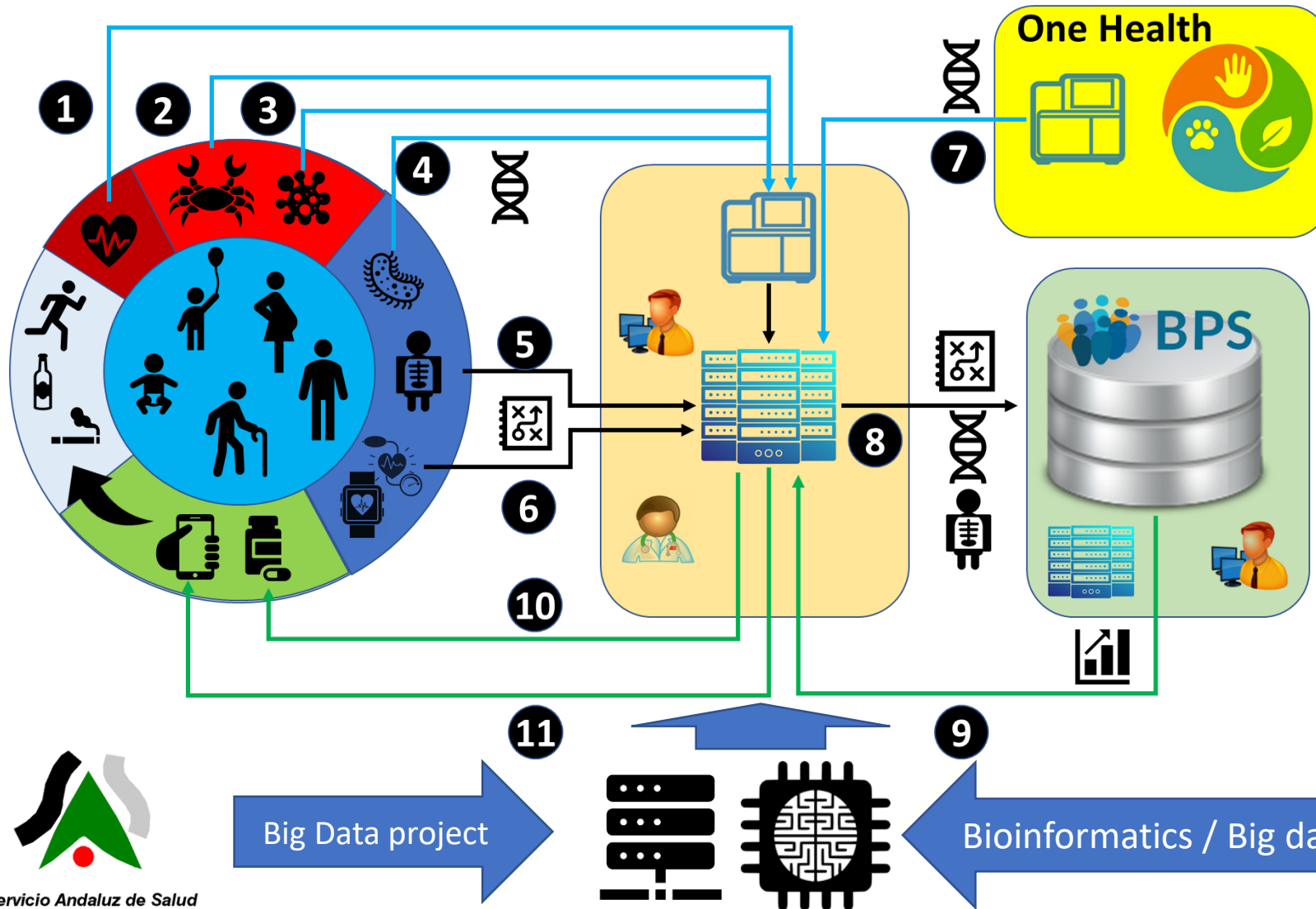
2. Learning from genomic big data

A new paradigm of knowledge generation



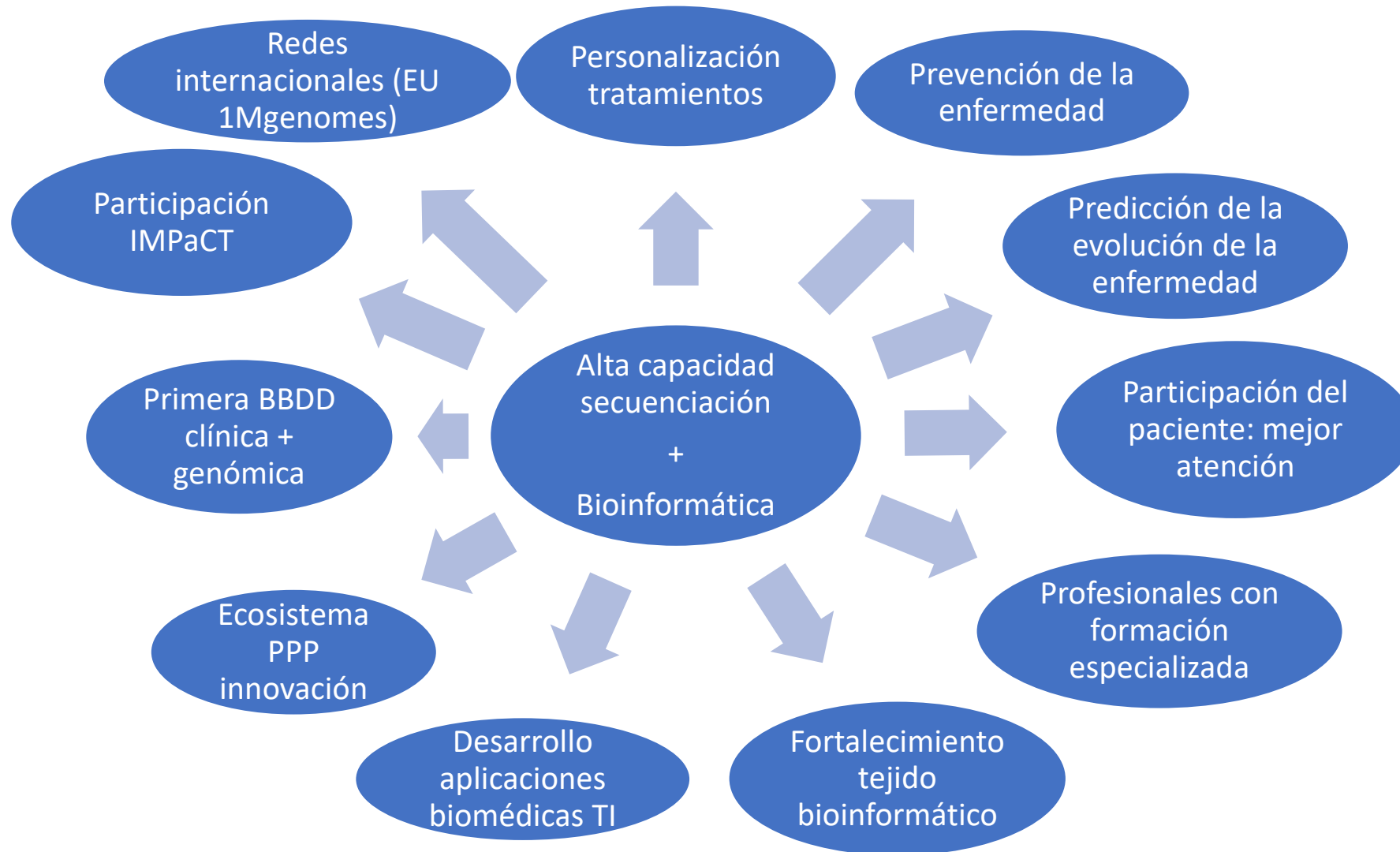
3. Real World Evidence Validation

# Integral data management plan



- 1 Hereditary diseases
- 2 Cancer
- 3 Infectious diseases
- 4 Microbiome
- 5 Medical image
- 6 Wearables
- 7 Environmental pathogens
- 8 Data for secondary use
- 9 Knowledge generated
- 10 Secondary findings
- 11 Life style recommendations

# Transformación del sistema de salud





# Clinical Bioinformatics Area, Fundación Progreso y Salud, Hospital Virgen del Rocío, Sevilla, Spain; Computational Systems Medicine, IBIS, Sevilla, Spain and ...the FPS/ELIXIR-ES and the BiER (CIBERER)



Follow us on twitter  
@xdopazo  
@ClinicalBioinfo



<https://www.slideshare.net/xdopazo/>

