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.clinbioinfosspa.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

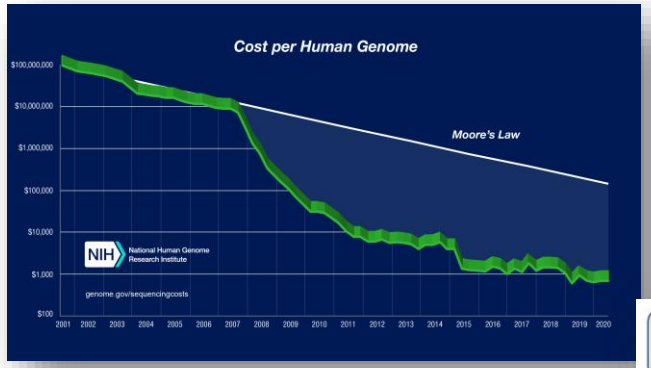
Author information ▸ Copyright and License information Disclaimer

This article has been cited by other articles in PMC.

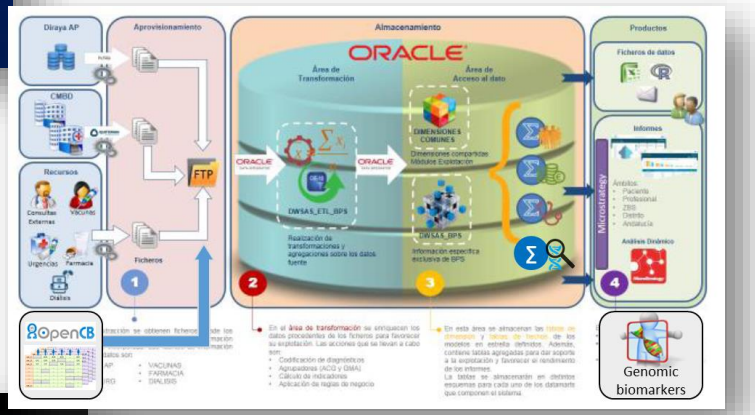
Abstract Go to:

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 confronti education. Key internal challenges include the focus on disease to the relati inpatient versus outpatient education, and implications of a faculty whose r molecular or submolecular level. External factors include the exponential g technologic ("disruptive") innovations, and societal changes. Addressing th institutional leadership with an eye to 2020 and beyond—the period in whi begin their careers. This paper presents a spiral-model format for a curricul on disease mechanisms, that addresses many of these challenges and incorp principles.

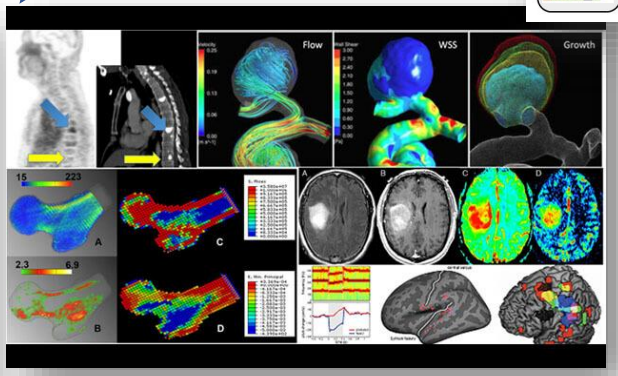


Whole genomes by < 600€



Clinical data:
The Population Health Database

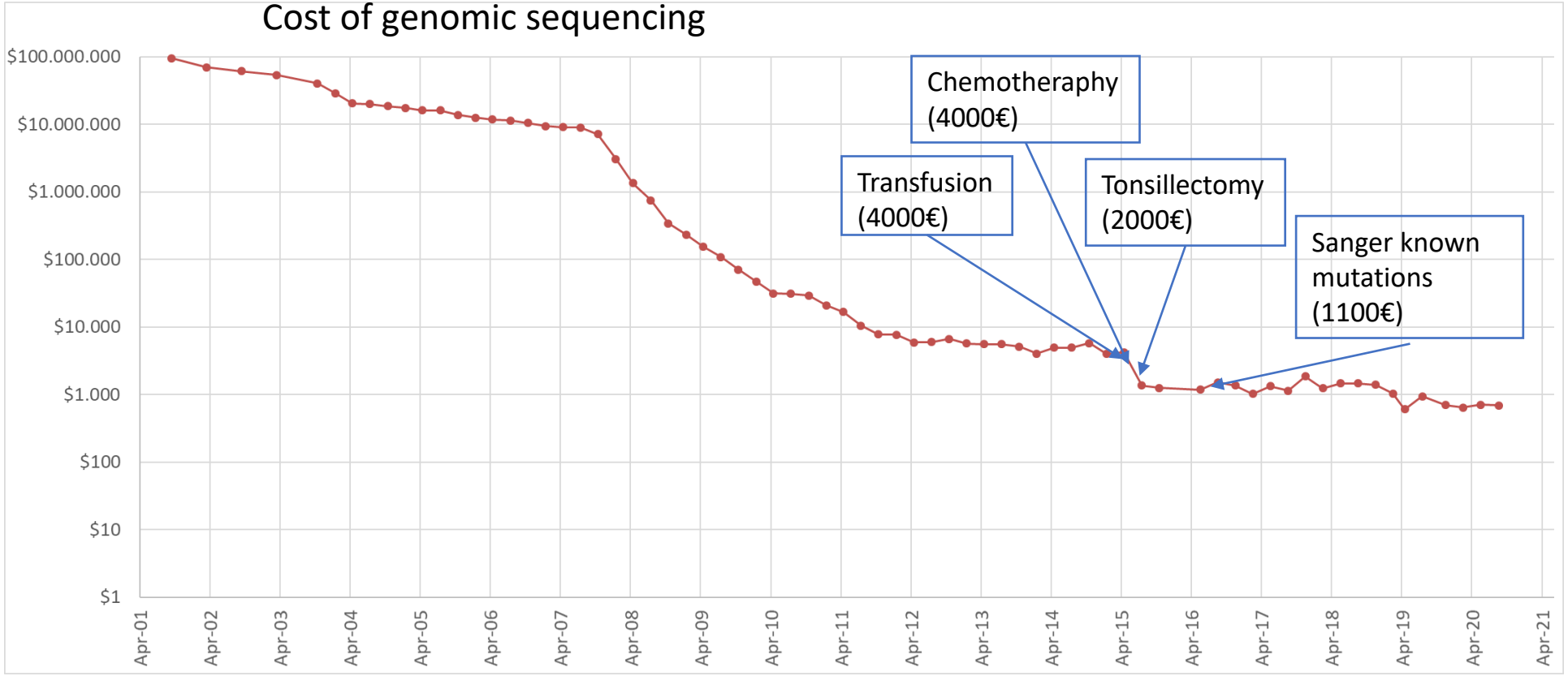
Even more data will come from portable medical devices



Medical image

TBs of medical image are produced on daily basis

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

Bioinformatician and biostatistician teams

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

Commercial software

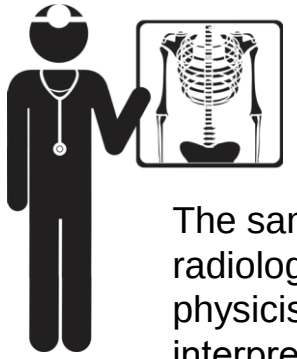
- Expensive (pay per use)
- **GDPR non compliant**
- Inequity

Our solution: corporative software

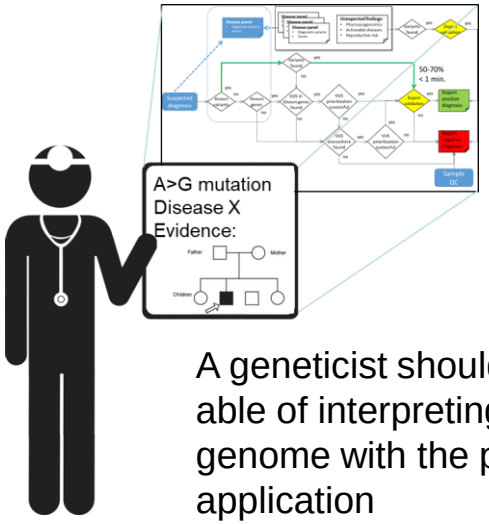
- **Equity**
- **Scalability**
- **Affordable**
- **End user: geneticist (clinician)**
- **GDPR compliant**

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**



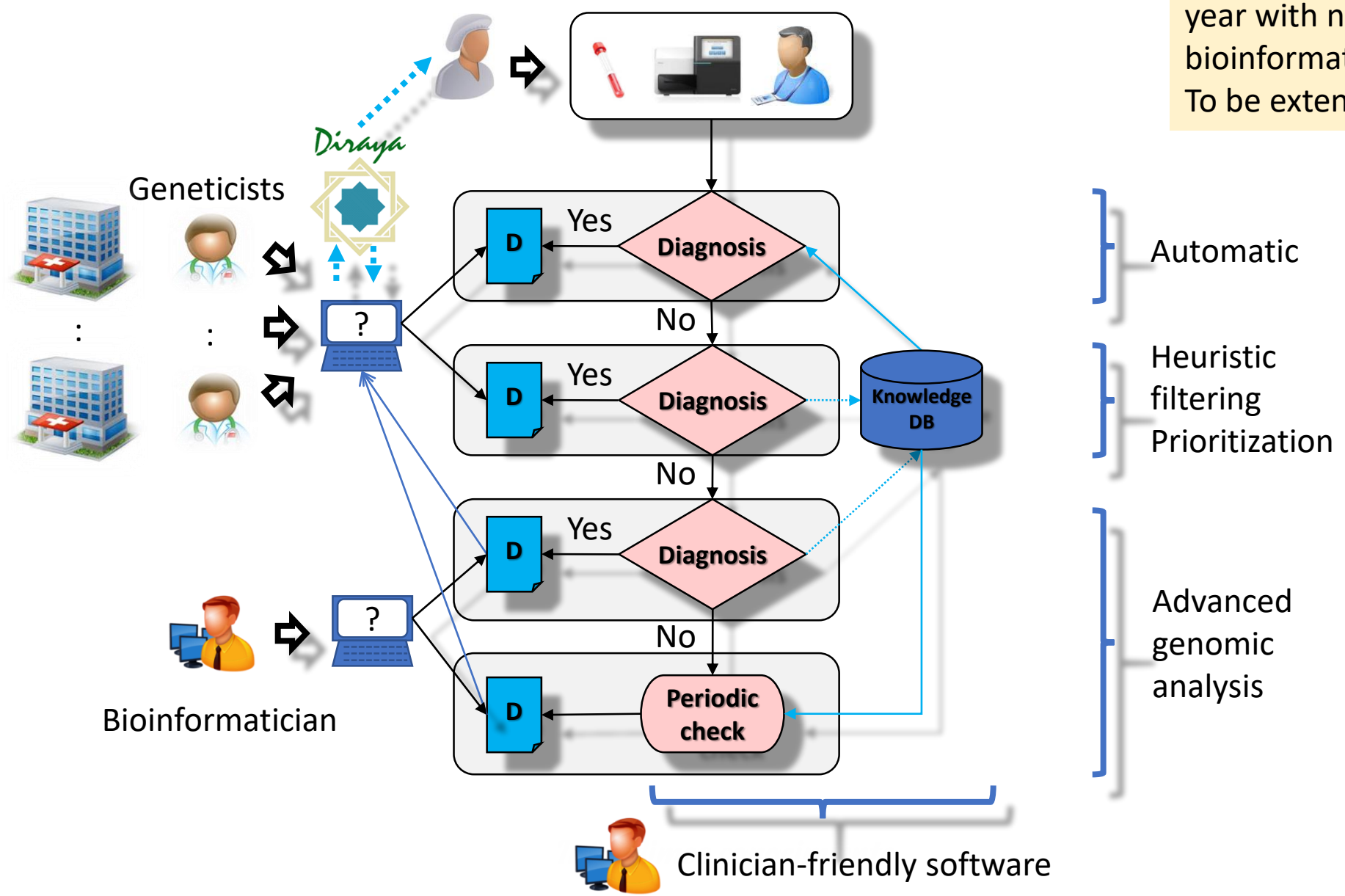
The same way that a radiologist do not need a physicist or an engineer to interpret a radiography…



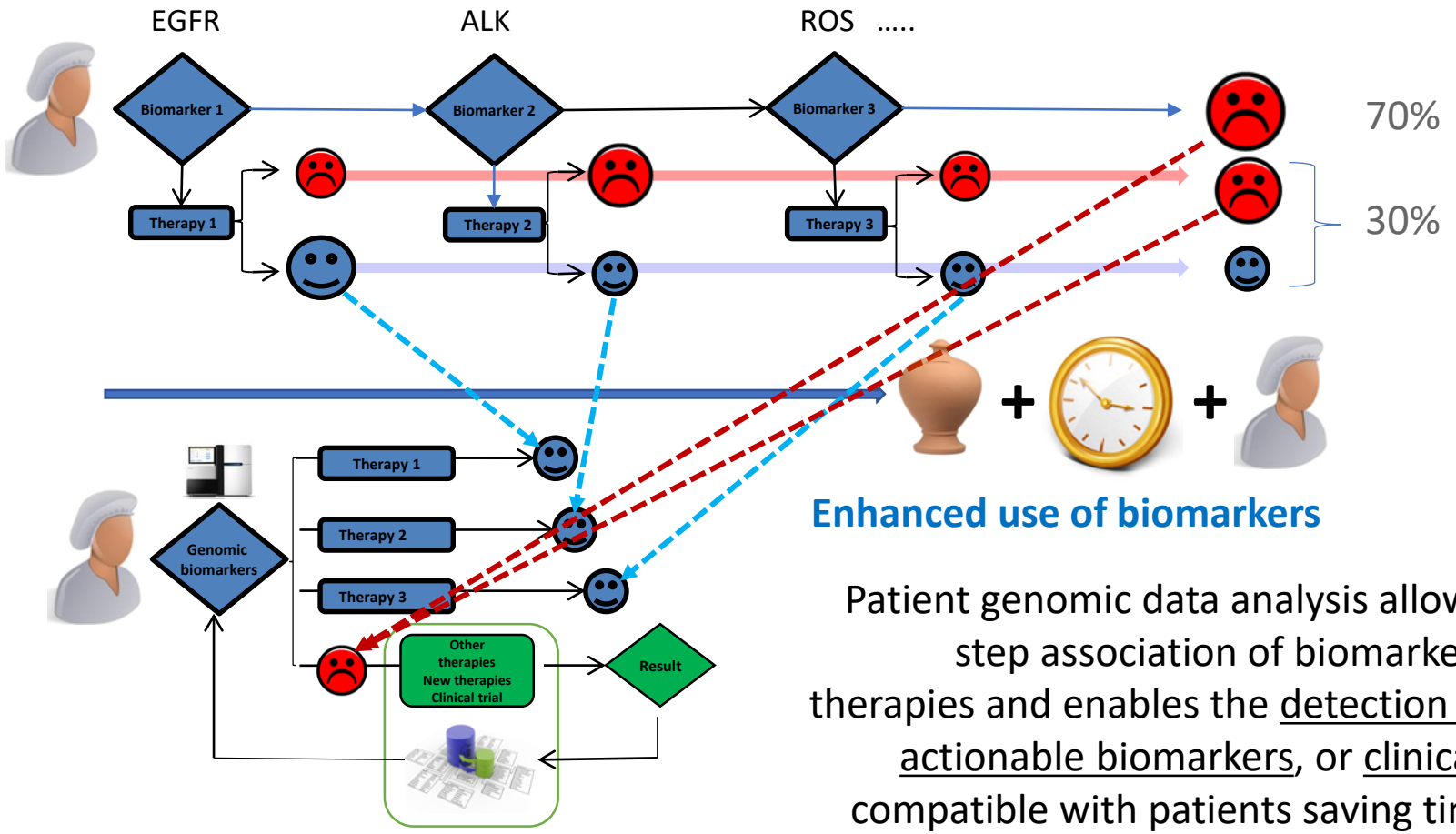
A geneticist should be able of interpreting a genome with the proper application

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



Enhanced use of biomarkers

Patient genomic data analysis allows one-step association of biomarkers with therapies and enables the detection of new actionable biomarkers, or clinical trials compatible with patients saving time and cost and increasing treatment success

Cheaper than the conventional clinical circuit!!

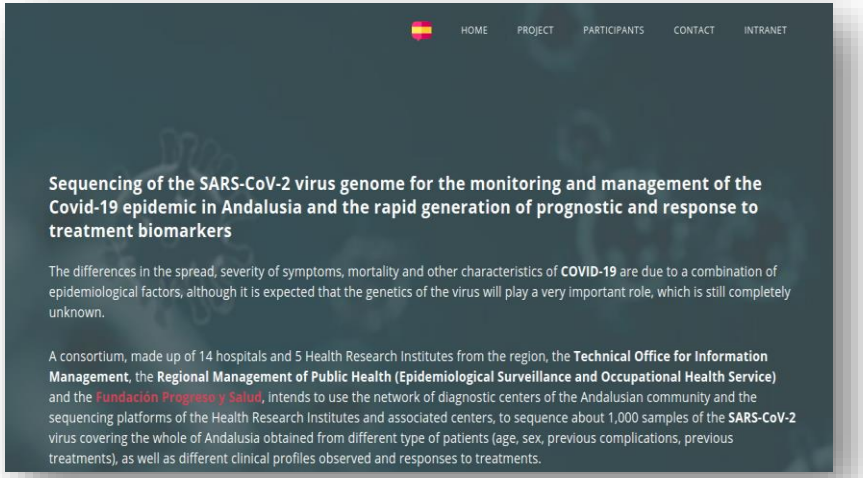
First wave: SARS-CoV-2 genomic sequencing initiatives in Spain and Andalusia.

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

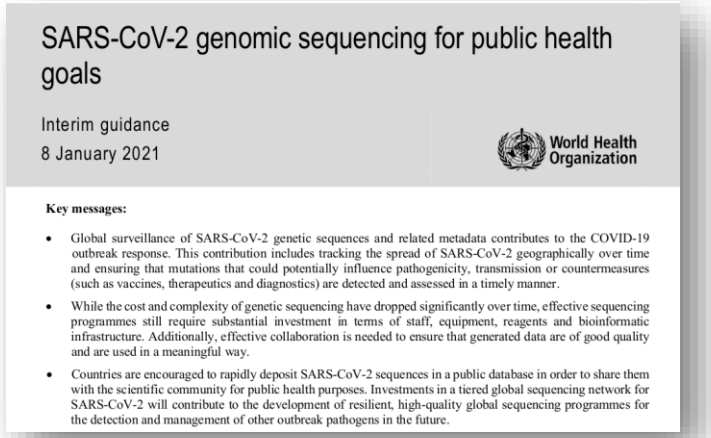
Previous efforts in Spain: *SeqCOVID*
http://seqcovid.csic.es/es/

Previous efforts in Andalusia: *covseq* project
http://clinbioinfosspa.es/projects/covseq



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

World Health Organization

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

SECREARÍA DE ESTADO DE SANIDAD. DIRECCIÓN GENERAL DE SALUD PÚBLICA

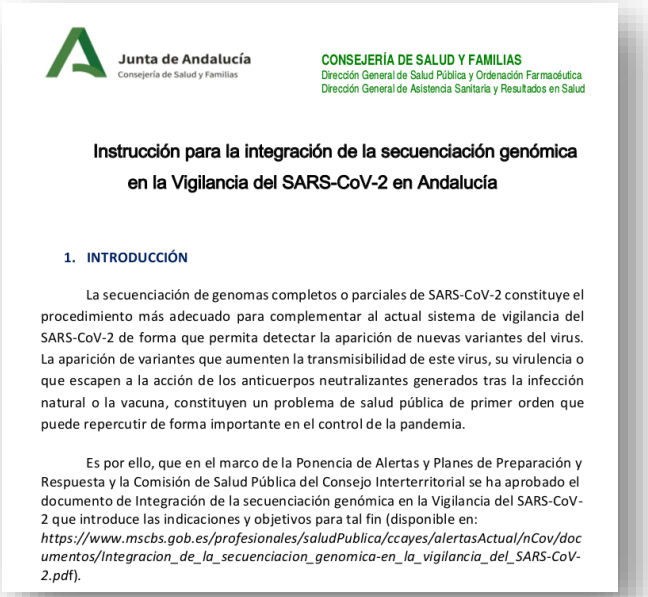
Centro de Coordinación de Alertas y Emergencias Sanitarias

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

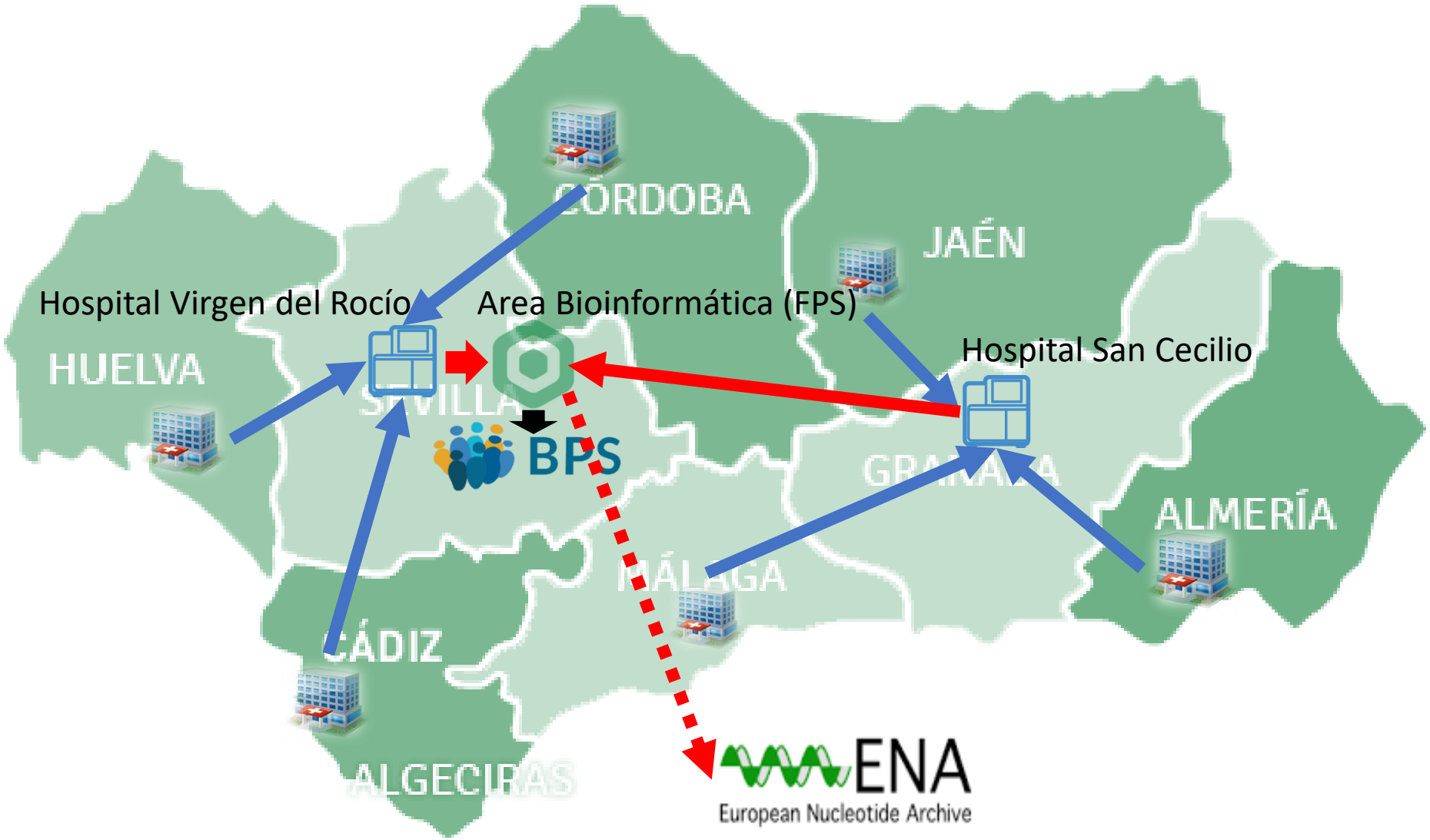
Instrucción para la integración de la secuenciación genómica en la Vigilancia del SARS-CoV-2 en Andalucía

1. INTRODUCCIÓN

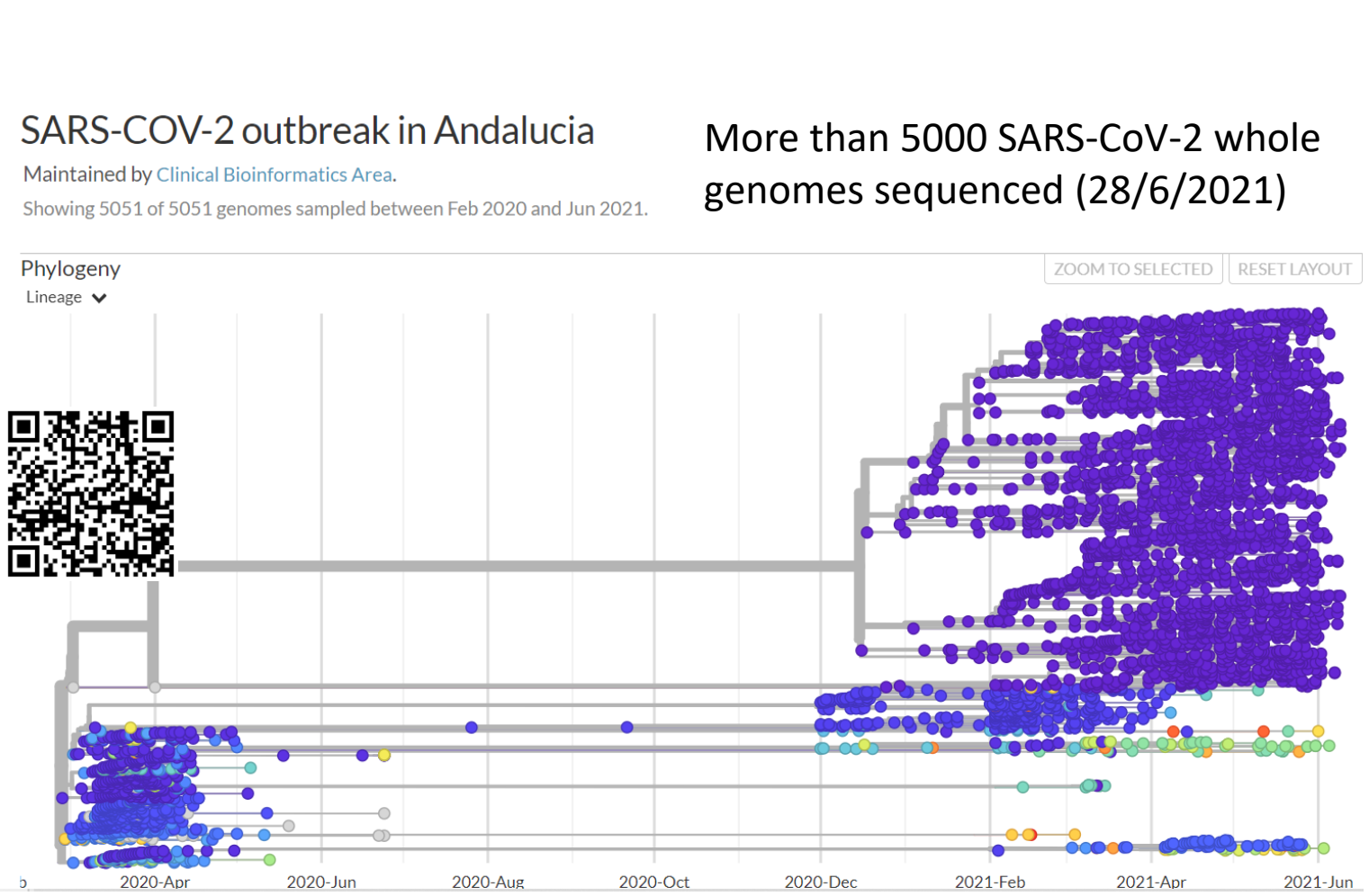
La secuenciación de genomas completos o parciales de SARS-CoV-2 constituye el procedimiento más adecuado para complementar al actual sistema de vigilancia del SARS-CoV-2 de forma que permita detectar la aparición de nuevas variantes del virus. La aparición de variantes que aumenten la transmisibilidad 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.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Integracion_de_la_secuenciacion_genomica-en_la_vigilancia_del_SARS-CoV-2.pdf*).

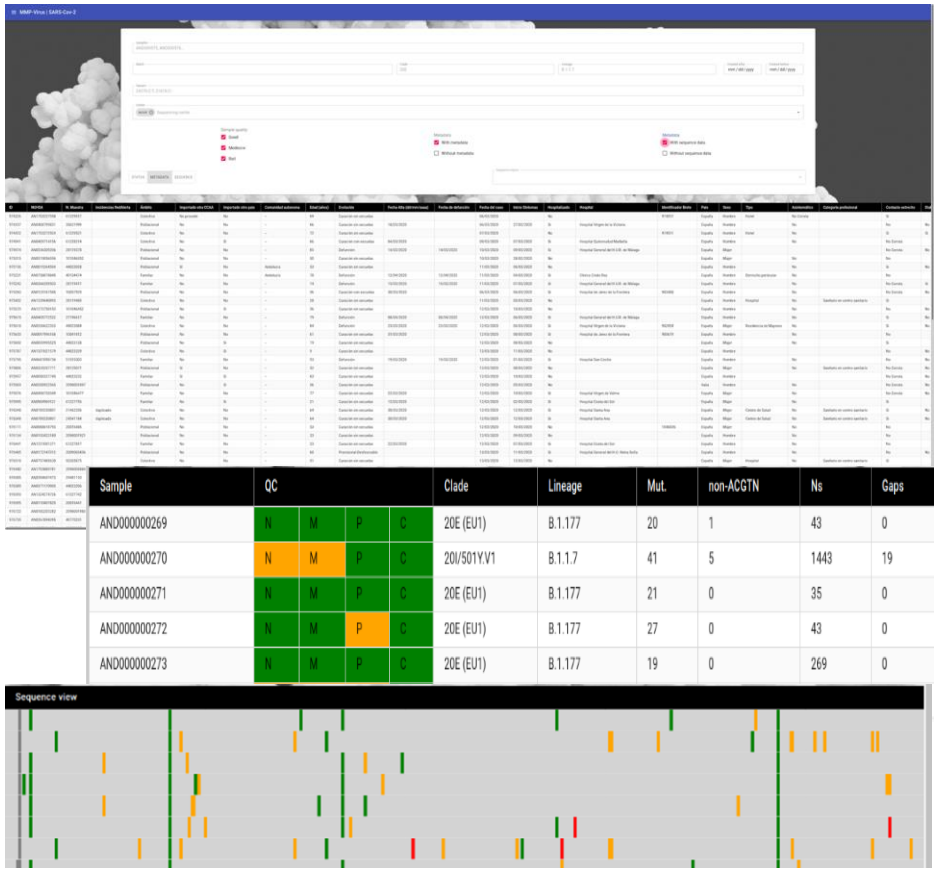
Beyond human genomes:
A clinical circuit of SARS-CoV-2 WGS



Results so far



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

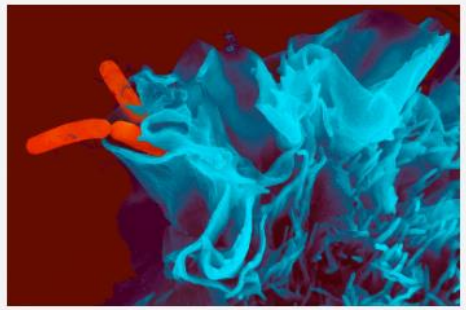
SIEGA: Sistema Integrado de Epidemiología Genómica de Andalucía

SIEGA
Sistema Integrado de Epidemiología Genómica de Andalucía

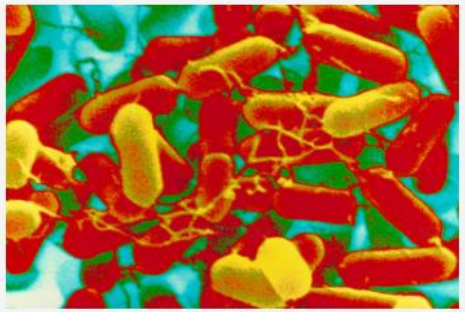
MAPAS EPIDEMIOLÓGICOS

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

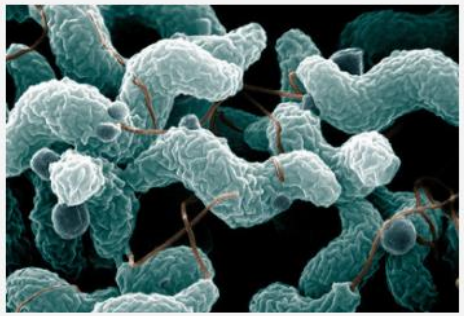
Salmonella enterica

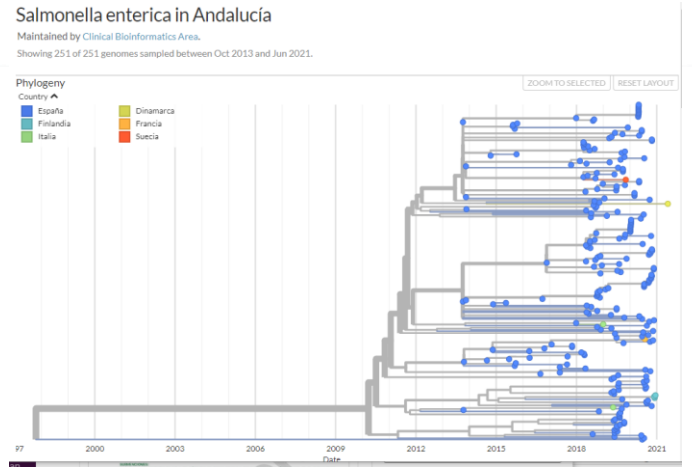


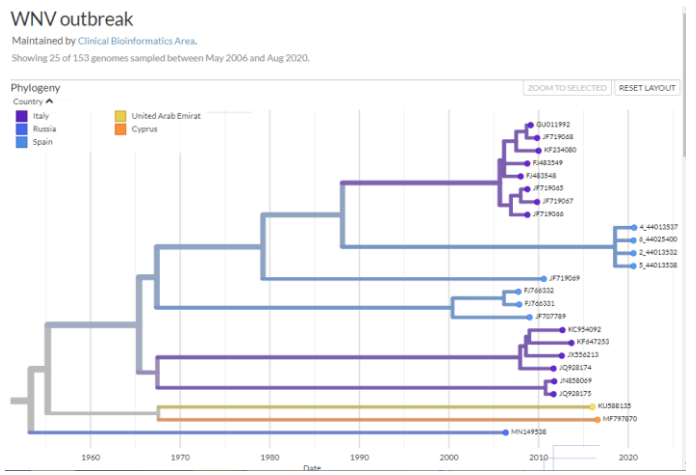
Listeria monocytogenes



Campylobacter jejuni



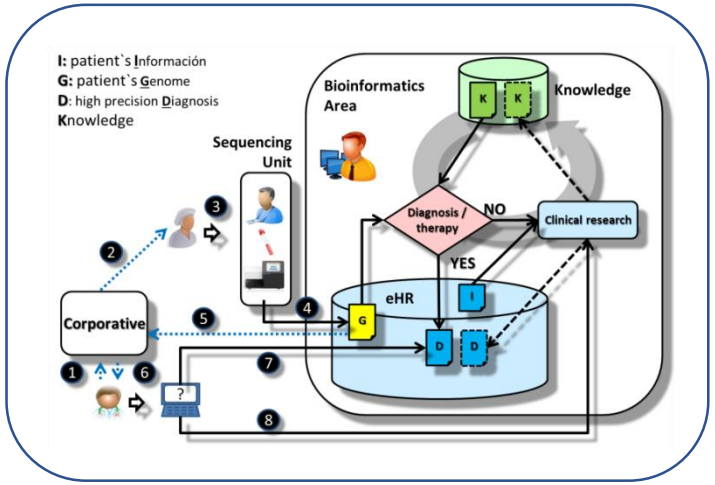


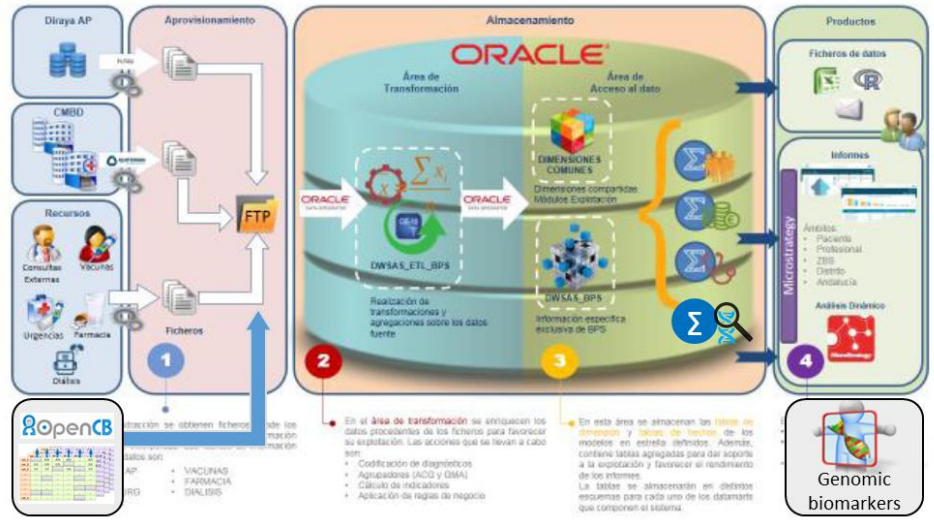


Integral One-Health control of environmental pathogens that can be transmitted to humans.
Currently includes Salmonella, Listeria and Campylobacter. Recently West Nile Virus has been added

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 tan 13 million patients of the Andalusian Public Health System.

BPS is used for secondary data analysis

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

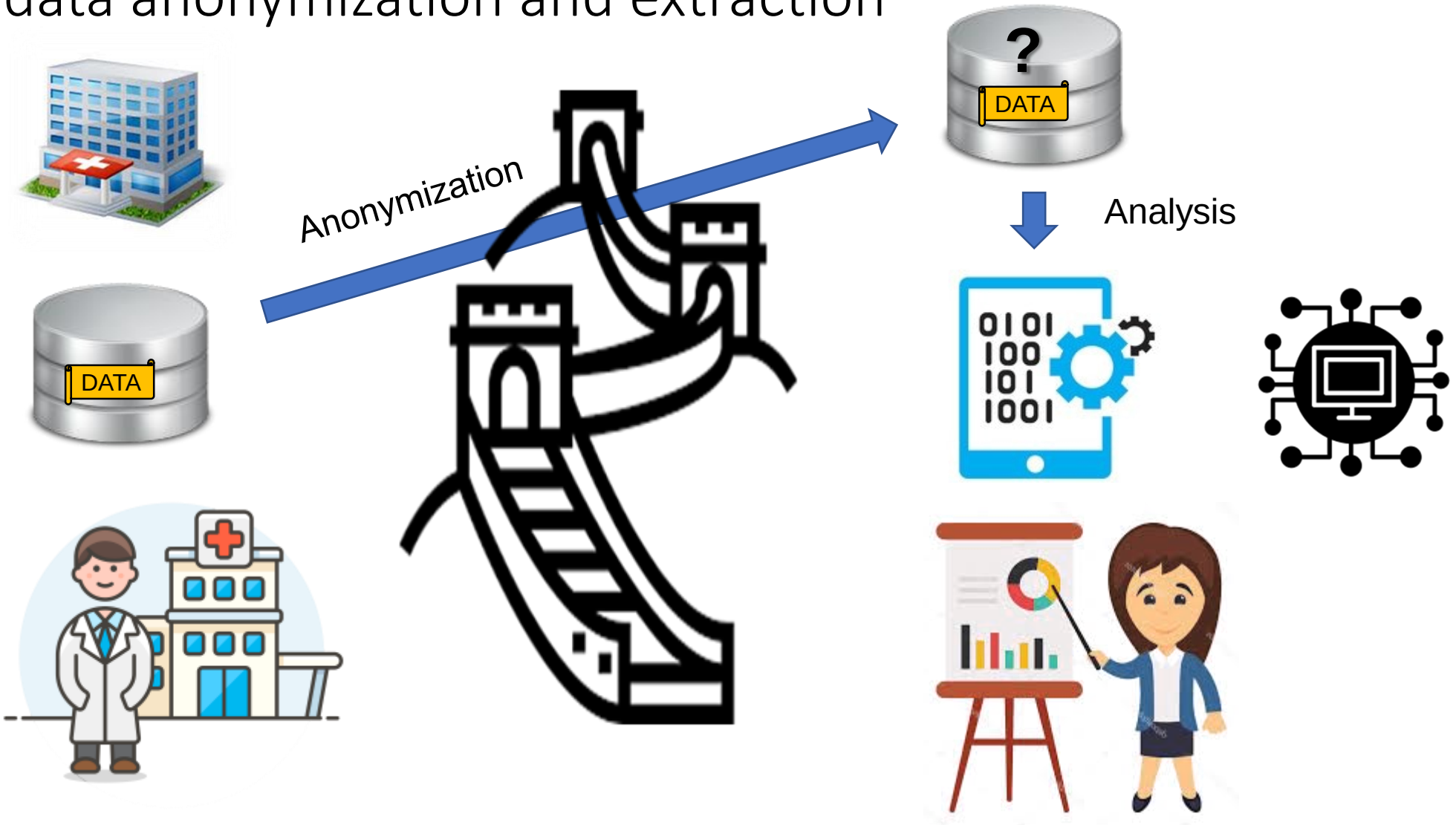
Health system

Firewall

World

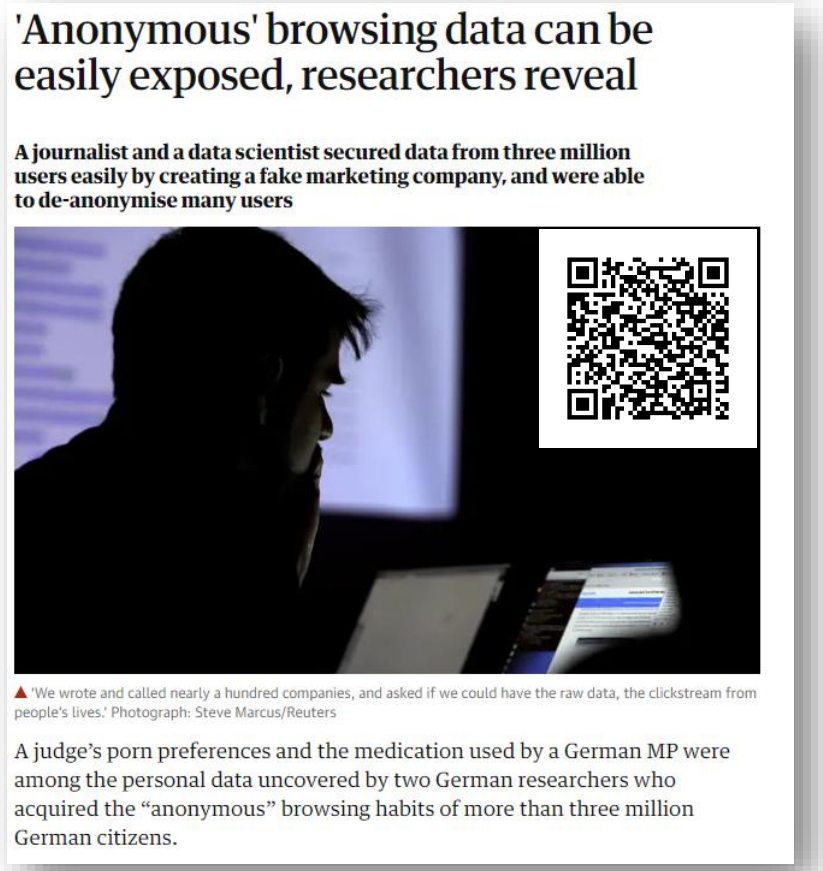
Problems with secondary analysis of clinical Big Data

Solution: data anonymization and extraction



Risk: patient re-identification.

Many examples show that anonymized data can be re-identified



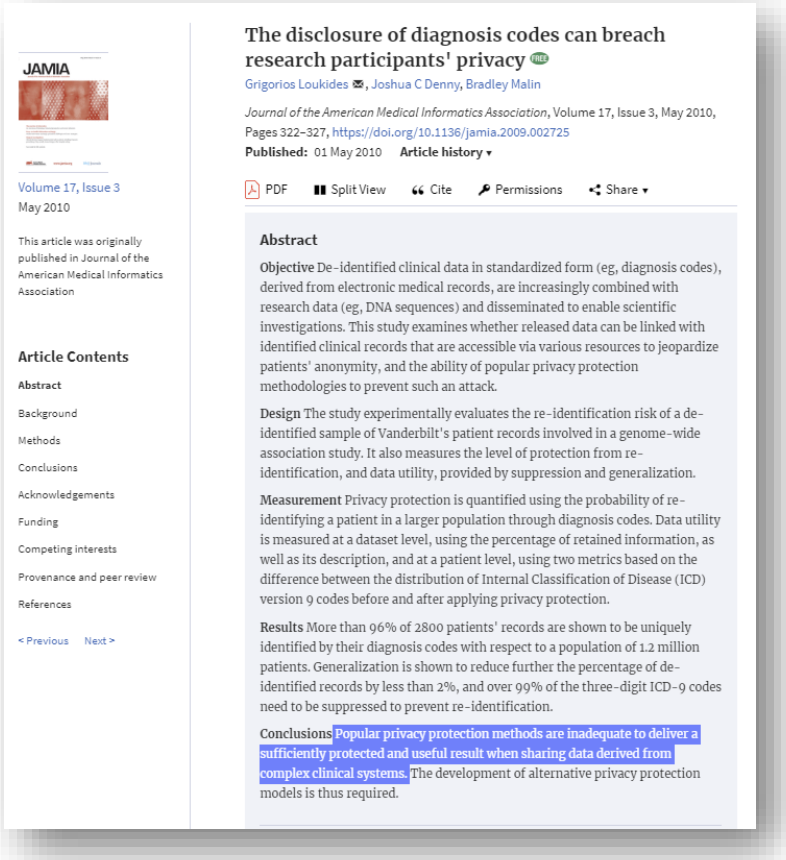
'Anonymous' browsing data can be easily exposed, researchers reveal

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



JAMIA

Volume 17, Issue 3
May 2010

This article was originally published in Journal of the American Medical Informatics Association

Article Contents
Abstract
Background
Methods
Conclusions
Acknowledgements
Funding
Competing interests
Provenance and peer review
References
< Previous Next >

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.

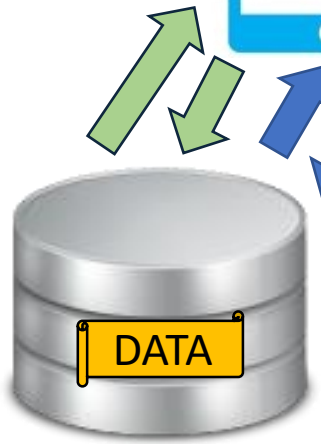
Conclusions Popular privacy protection methods are inadequate to deliver a sufficiently protected and useful result when sharing data derived from complex clinical systems. The development of alternative privacy protection models is thus required.

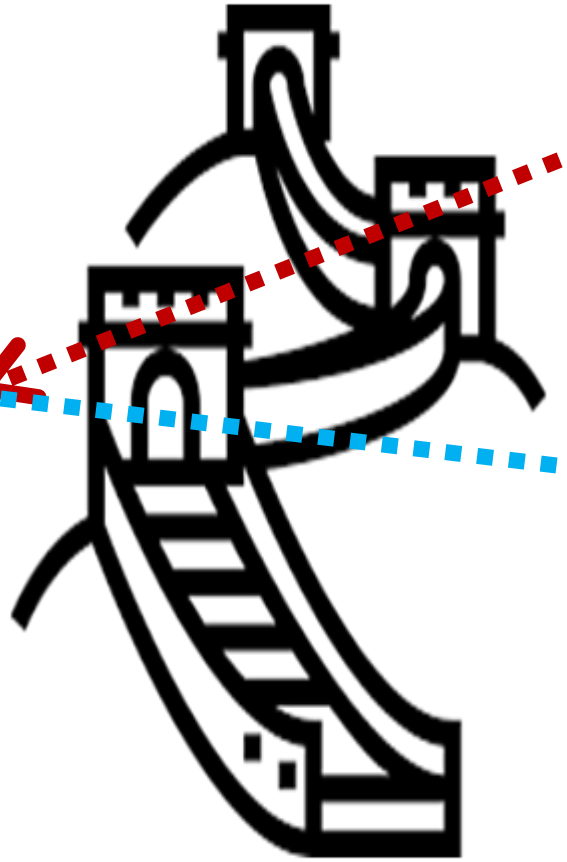
Studies have shown that de-identified hospital discharge data could be re-identified using basic demographic attributes.

Problems with secondary analysis of clinical Big Data

Solution: Bring the algorithm to the data

Algorithms and software using personal data



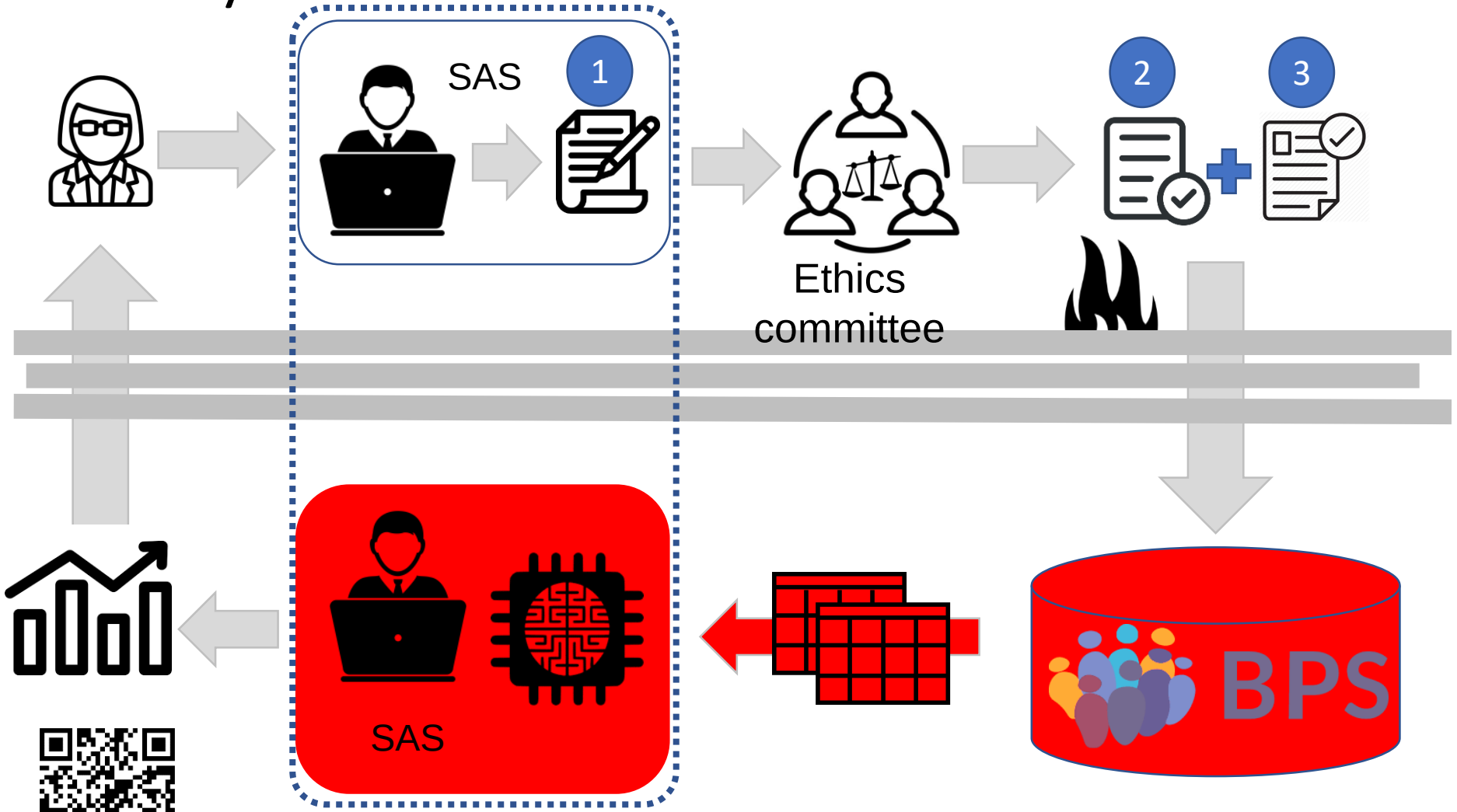


Algorithms and software generation

Aggregated results with no personal data

Bringing the algorithm to the data and not vice versa

Infrastructures for secure secondary data analysis



Requirements:

1. Research project
2. Approval ethics committee
3. Evaluation of impact in data protection (EIPD)

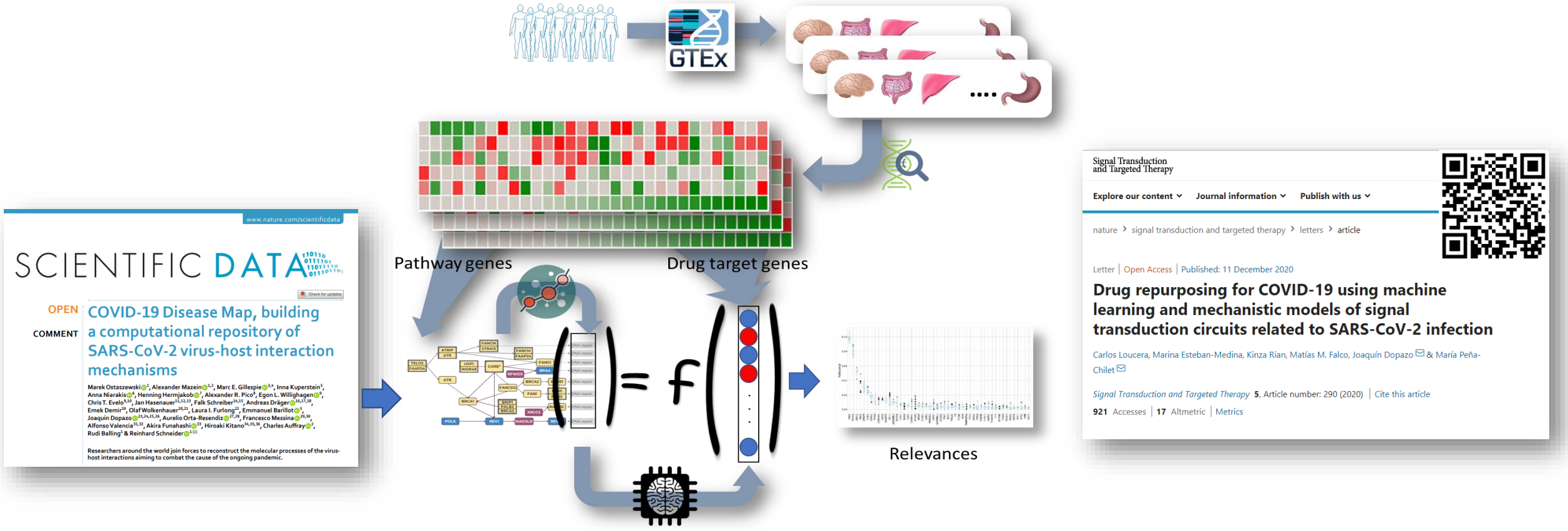
The current RWE Project and the future SAS Big Data Infrastructure

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

Machine learning and potential causal relationships



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 Ostaszewski[1], Alexander Mazein[1,2], Marc E. Gillespie[3,4], Inna Kuperstein[5], Anna Niarakis[6], Henning Hermjakob[7], Alexander R. Pico[8], Egon L. Willighagen[9], Chris T. Evelo[9,10], Jan Hasenauer[11,12,13], Falk Schreiber[14,15], Andreas Dräger[16,17,18], Emek Demir[19], Olaf Wolkenhauer[20,21], Laura I. Furlong[22], Emmanuel Barillot[5], Joaquin Dopazo[23,24,25,26], Aurelio Orta-Resendiz[27,28], Francesco Messina[29,30], Alfonso Valencia[31,32], Akira Funahashi[33], Hiroaki Kitano[34,35,36], Charles Auffray[7], Rudi Balling[1] & Reinhard Schneider[1]

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.

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

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

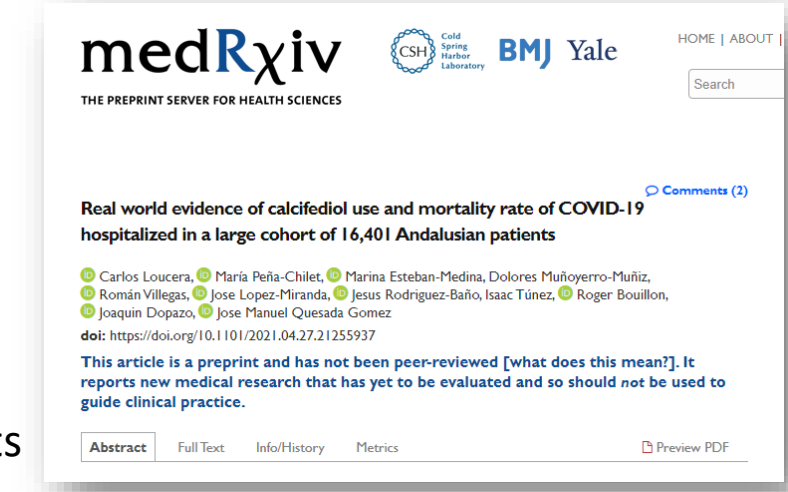
Systematic RWE validation of the protective potential of drugs in COVID-19

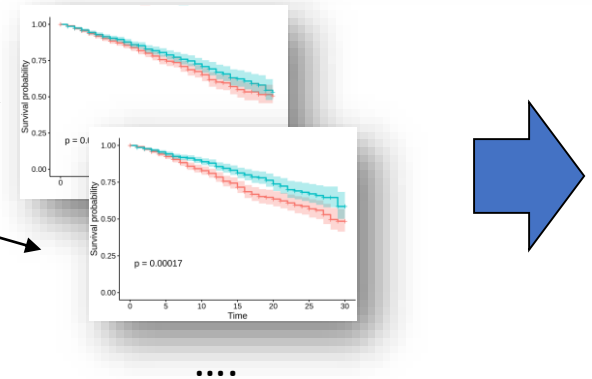
Real world evidence of the use of drugs prescribed for other reasons and mortality rate of COVID-19 hospitalized in a large retrospective cohort of 15,968 Andalusian patients hospitalized between February and November 2020

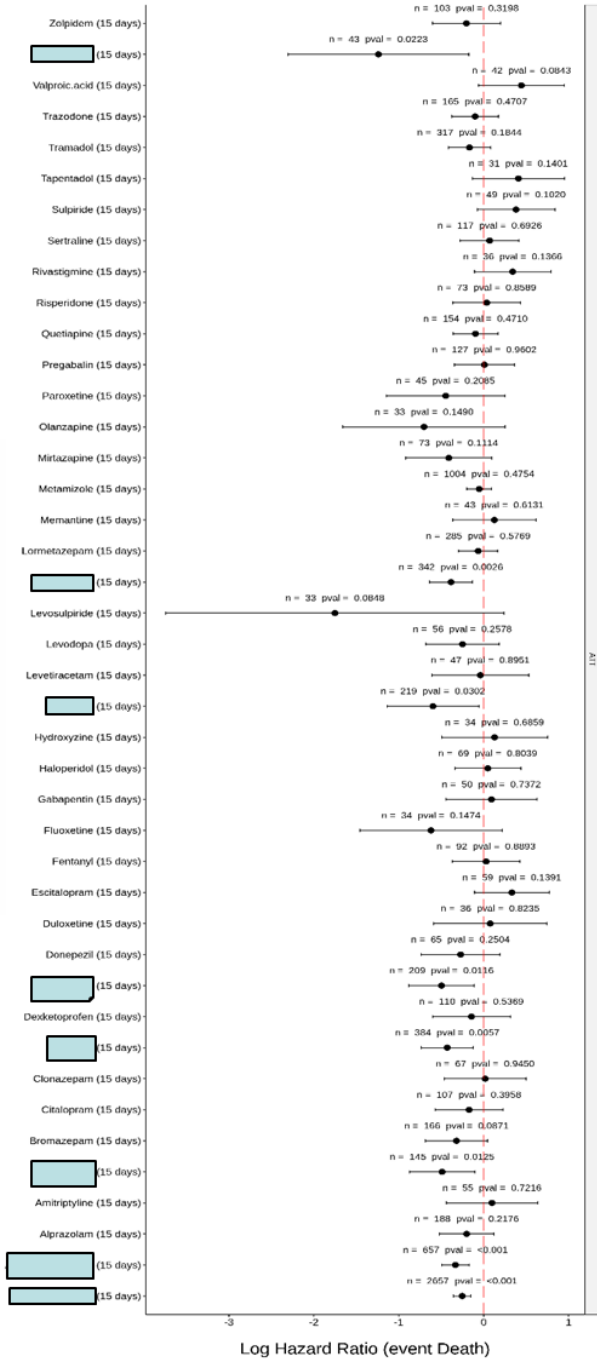
BPS

- Age
- Gender
- Comorbilities
- Vaccines
- Other treatments
- …

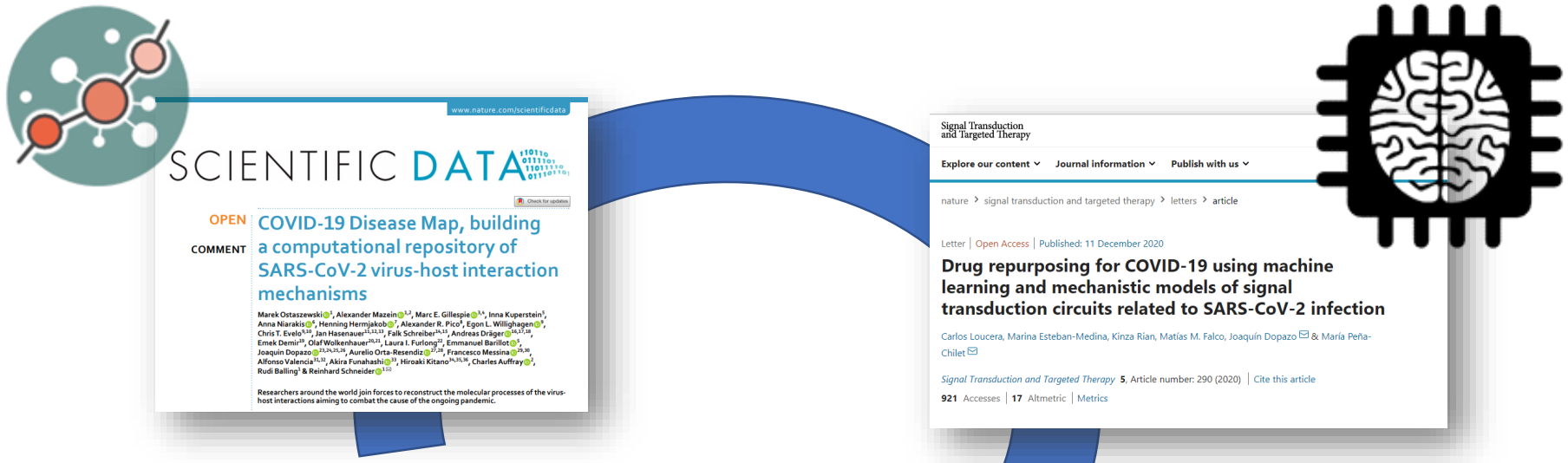
- Survival
- Drug 1
- Drug 2
- …







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



1. Modeling Biological knowledge

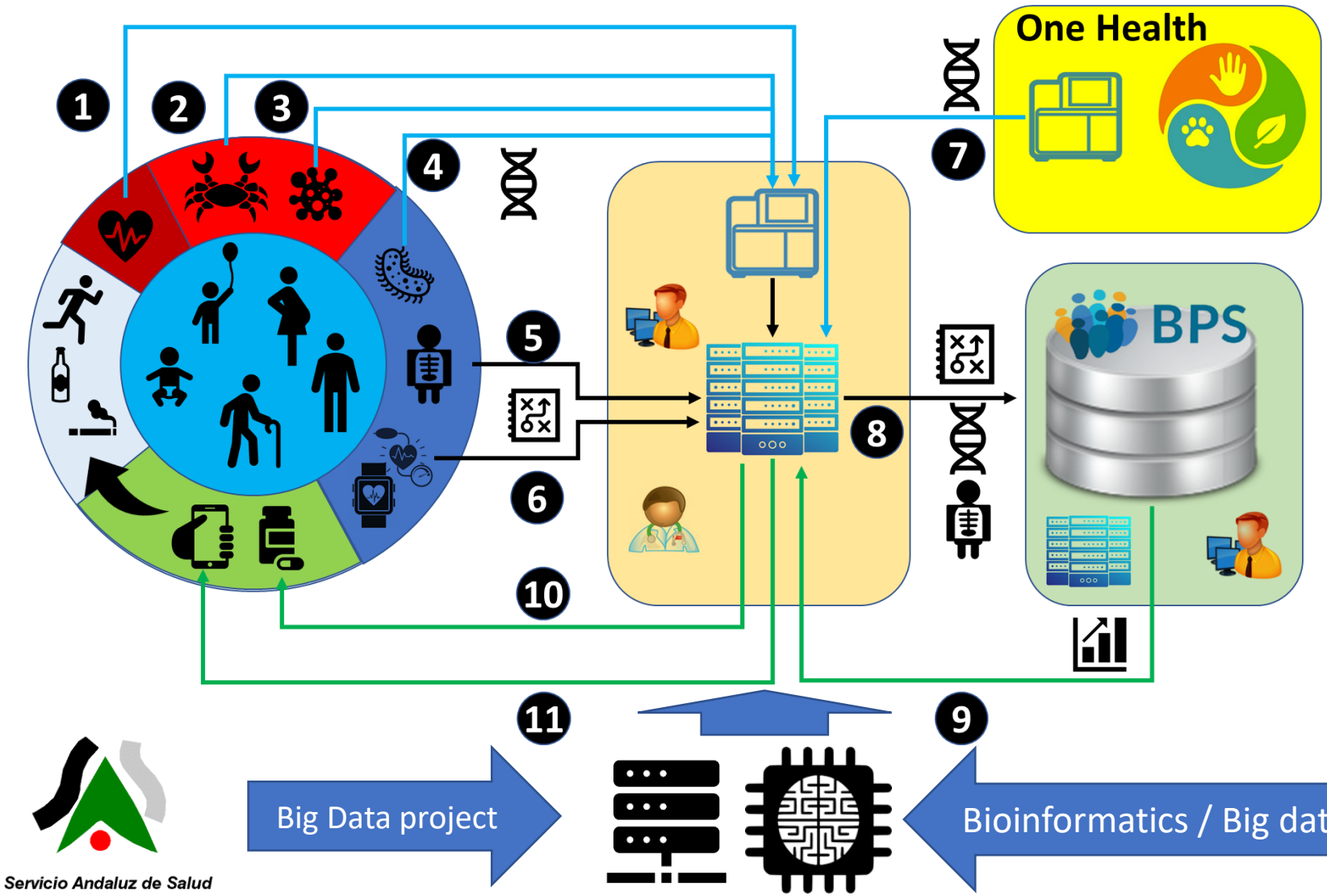
2. Learning from genomic big data

A new paradigm of knowledge generation

BPS

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

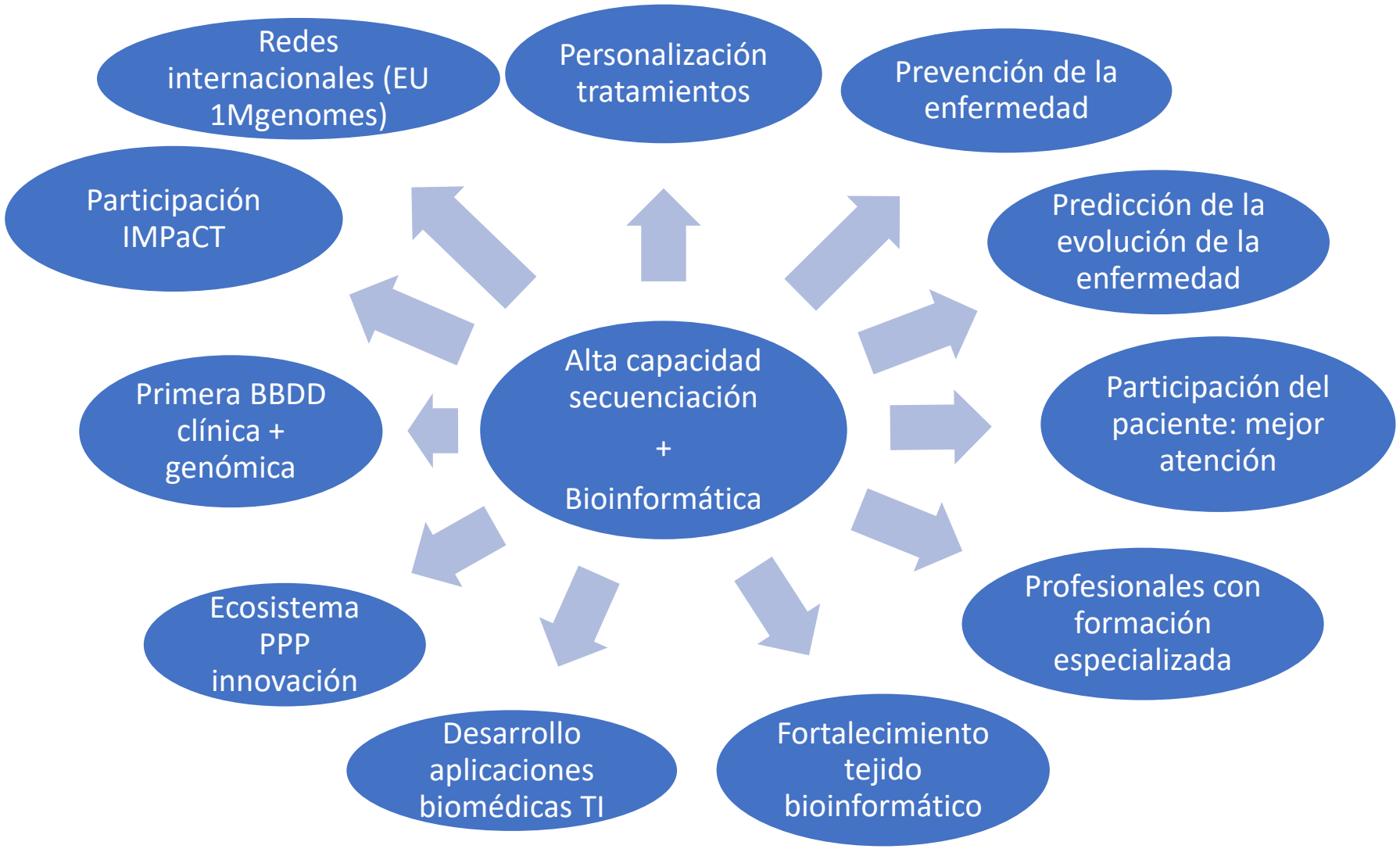
9 Knowledge generated

10 Incidental findings
11 Life style recommendations

Servicio Andaluz de Salud

Junta de Andalucía
Consejería de Salud y Familias
FUNDACIÓN PROGRESO Y SALUD

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/