E3.5.Comparativa de Procesos en Entornos Hospitalarios v 1.0

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Instituto de Salud Carlos III

IMPaCT

Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología

Program	IMPacT: Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología		
Project Name	IMPaCT-Data: Programa de Ciencia de Datos de IMPaCT		
Expedient	IMP/00019		
Duration	January 2021 – December 2024		
Website	impact-data.bsc.es/		
Work Package	WP3 – Genomics		
Task	T3.2 Adaptación, instalación y uso de software de código abierto para el análisis e integración de distintas fuentes de datos		
Deliverable	E3.5 Comparison of Processes in Hospital Environments		
Version	1.0		
Due Date	30/09/2024	***Approval Date***	18/12/2024
Responsible	CRG		
Dissemination Level	X	PU	Public
		CO-IMP	Confidential, only IMPaCT pillars members, including the evaluation commission from IMPaCT

Authors		
Organization	**Name**	**Role**
Acronym	*Name and Surname*	*Coordination / Author / Reviewer*
CRG	Jordi Rambla	Author
CRG	Amy Curwin	Author
IIS-FJD	Pablo Minguez	Reviewer
NASERTIC	Igor Ruiz de los Mozos	Reviewer

Versions History			
N.	**Date**	**Description**	**Author**
v 0	04/03/2022	Translated from Spanish version	L. López (BSC-CNS)
v 0.0	19/01/2024	Created	A. Curwin
v 0.1	15/04/2024	Table of contents and outline added	Amy Curwin, Jordi Rambla
v 0.1	15/05/2024	Overview draft	Igor Ruiz de los Mozos (NASERTIC)
v 0.2	24/05/2024	Genomics in Health Care draft, Genomic Data draft and Genomic data sections	Igor Ruiz de los Mozos (NASERTIC)
v 0.3	30/07/2024	Expanded various sections, added figures, tables	Amy Curwin
v 0.4	25/08/2024	Expanded more sections	Jordi Rambla
v 0.5	05/09/2024	Edits, addressing outstanding issues	Amy Curwin
v 0.6	04/10/2024	Final edits, added Annex, abbreviations, minor edits. Sent to reviewers	Amy Curwin, Jordi Rambla
v 1.0	16/10/2024	Addressed review comments. Final version	Amy Curwin

Content

IMPaCT Data
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Instituto de Salud Carlos III
Financiado por la Unión Europea
NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Executive Summary

The goal of IMPaCT is personalised medicine (PM). PM is based on getting access to patient profiles, or specific aspects of patient profiles (like genomic variants) obtained by research or secondary analysis of healthcare data. These data types are generated in research and healthcare institutions via processes that are similar, but their ability to be shared is limited by a number of factors, ranging from technical, ethical, social and legal. This document explains the state of the art for PM, internationally and in Spain, and explains the progress made within IMPaCT-Data to achieve the overall goals of PM in Spain.

IMPaCT Data
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Instituto de Salud Carlos III
Financiado por la Unión Europea
NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Introduction

Audience

Deliverable E3.5 “Comparison of Processes in Hospital Environments” is envisioned and written for those institutes working with healthcare data who would like to join a federation where the data is shared in accordance with the legal frame. This is the mission of IMPaCT-Data project that is setting the basis for the successful implementation of the Spanish personalised medicine program.

Topic

This deliverable is related to task 3.2 of WP3 “Adaptation, installation and use of open-source software for the analysis and integration of different data sources”.

Relation to other Deliverables

This is the fourth deliverable of WP3 and builds on E3.4 “Genomic Analysis in Healthcare Environments" where information was gathered across partner institutes regarding what genomic processes are performed, what diseases are most studied, what pipelines are used and their portability.

Also related to E5.1 and E5.3 by WP5, where they describe the technologies available for the integration of biomedical data and E5.6 Informe Final sobre la funcionamiento de la red IMPaCT de Beacons.

1 Overview

The IMPaCT program is at the forefront of advancing personalised medicine (PM), a healthcare approach that tailors treatments based on individual patient characteristics, including genetic profiles. Advancing PM depends critically on accessing detailed patient profiles, which include genetic mutations often identified through high-throughput sequencing techniques such as whole genome and exome sequencing. These profiles are enriched with secondary analyses of existing healthcare metadata and phenoclinic data, forming a robust foundation for personalised treatment plans.

However, the generation and processing of this data, vital for personalised medicine, face significant challenges across both research facilities and healthcare institutions. While these settings share similar processes, such as the collection, storage, and analysis of complex datasets including clinical, genomic, and imaging data, they differ markedly in their handling of patient privacy. Research environments often have more freedom in data usage, whereas clinical settings are tightly regulated to protect patient confidentiality, making interoperability a challenge.

Economic and Welfare Benefits of Personalized Medicine

The financial implications of integrating personalised medicine into the healthcare system are profound[1]. For every euro invested in personalised genomic medicine, three euros are saved by the national health system. These savings stem from various factors including reduced incidence of misdiagnosis, decreased use of ineffective therapies, and more targeted use of healthcare resources. Furthermore, personalised medicine significantly enhances patient welfare by minimising adverse drug reactions, reducing treatment side effects, and improving overall quality of life through tailored treatment plans. In this regard, genomic screening is cost effective and worth the investment[2].

Challenges in Hospital Data Management

Many hospitals in Spain are increasingly acquiring sequencing and imaging equipment; however, they may lack the necessary expertise, human resources or infrastructure to perform genomic analysis, imaging analysis, and other bioinformatics tasks. As a result, primary and secondary analyses are often outsourced to commercial platforms that rely on cloud environments. This approach can lead to issues with data privacy and compliance, as hospitals may not have the capability to interpret or analyse personalised data beyond standard clinical practices. Moreover, most hospitals lack specialised IT teams to manage high-throughput data, which is crucial for maintaining data integrity, privacy, and efficiency.

The compliance with the minimum requirements of RD 311/2022, which regulates the National Security Scheme (ENS), further complicates data management in healthcare settings

[1] https://pubmed.ncbi.nlm.nih.gov/37155986/

[2] https://www.nejm.org/doi/full/10.1056/NEJMoa2300792

(Deliverable E6.4. Aspectos de Seguridad en el Manejo de Datos Sensibles[3]). Hospitals struggle to meet these standards, which are critical for ensuring the secure and efficient handling of sensitive health data.

Box 1: Urgent Need for a Genetics Speciality in Spain

The integration of genomics into healthcare in Spain faces institutional hurdles, notably the absence of a recognized specialty in Genetics. This gap has been highlighted by concerted efforts from major scientific societies such as the Spanish Association of Human Genetics (AEGH), the Spanish Association of Prenatal Diagnosis (AEDP), the Spanish Society of Pharmacogenetics and Pharmacogenomics (SEFF), the Spanish Society of Clinical Genetics and Dysmorphology (SEGCD), and the Spanish Society of Genetic Counseling (SEAGEN). These groups, representing over 2,000 professionals, have repeatedly emphasised, during their biannual congress, the urgent need to establish Genetics as a recognized medical specialty in Spain—the only country in the European Union without such professional recognition.

Role of IMPaCT-Data

In this context, the role of IMPaCT-Data becomes crucial. By defining gold standards for process management and data handling, IMPaCT-Data is setting the stage for a healthy development of personalised medicine in clinical and hospital settings in Spain. The initiative aims to bridge the gaps between research and clinical data handling, advocate for the professional recognition of Genetics (see Box 1), and support hospitals in adopting advanced data analysis and dissemination technologies.

1.1 Genomics in healthcare

Healthcare processes assisted by genetic and genomics tests are applied to many different healthcare aspects, from epidemiology, screening, prevention, diagnosis and prognosis and to many different disease domains, specially in oncology and rare diseases (see Box 2: "NAGEN as an example of regional project") .

Genetic counselling plays a pivotal role in reducing the diagnostic odyssey that many patients with rare diseases endure. Historically, patients and their families have faced lengthy and frustrating quests for accurate diagnoses, averaging around 5.6 years[4]. Genetic counselling leverages genomic data to provide quicker, more accurate assessments, guiding patients more efficiently through their treatment options and reducing the emotional and financial strain of prolonged uncertainty.

[3] https://b2drop.bsc.es/index.php/f/2772869

[4] DOI: 10.1056/NEJMoa2035790

Box 2: NAGEN as an example of regional project

NAGEN in Navarra have accumulated over 3,600 genomes from patients with various diseases, significantly advancing genomic research and application in clinical settings. NAGEN's portfolio includes several specialised projects. NAGEN1000, started in 2017, conducts large-scale genomic sequencing to identify variants linked to rare diseases and cancer. PharmaNAGEN, focuses on pharmacogenomics to discover genetic predictors of drug responses and adverse reactions. NAGENcol, studies the genetic and environmental factors contributing to hypercholesterolemia. NAGENpediatrics investigates genetic determinants in complex clinical scenarios within pediatric care. NAGENmx aims to understand genetic variations affecting breast cancer screening across diverse populations. ReproNAGEN, explores genetic factors influencing reproductive health. NAGENdata, launched in 2023, this newest initiative manages the ecosystem for the use and reuse of genomic data from the Genoma Navarra project (NAGEN), with a focus on case studies related to neurological disorders. These projects exemplify the depth and breadth of contemporary genomic analysis, demonstrating its vital role in advancing personalised medicine by improving our understanding of genetic determinants in health and disease.

The landscape of genomic analysis in healthcare is rapidly evolving. It started with the analysis of chromosomal aberrations using techniques like fluorescence in situ hybridization (FISH), into chips designed for different types of diseases, to next generation sequencing (NGS) in their gene panel, clinical exome, whole exome or whole genome versions. These different techniques are usually applied in progressive order if the previous ones do not provide a diagnosis confirmation. New technologies such as long reads sequencing or optical genome mapping may also be incorporated into the diagnostic algorithms quite soon.

This variety of analysis options involves different protocols, different instruments, different analysis, and different expertise. Additionally the type and quantity of data that they generate differ in several orders of magnitude. In the deliverable E3.4 “Genomic Analysis in Healthcare Environments[5]”, a landscape analysis of IMPaCT-Data partners was conducted that exemplified this fact. The participating centres were diverse, performing a wide range of techniques, at different volumes, with varying disease focus. With respect to workflows and analysis, on the one hand, file formats used were quite homogenous and standardised (e.g. fastq, BAM, VCF), however, the workflows used varied significantly, ranging from standardised, commercially available and/or in-house developed pipelines, in combination or alone. Indeed, a need for more standard available workflows was apparent from this analysis.

[5] https://b2drop.bsc.es/index.php/f/2794025

The importance of collaboration and data sharing

The development of specialised human variation databases has greatly enhanced our, still very incomplete, understanding of genetic diversity and disease association. Population-specific catalogues, such as the 1000 Genome Project (1KGP) and the Genome Aggregation Database (gnomAD), provide invaluable resources for identifying rare variants that might be overlooked in smaller studies. These databases compile genetic data from diverse populations, offering a richer context for analysing human genetic variation.

Given that the representation of different human populations is still very limited in these global reference knowledge bases, national projects could contribute significantly to this field. For example, UK, USA, and Japan have established extensive genomic databases that support their respective healthcare systems and research communities. In Spain, back in 2010, the Medical Genome Project (MGP)[6] compiled genomic data from unrelated healthy individuals, providing a baseline for understanding common genetic variations among the Spanish population. The Collaborative Spanish Variant Server (CSVS)[7] includes this data together with WES/WGS data from patients with rare diseases. GCAT[8] (Genomes for Life, A Prospective Study of the Genomes of Catalonia) is more recent, with 20,000 people participating to help researchers answer important questions about health and the treatment of illnesses. IMPaCT-Genómica[9] and IMPaCT-Cohorte[10] pillars of IMPaCT have similar aims of improving health and disease treatments in the Spanish population. Therefore, local projects could contribute significantly to improve the efficacy of these knowledge bases and repositories that are then leveraged back to improve diagnoses. This diverse population resources are indeed more useful in regions where the habitants have diverse genetic ancestries and mixed populations, like Spain.

Given that genomics analysis in healthcare is recent, diverse, and quickly evolving, the protocols applied are not as mature as those for biochemical analyses. Therefore, collaboration among healthcare centres remains essential until standardised protocols are fully established. This collective effort helps unify disparate approaches to genomic analysis, ensuring that data handling and interpretation are consistent across different settings.

1.2 Standards in healthcare

For healthcare and research professionals to effectively use data from different sources and to properly share data, it is crucial that the data is homogeneous in terms of variables, terminologies, and vocabularies. Standardisation ensures that data collected from various sources can be integrated, compared, and analysed reliably, enhancing the quality and

[6] https://www.clinbioinfosspa.es/content/medical-genome-project

[7] DOI: 10.1093/nar/gkaa794

[8] http://www.gcatbiobank.org/

[9] https://genomica-impact.es/

[10] https://cohorte-impact.es/

IMPaCT Data
Barcelona Supercomputing Center Centro Nacional de Supercomputación
Gobierno de España Ministerio de Ciencia, Innovación y Universidades
Instituto de Salud Carlos III
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

efficiency of healthcare delivery and biomedical research. These concepts have been extended in the definition of FAIR principles[11].

Homogeneous data is essential for several reasons. Interoperability is a key benefit, as standardised data allows different healthcare systems and research institutions to exchange and use information seamlessly. This capability is crucial for collaborative research and integrated patient care. Additionally, data quality and integrity are significantly improved through standardisation, ensuring that information is accurate, complete, and consistent across various platforms. Efficient analysis is another major advantage, as standardised data facilitates the aggregation and analysis of large datasets, leading to more robust and reproducible research findings. Furthermore, patient care is directly enhanced through more accurate and timely information sharing, allowing healthcare providers to make better-informed decisions that improve patient outcomes.

Several standards have been established to address the need for homogeneity in healthcare data. **Health Level Seven International (HL7)[12]** provides a set of international standards for the exchange, integration, sharing, and retrieval of electronic health information. HL7 standards support clinical practice and the management, delivery, and evaluation of health services. **Fast Healthcare Interoperability Resources (FHIR)[13]**, developed by HL7, is a standard describing data formats and elements (known as "resources") and an API for exchanging electronic health records. FHIR aims to simplify the implementation process without sacrificing information integrity.

The **Picture Archiving and Communication System (PACS)[14]** is an imaging technology used primarily in healthcare organisations to securely store and digitally transmit electronic images and clinically relevant reports, eliminating the need to manually file, retrieve, or transport film jackets. **Digital Imaging and Communications in Medicine (DICOM)[15]**, specifically DICOM3, is the latest version of the standard for handling, storing, printing, and transmitting information in medical imaging. DICOM ensures that medical images and associated data can be consistently managed and communicated across various devices and systems.

Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)[16] provides a comprehensive clinical terminology that standardises the representation of clinical phrases, making it easier to share and analyse health information. The **Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM)[17]** is used to standardise the representation of healthcare data, enabling large-scale analytics and facilitating the integration

[11] https://www.go-fair.org/fair-principles/
[12] https://www.hl7.org/
[13] https://ecqi.healthit.gov/fhir
[14] https://doi.org/10.1148%2Fradiographics.12.1.1734458
[15] https://www.dicomstandard.org/
[16] https://www.snomed.org/
[17] https://ohdsi.github.io/CommonDataModel/

of diverse data sources into a common format for comparative effectiveness research and other analyses.

The **International Classification of Diseases (ICD)**[18], with its ninth, tenth and now 11th revisions, (ICD-9, ICD-10, ICD-11), is a globally used diagnostic tool for epidemiology, health management, and clinical purposes. These classifications provide a common language for reporting and monitoring diseases, enabling the comparison of health data across time and borders.

1.3 Genomic Data Standards

Given the complexity and specificity of genomic data, additional standards beyond those used for general healthcare data are essential. These standards are crucial for ensuring the interoperability, accuracy, and security of genomic information across different institutions and countries. They facilitate the responsible sharing of genomic and clinical data, enabling more effective research collaborations and personalised healthcare.

The advancement of genomic knowledge relies heavily on the efforts of various consortia and organisations that develop comprehensive reference materials and standards. These groups have been instrumental in mapping and understanding the human genome, providing critical resources that underpin modern genomic research and clinical applications.

1.3.1 Consortia and Organizations Developing Reference Human Genome Knowledge

The **Human Genome Project (HGP)**[19], completed in 2003, was a landmark international research initiative aimed at mapping all the genes of the human genome. Coordinated by the National Institutes of Health (NIH) and the Department of Energy (DOE) in the United States, along with international partners, the HGP provided the first comprehensive reference sequence of the human genome. This project laid the groundwork for subsequent genomic research by identifying approximately 20,000-25,000 human genes and sequencing the 3 billion base pairs of human DNA. The HGP has enabled researchers to understand the genetic basis of diseases, develop new diagnostics and treatments, and advance personalised medicine.

Building on the foundation of the HGP, the **Human Pangenome Reference Consortium (HPRC)**[20] aims to capture the full diversity of human genetic variation. Unlike the HGP, which provided a single reference genome, the Human Pangenome Project seeks to create a comprehensive reference that includes genomic sequences from multiple individuals. This approach acknowledges the genetic diversity across different populations and aims to provide

[18] https://www.who.int/standards/classifications/classification-of-diseases
[19] https://www.genome.gov/human-genome-project
[20] https://humanpangenome.org/

a more accurate representation of human genetic variation. By doing so, it enhances the understanding of genetic contributions to health and disease, particularly for populations that were underrepresented in the original HGP.

The **Telomere-to-Telomere (T2T) Project**[21] represents the latest advancement in human genome sequencing. This project aims to complete the remaining gaps left by the HGP by sequencing the entire human genome, including the highly repetitive regions that were previously inaccessible. The T2T Consortium, led by the National Human Genome Research Institute (NHGRI), achieved a complete and gapless sequence of a human genome in 2021. This comprehensive reference includes the repetitive centromeric regions and telomeres, providing an unprecedented level of detail. The T2T Project's complete sequence is crucial for understanding chromosome structure, function, and evolution, and it provides a more accurate template for studying genetic variation.

1.3.2 Human Genome Standards Consortia

The field of genomics relies on various consortia and organisations to establish and maintain data standards, ensuring interoperability, security, and accuracy in genomic research and clinical applications. Below is an overview of key organisations and initiatives that contribute to the development and implementation of human genome standards.

The **Global Alliance for Genomics and Health (GA4GH)**[22] is an international coalition aimed at accelerating the potential of genomic medicine to advance human health. GA4GH provides frameworks and tools for the responsible sharing of genomic and clinical data. One of its primary contributions is the Genomic Data Toolkit, which includes best practices for data sharing, privacy protection, and data security. GA4GH standards ensure that genomic data can be exchanged seamlessly across different systems while safeguarding patient privacy and complying with international data protection regulations.

ELIXIR[23] is an intergovernmental organisation that brings together life science resources from across Europe. Its **Human Data Communities** initiative focuses on the standardisation and integration of human genomic data. ELIXIR supports the development of data standards and provides infrastructure to ensure that genomic data can be shared and analysed efficiently across different research institutions and countries.

The **European Genomic Data Infrastructure (GDI) Project**[24] aims to build a secure and federated infrastructure for sharing genomic and clinical data across Europe. By integrating data from national and regional initiatives, the GDI project facilitates large-scale research and enhances the understanding of genetic diseases. This initiative is closely linked with the

[21] https://www.science.org/doi/10.1126/science.abj6987
[22] https://www.ga4gh.org/
[23] https://elixir-europe.org/
[24] https://gdi.onemilliongenomes.eu/

General Data Protection Regulation (GDPR), ensuring that data sharing complies with stringent privacy laws while promoting scientific innovation.

The **European Health Data Space (EHDS)**[25] is a proposal by the European Commission to create a common framework for health data sharing across Europe. The EHDS aims to facilitate secure access to health data for research, innovation, and policy-making. This initiative complements the GDI project by providing a broader context for health data integration, ensuring that genomic data can be used effectively alongside other types of health data to improve patient outcomes and drive healthcare innovations.

The **European Genome-phenome Archive (EGA)**[26] is a repository for securely archiving and sharing all types of genetic and phenotypic data. EGA provides researchers with access to data that is crucial for understanding the genetic basis of diseases. By supporting data standards and providing a secure environment for data sharing, EGA plays a critical role in facilitating international genomic research collaborations. The Federated EGA (FEGA) officially launched in 2022 to meet the challenges of increasing scale of data, introduction of national legislations or other legal issues that potentially restrict data movement. The current state of the FEGA is described in deliverable E3.3 Informe Final de un Nodo EGA[27].

The **Life Science Authentication and Authorization Infrastructure (LS AAI)**[28] provides a secure and federated system for authenticating and authorising researchers accessing sensitive genomic and health data. LS AAI ensures that only authorised individuals can access specific datasets, enhancing data security and compliance with ethical standards. This infrastructure supports the seamless integration of genomic data across different research platforms and institutions.

The **Genome Standards Consortium (GSC)**[29] is an international community-driven effort to develop and promote standards for genome sequence data. GSC's primary goal is to ensure that genomic data is consistently described and annotated, making it more accessible and useful for the research community. The consortium's standards facilitate the integration and comparison of genomic data from different sources.

The **Human Genome Variation Society (HGVS)**[30] focuses on the accurate and standardised description of genomic variants, characterization of genomic variations including population distribution and phenotypic associations. HGVS develops nomenclature guidelines for reporting genetic variations, ensuring that researchers and clinicians can communicate findings clearly and consistently. This standardisation is critical for the accurate interpretation of genetic data in both research and clinical settings.

[25] https://www.european-health-data-space.com/
[26] https://ega-archive.org/
[27] https://b2drop.bsc.es/index.php/f/3248200
[28] https://lifescience-ri.eu/ls-login.html
[29] http://www.gensc.org//
[30] https://www.hgvs.org/

The **American College of Medical Genetics and Genomics (ACMG)**[31] is a professional organisation that provides practice guidelines and standards for the clinical application of genomic data. ACMG's recommendations are widely adopted in the medical genetics community, helping to ensure the accuracy and reliability of genetic testing and the interpretation of results.

The **Genomic Data Commons (GDC)**[32] is an initiative by the National Cancer Institute (NCI) to provide a unified data repository for cancer genomics data. GDC supports data standards that facilitate the integration and analysis of genomic data across various cancer research projects. By providing a comprehensive and accessible data platform, GDC enhances the ability of researchers to discover new insights into cancer biology and treatment.

The **Genome in a Bottle Consortium (GIAB)**[33] is a public-private-academic consortium hosted by the National Institute of Standards and Technology (NIST). GIAB aims to develop the technical infrastructure necessary to enable the translation of whole human genome sequencing into clinical practice and foster innovations in genomic technologies. The consortium's priority is the authoritative characterization of human genomes for use in benchmarking. This includes analytical validation, technology development, optimization, and demonstration. GIAB's efforts are vital for providing reference standards, reference methods, and reference data that ensure the accuracy and reliability of genomic sequencing. By benchmarking and validating sequencing methods, GIAB helps set the groundwork for clinical applications of genomic data, facilitating advancements in personalised medicine.

These organisations and initiatives are interconnected through their shared goals of improving genomic data standards and facilitating data sharing. They work collaboratively to develop guidelines, tools, and infrastructures that support the seamless integration of genomic data across different platforms and institutions. The state of the art in genomic data standards is characterised by a strong emphasis on interoperability, security, and privacy, ensuring that genomic data can be used effectively and responsibly in both research and clinical settings.

The integration of these efforts has led to significant advancements in our understanding of the human genome and its implications for health and disease. By adhering to these standards, researchers and clinicians can collaborate more effectively, leading to improved patient outcomes and accelerated scientific discoveries. The continuous evolution of these standards is crucial for keeping pace with the rapid advancements in genomic technologies and their applications.

1.3.3 Germinal and Somatic Variant Calling

In genomic analysis, distinguishing between germinal and somatic variants is crucial. **Germinal variants** are inherited and present in every cell of the body, playing a role in

[31] https://www.acmg.net/
[32] https://gdc.cancer.gov/
[33] http://www.genomeinabottle.org/

inherited diseases and conditions. **Somatic variants**, on the other hand, are acquired mutations that occur in individual cells during a person's lifetime and can lead to diseases such as cancer. Accurate calling of these variants is essential for understanding their implications in health and disease.

1.3.4 Standard File Formats in Genomic Data

Various standard file formats are used in genomic data to ensure consistency and reliability in data storage and analysis:

- **BCL files**: These are raw data files generated by sequencing machines, containing base call data.
- **FASTQ/ORA**: FASTQ files store both nucleotide sequences and their corresponding quality scores. ORA is a Illumina proprietary compressed version of FASTQ that retains all the information but in ⅕ smaller file size.
- **SAM/BAM/CRAM**: The Sequence Alignment/Map (SAM) format is a text-based format for storing biological sequences aligned to a reference genome. BAM is the binary version of SAM, providing the same information in a compressed format. CRAM is another compressed format that provides more efficient storage by also compressing the reference sequence.
- **BED**: The Browser Extensible Data (BED) format is used to store genomic regions, often representing specific features or annotations.
- **VCF**: The Variant Call Format (VCF) is used to store gene sequence variations. This format includes information about the genomic location, reference allele, and observed variants.

1.3.5 Annotation of Variants

The annotation of genomic variants is a critical step in genomic analysis. It involves defining the genomic position of variants and predicting their possible effects. Annotation helps identify which variants are likely to be benign and which might be pathogenic, thereby aiding in the interpretation of genomic data in a clinical context. Tools and databases such as Ensembl, dbSNP, and ClinVar are commonly used for variant annotation, providing comprehensive information about known genetic variations and their associated clinical significance.

Normal Workflow for Genomic Variant Calling Analysis

1. **Genome Alignment**: Raw sequencing data in fastq/ORA files is aligned to a reference genome (GRCh38) using tools like BWA or HISAT2. Future workflows aim to utilise the HPRC or T2T references as algorithms are developed to support these comprehensive genomes.
2. **Alignment Co-Cleaning**: Processes such as IndelRealigner and BaseRecalibration (GATK) are employed to correct misalignments and recalibrate base quality scores.

IndelRealigner adjusts alignments around insertions and deletions, while BaseRecalibration recalibrates the quality scores of the sequenced bases.

3. **Somatic or Germinal Variant Calling**: Tools like GATK's HaplotypeCaller, MuTect2, and DRAGEN VCF identify somatic mutations (cancer-specific) and germline variants (inherited). These tools analyse the aligned sequences to detect variations from the reference genome.
4. **Variant Annotation**: Annotation tools such as VEP, ANNOVAR or SnpEff add functional information to the identified variants, indicating their potential impact on gene function and disease.
5. **Variant Aggregation**: Variants are aggregated to identify common mutations and patterns across samples. This step helps in understanding the genetic landscape of the studied population.
6. **Aggregated Variant Masking**: Filtering processes remove low-confidence variants and artefacts, ensuring that only high-quality, reliable data is used for further analysis.

Workflow Hubs and Pipelines for NGS Data Analysis

The IMPaCT-Data program leverages advanced workflow hubs and pipelines to enhance NGS data analysis for personalised medicine. All these workflows and tools are discoverable through the ELIXIR consortium, which provides access to a vast repository of bioinformatics resources and training materials, ensuring that researchers have the necessary tools and knowledge to conduct high-quality genomic analyses. Key platforms include:

- **WorkflowHub**[34]: An open platform supporting the sharing and execution of computational workflows. WorkflowHub enables researchers to share their analysis pipelines, ensuring reproducibility and fostering collaboration. It hosts a variety of workflows that can be adapted for different genomic studies.
- **nf-core**[35]: A community-driven project providing high-quality bioinformatics workflows. Built using Nextflow, nf-core pipelines are designed for scalability and reproducibility across different computing environments. These workflows cover a wide range of applications, from RNA-seq to whole-genome sequencing.
- **Sarek**[36]: An nf-core pipeline specifically designed for the analysis of whole-genome sequencing data. Sarek handles the complete workflow from raw data to variant calling and annotation, making it a robust tool for cancer and germline variant analysis. It integrates multiple tools for quality control, alignment, variant calling, and annotation, ensuring comprehensive data processing.

IMPaCT-Data has been actively working on improving these pipelines to ensure they meet the highest standards for accuracy, efficiency, and interoperability. By integrating these advanced tools, the program supports the seamless analysis and sharing of genomic data, paving the way for significant advancements in personalised medicine.

[34] https://workflowhub.eu/
[35] https://nf-co.re/
[36] https://nf-co.re/sarek/3.2.3/

IMPaCT Data · Barcelona Supercomputing Center Centro Nacional de Supercomputación · Gobierno de España · Ministerio de Ciencia, Innovación y Universidades · Instituto de Salud Carlos III · Financiado por la Unión Europea NextGenerationEU · Plan de Recuperación, Transformación y Resiliencia

The implementation of the Sarek nf-core pipeline within the IMPaCT-Data QC framework exemplifies these efforts, offering a standardised, efficient tool for genomic data processing that supports high-quality analysis and integration into clinical practice. This is a product of a combined effort amongst the groups and is based on the recommendations, needs and expertise of the genomics analysis experts in IMPaCT-Data. The workflow is designed for WES germline samples and a total of 40 metrics were identified and implemented using various tools. Complete details can be found in the supporting documentation, along with the pipeline itself, in the EGA github[37]. Moreover, the workflow has been registered in WorkflowHub and is freely searchable and usable by anyone[38].

1.4 Implementation of Standards in Healthcare Systems

Implementing these standards in healthcare systems requires a multifaceted approach. Training and education are essential to ensure that healthcare and research professionals are well-versed in using these standards. Continuous professional development and specialised training programs can help bridge knowledge gaps and foster a culture of standardisation. Technology integration is another critical component, involving the adoption of software and systems that support standardised data formats and communication protocols. Healthcare institutions need to invest in technologies that can seamlessly incorporate these standards into their workflows, thereby enhancing data interoperability and quality.

Moreover, establishing organisational policies and governance structures is vital for enforcing the use of standards and maintaining data quality. Clear guidelines and robust governance frameworks can help ensure that data is managed consistently and securely across the organisation. Compliance with regulatory requirements and industry best practices is also necessary to protect patient data and uphold ethical standards.

By adhering to these standards, healthcare institutions can improve the interoperability of their systems, enhance data quality, and facilitate more effective research collaborations. This standardised approach is critical for advancing personalised medicine and achieving the goals of initiatives like IMPaCT-Data. Standardisation not only supports better patient outcomes but also drives innovation and efficiency in healthcare delivery and research.

1.4.1 Healthcare and genomics standards in action

Figure 1 exemplifies a desirable data flow in healthcare institutions to manage the need for secondary use of the data for decision support or research, e.g. It is a general overview of operational and decision making systems roles and relationships.

The core data comes from the electronic health records (EHR) management system, which is used in day-to-day transactional operations, and that is not designed for the heavy queries

[37] https://github.com/EGA-archive/sarek-IMPaCT-data-QC
[38] https://workflowhub.eu/workflows/1030?version=2#projects

required for decision making processes. In the figure, the hospital's EHR is named IMASIS. As mentioned in the previous section, the genetic or genomic information coming from diagnosis assays is typically handled outside of the main EHR system, which is exemplified as “curated VCF/Omics data” in the figure. In order to facilitate clinicians’ interpretation of genomics data, this data is loaded into a cBioPortal instance, a dashboard-like tool for the visualisation and analysis of cancer patients data (cBioPortal is described in more detail below). These elements are hosting non-anonymized data for primary use.

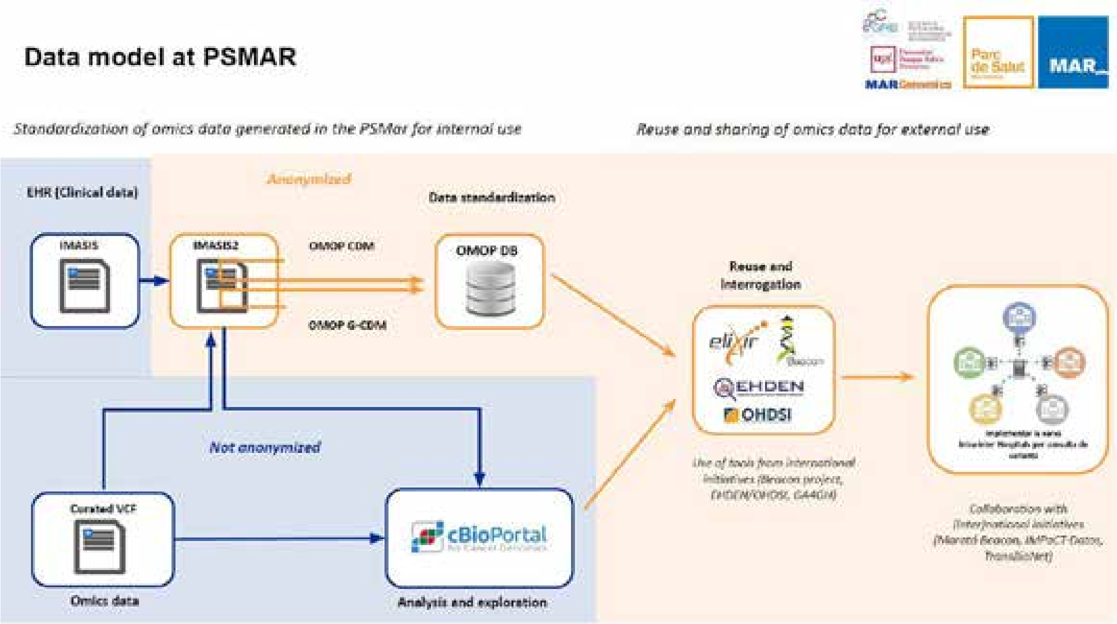
Data model at PSMAR
MARGenomics
Parc de Salut
MAR_
Standardization of omics data generated in the PSMar for internal use
Reuse and sharing of omics data for external use
EHR (Clinical data)
Anonymized
Data standardization
IMASIS
IMASIS2
OMOP CDM
OMOP G-CDM
OMOP DB
Reuse and Interrogation
elixir
Beacon
EHDEN
OHDSI
Use of tools from international initiatives (Beacon project, EHDEN/OHDSI, GA4GH)
Not anonymized
Curated VCF
Omics data
cBioPortal
Analysis and exploration
Collaboration with (inter)national initiatives (Marató Beacon, IMPaCT-Data, TransBioNet)

Figure 1 Data model at Parc de Salut Mar (PSMAR)

Data flows from the master EHR into an EHR copy which is pseudonymised in further steps. The figure exemplifies that, in some cases, the data in the EHR copy is leveraged back into the main system, in this case to the cBioPortal instance. Leveraging a decision support system to feed back operational systems is also a common business practice.

Although healthcare institutions rely on the central EHR for most of their operations, different hospital’s departments could be using specialised information management systems. A paradigmatic case would be image-based diagnose services (e.g. radiology) or biochemical analyses services. Generally, in order to allow analyses from heterogeneous data sources, data from these systems is transformed into a common data model. In the figure, OMOP-CDM (described above) and the OMOP extension for genomic data (OMOP G-CDM) were chosen.

From that point, pseudonymised or anonymised data could be used for research, and further leveraging the fact that it is hosted in a common data model to use available tools for that CDM.

In IMPaCT-Data, both OMOP and cBioPortal have been extended with a GA4GH Beacon API interface (this is described in more detail in the following sections), facilitating the discovery of available data both internally in the hospital or externally while keeping the security and privacy of sensitive data.

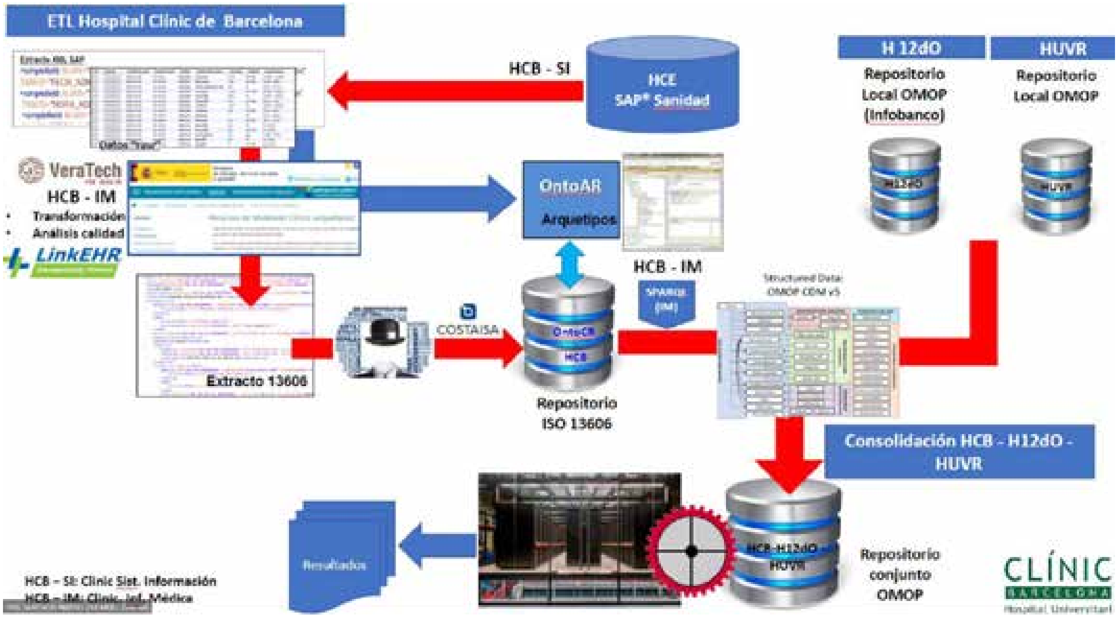

Figure 2. Implementation of standards in local healthcare systems

Figure 2 shows another example of how this common approach was leveraged by several hospitals (Hospital Clinic de Barcelona - HCB, Hospital 12 de Octubre - H12dO, and Hospital Universitario Virgen del Rocio - HUVR) to do a joint data analysis. Data is extracted from the respective EHR (or Historia Clínica Electrónica - HCE) into the OMOP-CDM, then consolidated in a secure processing environment, hosted by the Barcelona Supercomputer Center (BSC), in order to run the analysis.

The Hospital Clinic case is shown in higher detail as they perform a two step process: transforming from the EHR into archetypes (ISO13606) and from there to different destinations, OMOP-CDM among them.

2 IMPaCT-Data approach to the problem

The IMPaCT-Data approach is to leverage standardisation at as many levels as possible. WP3 has focused on the genomic standards, while WP4 has focused on both clinical and image standards. Recommendations for them could be found in the respective deliverables

IMPaCT Data
Barcelona Supercomputing Center Centro Nacional de Supercomputación
Gobierno de España Ministerio de Ciencia, Innovación y Universidades
Instituto de Salud Carlos III
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

(E4.1 Normas internacionales de información de HCE[39], E4.4 Normas Internacionales de Anotación de Información de Imagen Médica[40]).

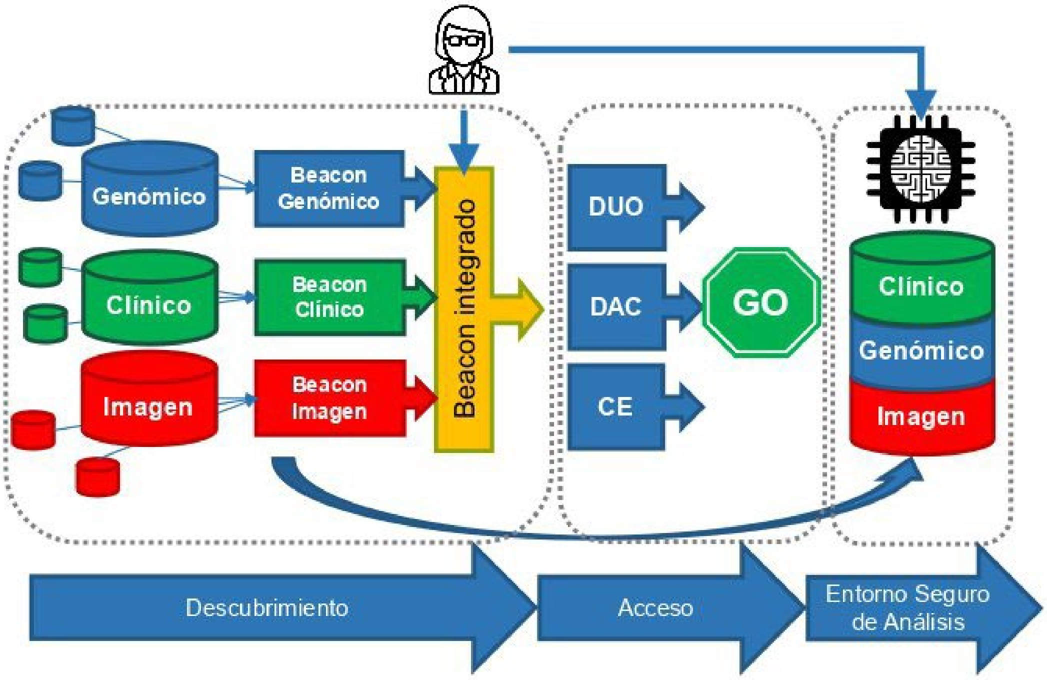

Figure 3. The IMPaCT-Data Platform proposal

However, just using the recommended standards is not enough to fulfill the personalised medicine needs. Decisions, assisted decisions or knowledge that helps to move forward in diagnostic or therapeutic processes require information that should be, first, located and, second, described in a meaningful way that allows their consumption. If the required information requires specific authorisation, the process to request access should be easily found and actionable.

WP3 and WP5 have worked together to develop a uniform way for the user to locate the relevant information. It is based on the GA4GH Beacon v2 standard[41] [42]. Beacon allows the discovery of sensitive data or knowledge obtained from it without disclosing any detail about such data. Beacon also includes functionalities that allow bridging the gap between locating data and requesting access to it.

Beacon v2 leverages standard vocabularies or ontologies to harmonise the questions sent to different beacons. For example, if two institutions have transformed their data to OMOP, Beacon would be able to use the OMOP vocabulary (the OMOP concepts) to dispatch the

[39] https://b2drop.bsc.es/index.php/f/2771857
[40] https://b2drop.bsc.es/index.php/f/2794024
[41] https://genomebeacons.org/
[42] https://doi.org/10.1002/humu.24369

same query to both. If an institution has standardised on ICD-10, e.g., and not OMOP, Beacon could leverage the mapping between both dictionaries to send each beacon a query adjusted to the supported vocabulary (ICD-10 or OMOP concept).

For genomic queries, the Beacon v2 specification includes recommendations on how to harmonise the data, with the data being stored as VCF or as any other format. Genomic variation representation is also harmonised two ways: 1) using a popular legacy approach or 2) using the GA4GH variant representation specification (VRS)[43].

For clinical data, which has been used as an example a few paragraphs above, the expectation is that each institution transforms their data into OMOP, and then uses the Beacon for OMOP[44] developed by BSC and CRG on top of OMOP-CDM v5.

For imaging data, two approaches have been discussed and one has been piloted (Note: in this text we use “imaging” as a synonym of “radiology” and “radiomics” (see Box 3[45]). Although GA4GH Beacon is focused on genomic and clinical data, it is designed to be extensible, and imaging is a natural extension to the specification itself. We leveraged the first approach (OMOP based, see below) to draft a radiomics (imaging) extension to GA4GH Beacon. Using that extension draft, we tested the second approach.

Box 3: What is Radiomics[45]?

- Radiomics refers to a range of techniques used to extract quantitative features from medical images, aiming to enhance the accuracy of diagnosis, prognosis, and prediction.
- By extracting and analysing the spatial distribution of signal intensities and pixel relationships, radiomics uses AI-based techniques to quantify textural information.
- Numerous studies across different imaging disciplines have demonstrated its potential to improve clinical decision-making.

The first imaging approach is to leverage the proposed radiology extension for OMOP-CDM, simply load the data into the OMOP-CDM, and query it as if it was a part of the CDM itself, using the same Beacon for OMOP technology cited above. As the suggested OMOP extension has 5 main elements (i.e: occurrences, features, conditions, measurements and devices), we created these additional endpoints as a Beacon extension. As, at the time of writing, we did not have any OMOP-CDM extension prototype available, we have not been able to pilot it.

[43] https://docs.genomebeacons.org/formats-standards/#variant-representation-standard-vrs

[44] https://gitlab.bsc.es/impact-data/impd-beacon_omopcdm

[45] https://doi.org/10.1186/s13244-020-00887-2

However, as a second approach, we loaded the imaging data into a database (mimicking the OMOP extension tables), developed the additional endpoints into the CRG Beacon Reference Implementation and tested it[46]. As this aspect is still immature, and the radiomics community has not reached a consensus yet, no further details will be provided in this document.

For genomic data there is no formal OMOP extension, therefore, we could not leverage OMOP to have all three data types. The OMOP G-CDM extension is still in development and not formally approved by OHDSI.

Having all the relevant data types (clinical, genomics, and imaging) available in the same system (Beacon), facilitates doing queries across the different data types, without having to send queries to disparate systems (e.g. OHDSI Atlas for phenoclinical data and Beacon for genomics data). To bridge the gap between beacons focused in one or two of the data types, an overarching beacon could be needed. This would be the only one visible to the end user or to the beacon networks (see below).

Using Beacon has an additional and critical benefit: Beacon is designed to be organised in networks. Beacon networks are communities of beacons that have something in common (e.g. focusing on a disease type, on data from a region, etc.) and that could be queried at once through the network. A simplified process could be as follows:

1. The user visits the Beacon network web page
2. The user posts the query
3. The query aggregator behind the web page dispatches the query to every beacon instance that is part of the network
4. The aggregator joins and organises the responses from the beacons and presents them to the user

This way, a user could get answers from multiple beacons in a single operation and could discard all those that have no data of interest, while focusing on the promising ones.

Using GA4GH Beacon as a data locator (or data discovery) standard is aligned with different initiatives, in particular the 1+Million Genomes, the European Genomic Data Infrastructure (GDI), the European Joint Program on Rare Diseases (EJP-RD), EOSC4Cancer, EUCAIM, Federated European Genome-phenome Archive (FEGA), etc. making IMPaCT-Data approach compatible with all these initiatives (Figure 4).

[46] https://github.com/EGA-archive/beacon2-ri-api-images

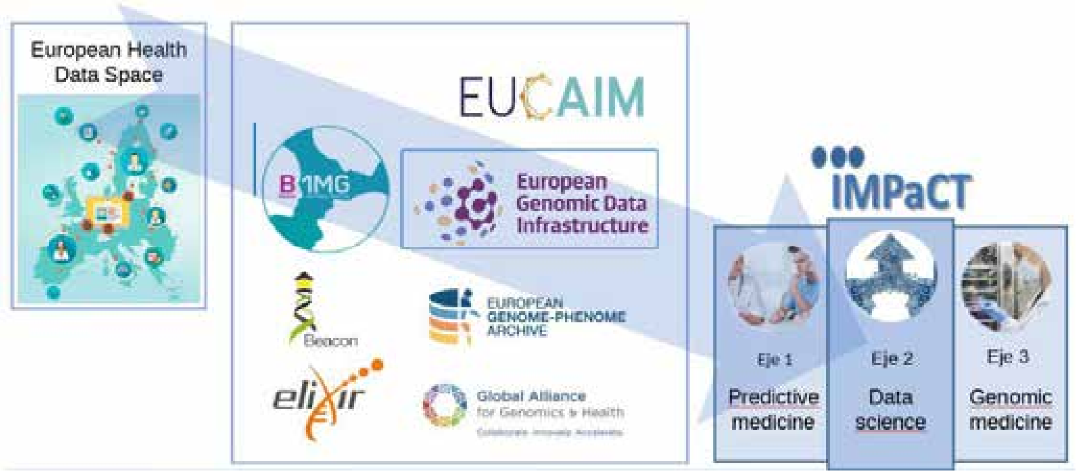
European Health Data Space
EUCAIM
B1MG
European Genomic Data Infrastructure
Beacon
EUROPEAN GENOME-PHENOME ARCHIVE
elixir
Global Alliance for Genomics & Health
IMPaCT
Eje 1
Predictive medicine
Eje 2
Data science
Eje 3
Genomic medicine

Figure 4. Alignment across international projects and initiatives

The text above describes how, through Beacon, data is made available to be discovered or how knowledge is available to be consumed. In order to make Beacon useful, the relevant data should flow from the original heterogeneous sources into the Beacon components or should be connected to Beacon. The next sections provide more details about these processes.

2.1 Making clinical data available to Beacon

Beacon is designed to be a programmatic interface on top of existing solutions, such as applications like RD-Connect GPAP[47] or cBioPortal[48], services like HL7 FHIR API, or databases like OMOP-CDM. Alternatively, Beacon implementations like the CRG Beacon Reference Implementation provide a complete solution that allows to light a beacon indeed if your data is not already available in any structured format.

How does this apply to clinical data in the IMPaCT context? Along this document we have mentioned that OMOP-CDM is the IMPaCT-Data recommended standard and, also, that we assume that for any IMPaCT project the clinical data is expected to be available in such format. Therefore, in these cases, Beacon for OMOP could be deployed, avoiding any further data processing or transformation. Additionally, every time that the OMOP data is updated, the associated beacon would make this new data available. Describing how the data is converted to OMOP is outside of this deliverable scope, however an example is shown in Figure 2.

In cases where there are no short term plans to provide the data in OMOP-CDM, an alternative mechanism is provided: loading the data into the Beacon Reference Implementation using

[47] https://platform.rd-connect.eu/
[48] https://www.cbioportal.org/

the Beacon Reference Implementation tools[49]. This mechanism is simple enough: make your data available in CSV format following a template provided with the tools and run a process to load it into a MongoDB database included in the Reference Implementation.

2.2 Making genomic data available to Beacon

The IMPaCT-Data program is dedicated to accelerate personalised medicine through the efficient processing and sharing of genomic data. This section delves into the genomic data workflow, from raw data processing to variant calling, and illustrates how this data integrates into the Beacon platform and cBioPortal for sharing and further analysis, fully adhering to international standards set by organisations such as GA4GH, GDI, EHDS, GSC, and GDC (see section 1.3.2).

Genomic data processing begins with the sequencing of DNA, generating raw data in the form of BCL files. These files are converted into fastq/ORA files, which contain the nucleotide sequences and their corresponding quality scores. The next steps involve aligning these sequences to a reference genome using tools like BWA (Burrows-Wheeler Aligner) or HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts). The alignment is typically performed against the HGP GRCh38 reference genome, with future transitions planned towards the more comprehensive HPRC (Human Pangenome Reference Consortium) or T2T (Telomere-to-Telomere) genome references, pending the development of algorithms capable of handling these new standards.

Once aligned, the data undergoes alignment co-cleaning processes to ensure high accuracy. This includes steps like IndelRealigner, which corrects misalignments around insertions and deletions, and BaseRecalibration, a process in GATK (Genome Analysis Toolkit) that adjusts the base quality scores based on known variants, improving the accuracy of the sequence data.

The aligned sequences are then subjected to variant calling, where tools such as GATK's HaplotypeCaller and MuTect2 or DRAGENVCF are used to identify both somatic and germinal variants. These tools analyse the sequence data to detect variations from the reference genome, distinguishing between inherited (germline) and acquired (somatic) mutations.

After variant calling, the resulting data, typically stored in VCF files, is ready for sharing and further analysis. Beacon Reference Implementation although other solutions are available, like RD-Connect GPAP (also mentioned above). As with the clinical data described above, genomic data flows to Beacon using the Beacon v2 Reference Implementation Tools, a toolset for populating the Beacon RI from VCFs (and phenoclinical information).

Technically speaking, the process is quite simple, run the RI Tools script to parse the desired VCFs, this will generate a Json file that is then loaded into the MongoDB database that supports the Beacon RI. All in one step. Some of the annotations in the VCF are also copied

[49] https://github.com/EGA-archive/beacon2-ri-tools-v2

into the Json, but not all given the broad diversity of annotations that VCF related tools produce. The CRG Beacon team is currently working on providing better tooling for moving desired annotations into the beacon instance.

However, similar to the phenoclinical data, if the data is already available in a platform (e.g., cBioPortal, CSVS or GPAP), the preferred solution could be to leverage that platform instead of duplicating the data into the Beacon RI backend.

2.3 cBioPortal use-case

An exemplary use case within the IMPaCT-Data framework is the integration with cBioPortal, an open-source platform designed for the visualisation and analysis of cancer genomic data. Data from genomic analyses conducted outside of clinical care, such as research studies, can be shared through cBioPortal. This integration makes clinically relevant genomic data accessible to a broader research community, facilitating the discovery of novel insights into cancer genomics and personalised treatment strategies.

We explored the use of cBioPortal and its integration with Beacon as part of WP3. The main issue with cBioPortal is a lack of interoperability across instances. Within specific cancer consortiums or projects, such as TCGA and ICGC, a good degree of commonality in the variables, ontologies, codes etc is observed. However, cBioPortal offers almost complete flexibility in choosing variables, with very few mandatory fields, and this leads to many differences when comparing available variables across different projects, consortia, studies, etc. Therefore, as part of our exploration of cBioPortal, we compared various cancer models, including TCGA, ICGC, and the newly developed cancer model from B1MG, to identify a minimal set of common variables amongst these popular cancer models. We also included in this comparison Phenopackets, a GA4GH open standard to share disease and phenotype information. This standard is the reference model for the representation of the clinical information in Beacon. Although it is not disease specific, it contains relevant fields to collect cancer-related data. Our findings and recommendations have been compiled in a recommendations document[50]. Herein we also describe a technical proof-of-concept following these recommendations, using publicly available data. A beacon was developed to deploy directly on top of a cBioPortal instance and this was tested in two sites. By following these recommendations, one can create a cBioPortal instance that would seamlessly enter a Beacon network.

In conclusion, the systematic approach to genomic data processing and sharing within the IMPaCT-Data framework exemplifies how advanced bioinformatics tools and collaborative platforms can drive forward personalised medicine. Through initiatives like Beacon and cBioPortal, IMPaCT-Data not only enhances the accessibility of genomic data but also ensures its quality and reliability, fostering a more collaborative and effective research environment. The commitment to adhering to international standards and using cutting-edge

[50] https://b2drop.bsc.es/index.php/f/3125209

technology ensures that IMPaCT-Data remains at the forefront of genomic research and personalised healthcare.

3 Beacon pilots in IMPaCT-Data

As part of IMPaCT-Data, so far 12 centres have taken on the challenge of being a pilot use-case for Beacon deployment and establishing the beginning of a Beacon Network that includes the IMPaCT Beacons[51]. The 12 centres, listed in Table 1, are at various levels of readiness. The technical details of the IMPaCT Beacon Network have been described as part E5.6[52]. In this deliverable, we focus on the challenges faced and the experiences of each centre in the process of deploying a Beacon. To this end, we asked each pilot to fill a short survey describing their experience thus far (the full results can be seen in the Annex.

Table 1. Beacon pilot participating centres

#	Centre	Syn	Real	URL?
1	Centro Nacional de Análisis Genómico (CNAG)	x		y
2	Navarrabiomed & NASERTIC	x	x	y
3	Institut Germans Trias i Pujol (IGTP)		x	n
4	IIS La Fe	x		y
5	Fundación Progreso y Salud (FPS)		x	y
6	Sant Joan de Déu Research Institute (IRSJD)	x		n
7	Institut d'Investigació en Atenció Primària Jordi Gol (IDIAPJGol)	x		n
8	Centro Nacional de Investigaciones Cardiovasculares (CNIC)		x	y
9	Biobizkaia	x		y
10	Fundación Pública Galega de Medicina Xenómica (FPGMX)		x	n
11	Centro Nacional de Investigaciones Oncológicas (CNIO)		x	n
12	Hospital del Mar and Hospital del Mar Research Institute (IMIM)	x		n

Syn = synthetic data; Real = real data; URL available? yes (y) or no (n); last updated 20-09-2024

The diversity across centres is apparent, but similarities exist. Most are research centres linked to hospitals, therefore performing diagnostic and research activities. Some centres (FPS, CNAG and CNIC) already had experience with Beacon, while most other centres were never exposed to the Beacon protocol prior to this. Additionally, three centres (CNAG,

[51] https://impact-beacon-network-demo.ega-archive.org/
[52] https://b2drop.bsc.es/index.php/f/3248197

Navarrabiomed & NASERTIC and FPGMX) are also the sequencing centres that participate in IMPaCT-Genomica[53]. These centres will use GPAP for IMPaCT-Genomica data and Beacons will be deployed separately with the aim of making IMPaCT-Genomica data discoverable.

In terms of challenges faced, the biggest hurdles seemed to be ELSI related, rather than technical. Most centres (7 of 12) reported Beacon was technically easy to deploy, although two mentioned it was a challenge learning the new technology. However, getting ethical or legal approval is a process that will take some time to resolve in most cases. Therefore, most centres (7 of 12) deployed a Beacon with synthetic data first, as a proof of technical readiness, while awaiting relevant approvals. Three centres have already deployed Beacons with real datasets (FPS, Navarrabiomed & NASERTIC, CNIC) and have the URLs publicly available as part of the IMPaCT Beacon Network.

Despite the challenges, most centres reported expected benefits of deploying a Beacon, such as accelerating discovery, fostering collaboration, exploring feasibility of secondary use, integration of EHR data, and increased visibility and interoperability in datasets. Indeed, some technical challenges mentioned included lack of interoperability in datasets, data curation and validation and dealing with large datasets. These could be overcome by implementing rigorous protocols and pipelines, meticulous data mapping and standardisation. Once such procedures are in place, they can be applied more easily to future datasets, further increasing interoperability.

Going forward, all centres plan to deploy Beacons on real data eventually, although some face more challenges than others. In the case of FPGMX, they will return only Boolean responses, as a way to overcome some of the ethical concerns, while others will only share aggregated responses or implement a subset of options, for the same reasons. Overall, there is consensus on the benefits of sharing data via Beacon and most, if not all, plan to add more datasets in the future and expand the IMPaCT Beacon Network. Additionally, more centres as part of IMPaCT, apart from the 12 pilots mentioned here, are also working towards implementing their own Beacons, therefore the network is expected to continue to grow.

4 Concluding remarks

In IMPaCT-Data WP3 we have focused on the management and discoverability of the genomic data. Clinical and imaging data were included in this document and in some of the performed tasks as they are part of the discoverability aspect and they share the same tooling (Beacon) as the genomic data.

In order to better understand the landscape of genomic data that the Spanish health care system is actually managing we conducted a series of surveys and working meetings, where

[53] https://genomica-impact.es/

this landscape has been profiled. The results are part of the deliverable E3.4 Genomic Analysis in Healthcare Environments (also described above in the Overview section).

To test how much the experience of some partners is portable to other partners, they selected a use case: quality control in the genomic pipeline that has been developed as an extension of the popular Sarek pipeline. These institutions that do not use Sarek could still benefit from the gathered knowledge and from the code itself as a reference to adopt it to their own environments. The pipeline extension has been published in a public repository, hence available to anyone inside and outside IMPaCT-Data.

To exemplify how to leverage popular sources to make the data discoverable, we have developed an implementation of Beacon on top of the OMOP-CDM and another on top of cBioPortal. We also provided a guideline on how to make cBioPortal more interoperable.

Moreover, a proof of concept for a Beacon for radiomics has been developed and discussed. It would be the subject of future work.

As the final goal of WP3, a series of pilots has been started in 12 different institutions and a demo of a Beacon Network has been deployed. These include the centres for IMPaCT-Genomica, although separate Beacons will be deployed specifically for the data of that project.

Altogether, the outcomes of WP3 is a set of tools, examples and documents that could be used directly as they are, could be used as reference for personalised solutions, or be the basis for further developments.

Acronyms and Abbreviations

In the following table there are some acronyms and abbreviations used in the deliverable

BED	Browser Extensible Data
CDM	Common Data Model
DICOM	Digital Imaging and Communications in Medicine
EHR	Electronic Health Records
ELSI	Ethical, Legal, Social Implications
FHIR	Fast Healthcare Interoperability Resources
FISH	Fluorescence in situ hybridization
GPAP	Genome-phenome analysis platform
HL7	Health Level Seven International
ICD	The International Classification of Diseases
NGS	Next Generation Sequencing
OMOP	Observational Medical Outcomes Partnership
PACS	Picture Archiving and Communication System
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
VCF	Variant Call Format
WES	Whole Exome Sequencing
WP	Work Package

Annex

IMPaCT-Data Use-cases for Beacon deployment

https://impact-beacon-network-demo.ega-archive.org/

Template to fill:

1. Name of Institute

2. Type of institute (hospital, research centre, clinical research centre, other?)

3. Name(s) of participants (and roles)

4. Briefly describe your institution’s profile (eg. mission, size, expertise)

5. Briefly describe your experience with deploying Beacon
   For example:
   a. Expected benefits
   b. Challenges faced and how you have adapted
   c. Choosing pilot data
   d. Plans for future beacons

6. Beacon URL

USE CASE 1: CNAG	
Name of Institute	RD-Connect Genome-Phenome Analysis Platform (GPAP; https://platform.rd-connect.eu), hosted by Centre Nacional d’Anàlisi Genòmica (CNAG; https://www.cnag.eu/)
Type of institute	CNAG: Research centre
Name(s) of participants (and roles)	Anastasios Papakonstantinou (Software Engineer), Davide Piscia (Lead Software Engineer), Sergi Beltran (Head of the Bioinformatics Unit)
Briefly describe your	CNAG’s Mission: “To carry out projects in genome analysis that will lead to significant improvements in people’s health

institution’s profile (eg. mission, size, expertise)	and quality of life, in collaboration with the Catalan, Spanish, European and International research and clinical community.” The RD-Connect GPAP is an IRDiRC recognised resource supporting collaborative research on genome-phenome data to accelerate rare disease diagnosis and gene discovery.
Briefly describe your experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	- Already deployed beacon v1 for https://beacon-network.org/#/ when it came out - Deployed a Beacon v2 based API for EJP-RD Virtual Platform (https://vp.ejprarediseases.org/). Specs are not fully compatible with EJP-RD - Currently implementing 3 endpoints on RD-Connect GPAP (g_variants, biosamples, individuals) for IMPaCT-Data and ELIXIR Beacon Network - Implementation on production data; return of counts; “range” (or upper range value) when less than X results.
Beacon URL	https://playground.rd-connect.eu/beacon2/api

USE CASE 2: NB & NASERTIC	
Name of Institute	Navarrabiomed (NB) & NASERTIC Sequencing Center
Type of institute	Public biomedical research center & Company of the Navarra Public Business Corporation (CPEN) group, and as a public entity belonging to the Government of Navarra (Sequencing Center)
Name(s) of participants (and roles)	María L Mansego Talavera (NB) - deployer Igor Ruiz de los Mozos (NASERTIC) - Lead Bioinformatics and Sequencing Unit Fernando Alvira Iraizoz (NB) – project manager Sara Ciria (NB) – data manager / data curation Josune Hualde Olascoaga (HUN) – NAGENpediatrics Principal Investigator Gonzalo Rodriguez Ordóñez (NASERTIC) - Director of Personalized Medicine Ángel Alonso (NB) – Principal Investigator of Genomics Medicine Unit
Briefly describe your	The partnership between Navarrabiomed and *Navarra de*

institution’s profile (eg. mission, size, expertise)	*Servicios y tecnologías* (NASERTIC) merges biomedical research with information technology, enhancing healthcare through advanced data analysis and technology utilization. Navarrabiomed focuses its efforts towards the implementation of advanced biomedical research in the Public Health System. NASERTIC offers supercomputing capacity and NGS services to all public local institutions, including research centres and the hospital; these resources allow accelerating scientific development. Navarrabiomed expertise in cutting-edge research in the field of precision medicine is complemented by NASERTIC proficiency in telecommunications and high-tech services such as high-throughput sequencing and supercomputing. Together, they expedite the development of innovative therapies, fostering interdisciplinary collaboration and knowledge exchange for improved healthcare outcomes.
Briefly describe your experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	Deploying Beacon involved utilizing the federated sharing of genomic and phenotypic data, leveraging the Beacon v2 Reference Implementation by Rueda et al. (2022). Initially, we worked with synthetic data using the CINECA dataset as a reference. Once the infrastructure was deployed, we transitioned to the NAGENPediatrics dataset (758 WGS samples) for our pilot study, with plans to eventually incorporate all available data from the NAGEN Program (NAGEN | Navarrabiomed). In order to include this dataset into the pilot study, we needed approval from the local Ethical Committee (CEIm). The process was relatively straightforward, although we had to go through the process twice since we decided to include additional variables shortly after setting off with the real-data use-case. Expected benefits: We envisaged that deploying Beacon would facilitate seamless sharing of genomic and phenotypic data among different institutions or research groups, enabling efficient collaboration and accelerating scientific discoveries in the field of genomics medicine and paediatrics. Challenges faced and adaptations: We encountered interoperability issues integrating diverse datasets, necessitating meticulous data mapping and standardization efforts. Managing millions of genomic variants posed a significant challenge, requiring efficient algorithms and data processing pipelines. Additionally, handling queries involving

	this vast number of variants demanded optimization of query performance and resource allocation within the Beacon infrastructure. Through iterative refinement, we adapted to manage these challenges effectively and ensure Beacon operation. Choosing pilot data: After setting up the infrastructure using synthetic data, we selected the NAGENpediatrics dataset for our pilot study due to its potential to yield valuable insights into paediatric genomics. This dataset offered a diverse range of genomic and phenotypic information (HPO), making it suitable for testing the effectiveness of our Beacon implementation. On top of that, using the NAGENpediatrics dataset forced us to deal with the anticipated issues associated with the use of real data. Plans for future beacons: Moving forward, we plan to expand our Beacon infrastructure to incorporate additional datasets from the NAGEN studies, thereby enhancing the breadth and depth of genetic and phenotypic information available for analysis.
Beacon URL	impact-beacon.nasertic.es/api/info impact-beaconped.nasertic.es/api/info

USE CASE 3: IGTP	
Name of Institute	Institut Germans Trias i Pujol (IGTP)
Type of institute	Health sciences research centre
Name(s) of participants (and roles)	R. de Cid (GCAT P.I.)/ N. Blay (GCAT bioinformatician) / A. Lymperidou (GCAT bioinformatician), L. Sumoy (local IMPaCT-Data coordinator / technical contact at IGTP)
Briefly describe your institution's profile (eg. mission, size, expertise)	IGTP is a biomedical research institute linked to a large hospital (Germans Trias i Pujol) aiming to improve patient care through translational research towards effective treatments and preventive medicine (~1000 staff, ~300 active projects, ~650 clinical assays, 8 spinoffs). Focus is in oncological, immune, cardiovascular, digestive, neurological, behavioral, endocrine, and infectious diseases, and community health. Expertise in biomarkers, device and drug testing, cohort studies, through multidisciplinary and interinstitutional collaborative research projects.
Briefly describe your experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	● Expected outcomes: Identifying possible limitations to patient data access for secondary use in research. By increasing visibility through beacon networks, we expect to multiply the use of the GCAT database as a resource for research. ● Challenges and adaptations: Initially we aimed to do a pilot with hospital data, but legitimacy and ethical constraints limited access to real world or synthetic patient data. The only possible way is through specific projects with defined research objectives and proper patient consent. We therefore chose the GCAT cohort data on 20.000 subjects with signed informed consents allowing use of their clinical and genomic linked information, including 808 WGS datasets. ● Pilot data: GCAT beacon focused on EHR diagnostic codes (ICD), genomics, and demographics (Gender, Age). Technically deploying of beacon was simple and hassle-free. ● Future beacons: ○ GCAT has participated sharing GCAT cohort

	data on beacon-SJD-rare. Catalan interhospital database of genetic variants to improve genetic diagnosis in rare diseases, coordinated from SJD hospital.To be published. ○ GCAT aims to federate específic datasets in the context of the IMPaCT T2D project (PI: Jorge Ferrer) for a large WGS dataset dedicated to T2D. To be used in within the consortia and beyond. ○ GCAT is now implementing the GCAT beacon V 2.0 for the GCATcohort database: GCAT cohort data on a beacon-network. To be published: https://beacon-network.org/#/
Beacon URL	To be published

USE CASE 4: IIS La Fe	
Name of Institute	The Health Research Institute Hospital La Fe.
Type of institute	Public Biomedical Research Center.
Name(s) of participants (and roles)	Victoria López Sánchez (Research group: eRPSS, Researcher) Encarnación Perez Martinez (Research group: eRPSS, Researcher) Maria Eugenia Gas-López (PI, Research group: eRPSS) Javier Ripoll Esteve (IT partner) Alfredo Marco Moreno (IT partner)
Briefly describe your institution's profile (eg. mission, size, expertise)	The Health Research Institute Hospital La Fe (IIS La Fe) is the biomedical research area created between the Hospital Universitari i Politècnic La Fe, the Universitat de València, the Universitat Politècnica de València, the Consejo Superior de Investigaciones Científicas, the Fundación para la Investigación del Hospital Universitario La Fe de la Comunidad Valenciana and the Fundación IVI. The IIS La Fe spearheads the research and innovation initiatives within the Health Department of Hospital La Fe. IIS La Fe is dedicated to fostering and advancing high-caliber translational research and health innovation, with the aim of

	tackling healthcare challenges, expanding scientific and economic knowledge, and enhancing societal well-being. Since 2009, IIS La Fe has been officially recognized by the Spanish Ministry of Science and Innovation as a distinguished "Health Research Institute," underscoring the excellence of its research and development endeavors. Key Scientific Areas: Advanced Therapies Translational Research into Prevalent Diseases and their underlying mechanisms Oncology and Oncohematology Reproductive Medicine, Perinatology, and Pediatric Health Drug Development and Safety Emerging Health Technologies Preventive and Sustainable Healthcare Systems
Briefly describe your experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	**Expected benefits**: Implementing Beacon, especially version 2 (v2), as the GA4GH standard provides significant advantages. This integration allows for the aggregation of global genomic and phenotypic data using a standardized API. By leveraging this standard, we enhance data interoperability, enabling seamless collaboration and accelerating scientific discoveries across institutions within the IMPaCT project. Furthermore, Beacon's compatibility with standards like openEHR, HL7 FHIR, and OMOP CDM ensures patient discovery across various data storage and exchange systems. This compatibility streamlines data access and sharing processes, facilitating research efforts. Additionally, the Beacon architecture's flexibility allows tailoring and mapping of the "individuals" model to existing standards, such as openEHR and OMOP CDM. This customization optimizes data management and retrieval, further enhancing research efficiency. Overall, Beacon integration offers a standardized and efficient approach to accessing and analyzing clinical and genomic data, promoting collaboration, and advancing scientific knowledge within the IMPaCT project and beyond. **Challenged faced and how you have adapted:** In addition to grappling with the intricacies of new technologies like Docker and MongoDB, understanding the nuances of writing effective queries to extract specific genomic information has posed a significant hurdle. This process demands a thorough grasp of both the underlying

	data structures and the intricacies of genomic data analysis. Overcoming these challenges requires a dedicated investment of time and resources into ongoing learning and skill development **Choosing pilot data:** Currently, we have only ingested the synthetic data. **Plans for the future beacons:** Following the successful implementation of Beacon with Synthetic Data, our next step is to explore the possibility of deploying the Beacon infrastructure within an IIS La Fe Cloud environment using real data. This expansion would allow us to harness the power of Beacon in a live clinical setting, enabling researchers and healthcare professionals to access and analyse real-world genomic and phenotypic data securely and efficiently. Additionally, we will continue to explore opportunities to collaborate with other institutions and organizations to expand Beacon networks and foster greater data sharing and collaboration in genomic research and healthcare.
Beacon URL	https://remote.iislafe.san.gva.es/api

USE CASE 5: FPS	
Name of Institute	Andalusian Platform for Computational Medicine. Fundación Pública Andaluza Progreso y Salud.
Type of institute	Public clinical research centre.
Name(s) of participants (and roles)	● Joaquín Dopazo. Head of the Computational Medicine Platform ● Javier Perez-Florido. Bioinformatician ● Gema Roldán. Developer ● José L. Fernández-Rueda. Developer
Briefly describe your institution's profile (eg. mission, size, expertise)	The Andalusian Platform for Computational Medicine, part of the Fundación Pública Andaluza Progreso y Salud, plays a crucial role in the Andalusian community's Personalized Medicine Plan. Its objective is to integrate cutting-edge genomic techniques into the everyday clinical practice of the Public Health System, both by developing advanced algorithms and high-quality software and putting these tools in the hands of clinicians.
Briefly describe your	The deployed beacon is built as a wrapper on top of our

experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	custom Collaborative Spanish Variant Server backend. This wrapper translates beacon queries to CSVS calls and the responses back to the beacon specification. The inclusion of CSVS in the beacon network's second version is expected to enable users to determine whether the variants they are studying are specific to the Spanish population. This consideration is crucial when prioritising variants for clinical or research purposes, a practice already established with the first version of the CSVS's beacon. The new version would allow researchers to make richer, more general queries. The overall experience has been great, as the beacon has been a collaborative effort with a focused direction. The biggest challenge in implementing this version of the beacon has been adapting the way our backends work to the huge number of combinations of requests and responses allowed by the beacon, and the complexity leap away from version 1. Perhaps a more closed version would have allowed for equally rich queries but simpler and more flexible implementations. An illustrative example of this is pagination: as some groups supported a offset/limit approach to pagination and others requested a token-based pagination, the specification opted to support both of them. This decision increases the complexity for both service providers and end-user clients without clear benefits for either party. To manage this complexity, it would have been highly desirable to have earlier and more complete conformity tests (as they do not cover all use cases), something we recommend for next versions. CSVS was chosen as our pilot experience due to its simplicity (containing only aggregated data) and our experience with the version 1 of the beacon, along with its overall usefulness. We expect to complete this experience in the next month with two more datasets, whose implementation we are currently testing with the latest compliance tests. These datasets are ENOD (https://www.ciberer.es/en/transversal-programmes/scientific-projects/undiagnosed-rare-diseases-programme-enod) and a dataset of variants associated with reproductive risk in women. In the medium term we hope to have most of our datasets published through the beacon.
Beacon URL	https://csvs.clinbioinfosspa.es/beacon/v2/api

USE CASE 6: IRSJD	
Name of Institute	Sant Joan de Déu Research Institute (IRSJD)
Type of institute	Biomedical Research Center
Name(s) of participants (and roles)	Nidia Barco-Armengol (Bioinformatician), Joe Kane (Developer), Dèlia Yubero (Geneticist), Guerau Fernandez (Principal Investigator of Clinical Bioinformatics Unit)
Briefly describe your institution’s profile (eg. mission, size, expertise)	IRSJD is a biomedical research institute linked to the Sant Joan de Déu Children’s Hospital (HSJD). HSJD offers a contact to specialized clinical services to more than 15,000 children with low-prevalence pathologies that are treated in the center, coordinates clinical care, facilitates genetic diagnosis, and promotes the research of these rare diseases. To overcome the diagnosis odyssey most rare disease patients suffer, we pursue, throughout new methodologies and technologies, the optimal strategy to identify the cause of the disease. HSJD advocates for FAIR data and we are eagerly looking forward to enabling genomic data sharing and standardisation.
Briefly describe your experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	**Expected benefits**: Implementing Beacon v2 will allow the search of specific variants and determine its correlation with specific phenotypes. Variants of unknown significance would be better categorised due to its presence or absence in a much larger cohort. **Challenges faced and how you have adapted**: Due to the increasing number of sequenced patients, updates of the MongoDB collections have to be done regularly. Phenotype, informed variants and diagnosis have to be merged in different time points. **Choosing pilot data**: Synthetic data. **Plans for the future beacons**: We plan to expand to other institutions the Catalan interhospital database of genetic variants to improve genetic diagnosis in rare diseases.
Beacon URL	To follow soon

USE CASE 7: IDIAPJGol	
Name of Institute	Institut d’Investigació en Atenció Primària Jordi Gol (IDIAPJGol)
Type of institute	Primary Care Research Institute
Name(s) of participants (and roles)	María Aragón (Head of SIDIAP) Clara Rodríguez (Responsible of SIDIAP’s Data Quality Unit) Jordi Carrere (Senior Data Scientist and Responsable of Develop and Innovation SIDIAP’s Unit) Marta García (Data Manager and Developer)
Briefly describe your institution’s profile (eg. mission, size, expertise)	Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) is a centre of reference in research and health promotion at the primary health care level that aims to promote and develop innovation, clinical, epidemiological and health services research in the field of primary health care. We also offer training to generate knowledge, disseminate results and transfer them to clinical practice. Our aim is to bring efficiency to the health system and promote the well-being of individuals.
Briefly describe your experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	The implementation of Beacon, v2, will help us to join genomic and biological data to the primary care electronic health records. For SIDIAP also is new the work methodology using the API, and it can be useful for the future. The volume of information has made us go with a test, hoping to include everything in the future. The need to have it out of our internal net also made it difficult for us. For the moment we have it only accessible from the Catalan Health Institute net. For the moment we are working with a synthetic test dataset for the future we believe we will have the approval of the Ethic Committee and the system ready to go outside.
Beacon URL	To follow soon

USE CASE 8: CNIC	
Name of Institute	Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC)
Type of institute	Cardiovascular Research Center
Name(s) of participants (and roles)	Fátima Sánchez Cabo (Head of Bioinformatics Unit) Jorge de la Barrera Martínez (Bioinformatician) Juan Ignacio Álvarez Arenas (Bioinformatician) Jose Javier Fuster Ortuño (Group Leader, Hematovascular Pathophysiology Laboratory) Miriam Díez Díez (Predoctoral Researcher - Hematovascular Pathophysiology Laboratory) Beatriz de las Mercedes López Ramos-Neble (Predoctoral Researcher - Hematovascular Pathophysiology Laboratory)
Briefly describe your institution's profile (eg. mission, size, expertise)	The CNIC is focused on translating research findings into societal benefits, improving public health, and generating economic opportunities. It prioritizes converting research results into clinical practice enhancements and business sector opportunities. Training is a fundamental activity at the CNIC, with the CNIC-JOVEN Training Plan supporting individuals from senior high school to postdoctoral levels to cultivate a strong base of biomedical researchers in Spain. The center employs over 400 staff, with 85% directly involved in research projects, either in technical units or research groups. CNIC researchers are internationally recruited through a competitive process and undergo periodic evaluation by an external committee of 10 internationally recognized scientists. The CNIC has been recognized as a Severo Ochoa Center of Excellence, placing it among the top 8 research centers in Spain.
Briefly describe your experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	Expected benefits: The implementation of Beacon v2 for our curated variants in clonal hematopoiesis of indeterminate potential (CHIP) aims to enhance accessibility and visibility of these variants. Researchers and clinicians will have the ability to search for specific variants and correlate them with CHIP phenotypes, facilitating better categorization of variants of unknown significance and contributing to a deeper understanding of CHIP. Challenges faced and how we have adapted: A significant challenge we encountered was the manual collection of variants from the literature. This process

	required extensive effort and attention to detail, as each variant had to be carefully curated. We adapted by establishing rigorous protocols and quality control measures to ensure the accuracy and reliability of the collected data. The database was an easy job due to the beacon JSON schemas published. Choosing pilot data: Our pilot data consisted of carefully curated variants identified as drivers in clonal hematopoiesis of indeterminate potential. These variants, extensively studied, were selected as ideal candidates for the initial implementation of the Beacon.
Beacon URL	https://bioinfo.cnic.es/chipdb/api/

USE CASE 9: Biobizkaia	
Name of Institute	Biobizkaia
Type of institute	Health Research Institute
Name(s) of participants (and roles)	● Naiara Garcia Bediaga(Head of the Bioinformatics Unit) ● Aitor Zarandona Garai (Bioinformatician) ● Alexander Moreno Lobato (Developer)
Briefly describe your institution's profile (eg. mission, size, expertise)	The Biobizkaia Health Research Institute unites a multidisciplinary team comprising more than 1,400 professionals, including a substantial contingent of clinical researchers associated with diverse healthcare institutions across Biscay. Furthermore, Biobizkaia serves as a hub for international Ikerbasque researchers and scholars from the University of the Basque Country. Biobizkaia's mission is to advance translational research and healthcare innovation to generate value and foster positive societal health outcomes by facilitating collaboration between clinical and basic research, and acting as a nexus between the healthcare system, industry, and the broader scientific and technological knowledge ecosystem.

Briefly describe your experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	**Expected benefits**: The adoption of Beacon v2, is poised to facilitate the seamless integration, retrieval, and sharing of genomic data alongside primary care electronic health records. This integration will enable 'discovery queries' across cohorts, thus helping us, researchers, to identify pertinent samples, patient data, and cohorts aligned with their specific research inquiries. **Challenges faced**: The deployment of Beacon2 and upload of the synthetic data has been relatively straightforward **Choosing pilot data and plans for future beacons:** Currently, we are only utilising a synthetic test dataset. Moving forward, we anticipate obtaining approval from the Ethics Committee and implementing Beacon in a new dataset.
Beacon URL	https://beacon2.iisbiobizkaia.eus/api

USE CASE 10: FPGMX	
Name of Institute	Fundación Pública Galega de Medicina Xenómica
Type of institute	Public Health Foundation
Name(s) of participants (and roles)	Jorge Amigo (Head of Bioinformatics Unit) Antón Ambroa (Bioinformatician) Iria Roca (Bioinformatician) Rubén Rodríguez (Bioinformatician)
Briefly describe your institution's profile (eg. mission, size, expertise)	The main aim of the FPGMX is to provide genetic diagnosis for the entire Galician population. Directed by Ángel Carracedo, over 60 professionals, from lab technicians to bioinformaticians and geneticists, deal with over 40000 annual samples. Starting in 2010, we have been increasingly using NGS to improve and accelerate our reports, and we have recently been included as 1 of the 3 sequencing IMPaCT nodes.
Briefly describe your	We initially planned to be part of the Beacon network

experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	through the GPAP implementation we have for IMPaCT data. We then decided to include at least boolean discovery of all the sequencing variation we have historically found.
Beacon URL	To follow soon

USE CASE 11: CNIO	
Name of Institute	Centro Nacional de Investigaciones Oncológicas (CNIO)
Type of institute	Cancer Research Center
Name(s) of participants (and roles)	Fátima Al-Shahrour (Head of Bioinformatics Unit) Francisco Javier Soriano Díaz (Bioinformatician)
Briefly describe your institution's profile (eg. mission, size, expertise)	The CNIO has over 400 highly qualified specialists dedicated to cancer research, aiming to make cancer no longer one of the leading causes of death in our society. Recognized globally, CNIO stands out for its contribution both in the quantity and quality of scientific publications, as well as in the development of innovations that may lead to new cancer treatments and drugs. CNIO's main focus is to understand the causes of cancer, believing that only by fully understanding these processes can the disease be effectively prevented and fought. Additionally, the center is committed to quickly translating research advancements into clinical practice by identifying promising compounds, leading clinical trials, and providing genetic diagnosis and counseling to affected families. This commitment is also reflected in the creation of biotechnological companies that expand the scope of research and generate employment. CNIO is a dynamic center that attracts talent globally, with a predominantly young research population and a solid technological infrastructure. Moreover, it collaborates closely with hospitals and pharmaceutical companies to

	enrich its expertise and ensure the continuity of research with additional sources of funding.
Briefly describe your experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	**Expected benefits:** Our goal is to implement Beacon 2 as a use case for further development of the tool on other projects. In the long term this will help us to make the data generated by our analysis available to anyone who wants it. **Challenges faced and how we have adapted:** It is our first approach to this technology, so the problems we have encountered are the usual ones that are encountered when using a technology for the first time, such as the difficulty with the deployment or the adaptation of the data to the appropriate format. Direct communication with the people involved in Beacon 2 has allowed us to overcome these problems. An unexpected difficulty that has not been solved for the moment is the obligation to enter metadata one by one by hand. **Choosing pilot data:** In this implementation we will use data from a cohort of brain metastasis data collected in the context of the RENACER project. Clinical and variant information is available.
Beacon URL	To follow soon

USE CASE 12 IMIM	
Name of Institute	Hospital del Mar and Hospital del Mar Research Institute
Type of institute	Clinical and Research Institution
Name(s) of participants (and roles)	Miguel Angel Mayer (Principal Investigator and Scientific Coordinator) Júlia Perera (Senior Bioinformatician) Juan Manuel Ramírez Anguita (Senior Data Scientist) Angela Leis (Post-Doctoral Data Scientist)
Briefly describe your institution's profile (eg. mission, size, expertise)	The Hospital del Mar is a general and university hospital, active and with research activity, which treats pathologies of medium and high complexity.

experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	through the GPAP implementation we have for IMPaCT data. We then decided to include at least boolean discovery of all the sequencing variation we have historically found.
Beacon URL	To follow soon

USE CASE 11: CNIO	
Name of Institute	Centro Nacional de Investigaciones Oncológicas (CNIO)
Type of institute	Cancer Research Center
Name(s) of participants (and roles)	Fátima Al-Shahrour (Head of Bioinformatics Unit) Francisco Javier Soriano Díaz (Bioinformatician)
Briefly describe your institution's profile (eg. mission, size, expertise)	The CNIO has over 400 highly qualified specialists dedicated to cancer research, aiming to make cancer no longer one of the leading causes of death in our society. Recognized globally, CNIO stands out for its contribution both in the quantity and quality of scientific publications, as well as in the development of innovations that may lead to new cancer treatments and drugs. CNIO's main focus is to understand the causes of cancer, believing that only by fully understanding these processes can the disease be effectively prevented and fought. Additionally, the center is committed to quickly translating research advancements into clinical practice by identifying promising compounds, leading clinical trials, and providing genetic diagnosis and counseling to affected families. This commitment is also reflected in the creation of biotechnological companies that expand the scope of research and generate employment. CNIO is a dynamic center that attracts talent globally, with a predominantly young research population and a solid technological infrastructure. Moreover, it collaborates closely with hospitals and pharmaceutical companies to

Briefly describe your experience with deploying Beacon *For example:* ● Expected benefits ● Challenges faced and how you have adapted ● Choosing pilot data ● Plans for future beacons	The implementation of Beacon, v2, will allow us to integrate genomic and biological data to the clinical information of our hospital-based electronic health records. The integration of genomic and clinical will allow us to transform from a one-size-fits-all approach to a more personalised and precise model. This shift not only has the potential to improve individual patient care but also to enhance the broader health outcomes of populations. We are considering the use of synthetic test dataset but the most interesting use case will be the use of clinical data after the approval of the Ethic Committee.
Beacon URL	To follow soon