Trabajo Fin de Master

MÁSTER EN DIRECCIÓN DE SISTEMAS Y TIC PARA LA SALUD

Y EN DIGITALIZACIÓN SANITARIA

Curso académico 2021/2022

Propuesta de diseño de un Data Lake Sanitario y Análisis Avanzado de Riesgo Cardio Vascular

04/10/2022

Tutor

Parra Calderón, Carlos Luis

Autores

López García, Ángela

González de Castro, Nuria

Oliva Pérez, Juan Carlos

PARRA CALDERON CARLOS LUIS - 28589927F Firmado digitalmente por PARRA CALDERON CARLOS LUIS - 28589927F Fecha: 2022.10.06 13:41:14 +02'00'	Gonzalez De Castro Nuria Digitally signed by Gonzalez De Castro Nuria DN: dc=com, dc=novartis, ou=people, ou=PH, serialNumber=882356, cn=Gonzalez De Castro Nuria Date: 2022.10.04 20:03:44 +02'00'	LOPEZ GARCIA, ANGELA (FIRMA) Firmado digitalmente por LOPEZ GARCIA, ANGELA (FIRMA) Fecha: 2022.10.04 18:07:43 +02'00'	OLIVA PEREZ JUAN CARLOS - 02639812X Firmado digitalmente por OLIVA PEREZ JUAN CARLOS - 02639812X Fecha: 2022.10.04 17:50:39 +02'00'
Firma Tutor	Firma Autor	Firma Autor	Firma Autor

INDICE

1. Resumen Ejecutivo

Los datos de una organización sanitaria son uno de sus principales activos, porque la prestación asistencial esta soportada por los mismos. Para extraer aún más valor de esta información, se puede avanzar en su uso secundario, destinándolos a la investigación para el impulso de la Medicina 5P (Poblacional, Preventiva, Predictiva, Personalizada y Participativa) y a la mejora de la toma de decisiones en todos los aspectos relacionados con la gestión de la salud, para lo que se requiere de sistemas de información especializados conocidos como Data Lakes Sanitarios.

Un Data Lake Sanitario es una solución que debe estar soportada por un modelo organizativo, procesos, estándares y herramientas, para de una forma ágil, segura y confidencial, proceder a la captura, consolidación, tratamiento y procesamiento masivo de la información con la que abordar análisis descriptivos, diagnósticos, predictivos, prescriptivos, simulaciones y el desarrollo de modelos con los que avanzar en la generación de nueva evidencia (Real World Evidence - RWE), y sistemas de soporte a la decisión (Decision Support System - DSS) a partir de los datos de vida real (Real World Data - RWD) del ámbito clínico-asistencial, medio-ambiental, demográfico, geográfico,… y de la experiencia (Patient-Reported Experience Measures - PREMS) y los resultados (Patient-Reported Outcome Measures - PROMS) reportados por los pacientes.

Este Trabajo de Fin de Master, en adelante TFM, constituye una propuesta de innovación práctica, que tiene por objeto ayudar en la transición digital del Sistema Nacional de Salud, mediante la definición de los requisitos que debe cumplir un Data Lake Sanitario en materia normativa, organizativa y técnica.

A lo largo de este TFM, se abordará un análisis de situación de las iniciativas de referencia en el ámbito internacional, europeo, nacional y regional, además de identificar los factores a tener en consideración para la gobernanza de estas soluciones y realizar una especificación general de los requisitos técnicos que deben cumplir las herramientas de analítica avanzada y para dar soporte a un caso de uso clínico sobre riesgo cardiovascular, uno de los problemas de salud de mayor prevalencia en España, primera causa de muerte y de un enorme impacto sanitario, social y económico.

Este TFM ha sido desarrollado bajo la tutoría de Carlos Luis Parra Calderón y se encuentra estructurado en tres capítulos con las siguientes autorías:

4. Capítulo I. Data Lake Sanitario. Juan Carlos Oliva Pérez

5. Capítulo II. Caso de uso de Riesgo Cardio-Vascular. Nuria González de Castro

6. Capítulo III. Soluciones de Analítica Avanzada. Ángela López García.

2.Justificación de los Objetivos

Los objetivos principales de este TFM son:

1. Extraer conclusiones a partir del análisis de situación de las iniciativas existentes identificando sus objetivos, característica y los retos a abordar para avanzar con espacio de datos de salud en el ámbito europeo. impulsar la I+D+i.

2. Identificar factores clave para crear una cultura en torno al dato en el entorno sanitario.

3. Definir los requisitos de un Data Lake Sanitario para:

   - Contar con un modelo de Gobernanza que le confiera máximas garantías.
   - Impulsar el desarrollo de la I+D+i y la mejora de los resultados en salud
   - Implementar un modelo técnico que soporte la Analítica Avanzada[1] y la Inteligencia Artificial[2]

4. Abordar un ejercicio "práctico" mediante el diseño de un caso de uso de Análisis Avanzado de datos en el ámbito del riesgo cardiovascular.

1 Analítica Avanzada (AA): Descubrimiento e interpretación de patrones en los datos para ser aplicados en la investigación o la gestión. Dentro de este concepto tiene cabida, tanto las soluciones capaces de ofrecer información consistente para optimizar la gestión, comúnmente conocidas como de "Analítica Tradicional", "Inteligencia de Negocio" (IN) o Business Intelligence (BI), pero también las soluciones de "Analítica Avanzada" o Knowledge Discovery in Databases (KDD), que son de aplicación en el ámbito de la Minería de Datos, conocido porque, de un modo similar a como se hace con los minerales, persigue encontrar algo valioso e inicialmente oculto y entre cuyas técnicas se encuentra el Aprendizaje Automatizado o Machine Learning (ML), un conjunto de diversos algoritmos con capacidades descriptivas, predictivas y prescriptivas, que se agrupan en tres modelos de aprendizaje conocidos como, no supervisado, por refuerzo y el que probablemente sea de mayor aplicación en el ámbito de la salud, supervisado, donde los sistemas aprenden mediante el entrenamiento y bajo la supervisión de las personas

2. Inteligencia Artificial (IA) o Artificial Intelligence (AI): La Comisión Europea define la IA como aquellos sistemas que manifiestan un comportamiento inteligente, al ser capaces de analizar el entorno y realizar acciones, con cierto grado de autonomía, con el fin de alcanzar objetivos específicos.

3.Referencia de los contenidos del temario que se han utilizado

Los temas del Master en Dirección de Sistemas y TIC para la Salud y en Digitalización Sanitaria, en adelante DSTICSDS, que han sido utilizados como referencia para la elaboración de este TFM son:

- Tema 2.5 La seguridad TIC. Legislación aplicable. Aplicación del Reglamento General de Protección de Datos. El papel del Delgado de Protección de Datos. Auditorias. Metodologías / Herramientas de Seguridad Magerit. SGSI.
- Tema 2.8 Cloud Computing. Big Data. Infraestructuras Centralizadas, CPD´s. Servicios de Housing, Hosting. Centros de backup.
- Tema 3.2 La aplicación de la normativa de Protección de Datos en el sector Salud
- Tema 3.3 La Interoperabilidad en al ámbito de la Salud
- Tema 3.9 Nuevos modelos asistenciales basados en las TIC: gestión de crónicos y relación con el ciudadano
- Tema 3.10 Estrategias, infraestructuras y aplicaciones avanzadas basada en datos para la Investigación en Salud y Biomedicina
- Tema 4.7 Aplicaciones en asistencia sanitaria e investigación
- Tema 4.8 Analítica y modelos predictivos en salud

4.Capítulo I. Data Lake Sanitario

4.1. Justificación de la Necesidad

En 1948 la OMS define la **Salud** como “un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades”, por lo que un buen estado de salud contribuye a que las personas vivamos más años y mejor, constituyendo España y su Sistema Nacional de Salud, referentes mundiales en esta materia de acuerdo a:

- El estudio de “Forecasting Life Expectancy” (The Lancet, 2018), que predice que los hombres y mujeres nacidos en España en 2040 tendrán la mayor esperanza de vida del mundo, respectivamente 83,6 y 87,4 años.

- El estudio de “Economies With the Most Efficient Health Care” (Bloomberg, 2018), que analizó los sistemas sanitarios de 200 economías, concluyendo que el español es el tercero más eficiente, siendo superado únicamente por Hong Kong y Singapur.

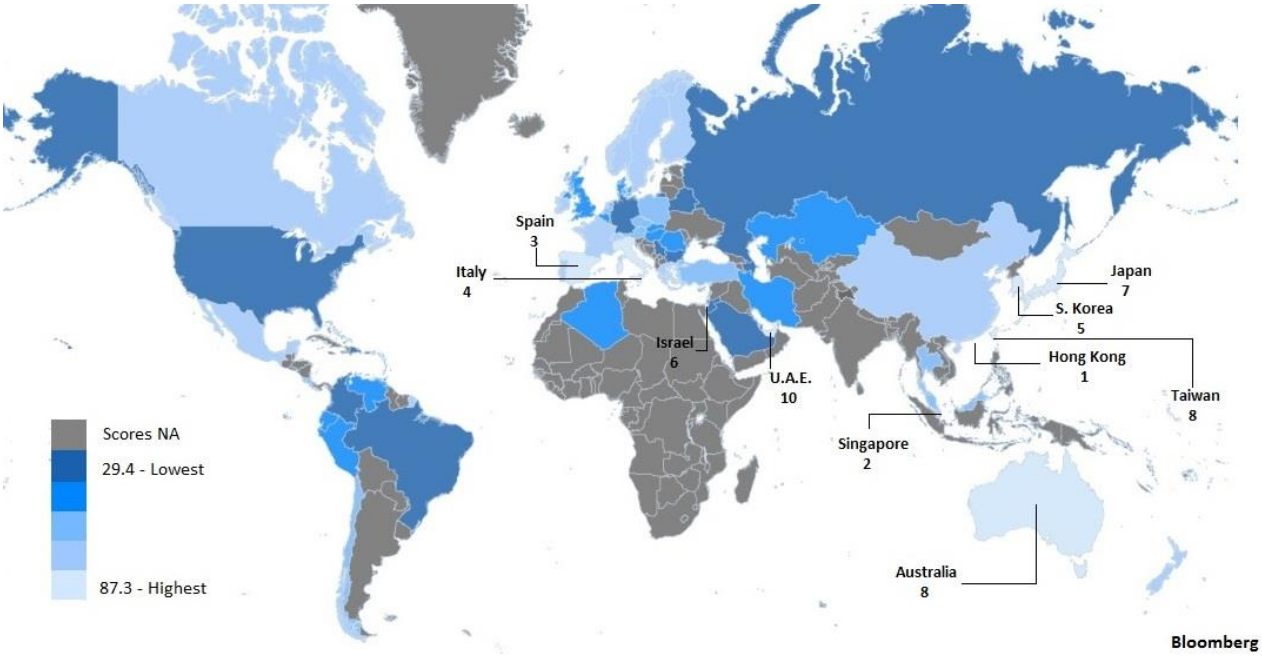


Ilustración 4.1. Health Care Efficience Score. Bloomberg

Para mantener y mejorar este estatus, debemos ser capaces de hacer frente al incremento de costes derivado de la atención socio-sanitaria y del envejecimiento progresivo de la población, para lo que se debe tener en consideración los siguientes factores:

- Multitud de estudios coinciden en sus conclusiones sobre los determinantes de la salud, de forma que los hábitos de vida relacionados con la alimentación, la práctica de ejercicio físico, el patrón de sueño, el estado emocional o los consumos tóxicos, impactan de una forma más notable en nuestra salud, que otros determinantes, como la propia atención sanitaria. Es por ello, que existe un consenso en evolucionar la prestación asistencial, desde una medicina reactiva, que actúa cuando ya ha aparecido la enfermedad, hacia una **medicina preventiva**, que nos ayude a estar más tiempo sanos.

  Estas conductas generadoras de salud, tienen que ser abordadas otorgando un papel protagonista a las personas, teniendo en consideración sus usos y costumbres y asumiendo que son pacientes activos, que quieren comprender su enfermedad, estar informados en todo instante, ser partícipes en la toma de las decisiones que impactan en su salud, avanzar en su autocuidado y ser atendidos en un entorno lo más amigable y próximo a su lugar de elección.

  Un ciudadano que no solo quiere leer, también quiere escribir en su historia clínica electrónica la información que es generada en su actividad diaria, o decidir, a que estudios de investigación son destinados sus datos sanitarios, lo que representa una nueva forma de relación conocida como **medicina participativa**.

Ilustración 4.2. Determinantes de la Salud. Neuropediatra.org. Glucómetro FreeStyle-Abbott

- Al margen de la idoneidad de estar más tiempo sanos, es inevitable que los problemas de salud aparezcan, y cuando esto ocurre es fundamental poder identificarlos en sus primeros estadios, ya que un diagnóstico temprano suele ir asociado a un tratamiento más favorable,

mas supervivencia y mejores resultados en salud con un coste menor, y es esta necesidad de anticiparse la que ha establecido como objetivo el desarrollo de la **medicina predictiva**.

- Aunque la medicina siempre ha sido personalizada, es un hecho contrastado que los diferentes tratamientos sólo actúan en un porcentaje de los casos, en base a una estadística conocida, además de existir un importante número de alternativas terapéuticas para hacer frente a cada problema de salud.

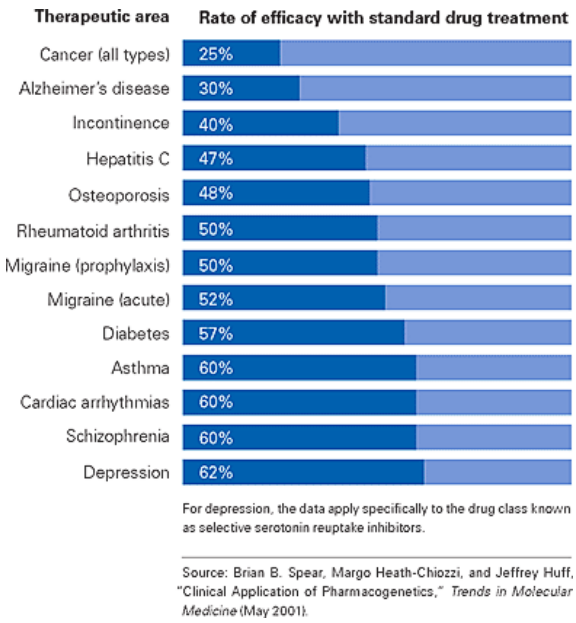


Ilustración 4.3. Eficiencia estadística de los tratamientos por área terapéutica

A modo de ejemplo, presentamos en este cuadro el coste de medicación mensual estimado de tratamiento para el cáncer colorrectal metastásico

Régimen	Medicación y programa de administración	Costes de medicación
Regímenes que contienen fluorouracilo		
Mayo Clinic	Bolo mensual de fluorouracilo más leucovorin	85 €
Roswell Park	Bolo semanal de fluorouracilo más leucovorin	180 €
LV5FU2	Fluorouracilo quincenal más leucovorin en una infusión de 48h	145 €
Regímenes que contienen irinotecán u oxaliplatino		
Irinotecán sólo	Bolo semanal	300 €
IFL	Bolo semanal de fluorouracilo más irinotecán	350 €
FOLFIRI	LV5FU2 con irinotecán quincenal	470 €
FOLFOX	LV5FU2 con oxaliplatino quincenal	780 €
Regímenes que contienen bevacizumab ó cetuximab		
FOLFIRI con Bevacizumab	FOLFIRI con Bevacizumab quincenal	6.000 €
FOLFOX con Bevacizumab	FOLFOX con Bevacizumab cada dos semanas	6.300 €
Irinotecán con Cetuximab	Cetuximab semanal más irinotecán	5.800 €
FOLFIRI con Cetuximab	FOLFIRI y Cetuximab semanal	6.000 €

*** Precios aproximados para cada tratamiento – actualizados año 2015**

Ilustración 4.4. Alternativas terapéuticas y coste para un mismo diagnóstico

Para hacer frente a esta variabilidad, se plantea la necesidad de avanzar en el desarrollo de nuevos bio-marcadores y en el análisis de la información clínica y las características moleculares y genéticas de cada persona, para el desarrollo de una **medicina personalizada y de precisión**[3], que establezca una respuesta precisa a cada tratamiento en términos de resultados de salud, eficacia, toxicidad etc., prescribiendo de una forma determinista para la combinación de factores única que se da en cada persona.

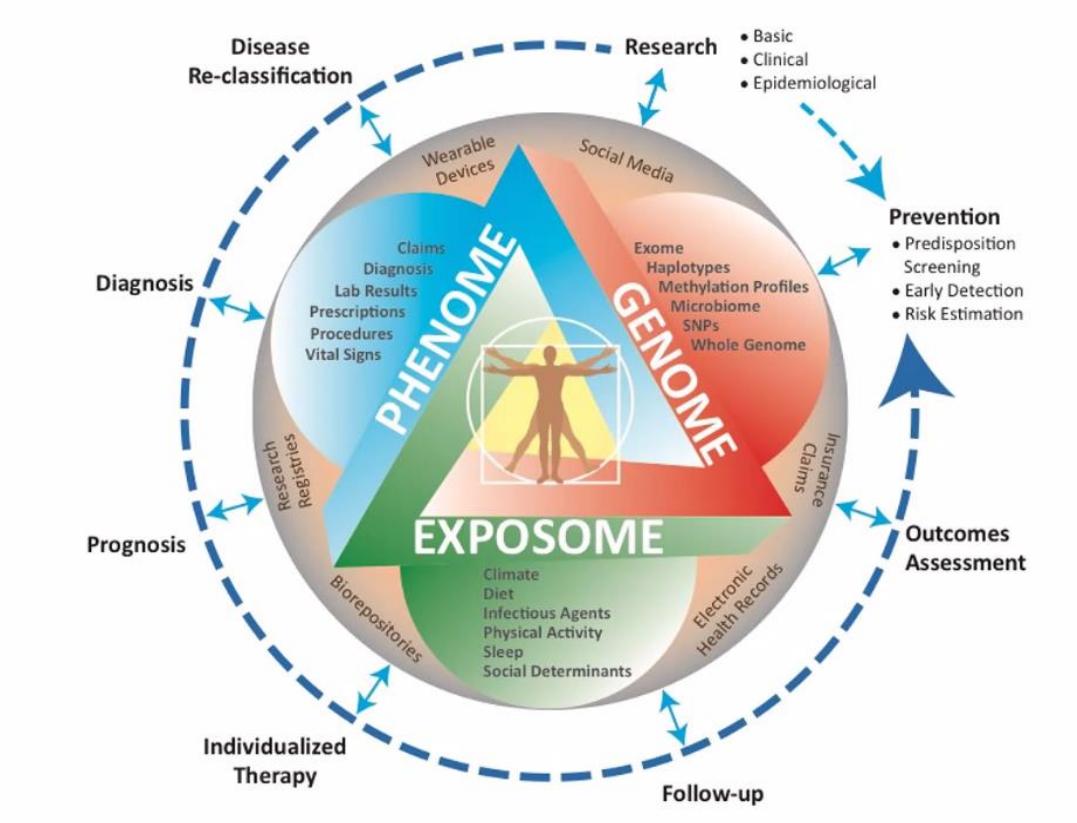


Ilustración 4.5. Fenotipo, Genotipo y Exposoma

3 **Medicina Personalizada:** aquella en la que los procesos de diagnóstico y tratamiento se hacen de forma expresa para cada persona, teniendo en consideración todas sus características individuales, fenotipo, genotipo (genoma, epigenoma, proteoma, metaboloma, etc.) y de la interacción con su propio entorno o exposoma.

Medicina de Precisión: ejecución de los actos asistenciales de la forma menos dañina posible para el paciente, con tecnologías mínimamente invasivas con microincisiones, o incluso no invasivas, como la radiocirugía, y técnicas futuras, como las nanoestructuras capaces de transportar un fármaco hasta la zona dañada.

Definidos los cinco principios rectores de la **Medicina 5Ps** (Preventiva, Participativa, Predictiva, Personalizada y de Precisión), recae en las instituciones sanitarias el deber de identificar y proveer los recursos necesarios para hacer posible su desarrollo.

Las organizaciones sanitarias se caracterizan por ser entidades complejas, que prestan asistencia mediante la combinación de un gran número de recursos heterogéneos, como son las grandes infraestructuras y las tecnologías diagnósticas y terapéuticas, aunque si tuviéramos que destacar un activo sobre todos los demás, éste sería sin duda el **conocimiento de los profesionales.**

Este conocimiento representa el factor clave para poder encontrar un equilibrio entre los resultados en salud y la sostenibilidad de las organizaciones sanitarias y constituye un gran reto, consecuencia del su enorme volumen y rápido crecimiento, consecuencia, entre otros, del continuo lanzamiento de nuevos tratamientos.

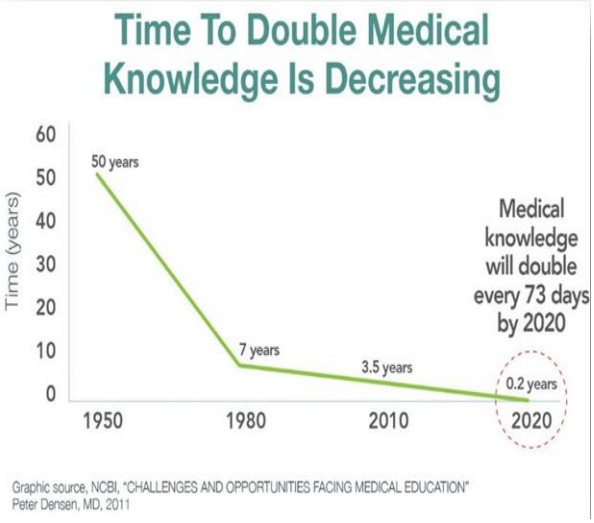


Ilustración 4.6. Incremento del Conocimiento

Proyectos	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	Total
Cáncer de pulmón	1	2	1	1	1	6	10	11	14	22	69
Leucemia	2	1	1	2	4	3	5	14	10	13	55
Cáncer de mama	1	3	1	3	1	1	1	17	15	11	54
Linfoma	2		2		2	1	5	12	11	6	41
Mieloma múltiple			1	3		4	5	4	3	7	27
Melanoma		1	1	2	1	6	2	5	2	3	23
Cáncer de ovario		1	1		1	1	1	5	3	4	17
Otro				2	1	4		3	1	4	15
Cáncer renal	2		1				3	5	1	2	14
Cáncer de próstata		2		3	1			2	3	2	13
Cáncer de cabeza y cuello								3	5	5	13
Cáncer colorrectal		1		2			4	3	1	2	13
Cáncer gástrico	1	1			2			1	3	5	13
Cáncer hepático								2	5	1	8
Cáncer de pancreas					1		1		3	2	7
Cáncer de vejiga								2	5		7
Tumores neuroendocrinos	1		1			1	3				6
Cáncer de tiroides			1		2	1			1		5
Sarcoma de Tejidos Blandos (ST			1				3				4
Mielofibrosis			1					2		1	4
Cáncer de esófago									2	1	3
Cáncer de próstata							2				2
Glioblastoma multiforme								1			1
Total	10	12	13	18	17	28	45	92	88	91	414

Ilustración 4.7. Incremento del 500% de nuevas terapias por tumor en una década

El aumento de la presión asistencial derivado del envejecimiento progresivo de la población, por una mayor prevalencia de patologías crónicas y la pandemia por SARS-CoV2, unido al incremento continuo del conocimiento, al que antes se hacía referencia, hace inviable en la práctica, que el personal clínico, asistencial y de gestión, pueda estar en disposición de la mejor evidencia científica y técnica, en términos de eficacia, eficiencia, seguridad, para cada una de las actuaciones de prevención, diagnóstico, tratamiento y gestión, a las que se enfrentan en su actividad diaria.

Esta situación se traduce en variabilidad clínica innecesaria, en diagnósticos más tardíos e imprecisos y también en terapéuticas menos favorables para el paciente y de mayor coste para el sistema sanitario, impactando negativamente en la calidad de la prestación asistencial y en la sostenibilidad de las instituciones. Por todo ello, se plantea la necesidad de evolucionar los sistemas de información actuales, de herramientas pensadas para el registro y consulta de información, a soluciones capaces de proveer soporte a la decisión (Decission Support Systems-DSS).

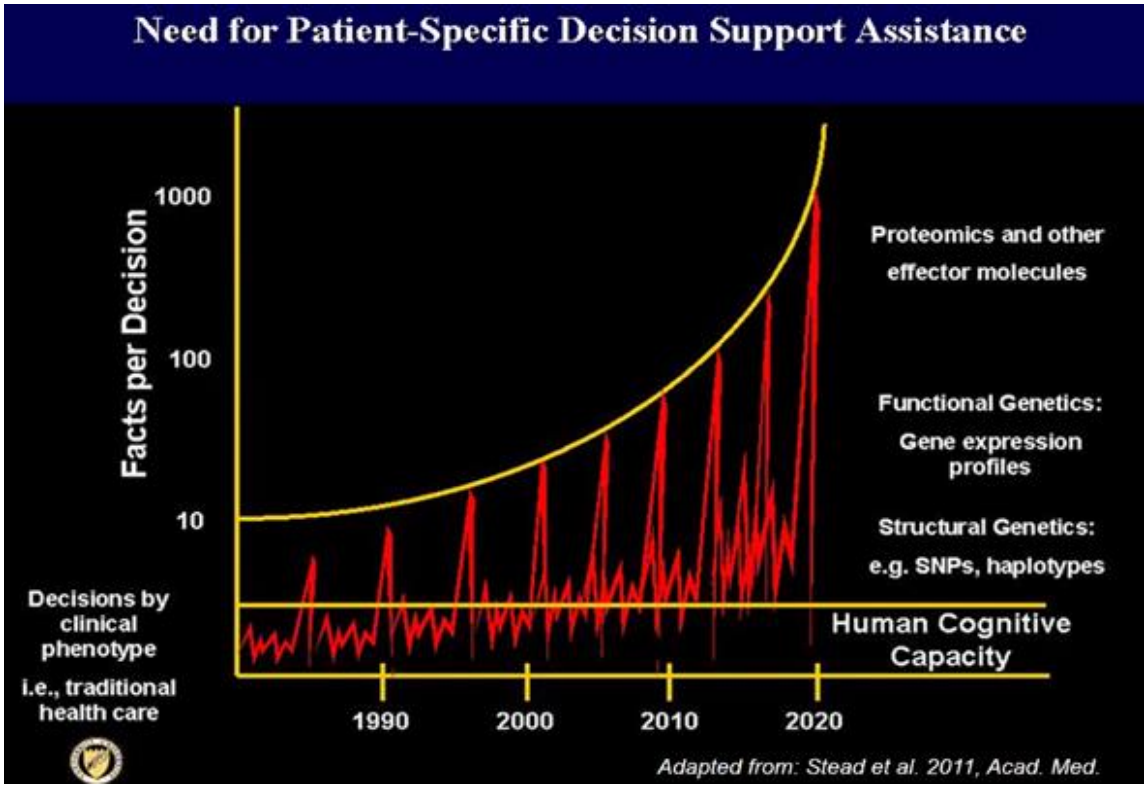


Ilustración 4.8. Necesidades de Soporte a la Decisión versus capacidad cognitiva humana

Para proveer soporte a la decisión, se requiere en primera instancia el desarrollo o la adquisición del mejor conocimiento y después, ser capaces de vehiculizarlo dentro de los procesos de atención y gestión, y es aquí donde la **Analítica Avanzada y la Inteligencia Artificial** soportadas por el **Big-Data**[4], son presentadas como herramientas de extraordinario valor, para ayudar a responder a preguntas complejas, soportar la evaluación sistematizada de tecnologías y procesos, evidenciar resultados en salud, acelerar la investigación clínica, desarrollar algoritmos para soportar la decisión de profesionales y ciudadanos y con todo ello, contribuir al desarrollo de la Medicina 5Ps, capaz de trasladar el foco desde la patología aguda y la actividad, a una visión holística y a largo plazo de resultados en salud, conocida como Value-Based HealthCare - VBHC (Porter, 2006).

	Servicio	Valor
Foco	**Patología Aguda**	**Visión integral de la salud**
Objetivo	**Más Actividad** y atención rápidas	**Mejores Resultados en Salud** a **largo plazo**
Evaluación	**Vertical** Para cada ámbito; servicio, unidad, profesional (agenda, derivación, I.T., prescripción, LEQ, LECEX, etc.)	**Transversal** Resultados en salud por cada problema de salud / opciones terapéuticas. GC+TA -> Continuidad de actuaciones (SP, AP, AE, social,..)
Indicadores	**Internos** Propios de cada organización Para consumos y pactos Comparación interna y anual	**Estandarizados** Habilita la mejora -> comparativa externa Calculados a partir de la información generada por profesionales y pacientes
Compra	Lo que se quiere **tener** ("cosas" -> servicios y suministros)	Lo que se quiere **conseguir** (resultados en salud)

Ilustración 4.9. Evolución asistencia: Actividad –> Valor. Fuente elaboración propia

El desarrollo de la analítica avanzada y la inteligencia artificial a partir de una cantidad ingente de datos, requiere de una combinación de conocimientos técnicos y un Data Lake Sanitario dotado de procesos y recursos para operarlo.

4 **"Big Data":** Se dice que es un concepto al que se recurre cuando los datos con los que vamos a trabajar, cumplen con las 3Vs. **Volumen,** el Big Data hace frente al problema del Volumen de datos, con el principio básico de divide y vencerás, en realidad recurre a paralelizar el tratamiento de datos en redes de ordenadores que se comportan como uno solo. **Velocidad**, hace referencia la enorme frecuencia en la generación, recogida y proceso de la información. Además, es capaz de hacer esto con multitud de herramientas para tratar una **Variedad** de datos (video, imagen, genoma, textos, etc.) y además con diferentes requisitos de tiempo.

4.2. ¿Qué podemos esperar?

La revolución industrial nos ayudó a realizar labores manuales y repetitivas, liberando a las personas de tareas penosas y acelerando nuestro desarrollo. Con la analítica avanzada y la inteligencia artificial, "los robots" pasarán de realizar labores mecánicas a intelectuales, y no requieren de brazos robóticos, si de algo más parecido a un cerebro, es por ello que la AA y la IA son soportadas por plataformas donde se combinen infraestructuras de procesamiento, almacenamiento y software con los que:

- **Aplicar el conocimiento existente**, la AA y la IA son capaces de replicar el conocimiento de los seres humanos mediante la aplicación de **reglas** predefinidas. También sobresalen cuando la definición de un patrón de actuación es difícil de expresar mediante la programación convencional y se procede a enseñar a un sistema mediante entrenamiento, alimentándolo con unos datos de entrada y unos resultados, para generar un software capaz de obtener los mismos resultados ante iguales entradas.

- **Descubrir nuevo conocimiento,** base de la investigación y donde la AA puede representar una gran ayuda gracias a su enorme capacidad observacional, identificando relaciones ocultas en los datos allí donde la mente humana no alcanza, consecuencia de un exceso de vectores. Una capacidad que también puede ser de aplicación en el ámbito clínico, asistencial y gestor, aunque teniendo siempre presente, que la correlación entre dos hechos no implica causalidad.

La analítica avanzada será de gran ayuda en modelar y presentar el conocimiento existente, pero no podemos pretender que estas soluciones sean creativas, formulen preguntas, planteen hipótesis o establezcan conclusiones, características inherentes al ser humano y que nos permiten llegar a lugares no alcanzables mediante la aplicación de simples reglas, imitación o copia.

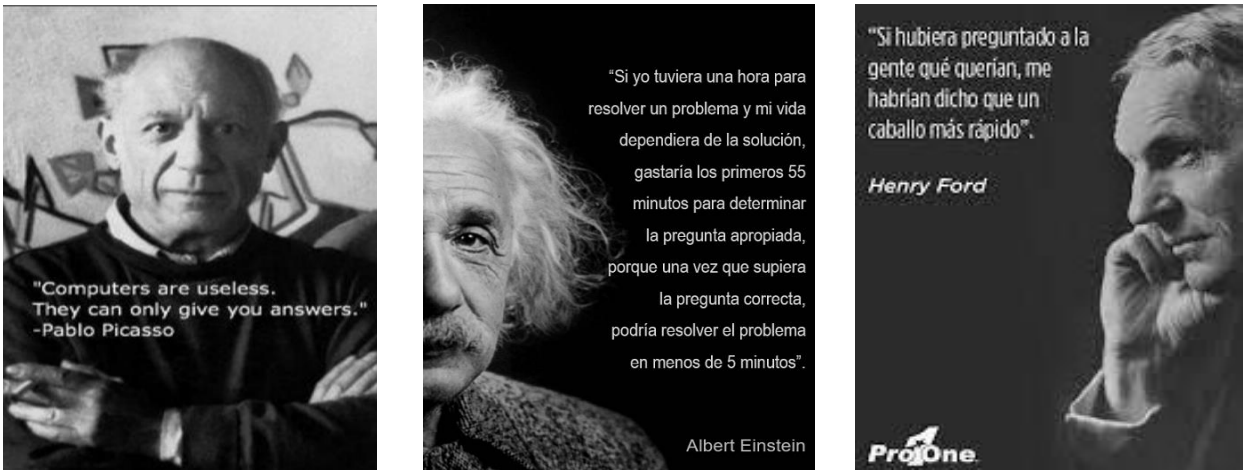


Ilustración 4.10. Preguntar, qué preguntar y creatividad.

4.3. ¿Por qué ahora?

Una vez expuesta la necesidad de un Data Lake Sanitario para soportar el desarrollo de la AA y la IA, y precisados el tipo de conocimiento y soporte que podemos esperar de estas soluciones, las siguientes cuestiones a plantearse son: ¿por qué nos planteamos en este instante una tecnología que se definió hace décadas? ¿realmente es una tecnología en fase productiva o estamos asistiendo a un pico de expectativas inflacionistas?

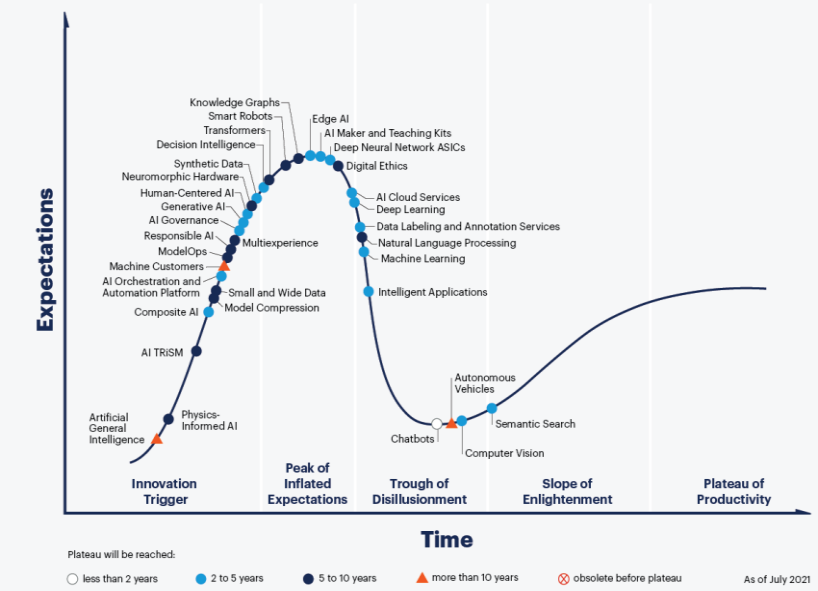


Ilustración 4.11. Hype Cycle for AI, 2021 Gartner

Multitud de estudios (Salud, 2018), auguran que la Analítica Avanzada y la Inteligencia Artificial serán capaces de ofrecer el mejor conocimiento allí donde sea requerido, habiendo sido identificado su potencial para mejorar los resultados en salud y permitir nuevas formas de atención, a la vez que se optimizan los recursos necesarios.

100 MEDIDAS QUE MEJORAN EL SECTOR DE LA SALUD — E&S ECONOMÍA Y SALUD FUNDACIÓN

"Potencial de las tecnologías para relacionarse con paciente, que permitan nuevas formas de atención y la capacidad para mejorar los procesos (incrementando la calidad y ahorrando en el uso de recursos)"

DESCRIPCIÓN	% VOTOS
Análisis avanzado o explotación inteligente de datos	32%
Sistemas de gestión remota de pacientes	18%
Biología sintética	18%
Plataforma de colaboración	12%
Captación masiva de datos	8%
Servicios digitales de valor añadido en el sector farmacéutico	6%
Desarrollo de nuevas tecnologías para el diagnóstico	6%

Ilustración 4.12. Tecnologías por potencial de mejora. Fundación Economía y Salud

Consecuencia de las limitaciones físicas y biológicas de los seres humanos, como el tamaño de nuestro cerebro o el número de neuronas y la velocidad máxima de comunicación entre ellas, además de nuestra necesidad de descansar, dormir, relacionarnos, etc. las personas necesitamos décadas de observación y trabajo para adquirir conocimientos mediante el aprendizaje y la experiencia.

Ilustración 4.13. Evolución del conocimiento de las personas con el tiempo

Por el contrario, un sistema de analítica avanzada o inteligencia artificial, que no es más que un software, se le puede exponer a un enorme volumen de información, recolectada durante décadas, mejor cuantos más datos sí estos disponen de calidad suficiente, e iterar con ellos, sin interrupción y a enormes velocidades, hasta replicar el mejor conocimiento de los seres humanos.

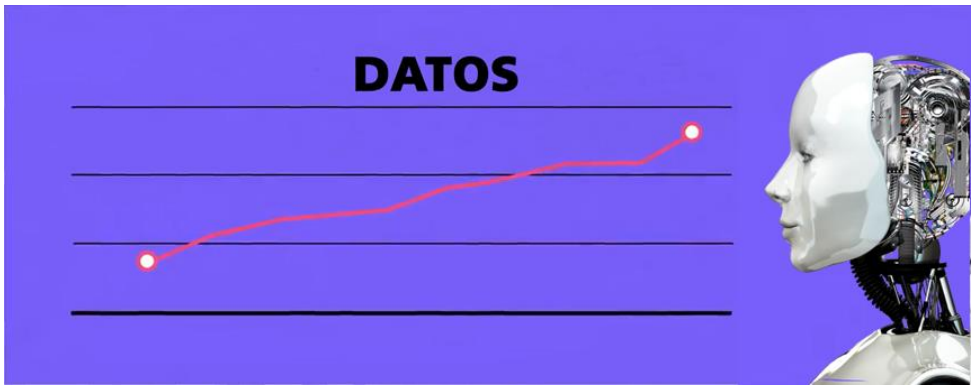


Ilustración 4.14. Evolución del conocimiento de la IA con los datos

Como se expone a continuación, parece qué en este instante estamos asistiendo a la confluencia de una serie de factores que favorecen la obtención de resultados satisfactorios en el desarrollo de actuaciones de Analítica Avanzada e Inteligencia Artificial, estos son:

1. **Datos:** Consecuencia de la digitalización progresiva de los procesos relacionados con la asistencia y la salud, se ha ido generando un ecosistema de sistemas de información, donde se almacenan enormes cantidades de datos, en diferentes formatos, que van desde el nivel molecular hasta el nivel poblacional, pasando por ámbitos de actuación tan diversos como la gestión, la atención primaria, hospitalaria, social, emergencias, la investigación biomédica y farmacéutica, el personal-health o las redes sociales.

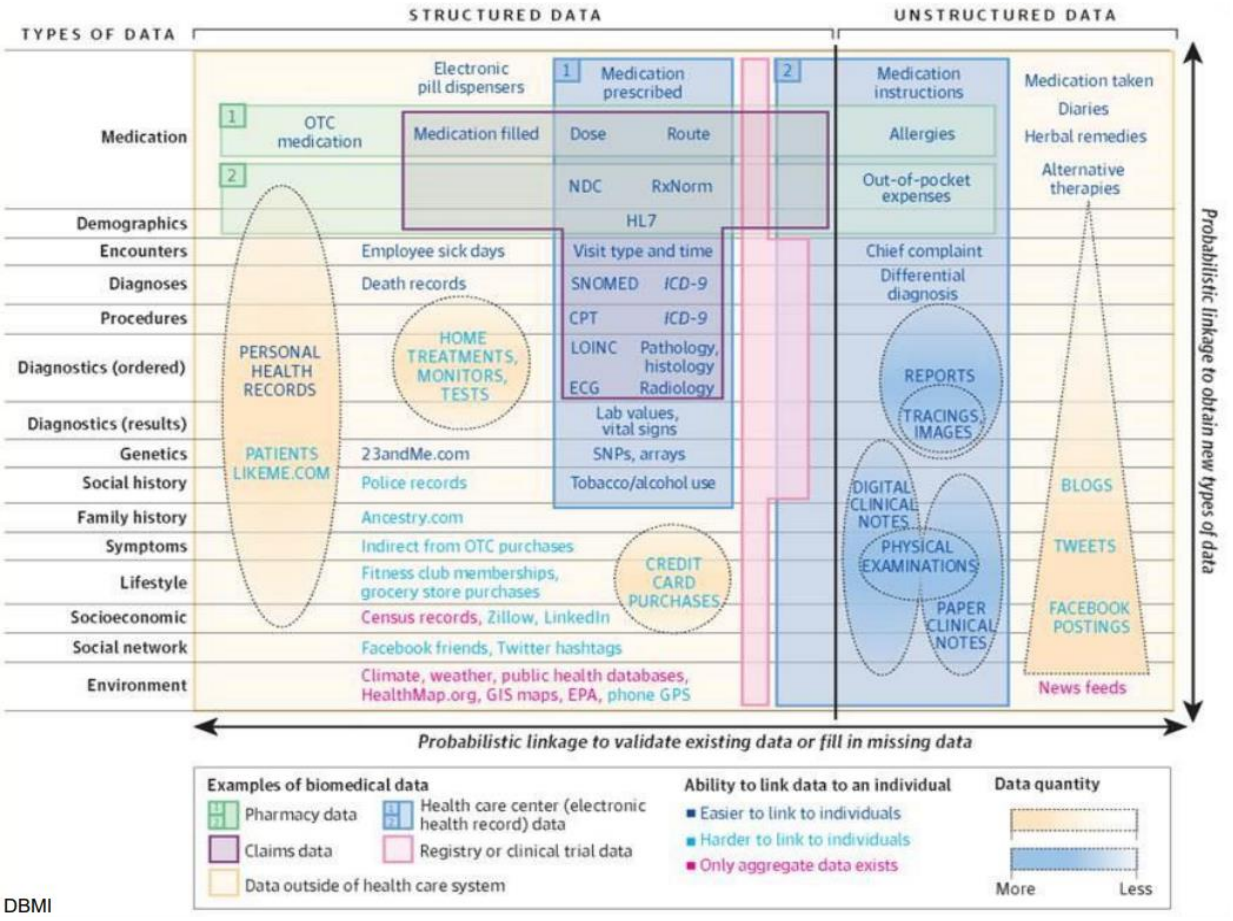


Ilustración 4.15. Tipos de datos por ámbito. Zak Kohane. Havard DBMI

Entre todos estos tipos de datos, hay algunos cuyo tratamiento requiere de un uso intensivo de recursos, como las imágenes de anatomía patológica, o las imágenes procedentes del radio-diagnóstico y la radio-terapia (TAC, RMN, PET/TAC, SPECT/TC, ALP, Eco-Cardiografía, OCT, etc.), los textos libres o los datos óhmicos, de genómica, proteóhmica y metabolóhmica.

Mención especial requieren los datos de genética, tradicionalmente exclusivos del ámbito de la investigación biomédica, aunque se estén extendiendo progresivamente al ámbito clínico, consecuencia de su utilidad diagnóstica en determinados perfiles patológicos y por la reducción de costes experimentada en las técnicas de secuenciación.

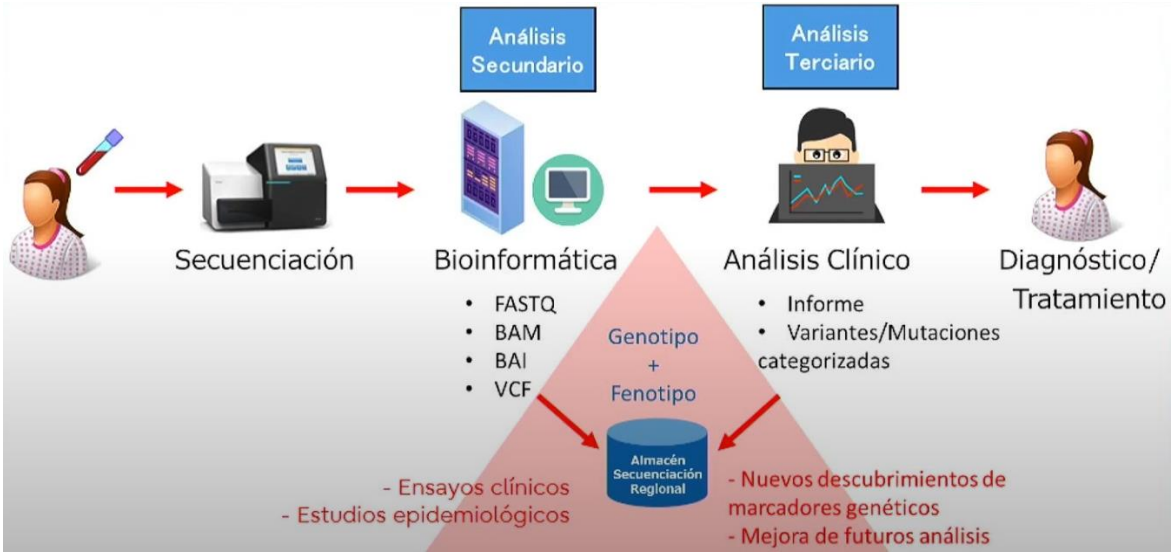
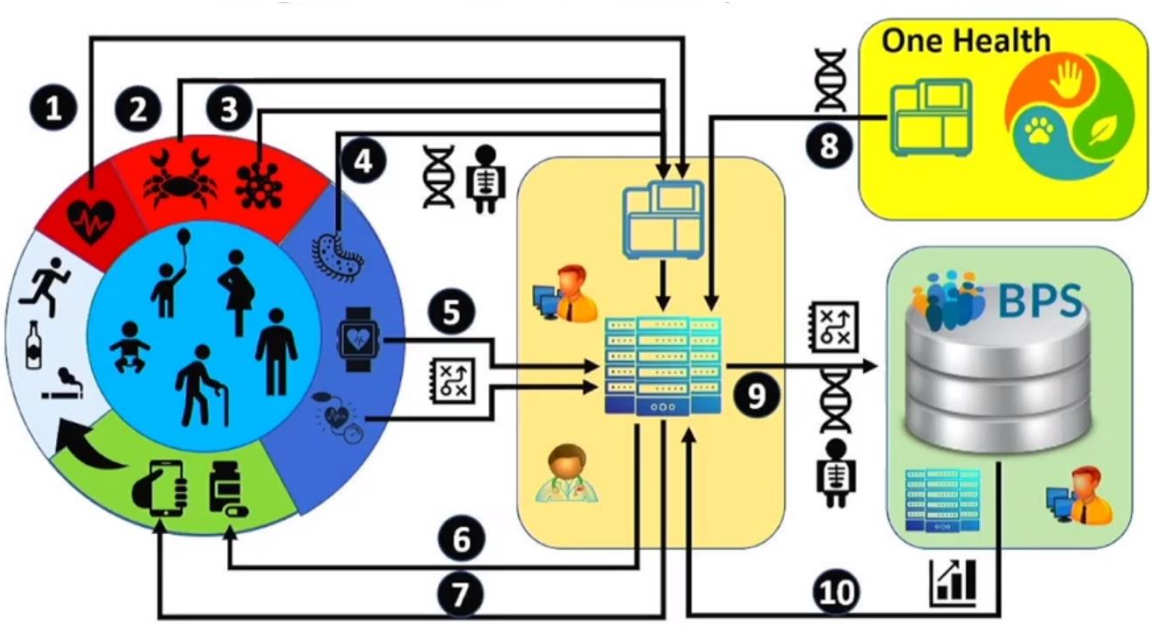


Ilustración 4.16. Flujo generador datos óhmicos en asistencia. XII Foro Interoperabilidad SEIS



1. persona 2. tumor 3. patógeno 4. microbiota y 8. patógeno ambiente, 5. Personal-health. 10 BPS del SSPA

Ilustración 4.17. Óhmicos. Fuente. XII Foro Interoperabilidad SEIS

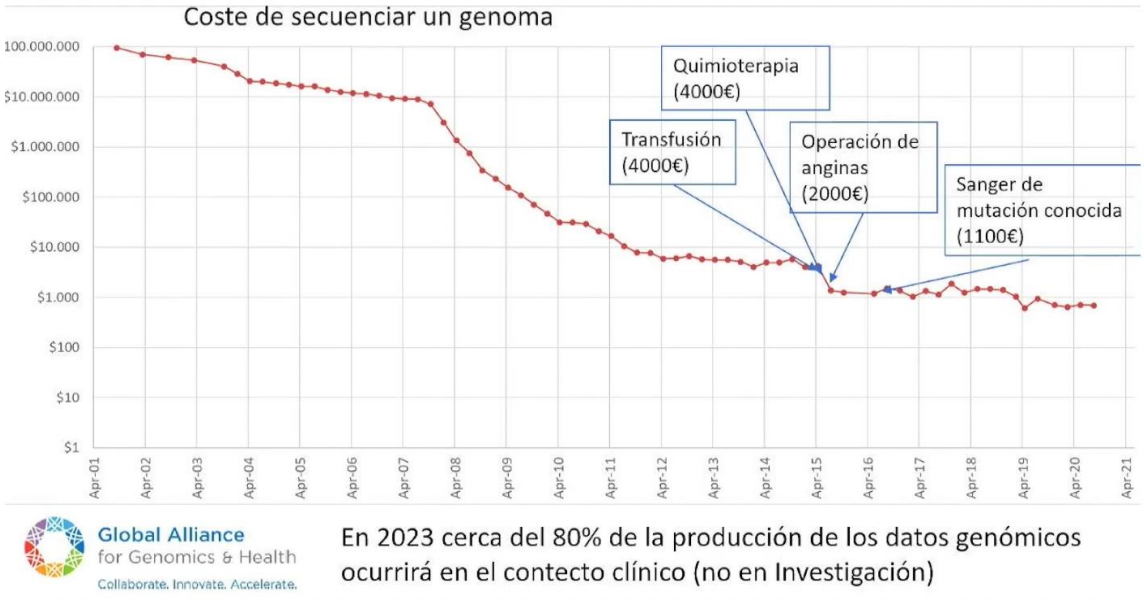


Ilustración 4.18. Evolución del coste secuenciación genoma. Fuente GA4GH

2. **Capacidad de Computo:** La Ley de Moore se basa en la observación de que la capacidad de cómputo de los procesadores, se duplica aproximadamente cada 18 meses. Esta ley, de carácter empírico, se ha cumplido de manera aproximada durante décadas, y ha venido acompañada de un abaratamiento exponencial de los costes de la electrónica digital.

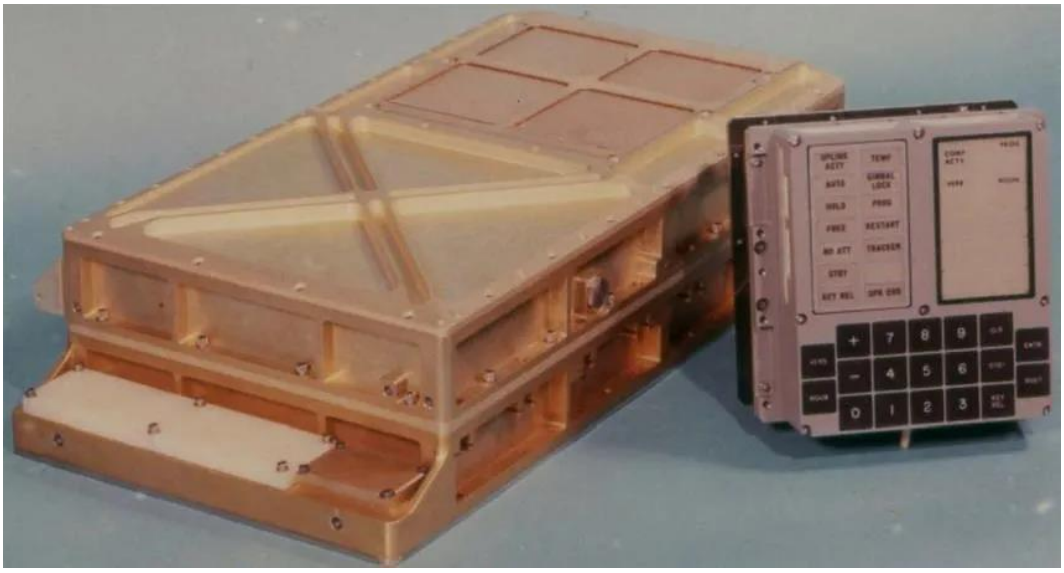


Ilustración 4.19. Block II. Primer Computador con semiconductores. Apolo XI-1969. CPU 2 MHz

3. **Capacidad de Storage:** De una forma similar al principio establecido por Moore, la Ley del Almacenamiento Masivo Digital, se ha traducido en una reducción continua del tamaño y del coste asociado al almacenamiento de los datos.

Ilustración 4.20. Disco Duro de IBM de 5 Mbytes y 1 Tonelada de peso. 1956

4. **Software especializado:** Además de datos, espacio para almacenarlos y capacidad para procesarlos, es necesario contar con herramientas que permitan extraerlos, transformarlos, modelarlos, presentarlos, etc., existiendo un enorme porfolio de soluciones para poder realizar estos trabajos, muchas de ellas han sido desarrolladas y liberadas por los referentes del mercado en el tratamiento de grandes volúmenes de datos (Superset de Air-bnb, Tensor Flow de Google, GPT-3 Open AI, etc.).

   Dada la enorme variedad de herramientas existentes para acometer el tratamiento y análisis de los datos, este TFM incluye un capítulo dedicado a su análisis y especificación de requisitos.

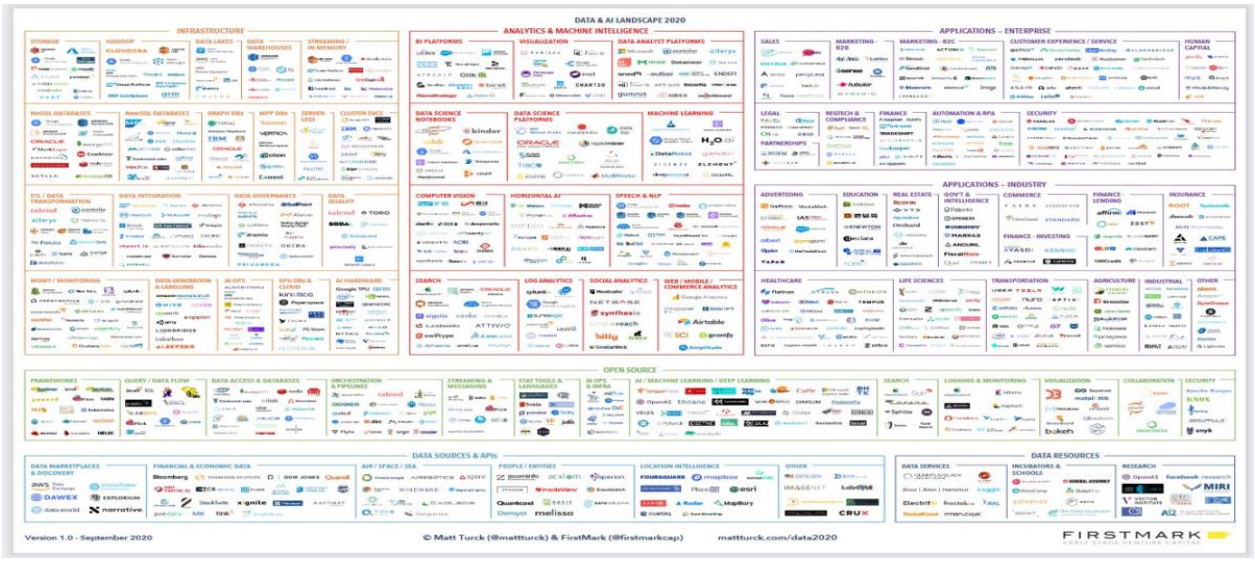


Ilustración 4.21. Mapa de soluciones para Datos e Inteligencia Artificial. Fuente Matt Turck

5. **Nube:** Aun existiendo una evolución suficiente de las tecnologías que requiere un Data Lake Sanitario, los recursos necesarios para movilizarlas pueden exceder, con mucho, los existentes en un departamento de informática sanitaria, ya que la propia naturaleza de los estudios observacionales, o el desarrollo de modelos de inteligencia artificial, puede llevar asociado un consumo muy dispar de recursos, con grandes demandas en instantes puntuales, seguidos de largos periodos de infra-utilización, lo que nos lleva a tener en consideración un consumo flexible de la tecnología como Infraestructure as a Service (IaaS), Platform as a Service (PaaS) y Software as a Service (SaaS), modelos de provisión típicos del Cloud.

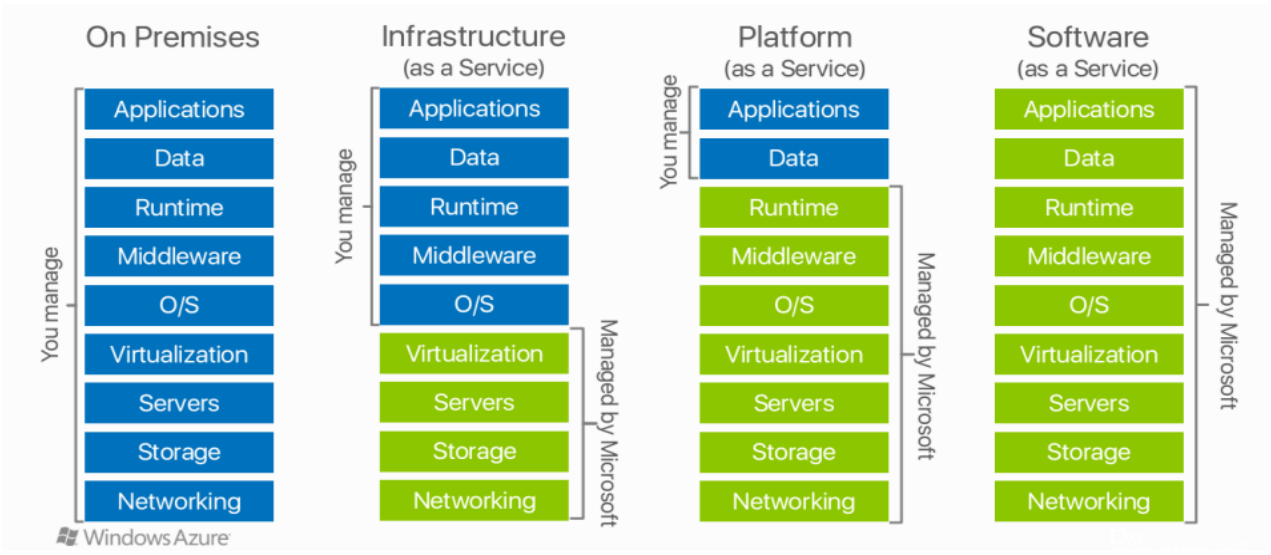


Ilustración 4.22. Comparativa de Modelos de Provisión de recursos. Fuente Microsoft Azure

6. **Financiación:** Un Data Lake Sanitario es una iniciativa que puede requerir de notables inversiones en recursos humanos y tecnológicos, representando los fondos europeos una gran oportunidad para su financiación.

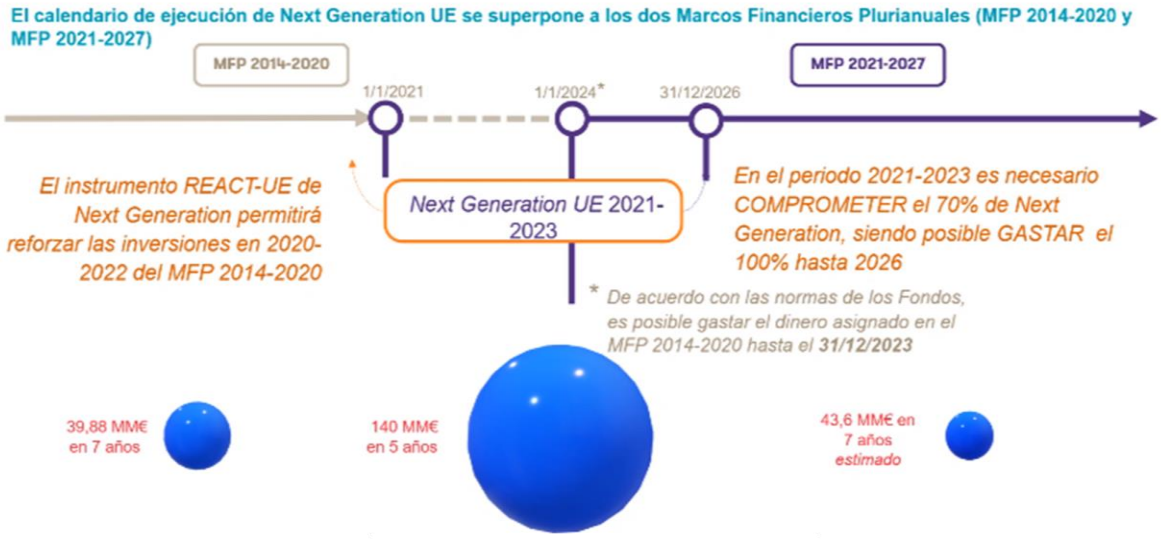


Ilustración 4.23. Fondos UE. MFP y Next-Gen. Fuente: Grant Thornton España

En el instante de elaboración de este TFM, estamos asistiendo a una situación coyuntural, de solapamiento de tres programas de la Unión Europea, los fondos Next-Generation, el MFP 2014-2020 y el MFP 2021-2027.

En el plazo 2021-2023 y únicamente del programa Next-Generation, a través de los subprogramas MRR y REACT, España espera recibir 140.000 millones de euros (72.700 M€ en subvenciones y 67.300 M€ en préstamos), de los que, según el Plan de Recuperación, Transformación y Resiliencia, "España Puede", el 33% deben ser destinados a la Transformación Digital, y aunque la ejecución de estos fondos puede acometerse en el periodo 2021-2026, se plantea comprometer el 70% del gasto entre el 2021 y el 2023.

7. **Voluntad:** Aunque son muchos los factores enumerados y que parecen dibujar un escenario favorable para el desarrollo de la AA y la IA, uno de los más relevantes es la voluntad de los altos cargos y en particular de los responsables de informática para priorizar la implantación de Data Lakes Sanitarios frente a otras actuaciones.

   Según la información recogida en el Índice SEIS del año 2021, los proyectos de Análisis de Datos y Generación de Conocimiento son los más prioritarios.

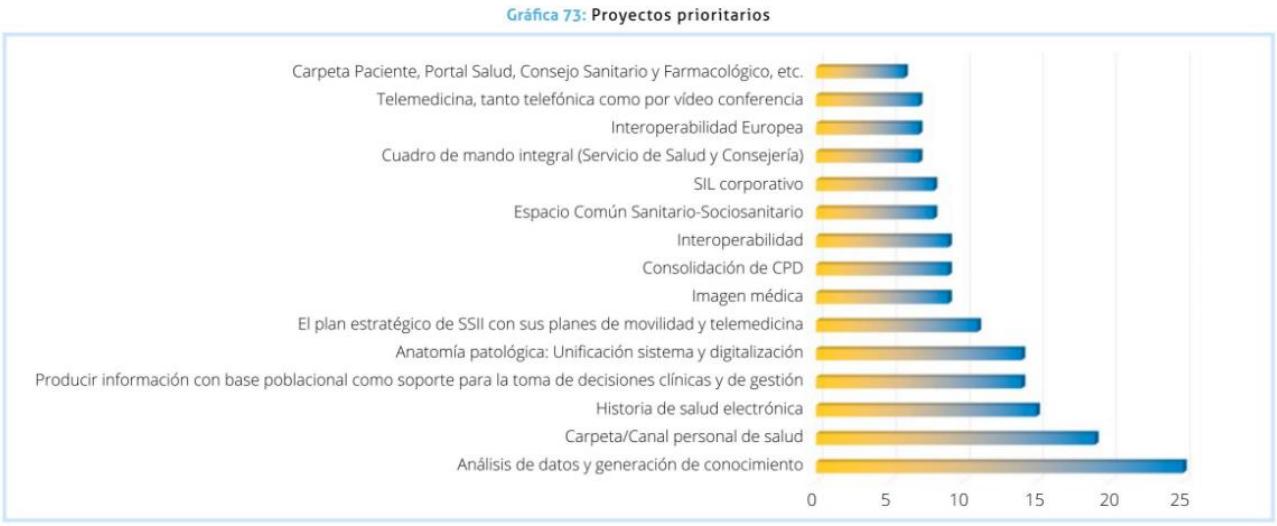


Ilustración 4.24. Proyectos Prioritarios. Índice SEIS 2021

8. **Transformación**: La aparición de Internet y la irrupción de los Smartphones ha supuesto una enorme transformación para muchos sectores, como la banca, las compras o la formación y aunque todavía no se ha experimentado un fenómeno similar en sanidad, éste podría darse con la irrupción de la Inteligencia Artificial y su capacidad para proveer soporte a la decisión, asistiendo a la aparición de un mercado, al que también podrán acceder los pacientes, como ya hacen con otros productos software y apps, pagando por su uso o accediendo a ellos de "forma gratuita" a cambio, normalmente, de sus propios datos, un comportamiento que, a priori, podría parecer un tanto temerario, pero que entendemos es:

    - lícito, porque los datos de salud son propiedad de los pacientes,
    - viable, porque los pacientes tienen acceso a sus datos de salud a través de las carpetas de paciente y las apps de salud,
    - loable, porque los pacientes quieren curarse, incluso aunque eso conlleve, que empresas con intereses económicos, acaben conociendo su identidad y sus problemas de salud,

    La generalización de esta forma de actuar, que ya se da en iniciativas nacionales (Fundación 29, 2022) e internacionales (Patients Like Me, 2022), puede materializarse en una transformación del sistema sanitario desde el lado del paciente, quién podría acabar acudiendo a la consulta médica provisto de la mejor evidencia para el tratamiento de su enfermedad, un cambio de paradigma en la acción prescriptora, que podría traducirse en enormes tensiones económicas para el Sistema Nacional de Salud.

4.4. Iniciativas de uso Secundario[5] de los datos de Salud

El 5 de julio de 2020, la ***Organización Mundial de la Salud*** publicó una propuesta general para la prestación de servicios de salud digital con el objetivo de mejorar la salud y el bienestar y entre sus objetivos estratégicos destaca el "establecimiento de ecosistemas nacionales de salud digital interoperables, así como la promoción del uso del **Big Data y la Inteligencia Artificial** bajo los principios éticos adecuados y una revisión de las regulaciones".

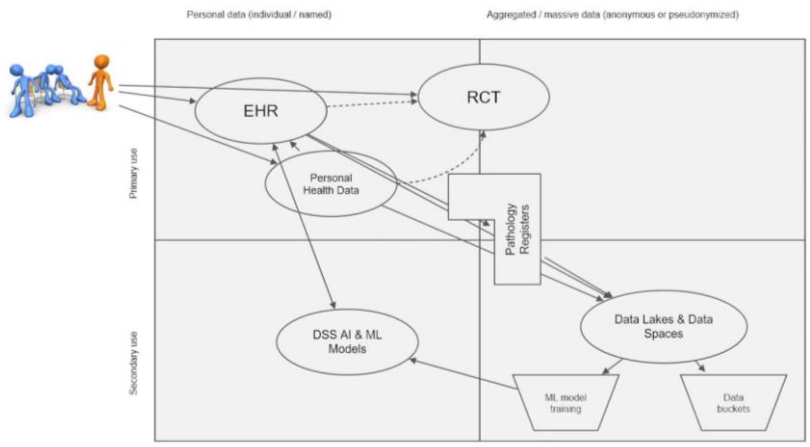
4.4.1. Ámbito Internacional

4.4.1.1. England – Data Saves Lives

La estrategia "Data Saves Lives: reshaping health and social care with data" publicada el 13 de junio de 2022 por el Department of Health and Social Care, establece planes ambiciosos para aprovechar el potencial de los datos, con el objetivo de mejorar el National Health System en:

5 **Uso primario de los datos:** Cuando son utilizados para cumplir con el objeto para el que fueron registrados.
Uso secundario de los datos: Cuando son destinados a otro fin.

En el caso que nos ocupa un dato en la HCE tendrá un uso primario cuando es utilizado para prestar atención sanitaria y podemos decir tendrá un uso secundario para cualquier otro, como su trasferencia a un Data Lake Sanitario. Un ensayo clínico sería una situación excepcional, en el que la información registrada en la HCE tiene como uso primario la investigación y no la asistencia.



Uso primario/secundario, dato agregado/personal.
XII Foro de Interoperabilidad de la SEIS

EHR = HCE. Uso individual e identificado primario

Personal Health Data: Apps de pacientes o similares

Randomize Clinical Trials-RCT = Ensayo Clínico es uso primario.

Data Lakes, uso secundario para:

- Entrenan modelos
- Segmentan en Data Buckets

Los modelos (DSS) se generan con el uso secundario de datos de una población y se consumidos desde la HCE para la asistencia individual, uso primario.

- La atención directa a particulares.
- La salud de la población a través de la focalización proactiva en los servicios.
- La planificación y mejora de los servicios.
- La investigación y la innovación que impulsarán nuevos tratamientos médicos.

Y establece tres compromisos para conseguirlo:

1. invertir en entornos de datos seguros para impulsar la investigación y los tratamientos
2. usar tecnología para permitir que el personal pase más tiempo de calidad con los pacientes
3. brindar a las personas un mejor acceso a sus propios datos

desarrollado una estrategia en siete capítulos:

1. Proporcionar confianza a los ciudadanos de cómo se manejarán sus datos para que estén de acuerdo en que sean utilizados para mejorar la atención que ellos y otros reciben
2. Brindar a los profesionales la información que necesitan para prestar la mejor atención
3. Mejorar los datos para la atención social de adultos
4. Proveer datos como herramienta de soporte a la decisión a los gestores
5. Capacitar a los investigadores en el uso de los datos que necesitan para desarrollar tratamientos, diagnósticos, modelos de atención e ideas que mejoren la vida
6. Fomentar la innovación mediante la colaboración con empresas
7. Desarrollar la infraestructura técnica adecuada, para que las arquitecturas de datos puedan trabajar en conjunto y hacer un uso más eficaz y eficiente de los datos.

Ilustración 4.25. Iniciativa Data Saves Lives

4.4.1.2. Estados Unidos. All of Us y las Big Tech

En enero de 2015, el presidente Barack Obama, anunció la Iniciativa de Medicina de Precisión (PMI), con la que desarrollar la infraestructura nacional necesaria para implementar la medicina de precisión en los Estados Unidos (USA, 2015).

La piedra angular de esta estrategia es el programa de investigación, All of US, liderado por la Agencia Estatal de Investigación, National Institutes of Health (NIH) del Departamento de Salud y Servicios Humanos de los EE. UU. y que tiene como principal objetivo la creación de una cohorte de un millón de voluntarios que refleje la diversidad de Estados Unidos, para contribuir con sus datos de salud y muestras biológicas a la investigación y desarrollo de la medicina personalizada.

A fecha de 23 de Mayo de 2022, el programa cuenta con 29 institutos, centros y oficinas de los Institutos Nacionales de la Salud y dispone de un Data Lake Sanitario con datos enriquecidos disponibles a través del portal "All of Us Researcher Workbench", una infraestructura en cloud, con aproximadamente 100.000 secuencias genómicas completas y 330.000 respuestas a los participantes encuestados, así como información de 214.000 registros de salud electrónicos, junto con mediciones físicas y datos de dispositivos portátiles.

Aun no tratándose de iniciativas de carácter público, es obligada la referencia a las tres Big Techs (Amazon Web Services, Microsoft Azure y Google Cloud y adicionalmente IBM Watson), quienes dan soporte a muchas de las iniciativas en torno al big data, la analítica avanzada y la inteligencia artificial con sus herramientas e infraestructuras en la nube y quienes disponen de algunos de las referencias más destacadas en este ámbito, aunque también con sonados fiascos, como la demanda colectiva a la que se enfrente Google Deep Mind por el uso de los registros de salud de 1,6 millones de pacientes en el Reino Unido, o la prescripción del "doctor Watson" de tratamientos contra el cáncer no seguros para los pacientes.

4.4.1.3. República Popular China

Conjuntamente con Estados Unidos y Europa, China está llamada a liderar el desarrollo de la analítica avanzada y la inteligencia artificial por medio de la explotación masiva de datos de salud, aunque a diferencia de sus competidores, China parte de una situación más ventajosa consecuencia de las facilidades jurídicas para la obtención masiva de datos.

Algunas iniciativas ya han trascendido al hermetismo que caracteriza al gigante asiático, como la existente en materia de seguridad, con un sistema de más de 170 millones de cámaras de video-vigilancia situadas en espacios públicos y capaces de identificar hasta 120 personas por segundo aplicando tecnologías de inteligencia artificial para el reconocimiento facial (LaRazón, 2018).

4.4.2. Unión Europea

4.4.2.1. Espacio de Datos Europeo, Ley de Datos y Ley de Gobernanza de Datos

Los Espacios de Datos son un elemento fundamental de la **Estrategia Europea de Datos**, que, entre otras cuestiones, busca impulsar la economía de la región a través de la creación de un mercado único europeo de datos, donde estos fluyan entre los diferentes Estados Miembros y entre sectores de actividad, de acuerdo a los valores europeos de autodeterminación, privacidad, transparencia, seguridad y competencia leal, impulsando el desarrollo de nuevos productos y servicios basados en datos.

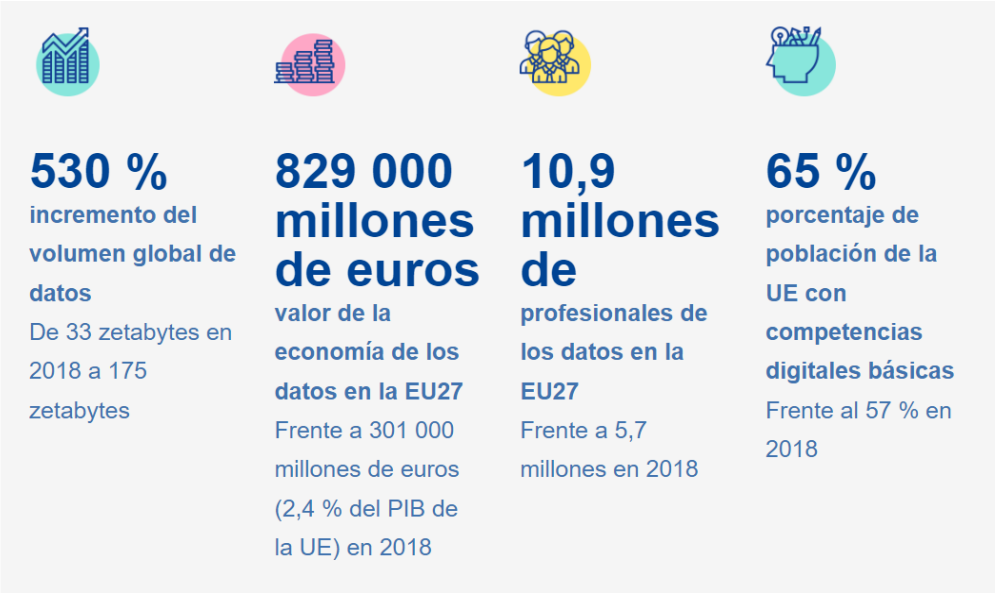


Ilustración 4.26. Cifras previstas por la UE para 2025

En dicha estrategia, la Comisión Europea anunció su interés en invertir y desarrollar espacios de datos comunes en sectores económicos estratégicos y de interés público, destacando los relacionados con la fabricación, la energía sostenible, la movilidad, el ámbito financiero, la energía, el sector agrario, las administraciones públicas y la salud y una vez desarrollados, se plantea su interconexión para que se puedan explotar de manera cruzada.

La creación de estos espacios de datos debe superar las barreras legales y técnicas ligadas a la compartición de datos, mediante normas, herramientas e infraestructuras comunes en un contexto de soberanía digital, es por ello qué según la estrategia europea de datos, el desarrollo de los espacios de datos debe ser realizado teniendo en cuenta los siguientes elementos:

- El despliegue de herramientas y servicios para el tratamiento, intercambio y compartición de datos, de forma justa, transparente proporcional y no discriminatoria, así como la federación de capacidades en la nube, seguras y eficientes desde el punto de vista energético y de sus servicios relacionados.

- El desarrollo de estructuras claras y fiables de gobernanza de los datos, en conformidad con la legislación de la UE, prestando especial atención a la protección de los datos personales, del consumidor y el derecho a la competencia.

- La mejora de la disponibilidad, la calidad y la interoperabilidad de los datos, tanto en ámbitos específicos como entre sectores.

Por su parte, la Comisión Europea adoptó el 23 de febrero de 2022 la propuesta de Reglamento sobre normas armonizadas para el acceso justo a los datos y su uso, también conocido como **Ley de Datos o "Data Act"**, que pretende ser un pilar clave de la estrategia europea para los datos ya que pondrá más datos a disposición y en beneficio de empresas, ciudadanos y administraciones públicas por medio de:

- Medidas para aumentar la seguridad jurídica de las empresas y los consumidores que generan datos, quién puede utilizar qué datos y en qué condiciones, e incentivos para que los fabricantes sigan invirtiendo en la generación de datos de alta calidad.

- Medidas para garantizar la equidad mediante el establecimiento de normas relativas al uso de los datos generados por los dispositivos del Internet de las cosas (IoT).

- Medidas para evitar el abuso de los desequilibrios contractuales que dificultan el intercambio justo de datos.

- Medios para que los organismos del sector público accedan y compartan datos con el sector privado, que sean necesarios para cumplir con fines específicos de interés público.

- Nuevas reglas que establezcan las condiciones marco adecuadas, para que los clientes cambien de manera efectiva entre diferentes proveedores de servicios de procesamiento de datos y desbloquear el mercado de la nube de la UE.

Complementa a la Ley de Datos el **Reglamento de Gobierno de Datos o "Data Governance Act",** de aplicación a partir de septiembre de 2023, como primer entregable de la estrategia europea de datos, para aumentar la confianza en el intercambio de datos, fortalecer los mecanismos para aumentar la disponibilidad de datos y superar los obstáculos técnicos para la reutilización de datos, así como el apoyo para la creación y el desarrollo de espacios de datos europeos comunes en dominios estratégicos, como la salud.

El reglamento de gobernanza de datos pretende garantizar el acceso a más datos para la economía y la sociedad de la UE y proporcionará un mayor control para los ciudadanos y las empresas sobre los datos que generan, lo que en el ámbito de la sanidad podría traducirse en la creación de **espacios de datos personales**, en el que los europeos obtengan un mayor control sobre sus datos y **decidan a nivel detallado quién tendrá acceso a sus datos y con qué propósito**.

Esta Ley también recoge la necesidad de desarrollar un formulario de consentimiento europeo común para **entidades sin ánimo de lucro dedicadas al altruismo de datos,** con el que permitir la recopilación de datos en todos los Estados miembros en un formato uniforme, quedando recogidas un registro público de nueva creación.

Otra figura que también se recoge en la ley es **el intermediario para el intercambio de datos,** entidades que serán controladas por las autoridades públicas y se encargarán de agrupar y organizar los datos de manera neutral para que las empresas compartan sus datos sin que esto implique una pérdida de ventaja competitiva o represente un riesgo de mal uso.

Por su parte, la **Directiva de Datos Abiertos,** de junio de 2019, establece las reglas para la reutilización de la información del sector público, teniendo en consideración el desafío que representa el encontrar formas de permitir que se extraiga el conocimiento de los datos, al tiempo que se preserva completamente la privacidad u otros derechos que pueden estar vinculados a los datos.

También se prevé la creación de un **Consejo Europeo de Innovación de Datos** para facilitar el intercambio de mejores prácticas por parte de las autoridades de los Estados miembros, en particular sobre el altruismo de datos, los intermediarios de datos y el uso de datos públicos que no pueden estar disponibles como datos abiertos, además de asesorar a la Comisión sobre la priorización de estándares de interoperabilidad intersectorial.

4.4.2.2. Espacio de Datos de Salud Europeo

Mediante la creación del **Espacio de Datos de Salud Europeo (European Health Data Space), en adelante EHDS,** los registros sanitarios digitalizados recopilados dentro del espacio europeo podrán contribuir a un mejor tratamiento de las principales enfermedades crónicas, como el cáncer y las enfermedades raras, pero también a la igualdad de acceso a servicios sanitarios de alta calidad para todas las personas.

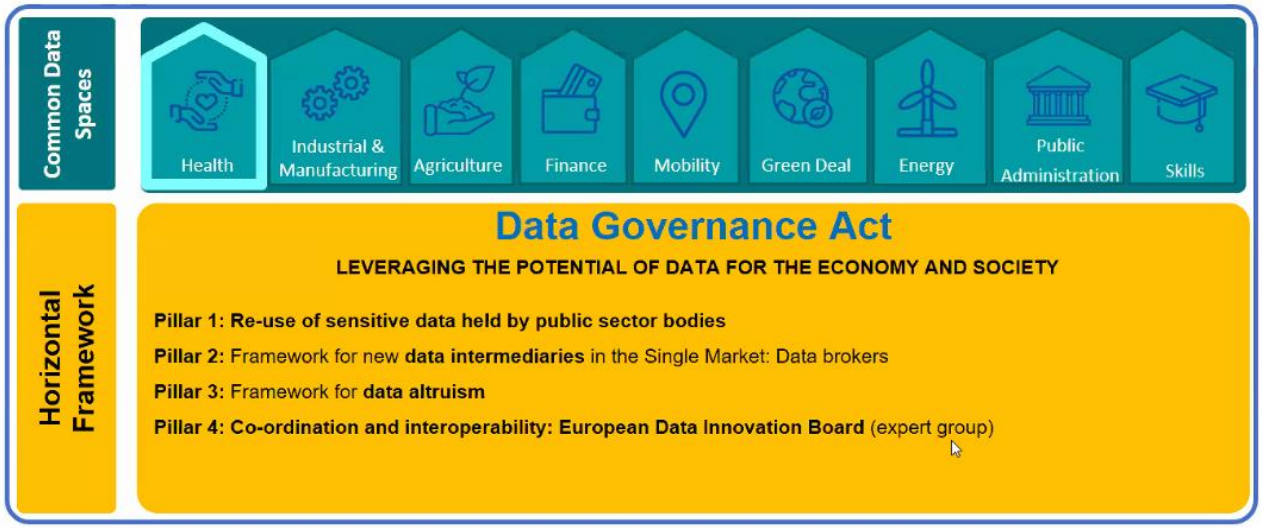


Ilustración 4.27. Pilares del Espacio Europeo Común de Datos de Salud

Con este propósito, el EHDS propone en una doble línea de actuación:

- **EHDS1**, un proyecto destinado a la interoperabilidad de las Historias Clínicas Electrónicas con el objetivo de prestar asistencia.

- **EHDS2**, para habilitar el análisis secundario de los datos con el fin de brindar mejores servicios de atención médica y el desarrollo de la medicina personalizada, el cuidado de las personas y promover la innovación, como el desarrollo de nuevos medicamentos y el impulso de la formulación de políticas basadas en el conocimiento.

El EHDS ayudará a la UE a dar un salto cualitativo en cuanto a cómo se prestan los servicios de atención sanitaria en toda Europa y en particular se prevé que:

- Permita a las personas controlar sus datos sanitarios, tanto si se encuentran en su país de origen como en otro Estado miembro, fomentando un auténtico mercado único de servicios y productos sanitarios digitales.

- Proporcione un marco eficiente, fiable y coherente para usar los datos sanitarios en investigación, innovación, salud pública, elaboración de políticas y reglamentación, a la vez que se garantiza el pleno cumplimiento de las estrictas normas de protección de datos de la UE.

- Las personas que trabajan en investigación e innovación, las instituciones públicas o la industria tendrán acceso, bajo condiciones estrictas, a grandes cantidades de datos sanitarios de alta calidad, que serán esenciales para desarrollar medicamentos, vacunas o tratamientos

capaces de salvar vidas, y que garantizarán un mejor acceso a la asistencia sanitaria y unos sistemas de salud más resilientes.

- Se aproveche del despliegue, actual y futuro, de bienes digitales públicos en la UE, como la inteligencia artificial (IA), la informática de alto rendimiento, la nube y los soportes intermedios inteligentes, contando con el apoyo de marcos normativos en materia de IA, identidad electrónica y ciberseguridad.

En este instante, la definición de este espacio de datos de salud europeo es liderado por la iniciativa **Towards European Health Data Space - TEHDAS**, integrada por entidades de 25 países y que trabaja en proponer una arquitectura, donde cada estado disponga de un nodo concentrador de la información nacional, creando una infraestructura europea federada, en la que también participarán grandes organismos europeos, como la Agencia Europea del Medicamento (EMA).

El proyecto TEHDAS está dividido en ocho paquetes de trabajo liderados por organizaciones de diferentes países. Los paquetes de trabajo de 1 a 3 están vinculados a la ejecución de la acción conjunta y los paquetes 4 a 8 son paquetes de trabajo temáticos vinculados al uso secundario de datos de salud.

1. **Coordinación:** Coordina y gestiona el proyecto y su ejecución.

2. **Difusión:** Comunica los resultados y entregables del proyecto.

3. **Evaluación:** Evalúa si el proyecto está alcanzando sus objetivos.

4. **Divulgación, compromiso y sostenibilidad:** Participa en el diálogo con las autoridades sanitarias nacionales de los países participantes y las partes interesadas internacionales e incorpora sus puntos de vista en el proyecto. Garantiza que los resultados del proyecto se integren en la futura legislación sanitaria de la UE, en particular en el EHDS.

5. **Compartir datos para la salud:** Desarrolla opciones de modelos de gobernanza para el intercambio y uso secundario de datos sanitarios entre países europeos, basados en la transparencia, la confianza, el empoderamiento ciudadano y el bien común. Proporciona recomendaciones para los países europeos sobre la planificación de la legislación nacional para permitir el intercambio transfronterizo y el uso secundario de datos de salud.

6. **Excelencia en la calidad de los datos:** Proporciona soluciones para el uso secundario confiable de datos de salud y atención médica, con el fin de promover la transformación digital de los sistemas de salud europeos. Desarrolla una guía para garantizar la calidad de los datos, la anonimización de los datos y el manejo de la disparidad de datos. El grupo de trabajo Data Quality Assurance Framework, es liderado por el Instituto Aragonés de Ciencias de la Salud.

7. **Uniendo los puntos:** Proporciona opciones para la interoperabilidad técnica en el uso secundario de datos sanitarios en el Espacio Europeo de Datos Sanitarios. Fomenta la participación de los futuros usuarios del Espacio Europeo de Datos Sanitarios, como investigadores y responsables políticos, y de los implementadores técnicos, como empresas e instituciones, en el diseño conjunto de los servicios.

8. **Ciudadanos:** Busca obtener una mejor comprensión de la actitud de los ciudadanos ante el intercambio de sus datos de salud. Identifica formas de informar a las personas sobre el uso de sus datos de salud y crear conciencia sobre los beneficios que ofrece el uso secundario de los datos.

4.4.2.3. EHDEN

La iniciativa **European Health Data and Evidence Network (EHDEN)**, forma parte del programa **Big Data for Better Outcomes (BD4BO),** y fue constituida en 2018 con fondos procedentes del programa Horizonte 2020 de la Unión Europea, en concreto de la Iniciativa de Medicamentos Innovadores, respaldada por la Unión Europea y la Federación Europea de Industrias y Asociaciones Farmacéuticas (EFPIA).

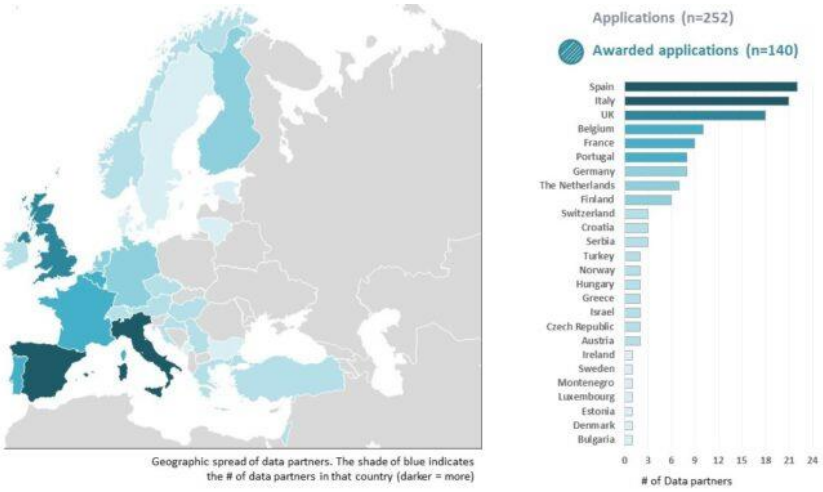


Ilustración 4.28. EHDEN en Europa

EHDEN financia la implementación de repositorios en el modelo **Observational Medical Outcomes Partnership - Common Data Model, OMOP-CDM** en adelante, para habilitar la interoperabilidad en la red EHDEN, además de ofrecer formación y otros recursos para el uso de las herramientas del proyecto **Observational Health Data Sciences and Informatics (OHDSI)**.

Coordinada por el centro médico de la Universidad Erasmus de Rotterdam, aunque el objetivo inicial del consorcio EHDEN era el análisis a gran escala de datos de salud en Europa, mediante la construcción de una red de datos federados que permitiese el acceso a los datos de 100 millones de ciudadanos de la UE, tres años después cuenta con más de 140 socios de datos repartidos en 26 países y armoniza más de 500 millones de registros de salud.

La red de EHDEN ofrece la posibilidad de participar en grandes estudios internacionales y facilita la interoperabilidad por medio del uso de un modelo de datos común, OMOP y la privacidad por diseño mediante un modelo de operación federado, además de ofrecer una serie de herramientas que permiten utilizar los datos y visualizarlos en un formato estandarizado de forma gratuita.

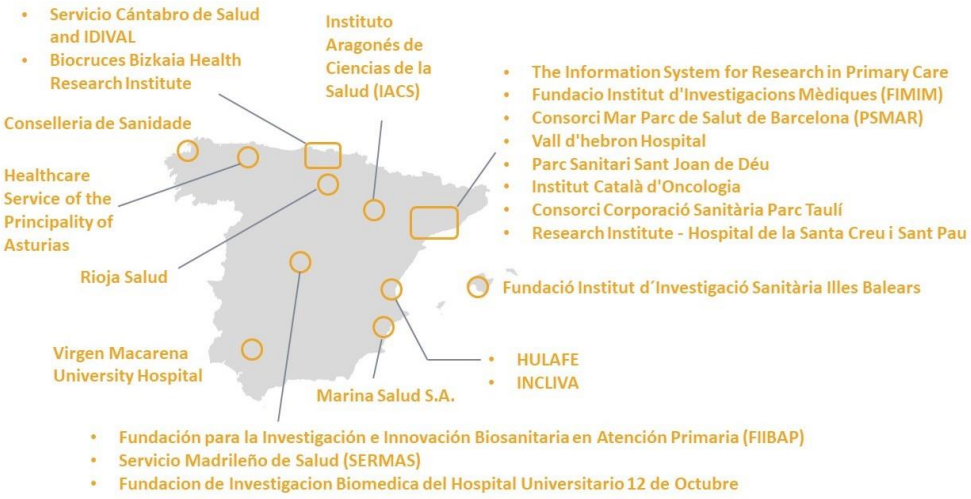


Ilustración 4.29. EHDEN en España

4.4.2.4. European Medicines Agency - EMA - Darwin

En febrero de 2022, la **Agencia Europea del Medicamento (EMA)** ha designado como coordinador del proyecto **"Darwin EU"** al Centro Medico de la Universidad Erasmus de Rotterdam, de igual modo a como también hizo el consorcio EHDEN.

Darwin EU facilitará a todos los reguladores nacionales "el acceso a pruebas fiables sobre datos del mundo real, sobre enfermedades, poblaciones de pacientes y el uso, seguridad y eficacia de los medicamentos".

Esta infraestructura puede constituir la primera implantación del Espacio Europeo de Datos de Salud y además es posible que **favorezca la confluencia del proyecto EHDEN con el EHDS2**, estableciendo OMOP-CDM como el modelo de datos de referencia para la interoperabilidad de datos en la Unión Europea.

4.4.2.5. HRIC, EOSC-Life, Healthy Cloud y otras iniciativas para I+D+i en Salud

La **Nube Europea de Investigación e Innovación en Salud (HRIC)** es otra de las futuras piezas que aspiran a ser fundamentales en la creación de un Espacio Europeo de Datos de Salud contribuyendo a la Nube Europea de Ciencia Abierta, **European Open Science Cloud (EOSC)**, una iniciativa para proporcionar a los investigadores, innovadores, empresas y ciudadanos europeos un entorno multidisciplinario federado y abierto donde puedan publicar, encontrar y reutilizar datos, herramientas y servicios para la I+D+i y con fines educativos.

EOSC-Life reúne las 13 infraestructuras de investigación biológicas y médicas del **European Strategy Forum on Research Infrastructures (ESFRI)**, denominadas **Biological and Medical Research Infrastructures (ESFRI BMS RI)**, para crear un espacio colaborativo abierto para analizar y reutilizar las enormes cantidades de datos producidos por las ciencias de la vida.

Mediante la publicación de datos y herramientas en una nube europea, EOSC-Life tiene como objetivo acercar las capacidades de los grandes proyectos científicos a la comunidad investigadora en general, y hará que los recursos de datos de BMS RI sean FAIR y los publicará en EOSC siguiendo pautas y estándares, con el objetivo de que los científicos de salud puedan encontrar, acceder e integrar datos de ciencias de la vida para su análisis y reutilización en la investigación académica e industrial.

Por su parte el proyecto **Healthy Cloud,** trabaja en elaborar una Agenda Estratégica con una hoja de ruta para implementar el ecosistema HRIC y se organiza en torno a cuatro objetivos fundamentales:

1. interacciones con las partes interesadas.
2. la inclusión de aspectos éticos, legales y sociales en el diseño del futuro ecosistema HRIC.
3. el acceso, uso y reutilización sostenible de datos relacionados con la salud considerando una adopción progresiva de los principios FAIR.
4. las soluciones tecnológicas en términos de instalaciones, para permitir el análisis de datos de salud distribuidos en toda Europa.

además de impulsar dos casos de uso, en cáncer y fibrilación auricular, para garantizar que las propuestas de los expertos sean técnica y éticamente sólidas y legales.

El consorcio Healthy Cloud reúne a 21 organizaciones, incluidas cinco infraestructuras de investigación:

- **ELIXIR**, una organización presente en 23 países europeos que colabora mediante un modelo de 'Hub and Nodes', donde se coordinan y desarrollan recursos para que los investigadores puedan encontrar, analizar y compartir datos, intercambiar conocimientos e implementar las mejores prácticas más fácilmente.

- **ECRIN**, es la red europea de infraestructuras de investigación clínica, una organización pública sin ánimo de lucro que vincula a socios y redes científicas de toda Europa para facilitar la investigación clínica multinacional, proporcionando a sus patrocinadores e investigadores asesoramiento, servicios de gestión y herramientas para superar los obstáculos de los ensayos multinacionales y mejorar la colaboración.

- **EATRIS**, es la infraestructura europea para la medicina traslacional, donde se reúnen recursos y servicios para comunidades de investigación puedan traducir los descubrimientos científicos en beneficios para los pacientes. Brindan acceso a una amplia gama de experiencia e instalaciones preclínicas y clínicas, que están disponibles en más de 127 centros académicos en toda Europa, enfocándose en mejorar y optimizar el desarrollo clínico, preclínico y temprano de medicamentos, vacunas y diagnósticos, y superar las barreras para la innovación en salud.

- **BBMRI-ERIC** (Biobanking and BioMolecular resources Research Infraestructure – European Research Infraestructure Consortium) tiene como objetivo establecer, operar y desarrollar una infraestructura de investigación distribuida paneuropea para facilitar el acceso a los recursos biológicos (bio-bancos), así como a las instalaciones y para apoyar la alta calidad en investigación biomolecular y biomédica.

- **Euro-BioImaging**, es una infraestructura de investigación que brinda acceso abierto, servicios y capacitación a una amplia gama de tecnologías de imágenes biológicas y médicas de última generación.

y tres acciones conjuntas relacionadas con la investigación en salud:

- **InfAct,** Information Action https://www.inf-act.eu/
- **iPAAC,** Innovative Partnership for Action Against Cancer
- **eHAction,** Joint Action supporting the ehealth Network

además de estar profundamente involucrado con iniciativas paneuropeas relacionadas con el programa de investigación **ELSI** (Ethical, Legal and Social Implications) **Research Program** desarrollado por el **NHGRI** (National Humane Genome Research Institute) de Estados Unidos en 1990, especialmente con el Código de conducta para la investigación en salud, liderado por BBMRI-ERIC y cuyo objetivo es unificar y aclarar los términos de los elementos relacionados con la salud del RGPD.

Por su parte el proyecto **PHIRI** (Population Health Information Research Infrastructure), plantea el despliegue de la infraestructura de investigación sobre información de salud de la población y tiene como objetivo facilitar y generar la mejor evidencia disponible para la investigación sobre la salud y el bienestar de las poblaciones afectadas por el COVID-19.

Además, al hacerlo, PHIRI está sentando las bases para construir una **Infraestructura Distribuida en Salud de la Población (DIPoH)** para ser utilizada en superar futuras crisis y asegurar la sostenibilidad del proyecto, además de apoyar la investigación en toda Europa a través de la identificación, el acceso, la evaluación y la reutilización de datos sanitarios y no sanitarios de la población para respaldar las decisiones de políticas de salud pública.

PHIRI se ha basado en los logros de los proyectos **BRIDGE Health** (BRidging Information and Data Generation for Evidence-based Health) y la Joint Action on Health Information (**Inf-Act**), y cuenta con 41 socios en 30 países durante un período de 36 meses (noviembre de 2020 - noviembre de 2023).

4.4.2.6. Un millón de Genomas (1+MG -> B1MG) y European Genome-phenome Archive

España es uno de los 23 países firmantes de **1+ Million Genomes (1+MG),** para acceder a un millón de genomas secuenciados, junto a otros datos clínicos en la Unión Europea, una iniciativa interoperable y transfronteriza de bases de datos de genomas que facilite la investigación, la prevención, el diagnóstico y el tratamiento de enfermedades.

Para cumplir con sus objetivos, cuenta con el soporte del proyecto **Beyond 1+ Million Genomes (B1MG),** qué en una primera fase, apoya y coordina a nivel operativo la implementación de la hoja de ruta sobre la configuración de la infraestructura, orientación legal y técnica, estándares de datos, y los requisitos y las mejores prácticas para permitir el acceso a los datos, además de plantear el desarrollo de una infraestructura sostenible para compartir datos.

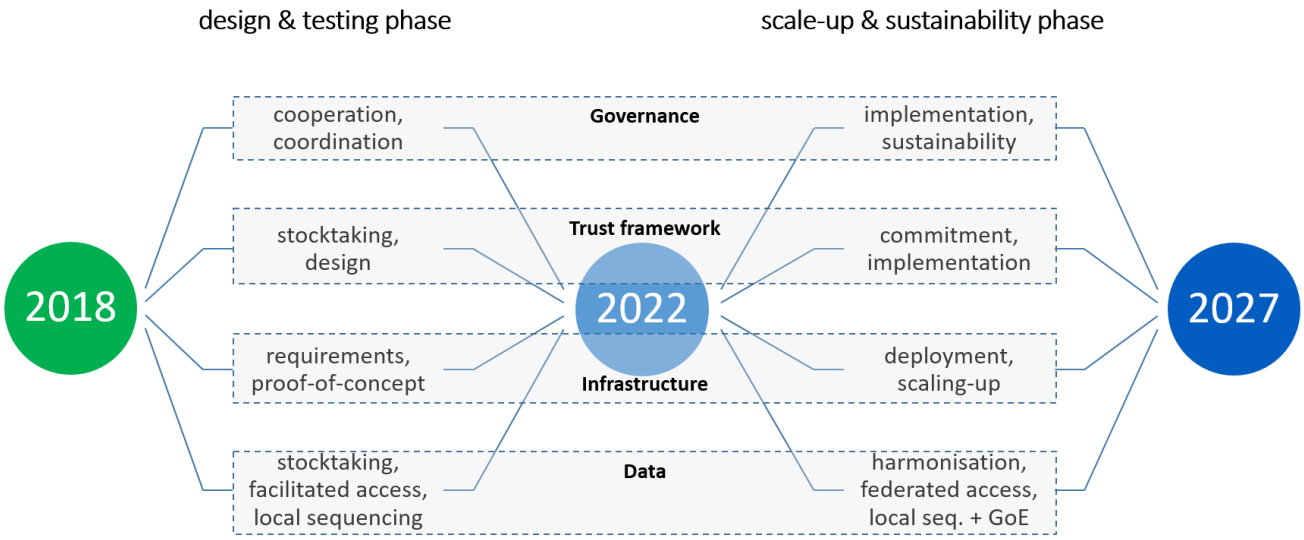


Ilustración 4.30.Hoja de Ruta B1MG

En una segunda fase B1MG, abordará la implementación financiado por el programa Digital Europe, estableciendo una infraestructura de datos federados para datos genómicos y clínicos en toda Europa que permitirá el aprendizaje distribuido para varios casos de uso y proporcionará una gobernanza de acceso a datos y un mecanismo de coordinación sostenible, y contribuirá a mejorar la interoperabilidad de los datos genómicos y clínicos disponibles para el acceso.

Por su parte el **European Genome-phenome Archive (EGA)**, (https://ega-archive.org) es un recurso para el archivo seguro a largo plazo de todo tipo de datos genéticos, fenotípicos y clínicos potencialmente identificables como resultado de proyectos de investigación biomédica.

Lanzada en 2008, la EGA ha crecido rápidamente y actualmente archiva más de 4.500 estudios de casi mil instituciones. Dado el tamaño y valor de los datos alojados, la EGA mejora constantemente su cadena de valor, facilitando su envío, descubrimiento, acceso y distribución, así como liderando el diseño e implementación de estándares y métodos necesarios para entregar la cadena de valor.

EGA se ha convertido en un proyecto impulsor clave de GA4GH, liderando múltiples esfuerzos de desarrollo e implementando nuevos estándares, herramientas y principios, como FAIR, y ha sido designado como recurso de datos básicos de ELIXIR.

4.4.2.7. IDSA y GAIA-X

Los espacios de datos facilitaran que diversos actores compartan datos de manera voluntaria, segura y sigan mecanismos comunes de gobernanza, organizativos, normativos y técnicos, para lo que se requiere del desarrollo de un ecosistema donde se definan modelos de referencia seguros para el intercambio de datos.

El modelo de arquitectura abierto de referencia **IDS-RAM (International Data Spaces Reference Architect Model)**, es una iniciativa elaborada por la IDS (International Data Space Association) y avalada por la Unión Europea y que actualmente integran 133 entidades y conecta con otras iniciativas europeas, incluyendo BDVA, FIWARE y Platform Industrie 4.0, participando en más de veinte proyectos de investigación europeos, principalmente a través del programa Horizonte 2020.

Otra de las actuaciones a considerar es **GAIA-X**, una iniciativa europea del sector privado para la creación de una infraestructura de datos abierta, federada e interoperable, constituida sobre los valores de soberanía digital, disponibilidad de los datos y el fomento de la economía del dato.

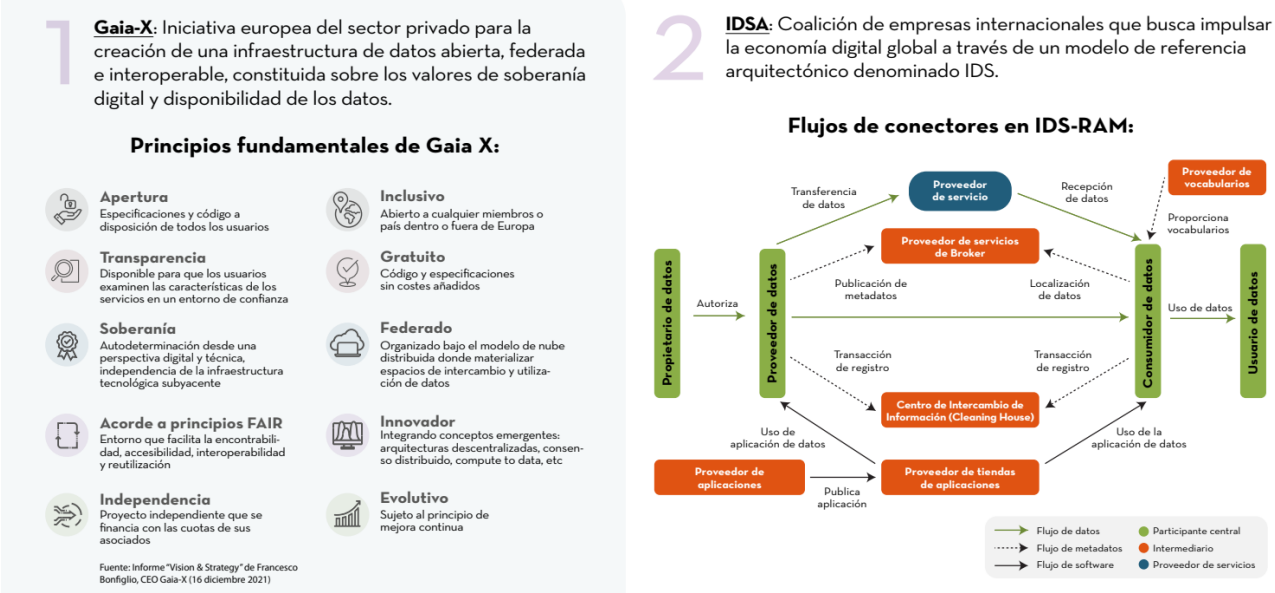


Ilustración 4.31. Comparativa Gaia-X e IDS. Fuente datos.gob.es

4.4.3. Ámbito Nacional

4.4.3.1. Estrategia de Salud Digital. Espacio Nacional de Datos de Salud (ENDS)

La **Agenda España Digital 2025** define su noveno eje estratégico como el de la "Economía del Dato e Inteligencia Artificial (IA)" y fija cuatro objetivos específicos:

1. Convertir a España en un referente en la transformación hacia una Economía del Dato.
2. Desarrollar un marco ético y jurídico para la IA basado en valores compartidos.
3. Preparar a España para las transformaciones socioeconómicas que origina la IA.
4. Fortalecer la competitividad a través de las actividades de I+D.

Por su parte, la **Estrategia de Salud Digital del Sistema Nacional de Salud** elaborada por el Ministerio de Sanidad para el periodo 2021 a 2026 y se encuentra estructurada en tres líneas de actuación, que sirven de ejes para articular los contenidos e iniciativas asociados a la misma:

1. Desarrollo de servicios sanitarios digitales orientados a personas, organizaciones y a los procesos que integran el sistema de protección de la salud, con un enfoque de equidad.
2. Generalización de la interoperabilidad de la información sanitaria.
3. Impulso a la analítica de datos relacionados con la salud, sus determinantes y el sistema sanitario.

Entre los objetivos de esta estrategia se encuentra la constitución de un **Espacio Nacional de Datos de Salud (ENDS)**, un espacio de datos y los servicios asociados a nivel nacional, alineados con el área de acción del EHDS, que permitirá la integración de las diferentes soluciones existentes en el ámbito del SNS, y permitirá trabajar con modelos compartidos de datos para el análisis avanzado, la simulación, la predicción y la personalización, con el doble objetivo de:

- "Mejorar la toma de decisiones clínicas en el SNS, dotándolo de una información interoperable y de calidad y de un Espacio de Datos que permita su uso secundario para la generación de conocimiento científico y para la evaluación de los servicios"

- "La reutilización de la información clínica, enlazada con otras grandes fuentes de datos supone una gran oportunidad para la mejora de la calidad de la promoción de la salud, la prevención de la enfermedad y la discapacidad, la asistencia sanitaria, la investigación, la educación y la vigilancia epidemiológica".

La arquitectura física del ENDS estará distribuida en las 17 CCAA y en una organización central, que facilite el acceso a los servicios compartidos y que además puede asumir la infraestructura en caso de que alguna CCAA así lo requiera.

De este modo se ofrecerá al conjunto del SNS (Ministerio de Sanidad, consejerías de sanidad, centros sanitarios, agencias sanitarias, institutos de investigación, sociedades científicas, profesionales, y otras partes interesadas) una Plataforma tecnológica de almacenamiento, archivo masivo (Data Lake Sanitario / Big Data), tratamiento y análisis con capacidades digitales avanzadas, para los datos procedentes de los sistemas de información del SNS y de otras fuentes, tanto clínicos como de gestión, epidemiológicos o de operaciones estadísticas relacionadas con la salud.

Según establece la Estrategia de Salud Digital la plataforma de datos en la nube debe ser diseñada con criterios de interoperabilidad, capacidad de crecimiento, calidad, protección de datos, seguridad, trazabilidad y auditoría por parte de agentes propios del SNS y externos (AEPD, CCN…), así como segmentación, de modo que permita a las CCAA y organismos asociados contar, si lo desean, con sus propios almacenes de datos y al SNS en su conjunto, disponer de datos agregados y consolidados.

Sobre la base del repositorio de datos y las herramientas digitales avanzadas para el análisis, la simulación y la predicción, la plataforma proporcionará servicios específicos para los sistemas de información del SNS, de las CCAA, agencias estatales, centros sanitarios e investigadores y los propios pacientes.

Asimismo, la plataforma se prevé de soporte a la interoperabilidad con proyectos de la UE como European Health Data Space y otros más específicos como Genómica (1+ Million Genomes initiative), Imagen Medica de Cáncer, HSD/HCE (Historia de Salud Digital/Historia Clínica Electrónica), entre otros, teniendo en cuenta la necesidad de cumplimiento de estándares de

confiabilidad del repositorio, así como a la interoperabilidad con otros proyectos de repositorios de datos de ámbito sanitario, como la infraestructura de investigación de medicina de precisión del ISCIII y otras que puedan desarrollarse en diferentes dominios e instituciones relacionados con la salud.

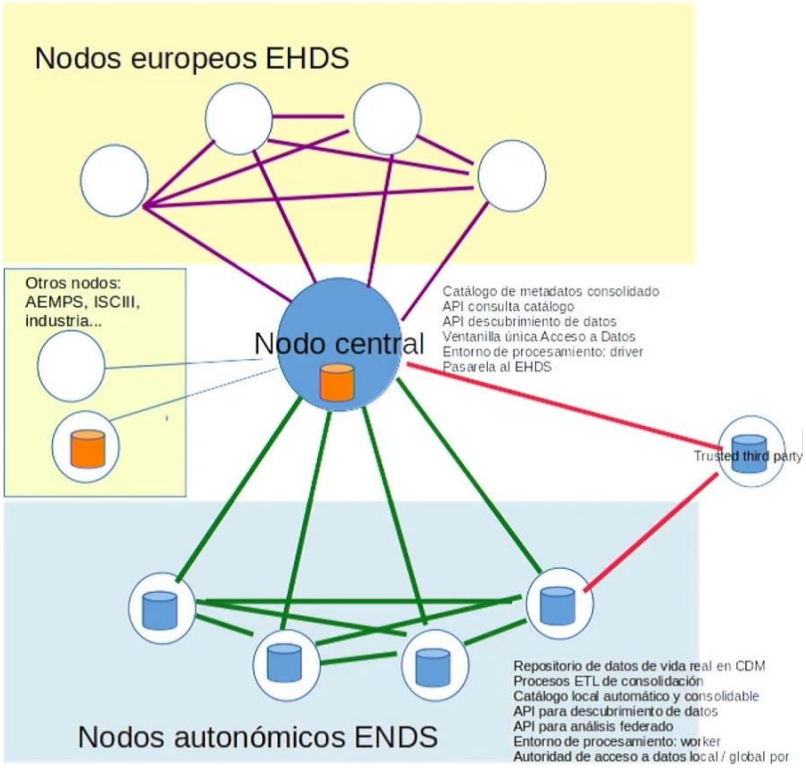


Ilustración 4.32. Nodos EHDS y ENDS. XII Foro de Interoperabilidad de la SEIS

Finalmente, también se recoge la posibilidad de definir escenarios para la cooperación público-privada que permitan acometer iniciativas de particular interés para el Sistema Nacional de Salud con la participación de organizaciones empresariales.

Entre las actuaciones particulares recogidas en la Estrategia de Salud del SNS relacionadas con la creación del Data Lake Sanitario que soportará el ENDS destacan:

1. Diseñar y poner en operación una plataforma cloud para el almacenamiento, procesamiento avanzado y análisis masivo de datos procedentes del SNS y organismos relacionados para su aplicación a la vigilancia en salud pública, a la práctica clínica, a la evaluación de servicios y a la investigación. La plataforma contará con espacios segmentados para las Comunidades autónomas y estará preparada para interoperar con el resto de nodos del European Health Data Space. Asimismo, la plataforma debe contemplar el uso de datos estructurados y no estructurados, incluyendo “real world data”

2. Constituir, por parte del Ministerio de Sanidad, una Oficina Técnica de Normalización y Calidad del Dato Sanitario, que será la única responsable, con la colaboración de las CCAA, de la definición, revisión y actualización de estándares y normas para asegurar la gestión de datos, la interoperabilidad en el conjunto del SNS y con terceros interesados, así como la correcta seudonimización o anonimización de los datos y sus casos de aplicación.

3. Establecer las condiciones habilitantes y los recursos facilitadores que permitan la generación y extracción de conocimiento aplicable a la prevención, diagnóstico y tratamiento, así como a la gestión del sistema sanitario y que definan un marco adecuado para la innovación tecnológica en colaboración con el sector privado, proporcionado periódicamente información a la ciudadanía sobre el funcionamiento del conjunto del sistema.

4. Proporcionar al conjunto del sistema herramientas de análisis y simulación (IA, ML, PNL, etc.) que permitan las actividades de vigilancia y control de riesgos para la salud, el seguimiento y control de los niveles de calidad de la asistencia sanitaria, así como la planificación, la toma de decisiones compartidas y la evaluación de las políticas públicas basadas en datos, incrementando la transparencia del sistema.

5. Implantar herramientas y servicios de explotación provistos por la plataforma contribuirán a la mejora de las habilidades de los profesionales sanitarios en las tecnologías digitales avanzadas, e impulsar la colaboración público-privada en el sector sanitario

4.4.3.2. PERTE Salud de Vanguardia

Entre los Proyectos Estratégicos de Recuperación y Transformación Económica, PERTE - Salud de Vanguardia, destacan:

- **Data Lake Sanitario** se corresponde con el Espacio Nacional de Datos de Salud - ENDS2, dotado presupuestariamente con 100 millones de euros, que se prevén sean contratados por el Ministerio de Asuntos Económicos y Transformación Digital de acuerdo a las especificaciones del Ministerio de Sanidad. Se trata de un proyecto de ámbito nacional que se abordará en colaboración con los gobiernos autonómicos, de un modo similar a como ya se ha realizado con anterioridad en la Receta Electrónica RE-SNS; Historia Clínica Digital HCD-SNS y EU Patient Summary, Código de Identificación Personal CIP-SNS, CMBD, BDCAP, Fondo de Cohesión, RegVacu o SERLAB-CoV.

- **Transformación digital de la asistencia sanitaria en atención primaria y comunitaria** con un presupuesto total de 230 millones de euros, vehiculizado a través de tres ámbitos de actuación y siete grupos de trabajo, planteando el último de ellos, el de **Soporte a la Decisión Clínica** en el área de Atención Personalizada, un marco de colaboración entre las CCAA para dotar al conjunto de sistemas sanitaros públicos de España de un modelo normalizado y un catálogo común de casos de uso de inteligencia artificial orientados al soporte a la toma de

decisiones en el ámbito de Atención Primaria y donde se han planteado diferentes casos de uso, para su desarrollo conjunto.

4.4.3.3. Infraestructura de Medicina de Precisión Asociada a la Ciencia y Tecnología (IMPACT)

IMPaCT es la Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología, un proyecto impulsado por el Ministerio de Ciencia e Innovación a través del Instituto de Salud Carlos III.

IMPACT aspira a ser una infraestructura científica de referencia que tiene como misión el establecimiento de los pilares para facilitar el despliegue efectivo de la Medicina de Precisión en el Sistema Nacional de Salud, asegurando la calidad científico-técnica, la equidad y la eficiencia en la utilización de los recursos científicos disponibles para dar respuesta a las necesidades de la ciudadanía.

El Plan estratégico IMPACT, se configura en torno a tres ejes, que a su vez disponen de acciones y paquetes de trabajo específicos, y de forma complementaria cuenta con dos líneas estratégicas transversales que aportan coherencia interna en aquellos aspectos tales como la ética de los datos y la internacionalización de la plataforma, que son comunes a los tres ejes estratégicos.

Ilustración 4.33. Ejes y líneas estratégicas de IMPaCT. Fuente ISCIII

El eje 1 de **medicina predictiva**, aborda el diseño y establecimiento de una cohorte de base poblacional representativa de la población residente en España, su variabilidad étnica, diversidad geográfica y ambiental, con la participación de todas las CC.AA. y seguimiento prospectivo.

El eje 2 de **ciencia de datos,** está orientado al desarrollo y validación de un entorno de integración y análisis conjunto de datos clínicos, moleculares y genéticos, para su uso secundario de forma coordinada con los ejes estratégicos 1 y 3. De igual manera, este eje generará modelos que permitan responder de forma eficiente a preguntas relevantes para el SNS promoviendo la generación de conocimiento de alto nivel basado en estas aproximaciones.

El eje 3 de **medicina genómica,** promueve el establecimiento de una infraestructura cooperativa distribuida en varios nodos para la realización de estudios genéticos de alta complejidad basado en tecnologías del ámbito de la investigación. Las capacidades científicas se orientan a contribuir al diagnóstico de enfermedades raras y otras enfermedades sin diagnóstico, de forma coordinada con las estructuras asistenciales del SNS y el Ministerio de Sanidad, atender las necesidades de secuenciación de la cohorte poblacional IMPaCT y contribuir a la iniciativa 1M+ Million Genomes**.**

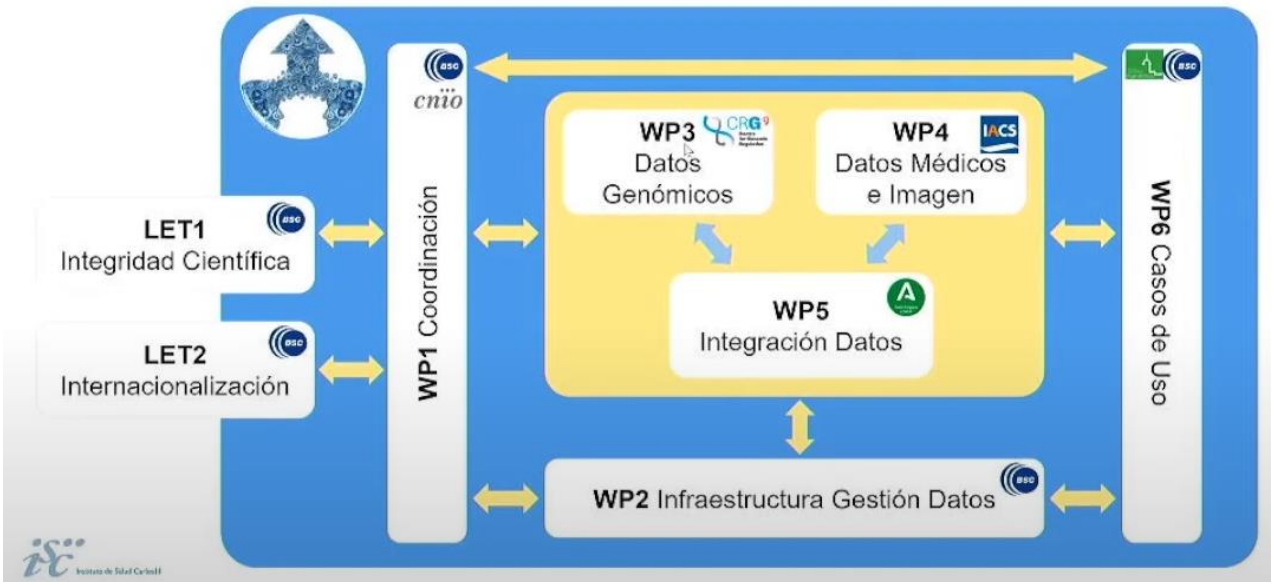


Ilustración 4.34. Impact-Data. Paquetes de Trabajo. Fuente ISCIII

La línea estratégica transversal 1, de **ética e integridad científica,** comprende las actuaciones para garantizar el cumplimiento de los principios de integridad científica y normas éticas, tanto en el tratamiento de datos y muestras biológicas como en la presentación de resultados

La línea estratégica transversal 2, de **internacionalización,** abarca las acciones dirigidas a posicionar a España como referente internacional en el ámbito de la Medicina de Precisión,

favoreciendo la participación y liderazgo de los grupos de investigación e instituciones de I+D+I en las actuaciones e infraestructuras europeas e internacionales en este ámbito.

4.4.3.4. Ministerios de Asuntos Económicos y Transformación Digital

En el marco de la Agenda España Digital 2025, la **Estrategia Nacional de Inteligencia Artificial 2021-2023** del Ministerio de Ministerio de Asuntos Económicos y Transformación Digital fija un plan de acción con seis ejes estratégicos, además de destacar la incidencia de la Inteligencia Artificial en el ámbito de la investigación multidisciplinar, y su alto potencial de aplicación en el ámbito sanitario; en el diseño de nuevos fármacos, reducción de errores de diagnóstico y mejora de la prevención y el tratamiento personalizado de las enfermedades más frecuentes.

Ilustración 4.35. Ejes estratégicos de la ENIA. Fuente Ministerio de Economía

Dentro de esta estrategia, entre otros actuaciones, la **Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA)** del Ministerio de Asuntos Económicos y Transformación Digital, procedió a convocar las Misiones de I+D en Inteligencia Artificial 2021, resultando ganador un consorcio público-privado, denominado Tartaglia, que tiene como objetivo el desarrollo de una **Red Federada de Inteligencia Artificial para Acelerar la Investigación Sanitaria** y en el que participan entre otros, entidades públicas pertenecientes a diferentes CCAA, planteando la implementación de una serie de casos de uso:

- Detección temprana de Alzheimer
- Adquisición de imagen diagnóstica de ultrasonidos guiada por IA
- Técnicas avanzadas de diagnóstico precoz del cáncer de próstata mediante IA
- Simulador para la predicción de enfermedades cardio-metabólicas con gemelo digital
- Cribado automático de retinopatía diabética mediante IA

La SEDIA también está encargada de representar al **hub nacional español de Gaia-X** a través del Consejo Gubernamental (GAIA-X Governmental Advisory Board), para contribuir a generar una infraestructura común europea de datos con un componente cloud que suponga una alternativa segura en el mercado y otorgue capacidad de control de acceso y reutilización para aquellos que producen los datos, para asegurar la disponibilidad de datos, la interoperabilidad, la portabilidad y que las empresas cumplan con los estándares y valores europeos.

4.4.3.5. Centro Nacional de Supercomputación

El Centro Nacional de Supercomputación con sede en Barcelona, también conocido como Barcelona Supercomputing Center, tiene por objeto investigar, implementar, gestionar y transferir tecnología y conocimiento en el área de High Performance Computing con el objetivo de facilitar el progreso en una variedad de campos científicos. Son múltiples los proyectos liderados por el BSC en el ámbito de la bio-informática, destacando, entre otros:

- Su papel como uno de los nodos del INB (Instituto Nacional de Bioinformática), plataforma tecnológica del ISCIII y coordinador de la Red de Bioinformática Traslacional (TransBioNet) un proyecto para la integración y desarrollo de los recursos bioinformáticos españoles en proyectos de las áreas de genómica, proteóhmica y medicina traslacional y que ha contribuido a la creación de una infraestructura computacional en el área de la bioinformática, participado en proyectos de genómica nacionales e internacionales, capacitado a usuarios y desarrolladores en bioinformática, actuando como una red de excelencia del Ministerio de Ciencia e Innovación, y trabajando en el mantenimiento y alineación del INB con:

    - la red Elixir, de la que actúa como nodo nacional
    - la *Global Alliance for Genomics and Health* (GA4GH), colaborando con esta iniciativa que tiene por objeto promover la compartición responsable de datos de genómica y de salud relativos a la genómica y que soporta, por poner un ejemplo, el estándar beacon , admitido por la B1MG, un estándar de interrogación normalizada a un repositorio genómico, a un conjunto de repositorios mediante las redes beacon, o incluso la posibilidad de que las preguntas se mapeen a los vocabularios de cada repositorio con query expansión o devuelvan sus resultados como OMOP y FHIR.

- la labor de liderazgo que realiza en el eje de ciencia del dato de IMPACT.

4.4.4. Ámbito Regional

4.4.4.1. Andalucía

La **Base Poblacional de Salud (BPS)** del Sistema de Salud Público de Andalucía (SSPA) es un sistema de información sanitaria que recoge datos clínicos y del uso de recursos sanitarios de cada una de las personas que reciben asistencia sanitaria en el Servicio Andaluz de Salud y en mayo de 2020, ha sido incluida en el **Repositorio de Prácticas Innovadoras en Envejecimiento Activo y Saludable (EIP on AHA)** de la Comisión Europea.

Con datos de más de 13.000.000 de pacientes, permite realizar estimaciones sobre la salud, el comportamiento de los usuarios en relación a los servicios sanitarios, estratificar la población para orientar la prestación de estos servicios, realizar estudios longitudinales, estimar la incidencia de las patologías y realizar proyecciones sobre el estado de la salud de la población, sus necesidades de recursos y analizar la eficiencia distributiva y en el uso de los recursos por los proveedores sanitarios.

La solución corporativa (plataforma software y hardware) de analítica avanzada, basada en tecnologías Big Data, para el SSPA; tiene por objeto el suministro de la solución corporativa, la administración, operación, soporte capacitación y el desarrollo de los siguientes casos de uso:

1. Definición de factores que inciden en la morbilidad y predicción de futuros riesgos de salud
2. Diseño de trayectorias óptimas y personalización en la prestación de los servicios sanitarios
3. Optimizar la distribución de cupos en Atención Primaria
4. Segmentación de crónicos, sobre la población andaluza, en base a niveles de cuidados
5. Comparativa de resultados de tratamientos farmacológicos
6. Modelos predictivos sobre grupos poblacionales respecto al consumo de recursos
7. Motor de recomendación para la optimización de la lista de espera quirúrgica
8. Identificación y prevención de interacciones entre fármacos en pacientes polimedicados
9. Análisis de imagen radiológica para asistir en el cribado de cáncer de mama
10. Identificación de pacientes objetivo de nuevos tratamientos farmacológicos
11. Procesamiento de textos clínicos con NLP para desarrollar codificador CIE10 y SNOMED

12. Detección de situaciones problemáticas, relativas a salud pública, analizando RRSS

13. Optimización de planes de choque hospitalarios

14. Predicción de demanda de servicios en centros que trabajan bajo concierto hospitalario

15. Búsqueda de factores que puedan predecir sepsis en pacientes

El **Grupo de Innovación Tecnológica (GiT)** de los Hospitales Universitarios Virgen Macarena y Virgen del Rocío, también desarrolla y mantiene la plataforma **ITC-Bio,** investigando en modelado e interoperabilidad de la información clínica basada en estándares, en el desarrollo de software, infraestructuras innovadoras para el soporte a la investigación, etc. apoyándose en software libre y técnicas de inteligencia artificial, e impulsando los principios FAIR.

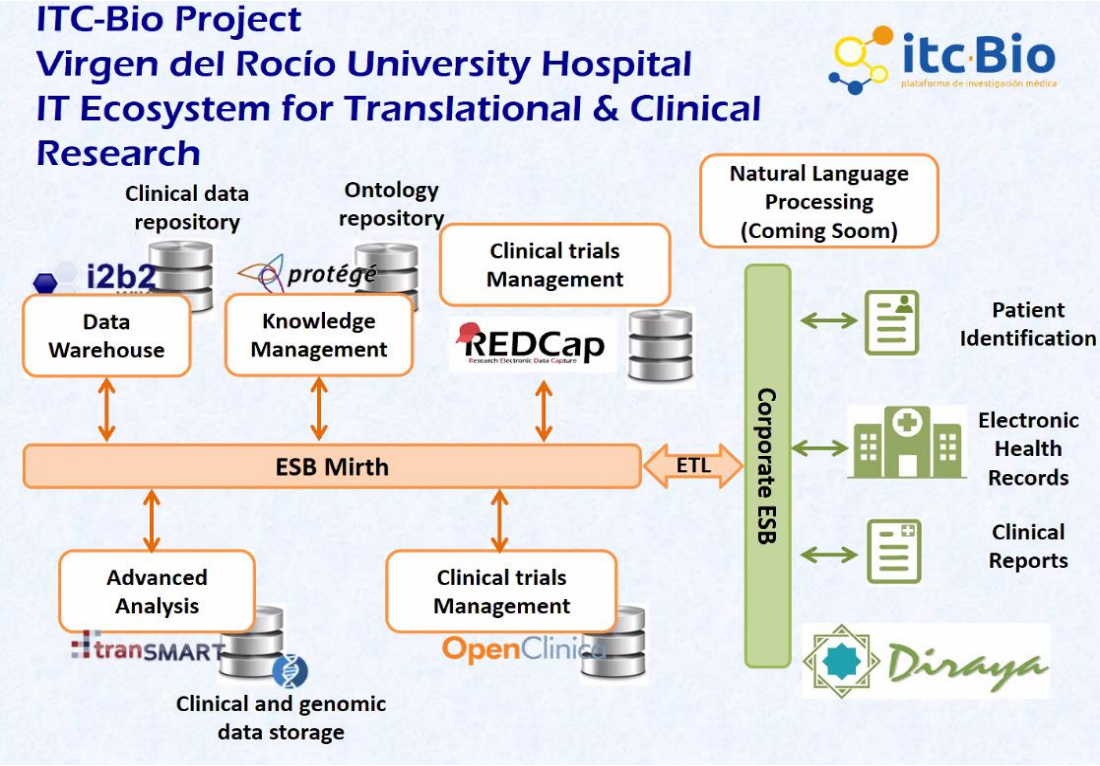


Ilustración 4.36. Plataforma ITC-Bio. Fuente Master DSTICSDS

Especialmente relevante en materia de Gobernanza es la Resolución conjunta 1/2021 de la secretaría general de investigación, desarrollo e innovación en salud de la Consejería de Salud y Familia y de la Dirección Gerencia del Servicio Andaluz de Salud por la que se aprueban instrucciones para la **ordenación del acceso** y uso de la información sanitaria contenida en los

sistemas de información del sistema sanitario público de Andalucía con fines de investigación e innovación por las entidades dependientes de la consejería de salud y familias, en el ámbito de los siguientes sistemas:

1. Los distintos módulos de la Historia de Salud Digital de Andalucía, en adelante Diraya

2. La Base Poblacional de Salud de Andalucía

3. Otros registros o bases de datos del SSPA, distintas a Diraya o BPS, con información sanitaria de los pacientes, una vez oficializadas

4.4.4.2. Aragón

Una de las iniciativas que cuenta con una trayectoria más dilatada es **BIGAN**, una plataforma que es parte del Sistema de Información de Salud de Aragón y está integrada por tres elementos, una infraestructura de integración de datos, una infraestructura tecnológica de procesamiento de dato masivo y una metodología de análisis avanzado para la evaluación y monitorización de los servicios sanitarios y de la salud de los ciudadanos.

Con la doble finalidad de integrar y analizar los datos relativos al Sistema de Salud de Aragón, para generar conocimiento aplicado a la mejora del funcionamiento del sistema sanitario, de la calidad de la atención sanitaria y, con ello, de la salud de la población de Aragón, trabajando en tres ámbitos de actuación:

- Gestión clínica
  - Contratos de gestión
  - Recomendaciones clínicas
  - Estrategias de salud

- Investigación clínica y sanitaria

- Formación de profesionales

Además del proyecto BIGAN, el **Grupo de Ciencia de Datos para la investigación en servicios y políticas sanitarias del IACS**, coordina diferentes actuaciones de ámbito internacional, como uno de los paquetes de trabajo del proyecto **TEHDAS**, el proyecto Health Research & Innovation Cloud – HealthyCloud, y también nacionales, como el **Atlas de Variabilidad de Práctica Médica (VPM)**, que tiene por objeto generar conocimiento que pueda accionar políticas sanitarias dedicadas a mejorar el sistema sanitario, en diferentes ámbitos temáticos como la calidad de los cuidados hospitalarios, las hospitalizaciones potencialmente evitables y la variabilidad en hospitalizaciones de diferentes entidades mediante la comparación de la información de sus Conjunto Mínimo Básico de Datos (CMBD).

4.4.4.3. Asturias

En 2022 el Servicio de Salud del Principado de Asturias ha sido admitido como Data Partner, para integrarse en el consorcio **EHDEN**.

4.4.4.4. Baleares

El 8 de julio de 2021, el Hospital Universitario Son Espases procedió a anunciar la implantación de un proyecto para el análisis y extracción de datos de salud no estructurados de las Historias Clínicas Electrónicas (HCE) y transformarlos en datos estructurados, con el objetivo de acelerar la investigación clínica.

4.4.4.5. Canarias

El Proyecto de **Medicina Personalizada y Big Data** - **MedP-Bigdata**, es una actuación de Investigación y Desarrollo (I+D), en el que participan conjuntamente la Conselleria de Sanidad Universal y Salud Pública de la Generalitat Valenciana y el Servicio Canario de la Salud, adscrito a la Consejería de Sanidad del Gobierno de Canarias.

El Proyecto MedP Bigdata plantea el desarrollo de los aspectos arquitectónicos, tanto software como hardware, para la implantación de una plataforma tecnológica que, de soporte a múltiples herramientas y acceso a los datos disponibles de los pacientes, con vistas a crear estudios que permitan tener una imagen de la población y su salud, además de una serie de casos de uso, agrupados en áreas temáticas:

- Interfaz paciente-sistema sanitario, para registro de estilos de vida y promoción de la salud:
    - Cuchara inteligente, para registro de los hábitos nutricionales y propuestas de alternativas saludables y atractivas sin la intervención profesional.
    - Contamos contigo, para registro actividad física, movilidad y ejercicio y propuestas de alternativas saludables y atractivas.
    - Sonrisa saludable, para registro de estado de ánimo, el afrontamiento positivo de los retos y dificultades de la vida.
    - Tabaco, alcohol y otras adicciones, para registro de adiciones incluyendo tanto el consumo de sustancias adictivas como la práctica de apuestas, dependencia, etc.

  - Mejor en compañía, registro de situaciones de soledad no deseada, aislamiento social y experiencias de alienación y prácticas y hábitos relacionales.

- Sistema analítico-predictivos, orientados al soporte a la decisión clínica e investigación:
  - Aplicación de procesamiento de lenguaje natural en el dominio de informes clínicos aplicando etiquetado semántico SNOMED-CT.
  - Descripción de la fisiopatología del dolor lumbar mediante la aplicación de técnicas analítico predictivas basadas en imagen médica con resonancia magnética.
  - Predicción del número de ingresos en urgencias en relación con la concentración de partículas en el aire.

- Además de plantear las siguientes actuaciones para una segunda fase:
  - Pre consulta inteligente, para optimizará la tarea de recabar datos de los pacientes.
  - Monitorización domiciliaria de las situaciones crónicas y de las altas hospitalarias.
  - Optimización terapéutica y detección de oportunidades de des-prescripción.
  - Segmentación de pacientes en las patologías de mayor relevancia.
  - Modelo de medida y predicción de la eficiencia de las unidades funcionales de atención primaria; impacto del clustering de enfermedades crónicas no comunicables.
  - Selección de pacientes para ensayos clínicos.
  - Selección de pacientes para búsqueda activa de enfermedades raras.
  - Dictáfono inteligente.
  - Pharmacia-covid análisis en tiempo real de la efectividad del tratamiento farmacológico frente a covid.
  - Predicción de reingresos no programados en el mes siguiente al alta.

4.4.4.6. Cantabria

El Servicio Cántabro de Salud y el Instituto de Investigación Valdecilla (IDIVAL) fueron el primer servicio regional de salud en ser elegidos como Data Partners, para formar parte de **EHDEN**.

Con fecha de 4 abril de 2022, el Servicio Cántabro de Salud ha publicado la Resolución por la que se dispone el Convenio para el tratamiento y **estructuración de datos clínicos** con el objetivo de desarrollar diversos proyectos de investigación médica, mediante la transformación de los datos de las HCE que se encuentran en texto libre (notas clínicas, curso clínico, evolutivos, informes médicos de pruebas diagnósticas, informes de alta, informes de intervenciones quirúrgicas, etc.) en una base de datos codificada utilizando vocabularios clínicos estándar (SNOMED, CIE-9, CIE-10, LOINC), para lo que se utilizarán procesos de inteligencia artificial y procesamiento del lenguaje natural.

4.4.4.7. Castilla la Mancha

La plataforma de ayuda al diagnóstico en Atención Primaria del Servicio de Salud de Castilla-La Mancha, conocida como **Sapiens**, está a disposición de todos los profesionales de Atención Primaria en los más de 200 centros de salud y 1.100 consultorios locales de la región y es capaz de avisar en tiempo real, en caso de que el paciente que está siendo atendido presente síntomas o alguna dolencia asociada a alguna de las 18 vías clínicas programadas, de manera que muestra, a médicos y enfermeras las recomendaciones a aplicar en cada proceso asistencial.

Además, Sapiens profundiza en el desarrollo de las estrategias de seguridad del paciente, ya que reduce posibles discrepancias entre la decisión del profesional y lo que marca la vía clínica correspondiente.

También permite calcular de forma automática el grado de adherencia de los profesionales de Atención Primaria a las vías clínicas creadas por el propio Servicio de Salud de Castilla-La Mancha y referidas a procesos asistenciales habituales en este ámbito asistencial.

Incluye un portal de explotación de toda la información procesada por Sapiens, con más de 160 millones de informes analizados disponibles de 2,9 millones de pacientes, datos que permiten su análisis desde todos los centros del SESCAM para obtener conclusiones sobre la mejora de la práctica clínica.

Aunque centrada en uso asistencial, el Servicio de Salud de Castilla-La Mancha (SESCAM), de forma conjunta con el Servicio Canario de Salud trabajan conjuntamente en el desarrollo de una Historia Clínica Interoperable y Multi-regional (**ISOHCE**), un proyecto que tiene como objetivo desarrollar un nuevo modelo de historia clínica electrónica interoperable, para facilitar la continuidad asistencial y mejorar la seguridad del paciente con los siguientes trabajos:

- Base de datos HCE estandarizada (ISO-13606).
- Extractor: transforma datos históricos en formato propietario a modelo estándar.

- APIs de lectura/escritura para interactuar con la B.D. estandarizada.
- Estación clínica modular y enriquecida.
- Integración nativa con servidores terminológicos.
- Integración con un MPI.

4.4.4.8. Castilla León

La Junta de Castilla y León contrató el sistema integral innovador para el desarrollo de una **plataforma de atención socio-sanitaria al paciente crónico y personas en situación de dependencia**, donde se solicita, entre otros, el desarrollo de una solución innovadora para la toma de decisiones basado en Inteligencia de Negocio (B.I.) y análisis de grandes volúmenes de datos (Big Data), con capacidades para construir, evaluar y explotar modelos predictivos y de aprendizaje automático a partir de información conjunta de salud y servicios sociales poniendo especial atención y seguridad en la privacidad y seguridad de los datos y en la transparencia y equidad de los modelos, ofreciendo a los usuarios el conocimiento de las técnicas de aprendizaje automático, el entrenamiento y validación de los modelos predictivos y a los usuarios avanzados el desarrollo y validación de modelos mediante APIs/frameworks del mercado (Spark, MLib, Tensorflow, Keras, etc.)

4.4.4.9. Cataluña

El Plan Director de Sistemas de Información 2018-2022 del Sistema Integrado Sanitario de Utilización Pública de Cataluña (SISCAT), establece la construcción de un nuevo sistema de información que será utilizado por todos los proveedores sanitarios del Servei Català de la Salut, para enfrentarse a una serie de retos sanitarios y sociales, como la atención crónica, la dependencia, la prevención de factores de riesgo, la aparición de nuevas enfermedades y la necesaria promoción de la salud, y el impulso de procesos analíticos basados en datos y que plantea el desarrollo de la **Historia Clínica Electrónica de Cataluña**, basada en el estándar openEHR, como un repositorio longitudinal, transaccional y compartido.

También en el ámbito de la Salud, el Departamento de Salud, el CatSalut y la Fundación Tic Salut Social presentaron el programa **Salud/IA**, que tiene como finalidad la promoción y desarrollo de la IA centrada en las personas y cuyos objetivos principales son:

- Promover la investigación, el desarrollo y la innovación (I+D+i) de la IA en salud para mejorar el bienestar de las personas, de acuerdo a las directrices éticas de la Unión Europea.

- Favorecer la participación e implicación de todo el Sistema de Salud de Cataluña para obtener resultados de mayor impacto.

- Asegurar que las soluciones desarrolladas estarán disponibles para todos, para homogeneizar la calidad asistencial y evitar desigualdades entre centros.

4.4.4.10. Comunitat Valenciana

De forma complementaria a la iniciativa **MedP-Bigdata**, abordada conjuntamente con Canarias y anteriormente desarrollada, es especialmente relevante el trabajo que viene siendo desempeñado por el grupo de Imagen y Tecnologías Aplicadas a la Salud del Instituto de Investigación Sanitaria del Hospital Universitario La Fe de Valencia, con trabajos de gran relevancia, en el desarrollo de la IA en imagen médica, como los trabajos dentro de la iniciativa europea **AI for Health Imaging (AI4HI),** PRIMAGE, EuCanImage, ProCancer o Chaimeleon, que tiene por objetivo crear un repositorio europeo centralizado con 20 millones de imágenes médicas, junto a otra información de la HCE, para avanzar en el diagnóstico del cáncer colon, mama, pulmón y próstata.

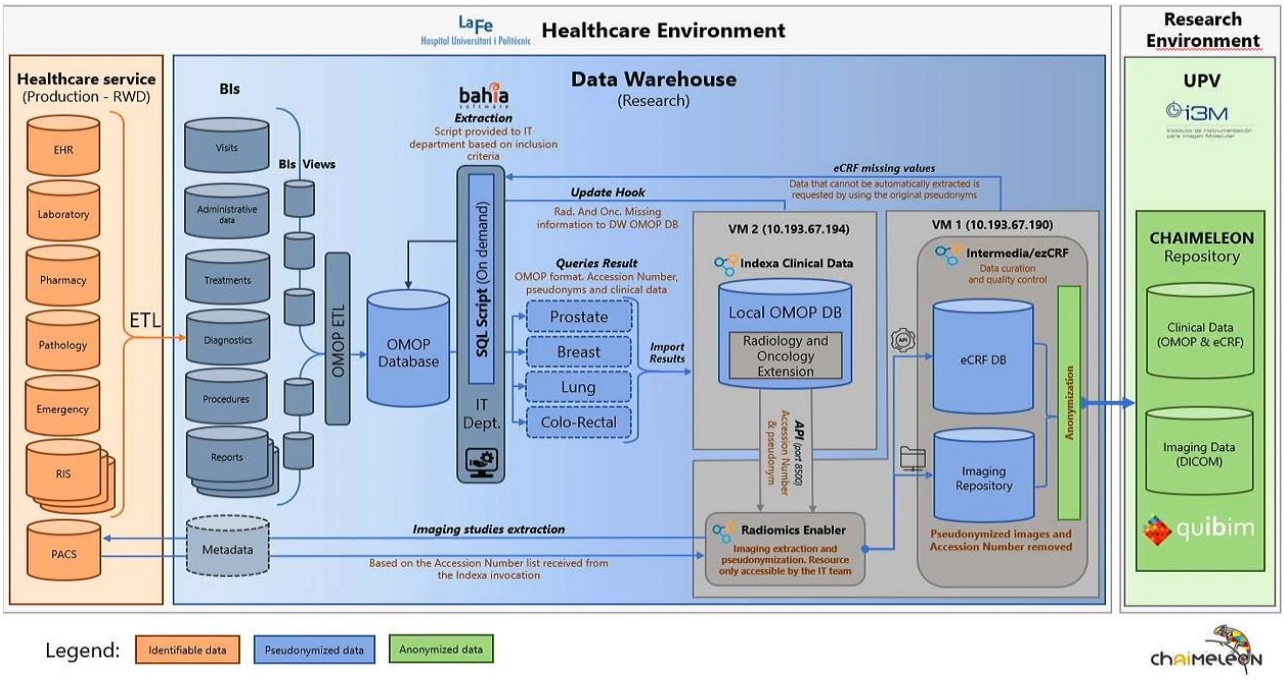


Ilustración 4.37. Arquitectura Chaimeleon - HULAFE. XII Foro de Interoperabilidad SEIS

4.4.4.11. Extremadura

El Servicio Extremeño de Salud impulsa el diseño y desarrollo de un sistema de prescripción personalizada, **MedeA**, para su validación en condiciones clínicas reales y que tiene por objeto ofrecer a los profesionales sanitarios un sistema de apoyo a la decisión clínica que integrará análisis farmacogenéticos / óhmicos, clínicos y farmacológicos con una metodología innovadora aplicable a cuatro ámbitos o subproyectos:

- Desarrollo de sistema de prescripción personalizada, sistema de decisión clínica.
- Desarrollo de sistemas para análisis fármaco-genéticos.
- Desarrollo de sistemas de evaluación de pacientes y/o voluntarios para ensayos clínicos.
- Desarrollo de sistemas para la evaluación de reacciones adversas a medicamentos.

4.4.4.12. Galicia

La Consellería de Sanidades de la Xunta de Galicia procedió a la evolución de su plataforma de analítica avanzada (SIACS), complementándola con la plataforma de Big Data **HEXIN** (Plataforma de Explotación de Información e Xestión de datos clínicos e Epidemiolóxicos), mediante una Compra Pública de Innovación, enmarcada dentro del programa Innova Saude.

HEXIN es una solución con multitud de capacidades, como el procesamiento semántico, indexación y búsqueda de información no estructurada, flexibilidad en la relación y navegación en información, exploración por facetas y jerarquías, facilidad de uso, rápida adopción, agilidad y flexibilidad.

HEXIN permite, entre otros, la explotación de información clínica disponible en los sistemas existentes con el propósito de:

- Facilitar la toma de decisiones clínicas.
- Proporcionar información para la identificación y clasificación de casos de epidemiología.
- Proporcionar información de gestión.

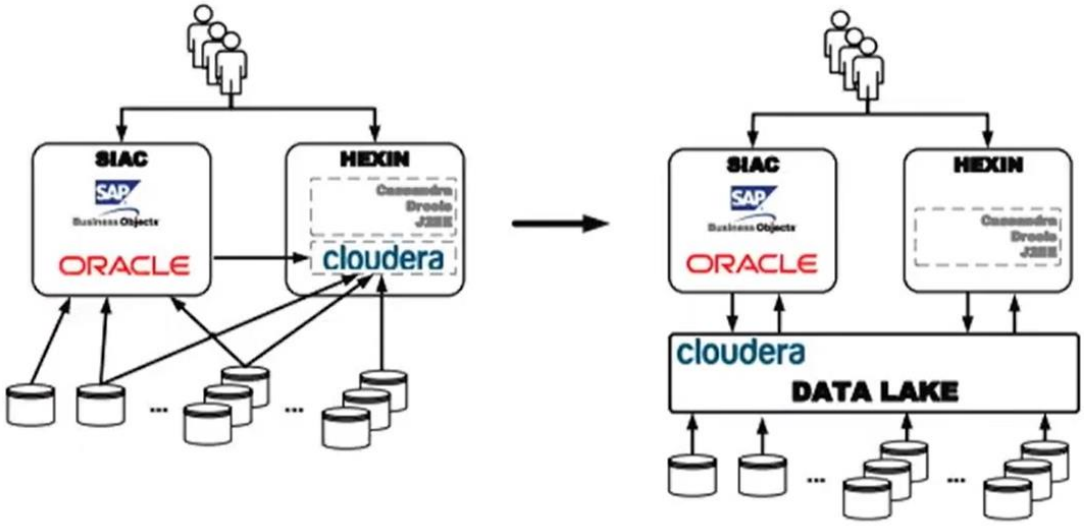


Ilustración 4.38. SIAC - HEXIN. XII Foro de Interoperabilidad SEIS

4.4.4.13. Madrid

La Dirección General de Sistemas de Información y Equipamientos Sanitarios de la Comunidad de Madrid lidera el desarrollo de un Data Lake Sanitario, que cuenta con más de 10.000 millones de registros sanitarios, para ser utilizados en dos actuaciones diferenciadas:

- **Delfos**, orientada a impulsar la planificación y gestión de los recursos del sistema sanitario.
- **Hipócrates**, busca impulsar la investigación en ciencia de datos.

El impulso de la medicina personalizada y de precisión a partir de los datos óhmicos es abordado por el Centro Madrileño de Análisis Genómico - **GMAC**, donde se consolidan todos los estudios de secuenciación de los hospitales públicos de la Comunidad de Madrid.

Por su parte, el Instituto de Investigación Hospital 12 de Octubre (Instituto i+12), lidera el proyecto **Infobanco,** para el desarrollo de un sistema innovador e inteligente de arquitectura de red regional de datos de salud para la Comunidad de Madrid y que tiene como finalidad impulsar el aprendizaje del sistema sanitario.

Infobanco se concibe como un repositorio normalizado de datos de salud, que consolidará la información generada por diferentes fuentes procedentes de los centros de la Consejería de Sanidad de Madrid y/o el Servicio Madrileño de Salud y de diferentes niveles asistenciales (atención primaria, hospitales, emergencias, farmacia), con una perspectiva tanto individual, en el ámbito de la medicina personalizada y de precisión, como poblacional y de salud pública, para la

mejora asistencial, la innovación, la atención sanitaria basada en el valor (VBHC), la investigación biomédica y otros usos secundarios.

Su arquitectura pretende tener un funcionamiento de plataforma que provea servicios a clínicos, gestores e investigadores, haciendo posible la combinación de datos de múltiples fuentes, y estará dotada de herramientas de gobernanza, ingesta, transformación, interrogación, visualización y análisis de datos para la obtención de conocimiento y el soporte a la toma de decisiones.

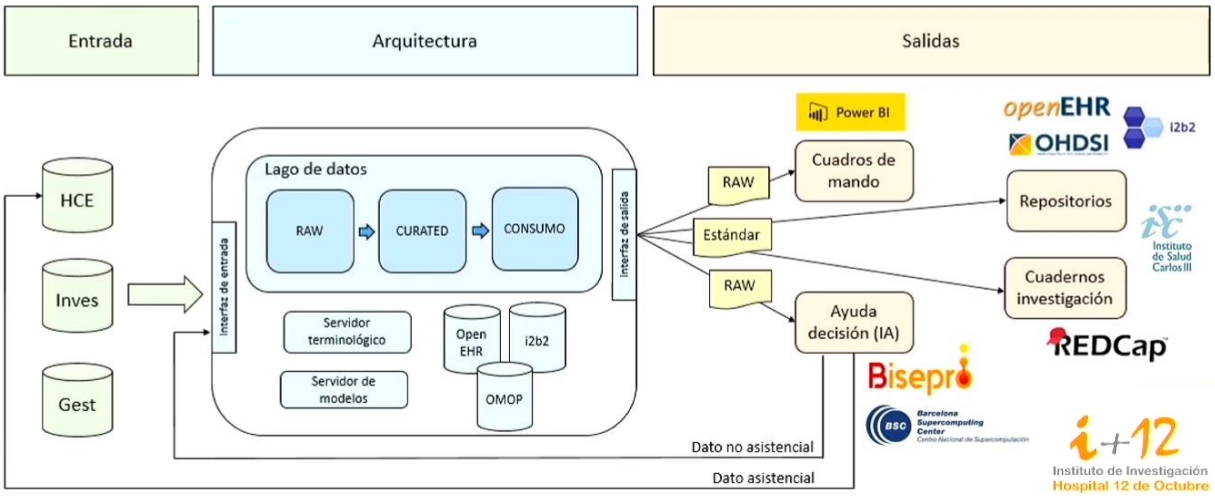


Ilustración 4.39. Arquitectura de Infobanco. Fuente i+12

También prevé su integración futura con el proyecto **Medigenomics** como plataforma y sistema experto de estudios genómicos e **Integra-CAM** como ecosistema tecnológico que permitirá la monitorización domiciliaria y el seguimiento de la capacidad intrínseca de las personas mayores.

4.4.4.14. Murcia

El Servicio Murciano de Salud ha procedido a contratar el Diseño, Implantación, Configuración y Desarrollo de una Plataforma Data Lake Sanitario para el Servicio Murciano de Salud - Proyecto **AZUD**, con la provisión de los siguientes servicios:

- Software y hardware necesario para desplegar el Data Lake y las soluciones de procesamiento Big Data y de análisis y predicción de información.
- Implantación de la infraestructura software.
- Configuración, parametrización y desarrollo de las soluciones de integración de datos.

- Configuración, parametrización y desarrollo de las soluciones de explotación de datos.
- Formación y soporte post-implantación
- Consultoría de diseño técnico y de diseño funcional de dos casos de uso de la plataforma.

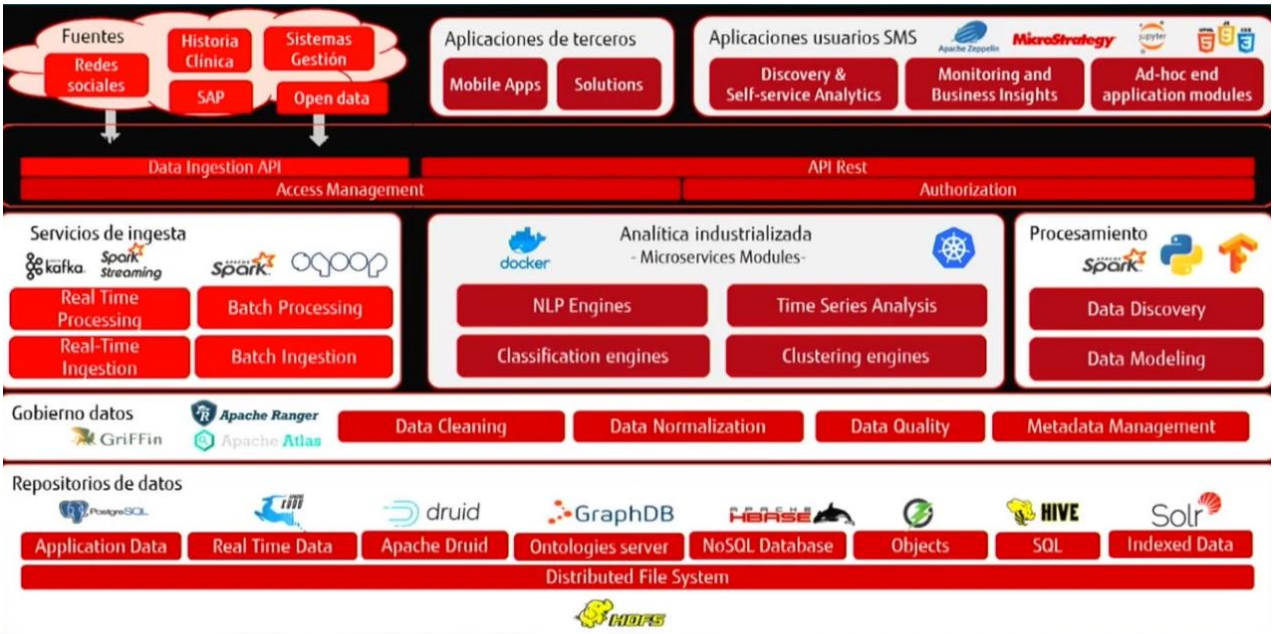


Ilustración 4.40. Arquitectura AZUD. XII Foro de Interoperabilidad SEIS

4.4.4.15. Navarra

La Base de Análisis de Resultados de Navarra, **Bardena**, del Servicio Navarro de Salud - Osasunbidea, es operada por un equipo multidisciplinar integrado por profesionales de la Salud de la Dirección General de Sistemas de Información y del Servicio de Tecnologías de Salud de la Dirección General de Transformación Digital, para recopilar información de Atención Primaria, hospitalizaciones, urgencias, quirófanos, salud mental, farmacia, diagnósticos y la TIS (tarjeta individual sanitaria) y tiene previsto incorporar gradualmente otros datos sobre laboratorios, anatomía patológica, enfermería hospitalaria, radioterapia, fisioterapia, banco de sangre, etc.

Su principal objetivo es la evaluación de actuaciones y la detección de problemas, realizando análisis predictivos de tendencias, desarrollo de indicadores que faciliten la gestión, tanto en un servicio, como en una zona básica de salud, o en toda Osasunbidea.

Entre los casos de uso implantados destacan:

- La elaboración de informes sobre la mortalidad en Navarra,
- Evaluación de la estrategia de crónicos,
- Evaluación del impacto de los golpes de calor en Urgencias
- Estadísticas sobre el consumo de drogas.
- Consumo de fármacos y personas candidatas para la revisión de los tratamientos,
- Estrategias de “no hacer” ante un problema de salud.
- Gestión en el ámbito del Trabajo Social
- Estudios de investigación en cáncer de mama e ictus.
- Evaluación de los circuitos de urgencias.

4.4.4.16. País Vasco

El Servicio Vasco de Salud impulsa la **Definición de un Modelo de Gobierno del Dato** para avanzar en el lanzamiento de iniciativas en el marco del Big Data y Analítica Avanzada, que permita incorporar sistemas inteligentes de analítica descriptiva y predictiva, capaces de optimizar la gestión de sus activos de información y mejorar la calidad de los servicios sanitarios que ofrece.

Con este proyecto Osakidetza pretende de organizar y coordinar todos los procesos de analítica de datos, con el fin de dar salida al gran potencial de sus sistemas de información, optando por definir un modelo de Gobernanza del Dato que garantice la disponibilidad, integridad, usabilidad y seguridad y que trabajará sobre la plataforma de **Big Data as a Service - BDAAS** de la Sociedad Informática del Gobierno Vasco - EJIE.

El proyecto será validado mediante el desarrollo de dos casos de uso:

- Comparador de modelos de intervención, análisis concluyente que permita confirmar las mejoras que se presuponen a la implantación de la metodología de atención al paciente en el proceso quirúrgico, denominado Fast-Track.
- Optimización del modelo identificador del parto activo, desarrollo de un modelo analítico basado en el estudio de historias clínicas y revisiones ginecológicas de mujeres embarazadas y en la capacidad predictiva de esta información, en referencia al momento del parto.

Adicionalmente, el Departamento de Salud del Gobierno Vasco y Osakidetza, a través de la Fundación Vasca de Innovación e Investigación Sanitarias, **Bioef**, participa en el proyecto europeo **MIDAS**, financiado por la Comisión Europea y cuyo objetivo es crear aplicaciones de big data compartidas por los sistemas sanitarios del País Vasco, Inglaterra, Irlanda, Irlanda del Norte, Finlandia y el Estado de Arizona (EEUU), un proyecto que permitirá a las autoridades sanitarias contar con nuevas herramientas para analizar el efecto de las políticas de salud pública y para elaborar previsiones sobre cómo abordar nuevos retos epidemiológicos, como por ejemplo la obesidad infantil.

4.4.4.17. La Rioja

Fundación Rioja Salud, como entidad del sector sanitario público de La Rioja, procedió a constituir una Unidad de Ciencia del Dato, Big Data e Inteligencia Artificial, **UCIDA**, integrada por un equipo multidisciplinar de profesionales con experiencia en matemáticas, informática, investigación y asistencia sanitaria.

Para acometer su trabajo se apoyan en una plataforma científico-técnica, regulada por la Orden SAL/41/2022, de 15 de julio, por la que se crea la plataforma analítica de salud de la Comunidad Autónoma de La Rioja, **PASCAL**[6], y se define su gobernanza en el sistema público de salud de La Rioja.

PASCAL ha de entenderse como una plataforma en continua evolución, un desarrollo específico para la explotación de los datos existentes en los sistemas de información que soportan los procesos de salud, y entre ellos, de forma particular, la historia clínica, recogida en la Ley 2/2002, de 17 de abril, de Salud, normativa donde se precisa que deber recoger aquellas actuaciones clínicas y sanitarias de los diferentes episodios asistenciales así como los datos administrativos de identificación, clínico asistenciales y sociales y estará a disposición de los enfermos y de los facultativos que directamente estén implicados en el diagnóstico y tratamiento del enfermo, así como para efectos de inspección médica o fines científicos.

PASCAL es una solución diseñada para disfrutar de disponibilidad, continuidad, adaptabilidad, crecimiento a demanda, mejora continua, seguridad, confidencialidad, trazabilidad y auditoria sobre los datos, además de segmentación, normalización e interoperabilidad para la colaboración en redes de investigación y estará integrada por infraestructuras, herramientas y procesos, con los que crear y mantener un archivo electrónico evolutivo, seguro y confidencial que ayude al cumplimiento de los objetivos del Sistema Público de Salud de La Rioja.

6 Acrónimo elegido en honor al matemático francés, Blaise PASCAL, al lenguaje de programación PASCAL y a la unidad de medida de la magnitud física de la presión (1 PASCAL = 1 Newton/m2), un excelente ejemplo de los tres ámbitos de conocimiento (ciencias matemáticas, ciencias de la computación y el conocimiento propio de un medio, en nuestro caso el de la Salud) que confluyen en la Ciencia del Dato.

Además, se procede a la creación de una Comisión de Control y Seguimiento, integrada por miembros de diferentes instituciones y cuya misión es supervisar la actividad de PASCAL.

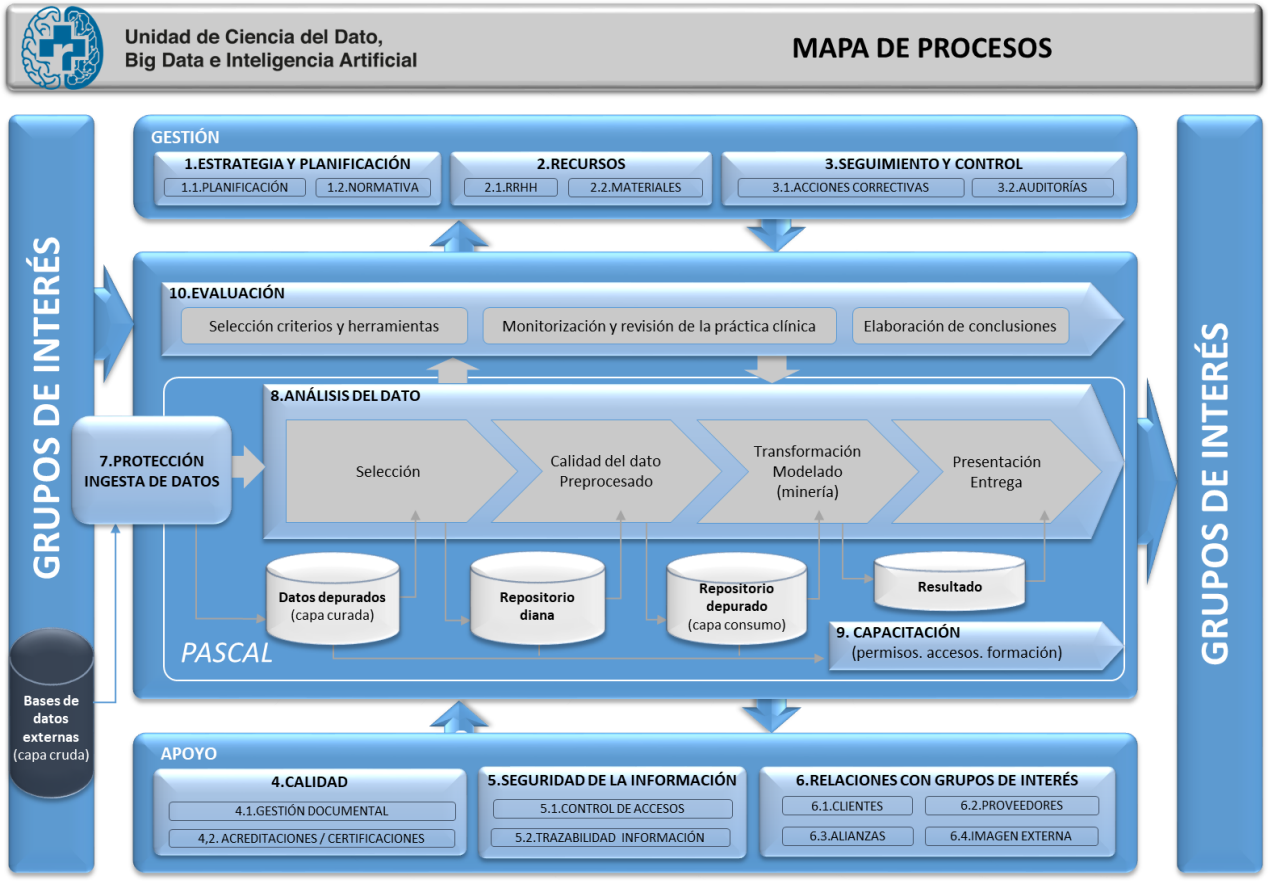


Ilustración 4.41. Mapa de procesos. UCIDA y PASCAL

El sistema público de salud de La Rioja, a través de Fundación Rioja Salud, también ha sido aceptado para forma parte del consorcio EHDEN como Data Partner y está integrado en el consorcio Tartaglia, para el desarrollo de la Red Federada de Inteligencia Artificial para Acelerar la Investigación Sanitaria.

4.5. Data Lake Sanitario

Este TFM **define** un Data Lake Sanitario como una plataforma segura diseñada para la ingesta periódica e incremental de datos des-identificados (anonimizado, seudonimizado, k-anonimizado) provenientes de múltiples fuentes, con los que construir un registro agregado con información de calidad, individual, longitudinal, multi-dimensional, multi-formato y normalizada.

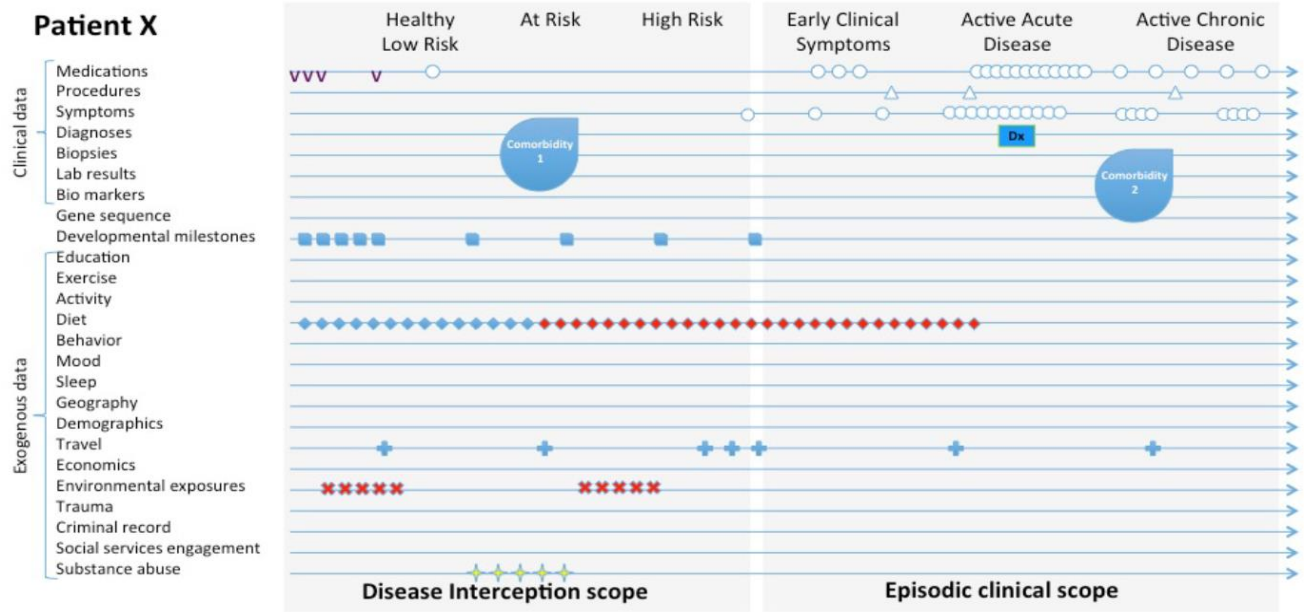


Ilustración 4.42. Registro longitudinal de datos de salud. Fuente IBM

Un principio clave de **diseño**, es que "el dato no viaje al profesional, sea él quien vaya al dato", por lo que un Data Lake debe tener la capacidad de segmentar los datos consolidados en múltiples sub-conjuntos o data-sets, a los que accederán los profesionales (clínicos, gestores e investigadores), de forma controlada, trazable y previa autorización, para acometer sus estudios haciendo uso de las herramientas, formación y servicios existentes para cada una de las cuatro tipologías de proyectos previstas (IA-DSS, RWE, CRF y BI).

El **objetivo** de un Data Lake Sanitario es que los resultados generados por los múltiples proyectos, o usos secundarios de los datos, impulsen:

- El desarrollo de la Medicina 5Ps
- La mejora de la atención, de la eficiencia y de los resultados en salud (VBHC).
- La generación de nuevo conocimiento (**I+D**+i) a partir de los datos y la colaboración en red.
- Una nueva economía (I+D+**i**) del dato y la IA, soportada por la colaboración público-privada.

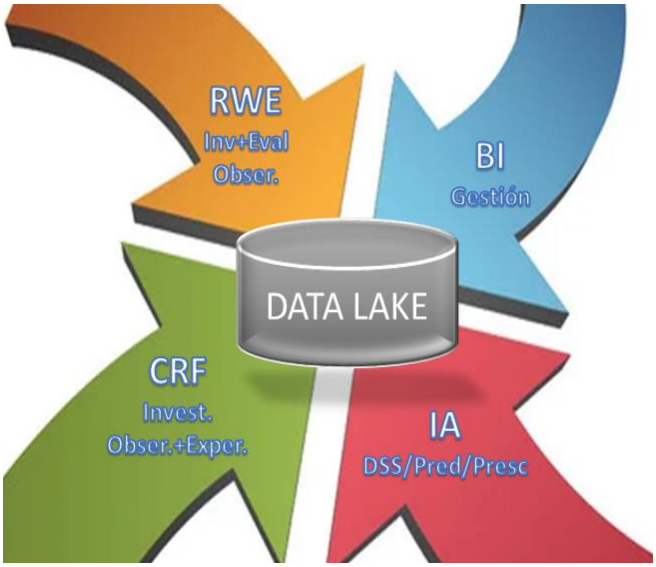


Ilustración 4.43. Consolidación en Data Lake Sanitario, datos entran no salen. Fuente propia

Producto	Ámbito	Análisis	Datos
I.A.: Soporte a la Decisión **(DSS)**	Clínica o Gestión	Modelos predictivos descriptivos y prescriptivos	Inteligencia Artificial a partir de datos (RWD)
Generación de Evidencia **(RWE)**	Investigación y/o Evaluación	Conocimiento a partir de estudio observacional	Repositorio RWD normalizado. Dato individual o agregado
Paneles visuales Cuadro de Mando **(BI)**	Gestión	Descriptiva, Diagnóstica	Dato Agregado (KPI)
Cuadernos de Investigación **(CRF)**	Investigación y/o Evaluación	Estudio Observacional y Experimental	Datos individuales.

Ilustración 4.44. Características de los 4 Outputs de un Data Lake Sanitario. Fuente propia

4.5.1. Gobernanza Organizativa, Legal y Ética

Son muchos los factores a tener en consideración en el diseño e implantación de un Data Lake Sanitario, pero dado que los aspectos legales, organizativos y éticos tienen fuertes dependencias y puntos en común, se procede a hacer un análisis conjunto con el objeto de dotar a esta propuesta de diseño de los mayores niveles de seguridad jurídica y desempeño.

Un Data Lake Sanitario consolida registros de ámbitos muy sensibles como el sanitario y el social, por lo que existe un riesgo inherente al tratamiento de información personal como el género, la raza, los hábitos de vida, los diagnósticos o tratamientos.

Datos calificados por los Comités de Ética Investigadora, en adelante CEIs, de alto riesgo de re-identificación, constituyendo el Reglamento General de Protección de Datos, en adelante RGPD, y la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y Garantía de los Derechos Digitales, en adelante LOPDGDD, el marco de referencia legal para su tratamiento.

De acuerdo a la calificación otorgada por herramientas especializadas como "Evalúa-Riesgo RGPD" (AEPD, s.f.), estos datos tienen una valoración de riesgo intrínseco y residual de nivel muy alto, por lo que, de acuerdo a lo establecido por la RGPD en su apartado 1, artículo 35, sección 3 del capítulo IV "Responsable del tratamiento y encargado del tratamiento", con carácter general, los responsables de los tratamientos de datos, tienen la obligación de realizar una Evaluación de Impacto de Protección de Datos, en adelante EIPD, con carácter previo a la puesta en funcionamiento de tales tratamientos cuando sea probable que éstos, por su naturaleza, alcance, contexto o fines, entrañen un alto riesgo para los derechos y libertades de las personas físicas**,** alto riesgo que, según el propio Reglamento se verá incrementado cuando los tratamientos se realicen utilizando "nuevas tecnologías", como las utilizadas en un Data Lake Sanitario, para finalizar este proceso implementando las recomendaciones de la EIPD.

Una EIPD debe analizar el modelo de operación de un Data Lake Sanitario y dado que uno de los principios de diseño, es que éste dé cabida a un registro longitudinal y agregado de datos de individuales y no siendo técnicamente viable la realización de cargas periódicas completas de toda la Historia Clínica Electrónica de cada paciente, la única alternativa factible, es la realización de cargas incrementales, para lo que resulta necesario contar con una referencia, a partir de la cual ir realizando la agregación periódica de datos individuales.

Para respaldar legalmente esta forma de operar, la disposición adicional decimoséptima de la LOGPDGDD introduce un elemento clave: la legitimidad del uso de datos sanitarios, sin consentimiento de los pacientes, siempre que se cumpla una serie de principios básicos:

- Seudonimizacion y compromiso de no seudonimización
- Separación técnica entre quien seudonimiza y quien hace el análisis
- Revisión por parte de un comité de ética investigadora

Dado que un Data Lake Sanitario constituye una plataforma singular, consideramos que éste debe estar respaldado por una **normativa que defina su modelo organizativo y de gobernanza**, confiriendo seguridad jurídica al proyecto, como disponen la Orden SAL/41/2022 para la creación y gobernanza de la Plataforma Analítica de Salud de la Comunidad Autónoma de La Rioja (PASCAL), o la Orden SAN/1355/2018 para la creación de la plataforma BIGAN como un

elemento del Sistema de Información de Salud de Aragón.

Un **modelo organizativo**, debe definir y delimitar el ámbito de actuación y las funciones desempeñadas por cada entidad participante respecto del origen, tratamiento y uso de los datos, estableciendo una división clara y precisa entre el uso primario y secundario de los datos, asignando las funciones y responsabilidades propias de cada ámbito, a profesionales pertenecientes a dos instituciones jurídicamente independientes y que en el caso de PASCAL y BIGAN, se materializa en:

1. Las entidades encargadas de la prestación asistencial, quedan encargadas del mantenimiento de datos en origen y su compartición (data holder y data sharing).

2. Las entidades responsables de los Data Lake Sanitarios PASCAL y BIGAN, respectivamente Fundación Rioja Salud y el Instituto Aragonés de Ciencias de la Salud, quedan encargadas de recepcionar los datos compartidos y realizar las tareas de curation, management, analysis, reporting y services.

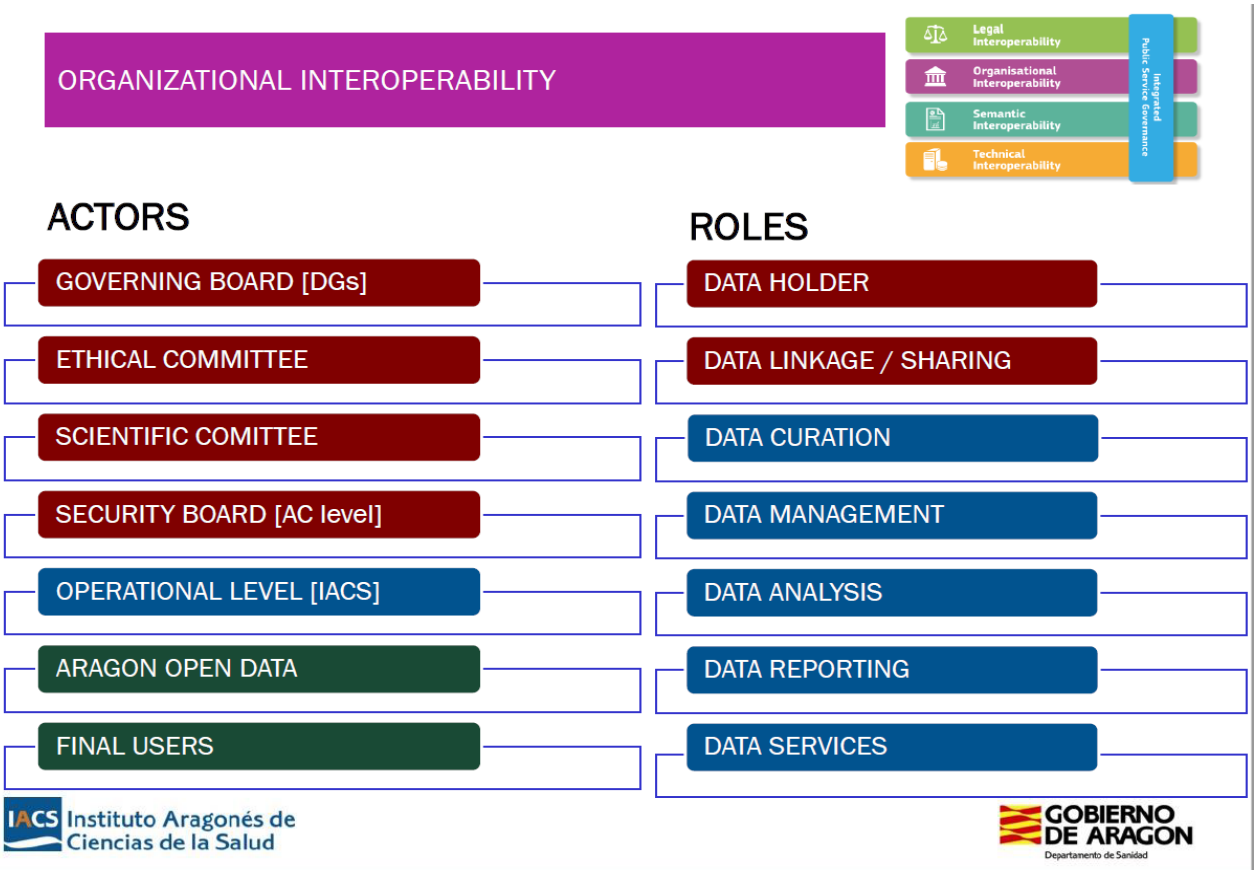


Ilustración 4.45. Modelo Organizativo de BIGAN en Aragón

Este **modelo organizativo se debe corresponder con un modelo de operación dotado de privacidad por diseño,** con garantías, incluso superiores a las requeridas por la LOPDGDD, acometiendo para ello una doble seudonimización[7], de forma que un profesional perteneciente a la entidad titular de los datos en origen, procede a realizar un primer seudonimizado de los datos, antes de que se pongan a disposición del Data Lake Sanitario, donde son seudonimizados por segunda vez, por parte de un profesional diferente, perteneciente a la entidad de destino, antes de proceder a la consolidación de los datos con el resto de registros del Data Lake Sanitario.

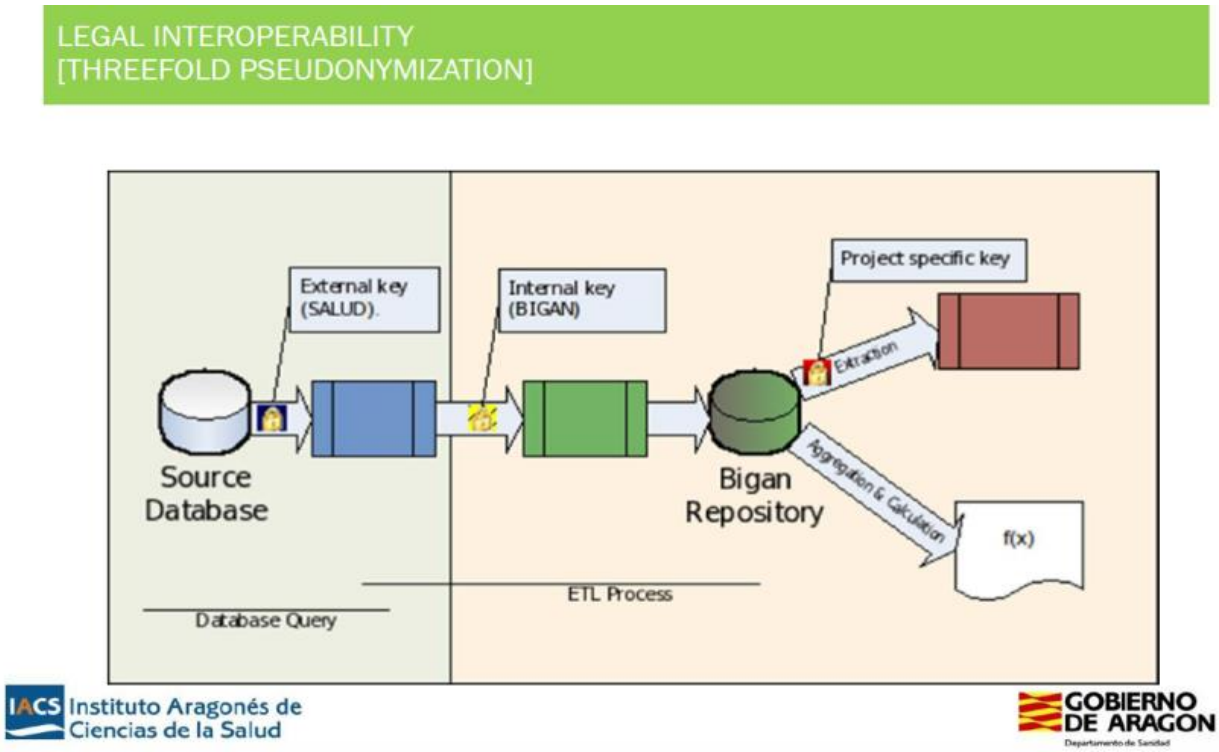


Ilustración 4.46. BIGAN. Privacidad por diseño. Esquema de doble y triple[8] seudonimizado

7 Siguiendo las "Orientaciones y garantías en los procedimientos de anonimización de datos personales" de la Agencia Española de Protección de Datos, para dificultar la re-identificación de un paciente se recomienda utilizar un cifrado de reutilización de clave basado en el algoritmo AES-256. El criptograma obtenido por este algoritmo de cifrado dispondrá de una clave conocida únicamente por el cifrador, quedando implícita en el hash creado y permitiendo su reversibilidad, aunque si la clave no es conocida, este mensaje podría corresponderse con prácticamente tantas claves, como pudieran idearse.

8 La ilustración de BIGAN, muestra cómo una vez se han consolidado los datos en el Data Lake Sanitario, se pueden destinar a realizar estudios de I+D+i y de generación de valor, o realizar un tercer seudonimizado de uno de los subconjuntos de datos (data sets o data buckets), cuando un estudio requiera su exportación, aunque de acuerdo a los principios de diseño de un Data Lake Sanitario, esta exportación debe ser tratada como una excepción a la regla y que deberá contar con una autorización previa y expresa.

Este proceso sólo será reversible, para los casos previstos en la LOPDGDD, mediante la intervención de dos profesionales pertenecientes a dos entidades independientes, que custodian las claves de forma independiente y confidencial, no siendo posible la re-identificación de un paciente y de sus datos, por parte de un único profesional e institución.

Para que los datos individuales puedan ser agregados dentro de un registro longitudinal, se requiere la elección de una referencia, que también se seudonimizará y que podría ser el CODSNS ya que dispone de significado en el ámbito regional y nacional, siempre que se encuentre mapeado con los identificadores asistenciales (números de historia clínica) de atención primaria y especializada.

En el ámbito europeo, y para el uso primario de los datos sanitarios con fines asistenciales (EHDS1), se podría recurrir como referencia al electronic IDentification, Authentication and trust Services, **eIDAS**.

En el ámbito europeo y para el uso secundario de los datos (EHDS2), la agregación de datos sobre una referencia se percibe como una actuación imperceptible en la práctica, ya que el histórico de la asistencia prestada a un ciudadano a lo largo de su vida, se encontrará registrada principalmente en un único país, por lo que el esfuerzo a realizar para referenciar estos registros a nivel europeo, probablemente no disponga de retorno de la inversión frente a la información que se conseguirá añadir en el EHDS2.

Al margen del proceso de doble seudonimizado, cada estudio requerirá realizar un análisis de **K-anonimidad** sobre el data-set objeto de la actuación. La K-anonimidad trabaja sobre los datos cuasi-identificadores[9] o seudo-claves y su objetivo es que no se pueda identificar una persona a través de ellos, ni relacionarla con sus datos sensibles, por ejemplo, el código postal, junto con la edad y la presencia de una enfermedad poco común puede hacer que el número de persona que corresponden a esa descripción sea bajo, y podría ser relativamente sencillo asociar a una persona con los datos con algún tipo de indagación adicional.

El respeto por los **principios de la bioética** (Bioethics, 2015), debe ser otro de los pilares en el diseño de un Data Lake Sanitario, ya que una falla en la privacidad o un mal uso de la información de los ciudadanos o del resultado de los trabajos, puede provocar un rechazo en la sociedad y bloquear las transferencias de datos a estas plataformas.

9 Los cuasi-ídentificadores serán definidos por el equipo encargado de la generación de cada data-set y se trataran de acorde al objetivo del estudio a realizar, asumiendo que una K-anomización completa es prácticamente imposible, por ello la anonimidad se formula como un concepto relativo ya que un conjunto de datos tiene la propiedad de ser k-anónimos si la información de todos los individuos en ese conjunto es idéntica al menos con otras k-1 personas que también aparecen en dicho conjunto. La AEPD recomienda una serie de herramienta para trabajar esta técnica, como, por ejemplo, ARX Data Anonymization Tool, un programa de código abierto que puede eliminar datos identificadores y aplicar criterios de uso para los cuasi-identificadores.

Por ello, además de las garantías legales en materia de privacidad definidas en este apartado, las diferentes investigaciones deberán respetar los principios fundamentales establecidos en la Declaración de Helsinki, en su revisión actual, de la Asociación Médica Mundial, en el Convenio del Consejo de Europa relativo a los derechos humanos y la biomedicina, en la Declaración Universal de la UNESCO sobre el genoma humano y los derechos humanos, así como cumplir los requisitos establecidos en la legislación española en el ámbito de la investigación biomédica, la protección de datos de carácter personal y la bioética.

Un Data Lake Sanitario plantea un importante reto ético en investigación y en general, en todas actividades relacionadas con las bases de datos de salud y los bio-bancos (Association, 2016), ya que todo lo realizado a partir de estos datos debería generar en beneficio de la sociedad y de los objetivos de salud pública, por lo que un Data Lake Sanitario debe contar con garantías de salvaguarda de la equidad y transparencia, tanto en la posibilidad de solicitar el acceso a los datos, como en la propiedad intelectual, en el retorno de los beneficios generados a la sociedad y evitando sesgos, como por ejemplo, que la brecha digital o la información previamente registrada, provoquen que unas poblaciones resulten más favorecidas que otras.

Como requisitos de actividad y atendiendo a su naturaleza, los proyectos acometidos en el Data Lake Sanitario deberán contar con las autorizaciones y/o informes legalmente establecidos:

- Informe de la Comisión de Investigación u órgano equivalente del centro al que pertenezca el investigador principal que deberá declarar la viabilidad de los proyectos en todos sus términos.

- Conjunto de informes y autorizaciones del Comité Ético de Investigación CEIs, y otros órganos colegiados responsables de velar por el cumplimiento de los convenios y normas existentes en materia de investigación, que se considere necesario.

- Autorización de la Agencia Española de Medicamentos y Productos Sanitarios del Ministerio de Sanidad cuando la legislación vigente así lo requiera

En el ámbito concreto de la Inteligencia Artificial, el informe de pautas éticas para una IA confiable (European Comission, 2019), concluye que la IA se debe fundamentar en tres principios que deben cumplirse a lo largo de todo su ciclo de vida:

1. debe ser legal, cumpliendo con todas las leyes y regulaciones aplicables

2. debe ser ética, asegurando el cumplimiento de principios y valores éticos

3. debe ser sólida, tanto desde una perspectiva técnica como social, ya que, incluso con buenas intenciones, los sistemas de IA pueden causar daños no intencionales.

Y define una lista de evaluación de siete requisitos claves, que los sistemas de IA deben cumplir para ser considerados confiables.

1. **Agencia humana y supervisión:** los sistemas de IA deben empoderar a los seres humanos, permitiéndoles tomar decisiones informadas y fomentando sus derechos fundamentales. Al mismo tiempo, se deben garantizar los mecanismos de supervisión adecuados.

2. **Robustez técnica y seguridad:** los sistemas de IA deben ser seguros, garantizar un plan de respaldo en caso de que algo salga mal, además de ser precisos, confiables y reproducibles.

3. **Privacidad y gobierno de datos:** además de garantizar el pleno respeto a la privacidad y la protección de datos, también se deben garantizar mecanismos adecuados de gobierno de datos, teniendo en cuenta la calidad de los datos, y asegurando el acceso legitimado.

4. **Transparencia:** los modelos comerciales de datos, sistemas e IA deben ser transparentes y los mecanismos de trazabilidad pueden ayudar a lograr esto. Además, los sistemas de IA y sus decisiones deben explicarse de manera adecuada y los seres humanos deben ser conscientes de que están interactuando con un sistema de IA y estar informados de las capacidades y limitaciones del sistema.

5. **Diversidad, no discriminación y equidad:** Los sistemas de IA deben ser accesibles para todos, e involucrar a las partes interesadas relevantes a lo largo de todo su ciclo de vida y debe evitarse el sesgo injusto, ya que podría tener múltiples implicaciones negativas, desde la marginación de grupos vulnerables hasta la exacerbación de los prejuicios y la discriminación.

6. **Bienestar social y ambiental:** los sistemas de IA deberían beneficiar a todos los seres humanos, incluidas las generaciones futuras, por lo tanto, debe garantizarse que sean sostenibles y respetuosos con el medio ambiente. Además, deben tener en cuenta el medio ambiente, incluidos otros seres vivos, y su impacto social debe considerarse cuidadosamente.

7. **Rendición de cuentas:** deben establecerse mecanismos para garantizar la responsabilidad y la rendición de cuentas de los sistemas de IA y sus resultados. La auditoría, que permite la evaluación de algoritmos, datos y procesos de diseño, juega un papel clave en esto, especialmente en aplicaciones críticas.

4.5.2. Gobernanza de la Colaboración y los Resultados

Un modelo de Gobernanza para la colaboración con **privacidad por diseño** aporta máximas garantías legales y se puede conseguir a través del aprendizaje federado, donde los datos permanecen en sus repositorios de origen, siendo las consultas o algoritmos los que viajan hasta los datos, una forma de operar propuesta por el proyecto EHDEN para las iniciativas de RWE, o por la red federada con IA para acelerar la investigación clínica y sanitaria propuesta por Tartaglia.

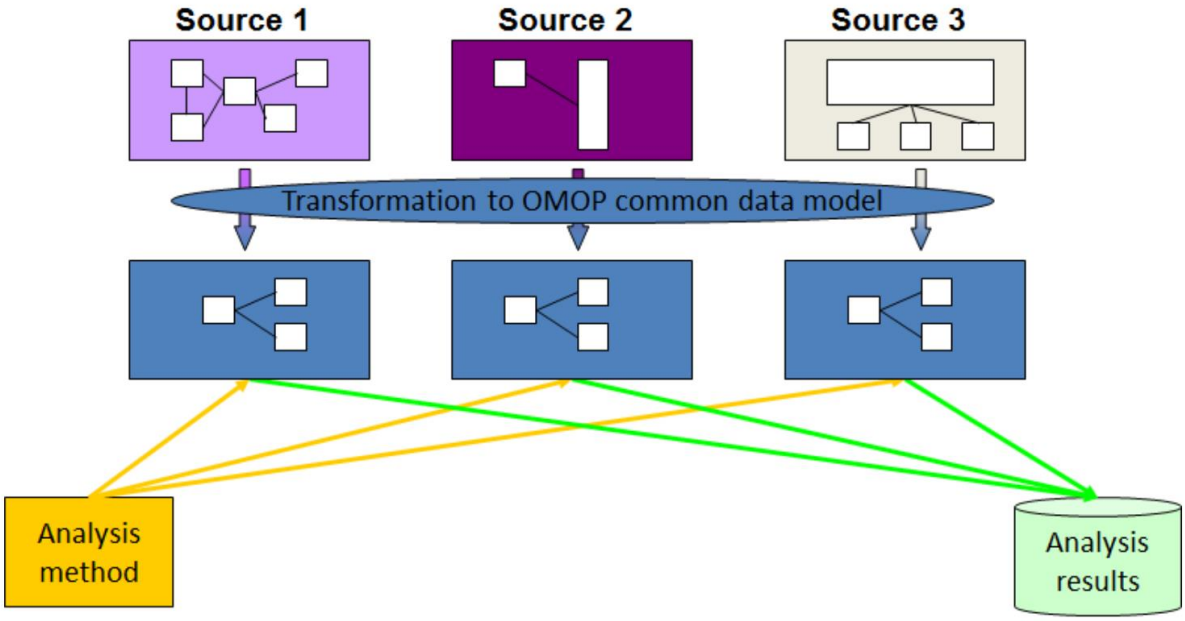


Ilustración 4.47. Arquitectura de aprendizaje federado en EHDEN. Fuente OHDSI

En EHDEN no se procede a la compartición de datos o registros individuales, únicamente se comparte el resultado de las consultas, que son comparables y agregables, porque previamente se han estandarizado las bases de datos a un modelo de datos normalizado conocido como OMOP-CDM, que además almacena conceptos expresados en vocabularios normalizados.

Este mismo planteamiento es extensible al aprendizaje automático federado, donde los datos no deben estar en una misma localización, aproximación centralizada, para poder entrenar una IA, sino que va a ser el algoritmo de la IA, el que viaje donde están los datos, desarrollando los algoritmos con un enriquecimiento progresivo mediante su entrenamiento en múltiples fuentes de datos, o aproximación descentralizada.

En este instante, buena parte de las herramientas con capacidad para proveer soporte a la decisión clínica, tienen su procedencia en la iniciativa privada, encontrando un buen ejemplo en el Plan de Inversión en Alta Tecnología, **plan invEAT**, ejecutado de forma conjunta por el Ministerio de Sanidad, Ingesa y las Comunidades Autónomas y donde se demandan sistemas de inteligencia artificial basados en aprendizaje profundo.

Ilustración 4.48. Funcionalidades de Deep Learning en contratación de RMN. Plan Inveat.

Aunque estos productos ya están siendo comercializados por empresas, a diferencia del software tradicional, el desarrollo de la analítica avanzada y de la inteligencia artificial, requiere del acceso a datos sanitarios, frecuentemente custodiados por entidades públicas, es por ello que para cumplir con los objetivos establecidos por la estrategia digital del Sistema Nacional de Salud y los Espacios de Datos Europeos, se van a requerir escenarios de cooperación público-privada, que permitan acometer iniciativas conjuntas de los Sistemas Públicos de Salud y las empresas.

Este tipo de colaboraciones son habituales en ensayos clínicos y proyectos de investigación y requieren de un soporte jurídico que las regule, en la forma de acuerdos de colaboración o similares, donde se definen las aportaciones realizadas por cada parte, la cotitularidad de la propiedad intelectual, e incluso la coparticipación de los beneficios generados por el resultado de los trabajos[10].

La firma de este tipo de acuerdos requiere de una flexibilidad jurídica, de la que normalmente carecen los Gobiernos Autonómicos o los Servicios Regionales de Salud, es por ello que la I+D+i tradicionalmente se está vehiculizando a través de instituciones de naturaleza fundacional, Institutos de Investigación o similares, con amplia experiencia en la participación de consorcios de colaboración público-privada.

10 La Comisión Europea en el programa H2020 y Horizon Europe, define un modelo donde la propiedad de un Resultado o Conocimiento resultante pertenece en exclusiva a la parte que haya sido responsable de su desarrollo y si el conocimiento resultante, total o parcialmente, ha sido desarrollado por dos o más partes y no es posible separar la contribución de cada parte al mismo, los derechos de propiedad sobre el mismo pertenecerán a cada parte proporcionalmente a su intervención en la generación del mismo, por lo que, en estos casos, cada uno de los cotitulares tendrá derecho a usar el conocimiento resultante de manera más conveniente y a otorgar licencias a terceros, sin derecho a sub-licencia.

En multitud de ocasiones, la constitución y participación en estos consorcios representa una obligación más que una opción, de lo contrario no es posible cumplir con los requisitos establecidos por concurrir a las convocatorias que están movilizando buena parte de los fondos europeos.

Por todo lo expuesto y coincidiendo con el apartado anterior, este TFM concluye que el modelo de doble institución, planteado por el Gobierno de Aragón y el Gobierno de La Rioja, de modo que el uso secundario de los datos es gestionado, respectivamente, por un instituto (IACS) y una fundación (Rioja Salud), es considerado el más indicado para disfrutar de la flexibilidad requerida para perfeccionar las colaboraciones público-privadas.

Cuando el resultado de los trabajos de un Data Lake Sanitario sean algoritmos empleados en el diagnóstico, prevención, seguimiento, predicción, pronóstico, tratamiento o alivio de una enfermedad, lesión o discapacidad, también requerirán de una gestión especial de acuerdo a lo establecido por el Reglamento (UE) 2017/745 sobre Productos Sanitarios, en adelante MDR.

El MDR establece las condiciones para garantizar la disponibilidad en el mercado de productos sanitarios eficaces, de calidad y seguros y al tratarse la IA de un software sanitario, es de aplicación la regla 11 categorías IIa, IIb o III, en función de los riesgos asociados al algoritmo en la decisión clínica, requiriendo un análisis de riesgos de producto, conforme a la norma UNE EN ISO 14791, cumplir con el Anexo I del MDR de requisitos generales de seguridad y funcionamiento, validar la plataforma conforme a la Norma UNE EN 62304, garantizar el cumplimiento del RGPD 2016/679 y por ultimo realizar una evaluación clínica bibliográfica o un ensayo clínico por cada algoritmo, antes de poder obtener el marcado CE como producto sanitario.

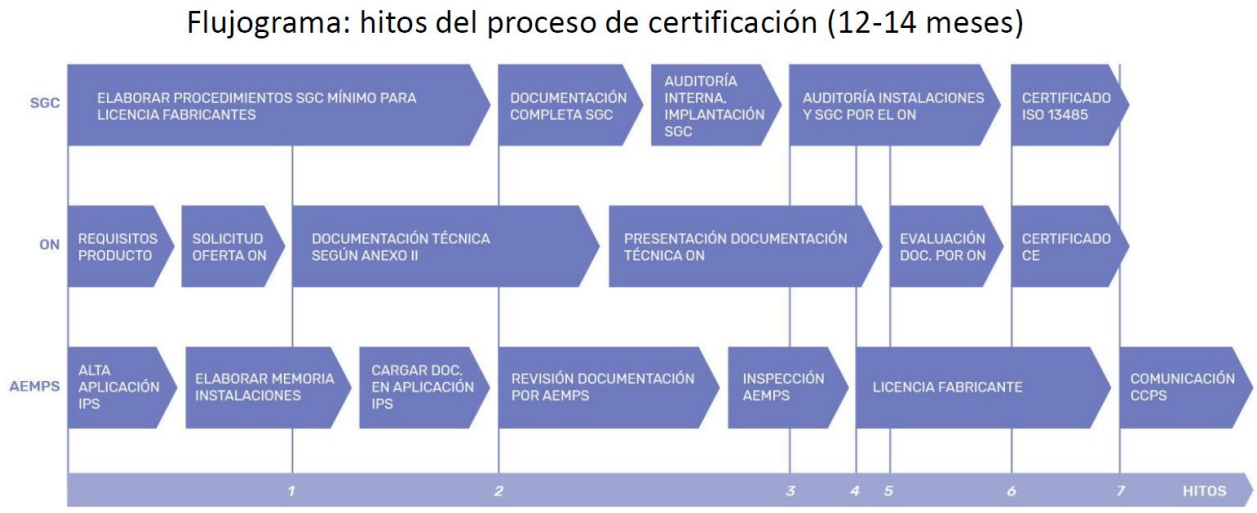


Ilustración 4.49. Procedimiento estándar obtención del marcado CE

4.5.3. Gobernanza del Dato

Para dar respuesta a la pregunta ¿cómo debemos gobernar los datos?, existen diversas metodologías internacionales, como la elaborada por Data Management Association, en adelante DAMA, una asociación internacional cuya principal misión es promover y facilitar el desarrollo de la cultura de gestión de los datos, convirtiéndose en la referencia para las organizaciones y profesionales en la gestión de la información, aportando recursos, formación y conocimiento sobre la materia, y presentando las mejores prácticas para garantizar el control sobre la información.

DAMA posiciona el Gobierno del Dato como la principal actividad, alrededor de la cual, se gestionan otras actividades, que son objeto de análisis especifico de este TFM, como la arquitectura, privacidad/seguridad, interoperabilidad, calidad de los datos, etc.

Ilustración 4.50. Gobierno de Datos. DAMA

DAMA define como objetivo del Gobierno del Dato, la definición de todas las funciones de la Gestión del Dato para asegurar que los datos son gestionados adecuadamente, de acuerdo con una serie de políticas y mejores prácticas, de forma que:

- Las organizaciones que establecen un programa formal de Gobierno del Dato están mucho más capacitadas para incrementar el valor que obtienen de sus activos de datos.
- Un programa de Gobierno del Dato debe ser sostenible, embebido y medible y requiere planificación para ser implementado.
- El Gobierno del Dato se focaliza en cómo se toman las decisiones acerca de los datos y cómo se espera que los procesos y las personas se comporten en relación con los datos.
- El Gobierno del Dato no es un fin en sí mismo, necesita estar directamente alineado con la estrategia de la organización.
- El Gobierno del Dato requiere de un programa continuado y focalizado en asegurar que la organización obtiene valor de los datos y reduce los riesgos relacionados con los datos.
- El Gobierno del Dato es diferente del Gobierno de TI, gestiona los datos como un activo.

Con el objetivo de intentar aclarar las dudas en esta materia, la Asociación Española de Normalización (UNE), también ha procedido a publicar diversos documentos de apoyo en torno a los que considera los cuatro pilares básicos para el gobierno del dato:

- Gobernanza
- Gestión
- Calidad
- Seguridad y privacidad de datos

poniendo a disposición de las organizaciones que quieran implementar un marco de gobierno sólido una serie de normas técnicas que proveen principios guiadores para garantizar que los datos de una organización son correctamente gestionados y gobernados, tanto internamente como por contrataciones externas.

En opinión de Javier Peris, presidente del subcomité UNE de Gestión de servicios TI y Gobierno de TI, a día de hoy las organizaciones que pretendan obtener beneficios y quieran adoptar un enfoque estratégico hacia el Gobierno del Dato pueden hacer uso de las normas UE para sacarle el máximo valor logrando sus objetivos estratégicos.

Pero no hay que olvidar que no es lo mismo gobierno que gestión; gestionar es recetar lo que nos pide el paciente, mientras que gobernar es recetar lo que verdaderamente necesita el paciente.

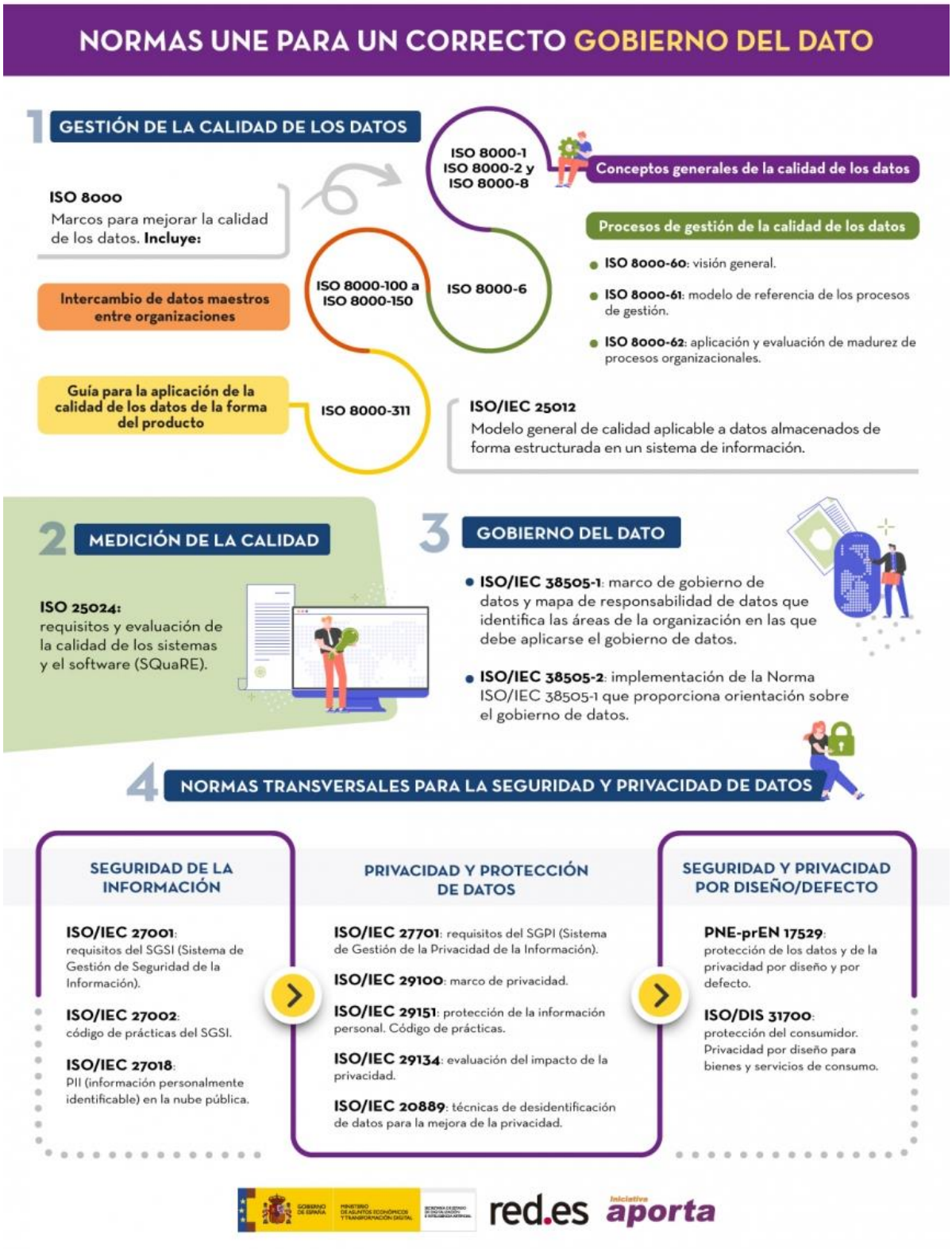


Ilustración 4.51. Normas UNE para Gobierno del Dato. Fuente datos.gob.es

4.5.3.1. Ciclo de vida de los Datos

Como recogen los contenidos del tema “Analítica y Modelos Predictivos en Salud” del master DSTICSDS, el ciclo de vida de los datos para el Descubrimiento de Conocimiento en Bases de Datos o Knowledge Discovery in Databases, en adelante KDD, y donde se enmarcan un conjunto de actividades como la Minería de Datos o Data Mining y el Aprendizaje Automático o Machine Learning, define un conjunto de actividades que forman parte de todo proceso de KDD:

- Determinar las fuentes de información que pueden ser útiles y dónde conseguirlas.
- Diseñar el esquema de un almacén de datos, data warehouse, que consiga unificar de manera operativa toda la información recogida.
- Implantación del almacén de datos que permita la “navegación” y visualización previa.
- Selección, limpieza y transformación de los datos.
- Seleccionar y aplicar el método de minería de datos apropiado, que servirá para obtener patrones de los datos.
- Evaluación, interpretación, transformación y representación de los patrones extraídos.
- Comunicación y uso del nuevo conocimiento.

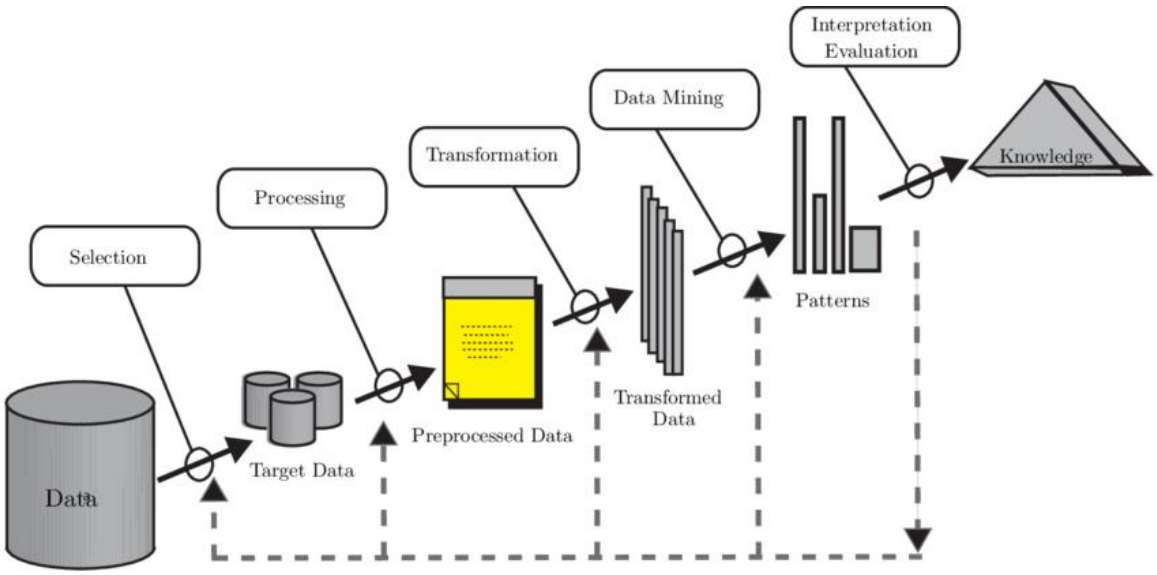


Ilustración 4.52. Knowledge Discovery in Databases. Fuente Brachman y Anand

El traslado de este proceso al ámbito sanitario, requiere de un análisis más preciso y detallado, ya que, en la extracción y consolidación de información a partir de múltiples fuentes, se deben tener en consideración más aspectos, como la confidencialidad y privacidad, abordados de forma explícita en este TFM, pero también la **transparencia**, que debe estar presente en todas las actuaciones que son acometidas con un presupuesto público y por el interés general, es por ello que este trabajo plantea la necesidad de abordar los proyectos de un Data Lake Sanitario a través de tres fases.

Ilustración 4.53. Tres fases del Ciclo de los Datos. Fuente eleboración propia

Un Data Lake Sanitario deberá incluir un site público, usable y accesible para la **Consulta de los Datos**[11] **y la Solicitud de Acceso** y donde se muestre:

1. **Listado de recursos abiertos**, modelos, software, herramientas y en general activos de dominio público, también conjuntos de datos completamente anonimizados e irreversibles cuando se decida compartir como open data, formación on-line, etc.

2. **Listado de resultados**, de los trabajos ya finalizados, que pueden ser modelos de inteligencia artificial, paneles visuales, reglas de inferencia, etc., y qué a diferencia del listado anterior, si estarían sujetos a derechos de propiedad intelectual.

3. **Listado de iniciativas activas,** que estén siendo abordadas en ese instante, indicando:

   a. Información contextual: objetivo, cohortes, plazos, financiación data-set, etc.
   b. Responsable: Investigador principal, profesional o gestor.
   c. Carácter de la iniciativa: local, consorcio, investigación en red.
   d. Posibilidades de colaboración y perfiles requeridos.
   e. Canal de solicitud para que los ciudadanos voluntariamente puedan ejercer sus derechos[12] y

11 No se plantea la publicación de los datos como tal**,** si de los metadatos asociados a estos datos, precisando información relevante, como su volumetría, rangos de fechas, origen, significado, normalizaciones disponibles y demás información de contexto.
12 De acuerdo con la normativa vigente, en particular la Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica

manifestar su voluntad de excluir ("opt-out") su información del Data Lake Sanitario, lo que se debería realizar, eliminando la información de estas personas en los procesos de extracción de datos desde las fuentes de origen.

4. **Listado de datos (metadatos)** existentes en el Data Lake Sanitario, procurando que esta información sea fácilmente localizable y disponga de un formato visual, usable y accesible, para que los investigadores, puedan plantear sus proyectos y estudios.

5. **Procedimiento de acceso a los datos:** Una vez los investigadores o profesionales hayan localizado una iniciativa en curso en la que quieren colaborar o un conjunto de datos de interés para un determinado proyecto o estudio, deben poder consultar información precisa sobre los pre-requisitos necesarios, y el procedimiento de solicitud y aprobación previo al acceso, y que según las características del estudio puede requerir su evaluación por parte de:

    a. **El Comité Científico,** para analizar la viabilidad del proyecto y velar por el interés público y científico

    b. **El Comité de Ética Investigadora**, para velar por el cumplimiento de los convenios y normas existentes en materia de investigación, que se considere necesario.

    c. **AEMPS**, para contar con su autorización en estudios con medicamentos y productos sanitarios, cuando la legislación vigente lo requiera.

    En el supuesto de que la solicitud sea aprobada y antes de conceder el acceso al investigador, éste deberá proceder a la firma de un acuerdo de confidencialidad[13], quedando obligado, entre otros, a reportar con inmediatez cualquier vulnerabilidad o falla en materia de privacidad, en el hipotético caso de que esta se produzca.

Este proceso de autorización, irá precedido de una serie de actuaciones que tienen por objeto mejorar la legibilidad y calidad de los datos que se almacenaran en el Data Lake Sanitario, y que este TFM ha denominado "**Ciclo de Preparación de los datos"** y que debería ser abordado de acuerdo a los principios FAIR (Wilkinson, 2016), para que los datos / metadatos sean:

- Findables o encontrables, de forma que el Data Lake Sanitarios y el descriptivo de su contenido pueda ser localizado.

- Accesibles, para que exista una formula conocida para acceder o solicitar el acceso a los datos.

- Interoperables, para que estén formateados y descritos conforme a ontologías, modelos de datos y vocabularios controlados, habilitando el trabajo en red y la consolidación y comparación con los datos de terceros.

13 Basado en las normas de buena práctica clínica, protección de datos y gestión de muestras biológicas

- Reusables, para que dispongan de licencias, documentación, etc. que permita su uso en múltiples estudios.

Pero la adopción de los principios FAIR no es suficiente para habilitar la generación de conocimiento y la mejora de resultados de salud, mediante la compartición y agregación de datos, dado que estos principios no tienen en consideración la calidad de los datos en origen, su procedencia, linaje o la gestión de factores éticos y de privacidad.

Es por ello que, con el objetivo de que un Data Lake Sanitario pueda operar más allá de su ámbito interno, se propone la ampliación de los principios FAIR desarrollada por la iniciativa FAIR4HEALTH (FAIR4HEALTH, 2018), que ha desarrollado indicadores para la evaluación del cumplimiento de las recomendaciones FAIR y una guía de implementación basada en la "fairificación" de GO FAIR (GO-FAIR, 2017) y adaptada a las especificidades del ámbito de los datos de salud y además ofrece herramientas basadas en el uso del estándar HL7 FHIR (FHIR4FAIR, 2021)

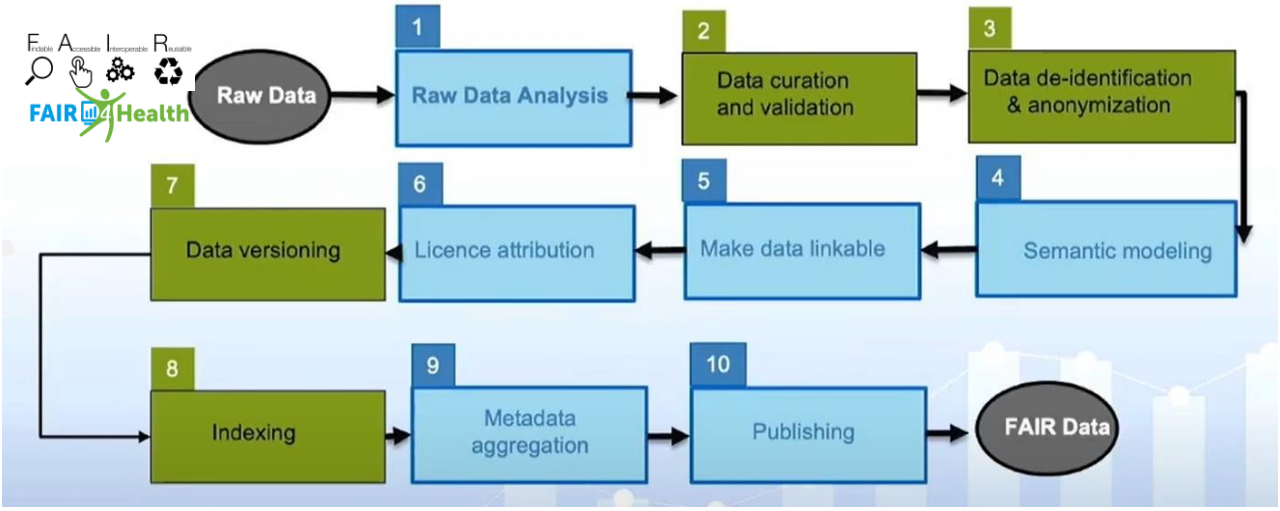


Ilustración 4.54. "Fairificación" de datos. Fair4Health. Fuente Master DSTICSDS

Para el desarrollo de los trabajos de preparación de los datos, además de las herramientas de la iniciativa FAIR4health, existen un gran número de soluciones tecnológicas, que serán objeto de análisis en otro capítulo de este TFM, para dar soporte a una serie de tareas:

1. Generación de un mapa con las fuentes de datos.
2. Des-identification: anonimizado, seudonimizado, k-anonimizado.

3. Ingesta de datos.
4. Análisis de datos crudos (raw)
5. Curación y Validación de Datos
6. Transformación de los datos raw, a un modelado con significado/semántica:
   a. Ontologías
   b. Modelos de datos
   c. Vocabulario
7. Procesar la información para que se pueda encontrar
8. Publicación de metadatos en listado de datos

Tras la preparación y el acceso a los datos, se acometerán otro conjunto de actuaciones, que este TFM ha denominado "**Ciclo de Producción con los Datos"** y que cuenta con las siguientes características:

1. En cumplimiento con el principio de minimización de datos, el investigador/profesional recibe la autorización para acceder a un data-set o subconjunto de datos del Data Lake Sanitario, que incluye los datos estrictamente necesarios para la realización del estudio aprobado, trabajos que además serán controlados mediante:

   a. Gestión de Usuarios
      i. Altas
      ii. Bajas
      iii. Modificaciones

   b. AAA
      i. Acceso: Quien accede
      ii. Autorización: A que puede acceder
      iii. Auditoria[14]: Quien accedió a que y que se hizo con los datos

   c. Trazabilidad del linaje del dato, como se transformó, por donde paso y quién lo hizo.

   Para cumplir con la máxima de diseño[15] de que "el investigador vaya al dato, evitando que los datos viajen hasta el investigador", conjuntamente con el data-set de trabajo, el Data Lake Sanitario pondrá a disposición del profesional/investigador un conjunto de herramientas (software, cuadernos jupyter, R-studio, DW/BI, servidores de modelos AI,

14 En cualquier instante, durante la ejecución de un estudio o una vez finalizado, debe existir la posibilidad de auditar la actividad, obteniendo información de la trazabilidad del linaje del dato, por donde paso, como se transformó, quien accedió a él y en que instante, resultado de los trabajos (algoritmo, panel de visualización, …), etc.

15 Durante la ejecución de estos trabajos, y salvo autorización expresa y de forma excepcional, no se debería permitir la extracción de los datos contenidos en el data-set de trabajo, únicamente el resultado s de los trabajos.

servidores terminológicos, etc.) típicas de la analítica, además de recibir soporte para su utilización, formación y/o servicios especializados en ciencia del dato, para el desarrollo análisis avanzados, paneles observacionales, algoritmos de soporte a la decisión, etc.

2. Una vez se disponga del resultado de los trabajos se debería:

    a. Proveer soporte a los procesos de protección intelectual/industrial, cuando sea de aplicación.

    b. En el supuesto de que el resultado sea un algoritmo de soporte a la decisión, se procederá a documentar su comportamiento (explicabilidad) y vehiculizar el mismo para su consumo de forma integrada, transparente y usable desde la Historia Clínica Electrónica.

    c. Siguiendo las recomendaciones de la UNESCO sobre ciencia abierta, se debería proceder a la publicación en el listado de resultados de trabajos del Data Lake Sanitario.

Existe un gran número de soluciones y herramientas tecnológicas para el desarrollo de los ciclos de preparación y producción con los datos, que serán abordados en otro capítulo de este TFM y con los que se pueden acometer diferentes tipos de estudios clasificables como:

- **Descriptivos,** se estudian las diferentes variables independientemente.
- **Exploratorios**, se buscan relaciones entre variables.
- **Inferenciales,** generalizar hipótesis a partir de los datos a una población general.
- **Predictivos**, se estiman o predecir una clase o magnitud a partir de datos.
- **Causales,** se establece la relación causa-efecto de ciertos hechos.
- **Mecanísticos**, se caracteriza de manera matemática los procesos naturales.

Estudios que también puede ser tipificados en base al valor que aportan en el soporte a la decisión, como descriptivos, diagnósticos, predictivos, prescriptivos y cognitivos.

Ilustración 4.55. 5 tipos de analítica. Fuente WeirdGeek

4.5.3.2. Calidad de los Datos

Tras la transferencia de las competencias en materia sanitaria en España, el desarrollo experimentado por las TIC ha sido tan notable, como carente de una estrategia transversal y consensuada que garantice aspectos tan críticos, como la calidad de los datos a lo largo de todo su ciclo de vida y la realización de auditorías periódicas para velar por la precisión y legibilidad de la documentación producida.

La calidad de los datos representa el factor de mayor impacto en el resultado de los proyectos de análisis de la información, motivo por el cual requiere una supervisión especial a lo largo de todo el proceso de trabajo con los datos, principalmente en el ciclo de preparación de los datos y metadatos y hasta que estos son publicados, pero también durante la fase de producción con los datos, ya que en ocasiones no se detecta un problema hasta que los datos no son analizados desde una perspectiva global o consolidada.

Los problemas de fiabilidad de la información, suelen tener su origen en el propio diseño de las aplicaciones y en la forma en la que son utilizadas, siendo frecuente que recojan diagnósticos genéricos, que los registros sean copiados o importados de otros lugares, se registre la información en campos donde no corresponde, no se respete la estructura normalizada de documentos clave (como los informes de alta), o incluso se llegue a mantener otros registros en paralelo para cumplir con las necesidades asistenciales, de gestión o de investigación.

Esta falta de fiabilidad de la información se puede materializar en problemas de eficiencia, eficacia, seguridad del paciente y calidad de la prestación asistencial, y en lo referente a los usos secundarios de los datos de RWD, una baja calidad de los mismos será irremediablemente arrastrada al Data Lake Sanitario, por lo que resulta fundamental mejorar la calidad de los datos en origen, y supervisar en el Data Lake que una baja calidad de los datos de entrada no arruine los resultados (GIGO: Garbage In - Garbage Out) de nuestros proyectos.

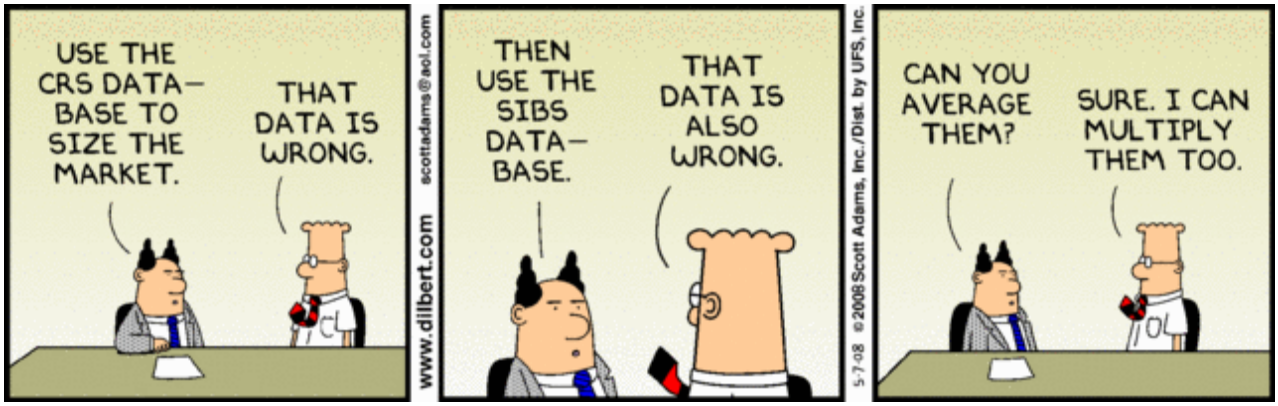


Ilustración 4.56. Dilbert by Scott Adams

Alcanzar una buena calidad en los datos, es una tarea crítica que forma parte del procesamiento inicial de los datos y puede llegar a consumir el 60% del tiempo total de un proyecto

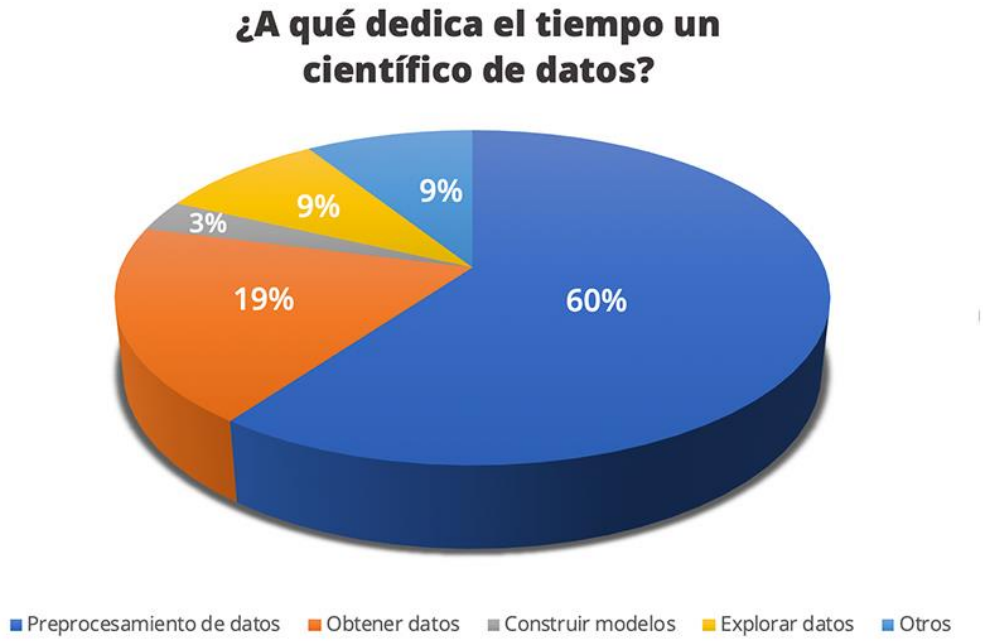


Ilustración 4.57. Distribución de tiempos Data Science Project. Fuente Xeridia

y que debe abordarse con un enfoque multidimensional:

- Precisión: hasta qué punto el dato representa la realidad
- Completitud: que no falten datos
- Consistencia: entre diferentes fuentes para un mismo dato
- Razonabilidad: que tenga sentido el valor del dato
- Temporalidad: que los datos estén referidos al tiempo correcto
- Unicidad: que no haya datos duplicados
- Validez: que el formato sea correcto.

De forma resumida, este proceso de calidad se puede estructurar en cinco tipos de actuaciones:

- Exploración general de los datos, incluyendo tipología de variables.
- Revisión resumida de estadísticos básicos (missings/min/max/negativos/medias, …).
- Analizar valores nulos/ceros/atípicos para descartar registros/variables/imputación.
- Detectar coherencia entre variables (anidaciones/visitas/fases/periodos, …).
- Proponer correcciones de mejora de los datos basada en la información global.

Las labores de depuración de los datos para su uso secundario, va a permitir identificar muchos problemas de calidad en origen y por tanto ayudará a revertirlos, por ejemplo, mediante el desarrollo de guías, pautas, nuevos interfaces de captura de datos, más usables y normalizados, para mejorar el registro de la información durante la asistencia y con ello la calidad de la información.

4.5.3.3. Interoperabilidad de los Datos

Si un Data Lake Sanitario contiene datos de elevada calidad, que proceden de los sistemas de información que dan soporte a la clínica y a la gestión de pacientes, ya se dispone de la materia prima necesaria para abordar análisis avanzados que cumplan con los objetivos gestores de mejorar los resultados en salud y la prestación asistencial, siempre que entre los requisitos del proyecto no se establezca la necesidad de comparar los resultados con terceras entidades.

En ese caso surgirá un problema, ya que cuando el trabajo con los datos no está destinado a un consumo exclusivamente interno; en éste supuesto, además de velar por la calidad de los datos, se debe acometer un trabajo de normalización de la información, para que ésta signifique lo mismo dentro y fuera de nuestra organización, habilitando la consolidación necesaria en investigación y la comparación de resultados para la mejora en materia de eficacia, calidad y especialmente de eficacia, por ejemplo para ayudar a desarrollar las políticas de “no hacer”, ya que no hay nada que resulte más ineficiente, que pretender optimizar aquello que no se debería hacer.

Dentro de los principios FAIR, pensados para desarrollar la legibilidad de los datos, la Interoperabilidad permite potenciar la investigación en red, ya que es reconocida como un factor clave para la escalabilidad multinivel (datos, investigadores, resultados, relevancia y financiación) de los proyectos de investigación, frente al uso de modelos observacionales tradicionales.

La interoperabilidad se basa en el uso de un lenguaje común, donde los datos son normalizados de acuerdo a un estándar que hay que elegir, aunque esto no resulte una tarea sencilla, dada la multitud de opciones existentes.

A pesar de ello, siempre es preferible optar por un estándar y complementarlo con nuevas versiones y mapeos a otros estándares, y así ir incrementando progresivamente su versatilidad y ámbito de actuación, que trabajar con formatos propietarios y/o locales, que carecen de sentido fuera de nuestra organización, dificultando enormemente la participación en proyectos de investigación en red y la mejora mediante la comparación de resultados.

Para disponer de una interoperabilidad semántica plena, que permita compartir y combinar con pleno significado los datos de salud entre sistemas heterogéneos, se requiere de un vocabulario normalizado, normalmente una terminología y también de un modelo de datos común, una estructura, también normalizada, donde se van a almacenar los vocabularios; existiendo modelos

de datos más indicados para el uso primario de los datos y otros, para el uso secundario.

Para evitar persistir en la inflexibilidad de los modelos tradicionales que dan soporte a la mayoría de Historias Clínicas Electrónicas, en la elección de un modelo o modelos para un Data Lake Sanitario, hay que tener presentes los modelos duales, que ofrecen importantes beneficios en la representación del conocimiento.

Partiendo de una ontología que recoge el conocimiento de forma general, como por ejemplo el existente en torno al accidente cerebro vascular isquémico, y que se podría encontrar definido en fuentes como la ISO13940 “System of Concepts to Support Continuity of Care” o en el Diccionario de Términos Médicos de la Real Academia de Medicina de España, se avanza en realizar un análisis de carácter epistemológico para la elección de una serie de términos o conceptos, que en su conjunto permitan representan ese conocimiento general, y que se materializan en un arquetipo, construido con las “piezas” existentes en el modelo de referencia para representar el conocimiento existente en un determinado instante y que podrá ser evolucionado a futuro si el conocimiento general es modificado o ampliado.

Ilustración 4.58. Modelos Clínicos Detallados o Duales. Fuente Master DSTICSDS

El gran beneficio de los modelos clínicos detallados o duales está en su flexibilidad, por ejemplo, si una sociedad científica decide que hay que tener en consideración el Tiempo en Rango Terapéutico de pacientes anti coagulados para el análisis del ictus, o si evolucionamos la HCE resumida estructurada introduciendo un nuevo campo al informe, mientras los modelos tradicionales se ven obligados a insertar un nuevo atributo en las tablas de nuestro SGBD, además de modificar el código fuente que soporta el formulario de entrada de datos, en los modelos duales se puede dar cabida a estos nuevos esquemas de conocimiento, sin más que proceder a evolucionar el arquetipo.

Para avanzar desde una historia clínica electrónica centrada en el registro y consulta de información, hasta un sistema de información capaz de proveer soporte a la decisión, los arquetipos como elementos de representación del conocimiento, se deben complementar con modelos de inferencias, donde se recogen reglas de actuación que habiliten mecanismos de generación de alertas, vías clínicas, etc. y que al igual que los arquetipos, deben tener la capacidad de evolucionar cuando lo haga el conocimiento asociado e estas reglas de actuación, por ejemplo, en la detección temprana de episodios de septicemia.

Estos modelos de representación del conocimiento pueden soportarse en servidores de terminologías, servidores de arquetipos y servidores de inferencias, en cualquier caso y como se avanzaba al principio de este apartado, es necesario decantarse por un Common Data Model, para lo que este TFM hará suyas las conclusiones del estudio "¿Pueden trabajar juntos OpenEHR, ISO 13606 y HL7 FHIR? (Jiménez, 2022).

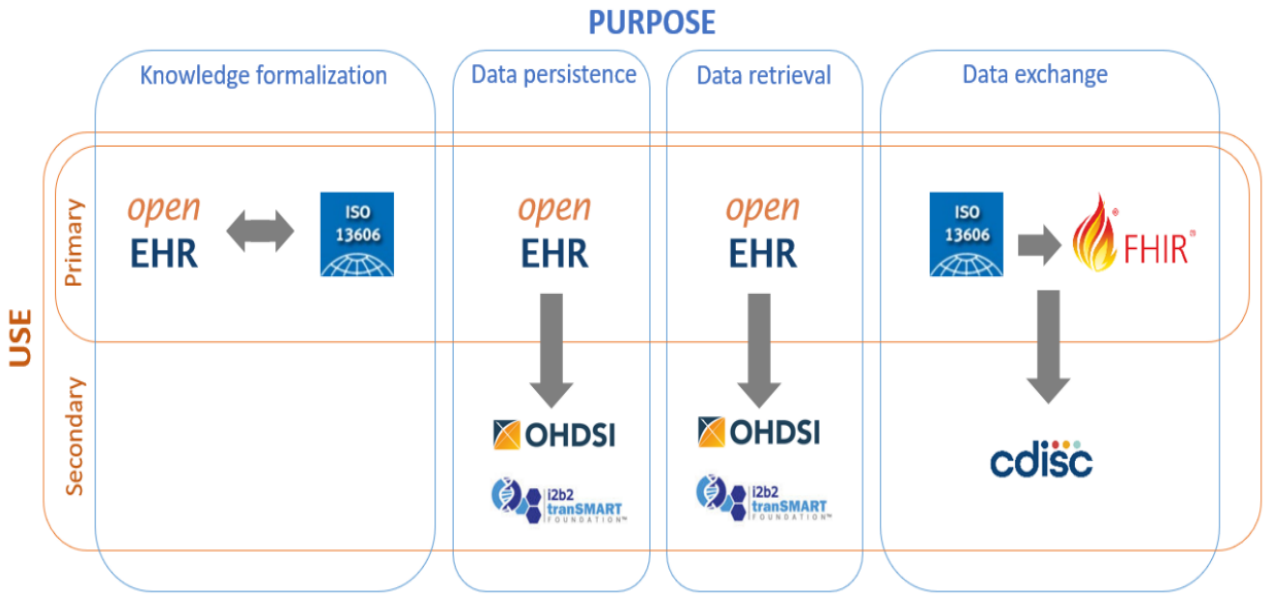


Ilustración 4.59. Agnostic perspective on selection and traslation of EHR standars

Aunque los modelos como Open EHR, ISO 13606 y HL7 FHIR fueron pensados para dar soporte a la Historia Clínica Electrónica, su diseño les permite ser utilizados conjuntamente para extender las necesidades de formalización de conocimiento, persistencia, recuperación e intercambio de datos con diferentes grados de complejidad, por lo que se podría implementar, por ejemplo, con los recursos que ofrece el modelo dual OpenEHR y compartir los datos con ISO 13606 o HL7 FHIR.

OpenEHR es un modelo que podría dar soporte también al uso secundario de los datos, aunque existen otros estándares específicamente desarrollados para tal fin, cómo lo son i2b2 (Informatics for Integrating Biology and the Bedside) y OMOP (Observational Medical Outcomes Partnership) para la persistencia de los datos y CDISC (Clinical Data Interchange Standards Consortium) para el intercambio, además de iniciativas, como la procedente de la FDA (Food and Drug Administration), trabajando en la armonización de estos modelos de datos a través de estándares orientados a la integración, como el propio CDISC o HL7 FHIR.

OMOP e i2b2 son modelos de datos versátiles que ofrecen diferentes aproximaciones:

- OMOP se caracteriza por ligar el modelo de datos con el modelo semántico y destaca por su capacidad para contener información clínica en un modelo común estandarizado, que le permite hacer uso de diferentes vocabularios estándar y además simplificar su mapeo,

- i2b2 dispone de un modelo de datos y una ontología para generar los conceptos clínicos que van a ser utilizados en los estudios, lo que le confiere mucha flexibilidad, aunque pueda acabar siendo diferente para cada estudio, lo que dificulta la interoperabilidad con terceros.

El estudio “Modelos de Datos para la Utilización Secundaria de Historias Clínicas: Experiencia de un Conector de OMOP a i2b2” (D. PEREZ-REY, 2018), desarrolló un conjunto de scripts capaces de cargar un repositorio OMOP-CDM a partir de la información de un repositorio i2b2 sin pérdida de significado.

Este estudio representa un excelente ejemplo de la versatilidad perseguida por un Data Lake Sanitario y que también la podemos encontrar en la arquitectura propuesta por el proyecto Infobanco, con persistencia de datos en Open EHR, OMOP e i2b2, servidores de terminología, servidores de modelos de arquetipos, ETLs e inferencias, y salidas de integración en CDISC, ISO13606 y HL7 FHIR.

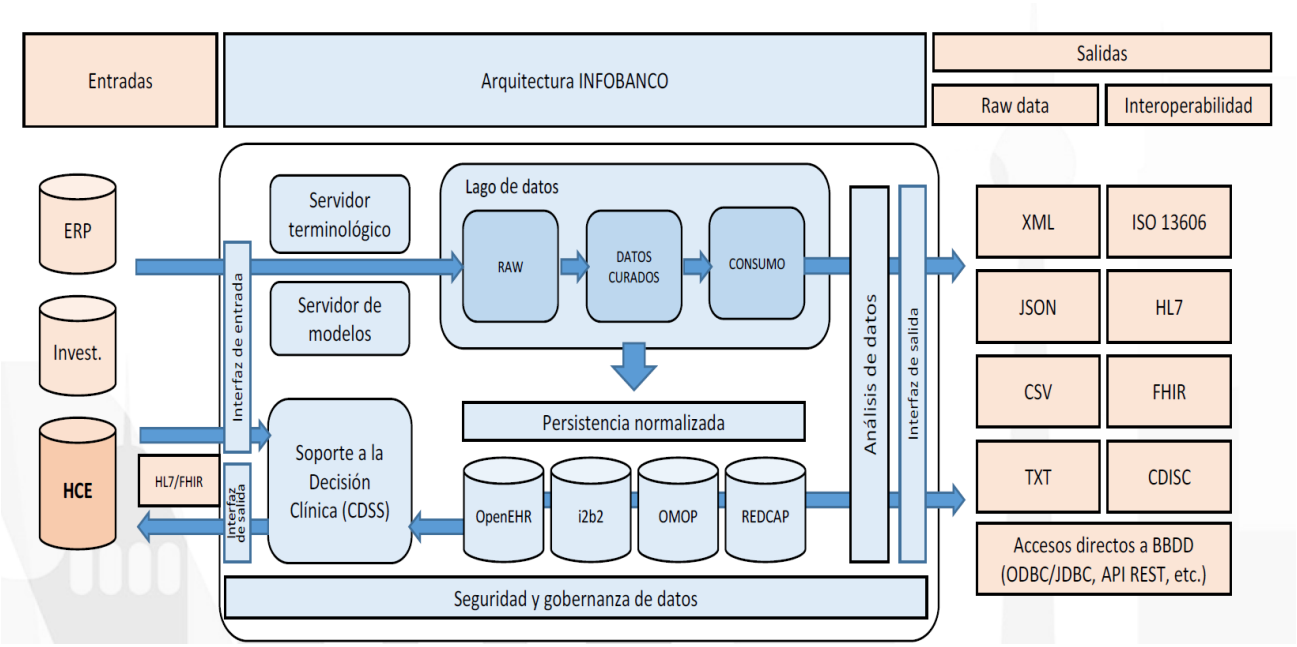


Ilustración 4.60. Arquitectura de Infobanco. Fuente i+12

Sí el mejor estándar es aquel que más se usa y tuviésemos que decantarnos en el instante de elaborar este TFM por uno, éste sería OMOP-CDM, creado en 2008 por la FDA estadounidense para el registro de reacciones adversas medicamentosas y que, a partir de 2014, es una iniciativa liderada por la Universidad de Columbia dentro de la comunidad Observational Health Data Sciences and Informatics (OHDSI), caracterizándose por ofrecer:

- Un Common Data Model (CDM) para uso secundario investigador de los datos y que probablemente contenga información de más pacientes en el mundo.

- Herramientas que permiten realizar control de calidad, detectando inconsistencias en las fuentes de datos, trabajar en red y obtener resultados de forma estandarizada y de gran utilidad para la gestión.

- Ayudar a crecer y evolucionar OMOP-CDM cuando éste no es capaz de cubrir el 100% de las necesidades de información de un proyecto, mediante el desarrollo de nuevas extensiones, codificaciones y elementos en nuevos ámbitos, como la imagen médica, la genética, etc.

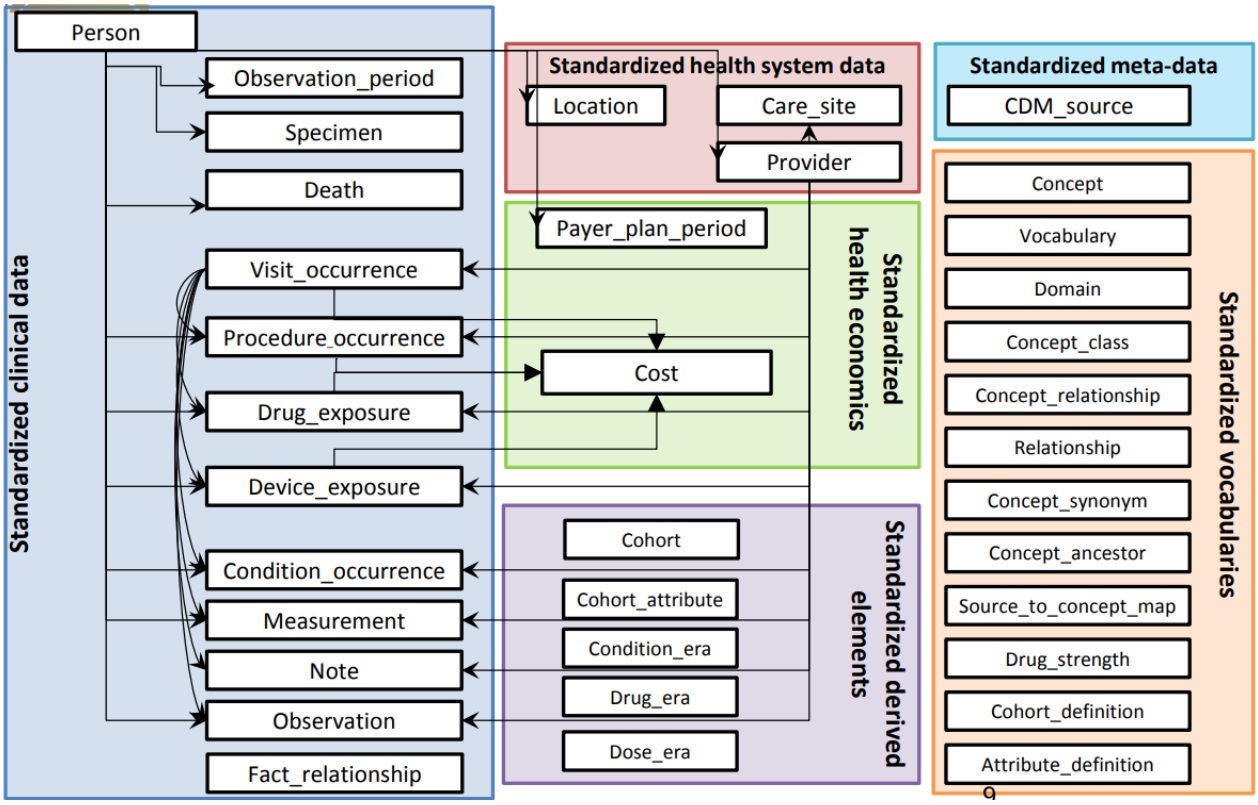


Ilustración 4.61. OMOP- Common Data Model versión 5.0.1. Fuente OHDSI

OMOP-CDM da preferencia a unos vocabularios estándar (RxNORM, ATC, SNOMED, LOINC), sobre otros en función de los dominios (mediciones, condiciones, …) y permite hacer mapeos entre estos vocabularios. También soporta como vocabularios clasificaciones, como es el caso de ICD / CIE, aunque es importante destacar que estas realizan agrupaciones de diagnósticos y procedimientos, algo de gran utilidad en estudios de evaluación de iso-consumos o listas de espera, pero que presentan enormes limitaciones cuando el ámbito de aplicación requiere mayor precisión, por ejemplo, para el soporte a la decisión clínica en diagnóstico y donde una terminología clínica, como SNOMED-CT, es más indicada por su mayor precisión.

La eficacia de OMOP-CDM quedó demostrada en el estudio acometido por el grupo de investigación en salud digital del Instituto de Salud Carlos III, y por medio del cual se concluye que un estudio de investigación trabajando sobre OMOP-CDM no pierde capacidad observacional, cuando es cargado con la información procedente de otros almacenes de datos observacionales, que trabajaban con codificaciones propietarias.

La réplica en OMOP de una gran cohorte es no inferior a la cohorte original, pudiendo localizarse todos los datos, extraerlos y replicar la misma evidencia que se había generado previamente, hipótesis que ha sido validada empíricamente mediante la traslación de tres trabajos de investigación de VIH de la Cohorte de la Red de Investigación en SIDA (CoRIS) al formato OMOP

y aunque el número de registros se vio triplicado, pasando de 4 a 16 millones, se confirmó que todos los estudios investigadores abordados en CoRIS se pudieron replicar en OMOP-CDSM con los mismos resultados investigadores.

Además de estos dos estudios donde se demuestra la versatilidad de OMOP-CDM para trasladar la información de estudios observacionales sin pérdida de eficacia investigadora, en este instante hay multitud de referencias soportadas por este estándar, destacando:

- Despliegue de la iniciativa EHDEN a nivel nacional y europeo.
- La valoración de OMOP-CDM por parte del grupo ciencia del dato de IMPACT.
- Trabajos previos de la EMA con OMOP y la elección del Centro Médico de la Universidad Erasmus de Rotterdam como coordinador de Darwin, probable primer nodo EHDS2.
- El trabajo realizado por la iniciativa Tartaglia o Chaimeleon para el desarrollo de extensiones OMOP que den soporte a los metadatos de la imagen médica y otros tipos de datos.
- Los proyectos de Hematology Outcomes Network in Europe (HONEUR) centrada en estudios de cáncer hematológico, la red ROADMAP para Alzheimer, HARMONY, BigData@Heart, PIONEER.
- El proyecto All of Us.
- La iniciativa OMOPonFHIR, que persigue la recuperación de datos almacenados en el esquema relacional de OMOP CDM como recursos FHIR.

Para la implantación de un proyecto de investigación en OMOP se requiere la ejecución de una serie de pasos, que se pueden resumir en:

- **Desplegar la Infraestructura** para soportar la instancia OMOP-CDM, que se materializa en una Base de Datos Relacional (SGBD) y como tal requiere la creación de su estructura.
- **Incorporar los vocabularios** desde el área de descarga de la herramienta **Athena** (https://athena.ohdsi.org/) a la instancia OMOP-CDM, para lo que resulta clave la elección de la terminología más adecuada para el ámbito de la investigación y además mantenerse al tanto de las actualizaciones de vocabularios, por si se incorporan conceptos necesarios y no soportados en las versiones anteriores.

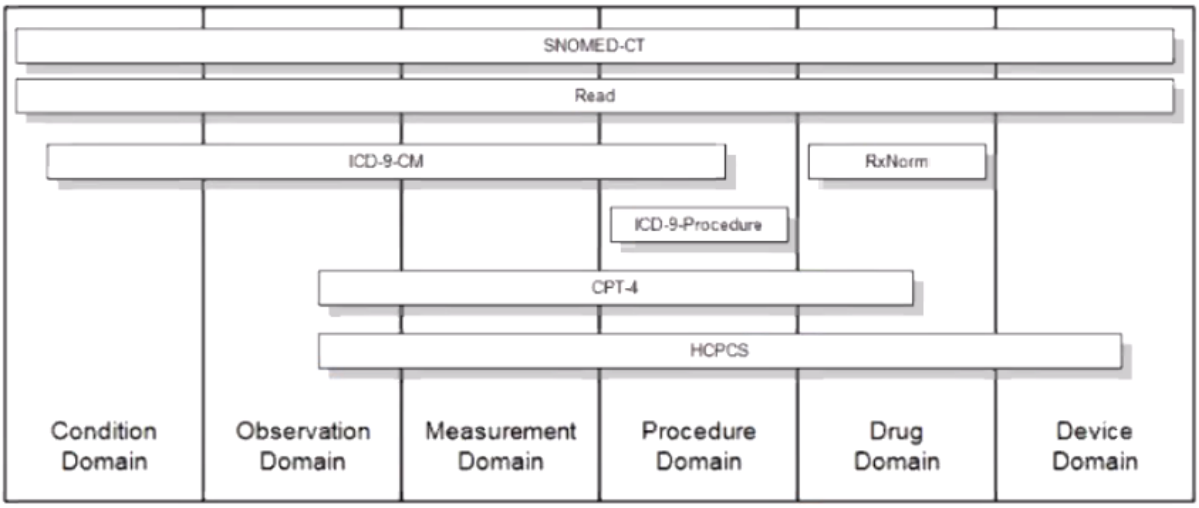


Ilustración 4.62. Vocabularios OMOP por dominio de conocimiento. Fuente OHDSI

- **Cargar la instancia** con los datos procedentes de las fuentes primarias, mediante "eventos" que se definen con tres componentes:
  - Quien genera, participa o recibe
  - Cuando se ha producido
  - A que se refiere el evento, mapeando un término del vocabulario estándar elegido con el hecho que se va a registrar.

- **Usar las herramientas OHDSI** (OHDSI, s.f.) para la explotación de resultados, como:
  - **Atlas,** ayuda a realizar estudios observacionales, inspeccionando instancias, conceptos, prevalencias, definir concept sets, análisis jerárquicos o definir cohortes, entre otros.
  - **Hades,** una colección de paquetes R de código abierto que ofrecen funciones para realizar un estudio de observación completo.
  - **Achilles**, paquete en R que permite la caracterización y visualización de una base de datos conforme al OMOP-CDM, avisando de que registros hay en la instancia OMOP que no están mapeados con la terminología descargada de Athena.
  - El modelo OMOP-CDM también puede ser interrogado utilizando herramientas convencionales como SQL, R, Phyton, Stata, etc.

Concluimos que un Data Lake Sanitario, debe ser diseñado preferentemente, como una solución que permita hacer un uso agnóstico de los modelos y la versatilidad suficiente para tener múltiples salidas de datos, aunque tras el análisis realizado, sí tuviésemos que optar por un único modelo para la explotación secundaria y masiva de datos con fines investigadores y/o la mejora de los resultados en salud, este sería **OMOP-CDM,** tanto por sus funcionalidades y nivel de desarrollo actual, como por su previsible crecimiento y capacidad de extensión.

Los trabajos acometidos para la normalización de los datos pueden ser de ayuda para identificar problemas de calidad de los datos en origen, pudiendo derivar en el desarrollo de guías, pautas, nuevos interfaces de captura de datos, más usables y normalizados, para mejorar el registro de la información durante la asistencia y con ello la calidad de la información.

4.5.4. Gobernanza de un Data Lake Sanitario

Las herramientas y software propias de este ámbito son muchas y variadas, y se abordarán con más detalle en otro capítulo de este TFM, a continuación, se analizarán aquellos factores que deben ser tenidos en consideración desde el punto de vista del gobierno de un Data Lake Sanitario.

4.5.4.1. Arquitectura y Aprovisionamiento

El análisis avanzado y/o el desarrollo de modelos de inteligencia artificial a partir de grandes volúmenes de datos, especialmente cuando se trabaja con imágenes de radiodiagnóstico, genética, anatomía patológica, lenguaje natural, etc., requieren la utilización de herramientas altamente específicas y gran escalabilidad de almacenamiento y procesamiento (CPUs/GPUs) soportando grandes picos de demanda de recursos durante ventanas de tiempo muy reducidas, lo que se traduce en la necesidad de provisionar y liberar estos recursos de forma prácticamente inmediata, algo inherente a los modelos de operación como servicio (PAAS, IAAS y SAAS) típicos de entornos cloud y donde se incurrirá en gastos, únicamente cuando se haga utilización de los recursos.

Esta forma de trabajar podría representar un escenario de difícil retorno de la inversión cuando se intenta provisionar con modelos tradicionales, on-premise, ya que requieren sobredimensionar la infraestructura para dar soporte a cargas de trabajo puntuales, incurriendo en grandes costes por CAPEX y OPEX, aun cuando el hardware y el software vaya a estar infrautilizado el resto del tiempo.

Aunque tradicionalmente el consumo de recursos a demanda y en la nube ha sido recibido con fuertes resistencias para su adopción por parte de las Administraciones Públicas, ya que no permite el control de la totalidad de las infraestructuras y soluciones, este TFM apuesta por la implantación de un Data Lake Sanitario en el cloud, con modelos de consumo de hardware, software y herramientas a demanda y de acuerdo a las siguientes recomendaciones de diseño, algunas de las cuales han sido extraídas del Pliego de Prescripciones Técnicas del proyecto Infobanco contratado por el i+12:

- **Operación:**

  - Se trabajará con proveedores de cloud que cuenten con soluciones para la gestión del ciclo de vida completo de los datos.

  - Debería contar con tres entornos independientes: Producción, Pre producción y Desarrollo y de un módulo específico para la monitorización y administración del funcionamiento y rendimiento de todos los elementos de los tres entornos.

  - PASCAL debería disfrutar de la funcionalidad de multi-tenants, para segmentar los datos de acuerdo a los diferentes casos de uso, estructuras organizativas, entidades, visualizaciones y conjuntos de datos, proporcionando gestión de recursos y seguridad a través de AAA para cada uno de los tenants configurados, con un control del acceso a datos (aplicación de reglas personalizadas de seguridad y negocio en el acceso a datos) y en espacios de trabajo: (segmentación relativa a la agrupación de contenidos, flujos de trabajo, modelos analíticos y visualizaciones con una gestión personalizada de seguridad).

  - Las diferentes soluciones deberían ser implementadas, en una arquitectura basada en microservicios y contenedores, como Docker[16], que maximice la escalabilidad y elasticidad facilitando el traspaso entre clouds (e incluso a un entorno “on premise”, si así se decide), proveyéndose las herramientas de orquestación necesarias, como Kubernetes[17].

  - La arquitectura de PASCAL debería facilitar configuraciones redundantes de alta disponibilidad tanto en software como en datos, que permitan la continuidad de las operaciones ante incidentes o desastres. Es conveniente definir tiempos de respuesta y de resolución a incidencias, disponibilidad de PASCAL y de recuperación ante incidentes (RPO y RTO) de acuerdo a la criticidad de PASCAL, y que debería ser inferior a la de los sistemas de información que soportan la operativa asistencial, lo que se debería traducir en una reducción de costes en este aspecto.

  - Dado que la persistencia del dato será realizada en el cloud, se definirá unos requisitos de ancho de banda para soportar los procesos ETL incrementales diarios, de forma que se adecuen a las ventanas de tiempo existentes.

16 Docker es un proyecto de código abierto que automatiza el despliegue de aplicaciones dentro de contenedores de software, proporcionando una capa adicional de abstracción y automatización de virtualización de aplicaciones en múltiples sistemas operativos

17 Kubernetes es un sistema de código libre para la automatización del despliegue, ajuste de escala y manejo de aplicaciones en contenedores, que fue originalmente diseñado por Google y donado a la Cloud Native Computing Foundation y que soporta diferentes entornos para la ejecución de contenedores, incluido Docker

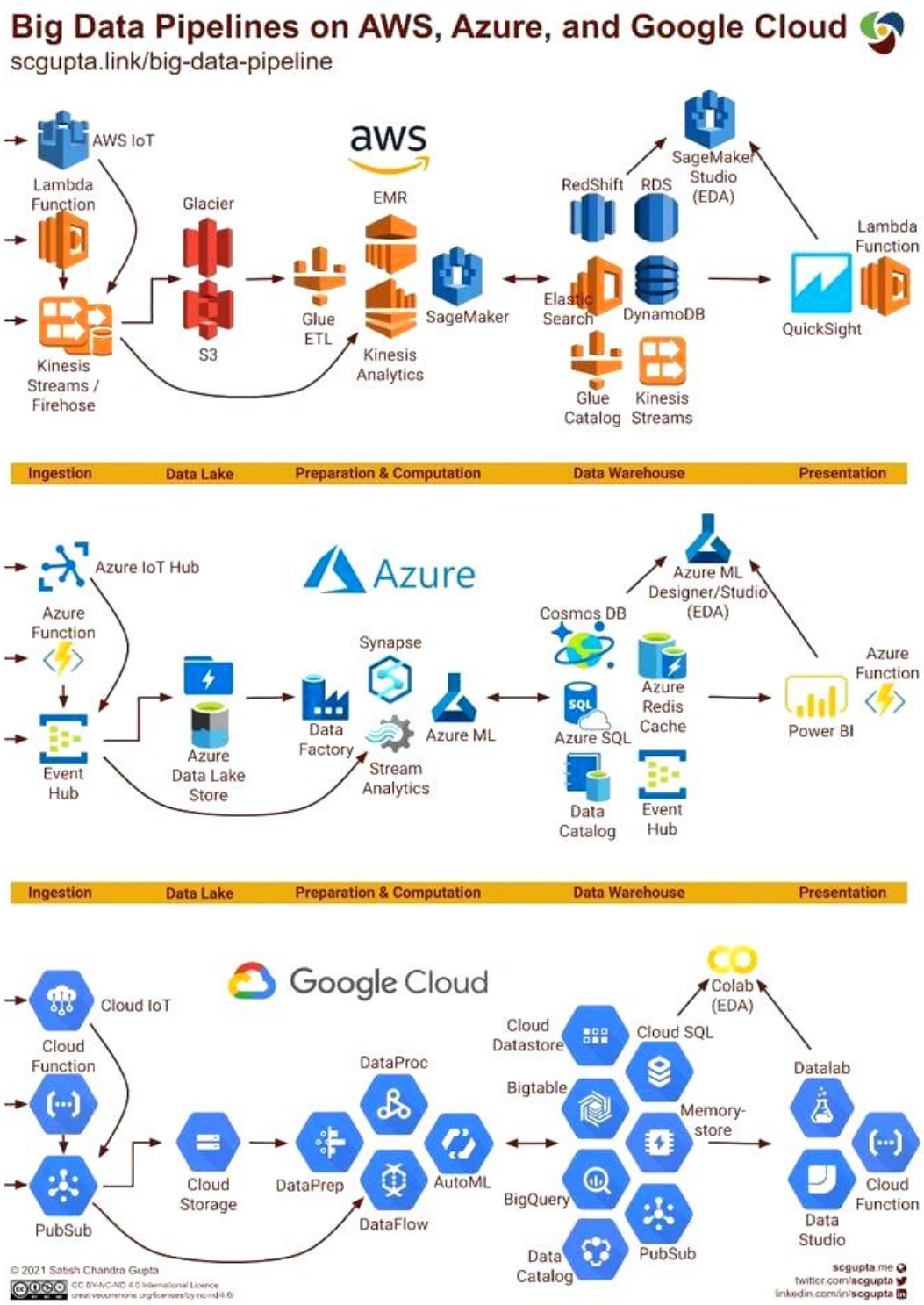


Ilustración 4.63. Comparativa de soluciones / herramientas en las tres "Big-Techs"

- **Seguridad:**

  - Los proyectos deberán incorporar desde el diseño el cumplimiento de medidas del Esquema Nacional de Seguridad (RD 3/2010), con especial hincapié en lo que respecta a la utilización de servicios en la nube y a la identificación de activos, su valoración de acuerdo a las cinco dimensiones que tiene en cuenta el ENS (disponibilidad, confidencialidad, integridad, autenticidad y trazabilidad), el análisis de riesgo y el establecimiento de medidas de protección de datos (RGPD y LOPDGDD).

  - Se garantizará la seguridad de los componentes software empleados y del resultado de los trabajos, mediante los oportunos mecanismos de homologación o certificación, también de los algoritmos de Inteligencia Artificial como producto sanitario, cuando así sea requerido.

- **Gestión:**

  - Todos los componentes, productos o servicios de la arquitectura se integrarán bajo un modelo de gobierno organizativo y tecnológico centralizado, contando con herramientas para la gestión de las infraestructuras y de los datos, que simplifiquen el cumplimiento de los objetivos de rendimiento, seguridad, calidad y resultado para cada proyecto.

  - Se recurrirá a servicios de asesoramiento para el diseño y despliegue del entorno cloud, asegurando que las aplicaciones que corran en la nube lo hagan con las máximas garantías de seguridad, rendimiento, a la vez que se optimizan los costes de operación, mantenimiento y gobierno de la plataforma.

  - Con el objeto de favorecer su reutilización y cesión, las soluciones software deberían ser de preferentemente de fuentes abiertas y disfrutar de soporte y una amplia adopción en el mercado.

  - PASCAL debería contar con actualizaciones para todas las herramientas y para el software de base y se debería garantizar la vigencia y soporte de todos los componentes durante un plazo mínimo de tres años.

  - Los proyectos acometidos en PASCAL deberían permitir la estandarización de los datos biomédicos y facilitar su interoperabilidad, acceso y gobernanza, para lo que es recomendable el uso de software con licencias de código abierto, cuyas fuentes estén en repositorios de control de versiones, como GitLab y GitHub.

  - En el caso de que se desarrolle software científico, idealmente se deberían seguir las recomendaciones de ELIXIR.

  - Los datos (metadatos) se deberían compartir bajo principios FAIR, lo que facilitará su localización, acceso y uso, para lo que los conjuntos de datos deberían estar descritos correctamente, incluyendo las taxonomías utilizadas y sus restricciones de uso.

- La tecnología y los estándares incorporados en PASCAL, deberían ser compatibles y estar alineados con muchas de las iniciativas de referencia, y al menos con:
    - European Health Data Space (EHDS),
    - Espacio Nacional de Datos de Salud (ENDS)
    - Infraestructura Medicina de Precisión Asociada a la Ciencia y Tecnología (IMPACT)
    - European Health Data & Evidence Network (EHDEN)
    - European Open Science Cloud for Health (EOSC-Health)
    - European Life-science Infrastructure for Biological Information (ELIXIR)
    - Global Alliance for Genomics and Health (GA4GH)
    - 1 Million Genomes (1+MG)
    - European Genome phenome Archive (EGA)

- Aunque comercialmente sean denominados como servicios, los contratos para su adquisición serán de suministros[18] y de esta forma aplicar, de forma más sencilla, como conceptos elegibles en convocatorias financiadas con fondos europeos, cuando así sea requerido.

- También con el objetivo de la financiación procedente de los fondos europeos, PASCAL debería velar por el cumplimiento de los principios Do No Significant Harm, DNSH, de su infraestructura, trabajando de forma preferente, con proveedores de cloud que sean capaces de ejecutar sus operaciones con el mayor porcentaje de energía generada libre de emisiones de CO_2.

4.5.4.2. Profesionales

Al igual que en los sistemas de información tradicionales, un Data Lake Sanitario requiere de la ejecución de multitud de tareas y servicios que puede ser acometidos con personal público o recurrir a su externalización.

En cualquiera de los dos casos, se requiere de una combinación de conocimientos propios de las matemáticas y la estadística, de las ciencias de la computación y, en nuestro caso, de las Ciencias de la Salud, que configuran un amplio dominio de conocimiento conocido como **Ciencia de Datos o Data Science.**

18 La calificación como contratos de suministros está respaldada por el informe correspondiente al expediente 13/2021, de Calificación jurídica de los contratos de prestación de servicios en la nube emitido por la Junta Consultiva de Contratación Pública del Estado del Ministerio de Hacienda, en el que se concluye que los contratos por los que las entidades públicas adquieren el derecho de uso de activos de software en la nube son contratos de suministro conforme a lo dispuesto en el artículo 16.3 b) de la LCSP, a menos que se trate del desarrollo de programas de ordenador a medida, que serán contratos de servicios.

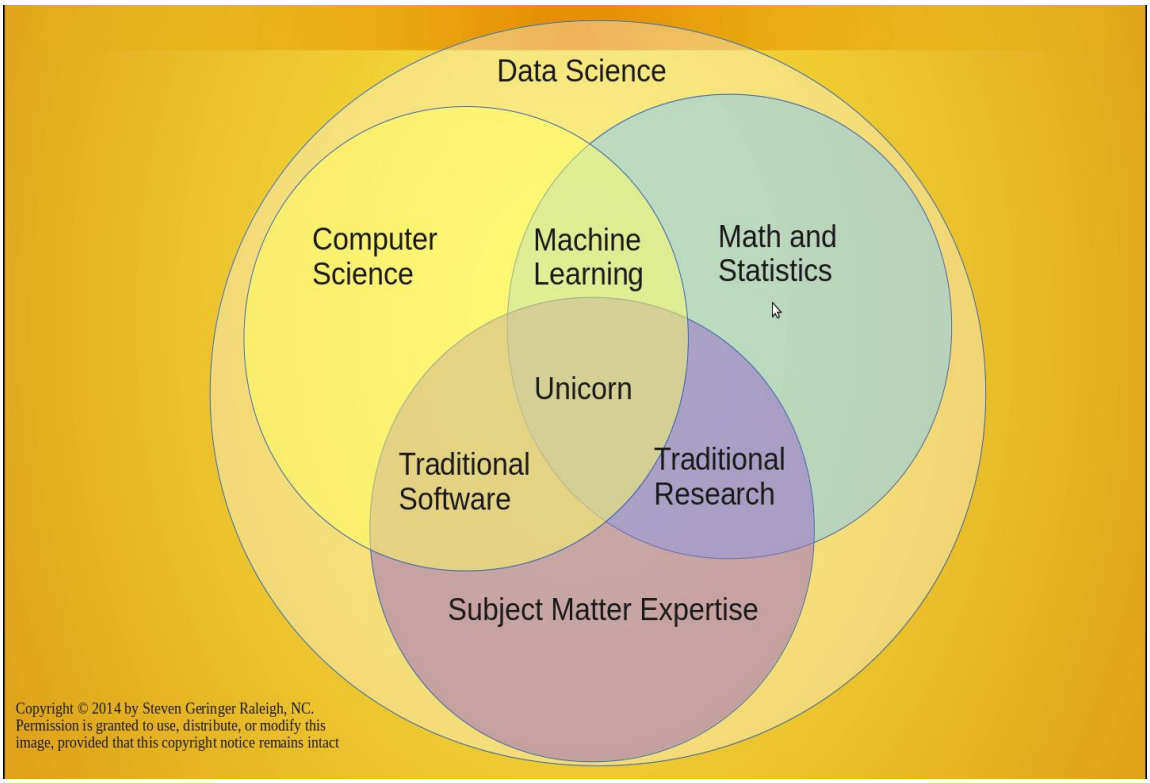


Ilustración 4.64. Data Science Venn Diagram

Por ello, cuando se acomete un proyecto de implementación y operación de un Data Lake Sanitario, se requieren **equipos multidisciplinares** en los que necesariamente deben estar presentes personal clínico-asistenciales, siendo altamente recomendable, identificar a aquellos profesionales que puedan actuar como **líderes digitales** ante el resto de la organización, ayudando a visualizar el éxito derivado de estas iniciativas y con ello impulsando la **creación de una cultura en torno al dato**.

Las organizaciones sanitarias cuentan con profesionales de experiencia contrastada en la operación de sistemas transaccionales, para registro y consulta de información asistencial y que han desarrollado su trabajo con una capacitación muy diferente a la requerida en un Data Lake Sanitario.

Por ello resulta necesario el desarrollo de planes formativos para la actualización de conocimientos en fundamentos de matemáticas, estadística, programación, aprendizaje automático, minería de texto, procesamiento de lenguaje natural, visualización de datos, tratamiento de imágenes, ingesta, gestión de grandes volúmenes de datos y de las herramientas específicas para dar soporte a estas tareas o proceder a su delegación en un prestados externo.

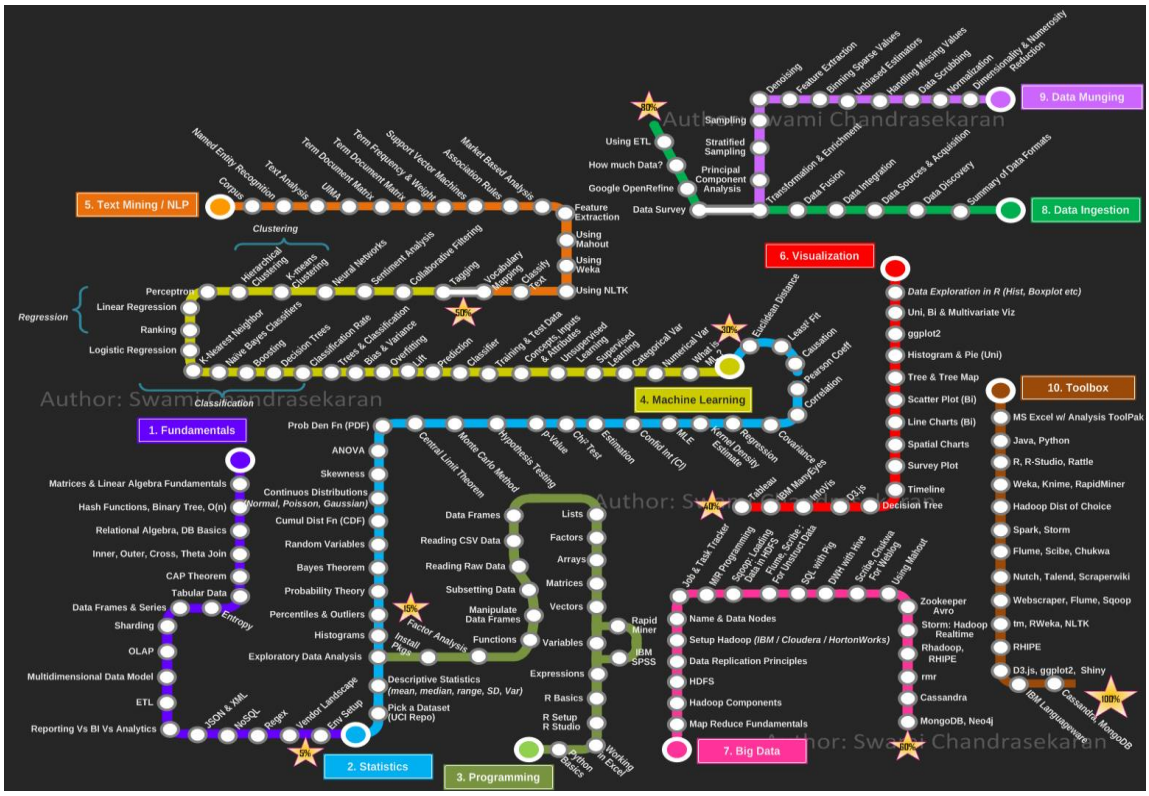


Ilustración 4.65. Infografía. Data Science Roadmap. Swami Chandrasekaran

Conocimientos o competencias que en función del tamaño y especificidad requeridos pueden distribuirse en diferentes perfiles profesionales:

1. **Analista de datos**, tareas de visualización, manipulación y procesamiento de grandes cantidades de datos.

2. **Ingenieros de datos**, para ingestar datos crear y probar ecosistemas de big data escalables antes de ejecutar un modelo de ciencia de datos.

3. **Administrador de base de datos**, responsables de la administración, copias de seguridad y recuperaciones de la base de datos.

4. **Científico de datos**, labores para encontrar, limpiar y organizar los datos, usando análisis y procesamiento de datos y comprendiendo los problemas asistenciales y realizando recomendaciones y presentación de resultados.

5. **Ingeniero de aprendizaje automático**, con conocimientos matemáticos y habilidades de

programación, gestionan data funnels y ofrecen soluciones de software.

6. **Científico de aprendizaje automático**, investigan nuevos enfoques de datos y algoritmos para sistemas adaptativos.

7. **Arquitecto de datos**, crean el diseño del plan para la gestión de datos con las mejores medidas de seguridad.

8. **Bio-estadístico**, organizan los datos con el fin de identificar tendencias y relaciones y ofrecer información valiosa.

9. **Analista clínico-asistencial**, para vincular el Big Data y la ciencia del dato con los conocimientos de los profesionales clínico-asistenciales y proveen soporte a la visualización de datos y la presentación de resultados.

10. **Responsable de equipo**, coordina tareas de ciencia de datos y la asignación de funciones al equipo de acuerdo con las habilidades, la experiencia y aplica técnicas de desarrollo agile adecuadas a los proyectos de un Data Lake Sanitario.

11. **Data Manager**, son perfiles más habituales en los ensayos clínicos que en ciencia del dato, encargados habitualmente de velar por el registro y calidad de los datos asociados a un estudio particular que se trabaja sobre software de gestión de ensayos clínicos tipo redCAP.

4.5.4.3. Control y Seguimiento

Para garantizar la independencia de la labor de auditoria de accesos y actuaciones realizadas por cada investigador/profesional sobre el set de datos de trabajo autorizado, una buena práctica es su gestión por parte de una **Comisión de Control y Seguimiento** del Data Lake Sanitario, que debería estar integrada por miembros pertenecientes a todas las instituciones donde se generen y consuman los datos y que, como mínimo, debería de asumir las siguientes funciones:

1. Planificar la actividad del Data Lake Sanitario.

2. Proponer, valorar y autorizar los sistemas de información y datos ingestados.

3. Valorar y autorizar los servicios que se presten para o en colaboración con terceras entidades públicas y privadas.

4. Aprobar las normas internas de organización y prestación de servicios del Data Lake.

5. Decidir sobre la política de accesos a la plataforma.

6. Supervisar la gestión segura de los datos y la plataforma.

7. Analizar el impacto del resultado de los trabajos en el uso asistencial o informativo de dichas aplicaciones, especialmente las dirigidas a los usuarios y pacientes.

8. Velar por la adecuación de los procesos de gestión y calidad de los datos y garantizar el cumplimiento de las normas éticas y legales.

9. Establecer un reglamento interno de actuación para la propia Comisión.

10. Identificar a los grupos de interés y determinar su participación

11. Aprobar los estándares utilizados.

12. Participar en la definición y evaluación de la estrategia del Data Lake.

4.6. Conclusiones del Capítulo I

Conclusión 1. Fragmentación

El análisis actual de iniciativas europeas, nacionales y regionales refleja que nos encontramos ante un conjunto de actuaciones inconexas y heterogéneas, que representan un importante desafío para su convergencia tecnológica, legal y organizativa, exponiendo a continuación las conclusiones de mayor relevancia:

1. La identificación de tres aproximaciones con diferentes enfoques, conocimientos, liderazgos e iniciativas:

Informática sanitaria	Bio-informática	Economía
Habituados al uso primario de los datos.	Habituados al uso secundario de los datos	No habituados al uso de los datos.
Usan las TICs para soportar la HCE con fines asistenciales	Usan las TICs y las matemáticas para soportar la I+D+i en bio-ciencias	TICs para impulso de la economía del dato y la I.A. con colaboración público-privada
Dirección General de Health & Food Safety (DG SANTE)	Dirección General for Research and Innovation (DG RTD)	Dir. General Communications Networks, Content and Technologies. (DG CONECT)
Ministerio de Sanidad,	Ministerio de Ciencia e Innovación	Ministerio de Asuntos Económicos y Transf. Digital
Servicios Regionales de Salud	Entidades dedicadas a la I+D+i	Entidades privadas
EHDS ENDS	HRIC, EOSC-Life, 1+MG IMPACT	Gaia-X, ENIA
EHDEN		

Ilustración 4.66. Clasificación de iniciativas

2. A pesar de que el desarrollo de la medicina 5Ps, requiere que los datos óhmicos se integren en el ámbito de la atención a la salud, se perciben una gran separación, a todos los niveles, entre las iniciativas centradas en la genética y el resto de actuaciones.

3. Se han identificado multitud de iniciativas de ámbito regional e incluso menor, con implantación de Data Lake Sanitarios y desarrollo de casos de uso similares, con riesgo de replicar escenarios de ineficiencia similares a los vividos con las HCE y que podrían llevar, por ejemplo, a desarrollar decenas de algoritmos públicos dotados de inteligencia artificial para solucionar un problema con idénticas requisitos en todas las CCAA.

Conclusión 2. Objetivos Comunes

Se han identificado objetivos coincidentes en la gran mayoría de las iniciativas, lo que debería favorecerá la búsqueda de puntos de encuentro y la compartición de datos en espacios globales para contar con los niveles de interoperabilidad organizativa, legal y técnica necesarios en:

- Desarrollo de la Medicina 5Ps
- Generación de nuevo conocimiento (**I+D**+i) a partir de los datos
- Impulso de una nueva economía (I+D+**i**) en torno al dato y la IA
- Mejora de la atención y VBHC mediante unas políticas de gestión centradas en el dato.

Conclusión 3. Eres tan fuerte como la calidad de tus datos

Se identifica la calidad de los datos en las fuentes de origen, como un factor crítico para la viabilidad de cualquier proyecto de analítica de datos.

Aunque existe una concienciación sobre este riesgo, que se materializa en la recomendación de destinar aproximadamente un 60% del tiempo del trabajo al análisis y depuración, no se dispone de un análisis objetivo y transversal, que permita dimensionar, cualitativa y cuantitativamente, el estado de la situación actual respecto de la calidad de los datos en el ámbito regional, nacional o europeo.

Conclusión 4. Mejor modelo de Gobernanza

Este TFM concluye que el modelo que ofrece más garantías legales, mejor desempeño en su operativa y que tiene mayor capacidad para maximizar los resultados obtenidos, esta soportado por una propuesta de Gobierno que separa el uso primario y el uso secundario de los datos, en dos entidades jurídicas diferentes e independientes:

- La primera, competente en materia de prestación asistencial encargada de compartir los datos que proceden del uso primario
- La segunda, competente en materia de I+d+i, encargada de ingestar, consolidar, segmentar y dar soporte a los diferentes estudios o usos secundarios de los datos que se acometan en el Data Lake Sanitario.

Este modelo de colaboración es el definido por CCAA como Aragón y La Rioja, y que presenta, entre otros, los siguientes beneficios:

- Salvaguarda más garantista de la privacidad al establecer una doble seudonimización, que no puede ser revertida mediante la actuación de un único profesional o de una única entidad.

- Se habilita la colaboración público-privada porque la entidad encargada del uso secundario dispone de mayor flexibilidad jurídica.

- La entidad encargada de gestionar el Data Lake Sanitario y de los usos secundarios de los datos, dispone de bio-informáticos y perfiles de investigación habituados a trabajar en el ámbito I+D+i, lo que se traducirá en una mayor eficiencia en las tareas y una mayor eficacia en la generación y protección de los resultados.

- Se favorece el cumplimiento de principios FAIR, habituales del ámbito de la I+D+i.

- Se favorecerá la colaboración entre entidades públicas porque están más habituados a estos modelos de trabajo, además de formar parte de redes de investigación consolidadas.

Pero este modelo de operación no solo debe ser implantado técnicamente, requiere de respaldos normativos, como los desarrollados por La Rioja y Aragón para la definición y separación de funciones entre entidades, o por Andalucía para regular el acceso a los datos con fines investigadores.

Adicionalmente y para evitar que la separación planteada, se perciba como una pérdida de visión sobre una iniciativa de gran visibilidad y carácter estratégico, este TFM plantea la constitución de una Comisión de Control y Seguimiento del Data Lake Sanitario integrada por miembros procedentes de las dos entidades.

El análisis de los datos para un uso secundario, es decir, con un fin diferente para el que fueron recabados, requiere el cumplimiento de una serie de requisitos legales por parte de todas las instituciones, también de las entidades que realizan el registro de los datos durante la asistencia. No debería abordarse un Data Lake Sanitario con datos incrementales, individuales y longitudinales de pacientes con fines investigadores, como si fuesen proyectos de agregación de datos para la mejora de la gestión, típicos Cuadro de Mando Integral[19], sin tener en consideración los elementos en materia de gobernanza organizativa, legal y ética analizados en este TFM.

19 Aunque estos también sean proyectos de consolidación y explotación de datos, incumplen la disposición adicional 17 de la LOPDGDD, la única habilitante para la exportación de datos incrementales de pacientes individuales con fines diferentes a los asistenciales, y que obliga a realizar, entre otros, una separación técnica entre quien seudonimiza los datos y quien los analiza, la revisión por parte de un comité de ética investigadora de las actuaciones y la validación del modelo mediante una EIPD.

Conclusión 5. Gobernanza de un Data Lake Sanitario

Dado el formato de consumo de recursos de la e analítica avanzada y el desarrollo de modelos de inteligencia artificial, este TFM ha concluido una serie de requisitos de diseño, siendo los más relevantes:

1. **Consolidación:** de la información en un único repositorio, con las herramientas y servicios necesarios para abordar todos los estudios de carácter secundario y que se pueden clasificar en cuatro categorías (BI, RWE/RWD, CRF y AI), siendo el investigador quien va al dato, el dato no viaja al investigador.

2. **Segmentación:** de los datos existentes en el Data Lake Sanitario en diferentes subconjuntos de datos, data-buckets o data sets, para trabajar de forma independiente, segura (AAA) y con trazabilidad cada uno de los casos de uso, conjuntos de datos, visualizaciones, entidades o espacios de almacenamiento.

3. **Adaptabilidad:** a un consumo dispar de recursos[20], con grandes demandas de forma muy puntual, seguidos de largos periodos de infra-utilización, lo que nos lleva a recomendar la utilización de soluciones que permitan flexibilizar el consumo de la tecnología, mediante su provisión como IaaS, PaaS y SaaS y en entornos cloud.

4. **Portabilidad:** mediante la implementación de las soluciones sobre una arquitectura basada en microservicios e implementada en contenedores, que maximice su escalabilidad y flexibilidad, permitiendo el traspaso entre diferentes entornos clouds, para evitar el vendor locker.

Conclusión 6. Aprendizaje federado en red

El análisis realizado concluye que la colaboración en red, se debe soportar por modelos de aprendizaje federado dotadas de privacidad por diseño, una arquitectura donde los datos permanecen en sus repositorios de origen, siendo las consultas y algoritmos los que viajan hasta los datos.

Consideramos que iniciativas, como EHDS o ENDS, que deben dar respuesta a diferentes regiones y países, también optaran por modelos federados, evitando la centralización de los datos, aunque quizá no, de las infraestructuras que soportan cada uno de los nodos del modelo federado, pudiendo usar el cloud y el multi-tenant para habilitar administraciones independientes.

20 Especialmente en proyectos federados que trabajan con imagen médica, textos libres, o genética, y donde se procesan grandes volúmenes de datos pudiendo requerir la provisión de enormes recursos en momentos muy puntuales, algo típicamente incompatible con la operación on-premise.

Conclusión 7. El mejor estándar el que se usa

Justificada la conveniencia de apoyarse en estándares para impulsar la investigación, o incluso para habilitar la mejora en la gestión mediante la comparación y tras el análisis de las diferentes alternativas, se concluye que se debe hacer un uso agnóstico de los estándares y avanzar de forma progresiva en su evolución y mapeo.

Aunque se ha evidenciado la flexibilidad y el valor aportados por los modelos duales para representación del conocimiento mediante ontologías, arquetipos e inferencias, este TFM también ha identificado aquellos estándares más relevantes por su extensión y versatilidad:

- Modelo de Datos:
    - Uso secundario de los datos:
        - **OMOP-CDM.**
        - Intercambio (futuro): CDISC versus OMOPonFHIR
    - Uso primario de los datos:
        - Intercambio: HL7 FHIR
        - Resto: Open EHR

- Vocabularios:
    - Terminológico: **SNOMED-CT**
    - Clasificación: ICD-10

Conclusión 8. Conocimiento y Cultura

Para poder hacer frente a proyectos complejos en Ciencia del Dato, es totalmente necesario la configuración de equipos multidisciplinares, con presencia de matemáticos, informáticos y personal clínico-asistencial.

La complejidad de un proyecto investigador para el descubrimiento de nuevo conocimiento o para el desarrollo de modelos de soporte a la decisión clínica, aplicando técnicas de analítica avanzada y explotación masiva de datos puede exceder, con mucho, la capacitación típica de los profesionales de informática sanitaria, teniendo que acometer importantes actuaciones en materia de formación y/o recurrir a la externalización de servicios especializados.

Para la creación de una cultura en torno al dato en las organizaciones sanitarias, resulta clave la identificación de personal asistencial motivado y con experiencia en proyectos de Ciencia del Dato que quieran actuar ante el resto de la organización como líderes digitales, comunicando y proyectando una visión del éxito que se puede conseguir.

Conclusión 9. Un nuevo Producto Sanitario

Este estudio ha intentado destacar las bondades y dificultades para el desarrollo de modelos de Inteligencia Artificial, concluyendo qué son un elemento clave para la transformación del Sistema Nacional de Salud siempre que:

- Los profesionales sanitarios perciban su adopción como la de cualquier otro producto sanitario, de forma que quien haga uso de la IA realizará mejor su trabajo.

- Si la IA es adoptada de forma masiva por parte de los pacientes, antes de que las instituciones sanitarias la incorporen a sus procesos asistenciales, se pueden generar importantes tensiones en la acción de prescripción.

Conclusión 10. Diccionario

Durante la elaboración de este TFM se ha identificado una enorme variabilidad en la terminología utilizada, en un ámbito, tan novedoso como carente de precisión, por lo que podría ser interesante el desarrollo de un diccionario por parte de la propia SEIS, de forma similar a lo realizado en el Glosario de la iniciativa THEDAS (THEDAS, 2021).

Conclusión 11. Transparencia

Recurriendo a las buenas prácticas impulsadas por los principios FAIR y la iniciativa FAIR4Health, este TFM segmenta el ciclo de trabajo de los datos, en dos fases, la de preparación y la de producción, separadas por un proceso intermedio para el cumplimiento de la gobernanza en materia de transparencia, de obligado cumplimiento en todas las iniciativas abordadas con presupuesto público y por el interés general:

1. **Preparación de los datos**
    a. Planificación y recolección
    b. Adecuación y Publicación
2. **Consulta, solicitud de acceso y aprobación**
3. **Producción con los datos**
    a. Uso de los datos
    b. Gestión de resultados

Conclusión 12. ¿Autorización?

Esta conclusión se ha dejado intencionadamente para el último lugar, puesto que en el instante de elaboración de este TFM no se puede saber de qué forma se van a materializar las diferentes regulaciones europeas en desarrollo.

Se percibe un riesgo inherente a que las diversas interpretaciones realizadas del RGPD, del Data Governance Act y del European Health Data Space por parte de los diferentes países de la UE, pudieran acaben otorgando a los ciudadanos el derecho a autorizar, de forma previa y expresa, la incorporación de su información a los Espacios de Datos de Salud para cada estudio o proyecto investigador, lo que podría traducirse en un freno para el desarrollo de este ámbito, frente a la versatilidad existente en la actualidad en España gracias a la disposición adicional decimoséptima de la LOPDGDD.

5. Capítulo II. Prevención Secundaria Riesgo Cardio-vascular en síndrome coronario

Objetivo del caso de uso: Aplicación de un data lake sanitario para el desarrollo de herramienta de estratificación del riesgo cardiovascular de los pacientes con síndrome coronario para su posterior aplicación en todas las fases de evaluación, seguimiento y monitorización del riesgo de los pacientes en el proceso asistencial

Esta herramienta aportará la capacidad de mejorar el seguimiento y resultados en salud al programa de prevención secundaria del riesgo cardiovascular en esta población de pacientes, y posteriormente el data lake nos permitirá el análisis del impacto de las mejoras de procesos con la incorporación de esta y otras herramientas digitales en la prevención secundaria de riesgo.

La necesidad del desarrollo del caso de uso y el marco de aplicación de la herramienta de estratificación de riesgo, está basado en las prioridades de la estrategia de salud cardiovascular del sistema nacional de salud, el análisis de las recomendaciones de la sociedad española de cardiología en evaluación del riesgo de los pacientes durante todo el proceso asistencial del síndrome coronario y las recomendaciones de automatización y digitalización de estos procesos en base a varios informes publicados, como el proyecto Anjana de la SEIS y nuestras experiencias profesionales en estos ámbitos.

5.1. Antecedentes ECV, síndrome coronario y gestión del proceso asistencial

5.1.1. ECV y estrategia de salud cardiovascular del SNS

- La enfermedad cardiovascular es la principal causa de muerte en nuestro país; con el síndrome coronario representando una de las patologías más comunes y de mayor riesgo.

- En el año 2019, la prevalencia de enfermedades cardiovasculares[1] (ECV) en España afectaba al 9,8% de la población, 52,6% mujeres y 47,4% hombres (1). La incidencia anual fue de 1 nuevo caso cada 100 personas. Para el mismo año, las ECV fueron la causa de defunción del 37,4% de la población de la Unión Europea, lo que supuso más de 2 millones de muertes. En España, constituyeron la primera causa de muerte, con un 27,9% del total, siendo los tumores (27,0%) y las enfermedades respiratorias (11,4%) la segunda y tercera causa respectivamente, así como la primera de ingreso hospitalario (Figura 3).

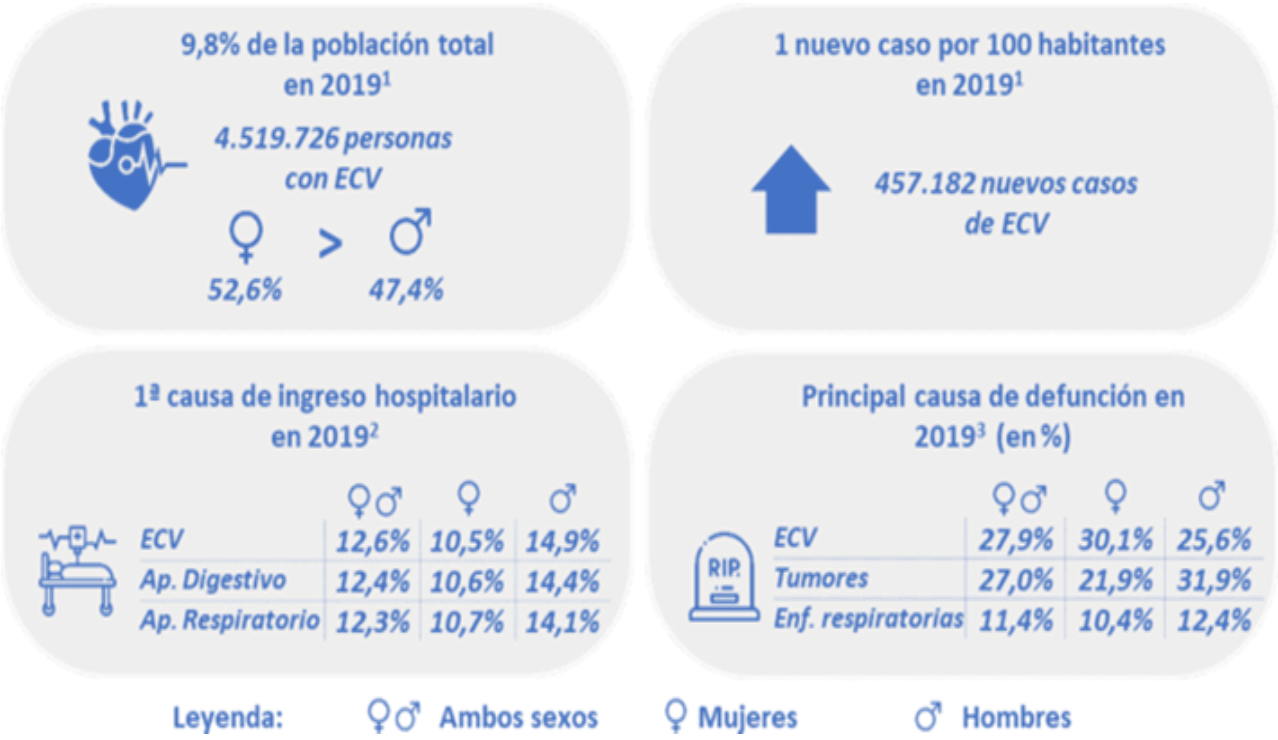
- El desarrollo de ECV y la aparición de eventos cardiovasculares se ve condicionada por los diversos determinantes individuales y sociales de la salud.

- Los asociados a recursos materiales, tales como un bajo nivel socioeconómico, que es un predictor de mayor riesgo de ECV o factores del entorno, como la influencia de la calidad

del aire sobre el riesgo de mortalidad de ECV

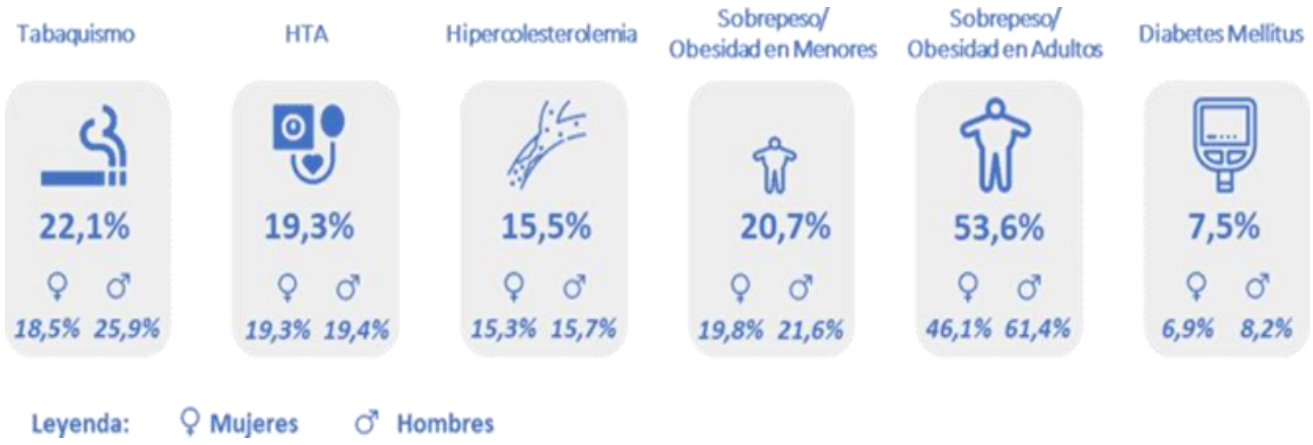
- Los factores de riesgo asociados a estilos de vida, como el tabaquismo, que es la principal causa de morbimortalidad cardiovascular (Figura 4)

- Los factores de riesgo biológicos, ya sean de origen metabólico, como la hipertensión arterial (HTA), el hipercolesterolemia, el sobrepeso/obesidad o la diabetes mellitus (DM), o no modificables, como la edad (Figura 4)



Fuentes.
1. Global Burden of Disease Study 2019 (GBD 2019), Institute for Health Metrics and Evaluation (IHME), 2020.
2. Encuesta de morbilidad hospitalaria. Año 2019. Instituto Nacional de Estadística (INE).
3. Estadística de defunciones según la causa de muerte. Año 2019. INE.

Ilustración 5.67. Figura 3. Situación de las enfermedades cardiovasculares en España.



Fuente: Encuesta Europea de Salud en España 2020. INE

Ilustración 5.68. Figura 4. Prevalencia auto declarada de los factores de riesgo más prevalentes en la población con mayor asociación con las enfermedades cardiovasculares

Los factores de riesgo relacionados con los estilos de vida y biológicos (Figura 11) son los rasgos, características o exposiciones de un individuo que aumentan su probabilidad de sufrir una enfermedad. En el ámbito de las ECV también se denominan factores de riesgo cardiovascular (FRCV), los cuales pueden ser innatos o adquiridos.

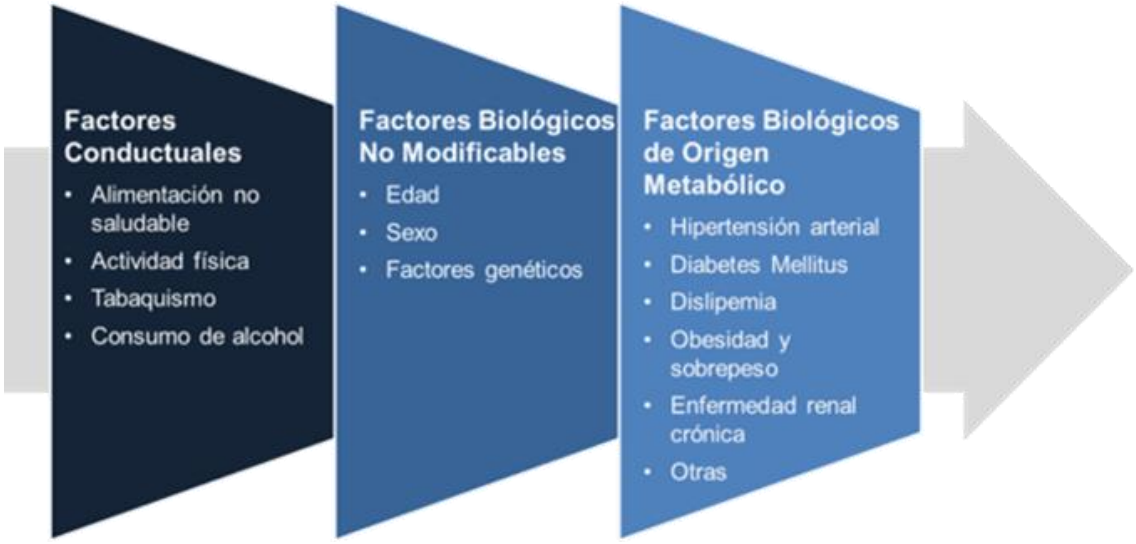


Ilustración 5.69. Figura 11. Clasificación de los factores de riesgo conductuales y biológicos

- La innovación tecnológica se ha incorporado de manera acelerada a todos los aspectos de la salud cardiovascular y la atención a las ECV. Su uso en la vida cotidiana de las personas incluye sistemas de monitorización de actividad física, alimentación, medida de pulso o de otras variables biológicas en dispositivos personales específicos o incorporados a teléfono móviles.

- En la atención sanitaria se emplea en todos los ámbitos, como la gestión clínica, los métodos diagnósticos y terapéuticos o el seguimiento y monitorización de pacientes, con espectaculares avances en el análisis de parámetros bioquímicos, genéticos, de señales (eléctricas…), en técnicas de imagen, la farmacología, los dispositivos y las intervenciones quirúrgicas. Sin embargo, el entusiasmo por la utilización de tecnología punta y los nuevos dispositivos en el cuidado de la salud y en la práctica clínica puede estar llevando a unas expectativas excesivas y a una incorporación acelerada de estos sin una evaluación adecuada de su efectividad y seguridad a largo plazo y sin un análisis detallado del coste-efectividad, lo que constituye un reto y una gran oportunidad en el ámbito cardiovascular.

- Especial atención debe prestarse a los sistemas de información, claves en cualquier estrategia e iniciativa de desarrollo. En España, los sistemas de información sobre salud están fragmentados y débilmente conectados. Hay instituciones que proporcionan información a nivel estatal: Instituto Nacional de Estadística, Centro Nacional de Epidemiología, Ministerio de Sanidad, con iniciativas y bases de datos de dato de ámbito estatal, como las Encuestas Nacionales de Salud en España (ENSE), el Conjunto Mínimo Básico de Datos (CMBD) de las altas hospitalarias del Sistema Nacional de Salud, o la Base de Datos Clínicos de Atención Primaria (BDCAP).

- No obstante, una gran parte de la información pertenece a las distintas comunidades autónomas o a las instituciones sanitarias locales –hospitales, centros de atención primaria– que utilizan frecuentemente herramientas distintas en cada centro (historias clínicas electrónicas de los hospitales) o por nivel asistencial (atención hospitalaria y atención primaria). Estas herramientas no se comunican o comparten información con dificultad por cuestiones técnicas, operativas o estratégicas, perdiéndose muchas oportunidades para el conocimiento global de la realidad epidemiológica y clínica de la ECV en España.

- Es esencial conseguir maximizar la posibilidad de explotar todos los datos disponibles de manera coordinada y utilizarlos como palanca de cambio para afrontar las necesidades de mejora y desarrollo de la salud cardiovascular y la atención a la ECV.

5.1.2. Estrategia de salud cardiovascular

La importancia de la prevención secundaria en pacientes con síndrome coronario/cardiopatía isquémica:

- El síndrome coronario y la cardiopatía isquémica o fase crónica del síndrome coronario: La incidencia en España de la Cardiopatía Isquémica (CI) en el año 2019 era de 361,4 nuevos casos por cada 100.000 habitantes, siendo mucho más elevada en hombres (463,4) que en mujeres (263,6). La prevalencia de CI en España en los últimos 10 años ha crecido paulatinamente, del 2,8% en 2009 al 3,3% en 2019, siendo del 4,2% en el caso de los hombres y del 2,4% de las mujeres (1).

- En las próximas décadas se espera que continúe la tendencia creciente de la prevalencia de la CI como consecuencia, en parte, del envejecimiento poblacional. En 2019 se estima que la CI causaba cerca de 1 millón de fallecimientos en la Unión Europea, lo que representaba el 48,5% de las muertes por ECV (1). En España, ese mismo año las muertes por CI eran de 29.247 personas, lo que suponía el 25,1% del total de muertes por ECV (2).

- Programas de rehabilitación cardíaca y prevención secundaria en atención hospitalaria y atención primaria según riesgo:

- La rehabilitación cardiaca está expresamente recomendada en la mayoría de las patologías cardiológicas debido a los beneficios funcionales, psicológicos y de pronóstico que produce, incluida la disminución de la morbimortalidad. Sin embargo, existe un déficit de unidades de referencia y de implementación de los programas de rehabilitación cardiaca en España (207). Debido al impacto de la pandemia en el último año se ha impulsado la telemedicina en el campo de la prevención y el desarrollo de programas domiciliarios e-supervisados, así como el acceso libre online de cualquier paciente que precise rehabilitación cardiaca.

- Junto al impulso al desarrollo de unidades de rehabilitación cardiaca en cualquier centro hospitalario, y en aquellos centros de AP que tengan las condiciones adecuadas, se considera que el impulso de un abordaje integral y multidisciplinar de las ECV –incluida la CI– en AP contribuiría a los resultados en salud. Dicho abordaje incluiría el seguimiento tras alta hospitalaria durante un período de un año, así como el desarrollo de cuidados individuales y familiares desde las consultas de medicina y enfermería, centrándose en el empoderamiento para lograr una adherencia al cambio de estilo de vida y al plan terapéutico de por vida (209).

- **ESCAV: Objetivos y acciones asociados a cardiopatía isquémica en ESCAV**

  Punto crítico CI1: Desarrollar programas de rehabilitación cardíaca y prevención secundaria hospitalarios y en atención primaria según el riesgo de los pacientes

OBJETIVOS GENERALES	CI-OG1. Mejorar el acceso a los programas de Rehabilitación Cardiaca (RC) y prevención secundaria de forma equitativa para hombres y mujeres tras IAM, tras revascularización coronaria mediante angioplastia o cirugía, haciéndolos extensivos a pacientes con enfermedad coronaria no revascularizables, estableciendo una red que incluya el propio hospital, los de referencia y la AP, organizaciones de pacientes, para mejorar la morbimortalidad y la calidad de vida
OBJETIVOS ESPECÍFICOS	• CI-OE1.1 Garantizar y facilitar la prestación de programas de Prevención Secundaria y RC hospitalaria en Fase II a pacientes con cardiopatía isquémica y riesgo moderado-alto lo antes posible tras sufrir un evento agudo (IAM, revascularización percutánea o cirugía cardiaca) • CI-OE1.2 Potenciar la continuidad asistencial mediante programas estructurados de Prevención Secundaria y RC en AP en Fase III y en Fase II en pacientes con riesgo bajo estableciendo indicadores que evalúen la eficacia y aprovechando las nuevas tecnologías • CI-OE1.3 Establecer medidas específicas para facilitar el acceso y la realización de programas completos de RC a las mujeres
ACCIONES	➢ CI-ACC1.1. Fomentar el desarrollo de Unidades de Prevención Secundaria y Rehabilitación Cardíaca (UPSRhC) en zonas donde aún no están implantadas con homogeneización de recursos y actividades ➢ CI-ACC1.2. Desarrollar programas ambulatorios (Fase II) interdisciplinares para pacientes de alto o moderado riesgo en hospitales terciarios o secundarios u otras instituciones sanitarias adaptadas al nivel de riesgo del paciente y para pacientes de moderado o bajo riesgo en hospitales comarcales, centros de salud o centros periféricos de especialidades coordinados con su hospital terciario o secundario referente potenciando la creación de redes de atención entre AP y AH ➢ CI-ACC1.3. Incorporar las nuevas tecnologías para reforzar el papel de los equipos de AP (medicina de familia, enfermería comunitaria y fisioterapia) en el abordaje de la fase II de bajo riesgo y fase III de la RC, estableciendo canales de comunicación ágiles y bidireccionales entre AP y hospitales (aplicaciones específicas, plataformas informáticas, e-consulta, historia electrónica compartida, App) Las herramientas tecnológicas de comunicación entre AP y AH deben ser versátiles y poder ser utilizada con diversos fines ➢ CI-ACC1.4. Designar centros de salud referentes en PSRhC en cada área o zona de salud ➢ CI-ACC1.5. Potenciar la continuidad asistencial mediante el diseño de protocolos y programas estructurados interdisciplinares de Prevención Secundaria y RC entre la AP y la AH que aseguren el seguimiento y los objetivos de la prevención secundaria estableciendo indicadores que evalúen la eficacia aprovechando las nuevas tecnologías ➢ CI-ACC1.6. Estratificación del riesgo, que permita prescribir y modificar el entrenamiento de forma individualizada y valorar el grado de supervisión que se debe realizar durante el programa de rehabilitación para dar seguridad al paciente ➢ CI-ACC1.7. Impulsar la elaboración y difusión de programas de Formación Continuada multidisciplinares de Rehabilitación Cardíaca y Prevención Secundaria entre todos los profesionales implicados en la atención del paciente con Cardiopatía isquémica

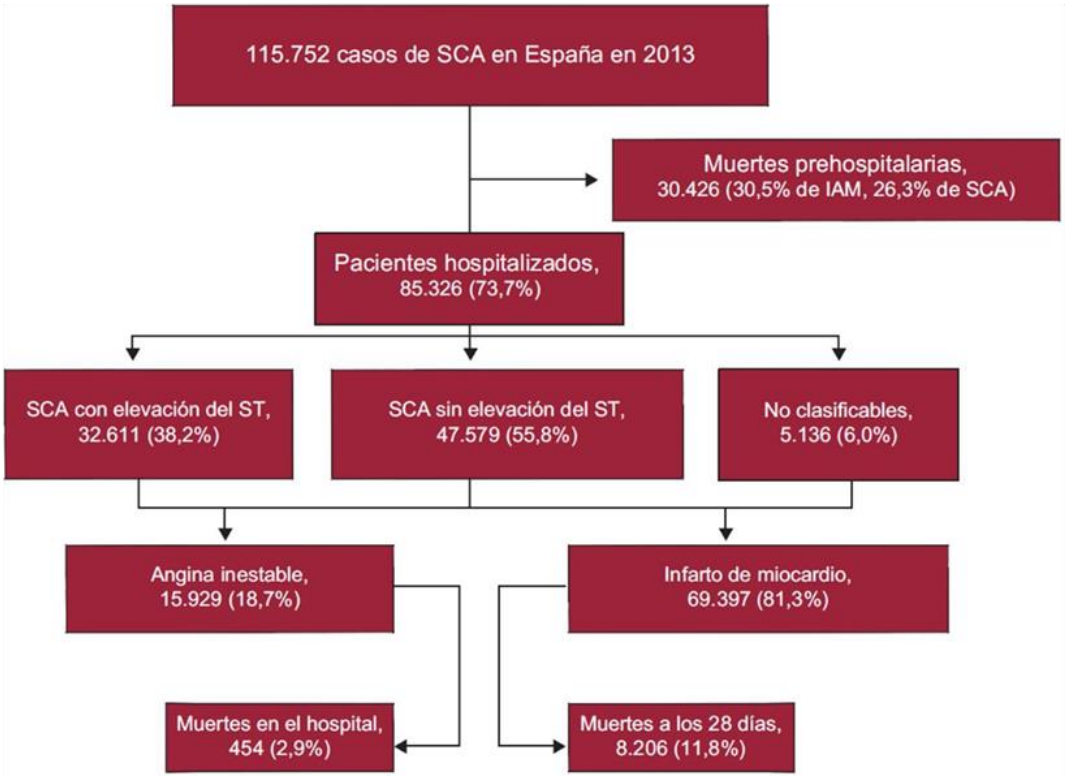
	percutáneo primario ➢ CI-ACC2.5. Definir estrategias de reperfusión para las áreas lejanas de los centros con realizar intervencionismo coronario percutáneo primario accesible las 24 horas todos los días y para los pacientes que acceden a hospitales sin realizar intervencionismo coronario percutáneo, evitando traslados intermedios ➢ CI-ACC2.6. Impulsar un programa de formación para todos los profesionales involucrados en el proceso de atención al IAM en los distintos niveles de atención con el objetivo de mejorar y homogeneizar el diagnóstico y el tratamiento del IAM ➢ CI-ACC2.7. Elaborar y poner en marcha un plan estratégico de educación, comunicación información y sensibilización a toda la población y a los profesionales, pero especialmente dirigidas a los colectivos más vulnerables (pacientes mayores y mujeres) para asegurar la equidad en el acceso al diagnóstico y a las terapias y promover un uso más eficiente de los sistemas de emergencias ➢ CI-ACC2.8. Establecer, junto con las CCAA, un mínimo de indicadores claves para que las redes de atención al IAM comuniquen información de forma homogénea ➢ CI-ACC2.9. Establecer, junto con las CCAA, un mínimo de indicadores clave sobre pacientes con diagnóstico de IAM que no han recibido tratamiento de reperfusión
OBJETIVOS GENERALES	CI-OG2. Mejorar la accesibilidad y el funcionamiento de las redes asistenciales de atenciónal infarto agudo de miocardio (IAM)
OBJETIVOS ESPECÍFICOS	• CI-OE2.1. Mejorar la morbimortalidad y el pronóstico del IAM, optimizando el funcionamiento de redes asistenciales específicas para la atención inmediata del máximo número de pacientes con sospecha de IAMCEST y garantizando el acceso a una estrategia invasiva en tiempo adecuado a los pacientes con SCASEST de alto riesgo, principalmente IAMSEST, ingresados en centros sin hemodinámica • CI-OE2.2. Incrementar el acceso a estas redes a las personas más vulnerables con SCASEST de alto riesgo (edad avanzada, con fragilidad y/o comorbilidades) y aumentar y acelerar las tasas de reperfusión en las mujeres con sospecha de SCACEST
ACCIONES	➢ CI-ACC2.1 Rediseñar las redes de atención al IAM (CIE-10: I21*; CIE-9: 410.0* a 410.7* [410.80 a 410.99 excluidos]) autonómicas y provinciales incorporando a los pacientes con criterios definidos de SCASEST de alto riesgo (CIE-10: I21.4* ó CIE-9: 410.7*) ➢ CI-ACC2.2. Promocionar el desarrollo de sistemas de registro y evaluación de calidad de la actividad asistencial mediante indicadores que aseguren buena práctica clínica, ausencia de variabilidad y equidad ➢ CI-ACC2.3. Establecer reuniones provinciales y autonómicas de la Red e identificar responsables provinciales y coordinador/a general de la red ➢ CI-ACC2.4. Definir protocolos de actuación y cuidados estandarizados (diagnóstico, tratamiento, transporte) en todo el ámbito geográfico la asistencia y la coordinación entre los sistemas de emergencias médicas y los hospitales capaces de realizar intervencionismo coronario

De las recomendaciones de la ESCV podemos concluir, para el diseño de nuestro caso de uso, que herramientas para la estratificación de pacientes según su riesgo cardiovascular es una de las líneas de trabajo claves que pueden aportar mucho valor para su implementación en el sistema sanitario.

5.1.3. Prevención secundaria del síndrome coronario

Recomendaciones de calidad del proceso asistencial del síndrome coronario de la sociedad española de cardiología

- SCA refleja la expresión clínica de la aterosclerosis coronaria y es la causa del desarrollo de la cardiopatía isquémica o síndrome coronario crónico. La gestión de esta entidad como un proceso no debe olvidar el abordaje de dicho sustrato. La actuación en la fase post hospitalaria, en forma de prevención secundaria es esencial para obtener un impacto significativo individual y poblacional.



- Si lo que pretendemos es ofertar una atención de calidad, no solo deberíamos incidir en los aspectos referentes al evento clínico, esenciales para la supervivencia del paciente en las mejores condiciones. También deberíamos aprovechar esa oportunidad para impactar sobre el sustrato que lo desencadenó. Es decir, sobre los aspectos preventivos que evitarán nuevos eventos clínicos con la mortalidad y morbilidad inherentes a ellos (Figura 5).

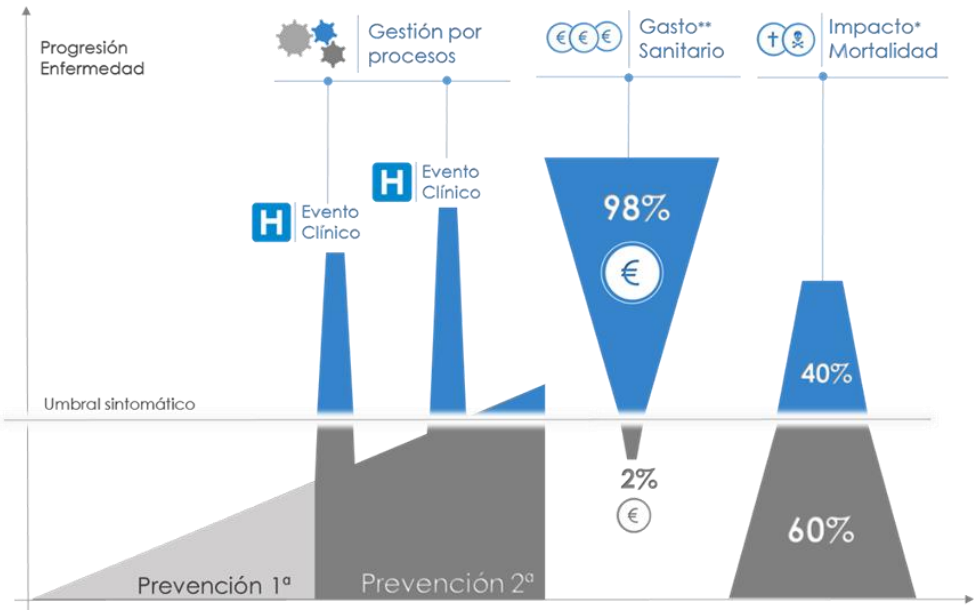


Ilustración 5.70. Figura 5. SCACEST como proceso.

- Desgraciadamente en España se dedican muy pocos recursos a este aspecto y aunque, en general, están bien resueltos los aspectos de atención sanitaria hospitalaria, especialmente para los eventos graves como el SCACEST, los aspectos de prevención primaria y secundaria, se encuentran descuidados tanto a nivel organizativo como presupuestario. De cada 100 euros gastados en sanidad en España, solo 2 los son en prevención. Y esta tendencia no ha variado en los últimos años (Figura 6). Todo ello es sorprendente sabiendo que dicha prevención es la responsable de evitar 6 de cada 10 muertes por enfermedad coronaria

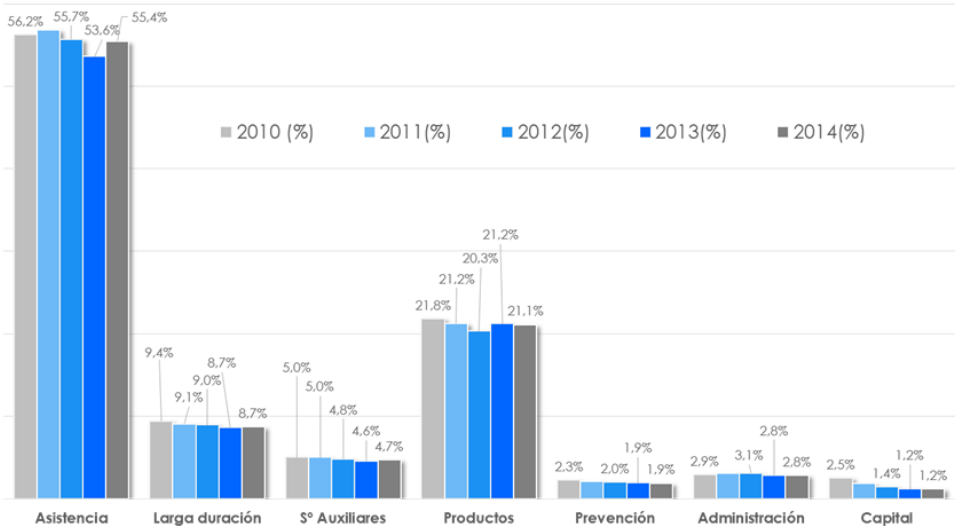
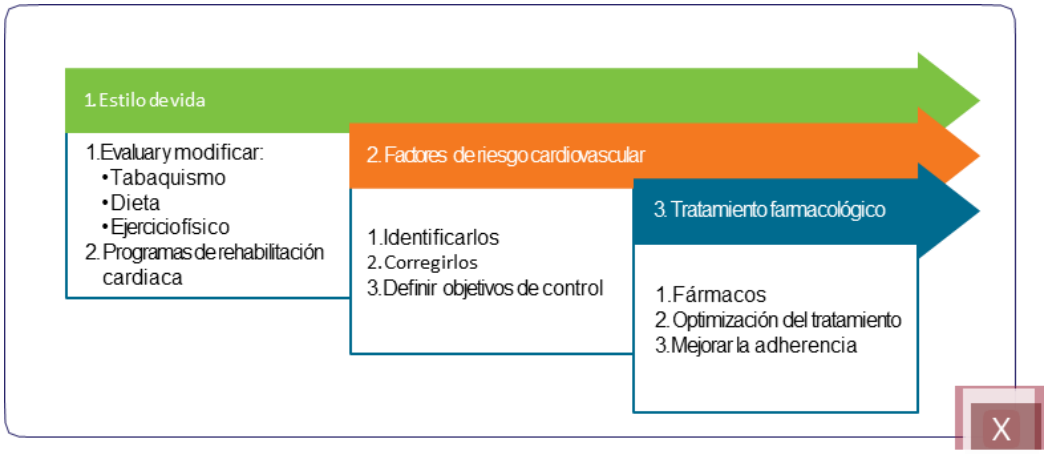


Ilustración 5.71. Gasto Sanitario en España.

- En los pacientes que presentan un síndrome coronario agudo (SCA) hay que establecer los objetivos de prevención secundaria, basándose en todas las recomendaciones cuya eficacia se ha comprobado (fig. 4.1).



- En todos los pacientes que han sufrido un SCA hay que identificar los factores de riesgo cardiovascular, para poder recomendarles medidas destinadas a corregir esos factores, en la medida de lo posible, y plantearles unos objetivos individualizados de control (tabla 4.4).

tabaquismo	Cese absoluto
Presión arterial	≤ 140/90 mmHg y evitar PAD < 60 mmHg
cLdL	< 70 mg/dL
HbA_1C	< 7%[a]
Ejercicio físico	Intensidad moderada 30 min/d, 5 días por semana
Peso corporal	IMC = 25 Perímetro abdominal en hombres < 102 cm y mujeres < 88 cm

[a] Hay que adecuar el objetivo a las características del paciente (edad, evolución de la diabetes, enfermedad vascular asociada, etc.).
cLDL: colesterol ligado a lipoproteínas de baja densidad; HbA_1C: hemoglobina glicosilada; IMC: índice de masa corporal; PAD: presión arterial diastólica.

Para que la prevención secundaria tras sufrir un SCA sea eficaz, es esencial optimizar el tratamiento médico; eso comprende elegir los fármacos de los que se ha visto que proporcionan una tasa mayor de supervivencia (fig. 4.4) y desarrollar todas las estrategias posibles para conseguir que el paciente siga lo más fiel y estrictamente posible las recomendaciones.

Ilustración 5.72. Figura 4.4. Tratamiento farmacológico para prevención secundaria del SCA

Proceso asistencial del paciente con síndrome coronario, proyecto ANJANA (Julio 2021)

- ANJANA es un proyecto colaborativo promovido por AMGEN (industria farmacéutica), en partenariado con SEIS y cuatro cardiólogos de referencia de España, y que tiene como objetivo la búsqueda de soluciones innovadoras para ayudar a mejorar la ruta del paciente cardiovascular a través de un análisis multifactorial que comprende varias áreas de intervención. El proyecto ANJANA tiene como objetivo principal la optimización del circuito asistencial para los pacientes que padecen un evento coronario agudo, desde la atención urgente hasta la prevención secundaria de nuevos eventos, a través de la implementación y optimización de soluciones TIC. A su vez, se busca mejorar la calidad asistencial del paciente con síndrome coronario agudo a través de los distintos niveles de cuidado que atraviesa.
- Durante la primera fase, representantes de la SEIS y cardiólogos pusieron en común las necesidades del paciente CV e ideas para la mejora de su itinerario, así como las

herramientas disponibles y posibles soluciones en el área de las TIC disponibles por parte de la SEIS. Durante la sesión se puso de manifiesto la existencia de diversos puntos críticos y/o líneas de mejora a lo largo de la ruta del paciente CV, pero también se evidenció el gran número de herramientas y posibles soluciones TIC para resolver dichos puntos. Todo esto puso de relieve la oportunidad de colaboración entre SEIS y cardiólogos con objeto de mejorar la atención al paciente CV. Como resultado de la sesión, los participantes definieron tres grandes áreas de trabajo relativas a las TIC a lo largo de la ruta del paciente:

1. Seguimiento de factores de riesgo CV
2. Estandarización del informe de alta
3. Telemedicina

- **El seguimiento de los factores de riesgo cardiovascular (FRCV),** engloba el análisis y tratamiento de todos aquellos factores que incrementen el riesgo de sufrir un evento CV. En el caso de los pacientes SCA, esta área de trabajo abarca distintos momentos de la ruta asistencial, desde la primera analítica completa al ingreso, hasta el seguimiento en AP.

- **Este punto clave será la base de trabajo de nuestro caso de uso, el desarrollo de la herramienta de evaluación del riesgo cardiovascular y su aplicación en la ruta/ proceso asistencial del paciente**

Análisis del proceso asistencial del síndrome coronario:

El proceso del síndrome coronario tiene dos fases de abordaje, una de la fase aguda desde que el evento coronario se produce hasta la estabilización del paciente en la hospitalización, en esta parte de entrada al hospital los puntos clave de intervención son:

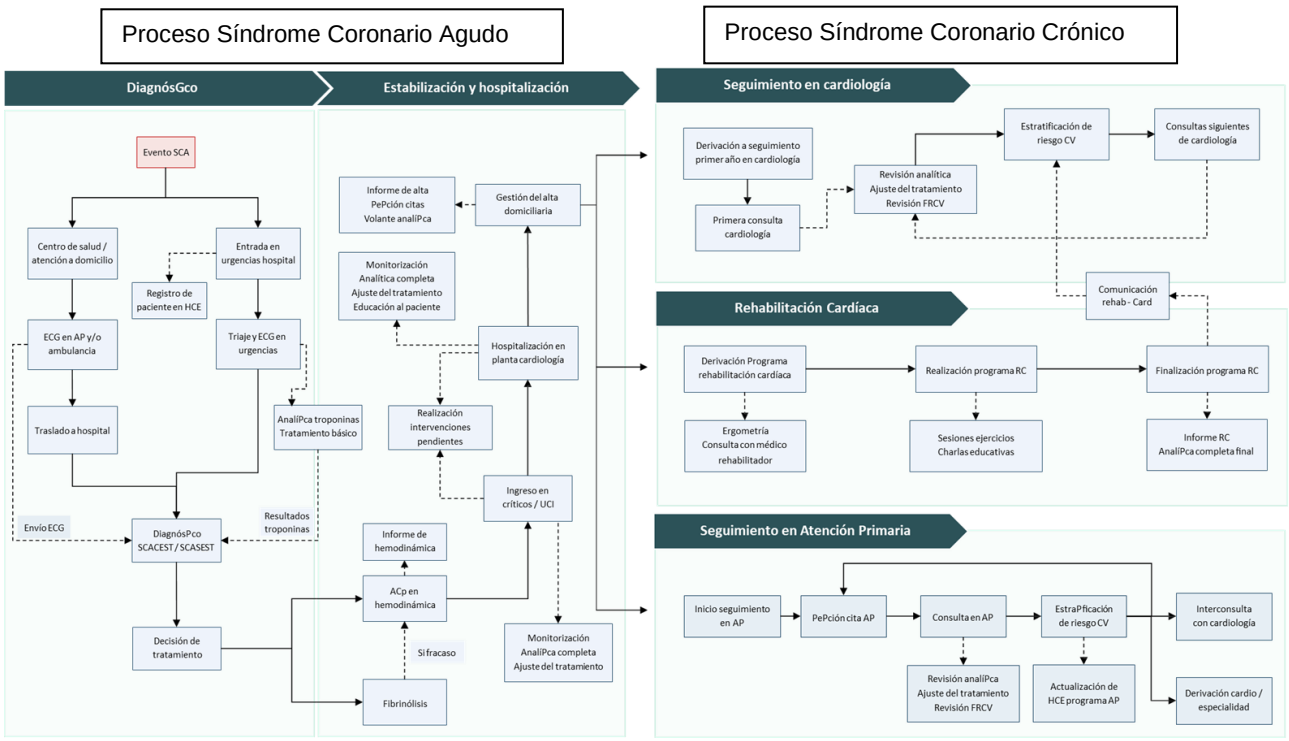
- Entrada en urgencias y registro del paciente en HCE
- Evaluación del paciente con analítica, electrocardiograma y tratamiento básico
- Confirmación diagnóstica y hospitalización en hemodinámica para posterior intervención y tratamiento, seguimiento en UCI y traslado a planta hasta estabilización del paciente
- Proceso de alta con codificación CIE10 e informe al alta en HCE y plan de control y prevención secundaria con seguimiento en cardiología y atención primaria

Posteriormente al alta del paciente se hace seguimiento durante el primer año en cardiología donde se pone en marcha el plan de prevención secundaria de riesgo CV con el objetivo de evitar un nuevo evento. La evaluación de los factores de riesgo durante el seguimiento del paciente es crítica en todos los niveles asistenciales por los que pasa el paciente.

1. Primera consulta en cardiología: se realiza analítica, ajuste de tratamiento y seguimiento de los factores de riesgo cardiovascular para ello se realiza una estratificación del riesgo del paciente y se pauta las visitas de seguimiento y la rehabilitación cardiaca

2. Rehabilitación cardiaca: el paciente comienza programa de autocuidado, y la formación y el control por el rehabilitador (pruebas de esfuerzo y ECG) así como la monitorización de la evolución de los factores de riesgo, finaliza el programa con informe de rehabilitación para seguimiento por cardiología durante el primer año y con la inclusión del informe en HCE.

3. Durante todo el proceso el paciente acude a su consulta de AP para seguimiento estrecho del programa de prevención, con analítica, ajustes de tratamiento y revisión de los factores de riesgo, coordinándose la atención en interconsultas programadas con cardiología durante el primer año y sucesivos si es necesario según su evolución



Áreas clave de intervención TIC para la digitalización del proceso asistencial para el desarrollo, seguimiento, monitorización y resultados en salud del plan de prevención secundaria del paciente con síndrome coronario.

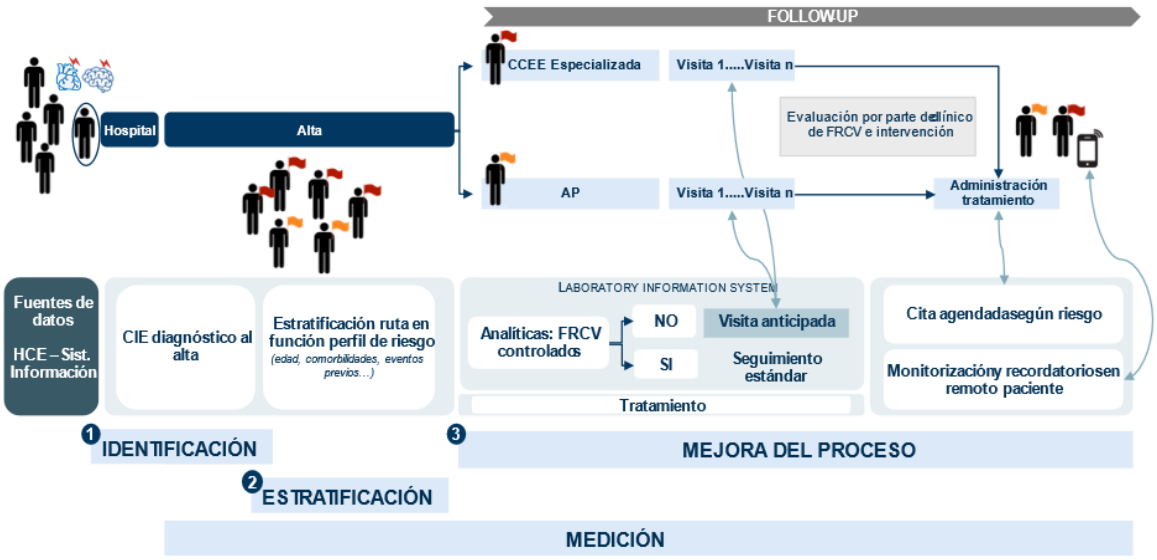
1. **Identificación de pacientes** durante el ingreso CIE10 y recogida de información del informe al alta en HCE

2. **Estratificación del riesgo** de los pacientes, desarrollo de algoritmo de riesgo en base a
   a. Factores demograficos: edad, sexo
   b. Estilo de vida: alimentación, ejercicio, tabaquismo
   c. Comorbilidades del paciente e indices de riesgo: Diabetes (Hba), hipertensión, (presión arterial) obesidad (IMC), hipercoleresterolemia (LDL) y (LpA), factores geneticos (Hipercoleresterolemia familiar)
   d. Tratamiento farmacológico: antiagregantes, estatinas, PCSK9, b-bloqueantes, IECA, Ara2, antidiabeticos

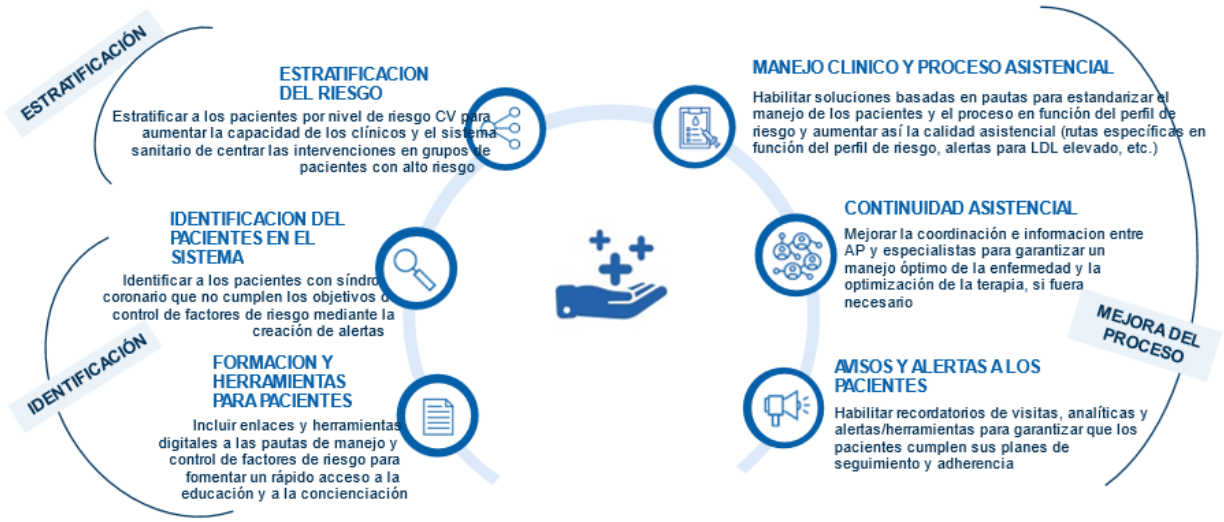
3. **Definición y automatización del proceso de seguimiento** en base al riesgo de nuevo evento en los pacientes, automatización de citas y alertas de adherencia al proceso.
   a. Programación de visitas en cardiología y AP segun riesgo del paciente cada 3, 6 o 12 meses
   b. Analíticas de seguimiento de los factores de riesgo y control del paciente en cada visita
   c. Impacto de programas presenciales y remotros de la rehabilitación cardiaca, estilo de vida, ejercicio y autocuidado
   d. Monitorización app remota del paciente de control de los factores de riesgo, comunicación paciente-cardiología y atención primaria

4. **Medición de seguimiento del proceso y análisis de resultados en salud e identificación de areas de mejora en el proceso de prevención secundaria**

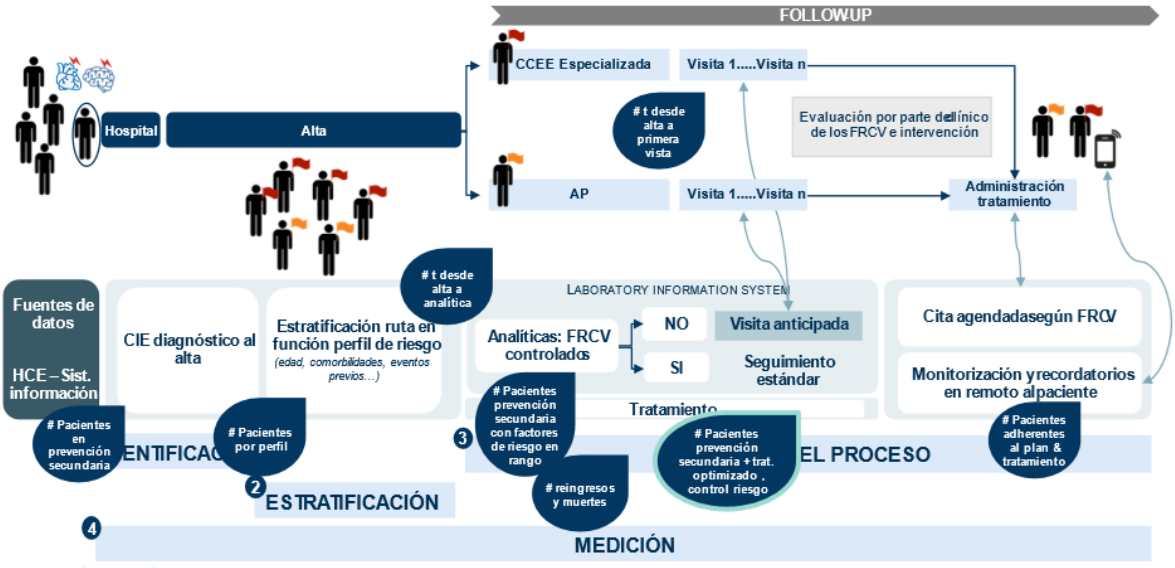
   - Proceso asistencial simplificado para identificación de áreas de intervención



- Lineas de desarrollo de herramientas TIC para la mejora del procesos y los resultados en salud de los pacientes con síndrome coronario



- Indicadores de seguimiento de la digitalización del proceso asistencial, control de los factores de riesgo CV y evaluación de correcta implementación y resultados



5.2. Caso de uso: Herramienta de A.A de RCV para la gestión del P.A. de prevención 2aria

5.2.1. Definición del objetivo: Objetivo del caso de uso

Aplicación de un data lake sanitario para el desarrollo de herramienta de estratificación del riesgo cardiovascular de los pacientes con síndrome coronario para su posterior aplicación en todas las fases de evaluación, seguimiento y monitorización del riesgo de los pacientes en el proceso asistencial

5.2.2. Desarrollo de la herramienta:

El objetivo de la herramienta es predecir la probabilidad de riesgo anual y hasta el año 10 de sufrir un nuevo evento en pacientes con síndrome coronario. Para ello se identificaran y cuantificarán estos pacientes en el sistema sanitario, comparando en la cohorte de pacientes que han sufrido un evento previo con síndrome coronario, un grupo serán los pacientes que hayan sufrido un segundo evento CV con hospitalización y/o muerte y el comparador será el otro grupo de pacientes con un primer evento pero que no hayan sufrido un segundo evento para la cuantificación del riesgo en prevención secundaria y su aplicación a modo de "calculadora de riesgo" en todas las fases de seguimiento del paciente en el proceso asistencial. (ingreso, trassición al alta, consultas en cardiología y consultas en atención primaria)

Identificación población:

Pacientes mayores de 18 años con diagnóstico de patologías incluidas en el síndrome coronario, para la identificación y cuantificación de estos pacientes en las fuentes de datos se utilizarán los códigos CIE10 de los grupos asociados con síndrome coronario:

Enfermedad CIE-10	Código CIE-10
Angina inestable	I20.0
Infarto agudo de miocardio	I21
Infarto agudo de miocardio subsiguiente con elevación de ST (IAMCEST) (IMEST) (STEMI) y sin eleva	I22
Complicaciones en curso, tras infarto de miocardio con elevación de ST (IAMCEST)(IMEST)(STEMI) y	I23
Otras enfermedades isquémicas agudas cardiacas	I24
Enfermedad isquémica crónica cardiaca	I25

Análisis de Variables de entrada en la población seleccionada: factores de riesgo cardiovascular (FRCV)

La recogida de estas variables se hará a través de la historia clínica electrónica, base de datos de laboratorio y datos procedentes de sistemas de monitorización remota de los pacientes (IoT)

Variables evalaución riesgo CV (FRCV)	Variables de entrada
Demografica	Edad
Demografica	Sexo
Demografica	Indice de masa corporal (peso/altura)
Demografica	Intensidad ejercicio fisico (escala tiempo/semana)
Clinica	Dislipemia (cLDL)
Clinica	Diabetes (HbA1C)
Clinica	Tabaquismo (SI/NO)
Clinica	Antecedentes genéticos familiares (HCF)
Clinica	hipertensión (presion arterial)
Clinica	Frecuencia Cardiaca
Imagen	Electrocardiograma

5.2.3. Resultados esperados de la herramienta de estratificación riesgo CV

Resultados esperados de la herramienta de estratificación riesgo CV y su aplicación en la digitalización en el proceso asistencial

Objetivo de la herramienta de evaluación de riesgo cardiovascular: desarrollo de herramienta de analítica avanzada que permita la evaluación del riesgo cardiovascular, riesgo de hospitalización y riesgo de mortalidad de pacientes con síndrome coronario y su validación para el uso en práctica clínica que permita a consecuencia de su uso la mejora de las intervenciones clínicas de prevención en estos pacientes.

Esta herramienta, a priori, podrá ser de aplicación en la digitalización del proceso asistencial de prevención secundaria de riesgo cardiovascular con intervenciones digitales complementarias (fuera del objetivo de diseño TFM)

La aplicación de la herramienta, posterior a su desarrollo, en el sistema sanitario para un mejor control y seguimiento de los pacientes y gestión de los procesos asistenciales tendrá varias áreas de aplicación en la digitalización del proceso asistencial:

1. Estandarizar en práctica clínica el uso del cálculo de riesgo individualizado y de población con síndrome coronario para adecuación del tratamiento y pautas de seguimiento clínico de los pacientes en prevención secundaria.

2. Modificación del proceso asistencial adaptado a los niveles de riesgo de los pacientes proporcionado por la herramienta en cada evaluación del paciente

(ingreso, alta y consultas de seguimiento)

3. Automatización y digitalización del proceso asistencial del síndrome coronario adaptado a los niveles de riesgo de los pacientes con generación de alertas.

4. Fomentar el seguimiento en remoto a través de uso de apps por los pacientes que permitan una mejor monitorización de los factores de riesgo

5. Aplicación del data lake para la evaluación del impacto de la digitalización del proceso asistencial de prevención secundaria en resultados en salud de los planes de prevención secundaria

5.3. Conclusiones del capítulo II

- **Salud cardiovascular y relevancia de las TIC:** La aplicación de herramientas TIC y los sistemas de información son muy relevantes y de gran valor para mejorar las enfermedades cardiovasculares que en nuestro país son la primera causa de muerte.
  Los avances tecnológicos y su aplicación en nuestro sistema sanitario, tanto en técnicas de intervención como de seguimiento de los pacientes en monitorización remota suponen un gran avance y oportunidad de mejora de la prevención, el control, manejo clínico y asistencial de las enfermedades cardiovascular y la eficiencia del sistema sanitario. Otro de los aspectos muy relevantes es la aplicación de las TIC y la analítica avanzada en la explotación de los datos tanto poblacionales como del sistema sanitario para seguir avanzando en el conocimiento, la investigación y la mejora de la salud cardiovascular

- El síndrome coronario es una de las manifestaciones más relevantes en la patología cardiovascular representando un 25% del total de muertes por ECV, una de las claves más relevantes para evitarlo es la prevención a través del control de los factores de riesgo cardiovascular durante todo el proceso asistencial del paciente La evaluación constante de evolución de los factores de riesgo es muy relevante para evitar los eventos cardiovasculares y crítica en los pacientes que ya han sufrido un primer evento, la analítica avanzada y la digitalización del proceso asistencial que permita el uso de estas y otras herramientas de control de los factores de riesgo sin duda aportarán gran valor en los resultados en salud y la mejora de la salud cardiovascular

- La creación de un data lake sanitario que nos permita la gestión de los datos poblacionales y sanitarios sin duda es de gran relevancia para el conocimiento de la salud de la población, el conocimiento de las enfermedades cardiovasculares y la mejora en la gestión clínica y asistencial del paciente, pudiendo además ser la base para el desarrollo de soluciones de analítica avanzada, como es nuestro ejemplo de caso de uso de una herramienta de evaluación del riesgo cardiovascular en pacientes con síndrome coronario, para su aplicación en el sistema sanitario.

6. Capítulo III. Soluciones de Analítica Avanzada

6.1. Introducción

El término "Data Lake" fue acuñado en 2010 por el director de Tecnología de Pentaho, James Dixon, para contrastarlo con el repositorio de almacenamiento de datos más refinado y procesado:

"Si piensas en un Data Mart como un almacén de agua embotellada, limpiada, empaquetada y estructurada para un fácil consumo, el Data Lake es un gran cuerpo de agua en un estado más natural. El contenido del data lake fluye desde una fuente para llenar el lago, y varios usuarios del lago pueden venir a examinar, bucear o tomar muestras".

Data Lake *Data Mart*

Ilustración 6.73. Analogía Data Lake – Mart (James Dixon)

Los Data Lakes son soluciones de gestión de datos, capaces de abordar los retos de Big Data y de alcanzar nuevos niveles de analítica en tiempo real. Proporcionan un repositorio de almacenamiento central, utilizado para contener una gran cantidad de datos granulares sin procesar (raw data), en formativo nativo. Facilitan la infraestructura necesaria para ML (Machine Learning) y AA (Analítica Avanzada).

Los avances en TI, almacenamiento masivo digital y “cloud computing” hacen cada vez más fácil y menos costoso almacenar grandes volúmenes de datos relacionados con la salud. El fin último, no es el almacenamiento, sino que éste es un medio para la analítica. Es el supuesto que

nos ocupa en este TFM, con la implementación de un Data Lake Sanitario.

El objeto de este apartado es analizar las herramientas necesarias para implementar un Data Lake que nos permita aplicar tecnologías de Big Data y Analítica avanzada (Inteligencia Artificial, Machine Learning, Deep Learning) para realizar una explotación masiva de información sanitaria, cumpliendo con garantías de coherencia y consistencia de los datos y con capacidad para llevar a cabo análisis predictivos, descriptivos y prescriptivos.

En puntos anteriores se habla en profundidad sobre los principios en los que se fundamenta un Data Lake Sanitario, qué debe aportar, normativa de protección de datos en la que se sustenta, la importancia de la gobernanza, y entre otros objetivos, el cumplimiento del principio **"el investigador va al dato, el dato no viaja al investigador"**.

Se propone un modelo de Gobernanza para la colaboración (4.5.2) con privacidad por diseño, aprendizaje automático federado, modelo de doble institución para colaboraciones público-privadas, así como, con un modelo específico de Arquitectura y Aprovisionamiento en cloud con modelos de consumo-provisión a demanda (4.5.4.1).

Como parte de esta sección "Soluciones de Analítica Avanzada", se aborda la Organización de un Data Lake por Capas Lógicas, en el que tiene cabida, tanto el modelo citado en el párrafo anterior, propuesto en este TFM, como cualquier otro que se decida implementar. Para cada una de estas Capas del Data Lake, se indican los objetivos y especificaciones que deben cumplir a modo de guía o mapa de ruta, que sirva de ayuda a la hora de buscar entre las herramientas disponibles (open-source o propietarias).

Previamente, veremos conceptos y métodos analíticos existentes y aplicables a grandes volúmenes de datos.

Finalmente desarrollamos los pasos necesarios para aplicar el caso de uso expuesto en este TFM dentro de nuestro Data Lake.

6.2. Analítica Avanzada

¿Qué es la Analítica? Es el conjunto de procesos para el descubrimiento, interpretación y comunicación de patrones en los datos.

En este caso, la analítica se plantea dentro de un modelo organizativo, donde los procesos analíticos tienen como fin la obtención de algún valor o mejora dentro de la organización, o bien un objetivo investigador – científico.

Tal y como se recoge en el tema "Analítica y modelos predictivos en Salud" del master DSTICSDS, la "analítica en salud" es un área que combina tecnología, métodos de análisis y

procesos de gestión, que conllevan la puesta en marcha de determinadas prácticas de adopción progresiva. El enfoque analítico es un cambio en la cultura de evaluación de la organización.

Los modelos de adopción o de madurez nos permiten la implantación progresiva y eficaz de un proceso, siempre que aporten elementos descriptivos y prescriptivos de manera coherente (Pöppelbuß and Röglinger 2011).

En este aspecto, vamos a exponer un Modelo de Madurez Analítico en salud que, aunque no sea una guía de buenas prácticas, permite conocer la conveniencia de implantar la analítica en función del grado de madurez de la organización en la madurez de sus datos. Se trata del Modelo de Adopción Analítico de 8 niveles de Healthcouldsolutions (fuente Master DSTICSDS)

Ilustración 6.74. Modelo de Adopción Analítico de 8 niveles Healthcloudsolutions

Nivel 0: Soluciones Fragmentadas. Nivel ineficiente. No hay procesos de gestión de los datos- Soluciones ad hoc en diferentes departamentos.

Nivel 1: Almacén de Datos Organizativo. Se utiliza tecnología. Se crea un almacén de datos integrado sobre el que se realizan los procesos de gestión de datos.

Nivel 2: Vocabulario estándar y registro de pacientes. Organización semánticamente coherente. Homogenización de códigos y vocabularios de las diferentes fuentes integradas y estructurarlos a un modelo de historia o registro del paciente.

Nivel 3: Informes automatizados internos. Permite construir información consistente y eficiente internamente. Construcción de KPI (Key Performance Indicators) e informes para gestión.

Nivel 4: Informes automatizados externos. Permite construir información consistente y eficiente al exterior. Informes para cumplimiento normativo – regulatorio. Requiere de una gestión y gobierno centralizado de los datos.

Nivel 5: Reducción de la variabilidad y el gasto. Creación del equipo analítico y búsqueda de mejora activa. Se aplican técnicas analíticas para identificar buenas prácticas clínicas que minimicen el gasto y reduzcan la variabilidad. Cuenta con un equipo analítico interdisciplinar de búsqueda activa de mejoras en calidad.

Nivel 6: Gestión de la salud y la analítica descriptiva. Desarrollo de métricas de gestión. Analítica descriptiva. Inclusión de fuentes de datos adicionales: aparatos de monitorización, monitorización tele-asistida, farmacia…. Se centra en métricas de calidad que se proporcionan a los profesionales.

Nivel 7: Intervención en riesgos clínicos y analítica predictiva. Analítica predictiva. Se centra en modelos predictivos de riesgos y costes basados en categorías diagnósticas y modelos de riesgo.

Nivel 8: Medicina Personalizadas. Personalización y apoyo a la toma de decisiones. La analítica se extiende a nivel individual y una gestión más del bienestar. En este punto se suele expandir la tipología de datos al texto y analítica de carácter prescriptivo como apoyo a la toma de decisiones de los profesionales.

En este punto es necesario mencionar también el Modelo de **Gartner Analytic Ascendancy**. En él se describen los cuatro tipos de análisis, incrementales en cuanto a dificultad y valor, que normalmente van en consonancia con el grado de madurez de la organización:

- **Análisis descriptivo:** se basa en la retrospectiva. Determina lo que ya ha sucedido en los datos.
- **Análisis de diagnóstico:** basada en la información, identifica por qué ocurrió un evento o cambio en particular en los datos.
- **Análisis predictivo:** Previsión y la determinación de lo que sucederá a continuación.
- **Análisis prescriptivo:** Determinar cómo hacer que el resultado deseado se haga realidad.

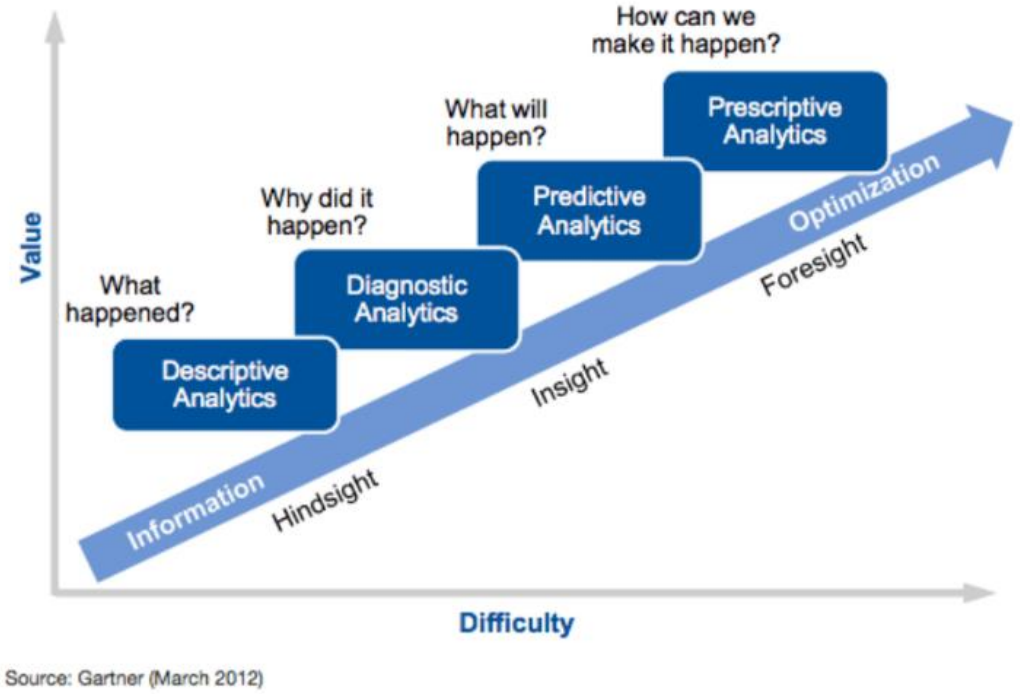


Ilustración 6.75. Cuadro de Gartnet (March 2012) Fuente Master DSTICS

6.3. Minería de Datos

Los procesos analíticos normalmente tener lugar como parte de un programa analítico, dentro de un esquema organizativo y ordenado, donde se suelen tomar como referencia los procesos de minería de datos.

La **minería de datos** es el proceso de encontrar patrones y extraer datos útiles de grandes conjuntos de datos. Es un conjunto de actividades orientada a descubrir conocimiento valioso y oculto (patrones, relaciones, hechos). La minería de datos es un marco analítico general, del que forma parte clave el aprendizaje automático (**Machine Learning**).

Incluye muchas actividades previas al trabajo de extracción de patrones (obtención de datos de fuentes externas, procesamiento de datos), precisando también de conocimiento del dominio y conocimientos técnicos. Se trata de un proceso iterativo, con evaluaciones precisas antes del uso de los modelos resultantes.

La minería de datos y el aprendizaje automático se encuadran en el marco de actividades de la disciplina KDD (Knowledge Discovery in Databases process) “Descubrimiento de conocimiento

en bases de datos", ya mencionada como parte de la gestión del Ciclo de Vida de los datos analíticos en el punto *4.5.3.1.* con la esquematización de las actividades en la ***Ilustración 49*** como un ciclo. Aquí tenemos otra representación de las fases de este ciclo:

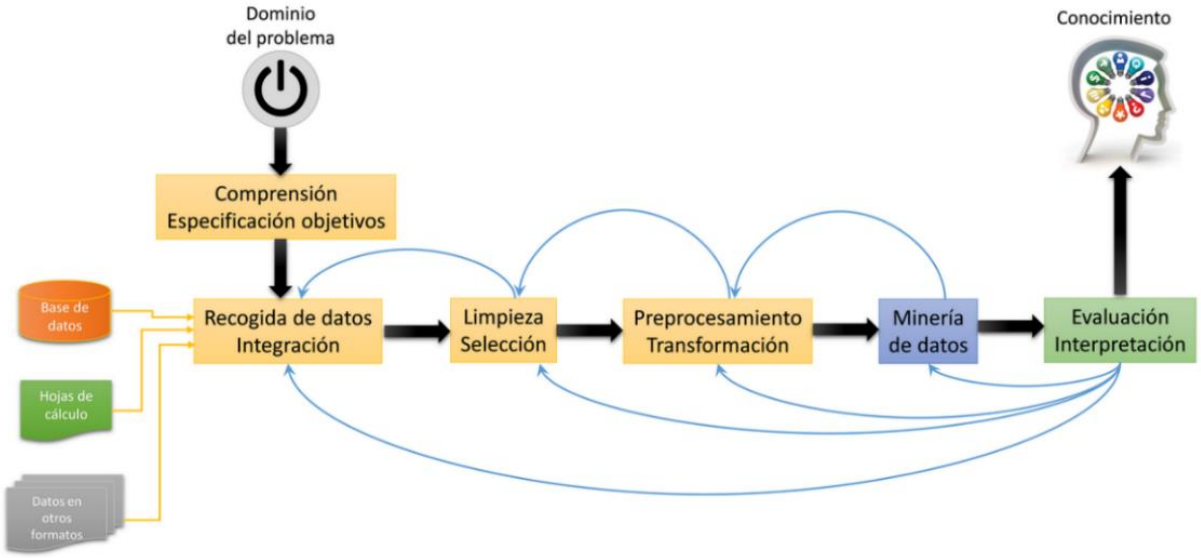


Ilustración 6.76. Ciclo de Actividades KDD

- Conocimiento del "Dominio del Problema" y compresión de la "Especificación de Objetivos".
- Recogida de datos e Integración de Fuentes en un almacén común.
- Selección y Limpieza de los datos.
- Preprocesamiento y transformación.
- Aplicación de las técnicas adecuadas de Minería de datos.
- Evaluación e Interpretación de los patrones extraídos.
- Comunicación y uso del nuevo conocimiento.

Modelo estándar CRISP-DM (Cross Industry Standard Process of Data Mining).

El estándar CRISP-DM, nos da una visión más amplia y generalizada en las actividades de minería de datos. Establece las fases necesarias para cubrir el ciclo de vida propio de las técnicas de minería de datos, que son comúnmente aplicadas al desarrollo de modelos analíticos.

Ilustración 6.77. Ciclo de vida CRISP-DM

CRISP-DM divide el proceso de minería de datos en seis fases principales. Establece un conjunto de tareas y actividades para cada fase del proyecto, pero no especifica cómo llevarlas a cabo.

La secuencia de las fases no es rígida: se permite movimiento hacia adelante y hacia atrás entre diferentes fases. El resultado de cada fase determina qué fase, o qué tarea particular de una fase, hay que hacer después.

El proyecto no se termina una vez que la solución se despliega. La información descubierta durante el proceso y la solución desplegada pueden producir nuevas iteraciones del modelo.

Fase I. Business Understanding Comprensión del negocio. Definición de necesidades del cliente.

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto. Después se convierte este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Fase II. Data Understanding. Estudio y comprensión de los datos.

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Fase III. Data Preparation. Preparación de los Datos Análisis de los datos y selección de características

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Fase IV. Modeling. Modelado

En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

Fase V. Evaluation. Evaluación (obtención de resultados)

En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde una perspectiva de análisis de datos. Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

Fase VI. Deployment. Despliegue (puesta en producción)

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.

6.4. Machine Learning

El aprendizaje automático o **Machine Learning (ML)** es una disciplina dentro de la **Inteligencia Artificial (IA)** en la que se trabaja en el desarrollo de modelos que permitan aprender de los datos, El resultado es una serie de técnicas algorítmicas preparadas para que los analistas de datos o data scientist las utilicen, aún sin conocer en profundidad cómo funcionan internamente.

Para obtener un buen modelo con el aprendizaje automático, es fundamental tener en cuenta los datos, sus distribuciones, la selección de variables relevantes y una correcta selección del algoritmo.

Hoy en día el Machine Learning juega un papel clave en muchas innovaciones en salud, incluido el desarrollo de nuevos procedimientos médicos, el manejo de datos de pacientes y tratamiento de enfermedades crónicas, Telemedicina, Ensayos Clínicos, Gestión de datos y privacidad, Prevención de Enfermedades, Ayuda con el diagnóstico, Ayuda con el tratamiento, entre otros

6.4.1. Modelos Predictivos

Los **modelos predictivos** son aquellos que permiten determinar alguna característica de un nuevo individuo con un cierto nivel de confianza. Permite tomar decisiones y aporta datos nuevos valiosos. De muchos algoritmos de aprendizaje automático se obtienen modelos que, tras evaluarse rigurosamente, pueden utilizarse como modelos predictivos. Los Tipos de modelos predictivos (también llamados Aprendizaje Automático Clásico):

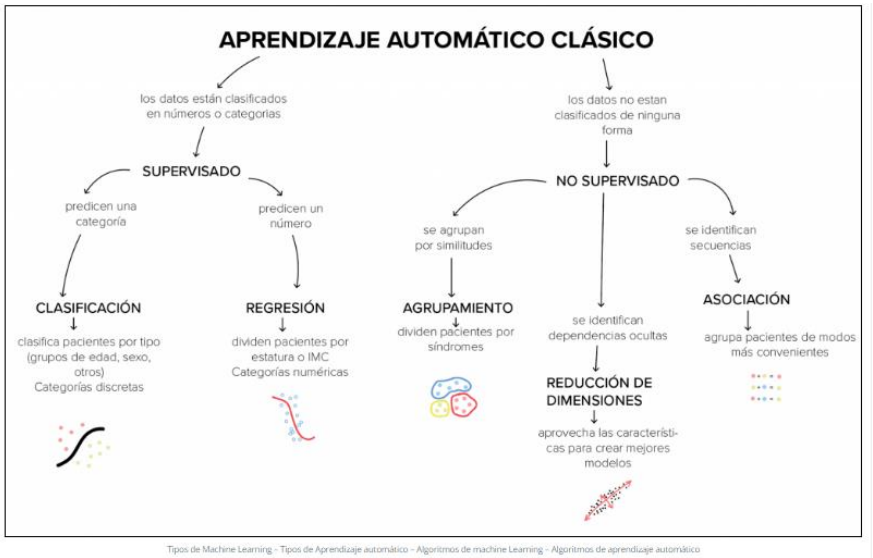


Ilustración 6.78. Tipos de Modelos Predictivos

Tipos de Algoritmos de Aprendizaje automático:

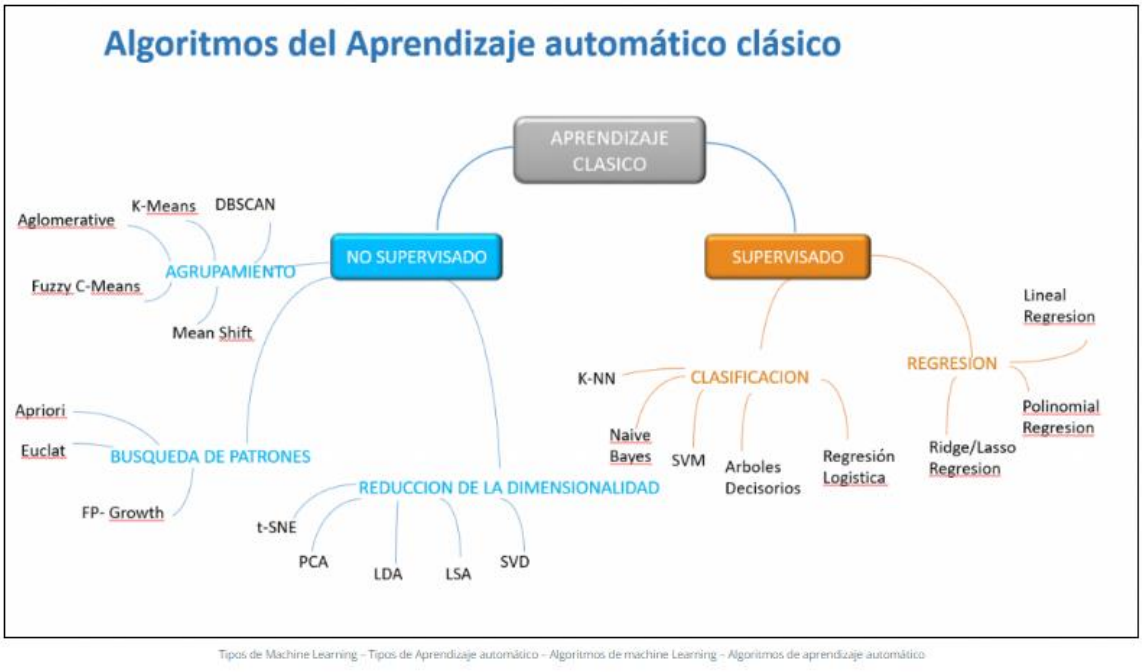


Ilustración 6.79. Tipos de Algoritmos de Modelos Predictivos

- **Supervisado:** Disponemos de datos etiquetados. Datos clasificados en números o categorías. Técnicas:

    - **Clasificación:** Busca la pertenencia o no a un conjunto de datos discreto. Categorías Discretas. Clasificación de pacientes por tipo.
        - Tipos de Algoritmos: Árboles de decisión, Regresión Logística, SVM, Naive Bayes, KNN

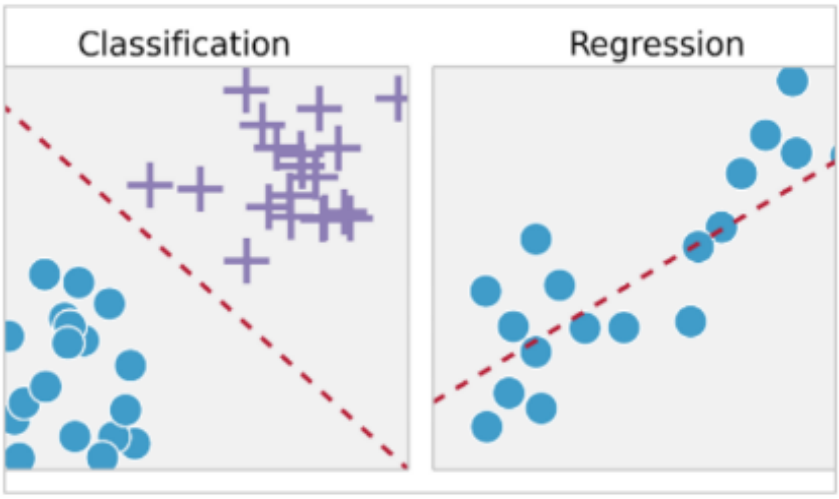
Ejemplos de Clasificación:

Clasificación de pacientes por Riesgo de Cardiopatía. Tendremos una serie de variables categóricas: Cardiopatía (etiquetado), Edad, Sexo, HTA, Diabetes, Obesidad, Tabaquismo, Cardiopatía... con una cantidad de pacientes suficientes para entrenar el algoritmo.

Desidentificación de pacientes, mediante clasificación: Objetivo, eliminar el proceso manual de desidentificación de registros médicos y automatizar el proceso. Un desafío importante es encontrar un conjunto de datos diverso para entrenar el modelo. El modelo debe garantizar la privacidad del paciente y al mismo tiempo crear una representación compartible del texto médico. Esto garantiza que la información se pueda utilizar cuando sea necesario para cualquier investigación. Se debe construir un clasificador de desidentificación robusto para separar la información privada de los datos a los que puede acceder públicamente.

- **Regresión:** Predicción de una variable continua. Predicen un número. Categorías Numéricas.
  - Tipos de Algoritmos: Regresión Lineal, Regresión Polinomial, Regresión Ridge/Lasso

Ejemplo de Regresión: Predicción de tiempo de Estancia de los pacientes en Hospitalización. Variables: Tiempo de estancia (Etiquetado), Edad, Sexo, HTA, Diabetes, Obesidad, Tabaquismo, Cardiopatía...



Fuente: "IPython cookbook", Cyrille Rossant

Ilustración 6.80 Representación de Clasificación – Regresión. Fuente Master

- **No supervisado:** Datos no etiquetados. Busca patrones, grupos o regularidades entre los individuos. Técnicas:

  - **Clustering:** Divide objetos en función de características desconocidas. La máquina clasifica a su mejor criterio.
    - Tipos de Algoritmos: K-Means, DBSCAN, Agglomerative, Fuzzy C-Means, Mean-Shift

  - **Búsqueda de patrones:** Busca patrones en el flujo de órdenes.
    - Tipos de Algoritmos: A priori, Euclat, FP-Growth

  - **Reducción de Dimensionalidad o aprendizaje de características:** Ensamblar variables específicas en nuevas de más alto nivel.
    - Tipos de Algoritmos: t-SNE, PCA, LDA, LSA, SVD

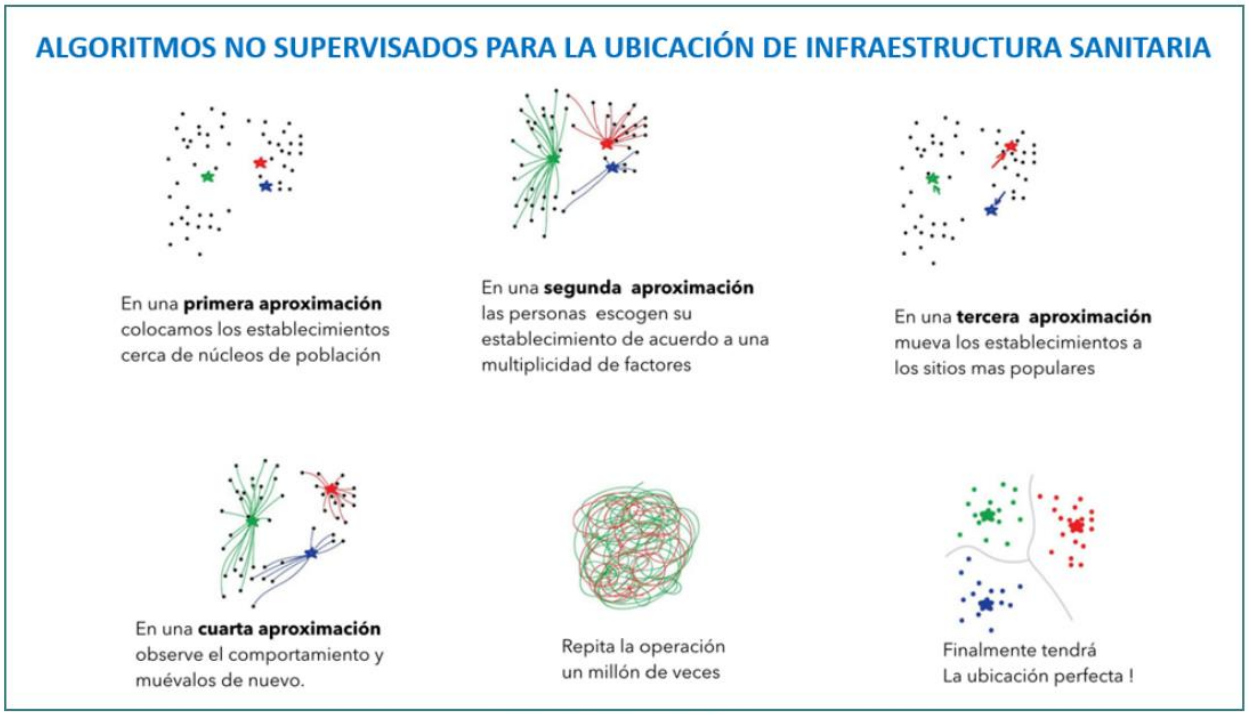


Ilustración 6.81 Uso de algoritmo K-means en el ámbito de salud

Un ejemplo sobre algoritmo de aprendizaje no supervisado: Predicción de Epidemias en un área específica. Se pueden construir modelos de aprendizaje automático que predicen la naturaleza de la propagación de una epidemia en un área y también determinan dónde es más probable que ocurra el próximo brote de una epidemia. Se deben tener en cuenta factores como la geografía, el clima, la demografía y la distribución de la población de un área afectada al entrenar el modelo de aprendizaje automático para que pueda identificar otras áreas propensas a los brotes.

En el artículo “Aplicaciones de la inteligencia artificial en cardiología: el futuro ya está aquí” (2019) (Ignacio Dorado, Jesús Sampedro, Victor Vicente y Pedro L. Sánchez) podemos encontrar referencias a varios estudios en los que se ha usado Aprendizaje Automático. Estos son algunos ejemplos extraídos de este artículo:

Contribuciones relevantes de la inteligencia artificial en las diferentes áreas de aplicación de la cardiología

Referencia (año)	Área	Aplicación	Técnica	Método	Resultados
Ebrahimzadeh et al.[22] (2018)	Arritmias	Predicción de FA paroxística a partir de la variabilidad de frecuencia cardiaca	Aprendizaje supervisado	Datos: 106 señales de 53 pares de electrocardiogramas para el entrenamiento Algoritmos: KNN, SVM, RN	Sensibilidad (100%), especificidad (95%), exactitud (98%)
Budzianowski et al.[23] (2018)	Arritmias	Predicción de recurrencia de FA tras crioablación de venas pulmonares	Aprendizaje supervisado	Datos: 118 pacientes con 56 señales clínicas, laboratorio y del procedimiento de cada paciente Algoritmos: GB, SVM, sobremuestreo	Identificación de 7 predictores en concordancia con análisis estadístico univariante
Eerikainen et al.[24] (2016)	Arritmias	Clasificación de alarmas por arritmias cardiacas en telemetría	Aprendizaje supervisado	Datos: PhysioNet/ Computing in Cardiology Challenge 2015 Algoritmo: RF	VP 95%, VN 83%
Nanayakkara et al.[25] (2018)	Arritmias	Predicción de mortalidad hospitalaria en pacientes con parada cardiaca resucitada a partir de un registro	Aprendizaje supervisado	Datos: Registro ANZICS, 39.566 pacientes Algoritmos: RL, GB, SVM, RN, RF, en conjunto (RF, SVM, GM)	Área bajo la curva del mejor algoritmo: 0,87 (frente al 0,80 de la escala APACHE III y 0,81 ANZROD)
Yildirim et al.[26] (2018)	Arritmias	Detección de hasta 17 tipos de arritmias a partir del ECG	Aprendizaje supervisado	Datos: 1.000 señales de ECG de la base de datos MIT-BIH Arrhythmia database Algoritmo: RN convolucional	Exactitud (91%)

Ilustración 6.82 Contribuciones de la IA en diferentes áreas de aplicación de la cardiología.

En todos los casos es muy importante elegir bien el algoritmo a utilizar. En el tema “Analítica y Modelos Predictivos en Salud”, vimos el modelo de ayuda a la selección de algoritmo scikit-learn.

En la siguiente ilustración, se muestra un modelo extendido, llamado por ”complicación” en el que se tiene en cuenta la fase previa, validez y acceso a los datos de origen, pretratamiento y limpieza de los mismos y planteamiento de las preguntas adecuadas para el caso de uso de analítica avanzada, que es con lo que nos encontramos en la vida real, antes de poder aplicar el algoritmo Scikit-Learn.

La aplicación de aprendizaje automático normalmente comienza en el nivel predictivo, dado que los estudios solamente exploratorios o descriptivos se suelen considerar como estudios previos a la investigación en sí. Y termina en el nivel predictivo, dado que los estudios causales o mecanísticos son propios de la ciencia.

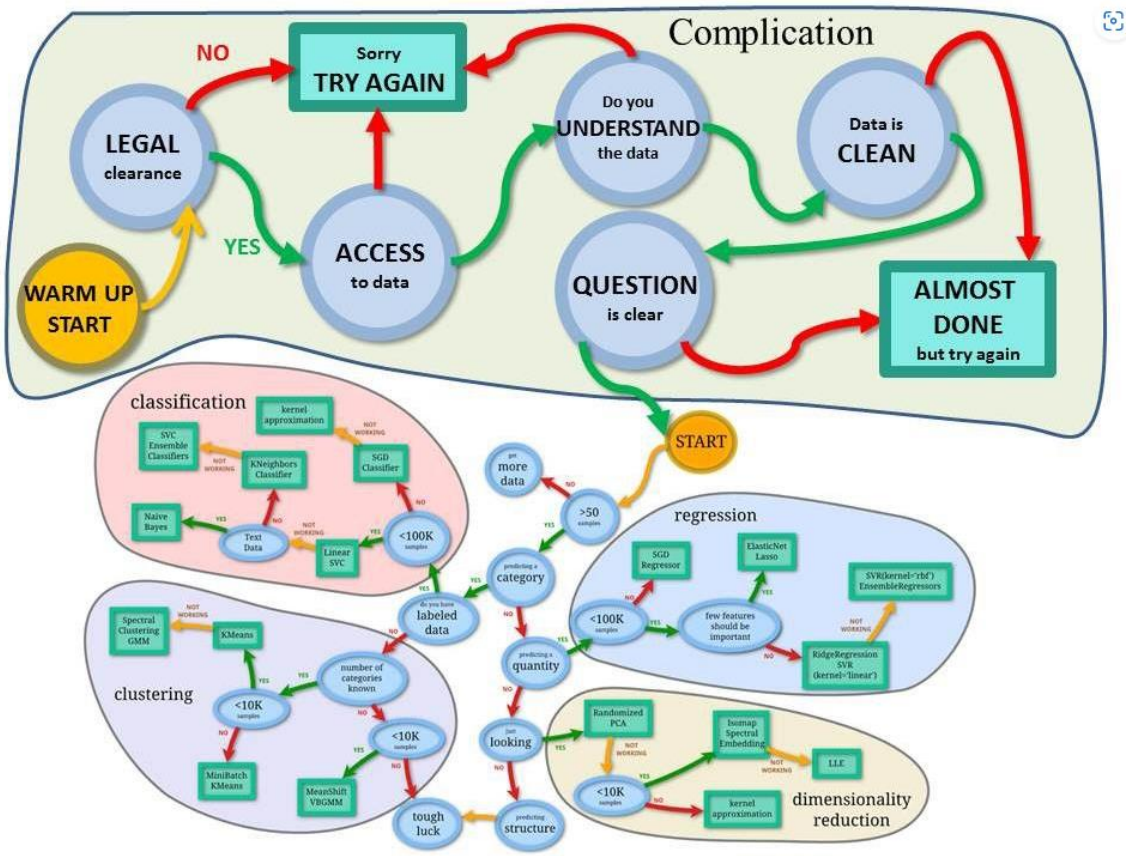


Ilustración 6.83 Extended Version of the Scikit-Learn Cheat Sheet

6.4.2. Aprendizaje Profundo (Deep Learning)

Es un conjunto de algoritmos de aprendizaje automático que intenta modelar abstracciones de alto nivel en datos usando arquitecturas computacionales que admiten transformaciones no lineales múltiples e iterativas de datos expresados en forma matricial o tensorial.

El aprendizaje profundo es parte de un conjunto más amplio de métodos de aprendizaje automático basados en asimilar representaciones de datos. Una observación (por ejemplo, una imagen) puede ser representada en algunas formas (por ejemplo: un vector de píxeles), pero algunas representaciones hacen más fácil aprender tareas de interés (por ejemplo, "¿es esta imagen una cara humana?") sobre la base de ejemplos, y la investigación en esta área intenta definir qué representaciones son mejores y cómo crear modelos para reconocer estas representaciones.

Ejemplo de Deep Learning en Salud: Segmentación de imágenes para el pronóstico de tumores cerebrales. Objetivo diagnosticar con precisión los tumores cerebrales para garantizar que el

paciente reciba el tratamiento adecuado. Puede utilizar redes neuronales complicadas para realizar la segmentación semántica. Las redes neuronales complicadas o CNN son un tipo de red neuronal artificial profunda que se utiliza ampliamente en la visión por computadora para la segmentación de imágenes. La mayoría de las CNN usan kernels bidimensionales, pero algunas pueden usar kernels tridimensionales y, por lo tanto, acceder completamente a la estructura tridimensional de las imágenes médicas. La segmentación de imágenes médicas plantea desafíos adicionales, como la escasez de datos etiquetados y la gran demanda de memoria de las imágenes médicas tridimensionales. (Proyect Pro)

Varias arquitecturas de aprendizaje profundo, como redes neuronales profundas, redes neuronales profundas convolucionales, y redes de creencia profundas, han sido aplicadas a campos como visión por computador, reconocimiento automático del habla, y reconocimiento de señales de audio y música, y han mostrado producir resultados de vanguardia en varias tareas.

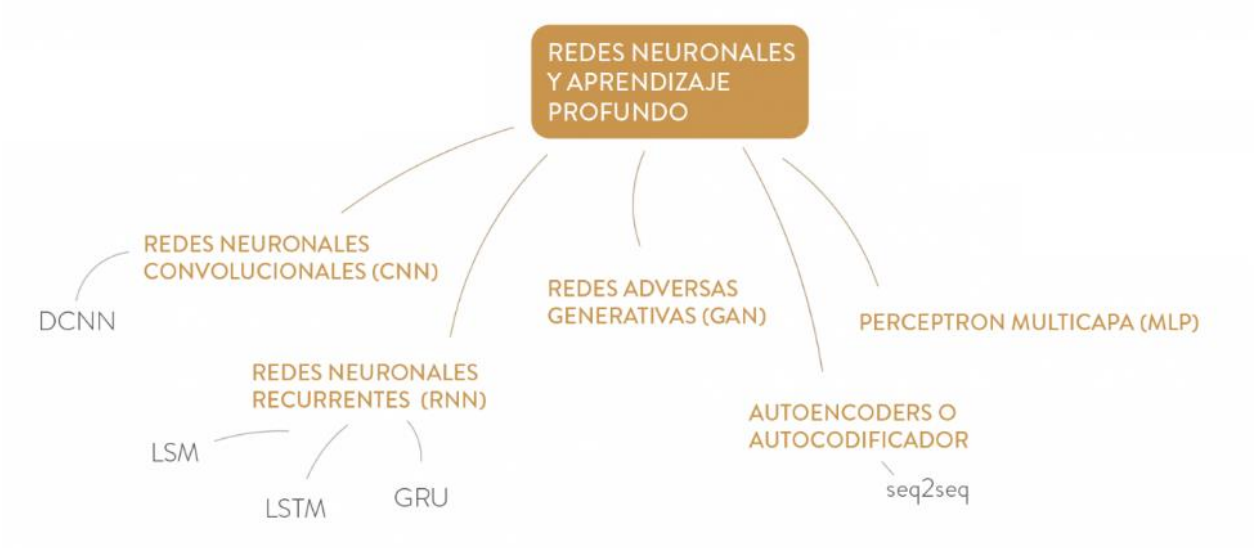
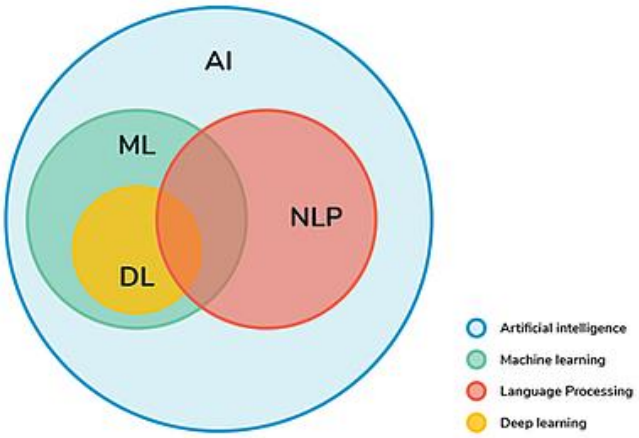


Ilustración 6.84 Redes Neuronales y Aprendizaje Profundo

6.4.3. Procesamiento de Lenguaje Natural (PNL)

El procesamiento del lenguaje natural (NLP, por sus siglas en inglés) es la interpretación del lenguaje humano por parte de una máquina. Como se puede ver en la Figura, tanto NLP como machine learning son parte de la inteligencia artificial y ambas ramas comparten técnicas, algoritmos y conocimientos en común.



AthenaTech LLC, 2019. AI, Machine Learning (ML) and Natural Language Processing (NLP). Recuperado

Ilustración 6.85 ML y PNL

Actualmente existen diferentes campos en los que se puede utilizar NLP. Entre los que se pueden mencionar:

- Reconocimiento del habla,
- Análisis de Sentimientos,
- Sistemas de Preguntas y Respuestas,
- Generación automática de resúmenes,
- Chatbots,
- Inteligencia de Mercado,
- Clasificación automática de textos,
- Revisión automática de gramática.

Ahora, es muy fácil encontrar modelos de machine learning pre-entrenados que facilitan a los desarrolladores utilizar NLP en diferentes aplicaciones.

Sin embargo, en el caso de que se quiera desarrollar un modelo con un conjunto de datos nuevos, sin utilizar un modelo pre-entrenado, dependiendo de la cantidad de datos, puede llevar bastante tiempo obtener un buen resultado.

PNL ya está bastante presente en el ámbito de salud: Chatbots, Análisis de artículos en Investigaciones Médicas, Codificación automatizada en base a la información contenida en estudios y documentos clínicos....

6.4.4. MLOps (Machine Learning Ops)

MLOps (Machine Learning Ops)

MLOps es un conjunto de mejores prácticas a la hora de desarrollar modelos de machine learning.

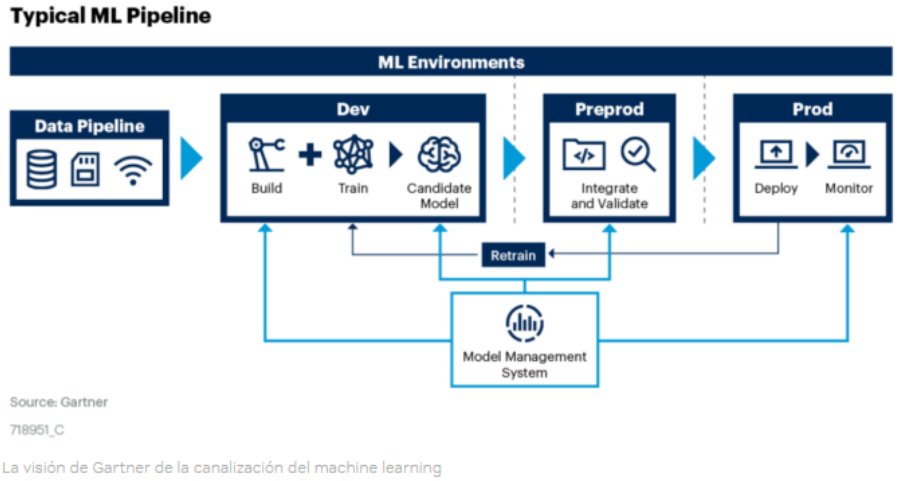


Ilustración 6.86 Visión de Gartner de la canalización del Machine Learning

Permite la colaboración y comunicación entre todos los implicados en el ciclo de vida del desarrollo de analítica avanzada. Aquí incluiría desde los usuarios de negocio, hasta los Data Scientists y los ingenieros de IT necesarios para el desarrollo de los modelos analíticos, provocando la agilización del proceso completo.

Seleccionan conjuntos de datos y crean modelos de inteligencia artificial que los analizan, para luego ejecutarlos a través de los modelos creados, de manera disciplinada y automatizada.

Ilustración 6.87 MLOps combina ML con desarrollo de aplicaciones y operaciones

Llevar a cabo un proceso de machine learning conlleva muchos pasos muy complejos:

- Ingesta de datos, preparación, entrenamiento de modelos y ajuste e implementación
- Supervisión de los modelos y su explicabilidad
- Coordinación de especialistas en ciencia de datos e ingenieros de ML.
- Rigor operativo para mantener todos los procesos sincronizados y trabajando a la par.

MLOps abarca todo este ciclo con el fin de que el proyecto llegue a buen puerto lo más rápidamente posible.

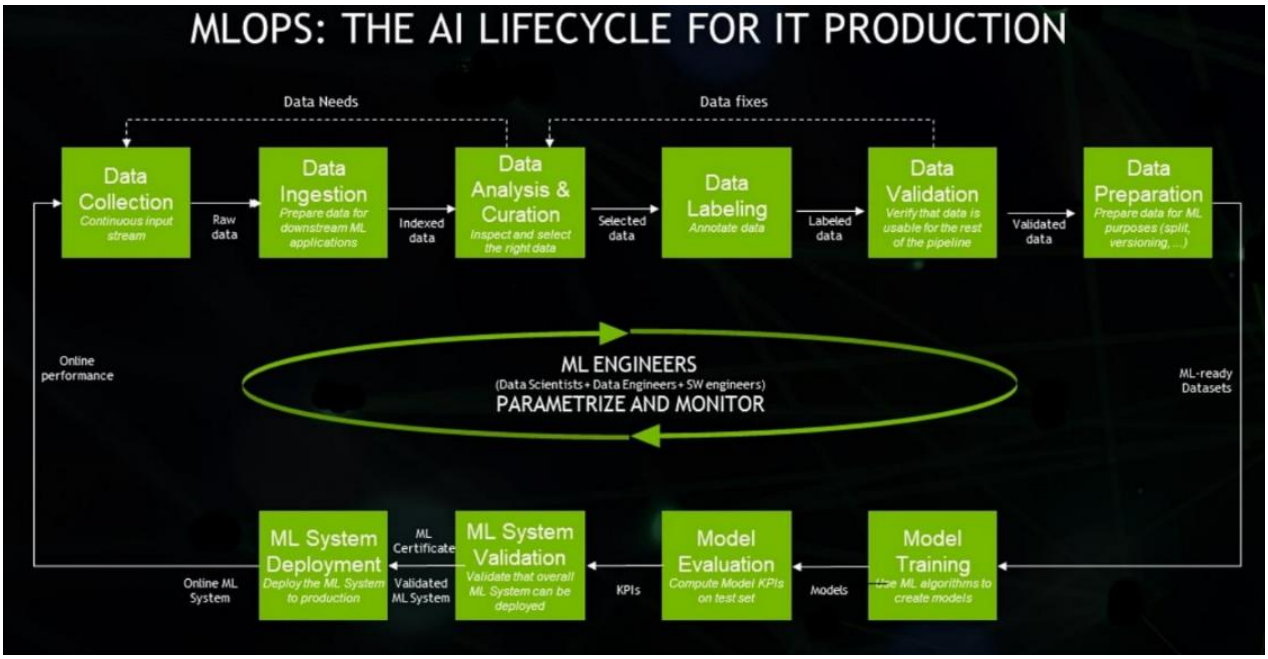


Ilustración 6.88 MLOps: The AI lilficycle por IT Production

6.5. Organización del Data Lake y Herramientas necesarias.

Existen multitud de proposiciones de arquitecturas a la hora de montar un Data Lake, muchas veces, influenciadas por requerimientos propios del sistema en el que se vaya a montar o por la necesidad de utilizar determinadas herramientas. Aquí se pretende mostrar un enfoque generalista que pueda adaptarse a las necesidades finales de la institución. Es por ello, que se muestra una arquitectura de Capas Lógicas que aseguren un flujo seguro de datos hasta llegar a la analítica y la capa final de visualización.

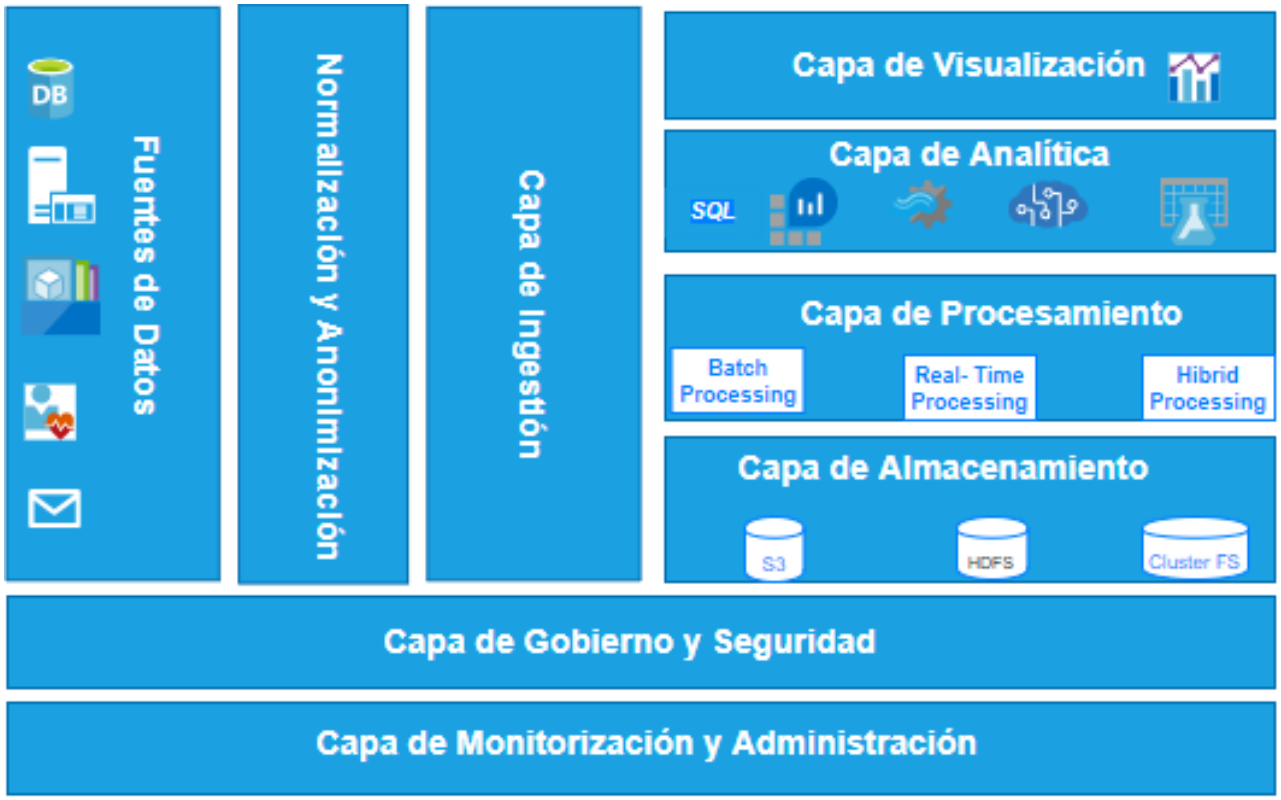


Ilustración 6.89 Arquitectura de Capas Lógicas para un Data Lake

Capas Lógicas que componen la Arquitectura del Data Lake:

- Capa de Normalización y Anonimización
- Capa de Gobierno del Dato y Seguridad
- Capa de Ingestión
- Capa de Almacenamiento
- Capa de Procesamiento
- Capa de Analítica
- Capa de Visualización
- Capa de Monitorización y Administración

6.5.1. Capa de Normalización y Anonimización

Esta capa previa a la ingestión de datos, hará que la información procedente de nuestras fuentes de datos, llegue al Data Lake habiendo pasado por un proceso de normalización y anonimización, cumpliendo así con las premisas necesarias para hacer una analítica avanzada con datos seguros y de calidad, que permita también la provisión de datos a sistemas externos.

Normalización:

Como hemos visto en el apartado dedicado a la Analítica Avanzada, obtener buenos resultados, depende en gran parte de la calidad de los datos. La viabilidad del uso del Machine Learning depende fundamentalmente de la calidad en el proceso de recopilación de los datos y de la fiabilidad de estos, rigiéndose por el principio de GIGO (garbage in, garbage out), si la calidad de lo que ingresa no es buena, el resultado normalmente tampoco es bueno. Sin este peldaño dentro de la IA, sería imposible obtener insights interesantes y acertados, que vienen tras el trabajo de un gran volumen de datos y la automatización.

Normalizar datos consiste en aplicar la misma norma a todos los datos que tenemos de diversas fuentes. Esto implica usar un único patrón a la hora de clasificarlos. Así evitamos la redundancia y le damos más valor a los datos que extraemos del mundo real.
Aquí nos servimos del trabajo con gran volumen de información, en tiempo real, a gran velocidad y en formatos muy diferentes. La normalización es clave porque nos permite aprovechar todo el potencial de la Analítica Avanzada de forma eficiente.

El trabajo que hay que llevar a cabo en la parte de normalización es tremendo. Gran parte de esta tarea, en un trabajo funcional puro y duro, que consume una gran cantidad de tiempo y recursos en los proyectos de análisis de datos.

Ya se ha hablado en el punto 4.5.3.3 de la Interoperabilidad de los Datos, el cumplimiento de los principios FAIR, la necesidad de normalización para la comparación y compartición de resultados con sistemas externos. También se describen los distintos estándares disponibles para la persistencia de datos, servidores de terminología, servidores de modelos de arquetipos: Open EHR, OMOP, i2b2, ISO13606, HL7 FHIR, CDISC.

Algunas herramientas que pueden ayudar en la ardua tarea de la normalización de datos son:

OHDSI: Esta herramienta de software libre en continua mejora, para la extracción de datos y transformación a OMOP CDM, así como para el análisis de datos ya normalizados en CDM. Ya ha sido descrita en el punto 4.5.3.3. “Interoperabilidad de los Datos”

I2b2: Herramienta de software libre para dar soporte a la normalización de datos de acuerdo al modelo i2b2 y para la investigación basada en este modelo/herramienta. Permite compartir, integrar, estandarizar y analizar datos heterogéneos de la atención médica y la investigación.

Ilustración 6.90 Herramientas y Estándares Normalización

Anonimización o seudoanonimización:

Las técnicas de anonimización y seudonimización de datos pretenden reducir la identificabilidad de los datos que pertenecen a una persona a partir de un conjunto de datos original determinado y descomponerlos a un nivel que no supere el umbral de riesgo preestablecido. Se realizará una primera anonimización en esta capa, pudiendo repetirse en el paso de desarrollo del análisis avanzado, un segunda anonimización-seudoanonimización, orientada a los resultados del mismo.

La Agencia Española de Protección de datos, en su Guia Básica de anonimización del 31 de marzo de 2022, en el Anexo E, hace referencia a una lista de herramientas open-source y propietarias, destinadas a tal fin:

Ilustración 6.91 Herramientas de anonimización

6.5.2. Gobierno del dato y Seguridad

Ilustración 6.92 Gobierno del dato y seguridad

El gobierno del dato es un eje fundamental en el éxito de un Data Lake en una organización. Es una labor transversal y de muy amplio espectro que requiere de la definición de una estrategia, objetivos, recursos humanos, procesos, tecnologías, regulación, etc.

El objetivo del gobierno de datos es mantener datos de alta calidad que sean seguros y fácilmente accesibles para extraer información de negocio más detallada.

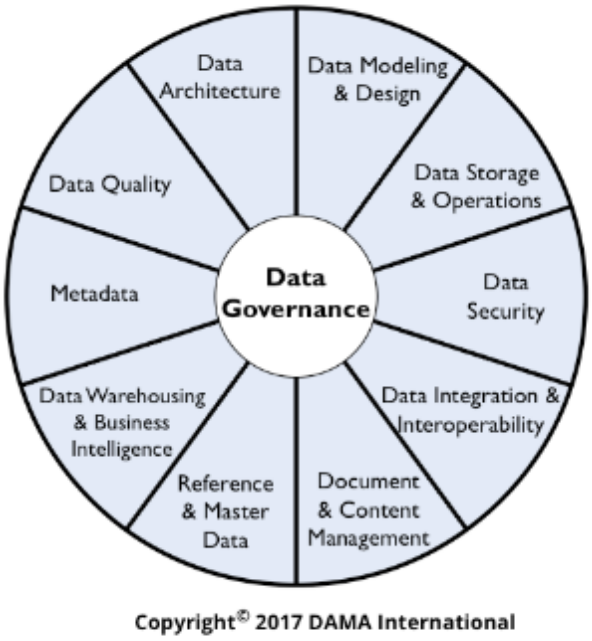


Ilustración 6.93 DAMA-DMBOK2 Data Management Framework

Los metadatos, son un aspecto clave en las herramientas de Gobierno de Dato: DAMA-I enumera en el DMBOK2 los principios de la gestión de los datos; entre ellos, destaca: **"se necesitan metadatos para gestionar los datos"**.

Catálogo de datos. Este permite a analistas de datos, científicos de datos, administradores de datos y otros profesionales de datos con acceso a datos corporativos, buscar a través de todos los activos de datos disponibles de una organización y ayudarse a sí mismos a obtener los datos más apropiados para sus fines analíticos o comerciales.

A lo hora de buscar una herramienta de Gobernanza de datos las **funcionalidades deseables** son las siguientes:

- **Gobierno del dato:** capacidades de flujo de trabajo predefinidas, flujos de trabajo configurables y personalizables, gestión de políticas, alertas y notificaciones y funciones de gobierno personalizables
- **Glosario de Negocio:** Registro de conceptos, identificación de relaciones entre ellos y visualización jerárquica de los mismos.
- **Perfilado de los datos,** permitiendo limpieza, filtrado y estandarización, así como la obtención de valores de tablas maestras.
- **Linaje de los datos**, de forma que se pueda conocer el flujo por el que estos han pasado, desde su origen y las transformaciones sufridas, antes de ser puestos a disposición para su consumo.
- **Versionado de los datos**, sobre los que se realizan las transformaciones, con el objeto de poder reutilizar los datos para los diferentes casos de uso que se desplieguen
- **Clasificación y etiquetado** de los datos.
- **Establecimiento de ontologías**, diccionarios o datos maestros, que faciliten las transformaciones y validaciones para garantizar la calidad del dato y su estandarización.
- **Descubrimiento de los datos** de forma que los usuarios puedan conocer qué datos tienen disponibles para el consumo.
- **Calidad del dato**, dimensionando aspectos como la disponibilidad, usabilidad, confiabilidad, pertinencia y la calidad de presentación
- **Seguridad** que centralice y facilite la gestión de la seguridad perimetral, gestión de autenticación, autorización y auditoría. Creación y difusión de políticas y reglamentos para realizar una gestión de políticas de autorización y auditoría de acceso de usuarios basada en roles y con una alta granularidad en la asignación de permisos a nivel de campos o metadatos específicos.
- **Anonimización:** Herramientas/métodos que permitan la anonimización, seudo-anonimización, enmascaramiento y cifrado de datos.
- **Gestión de servicio de datos:** Gestión de laboratorios de datos, areneros o servicios de datos
- **Integración:** Capacidades de integración con sistemas de gestión de la autenticación y autorización como LPAD, SAML, Kerberos, Oauth2, OpenID Connect

- **Capacidades IA:** Capacidades de AI relacionadas con los objetos del glosario de negocio y del catálogo de datos. Capacidades centradas en la prevención de errores y la mejora del rendimiento. Propuestas basadas en las acciones de los usuarios
- **Dashboard e informes:** Tableros de control personalizables, alertas y alarmas personalizadas y visualización de las series temporales

Algunos de los principales proveedores de dichos servicios de Gobierno del Dato y la Metadatación son los siguientes:

Ilustración 6.94 Algunas Herramientas de Gobierno del Dato

Mapeo de soluciones vs. dimensiones DAMA de AWS:

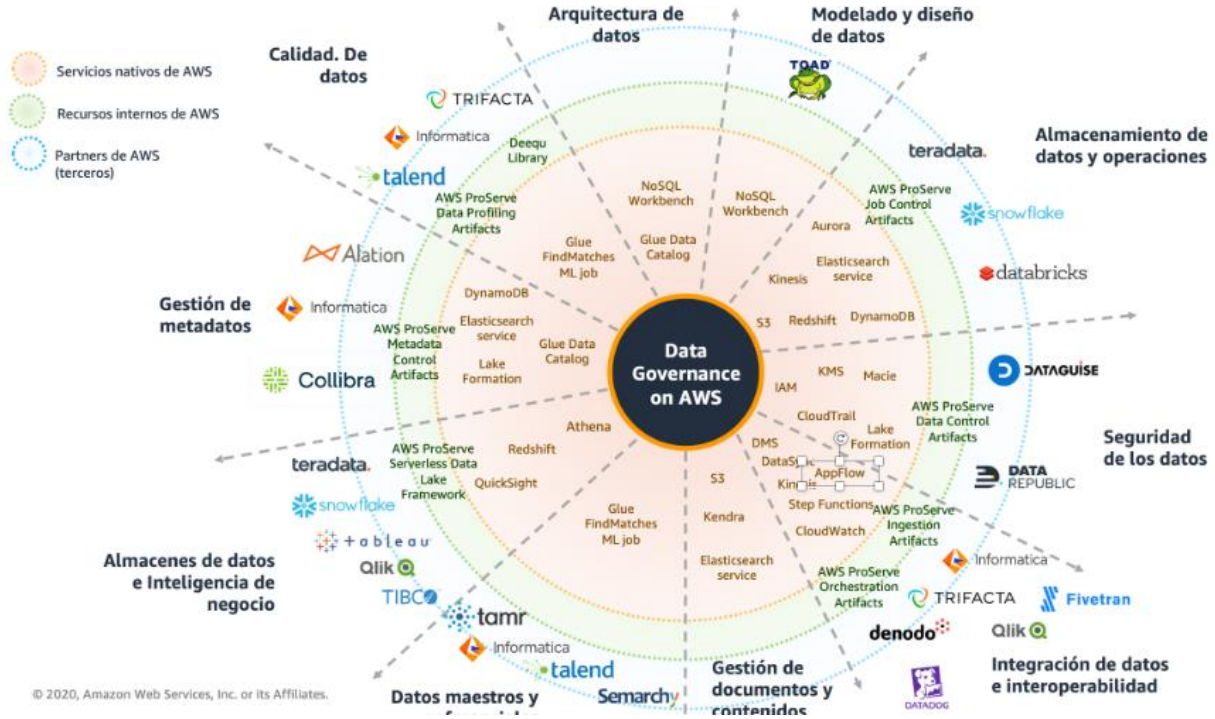


Ilustración 6.95 Mapeo de soluciones vs. Dimensiones DAMA en AWS

6.5.3. Capa de Ingesta

Ilustración 6.96 Tipos de datos en la ingesta

La ingesta de datos es el proceso mediante el cual se introducen datos, desde diferentes fuentes, estructura y/o características, tanto internas al sistema de salud, como externas, admitiendo grandes volúmenes de datos, con destino a nuestro sistema de almacenamiento o procesamiento de datos.

Buscaremos herramientas de **Ingesta** que cumplan con las **funcionalidades**:

- Capacidades de validación, limpieza, transformación y reducción de datos, con destino a las Capas de Almacenamiento.
- Mecanismos de ingesta mediante notificación de eventos para dispositivos IoT.
- Mecanismos de ingesta de procesos planificados mediante técnicas bach, micro-bach y streaming.

Es un proceso muy importante porque la productividad de un equipo va directamente ligada a la calidad del proceso de ingesta de datos. Deben ser flexibles y ágiles, ya que, una vez puesta en marcha, los analistas y científicos de datos puedan construir un pipeline de datos para mover los datos a la herramienta con la que trabajen.

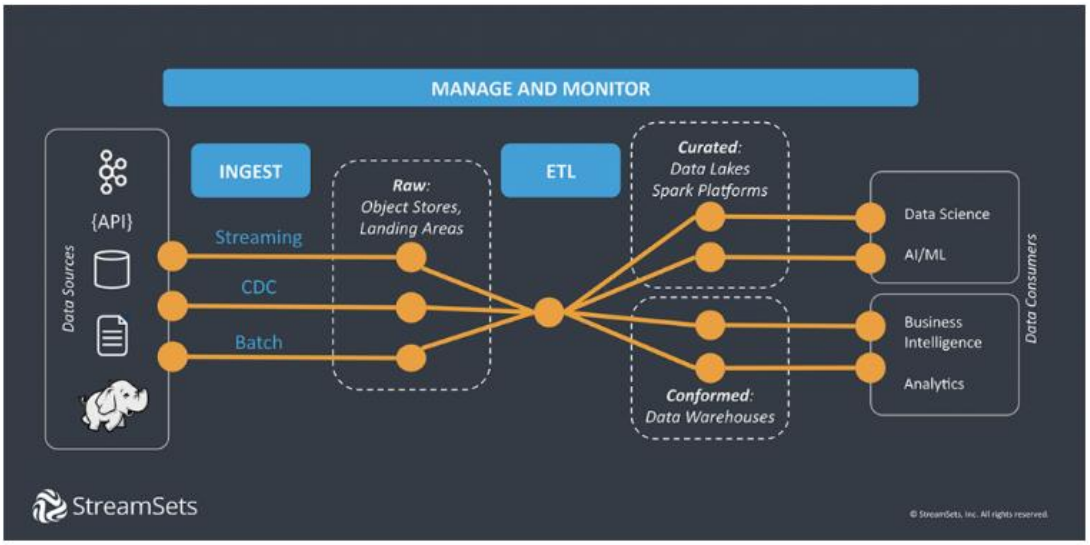
Un **pipeline de datos** es una construcción lógica que representa un proceso dividido en fases. Define el conjunto de pasos o fases y las tecnologías involucradas en el proceso de movimiento y procesamiento de datos.

Las ETLs (Extracción, Transformación y Carga) son un caso particular de pipeline de datos que involucran las fases de Extracción, Transformación y Carga:

- **Extracción:** Recopila los datos de los sistemas originales. Analizar que los datos sean veraces, que contiene la información que se espera, verificando que siguen el formato que se esperaba. Debe ser un proceso rápido, ligero, causar el menor impacto posible. Debe ser transparente para los sistemas operacionales e independiente de las infraestructuras. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.
- **Transformación:** realizar los cambios necesarios en los datos de manera que estos tengan el formato y contenido esperado. Transforma los datos para mejorarlos, incrementar su calidad, integrarlos con otros sistemas, normalizarlos, eliminar duplicidades o ambigüedades. Además, no debe crear información, duplicar, eliminar información relevante, ser errónea o impredecible. Una vez transformados, los datos ya estarán listos para su carga.
- **Carga**: Almacena los datos en el destino.

ELT (Extracción, Carga y Transformación): Técnica de ingestión de datos donde los datos que se obtienen desde múltiples fuentes se colocan sin transformar directamente en un data lake o almacenamiento de objetos en la nube. Con la separación de la extracción y la transformación, ELT permite que los analistas y científicos de datos realicen las transformaciones, ya sea con SQL, Python... Permite automatizar la carga del data lake y la posterior codificación de los flujos de datos.

La ingesta por dentro:



La ingesta de datos - StreamSets

Ilustración 6.97 La ingesta de datos - StreamSets

Algunas de las herramientas de ingesta de datos y sistemas de mensajería con funciones propias de ingesta:

Ilustración 6.98 Herramientas de Ingesta de datos y sistemas de mensajería -ingesta

6.5.4. Capa de Almacenamiento

Albergará la información obtenida en la capa de Ingesta, orientada a la persistencia de datos. Contendrá sistemas de almacenamiento con información tanto estructurada, semiestructurada, como no estructurada.

Según el tipo de estructura final que se use para la implementación de nuestro Data Lake, tendremos distintos sistemas de almacenamiento (de ficheros distribuidos HDFS, almacenamiento de objetos S3, archivo NFS, almacenamiento NoSql).

En cualquier caso, se buscará almacenamiento distribuido, alta disponibilidad, escalabilidad y tolerancia a fallos. Buscaremos capacidad de asignación de recursos de infraestructura y capacidad de procesamiento a demanda, en aquellos conjuntos de datos que requieran una frecuencia de acceso más alta en la realización de determinados procesos o análisis.

Zonas dentro del Data Lake:

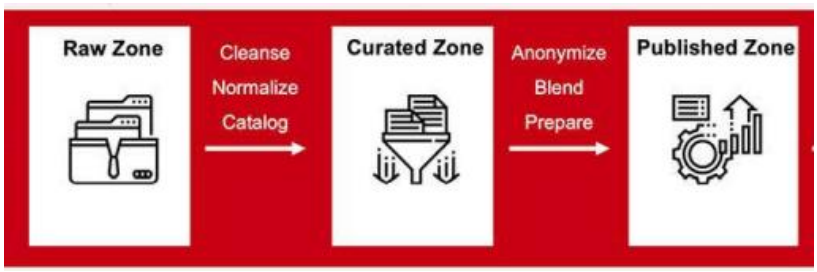


Ilustración 6.99 Zonas dentro del Data Lake

- Raw Zone – Datos en Crudo: Zona de almacenamiento de los datos en su formato nativo, sin ningún tipo de procesamiento.
- Curated Zone – Datos Curados: Zona de almacenamiento de los datos limpios y modificados de su formato original para que puedan ser consumidos en la siguiente zona. Permite agregaciones y cálculos sobre los mismos.
- Published Zone – Datos para Consumo: Zona de almacenamiento de los datos listos para el consumo por los analistas, data scientist, y usuarios permitidos, en las aplicaciones y servicios analíticos que los requieran.

El tipo de almacenamiento podrá ser segregado de la infraestructura de cómputo, o bien de almacenamiento y procesamiento unificado en la misma infraestructura.

Se buscarán herramientas que permitan la provisión de datos a sistemas externos.

6.5.5. Capa de Orquestación y Procesamiento

Capa encargada de la orquestación de procesos y el tratamiento de los datos. Las características deseables, que buscaremos en las herramientas que nos permitan el procesamiento y la orquestación, serán:

- Procesamiento de datos en batch y en streaming.
- Procesos ETL (extracción, transformación y carga).
- Construcción de flujo de datos y orquestación de procesos. Dentro de estos procesos necesitaremos capacidades de creación de flujos de control, preparación de los datos, evaluación de calidad, perfilado y correlación de múltiples fuentes.
- Capacidad de monitorización de los procesos, con estado y resultados de ejecución.
- Capacidad de inclusión de programas, procedimientos almacenados, uso de scripts, en los pasos de transformación de los procesos ETL.
- Capacidades de conexión e ingesta de distintas fuentes de datos (Hive, Impala, Oracle, MongoDB, PostgreSQL, Habase, S3....)
- Interfaz gráfica que permita de forma ágil y sencilla, permita la gestión y modelado de la orquestación de procesos, diseño de los mismos y gestión de flujos.
- Componentes de procesamiento distribuido.

Ejemplos de algunas herramientas de orquestación y procesamiento:

Ilustración 6.100 Herramientas de Orquestación y Procesamiento

6.5.6. Capa Analítica

Concentra las capacidades analíticas del Data Lake. Permite la construcción de modelos analíticos avanzados, cuadros de mandos e informes

Buscaremos herramientas:

- Para la construcción de modelos analíticos avanzados, con el ciclo de vida de las técnicas de Inteligencia Artificial, de los que ya hemos hablado en el punto 6.
- Aplicación de técnicas estadísticas para la preparación de los datos: construcción, entrenamiento, validación y comparación los modelos analíticos. Finalmente, se publican para su consumo.
- Gestión y gobierno del ciclo de vida de los modelos analíticos, incorporando capacidades para su control de acceso, auditoría, control de versionado, publicación automática, monitorización del rendimiento y mantenimiento de modelos en producción.
- Sistemas de publicación de modelos analíticos, mediante APIs a terceros, gestionados desde la plataforma.
- El consumo de los modelos analíticos se podrá realizar también a través de las capacidades que proporciona la capa de visualización y provisión.
- Interpretación y análisis de sesgo para los modelos de inteligencia artificial que se desarrollen
- La plataforma deberá proporcionar las técnicas de analítica avanzada más habituales:

    - Data Mining y Text Mining
    - Aprendizaje automático (Machine Learning)
    - Aprendizaje profundo (Deep Leaning)
    - Patrones y tendencias
    - Visualización y simulación
    - Análisis semántico
    - Estadística multivariada

  - Análisis de Grafos
  - Clustering y algoritmos de clasificación
- Herramientas o módulos de ciencia de datos, con componentes como Jupyter, Apache Zeppelin y Tensorflow.
- Debe permitir la incorporación de algoritmos propios desarrollados en R, Python, Java o Scala
- Permitir la construcción de estructuras de análisis multidimensionales, con acceso de baja latencia para ofrecer funcionalidad OLAP. Puede incluir un intérprete SQL propio para esta funcionalidad.

Ejemplos de herramientas de uso para el análisis avanzado

Ilustración 6.101 Herramientas de Analítica

6.5.7. Capa de Visualización

Visualización analítica de información, una vez está almacenada y procesada. Las facultades que debemos buscar en la herramienta usada en la Visualización son:

- Mecanismos de provisión de acceso a los conjuntos de datos (DataSets).
- Construcción de visualizaciones gráficas (grafico de barras, líneas, mapas de calor, geo mapas) que se podrán agrupar para realizar cuadros de mando para ser consumidos de forma individual, por aplicaciones o servicios de terceros en forma de componente visual.
- Construcción de informes que podrán ser distribuidos. Mediante herramienta gráfica se podrán construir y planificar su ejecución.
- Facilidades para la interacción con datos almacenados en el data lake mediante SQL, Hive, SparkSQL, Impala, Presto.... mediante una interfaz visual para la consulta y exploración de los datos.

Algunas herramientas de visualización de datos analíticos:

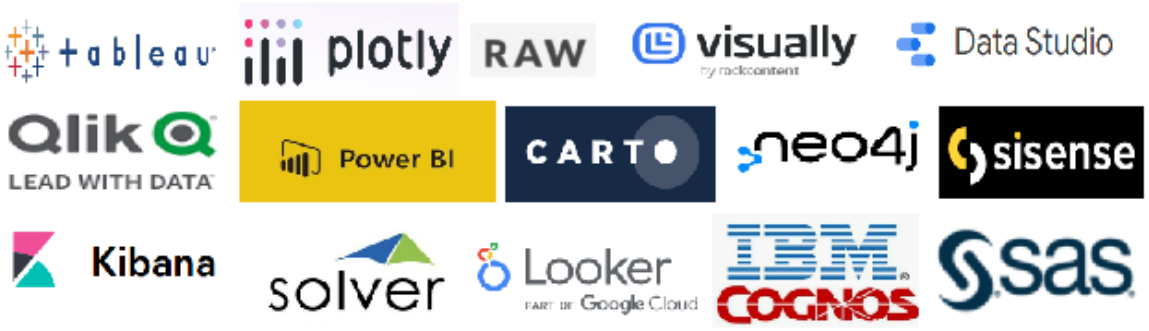


Ilustración 6.102 Herramientas de Visualización

6.5.8. Capa de Administración y Monitorización del Data Lake

Las características que buscaremos en las herramientas que nos permitan administrar y monitorizar el Data Lake son:

- Herramienta visual de administración que permita realizar acciones de gestión y operación sobre los diferentes nodos que conforman el clúster, sobre el propio clúster, o sobre la plataforma de orquestación de contenedores, en su caso.
- Capacidades de monitorización, de forma que se pueda conocer en todo momento, el estado de salud de la infraestructura y de la propia solución, en cuanto a recursos y estado de sus componentes, incorporando herramientas con interfaz de usuario que faciliten las diferentes actividades de monitorización.
- Monitorizar a nivel de clúster, host, servicios, tenants, y cargas de trabajo.
- Consultas de parámetros habituales como uso de CPU, memoria, parámetros de máquina virtual o contenedor, estado de almacenamiento, indicadores de I/O, througthput, o tasas de transferencia.
- Configuración de alertas de forma que se pueda notificar a los administradores cuando ciertos indicadores superen los umbrales configurados
- Configurar el registro de eventos de seguridad para dar cumplimiento a los requisitos de seguridad y protección de datos estipulados.
- Acceder a los logs de la plataforma para realizar tareas de troubleshooting.
- Distribuir de forma óptima y eficiente cargas de trabajo de “ejecución larga” como las generadas por lo módulos y herramientas de ciencia de datos, que requieren un alto tiempo de disponibilidad de recursos del clúster, tanto de forma manual como automática.

Se buscará una plataforma que sea compatible para su funcionamiento, tanto en entornos “on-premise”, como sobre Cloud Híbrida y Cloud Pública, buscando también la compatibilidad con el menos AWS, Microsoft Azure y Google Cloud.

6.6. Aplicación sobre los casos de uso

En el punto 5 se ha desarrollado y detallado el caso de uso que nos ocupa sobre “Riesgo CardioVascular”.

En el mismo se incluye, una definición, una primera aproximación a los resultados esperados y una valoración preliminar del impacto en el servicio prestado que supondría la integración del caso de uso en la operativa del servicio de salud. El éxito a la hora de implementar el caso de uso, dependerá de la calidad, disponibilidad, representatividad o profundidad histórica de los datos. Se definirán los indicadores que permitan valorar si se responde con éxito al problema que se va a plantear en el caso de uso.

6.6.1. Gobierno del dato para nuestro caso de uso

En consonancia con el gobierno del dato adoptado en nuestro Data Lake, daremos los siguientes pasos:

- Definición de las ontologías, diccionarios o datos maestros del caso de uso “Riesgo Cardiovascular”.
- Definición de las políticas de custodia, calidad y perfilado y linaje de los datos.
- Establecer los pasos para la repetibilidad y automatización de todo el ciclo de vida de desarrollo y despliegue del modelo analítico que dé respuesta a nuestro caso de uso.
- Para la construcción de este modelo, se definirán y establecerán los procedimientos operativos y elementos aceleradores (si se proponen algoritmos de partida, para acelerar el desarrollo del modelo analítico) y despliegue de los modelos, así como su monitorización y mantenimiento, una vez desplegados.
- Definición de las medidas de seguridad y privacidad: autorización, acceso, auditoría de los datos...
- Valoración de la viabilidad del caso de uso.
- Propuesta de planificación para el desarrollo y despliegue del caso de uso del proyecto

6.6.2. Ingesta de fuentes al Data Lake para el caso de uso.

1.- Localización de la información, desde los sistemas origen:

Según la información aportada en la definición del modelo, hemos identificado las siguientes fuentes de datos que serán necesarias para el análisis dentro de nuestro Data Lake

Procedentes de los distintos Sistema Información Hospitalaria:

- Datos demográficos del paciente.
- Ingresos Hospitalarios por Diagnóstico.
- Indicadores de Mortalidad Hospitalaria por Diagnóstico
- Resultados analíticos.
- Resultados de dispositivos IoT de Monitorización Cardíaca
- Resto de variables clínicas incluidas en la descripción del caso de uso.

Disponemos de datos históricos de los pacientes en nuestros sistemas de información hospitalaria, desde su puesta en marcha (10 años)

El conjunto de datos estará compuesto las variables correspondientes a la "población estudio" y a la "población de control".

Población Estudio: Serán pacientes, mayores de 18 años, con episodios codificados con diagnóstico CIE-10 de Patología Cardiovascular, y que cumplan con los indicadores de riesgo en los dispositivos IoT de monitorización cardíaca.
En nuestra BBDD, tenemos un total de 3000 pacientes cumplan con RCV (Patología y dispositivo IoT)

Población de Control: Serán pacientes que no tengan episodios con diagnósticos de Patología Cardiovascular y sin otras patologías asociadas, emparejados por edad y sexo, mediante un calibrador de vecindario de 5 años. Tenemos un total de 12000 pacientes

2.- Diseño e implementación de los flujos de planificación y sistematización de procesos de ingesta.

3.- Análisis inicial de las fuentes de datos.

4.- Descripción y perfilado de los datos.

5.- Evaluación de la calidad de los datos.

6.- Anonimización y Normalización de los datos.

7.- Realización de la ingesta a la capa cruda del Data Lake, en su formato nativo.

6.6.3. Actividades asociadas al desarrollo y despliegue del caso de uso.

Aplicación los siguientes modelos en el proceso de desarrollo:

- Modelo estándar CRISP-DM (Cross Industry Standard Process of Data Mining). Establece las fases necesarias para cubrir el ciclo de vida propio de las técnicas de minería de datos,

que son comúnmente aplicadas al desarrollo de modelos analíticos.

- Aplicación de MLOps (Machine Learning Ops) conjunto de mejores prácticas a la hora de desarrollar modelos de machine learning.

- Modelo SCRUM de entregas continuas, operativas, iterativas e incrementales, con mejoras continuas entre una entrega y otra (SPRINT) hasta completar el caso de uso.

Elaboración de un Plan de Trabajo, donde se recojan: Objetivo de negocio para la evaluación del caso de uso, Objetivos analíticos, Complejidad, Pasos a seguir y Técnicas analíticas a emplear para el desarrollo e implementación del caso de uso.

Conocimiento de los datos: Recopilación y descripción de los datos: se recogerán los datos iniciales del caso de uso y se adecuarán para el futuro procesamiento. Elaboración de informes con la lista de datos adquiridos, su localización, técnicas utilizadas en su recolección, volúmenes de datos, significado y formato de cada campo.

Metodología de aplicación de un modelo de aprendizaje automático:

- 1.- Obtención y preparación de los datos.
- 2.- Selección de las variables.
- 3.- Datos de entrenamiento, validación y Test.
- 4.- Ajustes de los algoritmos.
- 5.- Validación y selección del modelo.
- 6.- Test final del modelo.
- 7.- Utilización del modelo.

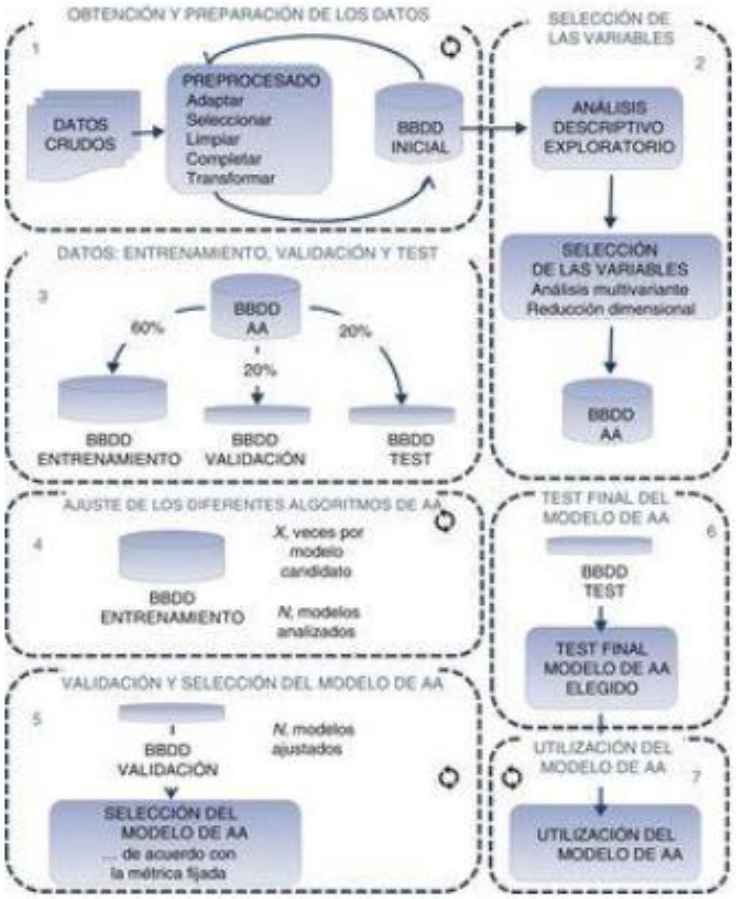


Ilustración 6.103 Metodología de aplicación de un modelo de aprendizaje automático

1. Obtención y preparación de los datos: se seleccionará el subconjunto de los datos adquiridos en la fase anterior compuesto por los pacientes anteriormente descritos como Población Estudio y Población Control (15000 pacientes en total)

    a. **Limpieza de datos:** se optimizará la calidad de los datos, al objeto de prepararlos para la fase de modelado analítico, mediante técnicas de normalización de datos, la discretización de datos numéricos, el tratamiento de valores ausentes o la reducción del volumen de datos, entre otras.

    b. **Estructuración de los datos:** se realizarán operaciones de preparación de los datos, como generación de nuevos atributos, a partir de los ya existentes o a transformación de valores para atributos existentes.

    c. **Integración de los datos:** se combinarán los datos que se encuentren en diferentes fuentes para tener una vista unificada de los mismos.

d. **Formateo e integración de los datos:** se realizarán las transformaciones sintácticas de los datos, sin modificar su significado, con la idea de permitir o facilitar el empleo de alguna técnica de modelado analítico particular.

2. Selección de las Variables: En este punto obtendremos las variables que nos servirán como base para la realización de nuestro estudio

   a. Análisis Descriptivo Exploratorio: Estudio de las variables origen, con el objetivo de seleccionar las más adecuadas para que formen parte del modelo
   b. Selección de las variables: Análisis multivariante y Reducción Dimensional
   c. Construcción del Data Set: Construcción del conjunto de datos, con las variables seleccionadas. En nuestro caso, tendremos variables etiquetadas con RCV SI o NO, Exitus=SI o NO, Reingreso: SI o No, en función de las condiciones anteriormente citadas

3. División del conjunto de datos en Conjuntos: Entrenamiento, Validación y Test (60 %, 20%, 20%)

POBLACION		**Entrenamiento (60%)**	**Test (20%)**	**Validación (20%)**
Estudio	**3000**	1800	600	600
Control	**12000**	7200	2400	2400
Total	**15000**	9000	3000	3000

4. Ajuste de los diferentes Algoritmos de Analítica Avanzada (en adelante AA): Utilizaremos el conjunto de datos de entrenamiento para ajustar los algoritmos seleccionados. Puede ser necesario el uso de técnicas de submuestreo, si hay una categoría con muchos más casos que otra. Ningún algoritmo, en principio es mejor que otro, su capacidad para realizar un buen ajuste, dependerá de las características de nuestros datos (número de variables, linealidad, normalidad, valores faltantes, variables continuas o categóricas, etc…)

En nuestro caso, probamos con distintos algoritmos de **Machine Learning** de aprendizaje automático **supervisado**, de **clasificación** (Árboles de decisión, Regresión Logística, SVM, Naive Bayes, KNN). En base a los resultados obtenidos, nos decantamos por una **Regresión Logística**, con un Intervalo de confianza del 92%

5. Validación y Selección del modelo de AA, de acuerdo con la métrica fijada. Usaremos el conjunto de datos de Validación para evaluar la calidad del modelo. Para ello trataremos de maximizar la métrica que más interese para nuestro caso particular: área bajo la curva de ROC, exhaustividad, precisión, exactitud u otras.
   Es normal que el proceso de Entrenamiento – Validación se repita un número de veces

aleatorizando ambos subconjuntos (k-folds), con el objetivo de optimizar los parámetros internos del algoritmo y evaluar la robustez.

6. Test final del modelo AA: Usaremos el conjunto de datos de test para comprobar que el modelo final se comporta según lo previsto con los datos que o han sido usados para su construcción y validación. Si el resultado difiere del obtenido en la validación, es probable que el set de datos de entrenamiento sea insuficiente y haya que ampliarlo.

7. Utilización del modelo AA

Despliegue: Plan de implantación: Se diseñará y elaborará el plan de implantación, incluyendo las tareas para la publicación o consumo del modelo analítico. Los resultados del modelo se almacenarán en el data lake, desde donde podrán ser consultados por los usuarios finales, o consumidos por otros sistemas o servicios.

- Desarrollo de APIs que permitan la publicación del modelo analítico, por parte de otros sistemas de información de la organización.
- Construcción de informes y visualizaciones ad-hoc, que faciliten la presentación de resultados y su consumo por parte de los usuarios.

Plan de monitorización y mantenimiento del modelo analítico: diseño y elaboración de estrategias de monitorización y mantenimiento del modelo, una vez desplegado, que permitirán valorar si el modelo está funcionando y siendo utilizado apropiadamente. Incluirá también las tareas de mantenimiento y control de versiones del modelo.

Revisión del proyecto analítico: evaluación de todo el proyecto de desarrollo y despliegue, identificando tareas aprendidas y acciones de mejora.

Generación del informe final o de conclusiones: Se generarán los Informes asociados al conocimiento de los datos.

- Construcción y uso del modelo analítico: proceso de elaboración, fundamentos, resultados, uso, consumo y conclusiones del modelo.
- Plan de implantación, monitorización y mantenimiento.
- Lecciones aprendidas y acciones de mejoras, asociadas a la fase de revisión del proyecto.
- Capacitación asociada al caso de uso.
- Elaboración de manuales y guías de usuario.

6.7. Síntesis de Conclusiones del Capítulo

6.7.1. Analítica Avanzada.

El poder de la analítica avanzada en general, y particularmente aplicada al ámbito de salud, es enorme. Disponemos de modelos de madurez y de ciclos de vida que sirven de guía para llegar a buen puerto, como KDD y CRISP-DM. Se sigue avanzando en utilidades que permitan la integración del Machine Learning con desarrollo de aplicaciones y operaciones como MLOps.

La teoría está bien definida respecto a los modelos analíticos existentes, y podemos ver cómo puede aplicarse en el ámbito sanitario, bien creando modelos predictivos en la detección de patologías potencialmente perjudiciales, ya sea a través de datos clínicos del paciente, imágenes radiológicas, resultados analíticos, o bien ayudando en las labores de documentación clínica con métodos PNL.

En todo esto, no podemos olvidar que para que estas técnicas hagan su labor, debemos partir de unos datos con calidad y cantidad suficiente para el entrenamiento de los modelos.

6.7.2. Organización del Data Lake y Herramientas.

Cuando pensamos en un Data Lake, en lo primero en lo que se piensa es en la capacidad de procesamiento, en el almacenamiento y la ingesta de datos. En todos estos aspectos, la oferta y distintas disposiciones a la hora de montarlo es amplia. Tenemos claro que necesitaremos sistemas de almacenamiento adecuados a los distintos tipos datos (estructurados, no estructurados y no estructurados), que la capacidad de procesamiento debe adaptarse a las necesidades del momento en el Data Lake, así como ingesta y procesamiento tanto en batch como en streaming.

Además, es necesario pensar en la ubicación “on-premise”, cloud, o sistemas híbridos, y en la administración del conjunto.

Tendremos que elegir unas herramientas de Visualización adecuadas y poner a disposición de los usuarios un acceso seguro.

Sin embargo, hay cuestiones que se deben tener en cuenta, y que no parecen estar tan presentes como deberían, y que son cruciales: La implementación de un modelo de Gobernanza y Seguridad, así como datos de origen de calidad normalizados y anonimizados.

Quizá sea esto último, lo que suponga una mayor dificultad a lo hora de llevar a cabo la implementación de un Data Lake

Necesitamos de este modelo de gobernanza, que nos asegure la calidad del dato, la trazabilidad y explicabilidad los modelos de analítica avanzada en nuestro Data Lake.

6.7.3. Aplicación del caso de uso.

Para desarrollar nuestro caso de uso, nos debemos basar en el modelo de gobierno del dato definido en nuestro Data Lake, desde el paso de la ingesta, hasta el procesamiento y publicación del mismo.

Los resultados de cualquier modelo que se pueda desarrollar dependerán desde el principio de la calidad de los datos de origen, de la elección de las variables correctas, y de la realización de las preguntas adecuadas. Para cumplir con estas premisas, es necesaria la participación activa y la colaboración de científicos de datos y de los expertos de salud.

6.7.4. Conclusiones del Capítulo III

Tecnológicamente estamos más preparados que nunca para abordar proyectos de Inteligencia Artificial, Analítica Avanzada, Machine Learning. Más cerca que nunca para desarrollar la tan ansiada Medicina Personalizada a través de Data Lakes. Pero esta tecnología, necesita de unos datos de calidad, normalizados e interoperables que no están tan alcance como nos gustaría.

A lo largo del desarrollo de la creación de este capítulo del TFM, me he encontrado con infinidad de soluciones comerciales y open-source, muchas propuestas de creación de Data Lakes y explotaciones masivas de datos, de las que muy pocas se han hecho realidad, porque se han quedado en el camino de la creación, o no han dado los resultados esperados. Mucha teoría y muy poca puesta en práctica.

Los que lo consigan, serán aquellos que tengan implantada una normalización de la información no solo para su uso primario de datos, sino también para para el uso secundario de los mismos. Tendrán un modelo de Gobierno del dato bien definido, con certeza de datos de calidad y una política de seguridad y acceso a los datos bien establecida. Serán estos, los que consigan que la información esté disponible para la mayor cantidad de usuarios posibles. Y será entonces, cuando tengamos modelos de éxito, y se pueda extender el uso de los Data Lakes y su aprovechamiento con la Analítica Avanzada.

La Analítica Avanzada, con los datos adecuados, variables correctas y las preguntas adecuadas, es muy potente. Parte de su potencia reside en la posibilidad de compartir los resultados del entrenamiento de sus modelos con otros sistemas externos, así como hacer uso propio de modelos ya entrenados.

Será un trabajo duro para muchos de nuestros sistemas de salud, pero estoy deseando que se materialice y ver el resultado que tanto puede aportar a la medicina.

7. Índice de gráficos, tablas e ilustraciones

8. Referencias bibliográficas

- *Fundación 29*. (2022). Obtenido de https://dx29.ai/
- AEPD. (s.f.). *Evalúa-Riesgo RGPD*. Obtenido de https://www.aepd.es/es/guias-y-herramientas/herramientas/evalua-riesgo-rgpd
- Association, W. M. (2016). *Declaration of Taipei on Ethical Considerations Regarding Health Databases and Biobanks.*
- Bioethics, N. C. (2015). *Linking and Use of Data in Biomedical Research and Heatlh care : Ethical Issues.*
- *Bloomberg*. (19 de 09 de 2018). Obtenido de Business: https://www.bloomberg.com/news/articles/2018-09-19/u-s-near-bottom-of-health-index-hong-kong-and-singapore-at-top#xj4y7vzkg
- D. PEREZ-REY, E. S.-O.-O.-J.-M.-B. (2018). Modelos de Datos para la Utilización Secundaria de Historias Clínicas: Experiencia de un Conector OMOP a i2b2. *Researchgate*.
- European Comission, D.-G. f. (08 de 11 de 2019). *Ethics guidelines for trustworthy AI*. Obtenido de https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1
- *FAIR4HEALTH*. (2018). Obtenido de https://www.fair4health.eu/
- FHIR4FAIR, H. (2021). Obtenido de http://build.fhir.org/ig/HL7/fhir-for-fair/
- Galicia. (s.f.). Obtenido de https://abertos.xunta.gal/busca-de-datos
- GO-FAIR. (2017). Obtenido de https://www.go-fair.org/
- Jiménez, M. P. (2022). “Can OpenEHR, ISO 13606 and HL7 FHIR work together? An agnostic perspective for the selection and application of EHR standards from Spain”. *TechRxiv*.
- Jimenez, R. (2017). *Four simple recommendations to encourage best practices in research software*. Obtenido de https://doi.org/10.12688/f1000research.11407.1
- OHDSI. (s.f.). *OHDSI*. Obtenido de https://www.ohdsi.org/software-tools/)
- Porter, T. K. (2006). *Redifining Healthcare.*
- Salud, F. E. (2018). *100 Medidas que mejoran la salud.*
- *The Lancet*. (16 de 10 de 2018). Obtenido de Global Health Metrics: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)31694-5/fulltext
- USA, E. O. (2015). *Big Data: Seizing Opportunities , Preserving Values.*
- Wilkinson, M. D. (2016). *The FAIR Guiding Principles for scientific data management and stewardship.*
- The eight levels of the Analytics Adoption Model (2016) : https://healthcloudsolutions.org/the-eight-levels-of-the-analytics-adoption-model
- Extracción de conocimientos de BBDD (2020): https://www.campusmvp.es/recursos/post/el-proceso-de-extraccion-de-conocimiento-a-partir-de-bases-de-datos.aspx
- Metodología CRISP-DM: https://www.sngular.com/es/data-science-crisp-dm-metodologia/
- Cross Industry Standard Process for Data Mining:https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
- Inteligencia Artificial y Salud. Un caso Práctico (2019) :https://www.juanbarrios.com/inteligencia-artificial-y-salud-un-caso-practico-borrador/
- https://www.revespcardiol.org/es-aplicaciones-inteligencia-artificial-cardiologia-el-articulo-S0300893219302507

- Sckit Learn Extendido (2014): https://medium.com/@chris_bour/an-extended-version-of-the-scikit-learn-cheat-sheet-5f46efc6cbb
- MLOPS (2020):https://artyco.com/que-son-los-mlops/ ; https://la.blogs.nvidia.com/2020/09/08/que-es-mlops/
- OHDSI (2022): https://www.ohdsi.org/software-tools/
- I2b2: https://community.i2b2.org/wiki/pages/viewpage.action?pageId=342684
- Guide to basic anonymisation. AEPD (2022): https://www.aepd.es/sites/default/files/2022-06/guide-to-basic-anonymisation-31-march-2022.pdf
- Como llevar a cabo una selección de herramientas para el gobierno del dato (2021): https://www.damaspain.org/como-llevar-a-cabo-una-seleccion-de-herramientas-para-el-gobierno-del-dato/
- Protege (2022):https://protege.stanford.edu/
- Metadatación (2022): https://metadatacenter.org/
- Herramientas de gobierno del dato: https://ifgeekthen.nttdata.com/es/herramientas-de-gobierno-del-dato
- Gobierno del dato en AWS: https://aws.amazon.com/es/blogs/aws-spanish/herramientas-de-gobierno-de-datos-en-amazon-web-services/
- Ingesta de datos en Big Data: https://aitor-medrano.github.io/bigdata2122/apuntes/ingesta01.html
- Herramientas de orquestación de datos: https://ceupe.com.ar/blog/herramientas-de-orquestacion-de-datos/
- Herramientas de Data Science: https://blog.bismart.com/las-10-mejores-herramientas-de-data-science
- Herramientas para análisis de datos: https://www.octoparse.es/blog/30-mejores-herramientas-de-big-data-para-datos-analisis#div2
- PPT Data Lake SAS (2020): https://contrataciondelestado.es/wps/wcm/connect/293845fa-b08e-4fed-91a1-57ee06f51357/DOC20210119104217PPT_Exp_067-20-SP.pdf?MOD=AJPERES
- PPT Data Lake MedP(2021):https://contrataciondelestado.es/wps/wcm/connect/6c6e5fca-a11d-4ec1-91e0-78220ae2b1d5/DOC20210921143353MPBD_OTAP_Doc_Informativo_210920.pdf?MOD=AJPERES
- https://www.projectpro.io/article/healthcare-machine-learning-projects-with-source-code/508#mcetoc_1firrba3ij

9. Anexos

9.1. Open Data en Salud

Repositorios de datos de salud abiertos identificados en el ámbito regional:

- Galicia: https://abertos.xunta.gal/busca-de-datos
- Asturias: https://transparencia.asturias.es/catalogo-de-datos/-/categories/695009?p_r_p_categoryId=695009
- Cantabria (ICANE): https://datos.icane.es/catalogo
    - Salud: [bajo filtro “Sociedad”] https://datos.icane.es/catalogo?groups=sociedad
- Navarra: https://gobiernoabierto.navarra.es/es/open-data/datos/lista-catalogo
    - Salud [filtrar por “Sanitario” o “Salud”]
- País Vasco: https://opendata.euskadi.eus/inicio/
    - Salud: <No parece haber filtro ni datos>
- La Rioja: https://web.larioja.org/dato-abierto
    - Salud: https://web.larioja.org/dato-abierto/datoabierto?filtros={%22tema_nti%22:%22Salud%22}#listado
- Castilla y León: https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html
    - Salud: https://www.saludcastillayleon.es/transparencia/es/datos-abiertos-sanidad/datos-abiertos-sanidad
- Aragón: https://opendata.aragon.es/datos/catalogo
    - Salud: https://opendata.aragon.es/datos/catalogo/temas/salud
- Cataluña: http://governobert.gencat.cat/es/dades_obertes/
    - Salud:
- Andalucía: https://www.juntadeandalucia.es/datosabiertos/portal/catalogo.html
    - Salud: https://www.juntadeandalucia.es/datosabiertos/portal/dataset?vocab_field_activity_sector=Salud&_vocab_field_activity_sector_limit=0
- Castilla La Mancha: https://datosabiertos.castillalamancha.es/search/type/dataset
    - Salud: https://datosabiertos.castillalamancha.es/search/field_topic/salud-39/type/dataset?sort_by=changed
- Madrid: https://datos.comunidad.madrid/catalogo/
    - Salud: https://datos.comunidad.madrid/catalogo/dataset?groups=salud
- Canarias: https://datos.canarias.es/portal/
    - Salud: https://datos.canarias.es/catalogos/general/organization/consejeria-de-sanidad

9.2. Base jurídica para transferencia de datos a DLS con fines de investigación

Art. 6.1 del Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo (RGPD)

- Art. 8, ap. 1 de la Ley 14/1986, de 25 de abril, General de Sanidad

Se considera como actividad fundamental del sistema sanitario la realización de los estudios epidemiológicos necesarios para orientar con mayor eficacia la prevención de los riesgos para la salud, así como la planificación y evaluación sanitaria, debiendo tener como base un sistema organizado de información sanitaria, vigilancia y acción epidemiológica.

- Art. 1 de la Ley 33/2011, de 4 de octubre, General de Salud Pública

La salud pública es el conjunto de actividades organizadas por las Administraciones públicas, con la participación de la sociedad, para prevenir la enfermedad, así como para proteger, promover y recuperar la salud de las personas, tanto en el ámbito individual como en el colectivo y mediante acciones sanitarias, sectoriales y transversales

- Art. 16, ap. 3 de la Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica modificado por la disposición adicional novena de la LOPDGDD.

El acceso a la historia clínica con fines judiciales, epidemiológicos, de salud pública, de investigación o de docencia, se rige por lo dispuesto en la legislación vigente en materia de protección de datos personales, y en la Ley 14/1986, de 25 de abril, General de Sanidad, y demás normas de aplicación en cada caso. El acceso a la historia clínica con estos fines obliga a preservar los datos de identificación personal del paciente, separados de los de carácter clínico asistencial, de manera que, como regla general, quede asegurado el anonimato, salvo que el propio paciente haya dado su consentimiento para no separarlos.

- Se exceptúan los supuestos de investigación previstos en el apartado 2 de la Disposición adicional decimoséptima de la Ley Orgánica de Protección de Datos Personales y Garantía de los Derechos Digitales.

Artículo 9.2. h) y j) del RGPD

h) el tratamiento es necesario para fines de medicina preventiva o laboral, evaluación de la capacidad laboral del trabajador, diagnóstico médico, prestación de asistencia o tratamiento de tipo sanitario o social, o gestión de los sistemas y servicios de asistencia sanitaria y social, sobre la base del Derecho de la Unión o de los Estados miembros o en virtud de un contrato con un profesional sanitario y sin perjuicio de las condiciones y garantías contempladas en el apartado 3;

j) el tratamiento es necesario con fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos, de conformidad con el artículo 89, apartado 1, sobre la base del Derecho de la Unión o de los Estados miembros, que debe ser proporcional al objetivo perseguido, respetar en lo esencial el derecho a la protección de datos y establecer medidas adecuadas y específicas para proteger los intereses y derechos fundamentales del interesado.

Disposición adicional 17a, ap. 2 b) e) f) y g) de la Ley 03/2018 LOPDGDD. Tratamientos de salud

2 El tratamiento de datos en la investigación en salud se regirá por los siguientes criterios:

- b) Las autoridades sanitarias e instituciones públicas con competencias en vigilancia de la salud pública podrán llevar a cabo estudios científicos sin el consentimiento de los afectados en situaciones de excepcional relevancia y gravedad para la salud pública.

- e) ...