LA EXPLOTACIÓN DE DATOS DE SALUD
Retos, oportunidades y límites

Coordinadores

Javier Carnicero Giménez de Azcárate
David Rojas de la Escalera

Autores

Alberto Andérez González
Juan Díaz García
Fernando Escolar Castellón
Pilar León Sanz

SOCIEDAD ESPAÑOLA DE INFORMÁTICA DE LA SALUD

Este documento ha sido elaborado por la Sociedad Española de Informática de la Salud (SEIS).

Sugerencia de cita de este documento:
– Cita del documento completo: Carnicero J. y Rojas D. (Coordinadores). La explotación de datos de salud: Retos, oportunidades y límites. Pamplona: Sociedad Española de Informática de la Salud, 2016. http://www.seis.es
– Cita del Capítulo I: Carnicero J. y Rojas D. *La explotación de datos de salud: Retos, oportunidades y límites*. En: Carnicero J. y Rojas D. (Coordinadores). La explotación de datos de salud: Retos, oportunidades y límites. Pamplona: Sociedad Española de Informática de la Salud, 2016. http://www.seis.es
– Cita del Capítulo II: Escolar F. *La importancia de la explotación de datos de salud*. En: Carnicero J. y Rojas D. (Coordinadores). La explotación de datos de salud: Retos, oportunidades y límites. Pamplona: Sociedad Española de Informática de la Salud, 2016. http://www.seis.es
– Cita del Capítulo III: León P. *Bioética y explotación de grandes conjuntos de datos*. En: Carnicero J. y Rojas D. (Coordinadores). La explotación de datos de salud: Retos, oportunidades y límites. Pamplona: Sociedad Española de Informática de la Salud, 2016. http://www.seis.es
– Cita del Capítulo IV: Andérez A. *Disposiciones legales aplicables*. En: Carnicero J. y Rojas D. (Coordinadores). La explotación de datos de salud: Retos, oportunidades y límites. Pamplona: Sociedad Española de Informática de la Salud, 2016. http://www.seis.es
– Cita del Capítulo V: Díaz J. *Organización y tecnología para la explotación de la información*. En: Carnicero J. y Rojas D. (Coordinadores). La explotación de datos de salud: Retos, oportunidades y límites. Pamplona: Sociedad Española de Informática de la Salud, 2016. http://www.seis.es

Secretaría Técnica: CEFIC
C/ Enrique Larreta, 5 - Bajo izda. 28036 - Madrid
Tel: 34 91 388 94 78 Fax: 34 91 388 94 79
cefic@cefic.com

ISBN: 978-84-608-8947-2

Índice

Capítulo I

La explotación de datos de salud: Retos, oportunidades y límites

Javier Carnicero Giménez de Azcárate
David Rojas de la Escalera

1. Introducción

Los sistemas de salud de los países occidentales deben hacer frente a la confluencia de varias circunstancias que amenazan seriamente su sostenibilidad, y que por lo tanto les exigen asumir una profunda transformación. Estas circunstancias son las siguientes:

- El envejecimiento de la población, que conlleva un aumento de enfermedades crónicas y degenerativas.
- La crisis económica, que supone la reducción del presupuesto público que se destina a financiar la actividad de los sistemas nacionales de salud.
- El aumento de los costes de las nuevas tecnologías médicas, entre las que se incluyen los medicamentos.
- Las crecientes demandas de los ciudadanos, que exigen la mejora de la calidad de los servicios.

La población de España ha pasado de 42,72 a 46,77 millones de personas entre los años 2003 y 2014. En ese mismo periodo de tiempo, el porcentaje de población mayor de 64 años ha ascendido desde el 17,03% al 18,05% del total de la población, y la tasa de dependencia, que relaciona la población mayor de 64 años con la comprendida entre los 15 y los 64, ha pasado del 24,75% al 26,99%. Por otra parte, el gasto sanitario público en España en 2003 era el 5,37% del PIB, alcanzó un máximo en 2009 del 6,77%, y cayó hasta el 6,26% en 2013. La evolución de estos indicadores durante esos períodos se recoge en la Tabla I.1.

Tabla I.1. Datos demográficos y de gasto sanitario público en España (2003-2014).

Año	Población total	Población 15-64 años	Población mayor de 64 años		Tasa de dependencia	Gasto sanitario público	
			Personas	% sobre total		M€	% PIB
2003	42.717.064	29.396.965	7.276.620	17,03%	24,75%	43.158,4	5,37%
2004	43.197.684	29.777.965	7.301.009	16,90%	24,52%	46.992,4	5,46%
2005	44.108.530	30.511.110	7.332.267	16,62%	24,03%	51.351,5	5,52%
2006	44.708.964	30.849.177	7.484.392	16,74%	24,26%	56.662,2	5,62%
2007	45.200.737	31.188.079	7.531.826	16,66%	24,15%	61.612,0	5,70%
2008	46.157.822	31.869.008	7.632.925	16,54%	23,95%	68.147,1	6,11%
2009	46.745.807	32.145.023	7.782.904	16,65%	24,21%	73.035,6	6,77%
2010	47.021.031	32.153.527	7.931.164	16,87%	24,67%	72.852,6	6,74%
2011	47.190.493	32.082.758	8.093.557	17,15%	25,23%	71.800,0	6,68%
2012	47.265.321	31.980.402	8.222.196	17,40%	25,71%	68.262,9	6,47%
2013	47.129.783	31.718.285	8.335.861	17,69%	26,28%	65.718,5	6,26%
2014	46.771.341	31.281.943	8.442.427	18,05%	26,99%	-	-

Fuentes: Datos demográficos, Sistema de Información Demográfica del Instituto Nacional de Estadística (INE). Datos de gasto sanitario público, OECD Health Statistics 2015.

En cuanto a la percepción de la calidad de la asistencia por parte de los pacientes, el Barómetro Sanitario de 2014[1] indica que la satisfacción media de los encuestados con el sistema sanitario público era entonces de un 6,31 en una escala de 1 a 10, con un 71,1% de las calificaciones en el intervalo 5-8. En 2015 el indicador es similar[2], con un resultado de 6,38 de media y un 71,5% de las calificaciones entre 5 y 8. En cuanto a las listas de espera, se preguntó a los encuestados su percepción sobre la evolución de las mismas. En 2014, un 38% no apreciaba cambios significativos, un 38,9% consideraba que había empeorado la situación y sólo un 7,8% afirmaba haber percibido una mejora. En 2015 estos resultados fueron de un 42,2%, un 33,3% y un 9,6% respectivamente.

Las Tecnologías de la Información y la Comunicación (TIC) se han incorporado de forma desigual al Sistema Nacional de Salud (SNS). Más del 90% de los médicos de atención primaria cuentan con sistemas de historia clínica electrónica y la receta electrónica es una realidad en la mayoría de las comunidades autónomas, aunque el grado de cobertura del servicio varía de unas a otras. La incorporación de la historia clínica electrónica en los centros hospitalarios ha sido más dispar. Con carácter general, los nuevos hospitales tienen una implantación completa, mientras en los hospitales con años de funcionamiento esta labor ha sido más dificultosa. Las comunidades autónomas no publican informes de resultados de los proyectos ni del impacto en la mejora de la calidad de la atención[3].

Precisamente la incorporación de las TIC a los sistemas de salud se ha considerado siempre como un facilitador para la transformación del sistema de salud, y por lo tanto como una de las estrategias fundamentales para afrontar los retos mencionados antes. Sin embargo, las TIC son instrumento, aunque imprescindible, para mejorar la calidad del sistema de salud, y nunca un fin en sí mismas.

Esta incorporación de las TIC al sistema de salud ha permitido disponer de grandes bases de datos con información clínica, tanto de tipo estructurado como de tipo no estructurado. Un dato estructurado es aquel que se registra de acuerdo con un formato homogéneo predefinido, lo que permite armonizar los distintos registros, controlar la calidad de los datos (por ejemplo, mediante la aplicación de rangos de validez) y realizar tratamientos avanzados de los mismos, como cálculos estadísticos y análisis comparativos de series. Un ejemplo de dato estructurado es el registro de una fecha o una hora. Por el contrario, los datos no estructurados no siguen estrictamente un formato concreto, lo que limita mucho sus posibilidades de tratamiento y explotación. Ejemplos de dato no estructurado son un texto libre o una imagen.

Por otra parte, la informática se ha introducido en los sistemas de organización y control de la asistencia sanitaria (Sistemas de Información de Hospitales –HIS– y de Atención Primaria –SIAP–) y en la gestión económico-financiera y logística del sistema de salud. La integración o relación de estos sistemas con los sistemas de información clínica y las bases de datos poblacionales nos permite, además de calcular los costes de la asistencia sanitaria, plantear la posibilidad de explotar grandes conjuntos de datos.

A pesar de la importancia que han alcanzado las TIC en el sistema de salud, esta inversión sólo tiene sentido si se consigue incorporar las TIC a su cadena de valor, de forma que su aportación sirva para mejorar los resultados de las organizaciones sanitarias, medidos en términos que tengan sentido tanto para los pacientes como para la sociedad. El objetivo de este capítulo es describir la importancia de la explotación de grandes bases de datos para la mejora de los resultados del sistema de salud.

[1] Ministerio de Sanidad, Servicios Sociales e Igualdad (2015).
[2] Ministerio de Sanidad, Servicios Sociales e Igualdad (2016).
[3] Carnicero y Rojas (2010).

2. El ecosistema de salud

Un sistema de salud no es un ente simple ni aislado, sino que engloba o interactúa con varias entidades públicas y privadas. Cada una de ellas tiene sus propios intereses, pero algunos de ellos son compartidos. El conjunto de todas estas entidades se conoce como *ecosistema* de salud, y entre ellas destacan las siguientes:

- Gobierno central y autoridades regionales y locales. Son los principales responsables de la regulación del sistema de salud, mediante el establecimiento de un marco legal específico y el control de su aplicación. En los sistemas públicos de salud les compete también la financiación de la provisión asistencial.
- Servicios de salud, entendidos como organizaciones responsables de la gestión de una red asistencial determinada, delimitada desde un punto de vista geográfico, con una cartera de servicios claramente definida, y con una plantilla y unas instalaciones –propias o ajenas– que prestan servicios a la población del área geográfica de actuación.
- Hospitales, dedicados a la prestación de asistencia especializada y urgente.
- Centros de atención primaria, que constituyen un primer nivel básico de asistencia.
- Servicios de emergencias extrahospitalarios.
- Farmacias, para la provisión de medicamentos y productos sanitarios.
- Centros de convalecencia y otros cuidados, para el apoyo en la recuperación de pacientes.
- Profesionales sanitarios que prestan sus servicios como proveedores externos del sistema de salud, sin estar integrados en su plantilla.
- Servicios de salud pública, cuyo cometido es velar por el estado de salud de la población desde una perspectiva comunitaria y no individual.
- Aseguradoras, mutualidades y otras entidades que financian de forma total o parcial el proceso asistencial de los pacientes afiliados a ellas.
- Facultades de medicina, enfermería y otras profesiones sanitarias, para la formación de nuevos profesionales.
- Centros de investigación, para la investigación de enfermedades y el desarrollo de nuevas técnicas diagnósticas y terapéuticas.
- Colegios y asociaciones profesionales.
- Fundaciones y sociedades científicas.
- Grupos de interés, como son las asociaciones de pacientes.
- Industria farmacéutica y de otras tecnologías sanitarias.

Las relaciones entre todos estos componentes generan una gran cantidad y diversidad de flujos de datos, implicando varios procesos de negocio y, por extensión, varios sistemas y subsistemas que deben compartir información. Todos estos flujos deben ser tenidos en cuenta a la hora de plantearse la explotación de grandes conjuntos de datos, a fin de garantizar que se trabaja con información completa y veraz. En la Figura 1.1 se representa la estructura de un ecosistema de salud.

3. La cadena de valor del sistema de salud

La cadena de valor es un instrumento metodológico que se emplea para el análisis interno de una organización, como puede ser un hospital o un sistema de salud (véase la Figura 1.2), y permite acciones como las siguientes:

- Identificar las distintas actividades separables y calcular su aportación a los objetivos finales.
- Configurar la actividad general como un conjunto de actividades económicamente distintas.
- Establecer las interrelaciones horizontales y verticales entre todos los elementos de la organización.

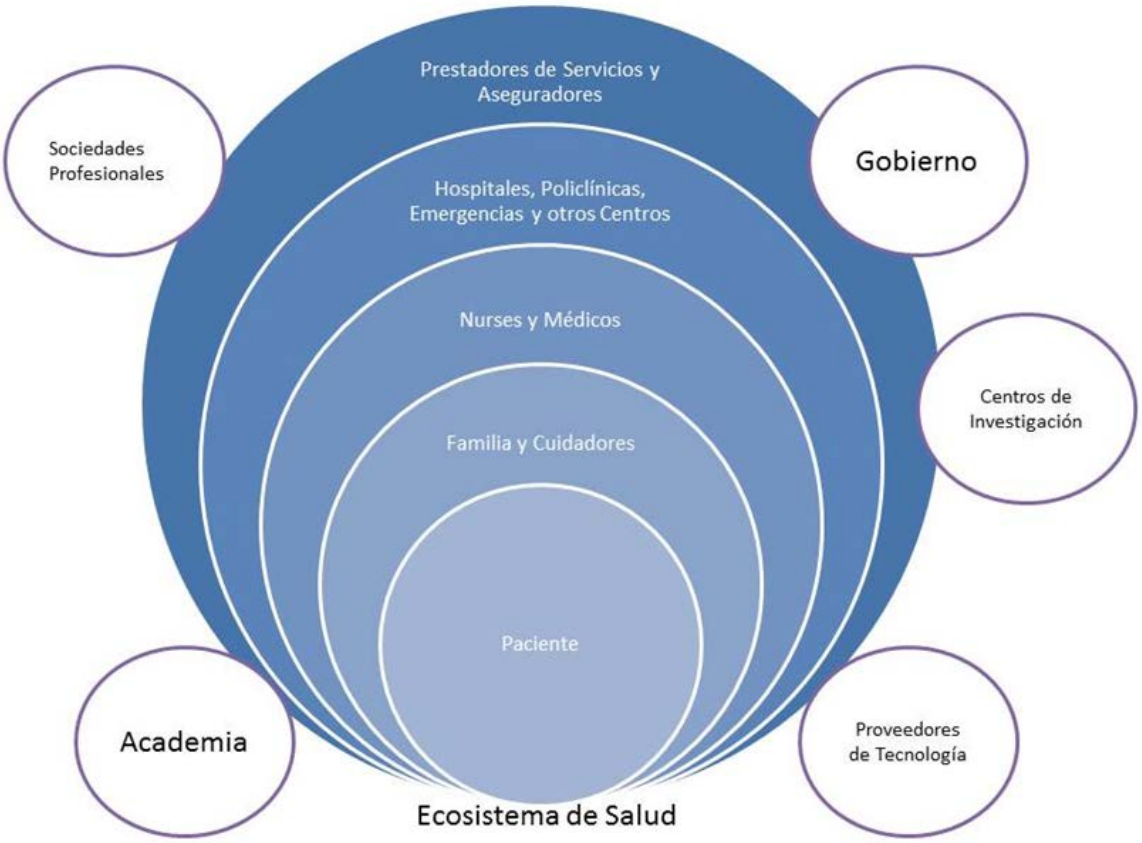

Figura 1.1. Componentes del ecosistema de salud.

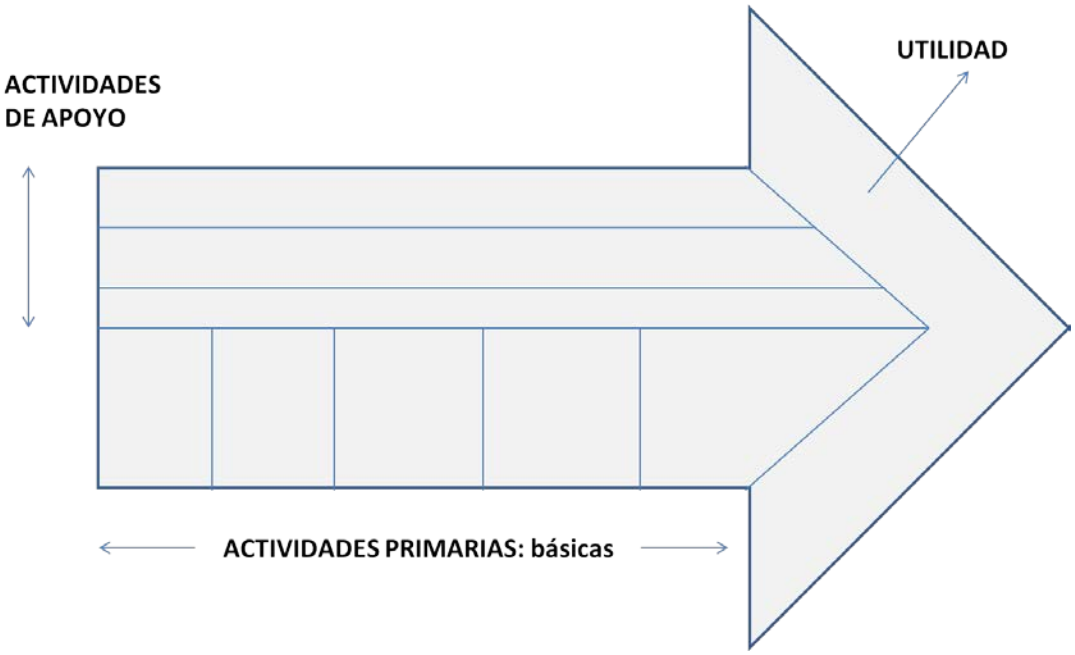

Figura 1.2. Cadena de valor del sistema de salud.

Definir la cadena de valor sirve para establecer cuantitativa y cualitativamente la contribución de cada actividad a la ventaja competitiva del centro sanitario o del sistema de salud. También permite identificar y comprender los eslabones que relacionan las distintas actividades, tanto la coordinación entre las internas (verticales y horizontales) como entre estas y las externas (cedidas o externalizadas, y las de la propia red del sistema de salud).

Otro criterio de clasificación distingue entre actividades primarias y de apoyo:

- Son actividades primarias:
  - La prestación de servicios de asistencia e investigación.
  - La entrada de inputs.
  - La logística interna.
  - La logística externa, como el marketing y la imagen.
  - Los servicios ulteriores, como las revisiones y controles posteriores a los pacientes.
- Son actividades de apoyo:
  - La gestión de infraestructuras, como el mantenimiento de instalaciones.
  - La administración de recursos humanos.
  - El desarrollo de tecnologías.
  - La gestión de suministros.

Todas las actividades, sean o no asistenciales, sean internas o externas, sean primarias o de apoyo, deben estar concatenadas y claramente orientadas a la consecución de resultados. Todos los miembros de la organización deben conocer cuáles son los objetivos finales de la misma, cuáles son los resultados que persigue y con qué criterios se van a evaluar.

No obstante, los sistemas de salud deben ir más allá de la mera evaluación de actividad y coste, y perseguir también resultados que deben tener sentido tanto para el paciente como para la sociedad, por intangibles o difíciles de ponderar que estos resultados puedan ser. Para garantizar la sostenibilidad del sistema es imprescindible superar los conceptos de eficacia y coste, y evolucionar hacia los conceptos de efectividad, eficiencia y calidad.

Aunque la calidad forma parte de la efectividad y la eficiencia, para su medición debe tenerse también en cuenta la valoración que los pacientes y los ciudadanos hacen del servicio recibido, no sólo en términos generales de satisfacción sino también con indicadores que valoren aspectos concretos del proceso por el que han sido atendidos. Por este motivo el sistema de salud debe fijar los resultados que se esperan de cada proceso asistencial. Por ejemplo, en el caso de cirugía de prótesis de cadera, además de medirse la supervivencia del paciente y el coste de la atención, también deben formar parte de los resultados aspectos tan importantes como la tasa de infección nosocomial, el índice de recuperación funcional, el alta laboral, el tiempo que el paciente sufre dolor, y las posibles secuelas. Como se ha explicado anteriormente, estos indicadores son de gran interés tanto para el paciente como para la sociedad. Estos resultados deben calcularse para cada paciente, y no sólo de forma acumulada o consolidada por servicios o departamentos. Para ello se requiere planificar, organizar, medir, controlar y evaluar el ciclo completo de atención para cada paciente y para cada grupo de pacientes con características similares[4].

Todo ello requiere un esquema de organización diferente al tradicional y habitual en el sector sanitario. El nuevo esquema contempla una organización que esté enfocada al paciente, que esté orientada

[4] Porter y Lee (2013).

a la consecución de los resultados que se hayan definido previamente en los objetivos generales, que elimine lo innecesario, que delimite claramente competencias y responsabilidades, y que reduzca costes allá donde sea posible y razonable. Este enfoque evita la fragmentación del proceso asistencial y exige –a la par que facilita– una coordinación tanto vertical como horizontal, una cooperación más estrecha entre los distintos participantes y una innovación tecnológica.

Una organización como la que sucintamente se ha descrito mejora la calidad, reduce los tiempos de atención y representa una oportunidad para relacionarse de otra manera con agentes externos, para alcanzar una mejor integración entre la atención primaria y la especializada, y en general para lograr una mejor coordinación de las entidades que forman parte del llamado ecosistema de salud, llegando así a constituir un auténtico clúster de salud.

Por otra parte, este planteamiento de organización que contribuye a la consecución de objetivos que importan tanto a la sociedad como a los pacientes, requiere financiación y dedicación de profesionales y de tiempo para la incorporación de tecnologías médicas y de la información, así como vencer la resistencia al cambio y disponer de un potente sistema de información que además se incorpore a la cadena de valor. No debe olvidarse nunca que la actividad sanitaria es muy intensiva en el tratamiento (tanto desde el punto de vista del consumo como de la generación) de información, sobre todo de aquella de naturaleza clínica, y por lo tanto exige contar con instrumentos de la debida potencia para satisfacer esta necesidad.

4. Algunas consideraciones sobre la incorporación de las TIC a la cadena de valor del sistema de salud

Como consecuencia de todo lo anterior, se concluye que el sistema de información de salud debe estar centrado en el paciente y en la consecución de los resultados específicos que se persiguen en cada proceso asistencial. Para conseguir ese objetivo se debe revisar la cadena de valor y favorecer las iniciativas innovadoras orientadas a la implantación de este esquema de organización y actuación. Debe tenerse en cuenta que sin un sistema de información no puede haber un control de gestión eficaz ni una evaluación precisa de los resultados, y sin control de gestión ni evaluación es imposible conseguir una mejora y mucho menos una transformación profunda del sistema de salud.

De la misma forma que el sistema de salud debe perseguir resultados que tengan sentido tanto para el paciente como para la sociedad, el sistema de información debe tener como objetivo propio la generación de valor para todas aquellas personas que hacen uso de los datos por él gestionados, independientemente de que participen en las actividades asistenciales o en las no asistenciales. Esto engloba a pacientes, profesionales, proveedores de servicios externalizados, servicios de salud pública y autoridades sanitarias, por lo que es imprescindible un acuerdo de todos los involucrados en el proyecto, sin que ello impida la búsqueda de resultados alcanzables mediante la formulación de propuestas realistas y viables, evitando de este modo incurrir en las utopías que suelen aparecer en procesos participativos como el que aquí se propone.

Hasta ahora, la incorporación de las TIC al sistema de salud ha tenido como resultado principal una mejora de la eficacia, consecuencia de la automatización total o parcial de los procesos. Los próximos pasos se deben dirigir a apoyar la gestión clínica, en pos de la mejora de los resultados de la atención sanitaria de forma personalizada en los pacientes. Una información más accesible y más fácil de interpretar podría mejorar los resultados y también reducir costes. Sin embargo, para conseguir estos objetivos se requiere, aunque resulte obvio decirlo, que los datos que se procesen sean relevantes y precisos. Una vez más debe

resaltarse que no se trata de acumular información si esta no puede estar disponible en el momento clave y con el nivel de detalle necesario[5].

5. La explotación de grandes bases de datos: oportunidades para el sistema de salud

Las tecnologías actualmente existentes permiten la explotación de grandes cantidades de datos que se han originado precisamente gracias a la incorporación de esas mismas tecnologías al sistema de salud. Esta explotación de datos debe concebirse y articularse como un apoyo expreso a la consecución de los objetivos generales del sistema de salud que, como ya se ha reiterado, deben tener significado en primer lugar para los pacientes y en segundo para la sociedad.

Por lo tanto, el primer objetivo de la explotación de grandes conjuntos de datos debería ser proporcionar la mejor información disponible a quienes toman decisiones relacionadas con la asistencia sanitaria, en especial los médicos, de manera que esta información les ayude a tomar la decisión más adecuada en cada situación. Esto implicaría un seguimiento menos estricto de protocolos estandarizados, sin que ello suponga en modo alguno que deba abandonarse esta práctica, para dar un mayor peso a la información sobre los resultados obtenidos por el profesional en su propia práctica, por su departamento en el entorno de su especialidad clínica, o en otros ámbitos. Ejemplos de este tipo de información serían datos tan diversos como la flora y resistencias bacterianas predominantes en su hospital, o los resultados obtenidos en casos similares con diferentes tratamientos.

Por otra parte, la explotación de grandes bases de datos de salud debe dirigirse también a prever las necesidades de los pacientes y planificar de forma anticipada los servicios que podrían requerir. El ejemplo más evidente de este tipo de análisis son los estudios dirigidos a detectar los pacientes crónicos, planificar su asistencia, gestionar el proceso de atención de forma personalizada y conseguir así una mejora de los resultados, lo que tendrá sentido tanto para el paciente como para el sistema de salud. Por ejemplo, las grandes compañías de venta por Internet explotan los datos de sus clientes para personalizar las ofertas comerciales que les envían. De modo similar, aunque desde luego con una finalidad bien distinta, los sistemas de salud deberían aprovechar las oportunidades que ofrecen las TIC para personalizar los servicios que necesitan sus pacientes[6].

El cambio de enfoque puede resumirse en que lo más importante no es disponer de la información, algo que ya sucede, sino ser capaces de formular las preguntas adecuadas en el momento oportuno, procesarlas para ofrecer sólo la información necesaria y relevante, y presentarla al profesional de modo que pueda interpretarla de forma correcta, rápida y sencilla para tomar una decisión acertada. Para ello, los clínicos también necesitan conocer el grado de cumplimiento de los resultados que se esperan de ellos, y si sus pacientes están recibiendo la atención apropiada y en tiempo oportuno.

Los directivos y gestores del área asistencial deben poder formular preguntas similares, de modo que puedan conocer el estado de situación, planificar la estrategia y los objetivos que debe alcanzar su área de gestión, evaluar los resultados, y tomar medidas preventivas o correctoras en caso necesario. Si estos directivos deben centrarse en alcanzar unos niveles mínimos de efectividad y eficiencia, los sistemas de información y la explotación de grandes volúmenes de datos deben permitirles y facilitarles la medición y el análisis de los indicadores correspondientes, proporcionándoles esta información en tiempo y forma para que puedan tenerla en cuenta durante el proceso de toma de decisiones.

[5] Harvard Business Review (2014).
[6] Davenport (2013).

Lo mismo puede aplicarse a cualquier otro profesional y directivo del sistema de salud. La aportación de valor de las TIC, y más en concreto de la explotación de grandes bases de datos, debe encaminarse a facilitar el proceso de gestión y de toma de decisiones, de manera que se contribuya a alcanzar los objetivos finales de la organización. En la Tabla I.2 se muestran algunas de las fuentes de datos más importantes para la incorporación de los sistemas de información a la cadena de valor del sistema de salud.

Tabla I.2. Principales fuentes de datos del sistema de información de salud.	
Ficheros maestros	Base de datos poblacional
Sistemas clínico-administrativos	Sistema de Información de Hospital (HIS)
	Sistema de Información de Atención Primaria (SIAP)
Sistemas clínicos	Historia Clínica
	Gestor de Peticiones Clínicas
	Sistema de Información de Laboratorio (LIS)
	Sistema de Información de Radiología (RIS)
	Sistema de Información de Anatomía Patológica
	Prescripción Electrónica de Medicamentos
	Receta Electrónica
Gestión logística	Farmacia hospitalaria
	Suministros
	Prótesis
	Material sanitario fungible
	Mantenimiento de infraestructuras y equipos
	Proveedores
Gestión económico-financiera	Contabilidad
	Costes
	Facturación
Explotación de datos	Conjunto Mínimo Básico de Datos (CMBD)

Por otra parte, como ya se ha expuesto antes, el sistema de salud forma parte de un ecosistema que puede y debe llegar a convertirse en un auténtico *cluster* de salud, circunstancia que debe ser tenida en cuenta tanto por el sistema de información en general como por la explotación de grandes conjuntos de datos en particular.

Un claro ejemplo de ello es la investigación, que aunque puede desarrollarse perfectamente en el ámbito interno de un hospital o un centro de I+D, tiende cada vez más a basarse en el trabajo colaborativo en red, surgiendo en consecuencia la necesidad de explotar información almacenada en bases de datos de diversa naturaleza, gestionadas mediante procedimientos particulares, soportadas por plataformas tecnológicas diferentes y correspondientes a distintos ámbitos. Lo mismo puede afirmarse de los procesos de innovación empresarial, que podrían beneficiarse notablemente de la explotación de información clínica en el marco de sus procesos de investigación, desarrollo e innovación.

6. Límites

La explotación de grandes conjuntos de datos en salud tiene un importante condicionante, que viene dado por la normativa vigente en materia de protección de datos, en virtud de la cual esta información goza de la máxima confidencialidad. Las leyes no son más que el reflejo de los valores, creencias y culturas imperantes

en la sociedad en un momento dado, por lo que primero deben tomarse en consideración los aspectos bioéticos de esta explotación de datos, y después los requisitos legales.

El segundo capítulo de este trabajo analiza el sistema de información de salud en su conjunto, reflexionando sobre la importancia del acceso a la información clínica y repasando sus posibles usos, que van más allá de la mera asistencia para englobar también la docencia, la investigación, la gestión de las organizaciones sanitarias, las actividades de salud pública y salud laboral, e incluso la validez legal de esta información en procedimientos jurídicos. En cada uno de ellos se plantean las distintas necesidades principales y sus motivos, y se remarca la obligación de llegar a unos compromisos que permitan un grado de satisfacción razonable de todas estas necesidades.

El tercer capítulo trata sobre los aspectos bioéticos de la explotación de grandes cantidades de datos sanitarios, subrayando el creciente protagonismo –y con él la relación de dependencia– de las TIC dentro del ecosistema de salud. Aunque la tecnología puede albergar las claves para la transformación del modelo de los sistemas de salud y el aseguramiento de su calidad y sostenibilidad, el aprovechamiento de este potencial no puede nunca estar reñido con el respeto a los principios fundamentales de la ética profesional del sector sanitario. En este capítulo se exponen cuestiones y retos relacionados con la explotación de datos a gran escala, y se analizan casos reales que ilustran los conflictos de intereses existentes dentro del ecosistema de salud. Todo ello lleva a proponer una regulación más exigente y exhaustiva, y una mayor formación y concienciación de los profesionales.

El cuarto capítulo se centra en el análisis del marco legal vigente, estudiando por separado la normativa general sobre protección de datos y la normativa propia del sector sanitario. Esto permite apreciar las colisiones que se producen en ocasiones entre una y otra, pero antes de eso el autor destaca dos hechos muy importantes: la inexistencia de un tratamiento legal específico para la explotación de conjuntos masivos de datos, y la consiguiente remisión a un marco normativo general que se aprobó en un momento muy anterior a la irrupción de esta disciplina.

Por último, el quinto capítulo constata la envergadura de los sistemas de tratamiento y explotación de cantidades masivas de información, y estudia las directrices organizativas que deben seguirse para su implantación y mantenimiento. Tras exponer las características y dimensiones fundamentales de los grandes conjuntos de datos, se detalla el proceso en todas sus fases, desde la identificación de estos datos y sus fuentes hasta la generación de indicadores y resultados, pasando por la definición de métodos de recogida, procesamiento y registro de los datos, la gestión de la calidad de la información, la presentación de datos al usuario o el apoyo a la toma de decisiones, entre otros. Asimismo, también se revisan las tecnologías aplicables a este ámbito actualmente existentes, con mención expresa de algunas herramientas específicas.

7. Conclusiones

Los sistemas de salud están obligados a evolucionar para conciliar las exigencias de resultados con la garantía de su sostenibilidad. Desde el primer momento, la incorporación de las TIC se ha considerado como un elemento imprescindible para acometer esta transformación. Tras una experiencia de varios años e incluso décadas en la utilización de las TIC dentro de los sistemas de salud, es también indiscutible que tanto las necesidades de las organizaciones como las posibilidades que ofrecen las tecnologías han cambiado. Se puede afirmar que, al mismo tiempo que las necesidades se han hecho más exigentes y específicas –y apremiantes en algunos casos–, las TIC han respondido a estos desafíos con la creación de herramientas cada vez más potentes.

En un marco de escasez de recursos para satisfacer una demanda cada vez mayor de más y mejor atención sanitaria, la aparición de sistemas de explotación de grandes conjuntos de datos resulta de especial interés para el desarrollo de sistemas de apoyo a la toma de decisiones clínicas y de gestión. En el ámbito clínico, esto puede suponer una mayor efectividad de los procesos asistenciales alineada con una mayor eficiencia en el aprovechamiento de los recursos sanitarios, tanto humanos como materiales y presupuestarios.

En el ámbito de la gestión, el análisis de grupos masivos de datos puede ayudar a las organizaciones a tener un conocimiento más preciso y rápido de las necesidades existentes, a diseñar estrategias y políticas mejor adaptadas a la realidad de estas necesidades, y a medir y evaluar razonadamente los resultados obtenidos. Es más, se puede realizar un seguimiento más estricto de la eficacia de estas medidas, detectar desviaciones e introducir las modificaciones que se consideren necesarias en un momento dado. Por último, se abre la puerta a la creación y utilización de modelos predictivos que permitan a las organizaciones sanitarias dar un paso más y anticiparse a los problemas y necesidades, en lugar de concentrarse en su detección y resolución.

Parece claro que los beneficios de este tipo de sistemas responden perfectamente a las necesidades actuales de los sistemas de salud, y que por lo tanto las organizaciones sanitarias deben empezar a incorporar estas herramientas en sus estrategias y prioridades. Sin embargo, debe tenerse en cuenta que se trata de actuaciones que entrañan una gran dificultad, como consecuencia de la complejidad de su implantación tanto desde el punto de vista técnico como organizativo, y de las limitaciones que establecen los principios éticos y los requisitos legales que corresponden al sector sanitario.

En otras palabras, el derecho de los pacientes a una asistencia de calidad incluye no sólo la resolución de sus problemas de salud, sino también el respeto a su intimidad y la protección de su información clínica frente a usos indebidos. En consecuencia, la urgencia de las necesidades de los sistemas de salud no puede llevar en ningún momento a obviar esta faceta de la calidad asistencial, máxime cuando una solución de compromiso es perfectamente alcanzable si se analizan debidamente las necesidades de la organización, los requisitos éticos y legales, y las posibilidades que ofrecen las TIC.

En conclusión, la situación actual de los sistemas de salud puede resumirse en que deben afrontar el reto inaplazable de compatibilizar la mejora de la calidad de sus servicios con una profunda transformación de su modelo que permita garantizar su sostenibilidad. Para ello deben tomar varias medidas cruciales, entre las que destaca el aprovechamiento de la oportunidad que les ofrecen las TIC, pero tampoco deben perder de vista en ningún momento los límites que vienen marcados por la bioética y por la legislación vigente, a fin de salvaguardar los derechos de las personas.

Bibliografía

- Carnicero J. y Rojas D. *Lessons learned from implementation of information and communication technologies in Spain's healthcare services: issues and opportunities*. Appl Clin Inform 2010; 1(4):363-76.
- Davenport, Thomas H. *Analytics 3.0*. Harvard Business Review 91, no. 12 (December 2013): 64–+.
- Harvard Business Review, 2014. *How Big Data Impacts Healthcare*.
- Ministerio de Sanidad, Servicios Sociales e Igualdad (2015). Barómetro Sanitario 2014. Resultados totales. Disponible en: http://www.msssi.gob.es/estadEstudios/estadisticas/docs/BS_2014/es8814mar.pdf

— Ministerio de Sanidad, Servicios Sociales e Igualdad (2016). Barómetro Sanitario 2015. Resultados totales. Disponible en: http://www.msssi.gob.es/estadEstudios/estadisticas/docs/BS_2015/Es8815mar.pdf

— Porter, Michael E., and Thomas H. Lee. *The Strategy That Will Fix Health Care*. Harvard Business Review 91, no. 10 (October 2013): 50–70.

Capítulo II

La importancia de la explotación de datos de salud

Fernando Escolar Castellón

1. Introducción

Los servicios de salud son grandes productores de información, en su mayoría procedente de personas concretas y en relación con su estado de salud. Esta información puede ser sobre aspectos relacionados directamente con la salud de las personas y sobre otros considerados administrativos y económicos. Los aspectos mecánicos y operativos de esta información (cómo se obtiene, se ordena, se almacena y se recupera) han sido tratados en otros documentos editados por la SEIS[7].

La información que procede de forma directa una persona concreta y concierne a su salud como individuo, se ordena y se almacena en la historia clínica o en la historia de salud. Éstas se utilizan de forma directa en la asistencia, docencia y en algunos casos como base jurídico-legal. Su uso indirecto y agregado es útil en investigación y gestión, así como en epidemiología y salud pública, aunque en algunos de estos casos también sea necesario el acceso a la información de individuos concretos.

También se producen grandes cantidades de otros tipos de información, como es la administrativa y económica, cuyo uso y tratamiento es diferente.

2. Asistencia

La asistencia sanitaria[8] es una de las funciones que más peso tiene en los servicios de salud, y consume la mayor parte de los recursos, también de información. Su objetivo principal es recuperar o conservar la salud de un individuo concreto.

Por su propia naturaleza el proceso de asistencia clínica es siempre personal. El "razonamiento clínico" utilizado en la práctica clínica es interpretativo, es decir, valora la información disponible de acuerdo con un contexto. Por muy sistematizado que esté el cuerpo doctrinal en el que se basa, en forma de guías y protocolos, éstos deben ser aplicados adaptándolos a la realidad individual y, por tanto, de manera "personalizada".

La información necesaria para prestar atención sanitaria no puede ser "anonimizada", sino que debe estar perfectamente clara de forma univoca e inequívoca la identidad del individuo a quien pertenece. Una información sobre la salud de una persona no puede ser utilizada cuando existan dudas sobre su pertenencia.

[7] Carnicero et al. (2002).
[8] Escolar y Martínez-Berganza (2004).

La mayor parte de la información utilizada en la asistencia estará incluida en la "historia clínica", que abarca los hechos e hitos asistenciales y clínicos de una persona en relación a sus patologías. El concepto de "historia de salud" es más amplio, ya que abarca todos los hechos en relación con la salud de la persona, además de las posibles patologías. La "historia clínica" podría considerarse como un subconjunto de la "historia de salud".

El conocimiento existente sobre el ser humano es incompleto, y la significancia que tendrá la información recogida sobre el proceso asistencial es difícil de establecer previamente. Por ello se tiende a recoger y registrar gran cantidad de información, que muchas veces refleja literalmente las aportaciones del paciente.

La información contenida en una historia clínica suele ser bastante heterogénea, con datos objetivos y subjetivos, cualitativos y cuantitativos, que siempre están sujetos a interpretación de acuerdo con el contexto y evidencias existentes. La historia clínica presta funciones de contenedor y vehículo de transmisión de información que podría ser relevante, entre los diferentes profesionales implicados en el proceso asistencial.

La aplicación de las tecnologías de la información ha facilitado la accesibilidad de las historias clínicas, pero también la trazabilidad de esta accesibilidad. No existía medio de averiguar los accesos a una historia en soporte convencional de papel. Sólo quedaba constancia de la salida y entrada en el archivo, pero no de quiénes la habían leído o de si se habían hecho copias. Se podría decir que la única protección real era el propio caos del documento. Por el contrario, en un documento electrónico puede quedar constancia detallada de las personas que acceden e incluso de qué partes del documento han visualizado.

Las tecnologías de la información también permiten limitaciones del acceso. Frecuentemente surge la cuestión: ¿Debe limitarse el acceso a partes de la información contenida en una historia clínica al personal sanitario implicado en el proceso asistencial, entendiendo por personal sanitario a los facultativos y a la enfermería (ATS-DUE, matronas, fisioterapeutas y auxiliares de clínica)? La imposibilidad de establecer previamente la relevancia de la información, así como el hecho de que esta información esté siempre sujeta a la interpretación de cada profesional implicado, hace que la ignorancia de informaciones referentes a la salud del individuo dé lugar a situaciones de riesgo no razonable. No hace falta mencionar tópicos concretos para vislumbrar el riesgo derivado del desconocimiento de patologías previas que puedan explicar los padecimientos actuales, o posibles efectos adversos o interacciones por desconocimiento de tratamientos prescritos por otros facultativos.

Además, todo el personal sanitario está obligado a registrar sus observaciones y acciones en las partes correspondientes de la historia clínica, teniendo que valorar anotaciones hechas por otras personas o cumplir instrucciones y prescripciones. Por razones puramente técnicas, en un proceso asistencial abierto estaría desaconsejado establecer limitaciones previas del acceso al personal sanitario (independientemente de su nivel) implicado en dicho proceso.

Además, la tecnología permite que el personal que accede a la información quede registrado y siempre estará obligado a la confidencialidad, tanto por obligación legal como deontológica, tal como se indica de forma explícita en los principios hipocráticos en los que se basa la ética médica.

También se plantea si debe limitarse el acceso a la información sobre la salud de una persona al personal no sanitario adscrito al servicio asistencial encargado de su atención, en especial al personal administrativo. Dependerá de la organización de la unidad, pero la gestión del movimiento de pacientes recae en personal administrativo, por lo que éstos deben tener acceso, al menos, a los datos demográficos y

a los sistemas que manejan las diferentes agendas. En la mayoría de las unidades asistenciales, la introducción de datos recae en última instancia en personal administrativo aunque se utilicen sistemas informatizados, ya que es frecuente el uso de dispositivos tipo dictáfono, formularios o inclusive el dictado directo. Además, es el personal administrativo el encargado de ordenar la documentación, confeccionar o "montar" informes externos, recibir información de pacientes que deben transmitir a los facultativos, gestionar correos y otras muchas funciones administrativas relativas a la salud de las personas.

Se deben crear perfiles de acceso a la información, dependiendo del puesto de trabajo y titulación, que no supongan un obstáculo al desempeño diario de las unidades. Recordemos que todo el personal que tiene acceso a datos de salud de las personas, por motivos de su trabajo, independientemente de si es sanitario o no, está obligado a la confidencialidad. Se ha demostrado útil la formación periódica a este respecto, para dar a conocer y concienciar de la obligación de confidencialidad por parte de todos los trabajadores de una institución sanitaria.

Todos los sistemas deben tener una trazabilidad que permita la comprobación del buen uso de los accesos, posibilitando auditorías dirigidas o aleatorias que sirvan para detectar casos de accesos no lícitos.

3. Docencia

El conocimiento experto necesario para el ejercicio de las profesiones sanitarias es adquirido fundamentalmente de forma empírica. Además del contenido doctrinal es necesario su contraste con situaciones reales. Las "prácticas con casos reales" son imprescindibles en el aprendizaje de la profesión médica.

En este sentido son necesarias prácticas clínicas "a la cabecera del paciente" en tiempo real. El estudiante debería acceder a la información que el profesor o tutor considere necesaria para su aprendizaje. En este acceso debe ser considerado como un profesional más implicado en el proceso y, por tanto, sujeto a los mismos registros de trazabilidad, aspectos legales y de confidencialidad. Por ello debería darse formación previa a los estudiantes, para que tuvieran un conocimiento cabal sobre los aspectos del acceso a la información clínica antes de comenzar con estas prácticas a la cabecera del paciente.

Las sesiones clínicas utilizan un caso concreto que por sus características puede servir de modelo docente. En esta situación, aunque la información utilizada corresponde a un caso real de un individuo concreto, no es necesaria la identificación del sujeto, debiendo eliminarse de la exposición los datos que puedan conducir a la identificación explícita de la persona, así como toda información que no se considere relevante para los fines docentes. En los casos cuya propia peculiaridad pudiera llevar a la identificación de la persona, se puede plantear la solicitud de autorización a la misma para su exposición, y asegurar que todos los profesionales y estudiantes que asisten a una sesión clínica están sujetos a la confidencialidad sobre el caso. Estas condiciones también tendrían que cumplirse si el caso va a ser objeto de publicación en una revista científica o expuesto en un congreso.

Las simulaciones basadas en la vida real creadas por profesionales expertos no corresponden a ningún individuo concreto.

4. Jurídico-legal

Las autoridades judiciales competentes pueden requerir la información perteneciente a un individuo concreto, bien al profesional o profesionales que le prestaron atención clínica, o bien al custodio de la misma

(como puede ser la dirección de un centro asistencial). Ante este requerimiento hay obligación de facilitar los documentos originales o copias exactas de los mismos. Generalmente, en el mismo requerimiento se indica si se solicita toda la información que se posea o sólo las partes relacionadas con un hecho o episodio concretos.

5. Investigación

La información acumulada en los servicios de salud posee una gran cantidad de datos que pueden ser de gran valor cuando se estudian ordenada y adecuadamente, pudiendo hacer aportaciones significativas al cuerpo de conocimiento de las ciencias biológicas y sociales.

La aplicación de las tecnologías de la información ha facilitado el estudio ordenado de la gran cantidad de información acumulada en los servicios de salud, favoreciendo la investigación clínica y epidemiológica[9]. Se pueden obtener directamente conjuntos de datos concretos relacionados, pero la heterogeneidad de la información clínica hará necesaria en muchas ocasiones la revisión individualizada de cada caso.

La información será útil para estudios retrospectivos, prospectivos, observacionales y ensayos:

- Los estudios retrospectivos necesitan examinar hechos pasados, donde no se realizó ninguna intervención en la introducción de los datos diferente a los procedimientos habituales existentes.
- En los estudios prospectivos se realiza algún tipo de intervención en la recogida e introducción de datos que aplica criterios homologables en todos los casos estudiados.
- Los estudios observacionales (bien retrospectivos o prospectivos) suponen la obtención de una serie de datos en un periodo de tiempo determinado, pertenecientes a una población aleatoria o concreta (en este último caso se denomina corte).
- Los ensayos clínicos suponen algún tipo de intervención en un grupo que se compara con otro, donde no se realiza intervención alguna o ésta es diferente de la del primer grupo. Deben ser estudios prospectivos, y pueden realizarse apoyándose en los sistemas de información existentes en un servicio de salud o de forma independiente de ellos. Al requerir intervenciones cuyo beneficio se trata de probar, deben de cumplir una serie de requisitos cuyo análisis se escapa del objetivo de este trabajo.

En cualquiera de los casos, esta información no es accesible de forma primaria a los investigadores, siendo necesaria la solicitud de autorización de acceso a los organismos custodios de la misma (generalmente las direcciones de los centros asistenciales). La solicitud, por parte de los investigadores, de autorización de acceso a datos de salud, debe obedecer a la existencia previa de un proyecto de estudio articulado, donde se formulen hipótesis y objetivos coherentes. Deben explicitarse los datos que se necesitarán, si pueden obtenerse en un conjunto o si también es necesario el acceso individualizado.

Si se dan estas condiciones, los servicios de salud deberían autorizar el acceso a estos datos, y el investigador deberá comprometerse a respetar el anonimato de los datos. En el caso de que el estudio requiera un acceso individualizado y por tanto el posible conocimiento de la identidad del sujeto, deberá comprometerse además a no facilitar información que pudiera conducir a la identificación de personas concretas y a utilizarlos sólo para el fin de investigación con el que fueron solicitados.

Los servicios de salud deberían establecer procedimientos ágiles que faciliten el acceso a los datos. Un trabajo de investigación no puede verse impedido por la complejidad de los procedimientos

[9] Fletcher y Fletcher (2009).

administrativos. La existencia de una obligación de confidencialidad por parte del investigador tampoco puede ser un impedimento a un proyecto que reúne todas las condiciones. Sí se le puede exigir y retirar la autorización, e incluso sancionar si así está dispuesto, en caso de que la quebrante.

6. Gestión

La gestión es el conjunto de acciones destinadas a la consecución de un fin u objetivo. Implica la máxima eficacia y efectividad posibles de la forma más eficiente. Da el soporte necesario[10] para facilitar las actuaciones que sobre la salud de las personas desempeña un servicio de salud, siendo la función de asistencia sanitaria la de más peso y complejidad, pero no menos importantes las acciones preventivas, sobre higiene, salud pública y otras acciones en coordinación con servicios sociales.

Gestionar significa realizar un plan de acción, es decir, planificar, tomar decisiones en consecuencia y evaluar, aplicando una garantía de calidad en todo el proceso. Para todo ello, en el mundo de la salud son necesarias grandes cantidades de información. La fuente principal de información para la gestión suele ser la generada por los propios servicios de salud, aunque también son necesarias fuentes externas a ellos, como son el censo y el Instituto Nacional de Estadística.

Con la información cuantitativa y cualitativa disponible, y utilizando preferentemente métodos estadísticos, se elaborarán indicadores que a modo de resumen informen sobre los atributos o un conjunto de parámetros determinados, que tienen que mostrar la imagen del desempeño de un servicio de salud, en sus diferentes aspectos funcionales y operativos, sobre lo ocurrido, el estado actual y las tendencias. En la medida en que reflejen la realidad donde se actúa, es decir, que sean pertinentes desde el punto de vista cualitativo, cuantitativo y en el tiempo, facilitarán la correcta toma de decisiones.

Los datos que serán la base de estos indicadores requieren una homogeneidad en los procedimientos de introducción, recuperación y procesado que permita la comparación.

Será necesario el conocimiento sobre costes y gastos, gestión de personal, mantenimientos, actividad sanitaria, estado de salud poblacional, morbilidad y casuística. Los indicadores que se elaboran a partir de información económica, administrativa y de salud pueden obtener sus datos de forma anónima.

Sin embargo, para cuantificar la morbilidad y la casuística y relacionarlas con el coste en recurso se pueden utilizar diversos métodos. Uno de los más utilizados en nuestro medio son los denominados “Grupos Relacionados por el Diagnóstico” o GRD, que lo realizan a través de la casuística. Se elaboran a partir del análisis individual y directo por un equipo de codificadores, que utilizan una metodología homogénea y validada que requiere al acceso a la historia clínica y a los informes médicos de cada episodio asistencial. La información final es agregada de forma anónima.

Con la información agregada y en forma de indicadores se construye el sistema de información y los cuadros de mando, que informen sobre los servicios prestados, costes, gastos, recursos, actividad sanitaria, morbilidad, casuística y otros, de forma relacionada y en periodos de tiempo que se determinen, mostrando la evolución y las tendencias.

Toda la información elaborada es la que da el conocimiento necesario sobre la situación de partida y los fines que marcan los objetivos, lo que permitirá realizar una planificación y deducir las acciones a realizar.

[10] Asenjo (2006).

Para evaluar, que significa comprobar el grado de cumplimiento de los objetivos propuestos, se utilizan los indicadores que mejor reflejen los aspectos del objetivo, sin cometer el frecuente error de “confundir” estos indicadores con los objetivos.

6.1. Garantía de calidad

La calidad es el conjunto de propiedades que permiten juzgar el valor de algo con respecto a otros. Implica auditoría y comparación. Los programas de garantía de calidad son aplicados en la búsqueda de la excelencia, bien de forma global o más frecuentemente en áreas o aspectos determinados: asistencia clínica, administración, funcionamiento global de una unidad concreta y otros.

La auditoría busca y obtiene información necesaria de acuerdo con estándares establecidos, que den homogeneidad y permitan la comparación. Para ello es necesario el acceso a datos anónimos y elaborados de los sistemas de información, pero también a historias clínicas e informes que contienen datos personales cuando la auditoría afecta a un área asistencial.

6.2. Gestión clínica

La gestión clínica supone acciones en torno a la asistencia, que interesan tanto a la atención prestada a una persona concreta como a la organización de la unidad asistencial. Comienza en el proceso clínico, que supone la toma de decisiones de acuerdo a la información derivada de las necesidades de la persona, su contexto, el conocimiento científico y los recursos disponibles, con objeto de que la atención aplicada sea eficaz y eficiente en la recuperación o en la conservación de la salud de esa persona.

En este proceso el conocimiento científico se obtiene externamente al servicio de salud, aunque éste pueda contribuir a él. La información sobre las necesidades de la persona y su contexto se encuentra en la historia clínica y en la de salud.

Los recursos técnicos existentes en un servicio de salud deben estar implícitos en la cartera de servicios ofertada por el servicio de salud, constituyendo el marco de actuación en la atención sanitaria. Si se quiere que la atención se eficiente además de eficaz habrá que conocer y tener en cuenta la información sobre los gastos y los costes.

El personal clínico tiene que tener conocimiento de sus resultados de actividad, tanto desde el punto de vista económico como sobre la salud, siendo necesarios indicadores sobre morbilidad, casuística, mortalidad y actividad, en tiempo suficiente para poder planificar su labor forma eficiente. Esta información debe ser proporcionada por el sistema de información del servicio de salud o de la institución donde desempeña su trabajo.

7. Salud pública

La salud pública se encarga de la protección y mejora de la salud de una población como colectivo. Para ello se necesita información epidemiológica sobre morbilidad y mortalidad, estilos de vida, medio ambiente, etc.

Estos servicios establecen programas de prevención poblacionales, como pueden ser la detección precoz de patologías como diversos cánceres, o los programas de vacunación, y necesitan información procedente de los servicios de salud asistenciales, del instituto nacional de estadística, meteorológicos y sobre agricultura y ganadería.

Además, los servicios asistenciales tienen que facilitarles la información relativa a personas concretas, que pudiera ser necesaria en los casos de declaración obligatoria y para la realización de mapas epidemiológicos y búsqueda de contactos.

8. Salud laboral

Los servicios de salud laboral se ocupan de la prevención de riesgos derivados del trabajo, y habitualmente utilizan sistemas de información e historias clínicas propias. Se discute sobre la conveniencia de establecer intercambios de información entre los servicios de salud asistenciales y los de salud laboral. En todo caso habrá que observar todas las precauciones sobre identificación, seguridad, trazabilidad y confidencialidad mencionadas.

9. Conclusiones y recomendaciones

- La información contenida en las historias clínicas y empleada en la asistencia sanitaria es abundante, heterogénea y de relevancia difícil de establecer previamente al cierre del episodio.
- Los sistemas de información sanitarios y de gestión de historias clínicas deben estar dotados de un sistema de trazabilidad, que permita la auditoría de los accesos a los mismos.
- Los profesionales sanitarios involucrados en la asistencia sanitaria no deben tener límites previos de acceso a la información del caso.
- Se debe permitir al acceso a historias clínicas al personal no sanitario involucrado en la asistencia sanitaria, de acuerdo a perfiles que no dificulten el desempeño diario.
- Todos los trabajadores de la salud, independientemente de su nivel y categoría y de si son sanitarios o no, están obligados a la confidencialidad.
- Es necesaria la formación sobre confidencialidad en las instituciones sanitarias.
- La historia clínica es un documento con valor legal que puede ser requerido por la autoridad correspondiente.
- La docencia y la investigación son dos funciones clave de las historias clínicas y, por extensión, del sistema de información.
- Debe permitirse el acceso a los datos necesarios para para poder realizar un proyecto de investigación articulado y coherente, estableciéndose los procedimientos formales para ello.
- Los sistemas de información basados en los propios datos del servicio de salud constituyen la base la información utilizada para la gestión administrativa, económica y clínica.
- Para el conocimiento de morbilidades y casuísticas es necesario el acceso a las historias e informes individualizados, con una metodología estandarizada y homologable.
- Para la gestión clínica es necesario proporcionar información sobre resultados de actividad y de salud.
- Los servicios de salud pública necesitan acceso a sistemas de información de los servicios de salud y a otros como servicios estadísticos, climáticos, agrícolas y ganaderos.
- Los servicios asistenciales tienen que facilitar la información relativa a personas concretas que sea necesaria en los casos de declaración obligatoria y para la búsqueda de contactos.
- Los sistemas de información de los servicios de salud laboral podrían integrarse con los de los servicios de salud correspondientes.

Bibliografía

- Asenjo Sebastián MA. Gestión diaria del hospital. 3ª edición. Ed. Masson. Barcelona 2006.
- Carnicero J, Chavarría M, Escolar F, et al. De la historia clínica a la historia de salud electrónica. 5 Informe SEIS. Ed. Sociedad Española de Informática de la Salud, 2002, Pamplona.
- Escolar F, Martínez-Berganza MT. Asistencia clínica en la cabecera del paciente. En: 6 Informe SEIS. El sistema integrado de información clínica. Ed. SEIS. Pamplona 2004. Pag: 95-122.
- Fletcher RH, Fletcher SW. Epidemiologia Clinica 4ª Edición. Lippincott Williams & Wilkins. 2009. México.

Capítulo III

Bioética y explotación de grandes conjuntos de datos

Pilar León Sanz

1. Introducción

La aplicación de la informática a la asistencia médica plantea un amplio número de cuestiones bioéticas. Una de ellas es la explotación de los grandes conjuntos de datos o *Big Data*. Cada vez es mayor el volumen y la variedad de datos almacenados relacionados con la salud. También han aumentado las posibilidades de la tecnología respecto al análisis de estos datos, lo cual es conocido por el término inglés *data mining* o minería de datos. Se trata de la aplicación de algoritmos a las grandes bases de datos, con el fin de descubrir patrones y tendencias hasta ese momento desconocidas.

La explotación de los grandes conjuntos de datos utiliza y combina métodos estadísticos, de aprendizaje automático, de reconocimiento de patrones y de gestión de base de datos. Se ha utilizado para desarrollar modelos predictivos, también en el ámbito de la salud. Los avances en este campo han dado lugar a la aparición de una nueva actitud hacia los datos, que son considerados como materia prima explotable para una variedad de propósitos diferentes a los que motivaron su recogida.

En general, en el ámbito de la salud hay un solapamiento en el origen de los datos: unos proceden de la asistencia médica, otros de la investigación, del área de la salud pública, del ámbito administrativo, o simplemente son incorporados como consecuencia del registro de actividades sociales. Todos ellos pueden tener interés en el ámbito sanitario en función de la aplicación de los algoritmos con los que son analizados[11].

Desde un punto de vista ético nos interesa señalar que el análisis de grandes masas de datos conlleva un proceso de objetivación de la información que pasa de un ámbito personal a otro colectivo, más amplio. En primer lugar, se ha dicho que puede dar lugar a una "des-individualización" de la información, puesto que las personas son tratadas como elementos, en lugar de como individuos.

En segundo lugar, aunque inicialmente los grandes datos se asociaban a las tres V (volumen, variedad y velocidad), cada vez es más difícil distinguir el tamaño de las fuentes de datos. El uso actual de "grandes datos" se refiere menos al tamaño de los conjuntos de datos involucrados y más al potencial para extraer información, ya sea directamente, ya sea por vinculación o combinación de diversos conjuntos de datos.

[11] Al debatir el nuevo proyecto de Reglamento de Protección de Datos de la Unión Europea se ha planteado si los "datos genéticos" debían tener una consideración especial, al margen de los datos de salud, debido al carácter identificativo y predictivo de esta información. Dada la dimensión del capítulo, no hemos entrado en la especificidad de esta cuestión. Cf. http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf [accedido 17 de mayo de 2016].

En tercer lugar, la repercusión ética de la minería de datos está en función del contexto, el momento y la finalidad del análisis de la información, elementos que pueden condicionar el significado y la repercusión de la información obtenida.

1.1. La valoración bioética

La explotación de grandes conjuntos de datos es una cuestión relativamente novedosa en el ámbito de la ética médica. Aunque hay algunos documentos marco, como "The collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues" del Nuffield Council on Bioethics (2015), o los documentos de la International Medical Informatics Association (IMIA) y de la American Medical Informatics Association (AMIA)[12], todavía es mucho lo que queda por hacer en este ámbito. Son, además, numerosos los aspectos en los que se mantiene un debate abierto.

En la valoración ética nos interesa distinguir entre lo que es ético y legal, cuestión esta última que es estudiada en otro capítulo, pero hay que advertir que, en general, las regulaciones y reglamentaciones aprobadas han tenido en cuenta los requisitos éticos derivados de la protección de datos de los pacientes y de los profesionales, aunque también se han puesto de relieve las notables diferencias entre las distintas legislaciones nacionales sobre esta cuestión[13].

Hay quien considera que la explotación de los grandes datos sería, en sí misma, éticamente neutra, como puede serlo cualquier otro tipo de metodología estadística[14]. Sin embargo, el uso de datos biológicos y de salud puede tener efectos tanto beneficiosos como perjudiciales. Así, por ejemplo, la explotación de los grandes conjuntos de datos puede llevar a comprender mejor los patrones de las enfermedades, de modo que puede facilitar el diagnóstico o el pronóstico y tratamiento médicos. Pero, al mismo tiempo, puede dar lugar a abusos en el respeto debido a la confidencialidad de las personas.

Se ha de promover la utilización de los datos de forma ética y responsable, que busque el interés público y que, al mismo tiempo, concilie los intereses relevantes de los individuos y otros grupos de personas, y respete sus derechos fundamentales.

Tabla III.1. Principios recomendados por el Nuffield Council
El principio de respeto a las personas
El principio de respeto de los derechos humanos
El principio de la participación de las personas con intereses moralmente relevantes
El principio de la responsabilidad de las decisiones
Fuente: Nuffield Council. The collection, Linking and Use of Data in Biomedical Research and Heatlh care: Ethical Issues 2015, p. 84.

El análisis ético-médico de los diversos usos de los grandes datos también ha de considerar las siguientes cuestiones: cuáles son los objetivos de la explotación de datos; quién debería utilizar los

[12] Ambas instituciones han organizado diversos seminarios y grupos de trabajo sobre la cuestión. Por ejemplo: 8th International Workshop on Biosignal Interpretation organizado por la IMIA en noviembre 2016; Data Mining and Big Data Analytics WG IMIA WG/SIG Report (2012). Se trata de un aspecto incluido en sus respectivos códigos éticos: Code of Professional and Ethical Conduct; Principles of professional and ethical conduct for AMIA members (2013); y Code of Ethics for Health lnformation Professionals, IMIA (2011).

[13] Verschuuren et al. (2008). Publicación realizada por el Work Group on Confidentiality and Data Protection of the Network of Competent Authorities of the Health Information and Knowledge Strand of the EU Public Health Programme 2003–2008.

[14] Seltzer (2005).

resultados de esos análisis; cómo y por quién se ha de hacer un balance entre los beneficios y los riesgos que esta tecnología puede suponer para los profesionales, para los pacientes y para la sociedad[15].

Con estos criterios, vamos a referirnos a algunos aspectos bioéticos específicos de la minería de datos. En primer lugar, ofreceremos un análisis ético-médico de dos cuestiones clave, comunes a los distintos usos de la explotación de los grandes datos: la fiabilidad de los análisis y el respeto a la privacidad o confidencialidad de los datos, tanto de los individuos como de los grupos de personas.

Posteriormente nos referiremos a la aplicación de la explotación de datos en el ámbito de la salud pública. Revisaremos la ética de algunos usos de la minería de datos en la asistencia clínica y en la gestión de la salud, y consideraremos dos cuestiones que han generado muchas suspicacias, como la explotación de datos de las prescripciones farmacéuticas o la utilización de los grandes datos por parte de las compañías de seguros.

2. Dos cuestiones éticas importantes

2.1. Precisión, validez y aceptación

En marzo de 2016, la revista *Anesthesiology* publicó el artículo titulado "A New Model for Predicting Postoperative Mortality"[16], en el que se analizaban datos de 5,5 millones de pacientes intervenidos en 958 hospitales franceses para valorar la influencia de factores no cardíacos en los índices de mortalidad en el postoperatorio de las cirugías cardíacas. Como resultado, se identificaron 15 factores con un valor predictivo significativo en ese tipo de pacientes.

Este es un ejemplo de cómo el análisis de grandes datos es una herramienta que proporciona información y, en este caso concreto, facilita la adopción de protocolos basados en las llamadas pruebas estadísticas o científicas. El artículo citado también muestra que entre los requisitos éticos fundamentales de la minería de datos se encuentran la corrección técnica del análisis de los datos, la precisión y el rendimiento estadístico[17]. De otra manera, la información obtenida puede estar sujeta a sesgos y a errores, lo que no permite conseguir estándares adecuados de atención clínica.

A propósito del estudio citado anteriormente, un grupo de especialistas del New England Journal of Medicine (Journal Watch, March 24, 2016) comentaba que el modelo predictivo propuesto podría resultar útil para informar y aconsejar a los pacientes acerca de los riesgos de una intervención, si están indecisos, pero el estudio estaba limitado porque el análisis no había reflejado el grado de gravedad de las enfermedades asociadas, y era necesario volver a validar los datos antes de generalizar su aplicación.

No existe un único método de clasificación de datos, por lo que se ha de encontrar el algoritmo de clasificación que mejor se adapte a los objetivos o al conjunto de datos que se va a analizar. De ahí la importancia de la corrección en el diseño de la investigación en la minería de datos que incluye cuestiones éticas y técnicas. El alcance de este punto es mayor si se tienen en cuenta las limitaciones derivadas de la incertidumbre en relación a la exactitud de los datos y respecto al poder estadístico de los análisis.

[15] Marckmann, Goodman (2006).
[16] Le Manach et al. (2016).
[17] Al-Sagaf, Tavani (2013).

La precisión intrínseca de los datos puede variar según su origen, o según el modo en que se han introducido, transcrito o manipulado. También influye la formación, experiencia e intencionalidad de los analistas porque quien lleva a cabo el análisis puede favorecer un algoritmo sobre otro; y hay que evitar el sesgo del experto en minería de datos que pone a punto un determinado algoritmo, en lugar de otros. Además, hay que evitar el intento de ajustar el rendimiento de cada algoritmo al conjunto de datos o a los objetivos de la investigación.

En la corrección ética del diseño de nuevos algoritmos influye, en primer lugar, la definición de los objetivos del estudio, de forma que los posibles beneficios justifiquen la manipulación de los datos, y que se demuestre la adecuación del grupo de datos incluido en el análisis, para que no se extrapolen los resultados más allá del alcance del estudio.

En segundo lugar, como veremos más adelante, en la ética de la explotación de los grandes datos es de gran importancia la confidencialidad y la privacidad de la información utilizada. Y en tercer lugar, hay que tener en cuenta la idoneidad y validez de los métodos empleados, por lo que hay que evaluar adecuadamente los algoritmos del análisis de los datos[18]. Esta es una cuestión que, aunque puede ser difícil en ocasiones, resulta imprescindible[19], sobre todo si se tiene en cuenta que con frecuencia los resultados de los estudios son usados para nuevos análisis. En cualquier caso, los desacuerdos sobre la precisión de los resultados obtenidos deberían ser abordados antes de su aplicación en la atención a los pacientes.

2.2. Conceptos innovados de privacidad, confidencialidad y consentimiento en el uso de los datos

La privacidad, el derecho a la confidencialidad y el modo de conservar la información han sido cuestiones importantes en la implantación de la tecnología de la información en Medicina. De ahí la presencia constante de estos aspectos en el desarrollo de la historia clínica electrónica y de cualquier aplicación informática relacionada con el cuidado de la salud[20].

La confidencialidad tiene por objeto garantizar que la información proporcionada por una persona no sea divulgada posteriormente sin su autorización (excepto en los casos establecidos)[21]. En el ámbito específico de la explotación de grandes conjuntos de datos, la privacidad y el derecho a la intimidad hacen referencia a la posibilidad de decisión sobre el flujo de la información personal. Es decir, a la capacidad de las personas para restringir el acceso y mantener el control sobre el uso y la circulación de su información personal, incluyendo la transferencia y el intercambio de dicha información.

Tradicionalmente se ha dicho que los datos o la información tienen carácter personal si están ligados a un nombre, es decir, si están identificados. Hasta ahora ha sido un punto clave en el manejo ético de datos, ya sea en el ámbito de la asistencia o de la investigación. Sin embargo, en la explotación de grandes conjuntos de datos hay que reconsiderar este criterio. Tampoco son de aplicación algunos criterios ético-deontológicos respecto a la confidencialidad y las medidas que se utilizan para salvaguardarla

[18] Seltzer (2005).
[19] Las publicaciones insisten en esta cuestión y aportan diversas alternativas. Anderson, Aydin (1997); Goodman (2015).
[20] León Sanz (2008).
[21] Al tratar este tema, el Nuffield Report insiste en la distinción entre términos “privacidad” y “confidencialidad”. La privacidad se referiría al interés de las personas respecto a quién tiene acceso a uno mismo, a sus hogares y a sus propiedades, o a la información sobre ellos. La privacidad llevaría a la restricción selectiva y voluntaria de la información propia, supeditada al buen uso por parte de quienes acceden a ella (pp. 46-49).

(anonimización, codificación, etc.); la posibilidad de obtener el consentimiento para el uso de los datos; y el modo en que se conserva la información médica[22].

Veamos a continuación algunos aspectos que han contribuido a la transformación comentada:

a) La reutilización de los datos y los usos secundarios. Aprovechar los datos evita el coste y los inconvenientes de volver a recabar la misma información para objetivos diversos. Así, por ejemplo, los datos clínicos pueden servir para la planificación de servicios de salud, para la investigación médica o, en el caso de las compañías de seguros, para fines actuariales, etc.

La reutilización o el uso o usos secundarios de las bases de datos y las vinculaciones o combinaciones de diversas fuentes implican nuevas oportunidades, pero desde un punto de vista bioético hay que advertir que el cambio de contexto o de finalidad puede llevar a que los datos adquieran también sentido y significado diferentes[23]. Por ejemplo, si las fuerzas del orden o de seguridad tuvieran acceso a bases de datos clínicos, los indicadores de salud o enfermedad se podrían convertir en "indicadores de culpabilidad"[24].

Por otra parte, es frecuente que una persona muestre una sensibilidad distinta respecto al uso de sus datos según la finalidad para la fueron proporcionados. En este sentido, sería diferente si han sido facilitados para fines clínicos o para una investigación. En otros casos, la información personal (de salud sexual, salud mental,…) puede ser más o menos delicada, según las circunstancias o el contexto social.[25]

Además, la combinación de conjuntos de datos puede dar lugar a enlaces más o menos temporales, e incluso permanentes. Un ejemplo podría ser la vinculación de los datos de registros de enfermedades con la localización de contaminantes ambientales para examinar o vigilar algún vínculo, incluso futuro, entre ellos.

Cada vez se diseñan algoritmos más sofisticados que permiten correlacionar y "extraer", de las bases de datos existentes, nuevos puntos de vista y nueva información. Como resultado, la utilidad potencial de un determinado conjunto de datos es también más imprevisible.

Estas posibilidades técnicas implican que no siempre sea posible obtener el consentimiento o la autorización de las personas individuales para el uso secundario de los datos médicos. Además, también puede ser ardua para un participante no experto la comprensión adecuada de las posibilidades técnicas de la información, lo que también condicionaría el consentimiento.

Por otra parte, esta realidad (los usos secundarios no previstos de los datos, a través de la combinación de distintos parámetros y la aplicación de nuevos algoritmos a la información) lleva a que no se puedan garantizar a las personas los derechos de acceso, rectificación, cancelación y oposición. Tampoco se puede establecer fácilmente la forma en que puedan ser retirados de los proyectos de minería de datos.

[22] Los NHI han establecido un programa de conocimiento (BD2K) que tiene como objetivo formar y ayudar a los investigadores del área biomédica en el buen uso de los grandes volúmenes de datos: http://bd2k.nih.gov/about_bd2k.html#bigdata.

[23] Goodman (2015), p. 123,

[24] Nuffield Report (2015), p. 18.

[25] Un análisis de las amenazas y de los posibles daños derivados del mal uso de los datos, en Laurie et al. (2014).

b) En la minería de datos, la codificación o la anonimización de los datos puede no ofrecer suficiente protección. En general, hasta ahora se consideraba que si los datos eran de dominio público, eran anónimos o estaban anonimizados, no era necesario requerir la aprobación de los interesados para su utilización. Sin embargo, la combinación de bases de datos puede llevar a identificar a personas singulares o a grupos que estaban de forma anónima en alguna de las colecciones de datos.

Por ejemplo, con fines de investigación se permitió la combinación de una base de datos anonimizada proporcionada por la Group Insurance Commission de Boston (no constaban los nombres, las direcciones, los números de seguridad social, ni cualquier otro tipo de información identificativa), con la base de datos de los votantes del Estado (que incluía nombre, código postal, dirección, sexo, fecha de nacimiento) que es de dominio público. Tras la combinación fue posible identificar a ciudadanos concretos, y de hecho se publicaron los datos médicos del entonces gobernador de Massachusetts[26].

Otro caso más frecuente y que revisaremos más adelante ha sido la venta de datos de prescripción médica a la industria farmacéutica, con el fin de conocer los hábitos de prescripción de los médicos y evaluar la eficacia de las estrategias de mercadotecnia de la venta de medicamentos.

Por lo que los expertos indican, es difícil garantizar que no se vaya a identificar una persona o un grupo de personas cuando se combinan bases de datos de diferente procedencia. Depende de qué herramientas se utilicen y de qué otra información esté disponible. De ahí la importancia de que se desarrollen técnicas informáticas y una regulación adecuada que sirvan para preservar la privacidad de datos tan sensibles como los relativos a la salud de las personas[27].

c) El planteamiento del 'opting out', o de la necesidad de excluirse, en la cesión de datos. Se tiende a dar por supuesto que, cuando una persona cuelga una página en la red, publica información en las redes sociales, utiliza una aplicación del móvil o se conecta con otros a través del correo electrónico, los datos que maneja pueden ser utilizados por otros en un futuro[28].

El planteamiento general respecto a la cesión automatizada de datos, también en el ámbito de la salud, es que, si no se hace una manifestación en contra, se considera que pueden utilizarse los datos que quedan registrados como consecuencia de la actividad informática. Se trata, como reconoce el informe del Nuffield Council (n. 6.32), de una cuestión de debate actual que está lejos de estar resuelta. El hecho de que este procedimiento se esté generalizado no implica que sea el modo más idóneo para proteger a las personas.

d) La comunicación voluntaria de datos y el fenómeno del Crowdsourcing. A través de encuestas y estudios de campo se percibe que ha cambiado la opinión de algunos sectores de la sociedad sobre la utilización de los datos con interés público. La novedad supone primar el interés del conjunto por encima del derecho individual a la privacidad.

Así, por ejemplo, cuando los ciudadanos de Australia Occidental fueron consultados sobre la utilización, para fines de política y gestión sanitaria y otras investigaciones, de la base de datos de más de

[26] Nuffield Report (2015), p. 67; Sweeney (2002); Wel, Royakkers (2004).

[27] El problema de preservar la anonimización es mayor conforme aumenta la capacidad de almacenar datos personales y se hacen más sofisticados los algoritmos de minería de datos. Se han sugerido diversas técnicas como la aleatorización y k-anonimato. Las líneas de trabajo de los diversos grupos son similares. Cf. Aggarwal, Yu (2008), pp. 11-52. También: Ohm (2009); Wel, Royakkers (2004).

[28] Al-Saggaf, Islam (2015).

tres decenios, que incluía todo tipo de registros personales relacionados con la morbi-mortalidad de la población, no sólo la respuesta mayoritaria fue de apoyo, sino que además se planteó por qué no estaba ya en uso para la investigación[29]. El informe del Nuffield Council, por su parte, destaca que en el Reino Unido también existe un amplio apoyo social a la utilización para fines secundarios de la información contenida en las grandes bases de datos, si tal uso contribuye a la mejora de la investigación o de la atención en el ámbito de la salud[30].

Otro ejemplo de este cambio de mentalidad es el llamado *crowdsourcing*, término aplicado desde 2005 al proceso por el que se pueden obtener ideas, datos, trabajos, dinero, etc., mediante el uso de Internet. De manera voluntaria, las personas responden a solicitudes de información y ponen datos a disposición de terceros, incluyendo los relacionados con la salud o las enfermedades que padecen[31].

La amplia disponibilidad de plataformas de redes sociales ha facilitado la investigación mediante una dinámica social diferente a la investigación institucional más formal. Este tipo de recogida de datos exige el compromiso de asegurar la protección de los intereses individuales, también en el proceso de traslación de los resultados a productos y prácticas clínicas.

Desde un punto de vista ético, es importante que se promueva el bien público para el conjunto de la sociedad, pero al mismo tiempo la propia sociedad también está preocupada respecto a mantener la privacidad y confidencialidad de datos personales tan sensibles como son los sanitarios. Por eso se ha subrayado que se ha de procurar salvaguardar ambos ámbitos. Por otra parte, el "interés público" no tiene por qué ser siempre contrario a los "intereses privados".

e) El almacenamiento de los grandes datos de la salud. Legal y deontológicamente ha habido a lo largo del tiempo una exigencia de custodia, por parte de los gestores de los centros de salud, de la información relacionada con la asistencia médica. En este aspecto hay que señalar que, hasta ahora, la deontología médica afirma que "es muy recomendable que el responsable de un servicio de documentación clínica sea un médico" (Código de Deontología, 2011, art. 19.3) y que "la historia clínica electrónica sólo es conforme a la ética cuando asegura la confidencialidad de la misma, siendo deseables los registros en bases descentralizadas" (Código de Deontología, 2011, art. 19.3; 19.9). Sin embargo, ninguno de los dos criterios son compatibles con los nuevos sistemas de almacenamiento de datos, como los "espacios en la nube", en los que se encuentran gran número de bases de datos[32].

[29] Meslin, Goodman (2014); cf. sobre esta cuestión: Nuffield Report, p. 132

[30] Nuffield Report, p. 56; la p. 133 cita los resultados de una encuesta europea que mostró que había una menor preocupación por la privacidad de los datos que por la posibilidad de controlar la información relacionada con el material biológico. Por su parte, Willison et al. (2003) encontraron que hay personas que quieren dar el consentimiento si la información personal va a ser utilizada para un segundo propósito.

[31] Hay diversas webpages y aplicaciones desarrolladas con esta finalidad. La iniciativa *PatientsLikeMe* fundada en 2004 cuenta con más de 400.000 seguidores. Las iniciativas de salud "participativos" implican el compromiso del buen uso de los datos por parte de los investigadores. Swan (2012).

El crowdsourcing se ha utilizado, sobre todo, en el área de la salud pública. Por ejemplo, en 2013 profesionales de la salud pública de la Universidad del Estado de Colorado, en colaboración con la Escuela de Salud Pública y el Departamento de Salud Pública y Medio Ambiente, creó una iniciativa (en formato wiki) para recabar información sobre las prácticas de producción de alimentos y los sistemas de distribución de los comestibles.

[32] El procedimiento se utilizó durante la década de 1990 en el ámbito bancario para las redes de cajeros automáticos. En 2006, Eric Schmidt, CEO de Google, comenzó a usar el término, que se hizo popular en su significado actual. Empresas como Gmail, iCloud y Salesforce ofrecen sus servicios a bancos, industrias farmacéuticas, compañías de seguros, empresas de marketing, consultoría e investigación, etc. Cf. Bruin, Floridi (2016).

El almacenamiento en la nube permite reducir los costes de hardware y soporte de los servicios informáticos: no requiere instalación ni actualizaciones y la potencia de cálculo supera con creces la de una instalación con ordenadores o servidores propios, por lo que cada vez se está generalizando más. En el caso de los datos de salud, facilita además el acceso simultáneo desde instalaciones y centros sanitarios diversos, lo cual es interesante puesto que cada vez son más las personas y entidades que participan en la prestación de asistencia sanitaria, y que necesitan acceder directamente a los registros de pacientes. También se está generalizando su uso en proyectos de investigación, sobre todo de carácter multicéntrico.

Este tipo de almacenamiento facilita las tareas informáticas de alta complejidad mediante la combinación de innumerables procesadores repartidos por todo el mundo, lo cual es un fenómeno nuevo desde el punto de vista ético médico. En febrero de 2016, el holandés Boudewijn de Bruin (Universidad de Groningen) y el inglés Luciano Floridi (Universidad de Oxford) reclamaban para sí la autoría del primer artículo que analizaba la informática de la nube desde el ámbito de la ética empresarial (2016)[33]. Ambos investigadores señalaban los riesgos éticos de la computación en nube, como por ejemplo: la privacidad del consumidor o usuario de la nube; la fiabilidad de los servicios; la propiedad de los datos; y la explotación de las bases de datos depositadas en la nube por parte de empresas de marketing.

Para estos autores la clave de la ética del uso de la nube, como modo de almacenamiento, sería la transparencia. De hecho, muchos usuarios de este sistema ("clouders") no son conscientes de lo que supone depositar los datos en la nube. Las empresas de alojamiento, por ejemplo, pueden mover los datos depositados por los clientes de un centro de datos a otro, con el fin de permitir un uso más eficiente del espacio de almacenamiento[34]. Además, es frecuente que muchas empresas que ofrecen servicios de computación en nube estén localizadas en países diferentes del que las utiliza, por lo que la regulación deontológica y legal de protección de los datos también varía.

En estos momentos se debate si éticamente es aceptable utilizar estos sistemas de almacenamiento en el caso de información especialmente sensible, como es el caso de despachos de abogados, de datos militares o de datos médicos. En esta discusión, la opinión de Bruin y Floridi sería negativa.

3. Retos éticos en la aplicación del análisis de los grandes datos a la asistencia médica

3.1. Explotación de los grandes datos para uso epidemiológico y de salud pública

Una de las aplicaciones más importantes de la explotación de las grandes masas de datos en medicina es la salud pública y la epidemiología. El análisis de los datos masivos permite identificar correlaciones entre condiciones ambientales, estilos de vida y comportamientos sociales, por un lado, y morbi-mortalidad, por otro. Además, el diseño de estos estudios implica en muchos casos el establecimiento de redes nacionales e internacionales con el fin de agrupar el mayor número posible de datos.

Revisemos con un caso reciente las grandes oportunidades y algunas consecuencias y limitaciones éticas del análisis de las grandes bases de datos en esta área. El 10 de abril de 2016 se publicó un estudio en la revista JAMA sobre la asociación entre ingresos y esperanza de vida en los Estados Unidos (2001-2014)[35]. Se trata de una investigación dirigida por David Cutler (Harvard University) que incluyó más de 1,4 billones de observaciones, sobre personas entre 40 a 76 años, y año. El estudio compara los ingresos promedio por

[33] Esta cuestión está analizada también en el Nuffield Report (2015), p. 142-144.
[34] Bruin, Floridi (2016), p. 10.
[35] Chetty et al. (2016).

hogar entre las personas que trabajan, obtenidos de las declaraciones de impuestos de forma anonimizada, por un lado, con las cifras de mortalidad obtenidas de los registros de mortalidad de la Seguridad Social, por otro. El análisis tuvo también en cuenta la raza, el sexo y el área geográfica, con el fin de evaluar los factores asociados con las diferencias en la esperanza de vida.

Los resultados han sido tan sorprendentes (diferencias de esperanza de vida entre 10 y 15 años, según fueran mujer u hombre, en el mismo país), que los autores han publicado la investigación con acceso libre, con el fin de que sirva a gobernantes, especialistas en salud pública, agentes sociales, etc., para diseñar políticas asistenciales –sociales y médicas– que puedan contribuir a disminuir las desigualdades respecto a la esperanza de vida.

Del estudio también se derivan algunas advertencias sobre el uso no matizado de la información obtenida del análisis de los grandes datos. En efecto, la conclusión inicial del estudio de Cutler coincide con otros trabajos anteriores: la esperanza de vida aumenta con el mayor nivel de riqueza. Pero el nuevo análisis ha puesto de manifiesto que las correlaciones entre esperanza media de vida, riqueza y estilos de vida son más complejas de lo que se pensaba. Es decir, el simple análisis de millones de datos no acerca a la realidad si aquellos no incluyen los algoritmos adecuados.

En el ámbito de la Salud Pública identificamos algunas cuestiones bioéticas problemáticas. Como se ha comentado anteriormente, una preocupación ético-médica importante es la correcta utilización y custodia de los datos clínicos. En la mayoría de los casos se trabaja con datos anónimos o anonimizados. Se han comentado ya los riesgos que conlleva la combinación de las bases de datos. También hay estudios que incluyen amplias cohortes de pacientes o de sujetos sanos. En estos casos, y si el diseño prevé la actualización de datos personales y de salud, se ha de advertir a los participantes cómo se va a llevar a cabo la custodia de los datos. Estas precauciones serán mayores si se agregan datos médicos, psicológicos o psiquiátricos, genéticos, de estilo de vida, etc., junto con otros de tipo social, geográfico o económico. Además del compromiso de los investigadores y la supervisión de las instituciones para salvaguardar la confidencialidad, será necesario contar con el consentimiento de los pacientes, lo que les convierte en protagonistas activos en esos trabajos[36].

En segundo lugar, hay que advertir que los perfiles obtenidos por la minería de datos pueden aportar informaciones relevantes para establecer recomendaciones y políticas de salud pública, pero también pueden llevar a discriminar algún grupo de población, ya sea por nivel socio-económico, por riesgo de desarrollar enfermedades, por una estimación de un menor rendimiento o eficiencia de las medidas a adoptar, etc. Desde el punto de vista ético, resultan especialmente sensibles los análisis que correlacionan información sobre discapacidades, enfermedades mentales, adicciones, delincuencia juvenil, cuestiones políticas o religiosas, etc.

En tercer lugar, existe el riesgo ético de que los resultados de los análisis de grandes datos lleven a implantar políticas que, con el fin de conseguir una población más saludable, condicionen la libertad de actuación y la vida de las personas imponiendo modelos de vida opcionales como obligatorios.

Por último, los estudios de Salud Pública tienen también consecuencias sociales y económicas que exceden el ámbito sanitario. Basta recordar las repercusiones tan negativas, que para la industria cárnica

[36] Willison et al. (2003).

mundial, tuvo el informe de la Organización Mundial de la Salud sobre el posible efecto carcinogénico del consumo de carne roja y de carne procesada (2015)[37].

3.2. Apoyo a la decisión clínica: entre la subjetividad de la decisión y la objetividad de los datos

El Instituto de Medicina de EE.UU publicó en septiembre de 2012 el informe "Best Care at Lower Cost: The Path to Continuously Learning Health Care in America", en el que señalaba el gran potencial que tenían para la medicina los avances informáticos y los análisis de los datos de salud, por las mejoras que introducen en la práctica de la clínica. Uno de esos beneficios es el apoyo que proporcionan en la toma de decisiones clínicas.

La búsqueda de algoritmos computerizados para tomar decisiones clínicas proporciona resultados que llevan a elaborar guías, recomendaciones útiles y sólidas desde el punto de vista estadístico. Por mucha experiencia que tenga un profesional, siempre estará basada en un número limitado de casos, mientras que la minería de datos facilita cierta "objetivación". Se trata de una cuestión que conecta con el debate sobre la subjetividad u objetividad en que se ha de basar la decisión clínica.

En este sentido, hay que advertir que los métodos estadísticos no eliminan la responsabilidad del profesional. Tampoco eliminan totalmente la incertidumbre propia de la decisión clínica. El razonamiento médico ha de tener en cuenta los valores, las necesidades y las prioridades de los pacientes individuales, lo cual no es una habilidad computable.

Por lo tanto, la búsqueda de algoritmos computerizados para tomar decisiones clínicas es una estrategia importante, pero no debe ser sobreestimada porque no puede determinar, de antemano, una decisión particular. Lo que sí ofrece son pautas basadas en evidencias estadísticas que aunque han de ser muy tenidas en cuenta, en ocasiones pueden no ajustarse a los casos concretos.

3.3. Bioética, gestión sanitaria y grandes conjuntos de datos

En el ámbito sanitario, las bases de datos que incluyen datos de la actividad de los profesionales y de los pacientes permiten crear perfiles de práctica médica. Esta información es utilizada por quienes organizan o sufragan la asistencia sanitaria, ya sean sistemas públicos o privados, agencias reguladoras autonómicas, estatales, o de un ámbito internacional.

Los gestores de los servicios de salud y de las instituciones sanitarias (públicas o privadas) están muy interesados en una información que permita implantar sistemas eficaces de contención de costes, de gestión de riesgos y de programas de seguridad y de garantía de calidad[38]. Es indudable la utilidad de la información de los grandes datos para la seguridad de los pacientes. Facilita el rigor y la evidencia estadística, establece programas para evitar la iatrogenia y aumentar la calidad asistencial. La American Medical Informatics Association, por ejemplo, ha señalado repetidamente la ayuda que supone la explotación de los grandes datos para reforzar la seguridad de los fármacos y evitar los efectos adversos[39].

[37] WHO, *Q&A on the carcinogenicity of the consumption of red meat and processed meat*. October 2015. http://www.who.int/features/qa/cancer-red-meat/en/ [Accedido 2 de mayo de 2016].
[38] Al-Saggaf (2015).
[39] AMIA: Medical Data Mining Strengthens Drug Safety Monday, May 16, 2011.

Sin embargo, también son numerosas las voces que denuncian que los hospitales, públicos y privados, y las organizaciones de atención médica, como las compañías de seguros, utilizan esa información casi exclusivamente para controlar los costes y evaluar el rendimiento de los profesionales, en lugar de para garantizar la calidad de la atención[40]. Esto puede dar lugar a que en los Centros de Salud o en los Departamentos de los hospitales se incentive un determinado perfil de prescripción, tanto diagnóstica como terapéutica, de acuerdo con estándares establecidos por los perfiles más prevalentes o más deseables, según la política de salud establecida. Una crítica importante apunta que uno de los principales fallos de las decisiones así tomadas es el carácter inexacto o sesgado que pueden tener los análisis: pueden no haberse incluido todos los datos relevantes o faltar un análisis de la variabilidad, en un campo tan complejo[41].

4. Dos casos especiales

4.1. La minería de datos en el contexto de las compañías de seguros

Se ha dado gran importancia al análisis de grandes datos en el ámbito de las compañías de seguros y mutuas. Se trata de entidades privadas que conservan grandes cantidades de información personal sobre sus asegurados puesto que, cuando una persona concierta un seguro de salud, la empresa recoge sobre ella mucha y variada información. Parte es de carácter administrativo (edad, sexo, estado civil, lugar de residencia, trabajo, etc.) y se complementa con información sobre la salud y las enfermedades padecidas, información genética, información sobre la salud mental, etc. Con estos datos, las entidades aseguradoras establecen exclusiones, valoran los riesgos de las solicitudes y proponen la cuantía de las pólizas o de las primas.

Además, es habitual que la firma de una póliza incluya la cesión de los datos con fines de reaseguro. El volcado de datos de las diversas compañías crea bases de datos de mayores dimensiones y pueden generar nueva información[42]. El asegurado queda en la ignorancia tanto de cómo se van a utilizar los datos, como de los resultados que se pueden obtener.

En este ámbito, la minería de datos se relaciona fundamentalmente con la definición de perfiles de asegurados o de posibles clientes. De esta manera se podrían identificar las personas con más riesgo, por lo que los análisis de los grandes datos pueden influir en la posible discriminación en la selección de los asegurados. También se ha comentado que la minería de datos puede servir para hacer valoraciones estimadas de posibles asegurados en función de las circunstancias económicas, con el fin de ofrecer (o imponer) a los usuarios pólizas de diferentes precios, o bien para restringir el acceso a las compañías.

Las entidades aseguradoras defienden la realización de estos estudios porque –según explican– tienen efectos beneficiosos, ya que daría la oportunidad a los seguros de planificar estrategias de intervención y prevención adecuadas para los asegurados. Esta “discriminación justa” serviría para calcular los riesgos y adecuar el importe de las pólizas y de los fondos de reservas económicas.

En la práctica, la minería de datos de los seguros de salud ha llevado a establecer correlaciones inesperadas. Por ejemplo, se ha demostrado la eficacia para la detección del fraude. Al-Saggaf indica que en 2003, en los Estados Unidos, el coste del fraude en los seguros de salud se estimó en 170 billones de dólares,

[40] Anderson (2002).
[41] Goodman (1999), p. 63.
[42] Borna, Avila (1999).

y la aplicación de algoritmos de datos diseñados para detectar el fraude, llevó a una disminución del fraude de 11,5 millones de dólares en un año[43].

También en este ámbito existe una corriente mayoritaria que reclama la elaboración y promulgación de controles y restricciones, en forma de leyes o reglamentos, para el uso de la explotación de datos por parte las compañías de seguros de salud.

4.2. La minería de datos de prescripciones de fármacos y la protección de los intereses de los pacientes

Son conocidos dos casos judiciales sobre la venta de datos de prescripción para la comercialización de productos farmacéuticos que se produjeron en Estados Unidos (Sorrell versus IMS Health Inc. et al. en 2011) y en el Reino Unido (Rv. Department of Health, Ex Parte Source Informatics Ltd., 2000)[44]. En el caso del Reino Unido se permitió la venta de datos de prescripción de fármacos porque los datos de los pacientes (o compradores de fármacos) se habían anonimizado, por lo que se entendió que no les causaba ningún perjuicio. Además, como fueron los farmacéuticos los que enviaron los datos tampoco se consideró que se lesionaba el acceso a su identidad[45].

En ambos casos se demostró que la venta de datos de prescripciones supuso un beneficio económico directo para las farmacias, que habían recibido la bonificación; para las empresas de minería de datos que hicieron los análisis; y sobre todo para la industria farmacéutica, puesto que permitió orientar las actividades de mercadotecnia de los fármacos y el trabajo de los representantes de los laboratorios que conocían de antemano las tendencias de prescripción de los profesionales que visitaban.

Se trata de una cuestión en la que se enfrentan intereses múltiples tanto de orden público, como privado: intereses de los pacientes, de los profesionales y de la industria y que, además, tiene consecuencias respecto a los costes de salud.

Como una primera valoración bioética, se puede afirmar que utilizar con fines lucrativos los datos obtenidos de las actividades asistenciales o de investigación, erosiona la confianza de la sociedad en el ámbito biomédico.

4.2.1. Datos necesariamente informatizados

La valoración de la divulgación y la venta de datos de prescripción ha de tener cuenta que se trata de datos recogidos de forma obligada. Por ley, es necesaria la receta médica para acceder a muchos medicamentos. Se trata de una información en la que obligatoriamente quedan nominalmente identificados los pacientes y los médicos. Por ley, los farmacéuticos que dispensan los medicamentos han de conservar la información de las prescripciones que han distribuido. Como toda esa información se recoge por medios informáticos, es fácil de agregar, procesar y vender.

[43] Al-Saggaf (2015), p. 282. La misma cifra es estimada en Yoo et al. (2012), p. 2441. Cf. también Kuo-Chung, Ching-Long (2012).
[44] Kaplan (2015); Orentlicher (2010).
[45] En el caso de Sorrell versus IMS Health Inc., el Tribunal Supremo de Estados Unidos revocó una ley que decía que "las empresas de minería de datos, para obtener datos de los proveedores individuales de registros de recetas necesitaban tener autorización de los particulares" (Petersen et al. 2013, 35).

Cuando las personas desarrollan relaciones con médicos y farmacéuticos, tienen derecho a la seguridad de la información sobre su condición médica. A veces las recetas proporcionan sólo pruebas indirectas de la salud de un paciente, pero en otros casos señalan directamente a un diagnóstico. La prescripción de efavirenz o tenofovir conlleva un diagnóstico infección de VIH, y si un paciente está siendo tratado con olanzapina (Zyprexa) se puede sospechar razonablemente que puede tener una enfermedad mental.

Otros ejemplos del riesgo que suponen para salvaguardar la confidencialidad de los datos clínicos son las colaboraciones que se han dado entre organismos públicos y privados con ánimo de lucro. Así, en Canadá, el Consejo de Inversiones del Plan de Pensiones de Canadá y TPG Capital adquirieron IMS Health en 2010[46]. O bien cuando la agencia gubernamental eSalud (Ontario) explotó a lo largo del tiempo los registros de diabéticos a través de una base de datos que integraba directamente los valores de laboratorio de los pacientes. Ante estos casos, hay que recordar que retirar la identificación de los datos del paciente, o añadir un código para permitir el seguimiento temporal del enfermo no resuelve el problema de la privacidad porque, como se ha comentado, la combinación de bases de datos puede facilitar la identificación de los pacientes o de los profesionales.

Además, estos episodios han reabierto de nuevo el debate sobre la propiedad de los datos de salud.

4.2.2. Algunas actitudes de los médicos ante la minería de datos

La comercialización de los datos de salud afecta también a las normas profesionales y tiene una implicación directa sobre la prescripción médica. Conocer los perfiles de venta de fármacos influye tanto positiva como negativamente en las prácticas de prescripción, y lleva a pensar en la vulnerabilidad de los prescriptores y de los pacientes frente a la industria farmacéutica. La industria puede modificar los precios de los medicamentos en función de los análisis de mercado, o puede plantear una comercialización agresiva que lleve al aumento de precios de los medicamentos y a modificar las formas de publicidad. Esto puede a su vez tener efectos perversos adicionales, porque el aumento del coste farmacéutico aumenta la inequidad y la discriminación en el acceso a la atención médica.

Por ello resulta sorprendente la pasividad de ciertos sectores de las profesiones de la salud ante el intercambio comercial de los datos. En 2006, La Asociación Médica Americana (AMA) estableció el llamado “AMA’s Physician Data Restriction Program” (PDRP). Se trata de un acuerdo de cesión del fichero general de datos de los médicos a la industria farmacéutica[47]. El Programa establece un sistema “opting-out”, según el cual, salvo que un médico rechace su participación, se facilita a las compañías farmacéuticas el acceso a los datos del profesional para fines comerciales y de investigación. En palabras del vicepresidente de la AMA, Roberto Musacchio, la cesión a la industria farmacéutica iba a “beneficiar a los médicos porque podrán recibir visitas de representantes farmacéuticos para la presentación de productos terapéuticos en los que realmente estén interesados”[48]. Parecen lógicas las críticas que se hicieron sobre el PDRP en la siguiente reunión anual de la Asociación Médica Americana (2007). Sin embargo, y pese a esas vigorosas protestas, en la práctica han sido muy pocos los médicos que realmente se han preocupado de la cuestión[49].

[46] IMS Health es una compañía de tecnología de la información sobre salud que facilita información sobre enfermedades, tratamientos y costos a instituciones públicas y privadas de más de 100 países.
[47] Kaplan (2015).
[48] Barclay (2007)
[49] Barclay (2007), p. 3.

La venta de datos de prescripciones farmacéuticas es un asunto que tiene dimensiones internacionales o supranacionales, puesto que, en el caso de IMS Health Inc., la empresa que manejó los datos era subsidiaria de Wolters Kluwer Pharma Solutions, radicada en otro país.

Las sentencias de los dos casos comentados en el inicio de este apartado han abierto un debate que insiste en la necesidad de que haya transparencia en las relaciones entre la industria y los gestores sanitarios.

5. Conclusión: la propuesta de una mayor regulación y formación de los profesionales

La creciente dependencia de las tecnologías de la información constituye una de las tendencias más notables en la atención de la salud durante los últimos años. En ese marco, la explotación de las grandes bases de datos o *Big Data* son algo más que un gran número de fuentes de datos. Es un término que hace referencia a la complejidad, a los desafíos y a las nuevas oportunidades que presenta el análisis combinado de los datos.

En general, se puede decir que el buen uso y las buenas prácticas estarán en función de las mejoras que pueda introducir en la atención al paciente individual y en la salud de la sociedad en general. Es innegable que el análisis de los grandes datos tiene consecuencias para ambas esferas. Hemos visto que la explotación de las grandes bases de datos es una práctica social compleja, donde existen tensiones y posibles conflictos de intereses[50].

Hay una corriente mayoritaria que reclama la elaboración y promulgación de más controles, en forma de leyes, reglamentos o directrices que proporcionen confianza social y seguridad en relación con la minería de datos. Sería un modo de reducir el posible daño a las personas y proteger los derechos humanos básicos.

Antes de concluir, hay que hacer referencia a un último aspecto: la necesidad de la formación de los profesionales de la salud en esta área. La explotación de los grandes datos en medicina es relativamente nueva. Es un campo en la que hay continuos avances, por lo que hay que subrayar el imperativo ético de procurar una formación, también bioética, adecuada.

Desarrollar guías de buenas prácticas médicas sobre el uso de las grandes bases de datos en sanidad es importante, pero no suficiente. Los profesionales tenemos que intentar entender los métodos y los usos de estas nuevas técnicas. Sólo así podremos valorar los diseños de aplicación, los resultados, también los "hallazgos incidentales" o inesperados, o qué efectos tienen sus informes sobre los pacientes, las familias y los colegas profesionales. Sólo con ese entendimiento se podrá salvaguardar bien la confidencialidad, el consentimiento o los intereses de los pacientes y de la sociedad. Tenemos por delante un largo camino por recorrer.

Además, se han de construir espacios de reflexión ética que permitan dar razón y proponer modos de hacer basados en la búsqueda de soluciones buenas, no sólo de las consideradas aceptables.

A pesar de los numerosos logros obtenidos mediante la aplicación de la informática a los cuidados de la salud, Kenneth Goodman ha recomendado a lo largo de los años mantener un equilibrio entre el entusiasmo servil de los partidarios de las tecnologías de la información y el escepticismo hipercrítico de

[50] Anderson, Aydin (1997).

quienes rechazan cualquier avance en esa dirección. Este autor propone una postura de "progressive caution", en la que el profesional de la salud incorpore la utilización de nuevas herramientas de análisis, sin sobrepasar los límites éticos. Sólo así se podrían prever y resolver las posibles contradicciones[51].

Tabla III.2. Objetivos de la IMIA para la explotación de los grandes datos en el área Biomédica.
1. Difundir la aceptación de la aplicación de la Inteligencia Artificial (IA) a la minería de datos.
2. Fomentar el debate y la difusión de nuevos métodos de IA. Promover plataformas y soluciones estandarizadas.
3. Proporcionar un foro para presentación nuevas implementaciones y revisar mejores prácticas.
4. Centrarse específicamente en: — Predicción en medicina clínica — Genómica funcional — Investigación de fenotipos moleculares — Evaluación de riesgo clínico — Minería de datos temporal en medicina y bioinformática — Computación evolutiva en el avance del conocimiento biomédico
Fuente: Data Mining and Big Data Analytics WG de la IMIA, 2012. http://www.imia-medinfo.org/new2/sites/default/files/wg-datamining-aug13ga.pdf

La sociedad reconoce el enorme potencial del análisis de los grandes datos. Necesitamos esos análisis para desarrollar y hacer más eficientes las prestaciones asistenciales, para mejorar la gestión y favorecer la salud pública, y para orientar las políticas locales, nacionales y mundiales de salud. La explotación de los grandes datos relacionados con la salud se ha de desarrollar de tal manera que se maximicen los efectos positivos y se reduzcan al mínimo los negativos. Hemos de ser conscientes de la necesidad de que su utilización salvaguarde los derechos de las personas y los valores de la sociedad.

Bibliografía

— Al-Saggaf Y. The use of data mining by private health insurance companies and customers' privacy. Camb Q Healthc Ethics 2015; 24(3):281-292.

— ---, Islam MZ. Data Mining and Privacy of Social Network Sites' Users: Implications of the Data Mining Problem. Sci Eng Ethics 2015; 21(4):941-966.

— American Medical Informatics Association, Code of Professional and Ethical Conduct; Principles of professional and ethical conduct for AMIA members. November, 2011. J Am Med Inform Assoc. 2013; 20(1): 141–143.

— Anderson JG. Ethics and information technology: a case-based approach to a health care system in transition. New York: Springer, 2002.

— ---, Aydin CE. Evaluating the Impact of Health Care Information Systems. Int J Technol Assess Health Care 1997; 13(2): 380-393.

— Aggarwal ChC., Yu PS., eds. Privacy-Preserving Data Mining. Models and Algorithms. Boston: Springer, 2008.

— Bradley AP. Ethics and Data Mining in Biomedical Engineering. En: Jong Yong Abdiel Foo, Stephen J. Wilson, Andrew P. Bradley, Winston Gwee, Dennis Kwok-Wing Tam, Ethics for Biomedical Engineers, Boston, Springer, 2013, pp.77-97.

— Barclay L., AMA Discloses Masterfile Physician Data to Pharmaceutical Companies, Medscape Medical News MediaWatch, July 12, 2007. http://www.medscape.com/viewarticle/559704#vp_1 [accedido 28 de abril de 2016].

[51] La recomendación hecha por Goodman en 1999 (p. 1), es reiterada en el libro publicado en 2015 (p. 140).

- Borna S., Avila S. Genetic information: Consumers' right to privacy versus insurance companies' right to know a public opinion survey. Journal of Business Ethics 1999; 19: 355-362.
- Bruin B. de, Floridi L. The Ethics of Cloud Computing. Sci Eng Ethics 2016: 1-19.
- Chetty R., Stepner M., Abraham S, et al. The Association Between Income and Life Expectancy in the United States, 2001-2014. JAMA 2016; 315(16): 1750-1766.
- Consejo General de Colegios Oficiales de Médicos, Código de Deontología Médica. Madrid, 2011.
- Goodman KW., ed. Ethics, computing, and medicine: informatics and the transformation of health care. Cambridge: Cambridge University Press, 1999.
- ---, Ethics, medicine, and information technology: intelligent machines and the transformation of health care. Cambridge: Cambridge University Press, 2015.
- ---, Meslin EM. Ethics, information technology and public health: Duties and challenges in computational epidemiology. En: Magnuson, J A., Fu, PC., eds., Public Health Informatics and Information Systems, London: Springer-Verlag, 2014, pp. 191-209.
- International Medical Informatics Association. Code of Ethics for Health lnformation Professionals, 31 January, 2011. [http://www.imia-medinfo.org/new2/node/39]
- Kaplan B. Selling health data: de-identification, privacy, and speech. Camb Q Healthc Ethics 2015; 24(3):256-271.
- Kuo-Chung L., Ching-Long Y. Use of Data Mining Techniques to Detect Medical Fraud in Health Insurance. International Journal of Engineering and Technology Innovation (IJETI) 2012; 2(2): 126-137.
- Laurie G., Jones KH., Stevens L., Dobbs C. A review of evidence relating to harm resulting from uses of health and biomedical data, 2014: www.nuffieldbioethics.org/project/biological-health-data/evidence-gathering/
- Le Manach Y. et al. Preoperative score to predict postoperative mortality (POSPOM): Derivation and validation. Anesthesiology 2016; 124:570.
- León Sanz P. Aspectos éticos de la seguridad de la información en los entornos sanitarios. En: Carnicero Giménez de Azcárate, J., et al., Seguridad de la información en entornos sanitarios, Sociedad Española de Informática Sanitaria y Navarra de Gestión para la Administración, Pamplona, 2008, pp. 25-42.
- Marckmann G., Goodman KW. Introduction: Ethics of Information Technology in Health Care, International Review of Information Ethics (IRIE) 2006; 5: 2-5.
- Nuffield Council on Bioethics. The collection, Linking and Use of Data in Biomedical Research and Heatlh care: Ethical Issues, 2015. Disponible en:
  http://nuffieldbioethics.org/wp-content/uploads/Biological_and_health_data_web.pdf
- Ohm P. Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Review 2009; 57: 1701-1777.
- Orentlicher D. Prescription data mining and the protection of patients' interests. J Law Med Ethics 2010; 38(1):74-84.
- Seltzer W. The promise and pitfalls of data mining: ethical issues. In Proceedings of the American Statistical Association, Section on Government Statistics, Alexandria, VA: American Statistical Association 2005: 1441-1445.
- Swan M. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. J Med Internet Res. 2012; 14(2): e46.
- Sweeney L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 2002; 10(5): 557-570.
- Tavani HT. Ethics and technology: Controversies, questions, and strategies for ethical computing. 4th ed. Hoboken: John Wiley, 2013.

— Verschuuren M., Badeyan G., Carnicero J., Gissler M., Pace Asciak R., Sakkeus L., Stenbeck M., Deville W. The European data protection legislation and its consequences for public health monitoring: a plea for action. European Journal of Public Health 2008; 18 (6): 550–551.

— Wel L. van, Royakkers L. Ethical issues in web data mining. Ethics and Information Technology 2004; 6: 129-140.

— Willison DJ., Keshavjee K., Nair K., Goldsmith C., Holbrook AM. Patients' consent preferences for research uses of information in electronic medical records: Interview and survey data. BMJ 2003; 15: 326-373.

— Yoo I., Alafaireet P., Marinov M., et al. Data mining in healthcare and biomedicine: A survey of the literature. J Med Syst 2012; 36: 2431-2448.

Capítulo IV

Disposiciones legales aplicables

Alberto Andérez González

La utilización y explotación masiva de datos, lo que actualmente se conoce como *Big Data*, se revela como un campo que, aun vinculado específicamente al desarrollo tecnológico y de los sistemas de información, se ve precisado de una especial atención también desde un punto de vista legal. Esta consideración adquiere una dimensión si cabe mayor en el ámbito sanitario, en razón a la protección singular que la legislación tanto comunitaria como interna dispensa al tratamiento, uso y cesión de los datos relativos a la salud de las personas.

A pesar de ello, y anticipando la conclusión que se desprende del análisis que seguidamente se efectúa, debe resaltarse la ausencia de un tratamiento legal específico de esta figura, tanto con carácter general como de modo particular en lo referido a la información sanitaria; lo que determina en este último caso la necesidad de remitir a la aplicación del marco general en la materia, cuya aprobación y promulgación se sitúa en un momento notablemente anterior en el tiempo a la aparición del fenómeno.

Este marco normativo, como es conocido, se integra por dos regulaciones sectoriales principales: por un lado, la regulación legal en materia de protección de datos de carácter personal y, por otro, la legislación sanitaria; si bien dentro de esta última podemos considerar tanto la normativa general en materia de salud y derechos de los pacientes, como la dictada en relación con determinadas actividades de investigación. Es precisamente el campo de la investigación sanitaria el que debe ser objeto principal de análisis en el presente estudio, habida cuenta que la utilización con tal objeto constituye uno de los principales usos o aplicaciones del *Big Data* en este campo.

En todo caso, y aun cuando los dos ámbitos sectoriales (sanitario y de protección de datos de carácter personal) convergen en gran medida en el tratamiento dispensado, resulta aconsejable proceder a su examen por separado.

1. Normativa en materia de protección de datos de carácter personal

Los principios y reglas que disciplinan el tratamiento de los datos de salud conforme a este marco legal son bien conocidos. No en vano la regulación legal en la materia (con origen en la previsión del artículo 18.4 del Texto Constitucional: *"la ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos"*) cuenta con bastantes años de aplicación y, además, en relación con la interpretación de la misma existe en la actualidad un cuerpo de doctrina consolidado, que emana tanto de los Tribunales como de la Agencia Española de Protección de Datos y demás entes autonómicos con competencias en la materia.

La rigidez de este marco normativo no resulta únicamente del rango orgánico de la regulación legal en que se contiene, sino del hecho de constituir la misma trasposición obligada de la Directiva 95/46/CE del Parlamento y del Consejo, al margen de las propias previsiones del Convenio 108 del Consejo de Europa para

la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal, hecho en Estrasburgo el 28 de enero de 1981, ratificado por España en fecha 27 de enero de 1984. Ambos instrumentos internacionales, así, son expresamente invocados por la Agencia Española de Protección de Datos como fundamento y motivación de sus resoluciones y dictámenes (véase, por ejemplo, el informe jurídico 0471/2008).

En orden a la aplicación de esta regulación en relación con el uso y explotación masivo de datos sanitarios interesa, ante todo y en primer lugar, delimitar este último concepto, labor que realiza el artículo 5.1 del Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal. En concreto, su apartado g) define los datos de carácter personal relacionados con la salud como aquellas *"informaciones concernientes a la salud pasada, presente y futura, física o mental, de un individuo"*, añadiendo que *"en particular, se consideran datos relacionados con la salud de las personas los referidos a su porcentaje de discapacidad y a su información genética"*.

Partiendo de esta noción, el artículo 7 de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal incluye los datos de salud entre los tributarios de una especial protección, disponiendo en su apartado tercero que:

"Los datos de carácter personal que hagan referencia al origen racial, a la salud y a la vida sexual sólo podrán ser recabados, tratados y cedidos cuando, por razones de interés general, así lo disponga una ley o el afectado consienta expresamente."

Esta regla se exceptúa en los términos que se contienen en el apartados sexto del citado artículo 7 y en el artículo 8 de la misma Ley, a cuyo tenor respectivamente:

"No obstante lo dispuesto en los apartados anteriores, podrán ser objeto de tratamiento los datos de carácter personal a que se refieren los apartados 2 y 3 de este artículo, cuando dicho tratamiento resulte necesario para la prevención o para el diagnóstico médicos, la prestación de asistencia sanitaria o tratamientos médicos o la gestión de servicios sanitarios, siempre que dicho tratamiento de datos se realice por un profesional sanitario sujeto al secreto profesional o por otra persona sujeta asimismo a una obligación equivalente de secreto.

También podrán ser objeto de tratamiento los datos a que se refiere el párrafo anterior cuando el tratamiento sea necesario para salvaguardar el interés vital del afectado o de otra persona, en el supuesto de que el afectado esté física o jurídicamente incapacitado para dar su consentimiento."

(…)

"Artículo 8. Datos relativos a la salud.

Sin perjuicio de lo que se dispone en el artículo 11 respecto de la cesión, las instituciones y los centros sanitarios públicos y privados y los profesionales correspondientes podrán proceder al tratamiento de los datos de carácter personal relativos a la salud de las personas que a ellos acudan o hayan de ser tratados en los mismos, de acuerdo con lo dispuesto en la legislación estatal o autonómica sobre sanidad."

De estas normas se deduce el criterio que exige con carácter general el consentimiento del interesado (entendiendo por este, según el artículo 3 de la Ley Orgánica, la *"persona física titular de los datos que sean objeto del tratamiento"*), el paciente en este caso, para el tratamiento de sus datos de salud, salvo en lo que concierne a los datos necesarios para la aplicación del diagnóstico y tratamiento que motiva su atención en el centro sanitario en los términos de la regulación legal en materia de historia clínica y derechos de los pacientes, a la que posteriormente nos referimos.

Dicho consentimiento, además, debe ser prestado de unas determinadas condiciones, como son las que se deducen del apartado h) del artículo 3 de la Ley Orgánica, en cuanto exige que la manifestación de voluntad sea libre, inequívoca, específica e informada. Sobre este último aspecto incide el artículo 5.1 de la norma legal, que define en términos ciertamente rigurosos el contenido y alcance de la información que debe facilitarse, cuya ausencia vicia el consentimiento prestado, y que comprende:

a) La existencia de un fichero o tratamiento de datos de carácter personal, la finalidad de la recogida de éstos y los destinatarios de la información; sobre este punto, el artículo 12.1 del Reglamento de desarrollo de la Ley Orgánica especifica que la solicitud del consentimiento deberá ir referida a un tratamiento o serie de tratamientos concretos, con delimitación de la finalidad para los que se recaba, así como de las restantes condiciones que concurran en el tratamiento o serie de tratamientos.

b) El carácter obligatorio o facultativo de su respuesta a las preguntas que les sean planteadas.

c) Las consecuencias de la obtención de los datos o de la negativa a suministrarlos.

d) La posibilidad de ejercitar los derechos de acceso, rectificación, cancelación y oposición.

e) La identidad y dirección del responsable del tratamiento o, en su caso, de su representante.

Es preciso tener en cuenta, por otro lado, que el tratamiento precisado de consentimiento se define en términos notoriamente amplios por el artículo 3.c) de la propia Ley Orgánica como todas aquellas *"operaciones y procedimientos técnicos de carácter automatizado o no, que permitan la recogida, grabación, conservación, elaboración, modificación, bloqueo y cancelación, así como las cesiones de datos que resulten de comunicaciones, consultas, interconexiones y transferencias"*.

Junto con este criterio, el segundo principio fundamental en la materia, y que incide especialmente sobre el objeto del presente análisis, es el contenido en el artículo 11 de la Ley Orgánica que dispone las condiciones a que se sujeta la cesión o comunicación de datos, definida en el artículo 3.i) como *"toda revelación de datos realizada a una persona distinta del interesado"*. La cesión, conforme a este precepto, únicamente se autoriza para el cumplimiento de fines directamente relacionados con las funciones legítimas del cedente y del cesionario y se condiciona, asimismo, al previo consentimiento del interesado salvo determinadas excepciones.

Entre estas excepciones (además de la de carácter general, esto es, cuando la cesión venga autorizada por una ley) cabe citar una referida específicamente a los datos de salud, cual es la que faculta para el acceso a los mismos para realizar los estudios epidemiológicos en los términos establecidos en la legislación sobre sanidad estatal o autonómica. Del mismo modo que, a su vez, el artículo 10 del Reglamento, en coherencia con lo que dispone el artículo 8 de la Ley Orgánica, exceptúa el consentimiento del interesado para la comunicación de datos personales sobre la salud, incluso a través de medios electrónicos, entre organismos, centros y servicios del Sistema Nacional de Salud pero únicamente cuando la cesión se realice para la atención sanitaria de las personas.

Por el contrario, y dado el sentido de la normativa sanitaria a la que nos referimos con posterioridad, la cesión de datos con fines de investigación clínica sin recabar el consentimiento del interesado no encuentra respaldo en la previsión de la Ley Orgánica relativa a la comunicación entre Administraciones públicas con objeto del tratamiento posterior de los datos con fines, entre otros, científicos. La remisión que en este punto efectúa el artículo 9 del Reglamento de desarrollo de la Ley Orgánica a la regulación contenida en la Ley 13/1986, de 14 de abril, de Fomento y coordinación general de la investigación científica y técnica (actualmente derogada por la Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación) confirma la conclusión expuesta.

Sí reviste, sin embargo, especial importancia la mención del apartado 6 del artículo 11 citado, que exceptúa del régimen general expuesto (y, por tanto, de la necesidad de consentimiento del interesado) la cesión de datos cuando se realice previo un procedimiento de disociación, esto es, y en los términos del artículo 3.f) de la Ley Orgánica, cuando los datos personales hayan sido tratados de modo que la información que se obtenga no pueda asociarse a persona identificada o identificable.

Por lo demás, el consentimiento para la cesión de datos exige, bajo sanción de nulidad en otro caso, que se facilite al interesado la finalidad a que destinarán los datos cuya comunicación se autoriza o el tipo de actividad de aquel a quien se pretenden comunicar; y además, el mismo tiene en todo caso carácter revocable (apartados 3 y 4 del artículo 11 de la Ley Orgánica).

La aplicación de estas exigencias normativas encuentra su reflejo en el criterio reiteradamente expresado por la Agencia Española de Protección de Datos.

Con carácter general, el Informe 0471/2008 examina si la recogida de datos de salud requiere el consentimiento escrito de los pacientes de acuerdo con las previsiones contenidas en la Ley Orgánica 11/1999, de 13 de diciembre de Protección de Datos de Carácter personal, señalando al respecto que:

"La especial protección conferida a los datos relacionados con la salud de las personas no es arbitraria, sino que resulta de lo dispuesto en las normas Internacionales y Comunitarias reguladoras del tratamiento automatizado de datos de carácter personal. En este contexto, tanto el artículo 8 de la Directiva 95/46/CE del Parlamento y del Consejo, así como el artículo 6 del Convenio 108 del Consejo de Europa para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal, hecho en Estrasburgo el 28 de enero de 1981, ratificado por España en fecha 27 de enero de 1984, hacen referencia a los datos de salud como sujetos a un régimen especial de protección.

En este sentido, el artículo 8 de la Directiva 95/46/CE limita el tratamiento de datos a supuestos y finalidades concretos en los que será preciso el consentimiento, que además deberá ser expreso, del afectado o la necesidad del tratamiento con fines de asistencia sanitaria o atención de un interés vital del afectado. Esta cuestión ha sido especialmente analizada por el Grupo de Autoridades de Protección de Datos creado por el artículo 29 de la citada Directiva en su Documento de trabajo sobre el tratamiento de datos personales relativos a la salud en los historiales médicos electrónicos (Documento EP131), en el que se indica expresamente que "todos los datos contenidos en documentos médicos, en historiales médicos electrónicos y en sistemas de HME son "datos personales sensibles". Por tanto, no sólo están sujetos a todas las normas generales sobre protección de datos personales de la Directiva, sino también a las normas sobre protección de datos especiales que rigen el tratamiento de la información sensible, contenidas en el artículo 8 de la Directiva."

En análogo sentido se pronuncia el Informe 0081/2009, en el que se examina si resulta conforme a la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal, la remisión que efectúan las farmacias, al solicitar el correspondiente pedido al laboratorio consultante, del formulario de solicitud del tratamiento que contiene datos personales de aquéllos para quienes se prepara una vacuna personalizada, y en el que se afirma:

"Tratándose, en el presente caso de datos de salud, debe recordarse que el tratamiento y cesión de datos de carácter personal, cuyo régimen aparece recogido con carácter general en los artículos 6 y 11 de la Ley Orgánica 15/1999, se encuentra, por vía de excepción, sometido a particulares restricciones en lo que a los datos de salud respecta, por el artículo 7 de la citada Ley Orgánica 15/1999, cuyo apartado 3 establece como regla general que "Los datos de carácter personal que hagan referencia al origen racial, a la salud y a la vida sexual sólo podrán ser recabados, tratados y cedidos cuando, por razones de interés general, así lo disponga una Ley o el afectado consienta expresamente". Esta regla únicamente es matizada por la Ley Orgánica en sus artículos 7.6 y 8.

Por consiguiente, será precisa la existencia de una ley que ampare la cesión y el tratamiento de los datos o que el interesado consienta tanto el tratamiento como la cesión de sus datos.

Cabe aquí analizar las características que el consentimiento debe reunir conforme a la Ley Orgánica 15/1999. El artículo 3.h de dicha norma señala que se trata de una "manifestación de voluntad, libre, inequívoca, específica e informada, mediante la que el interesado consienta el tratamiento de datos personales que le conciernen", a ello debe añadirse que, en el presente caso, debe ser expreso, tal y como indica el artículo 7 de la Ley Orgánica 15/1999.

Esta Agencia ha venido describiendo en sus informes dichas características de manera que se entiende por consentimiento libre aquel que ha sido obtenido sin la intervención de vicio alguno del consentimiento en los términos regulados por el código civil. El consentimiento específico viene referido a una determinada operación de tratamiento y para una finalidad determinada, explícita y legítima del responsable del tratamiento, tal y como impone el artículo 4.2 de la Ley Orgánica 15/1999. Para que pueda hablarse de consentimiento inequívoco se exige la realización de una acción u omisión que implique la existencia del consentimiento. En cuanto al requisito de la información, supone que el afectado conozca con anterioridad al tratamiento la existencia del mismo y las finalidades para las que el mismo se produce.

(…)

A este respecto será preciso, que se facilite al interesado la información a que hace referencia el artículo 5.1 de la Ley Orgánica 15/1999…"

Especial interés reviste también el Informe 0509/2009 en relación con los datos de pacientes afectados por la correspondiente patología y que se recaban para su comunicación al fichero denominado "Registro Español de Poliposis" cuyo responsable es la Asociación Española de Gastroenterología. Dicho Informe, tras recordar que el supuesto se configura como una cesión de datos a efectos de la Ley Orgánica 15/1999, de 13 de diciembre, concluye que el documento en virtud del cual se recaba el consentimiento de los pacientes debe informar con más claridad de la finalidad del fichero y los distintos cesionarios que podrán acceder a la información contenida en el correspondiente fichero.

E igualmente son interesantes las consideraciones del Informe 0654/2009 en relación con el concepto de datos disociados a efectos de la aplicación de las disposiciones legales sobre protección de datos de carácter personal, cuestión que en el caso allí examinado se suscita respecto de un fichero en el que, para un proyecto de investigación, se contienen datos relativos a reacciones alérgicas, test realizados y datos relativos al paciente, al que se le identifica por un código numérico. Sobre este particular señala el citado informe:

"La cuestión planteada ha sido objeto de informe en diversas ocasiones por esta Agencia, por todas ellas cabe aquí reiterar lo indicado en informe de 22 de septiembre de 2008 en el que se señalaba lo siguiente:

"La cuestión a dilucidar en este caso es la de si el tratamiento al que se refiere la consulta se encuentra sometido a lo dispuesto en la vigente normativa de protección de datos, dado que el párrafo primero del artículo 2.1 de la Ley Orgánica 15/1999 dispone que "La presente Ley Orgánica será de aplicación a los datos de carácter personal registrados en soporte físico que los haga susceptibles de tratamiento, y a toda modalidad de uso posterior de estos datos por los sectores público y privado", siendo datos de carácter personal, conforme al artículo 3 a) de la propia Ley "cualquier información concerniente a personas físicas identificadas o identificables".

Esta definición se complementa con la de persona identificable, a la que se refiere el artículo 5.1 o) del Reglamento de desarrollo de la Ley Orgánica, que dispone que lo será "toda persona cuya identidad pueda determinarse, directa o indirectamente, mediante cualquier información referida a su identidad física, fisiológica, psíquica, económica, cultural o social. Una persona física no se considerará identificable si dicha identificación requiere plazos o actividades desproporcionados".

A título meramente ilustrativo, cabe tener en cuenta las definiciones previstas en las letras p) a r) de la Ley 14/2007, de 3 de julio, de Investigación biomédica, que permiten delimitar los supuestos en los que, ciertamente en su ámbito de aplicación, será o no de aplicación lo dispuesto en la legislación de protección de datos. Así, se distinguen los siguientes conceptos:

- *«Muestra biológica anonimizada o irreversiblemente disociada»: muestra que no puede asociarse a una persona identificada o identificable por haberse destruido el nexo con toda información que identifique al sujeto, o porque dicha asociación exige un esfuerzo no razonable.*
- *«Muestra biológica no identificable o anónima»: muestra recogida sin un nexo con una persona identificada o identificable de la que, consiguientemente, no se conoce la procedencia y es imposible trazar el origen.*
- *«Muestra biológica codificada o reversiblemente disociada»: muestra no asociada a una persona identificada o identificable por haberse sustituido o desligado la información que identifica a esa persona utilizando un código que permita la operación inversa.*

Mientras los dos primeros supuestos podrían quedar excluidos de la aplicación de la Ley Orgánica 15/1999, dicha Ley sí será de aplicación en el supuesto de tratamiento de datos "codificados o reversiblemente disociados, toda vez que a partir de la información de que se tiene conocimiento será posible realizar la "operación inversa" a la codificación.

De este modo, si los datos relacionados con el seguimiento del ensayo se encuentran asociados a datos que pudieran permitir la asociación de los mismos al concreto sujeto del mismo, como sucederá en caso de que aquéllos aparezcan asociados a un código establecido por el investigador, cabrá entender que el fichero se encuentra sometido a lo dispuesto en la Ley Orgánica 15/1999, debiendo implantarse en el mismo las medidas de seguridad previstas en dicha Ley y su Reglamento de desarrollo. Este suele ser el procedimiento seguido en el ámbito de los ensayos clínicos, en los que será posible la identificación del sujeto del ensayo, incluso cuando alguno de los sujetos intervinientes en el mismo únicamente pueda acceder, en principio, a datos codificados."

Conforme al criterio de esta Agencia, expuesto en dicho informe, la Ley Orgánica 15/1999 es de plena aplicación al presente caso, puesto que el paciente es identificable a través de un código numérico, lo que impide entender que este dato constituya un dato anónimo o anonimizado, en la terminología de la Ley 14/2007, o un dato disociado en la definición dada por el Reglamento de protección de datos de carácter personal, esto es, "aquél que no permite la identificación de un afectado o interesado". En definitiva, no habiéndose producido un procedimiento de disociación que impida la asociación del dato con una persona identificada o identificable, la aplicación de la Ley Orgánica 15/1999 no puede quedar excluida."

Como puede observarse, el dictamen transcrito en último lugar ilustra sobre la interrelación entre la legislación sobre protección de datos y la propia normativa sanitaria, en la medida en que la Agencia acude a esta última en orden a delimitar los conceptos y exigencias que impone la primera.

2. Normativa de salud e investigación sanitaria

Los criterios expuestos en el apartado precedente anticipan en gran medida las conclusiones que se desprenden de la regulación integrante del segundo de los ámbitos normativos de aplicación ya señalados.

La primera referencia en este ámbito viene constituida por 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, y en concreto las previsiones contenidas en sus artículos 7 y 16.

El primero de ellos, bajo el título de *"el derecho a la intimidad"*, dispone que:

"1. Toda persona tiene derecho a que se respete el carácter confidencial de los datos referentes a su salud, y a que nadie pueda acceder a ellos sin previa autorización amparada por la Ley.

2. Los centros sanitarios adoptarán las medidas oportunas para garantizar los derechos a que se refiere el apartado anterior, y elaborarán, cuando proceda, las normas y los procedimientos protocolizados que garanticen el acceso legal a los datos de los pacientes."

Por su parte, el artículo 16, en su apartado 3, señala que:

"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 Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal, 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 de la autoridad judicial en los que se considere imprescindible la unificación de los datos identificativos con los clínico asistenciales, en los cuales se estará a lo que dispongan los jueces y tribunales en el proceso correspondiente. El acceso a los datos y documentos de la historia clínica queda limitado estrictamente a los fines específicos de cada caso.

Cuando ello sea necesario para la prevención de un riesgo o peligro grave para la salud de la población, las Administraciones sanitarias a las que se refiere la Ley 33/2011, General de Salud Pública, podrán acceder a los datos identificativos de los pacientes por razones epidemiológicas o de protección de la salud pública. El acceso habrá de realizarse, en todo caso, por un profesional sanitario sujeto al secreto profesional o por otra persona sujeta, asimismo, a una obligación equivalente de secreto, previa motivación por parte de la Administración que solicitase el acceso a los datos."

Es, por tanto, la propia regulación sanitaria la que, a efectos del tratamiento de datos de salud con fines distintos a la propia prestación de asistencia, remite al régimen general contenido en la legislación sobre protección de datos de carácter personal y, en coherencia con el sentido de esta última, impone en tales casos el carácter anónimo de los datos, salvo consentimiento expreso del paciente. En definitiva, nos encontramos ante la plasmación de los mismos principios en los que se inspira la regulación de la Ley Orgánica 15/1999 y en su Reglamento de desarrollo.

Estas previsiones, por otro lado, se reiteran en la regulación legal dictada en materia de investigación sanitaria, específicamente en relación con un sector de la misma como es el de la investigación biomédica, regulada actualmente en la 14/2007, de 3 de julio (que deroga los artículos 106 a 110 de la Ley General de Sanidad sobre este punto).

Efectuamos esta primera precisión habida cuenta que esta norma legal no tiene por objeto una regulación general de la actividad investigadora en el campo sanitario, sino tan solo aquellos ámbitos de la misma relacionados en el artículo 1 de la Ley, entre ellos:

a) Las investigaciones relacionadas con la salud humana que impliquen procedimientos invasivos, definidos estos últimos como aquellas intervenciones realizadas con fines de investigación que impliquen un riesgo físico o psíquico para el sujeto afectado.

b) La donación y utilización de ovocitos, espermatozoides, preembriones, embriones y fetos humanos o de sus células, tejidos u órganos con fines de investigación biomédica y sus posibles aplicaciones clínicas.

c) El tratamiento de muestras biológicas, así como su almacenamiento y movimiento.

d) La realización de análisis genéticos y el tratamiento de datos genéticos de carácter personal, definiéndose el dato genético de carácter personal como aquella información sobre las características hereditarias de una persona identificada o identificable obtenida por análisis de ácidos nucleicos u otros análisis científicos.

Sin embargo, quedan excluidos del ámbito de aplicación de la Ley y remitidos a su normativa específica los ensayos clínicos con medicamentos y productos sanitarios, así como las implantaciones de órganos, tejidos y células de cualquier origen que se rigen por lo establecido en la Ley 30/1979, de 27 de octubre, sobre extracción y trasplante de órganos.

Entre los principios y garantías de la investigación biomédica que enumera el artículo 2 de la Ley se sitúan la garantía de los derechos y libertades fundamentales de la persona y específicamente la garantía de la confidencialidad en el tratamiento de los datos de carácter personal y de las muestras biológicas, en especial en la realización de análisis genéticos, encomendándose al Comité de Ética de la Investigación correspondiente al centro la función, entre otras, de velar por la confidencialidad en el desarrollo de esta actividad (artículo 12 de la Ley).

Partiendo como premisa de la necesidad del consentimiento de la persona vaya a participar en una investigación biomédica (artículo 4 de la Ley: *"se respetará la libre autonomía de las personas que puedan participar en una investigación biomédica o que puedan aportar a ella sus muestras biológicas, para lo que será preciso que hayan prestado previamente su consentimiento expreso y escrito una vez recibida la información adecuada"*), el cual es a su vez revocable, el artículo 5 de la norma regula la protección de datos personales y las garantías de confidencialidad, estableciendo al efecto que:

"1. Se garantizará la protección de la intimidad personal y el tratamiento confidencial de los datos personales que resulten de la actividad de investigación biomédica, conforme a lo dispuesto en la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. Las mismas garantías serán de aplicación a las muestras biológicas que sean fuente de información de carácter personal.

2. La cesión de datos de carácter personal a terceros ajenos a la actuación médico-asistencial o a una investigación biomédica, requerirá el consentimiento expreso y escrito del interesado.

En el supuesto de que los datos obtenidos del sujeto fuente pudieran revelar información de carácter personal de sus familiares, la cesión a terceros requerirá el consentimiento expreso y escrito de todos los interesados.

3. Se prohíbe la utilización de datos relativos a la salud de las personas con fines distintos a aquéllos para los que se prestó el consentimiento.

4. Quedará sometida al deber de secreto cualquier persona que, en el ejercicio de sus funciones en relación con una actuación médico-asistencial o con una investigación biomédica, cualquiera que sea el alcance que tengan una y otra, acceda a datos de carácter personal. Este deber persistirá aún una vez haya cesado la investigación o la actuación.

5. Si no fuera posible publicar los resultados de una investigación sin identificar a la persona que participó en la misma o que aportó muestras biológicas, tales resultados sólo podrán ser publicados cuando haya mediado el consentimiento previo y expreso de aquélla."

Cabe observar la plena correspondencia entre la norma transcrita y los criterios y principios contenidos en la Ley Orgánica de Protección de Datos de Carácter Personal ya examinados anteriormente.

A estos efectos, revisten importancia las distintas definiciones del artículo 3 de la Ley relacionadas con el carácter anónimo o disociado de los datos, como son en particular las siguientes:

- «Anonimización»: proceso por el cual deja de ser posible establecer por medios razonables el nexo entre un dato y el sujeto al que se refiere. Es aplicable también a la muestra biológica.
- «Dato anónimo»: dato registrado sin un nexo con una persona identificada o identificable.
- «Dato anonimizado o irreversiblemente disociado»: dato que no puede asociarse a una persona identificada o identificable por haberse destruido el nexo con toda información que identifique al sujeto, o porque dicha asociación exige un esfuerzo no razonable, entendiendo por tal el empleo de una cantidad de tiempo, gastos y trabajo desproporcionados.
- «Dato codificado o reversiblemente disociado»: dato no asociado a una persona identificada o identificable por haberse sustituido o desligado la información que identifica a esa persona utilizando un código que permita la operación inversa.
- «Muestra biológica»: cualquier material biológico de origen humano susceptible de conservación y que pueda albergar información sobre la dotación genética característica de una persona.
- «Muestra biológica anonimizada o irreversiblemente disociada»: muestra que no puede asociarse a una persona identificada o identificable por haberse destruido el nexo con toda información que identifique al sujeto, o porque dicha asociación exige un esfuerzo no razonable.
- «Muestra biológica no identificable o anónima»: muestra recogida sin un nexo con una persona identificada o identificable de la que, consiguientemente, no se conoce la procedencia y es imposible trazar el origen.
- «Muestra biológica codificada o reversiblemente disociada»: muestra no asociada a una persona identificada o identificable por haberse sustituido o desligado la información que identifica a esa persona utilizando un código que permita la operación inversa.

Sobre la importancia de estas definiciones ya hemos señalado cómo es la propia Agencia Española de Protección de Datos la que acude a las mismas en orden a delimitar la aplicación de las previsiones de la Ley Orgánica 15/1999 a la toma y comunicación de datos a efectos de investigación clínica, que se afirma expresamente en aquellos casos en que el tratamiento se realice sobre datos codificados o reversiblemente disociados (Informe 0533/2008).

Estos criterios generales de la Ley sobre de confidencialidad y protección de datos se reiteran en el artículo 45 en relación con los análisis genéticos y las muestras biológicas.

De especial interés resultan las normas referidas a análisis genéticos. Este es el caso, en primer lugar, del artículo 47 de la Ley relativo la información escrita que ha de suministrarse al interesado previamente a la realización de análisis genéticos con fines de investigación en el ámbito sanitario y que debe abarcar los siguientes aspectos:

1.º Finalidad del análisis genético para el cual consiente.

2.º Lugar de realización del análisis y destino de la muestra biológica al término del mismo, sea aquél la disociación de los datos de identificación de la muestra, su destrucción, u otros destinos, para lo cual se solicitará el consentimiento del sujeto fuente en los términos previstos en esta Ley.

3.º Personas que tendrán acceso a los resultados de los análisis cuando aquellos no vayan a ser sometidos a procedimientos de disociación o de anonimización.

4.º Advertencia sobre la posibilidad de descubrimientos inesperados y su posible trascendencia para el sujeto, así como sobre la facultad de este de tomar una posición en relación con recibir su comunicación.

5.º Advertencia de la implicación que puede tener para sus familiares la información que se llegue a obtener y la conveniencia de que él mismo, en su caso, transmita dicha información a aquéllos.

6.º Compromiso de suministrar consejo genético, una vez obtenidos y evaluados los resultados del análisis.

Sobre la base de esta información, el artículo 48 de la Ley requiere con carácter general la prestación de consentimiento expreso y específico por escrito para la realización de un análisis genético.

La regulación del acceso a los datos genéticos por parte del personal sanitario sigue un mismo criterio restrictivo en el artículo 50, conforme al cual:

"1. Los profesionales sanitarios del centro o establecimiento donde se conserve la historia clínica del paciente tendrán acceso a los datos que consten en la misma en tanto sea pertinente para la asistencia que presten al paciente, sin perjuicio de los deberes de reserva y confidencialidad a los que estarán sometidos.

2. Los datos genéticos de carácter personal sólo podrán ser utilizados con fines epidemiológicos, de salud pública, de investigación o de docencia cuando el sujeto interesado haya prestado expresamente su consentimiento, o cuando dichos datos hayan sido previamente anonimizados.

3. En casos excepcionales y de interés sanitario general, la autoridad competente, previo informe favorable de la autoridad en materia de protección de datos, podrá autorizar la utilización de datos genéticos codificados, siempre asegurando que no puedan relacionarse o asociarse con el sujeto fuente por parte de terceros."

En último lugar, el artículo 51 de la norma contempla el deber de confidencialidad y el derecho a la protección de los datos genéticos estableciendo que *"sólo con el consentimiento expreso y escrito de la persona de quien proceden se podrán revelar a terceros datos genéticos de carácter personal"* y que *"si no es posible publicar los resultados de una investigación sin identificar a los sujetos fuente, tales resultados sólo podrán ser publicados con su consentimiento"*.

Esta última previsión coincide, por lo demás, con lo que respecto de los ensayos clínicos con medicamentos dispone el artículo 42 del Real Decreto 1090/2015, de 4 de diciembre, por el que se regulan los ensayos clínicos con medicamentos, los Comités de Ética de la Investigación con medicamentos y el Registro Español de Estudios Clínicos, que dispone que en la publicación de los resultados de los ensayos clínicos se mantendrá en todo momento el anonimato de los sujetos participantes en el ensayo.

Respecto de la garantía del anonimato en la publicación o divulgación científica de datos de salud de carácter personal con relevancia desde el punto de vista de la investigación clínica resulta interesante el pronunciamiento de la Sala de lo Contencioso Administrativo del Tribunal Superior de Justicia del Principado de Asturias en su sentencia de 30 de septiembre de 2008, que declara la existencia de responsabilidad patrimonial de la Administración sanitaria por la publicación en revista especializada de un artículo sobre la enfermedad padecida por un menor no identificado que, aun revistiendo un notable interés científico, se realiza sin contar con el consentimiento de los padres; pronunciamiento que se adopta frente al criterio expresado por el Consejo de Estado y no obstante el archivo de las actuaciones incoadas por denuncia de los recurrentes ante la Agencia Española de Protección de Datos.

3. Reflexiones finales

Desde una perspectiva exclusivamente jurídica, las consideraciones expuestas permiten efectuar, a modo de conclusiones generales, las reflexiones siguientes.

Cabe señalar, en primer lugar, que el marco normativo de aplicación a una figura de aparición relativamente reciente como la que se analiza sigue estando constituido por un conjunto de disposiciones legales cuya vigencia, en general, se remonta a fechas notoriamente anteriores en el tiempo, lo que supone que la explotación masiva de datos sanitarios quede sujeta a las mismas reglas, criterios y principios contenidos en aquellas normas.

A su vez, este marco regulatorio presenta escaso o nulo margen para su modificación o modulación en el ámbito estatal, y ello no solo en razón al rango orgánico de la ley reguladora de la protección de datos, sino también debido a que esta última, en definitiva, constituya necesaria y obligada trasposición de normas de Derecho comunitario.

Por lo demás, las pautas de interpretación de estas normas por parte de los operadores jurídicos (tanto los órganos judiciales como las agencias e instituciones competentes en materia de protección de datos de carácter personal) revelan un rigor elevado en la defensa y garantía de los principios básicos de confidencialidad de la información sanitaria y exigencia del consentimiento para su tratamiento y cesión.

Esta última consideración puede llevar a que por parte de profesionales y responsables sanitarios se pueda percibir en ocasiones una cierta desatención legal a otros fines u objetivos distintos, pero de indudable relevancia en el ámbito de la salud individual y colectiva, y para cuya satisfacción resultan necesarios el acceso y la utilización de información sanitaria. La práctica plantea en este sentido cuestiones como:

- La revisión por los profesionales de los datos de pacientes para la evaluación de resultados, con fines de mejora de la calidad asistencial, evaluación de su eficiencia o investigación.
- El acceso a esa misma información por parte de los responsables de los servicios clínicos y de los centros asistenciales, con idéntica finalidad o para la evaluación del desempeño.
- La utilización de datos de salud para investigación, innovación y desarrollo de nuevos productos.
- La cesión de datos entre centros de investigación para trabajos en red.

La respuesta legal a estas cuestiones puede formularse sobre los siguientes criterios generales:

1. El acceso a la información contenida en la historia clínica por parte de los profesionales que intervienen en el proceso asistencial del paciente se habilita directamente por la Ley, y ello ha de amparar razonablemente la utilización, por los propios profesionales, de dicha información (y del conocimiento adquirido a través de ella en su práctica profesional) para la evaluación y mejora de la calidad asistencial.
2. Las funciones de evaluación e inspección con fines, entre otros, de comprobación de la calidad de la asistencia, autorizan igualmente el mero acceso a los datos clínicos por parte de los profesionales sanitarios autorizados.
3. El acceso por parte de otros profesionales con fines de investigación y docencia únicamente se habilita previa anonimización de los datos.
4. Asimismo, en todo caso las actuaciones que, más allá del mero acceso, comporten cualquier grado de elaboración, modificación o recogida en un fichero o registro específico de información contenida en la historia clínica deben reputarse legalmente como tratamiento de datos; lo que, respecto de los datos de salud, exige recabar el consentimiento del interesado o titular de los datos para todo fin distinto a la prestación de asistencia, en particular los de investigación y docencia.
5. Cualquier comunicación de información clínica realizada a terceros distintos a quienes legalmente están autorizados para el acceso a la información sanitaria constituye una cesión de datos que precisa también del consentimiento del titular de los mismos.

Capítulo V

Organización y tecnología para la explotación de la información

Juan Díaz García

1. Organización de la Información para el *Big Data*

En este capítulo se estudian los requisitos de la información, en lo referente tanto a sus características como a su tratamiento, para que pueda ser considerada *Big Data*. Esta denominación se aplica a los conjuntos de datos que, por su complejidad o volumen, no pueden ser procesados adecuadamente mediante las herramientas y las bases de datos convencionales.

Para poder afrontar la gestión de la información con los requerimientos específicos del *Big Data* es importante tener una visión estratégica de este tema y aplicar las diferentes metodologías que recogen las mejores prácticas para afrontar estos retos[52].

Se repasan también los recursos necesarios para afrontar el manejo de esta información y los procesos que son necesarios a lo largo del ciclo de vida de la misma, desde su recolección hasta la generación de conocimiento por parte de las organizaciones, y se revisan las últimas tendencias, donde las máquinas (algoritmos) piensan de forma autónoma o bajo supervisión humana.

1.1. Dimensiones del *Big Data*

El especial tratamiento de la información en el ámbito del *Big Data* se debe a una serie de características, entre las que destacan las siguientes, conocidas habitualmente como "10 V":

- **Volumen**: el manejo de una ingente cantidad de datos supone un gran reto y está cada vez más presente dada la evolución de los sistemas de información, que generan datos siguiendo un crecimiento exponencial. No hay más que fijarse en las unidades de medida de este volumen: Gigabytes (10^9 bytes), Terabytes (10^{12} bytes), Petabytes (10^{15} bytes), etc.
- **Variedad**: hoy tenemos a nuestra disposición una gran variedad de datos que a su vez pueden proceder de diversas fuentes, como son por ejemplo los datos asociados a un paciente, lo que supone una mayor complejidad de los procesos de tratamiento de la información. Un ejemplo de esta complejidad es el análisis de miles de genes relacionados con una enfermedad.
- **Velocidad**: otro hecho significativo es la actual capacidad de los sistemas para generar datos a una mayor velocidad, lo que requiere una gran capacidad de procesamiento para la información que se genera y se trata en los sistemas en tiempo real, es decir, aquellos que además de analizar su entorno físico y determinar la respuesta necesaria, son capaces de garantizar que esta última se lleva a cabo dentro de un plazo máximo de tiempo predeterminado.
- **Veracidad**: para poner a prueba las diferentes hipótesis deben identificarse los datos relevantes que sean necesarios y suficientes. Del mismo modo, la construcción de modelos y su validación posterior requieren la identificación de las variables de calidad precisas para, desde la pequeña a la gran escala,

[52] Labrinidis y Jagadish (2012).

hacer una extrapolación que permita extender el análisis a grandes muestras, como por ejemplo toda la población.

- **Validez**: para asegurar la fiabilidad de los datos debe garantizarse su calidad, lo que obliga al seguimiento de protocolos para su gobierno, y en especial para la gestión de datos maestros procedentes de diversas fuentes, que son por lo tanto de tipo masivo, distribuido y heterogéneo.
- **Valor**: la gestión de grandes volúmenes de datos tiene potencial para ayudar a transformar la organización, desde sus procesos básicos hasta las estrategias institucionales, lo que tiene un impacto claro en el valor del negocio y en el retorno de las inversiones.
- **Variabilidad**: la información no es estática sino dinámica. Los datos pueden cambiar según evolucione el comportamiento de las distintas fuentes de las que proceden, sin que estén necesariamente armonizados. Por ejemplo, pueden cambiar en el tiempo, dando lugar a la necesidad de analizarlos como series temporales correspondientes a periodos concretos.
- **Variedad de fuentes**: los datos se originan en diferentes fuentes distribuidas, a través de múltiples plataformas tecnológicas, pertenecientes a diferentes organismos u organizaciones, con diferentes requisitos de acceso y de formato, y pueden estar ubicados en plataformas locales, centralizadas o distribuidas en la nube (*cloud computing*).
- **Vocabulario**: es crítico mantener la coherencia, sentido y significado de la información manejada, estableciendo esquemas, modelos de datos, semántica, ontologías, taxonomías, metadatos e información basada en el contexto del contenido, para describir y controlar la estructura, sintaxis, contenido y procedencia de los datos.
- **Vaguedad**: *Big Data* es un término difuso, que no establece claramente requerimientos, límites ni potenciales resultados. Está presente en muchos entornos, soluciones, productos y a veces es considerado como un fin en sí mismo. Todo esto termina generando confusión sobre el significado y la naturaleza del tratamiento de masivo de datos.

1.2. Recursos necesarios

En el estudio del *Big Data* se deben contemplar diferentes elementos para tener una visión completa de lo que supone su ciclo de vida, desde la información de la que se parte, los procesos necesarios para su tratamiento, la gestión de los datos y la información, hasta el conocimiento que aporta mediante la materialización a través de diversas interfaces.

1.2.1. Información

La información es la materia prima del *Big Data*. En algunos casos se encontrará dispuesta de una manera estructurada y organizada, y en otros muchos casos no. Dependiendo de su relación, evolución y madurez, esta información servirá en mayor o menor medida como soporte para los procesos administrativos y asistenciales, la automatización de procesos, el apoyo a la toma de decisiones y la predicción.

La información se encontrará en diversos entornos, desde plataformas operacionales que dan soporte al tratamiento diario de la información en tiempo real, hasta modelos especializados como los almacenes de datos *Data Warehouse* y los *Data Mart*, que son específicos de cada área de conocimiento[53].

[53] Wang et al. (2014).

1.2.2. Datos estructurados y no estructurados

Un criterio inicial para clasificar los tipos de información es su organización, que puede ser de dos clases:

- **Estructurada**: la información estructurada se compone de tipos de datos básicos con un formato homogéneo predefinido, como por ejemplo números, caracteres o tipos especiales para la hora y fecha, etc. Abarca también formatos compuestos, como vectores y matrices (*array*), cadenas de caracteres (*string*), registros y uniones, etc.

  Los datos de tipo estructurado se suelen almacenar en bases de datos de tipo relacional, un modelo estándar orientado al procesamiento eficiente y optimizado para el soporte de la información de las organizaciones.

- **No estructurada**: los datos no estructurados poseen una estructura formal definida que, no obstante, no es adecuada para el desarrollo de ciertas tareas de procesamiento directo, por lo que deben ser interpretadas por un algoritmo concreto. Ejemplos de este tipo son las imágenes, vídeos, música, documentos en diferentes formatos, etc.

  En forma individual estos datos poseen una estructura variable, aunque pueden encontrarse empaquetados en objetos que sí tienen una estructura uniforme, como archivos, documentos multimedia, páginas web, etc. Asimismo, algunos datos no estructurados presentan una organización interna que facilita su tratamiento, tales como documentos XML (*eXtensible Markup Language*) y datos almacenados en bases de datos NoSQL (*Not Only Structured Query Language*). Normalmente este tipo de datos se gestionan mediante herramientas específicas, como gestores documentales, de imágenes o de vídeo, y crecen a un ritmo exponencial que provoca la necesidad de una gran capacidad para su almacenamiento[54].

1.2.3. Bases de datos

También llamadas *Operational Data Store* (ODS), las bases de datos se pueden definir como un conjunto de información relacionada que se encuentra agrupada o estructurada. Constituyen los cimientos de los sistemas de información que maneja cualquier organización, incluidas las instituciones sanitarias, para su funcionamiento diario. Pueden organizarse como un gran sistema integrado o como múltiples subsistemas de información especializados, lo que puede suponer la existencia de estructuras variadas para el almacenamiento de la información.

Independientemente de cómo estén organizadas, son las fuentes de datos operativas de la organización. Además de estructurados, los datos de un ODS están frecuentemente indexados, lo que significa que se dispone de varios criterios de ordenación de la información para facilitar su consulta y análisis. Esto permite realizar operaciones muy rápidas tanto de consulta como de registro de información, un factor crítico teniendo en cuenta la necesidad de accesos múltiples y continuos a grandes volúmenes de datos que se derivan de la actividad diaria de las organizaciones sanitarias, tanto en el ámbito de la asistencia como en el de la gestión.

Por ejemplo, deben existir unos registros estructurados que almacenen la información generada por la actividad de hospitalización, consultas ambulatorias o urgencias, y debe poderse acceder a estos registros mediante aplicaciones específicas. Estas aplicaciones están diseñadas para adaptarse a los flujos de trabajo, y no suelen tener en cuenta el análisis de eventos en tiempo real ni concebirse como apoyo a la toma de

[54] Jung y Lee (2015).

decisiones estratégicas, puesto que estos procesos requieren normalmente métodos diferentes de estructura e indexación de los datos.

1.2.4. Data Warehouse y Data Marts

Los *Data Warehouse* (DW) o *Enterprise Data Warehouse* (EDW) se pueden definir como almacenes de datos procedentes de múltiples sistemas o aplicaciones, con información histórica o consolidada, para la generación de informes analíticos específicos de las diferentes áreas de negocio de una organización. Los DW representan el primer paso hacia la creación de un sistema de inteligencia de negocio o inteligencia empresarial, comúnmente llamado *Business Intelligence* (BI)[55].

Los DW se rellenan fundamentalmente con la ayuda de procesos de Extracción, Transformación y Carga que recogen los diferentes tipos de datos y estructuras desde las bases de datos operativas ODS, los transforman a estructuras y relaciones orientadas al análisis de la información y los almacena como datos estructurados, indexados, en formato consistente, y todos disponibles en un solo lugar.

Por su parte, los *Data Mart* (DM) son subconjuntos especializados del DW orientados al análisis y generación de informes sobre un área específica de conocimiento, negocio o gestión de la organización. Están basados en modelos relacionales y optimizados para la realización de operaciones de lectura eficientes sobre datos indexados.

1.2.5. Procesos

Desde el punto de vista lógico, es necesario establecer procesos que definan los flujos y etapas en el tratamiento de la información para conseguir un modelo integral de un *Big Data*. En este apartado se analizarán algunos considerados básicos.

1.2.6. Almacenamiento

El almacenamiento de datos es uno de los primeros aspectos a diseñar y planificar en el tratamiento de la información del *Big Data*. Para ello es necesario tener claros los procesos que generan la información, con sus fuentes, flujos y necesidades para el tratamiento eficiente de los datos. De este modo se podrán definir con exactitud los requisitos de capacidad de proceso y almacenamiento de datos del sistema.

Otro punto importante es la velocidad de acceso a la información, que dependerá del soporte físico donde se almacene. Existen dispositivos muy rápidos, que son adecuados para el procesamiento en tiempo real, y dispositivos más lentos, que son más apropiados para el acceso a datos históricos, puesto que ofrecen una gran capacidad de almacenamiento. Ejemplos de dispositivos rápidos son la memoria RAM (*Random Access Memory*) de los procesadores, los discos en estado sólido (SSD, *Solid-State Drive*), o las herramientas del tipo *in-memory analytics*, en las que la información reside en la memoria de los procesadores, permitiendo así el tratamiento ultra-rápido de los datos para conseguir resultados de forma casi instantánea. Ejemplos de dispositivos lentos son los discos y cintas magnéticas, o el más reciente almacenamiento en la nube (*cloud*).

Debe contarse también con una previsión de la obsolescencia de los datos que permita estimar su validez o vigencia, ya que esto tendrá un gran impacto en el diseño de la solución para la construcción de

[55] Hurwitz et al. (2013).

una plataforma escalable con unos costes razonables. La escalabilidad se define como la capacidad de crecimiento y adaptación de la plataforma, de forma que puedan incorporarse los cambios necesarios para, con el paso del tiempo, incrementar su capacidad de almacenamiento de datos y de cálculo y poder responder a nuevas necesidades de tratamiento masivo de los datos. Esto supone la adaptación tanto de los algoritmos de proceso como de la infraestructura tecnológica de base del sistema[56].

1.2.7. Acceso

El siguiente paso es definir los roles funcionales y los criterios de seguridad en el acceso a la información, de acuerdo con las necesidades de la organización. Para ello hay que establecer los procesos de autorización de acceso a los diferentes conjuntos de datos u objetos de información, así como los mecanismos necesarios para el registro, trazabilidad y auditoría de los accesos producidos. Esto permitirá tener un conocimiento detallado de lo que ocurre con la información: quién accede, a qué datos accede, cuándo accede, y qué tratamiento y uso hace de estos datos[57].

La precisión y eficacia de estos controles depende en gran medida de la naturaleza de los conjuntos de datos y de las tecnologías empleadas para el almacenamiento de la información. Los controles en modelos de datos estructurados, como los de las bases de datos relacionales, serán más estrictos y eficientes que los correspondientes modelos no estructurados, como las tecnologías NoSQL, orientadas a objetos complejos, o el almacenamiento de tipo *cloud*, donde la granularidad de la información[58] depende de cada proveedor.

1.2.8. Orquestación

Cuando se habla de *Big Data* se presupone una complejidad en la gestión de la información que obliga a coordinar los distintos procesos de su tratamiento, para que no se interfieran y sean coherentes en el tiempo. En muchos casos se procesa información procedente desde diferentes fuentes y cada una de ella requiere procedimientos diferentes, con hitos y fases consensuadas para garantizar la integridad de la información y el correcto funcionamiento de los procesos de análisis y consolidación de los datos.

La coordinación de todos estos procesos y etapas se conoce como orquestación, y trata de alinear los requerimientos funcionales con los recursos existentes: datos, aplicaciones e infraestructura. A través de la orquestación se definen las políticas y los niveles de servicio para crear una plataforma de sistemas de información perfectamente escalable en función de las necesidades de consumo de recursos de cada aplicación: capacidad de procesamiento, almacenamiento de datos, licencias u otros costes de los sistemas informáticos[59]. Para ello se utilizan flujos de trabajo automatizados que proporcionan una gestión completa de los recursos, incluyendo su medición, control y evolución.

1.2.9. Búsqueda

La existencia de mucha información compleja, como la que supone el *Big Data*, supone procesos complejos de análisis que implican búsquedas repetitivas, enlazadas y compuestas. Por lo tanto, es necesario establecer mecanismos de acceso a la información que funcionen de manera eficiente, planificada y

[56] Keen y Moore (2015).
[57] Ye et al. (2013).
[58] Granularidad, del inglés *granularity*, no tiene aún una definición aceptada por la Real Academia Española. En almacenamiento de datos puede definirse como “la escala o nivel de detalle de un conjunto de datos”.
[59] Chang et al. (2009).

controlada. Estos mecanismos pueden afectar a la organización o al almacenamiento, procesamiento o uso de la información, de forma que la gestión de búsquedas permita optimizar el rendimiento del sistema.

Una forma de optimizar las búsquedas es crear conjuntos reducidos de datos que sean estadísticamente representativos del conjunto de datos a analizar, acotando en consecuencia el alcance de estas búsquedas. También se pueden mejorar los accesos mediante el desarrollo de algoritmos específicos para la búsqueda de información no estructurada, como por ejemplo reconocimiento de textos, análisis multidimensionales o técnicas de visualización de datos.

Asimismo, pueden ser necesarios procesos de depuración de los datos, reubicándolos en función de su uso para que los más consultados estén más rápidamente accesibles, o pasándolos a histórico[60] o incluso borrándolos para liberar espacio en las bases de datos y así poder incorporar nueva información.

1.2.10. Visualización

Como en todo sistema, una de las necesidades más importantes en el *Big Data* es la interpretación de la información que se maneja. Debido a la ingente cantidad de datos que esto supone, la capacidad y facilidad para visualizarlos eficientemente es un factor crítico. De hecho, el uso de herramientas potentes de visualización de datos es clave para una rápida exploración de los datos que permita entender su significado y convertirlos en conclusiones y conocimiento. Estas herramientas permiten representar ideas complejas de modo relativamente sencillo y comunicarlas de forma amigable, mostrando su evolución y las dependencias y correlaciones entre las diferentes dimensiones y magnitudes que sean objeto de análisis.

Una correcta visualización requiere una combinación de análisis de la información, estadística y experiencia sobre los datos presentados, mostrando aquellos que son relevantes para la toma de decisiones dentro de la organización.

1.2.11. Gestión de la Información

La gestión de la información permite garantizar la correcta interoperabilidad de los diferentes sistemas, la calidad de los datos y, por extensión, la fiabilidad en su uso.

La Data Management International[61], una asociación independiente de proveedores que analiza y estudia los conceptos de la gestión de datos, propone las siguientes funciones para la Gestión de Datos:

- **Gobierno de los Datos**: se ocupa de la planificación, supervisión y control en la gestión y uso de datos.
- **Arquitectura de Datos**: encargada de establecer los modelos, políticas y reglas para gestionar los datos.
- **Diseño y Modelado de Datos**: diseña la base de datos, implementación y soporte.
- **Almacenamiento de Datos**: función que determina cómo, cuánto y qué se almacena.
- **Seguridad de los Datos**: se encarga de todo lo relativo a la privacidad y confidencialidad, y de garantizar un acceso apropiado.

[60] El paso a histórico consiste en trasladar la información de una cierta antigüedad a otros dispositivos de almacenamiento distintos de los principales. Aunque esto supone un tiempo de acceso sensiblemente mayor, se asume que esta información no va a ser necesaria, o al menos no lo va a ser de manera urgente, lo que hace que el espacio liberado suponga una ventaja mayor que el inconveniente causado por el incremento en el tiempo de acceso a estos datos.

[61] www.dama.org

- **Integración e Interoperabilidad de los Datos**: responsable de definir la integración y transferencia de los datos.
- **Documentos y Contenidos**: establece las reglas aplicables a los datos fuera de las bases de datos.
- **Referencias y Patrones de Datos**: busca aportar una visión completa de la información.
- **Repositorios de Datos e Inteligencia de Negocios**: se ocupa de lo referente a datos históricos y analíticos.
- **Metadatos**: trata de integrar, controlar y proporcionar los metadatos de la información.
- **Calidad de Datos**: define los procesos de control y mejora de la calidad de los datos.

1.2.12. Virtualización

La virtualización es la capacidad de aislar ciertas propiedades de los sistemas de información para flexibilizar, asegurar, escalar y garantizar su evolución a lo largo del tiempo, así como su expansión dentro de las organizaciones. De este modo se pueden definir diferentes servicios de almacenamiento, procesado y análisis de la información, y virtualizarlos mediante la simulación de componentes de servicio sobre una infraestructura física común, de modo que se pueda evaluar y validar individualmente el comportamiento y rendimiento de cada uno de ellos, y también la forma en la que varios de ellos interactúan como partes de una aplicación o sistema de información más complejo.

La virtualización de servicios permite ajustar las capacidades tecnológicas a los requerimientos de las organizaciones en función de su uso previsto y también reaccionar rápidamente a nuevas demandas, como pueden ser almacenamientos más rápidos o voluminosos, o una mayor exigencia de velocidad de procesamiento en función de la carga de trabajo o del uso de nuevas herramientas analíticas. Al tratarse de componentes simulados, un cambio en su configuración permite un ajuste de los recursos asignados mucho más rápido que el tradicional despliegue de nuevos dispositivos físicos.

1.2.13. Integración

La integración de datos es de gran importancia debido a la variedad, volumen de datos y sistemas de información presentes en organizaciones complejas como las sanitarias. Una correcta integración se convierte en una ventaja estratégica ante los nuevos escenarios y retos presentes, pues permite la interoperabilidad de sistemas, evita registros duplicados de información, elimina errores de transcripción y facilita la trazabilidad de los datos.

Integrar es necesario para la migración y sincronización de datos entre las aplicaciones operativas, para la consolidación y análisis de datos históricos, para el intercambio de datos en una arquitectura SOA (*Service Oriented Architecture*, Arquitectura Orientada a Servicios) o entre organizaciones, para el manejo de datos en la nube o para la integración de subsistemas de cara a su explotación en los *Big Data*[62].

Existen herramientas especializadas para integrar los diferentes esquemas de datos y servicios, normalizando la codificación de los datos mediante el uso de ontologías para así mantener el significado de la información en los diferentes sistemas.

[62] Das et al. (2010).

1.2.14. Calidad

La calidad de la información es un concepto que combina su correcto almacenamiento, uso, tratamiento y difusión de acuerdo a los condicionantes del negocio (en este caso, del sector sanitario), incluyendo además las obligaciones legales relativas a la gestión de los datos para garantizar que su tratamiento sea adecuado, pertinente y en modo alguno excesivo. La incorporación de estos requisitos establece unas garantías y genera cultura y confianza sobre el correcto uso de la información[63].

Por todo ello, se debe establecer unos procedimientos específicos para garantizar la calidad de la información, identificando y midiendo los indicadores que permitan definir los objetivos de calidad a alcanzar. Estos procedimientos se deben extender a toda la cadena de la información (fuentes, codificaciones, integraciones, algoritmos, explotación, visualización, etc.) para poder hacer un seguimiento de los objetivos establecidos en cada una de sus fases. Estableciendo estas medidas se minimizan riesgos, se ahorran tiempo y recursos, y se mejora el rendimiento de la infraestructura tecnológica, ya que se evitan duplicidades, incoherencias, errores o incluso datos superfluos en la información almacenada[64].

1.2.15. Gobierno de Datos

En cualquier proyecto de sistemas de información es necesario coordinar las necesidades de las organizaciones, los procesos establecidos y la tecnología empleada para convertir la información generada en un recurso de gran valor. Esta labor se denomina gobierno de datos. En el caso del *Big Data*, las dimensiones de los datos (las 10 V explicadas con anterioridad) hacen que este gobierno sea crítico, debido al gran impacto y coste en asignación de recursos y capacidad de respuesta que cualquier error de diseño, dimensionamiento o ejecución puede suponer. Obviamente, el gobierno de los datos será una tarea más exigente cuanto mayor sea la complejidad de la organización y sus sistemas de información.

En los *Big Data* se suelen integrar múltiples sistemas y subsistemas de información que normalmente se han desarrollado en momentos distintos, utilizando tecnologías de naturaleza y madurez diferentes, y sobre los que se han incorporado gradualmente nuevas funcionalidades, desarrollos y transformaciones o traspaso de datos. Esto obliga a establecer claramente un mapa evolutivo de cada conjunto de información, especificando su persistencia en el tiempo y, si es necesario, planificar su desuso o destrucción, cerrando así su ciclo de vida.

Otro aspecto importante es procurar que los proyectos de desarrollo de sistemas de información sean lo suficientemente dinámicos para poder adaptarse a nuevas necesidades de datos, previendo y facilitando la incorporación de nuevas fuentes, la aplicación de nuevas técnicas de análisis, la incorporación de herramientas o el uso de nuevos mecanismos de visualización de información.

Las etapas que se pueden definir para el gobierno de datos no difieren en gran medida de las de un sistema de información en general:

- Establecer metas de la organización sobre la información, que permitirán definir los principios que guían la operación y desarrollo de la cadena de suministro de la información.
- Definir métricas viables para evaluar la efectividad del programa de desarrollo de los sistemas de información y los procesos de gobierno asociados.

[63] Cormode y Srivastava (2009), Lorch et al. (2013), Ghoting et al. (2009), Chen et al. (2004).
[64] Wu and Zhu X. (2008).

- Tomar decisiones efectivas, que permitan que la estructura organizacional y el modelo de sistemas de información sean facilitadores del cambio e instrumentos de mejora de las organizaciones.
- Comunicar los cambios y políticas sobre los sistemas de información, de modo que la organización esté alineada con sus objetivos y metas.
- Establecer métricas sobre los resultados de los análisis facilitados por el *Big Data*, para poder comparar los resultados de las políticas con las metas establecidas.
- Auditar los sistemas, procesos y resultados, para verificar de manera objetiva los resultados obtenidos y compararlos con los estándares de buenas prácticas (ontologías, codificaciones, algoritmos, costes, etc.).

1.2.16. Inteligencia

La inteligencia aplicada a una organización o sistema de información hace referencia a la capacidad de utilizar los datos para apoyar la toma de decisiones en los diferentes niveles jerárquicos de la institución.

La base de la pirámide de información está compuesta por los datos directamente derivados del ejercicio de la práctica asistencial: listas de trabajo, planes de cuidados de enfermería, listados de pacientes citados, plan de medicación, etc. En otras palabras, cualquier dato relacionado con la realización o apoyo a la asistencia sanitaria diaria.

El siguiente nivel está constituido por los cuadros de mando que resumen, clasifican y ponderan las actividades o indicadores necesarios para cada área de negocio o conocimiento. En este nivel debe incorporarse una continuidad temporal, permitiendo hacer un análisis no sólo del estado de situación sino también de su trayectoria en un período determinado, de manera que se puedan calcular las desviaciones respecto de los objetivos definidos para ese mismo período, o bien la evolución con respecto a períodos anteriores. Al tratarse de datos que permiten diagnosticar la situación y su tendencia actual, además de las diferencias de estado entre diferentes momentos, se puede hablar de información de tipo descriptivo.

En el tercer nivel se incorporan buenas prácticas, métricas, objetivos y el conocimiento previo de la organización, con el fin de que el sistema pueda ofrecer recomendaciones sobre las decisiones o medidas que deben tomarse para corregir el rumbo de la actividad y cumplir las metas establecidas. Todo ello da como resultado la creación de herramientas o sistemas de apoyo a la toma de decisiones. En el ámbito sanitario, y más concretamente en el entorno asistencial, un sistema de este tipo puede alimentarse de guías clínicas, guías de medicamentos, procedimientos de la organización o criterios de seguridad del paciente para, por ejemplo, aconsejar al médico sobre la dosis más adecuada del medicamento que va a prescribir a un paciente en concreto, o alertarle sobre posibles efectos adversos teniendo en cuenta las características particulares de ese mismo paciente. En este nivel se puede hablar –valga la redundancia– de información prescriptiva.

En el nivel superior se utilizan los datos para realizar previsiones del estado de situación a corto, medio o largo plazo. Esta información es más valiosa cuanto más acertada resulte, puesto que permite a la organización prever situaciones indeseadas y evitar las consecuencias negativas que se pueden derivar de ellas, como desperdicio de recursos, falta de medios o sobrecostes, entre otros[65]. Establecer los requisitos necesarios para poder realizar estas predicciones es uno de los elementos clave para el avance de las organizaciones sanitarias, siendo uno de los máximos exponentes de ello la medicina preventiva, tanto por el impacto que puede suponer en la salud de los pacientes como por el seguimiento de unos métodos de actuación, basados a su vez en la definición de unos requerimientos de información, la formulación y posterior confirmación de hipótesis de trabajo y el establecimiento de una serie de actuaciones, cada una de ellas con la debida prioridad. En este nivel procede habla de información predictiva.

[65] Bollen et al. (2011).

Tradicionalmente, todos estos niveles se han basado en el análisis de información y la aportación de conocimiento y experiencia por parte de un ser humano, con sus correspondientes limitaciones. La aplicación del *Big Data* en esta etapa supone la aparición de sistemas de aprendizaje automático (*Machine Learning*), basados en modelos de analítica predictiva[66]. En estos modelos se prescinde de la participación de un experto humano, y son varios algoritmos de clasificación, agrupación y correlación los que se encargan de generar hipótesis, depurar las posibles dependencias o errores mediante la inclusión de nuevos datos, armonizar las diferencias encontradas entre los distintos grupos o clasificaciones, y finalmente obtener predicciones individualizadas para cada caso, todo ello de manera automática. Este proceso de aprendizaje puede ser supervisado por expertos humanos, que dirigen los procesos de entrenamiento y aprendizaje del sistema, o puede ser totalmente autónomo, en cuyo caso es el propio sistema el que genera sus propias reglas, realimentándose continuamente para mejorar la precisión de sus predicciones[67].

En el campo del *Big Data*, debido al volumen, complejidad y variabilidad de la información que se maneja, la aplicación de herramientas de aprendizaje automático resulta de gran interés, puesto que puede hacer viable el análisis predictivo de grandes cantidades de datos. En el entorno sanitario, esto puede suponer un impulso muy importante para la toma de decisiones estratégicas, de gestión y clínicas.

1.3. Etapas de Tratamiento

1.3.1. Gestión de datos

El primer paso para el tratamiento de los datos es una correcta gestión de los mismos, entendidos como cada uno de los elementos que establece una característica específica y mensurable de un hecho determinado y que se recoge en forma empírica y objetiva[68]. Para planificar la gestión de los datos deben tenerse en cuenta las diferentes fuentes de información que se manejan.

En el caso de las organizaciones sanitarias existe una gran cantidad de fuentes y métodos de generación, que a su vez pueden ser muy complejos y heterogéneos. Buena muestra de ello son los diferentes sistemas de información que dan soporte a la actividad asistencial, a los procesos de aprovisionamiento o a la gestión de recursos humanos, entre otros. La consecuencia directa de todo esto es una gran variabilidad de los tipos de datos y de las tecnologías implicadas en su generación, como por ejemplo la información generada por los diversos equipos utilizados por los laboratorios clínicos, los servicios de radiología o los tratamientos de radioterapia.

El siguiente paso es la extracción de los datos desde los dispositivos y sistemas de información que se emplean durante la actividad diaria, de modo que se seleccione y recolecte información relevante y consolidada. Para ello se debe definir previamente qué información se debe registrar, el modo en que se registra, y en qué momento o momentos del proceso (durante su realización, tras su conclusión o tras una verificación final, por ejemplo). Esto permite establecer el ciclo de vida de los procesos y también estimar otros requisitos adicionales, como por ejemplo la capacidad de almacenamiento necesaria, los momentos más apropiados para no penalizar el rendimiento de los sistemas[69].

[66] Wu et al. (2013).
[67] Ghoting et al. (2009).
[68] Indarte y Vero (2014)
[69] Alam et al. (2012).

En algunos casos puede ser necesario depurar la información, simplificando y unificando los datos recogidos, codificándolos, verificando su calidad y optimizando la integridad referencial[70] de las bases de datos que los almacenan. Cada una de estas tareas de depuración puede requerir un proceso o subproceso específico.

Una vez verificada la calidad de la información, el siguiente paso es la gestión de su almacenamiento y de su ciclo de vida, no sólo para su análisis y explotación sino también para establecer los modelos de pruebas, carga o rendimiento de los sistemas. Al igual que en otros sistemas de información, normalmente son necesarios tres entornos de trabajo: uno de desarrollo de la solución, en el que se introducen las distintas modificaciones según las necesidades existentes en cada momento; otro de preproducción, donde se verifica el correcto funcionamiento de las nuevas funcionalidades y se realizan pruebas para evaluar la capacidad de procesamiento; y otro de producción, que funcionará con datos reales y será sobre el que trabajen los usuarios finales. Asimismo, será necesario definir unos protocolos de copia de seguridad de la información, de manera que los datos puedan restaurarse en caso de pérdida o deterioro de los mismos.

Durante el mantenimiento de los datos, es posible que se dé la necesidad de consolidar información u optimizar el rendimiento del sistema. Esto obliga a acometer una reestructuración del modelo de base de datos, agrupando tablas o creando tablas resumen para poder simplificar y optimizar las búsquedas, incrementando así la eficiencia de las consultas de datos[71].

1.3.2. Ciencia de los datos

La ciencia de los datos trata el análisis y conocimiento profundo de sus diferentes dimensiones, desde las teorías sobre el análisis de datos fundamentado en la estadística hasta las herramientas que lo potencian y facilitan[72]. El objetivo es poder definir, en un marco teórico, las interfaces de comunicación con las diferentes fuentes de información, normalizando su ciclo de vida y validando la calidad de los datos[73]. Asimismo, debe establecerse también el modelo lógico de datos del sistema *Big Data* que se desea construir, guardando la debida coherencia entre los modelos de los distintos sistemas de información que se integren con él. Obviamente, el modelo de datos de cada sistema dependerá sobre todo del tipo de datos con los que trabaje, de la plataforma tecnológica sobre la que funcione y, en el caso del *Big Data*, de las herramientas empleadas para el tratamiento masivo de datos.

En el ámbito de la ciencia de los datos se incluye también la presentación de la información, que tiene unos requisitos propios tal y como se explicó en el apartado sobre el concepto de visualización de los datos. Actualmente, las prestaciones de los sistemas informáticos han evolucionado hasta el punto de permitir el uso de herramientas interactivas bastante potentes para facilitar su mejor comprensión. No obstante, los requisitos para una visualización eficiente de la información pueden influir en la organización de los datos, por lo que deben tenerse en cuenta a la hora de definir el modelo correspondiente[74].

Otra de las grandes líneas de actuación es, como se describió en el anterior apartado sobre inteligencia, la generación de conocimiento para la construcción de herramientas de tipo prescriptivo, generando recomendaciones sobre la toma de decisiones, o predictivo, para la anticipación de los hechos

[70] Propiedad de las bases de datos que asegura que las posibles relaciones entre sus distintos registros son correctas. En otras palabras, cuando un registro de la base de datos se relaciona con otros, la integridad referencial garantiza que estos registros existen, que no hay redundancias ni incoherencias y, en consecuencia, que los datos son correctos.
[71] Luo et al. (2012).
[72] Schroeck et al. (2012).
[73] Bughin et al. (2010).
[74] Chen et al. (2004).

futuros en función de las circunstancias actuales. En problemas tan complejos como la evolución de las enfermedades, los tratamientos personalizados o la planificación de recursos, los procesos *Big Data* pueden llegar a ser muy importantes para las organizaciones. En algunos casos pueden ser necesarios sistemas de tiempo real, como sucede por ejemplo en herramientas de apoyo al diagnóstico o al soporte vital[75].

1.4. Metodologías

Los métodos puestos en práctica en el *Big Data* se basan en estándares en diferentes campos y componentes del manejo de la información. La decisión sobre las diferentes metodologías aplicables debe tomarse teniendo en cuenta la visión global de la organización, los sistemas de información implicados, las tecnologías disponibles, el tratamiento de la información, y regulatorias o las específicas sobre la seguridad y privacidad de la Información.

Estas metodologías pueden clasificarse en dos grandes grupos: las dependientes del área de conocimiento y los algoritmos puestos en marcha.

1.4.1. Visión del Big Data

El tratamiento de la información en el *Big Data* requiere tener una visión estratégica para conseguir implantar un modelo útil de análisis de la información. Este modelo debe estar alineado con la estrategia de la organización a la que se aplica, de modo que los objetivos específicos del análisis de información sean coherentes con los objetivos generales de la institución.

Otro aspecto a tener en cuenta son las necesidades funcionales de los usuarios en lo referente al análisis de datos, de modo que se consideren los diferentes perfiles de consumo funcional de la información y los plazos de tiempo requeridos[76]. Lo mismo puede decirse de la tecnología existente o requerida para proveer las funcionalidades necesarias.

1.4.2. Algoritmos

El otro gran grupo de metodologías para la creación de un sistema *Big Data* trata sobre los diferentes algoritmos aplicados a los diferentes elementos y fases de su desarrollo[77].

Es fundamental tener un conocimiento detallado de los procesos de gestión de la información específica de la organización, y en especial de la semántica asociada. En el sector sanitario se maneja gran cantidad de información heterogénea, con circuitos muy complejos en los que participan multitud de agentes e interlocutores, lo que dificulta su recolección. Ante esta situación se puede plantear la aplicación de modelos para el análisis semántico de las diferentes fuentes de datos, la utilización de ontologías y el uso de agentes específicos para el tratamiento de datos con un formato concreto, como por ejemplo herramientas de análisis de imágenes, reconocimiento de voz, clustering, etc[78].

Por otra parte, el conocimiento asociado a los procesos se encuentra normalmente disperso entre los usuarios de los diferentes sistemas, puesto que cada uno de ellos se ciñe a las etapas en las que participa,

[75] Rana et al. (2015).
[76] Raghupathi y Raghupathi (2014).
[77] Chen et al. (2012).
[78] Barry et al. (2015).

y es raro que exista documentación formal que recoja detalladamente la naturaleza de cada proceso en forma de algoritmos concretos, privando a la organización de una base formal de aplicación. Un ejemplo de este tipo de documentos son las Guías de Práctica Clínica.

Si a todo esto añadimos que estos procesos cambian en el tiempo, pudiendo perder algunas de sus referencias históricas más importantes, resulta aún más difícil definir y mantener un modelo coherente de datos y, por extensión, conservar el significado de los mismos. Un ejemplo de ello es un cambio de técnica de laboratorio que implica una nueva escala para la obtención de resultados o una modificación de los rangos de normalidad de las determinaciones[79].

En cuanto a los algoritmos para el intercambio de datos entre sistemas y subsistemas de información, es habitual que participen procesos y tecnologías diferentes (SOAP, REST, DICOM, XML, etc.), y además suele tratarse de sistemas dinámicos con diferentes niveles de madurez y estadio en su ciclo de vida[80]. Toda esta complejidad debe tenerse en cuenta a la hora de definir los modelos de datos, los procesos de extracción y depuración, y los algoritmos de análisis del *Big Data*.

Además, como se explicó en un apartado anterior, es necesario establecer también políticas de acceso y uso de la información a nivel de organización, sistema y subsistema. Cada uno de estos entornos (historia clínica, investigación, biobancos, etc.) tendrá unos criterios de restricción de acceso y uso diferentes, que por extensión condicionan el tratamiento de los datos en el sistema *Big Data*. Los usuarios habituales suelen pertenecer a perfiles de investigación, gestión y control.

Estas políticas de control de acceso y uso se definen con bases en criterios específicos del sector sanitario, incluyendo el marco legal correspondiente. Los diferentes niveles y permisos de acceso deben establecerse en función del tipo de datos, de la etapa del proceso (recolección, tratamiento, transmisión, etc.), del ámbito de la información y del nivel de detalle de los registros o vistas definidas[81].

1.5. Futuro del *Big Data*

1.5.1. Lago de Datos (Data Lake)

La invención del término *lago de datos* se atribuye a James Dixon[82], que lo describió en su blog: *"Si se piensa en un Data Mark como depósito de agua embotellada –limpiado y empaquetado y estructurado para el consumo fácil–, el lago de datos es una gran masa de agua en un estado más natural, el contenido del lago de flujo de datos desde una fuente para llenar el lago, y varios usuarios del lago pueden llegar a examinar, bucear en él, o tomar muestras"*[83].

La idea del lago de datos es tener un único almacén de todos los datos de la organización que van desde los datos en bruto, es decir, una copia exacta de los datos del sistema de origen, hasta datos transformados que se utilizan para diversos fines, incluyendo informes, visualización, análisis y aprendizaje automático. El lago de datos incluye datos estructurados (extraídos de bases de datos relacionales), datos semiestructurados (CSV, registros, XML y nuevos formatos como JSON) y datos no estructurados (correos electrónicos, documentos, archivos PDF, imágenes, audio, vídeo), creando así un almacén centralizado que

[79] Estape et al. (2016).
[80] Ahmed y Karypis (2012).
[81] Sariyar et al. (2015).
[82] http://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/
[83] http://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

admite prácticamente cualquier formato de datos. En este sentido, existe la opinión bastante extendida de que el lago de datos no es más que una nueva denominación del clásico repositorio de datos, siendo su ejemplo más claro la tradicional carpeta de ficheros.

Esta flexibilidad conlleva riesgos sustanciales, como la falta de supervisión y gobierno, y la dificultad o en algunos casos imposibilidad para gestionar la calidad de los datos y su ciclo de vida. El lago de datos carece de un mecanismo de metadatos[84], lo que puede llevarlo a convertirse más bien en un "pantano". Sin la información de los metadatos se pierden la trazabilidad de los procesos y la posibilidad de aprovechar el trabajo realizado anteriormente.

Otro gran riesgo es la dificultad para implantar medidas de seguridad y control de acceso y uso de la información. Los datos pueden ser almacenados y consultados prácticamente sin restricciones ni supervisión. Teniendo en cuenta que cuando la privacidad y los requisitos legales imponen mecanismos estrictos de acceso y de trazabilidad[85], la aplicación del lago de datos al entorno sanitario se reduce a datos de acceso libre.

1.5.2. Datos Rápidos (Fast Data)

El tratamiento de *datos rápidos* se puede definir como una variante del *Big Data* en la que los datos se generan a gran velocidad, y por lo tanto se requiere una velocidad de procesamiento equivalente que permita su tratamiento sin saturar la capacidad del sistema. Esto suele suceder cuando existen múltiples fuentes de datos, o cuando los procesos generan grandes volúmenes de información a un ritmo muy elevado. Esta situación será cada vez más frecuente, al estimarse que la información generada está creciendo hasta el punto de duplicarse aproximadamente cada dos años.

Los procesos estándar de *Big Data* no tienen capacidad para dar respuesta a esta necesidad, puesto que su velocidad de procesamiento no es lo suficientemente rápida para asumir este flujo masivo de nuevos datos. En el caso de sistemas de tiempo real, donde las decisiones son críticas[86], debe identificarse y filtrarse toda la información significativa almacenada en los grandes volúmenes de datos existentes, y después debe procesarse a una velocidad que le permita servir de apoyo a la toma de decisiones en tiempo y forma.

1.5.3. Análisis Masivo (Big Analytics)

El análisis de grandes volúmenes consiste en convertir la información en conocimiento utilizando una combinación de enfoques nuevos y existentes, desde las herramientas estándar de análisis de información mediante paquetes estadísticos, como MATLAB, SAS, y R, hasta los sistemas de específicos para el *Big Data* y la incorporación del aprendizaje automático. El análisis masivo acumula y consolida la información resultante de la ejecución de los algoritmos parciales, e intenta llegar a conclusiones para la detección de tendencias ocultas o dependencias complejas entre datos.

Cuanto mayor es la cantidad de información y la envergadura de los análisis realizados, más se tiende a automatizar la generación de conclusiones, prescindiendo por lo tanto de la intervención de usuarios expertos que realicen preguntas o formulen hipótesis.

[84] Los metadatos son una herramienta que propociona información acerca de los conjuntos de datos y los procesos realizados sobre ellos: naturaleza, autoría, integridad, control, etc. Es decir, son datos cuya finalidad es describir otros datos.
[85] Duncan (2007), Schadt (2012).
[86] Manyika et al. (2011).

1.5.4. Búsquedas profundas (Deep Analysis).

Las búsquedas o análisis profundos son el resultado de aplicar fuentes de datos, procesos y algoritmos específicos creados para un problema concreto, debido a que las herramientas generalistas de análisis de grandes volúmenes de datos no son suficientes para ese caso. En otras palabras, pueden entenderse como una especialización del *Big Data*.

Estas búsquedas profundas pueden resultar de interés para las organizaciones que deseen mejorar el conocimiento de áreas específicas de gran impacto[87].

2. Tecnologías aplicadas al *Big Data*

2.1. Arquitecturas

El primer elemento para la construcción de un sistema *Big Data* es la definición de las arquitecturas de las plataformas tecnológicas que soportan su alojamiento y su procesamiento. Esta arquitectura combina componentes hardware, software y comunicaciones.

Las arquitecturas han evolucionado de modelos estáticos, con soluciones finalistas basadas en la instalación de servidores dedicados a una aplicación concreta, a otros dinámicos, con sistemas centralizados que concentran y homogeneizan los recursos para reducir costes y optimizar los despliegues. Actualmente las arquitecturas están evolucionando hacia soluciones de virtualización, que como se explicó en un apartado anterior permiten simular componentes y reasignar recursos con una gran flexibilidad, y hacia modelos en la nube, que ofrecen una flexibilidad similar y además externalizan la función de alojamiento de los sistemas.

Los modelos en la nube plantean tres niveles en su diseño:

- **Infrastructure as a Service (IaaS).** Este nivel es el que más se aproxima al tradicional esquema de plataforma propia de la organización. Consiste en arrendar la plataforma remota de un proveedor, y permite adecuar el dimensionamiento de estos recursos tecnológicos a cambios en las demandas de procesamiento y almacenamiento. El principal inconveniente es que los sistemas de información se deben adecuar para poder funcionar bajo esta modalidad. Un ejemplo es el almacenamiento remoto de copias de seguridad.
- **Software as a Service (SaaS).** En este nivel se utiliza el software de un proveedor configurado en función de las necesidades, contratándose licencias de uso dinámicas y pudiendo incorporarse nuevas funcionalidades en caso necesario. Un ejemplo es el uso de software ofimático en la nube, sin necesidad de desplegarlo en los equipos de los puestos de trabajo de cada usuario.
- **Data as a Service (DaaS).** En este nivel se contrata el uso de sistemas de información completos a los que se accede mediante los navegadores disponibles de forma estándar en cualquier dispositivo informático, tanto ordenadores como dispositivos móviles. De este modo la organización obtiene un servicio finalista de los datos en áreas específicas de conocimiento o funcionalidad. Un ejemplo son las herramientas analíticas, de visualización o el aprendizaje automático.

Al margen del modelo de arquitectura por el que se opte, es fundamental definir una plataforma tecnológica que satisfaga los siguientes requisitos[88]:

[87] Hitz y Katsanis (2014).

- **Capacidad de procesamiento y almacenamiento de datos**, combinando criterios de capacidad de cálculo, entrada y salida de datos a gran velocidad, copias de seguridad, y control de los accesos, entre otros. Cuando se trabaje con procesos analíticos que entrañen una gran complejidad o requieran sistemas de tiempo real, puede recurrirse a modelos de procesamiento sobre la propia base de datos, en los que la información se almacena directamente en memoria (*in-memory analytics*, como se explicó anteriormente) para eliminar los retrasos ocasionados por los procesos de transferencia de información entre los dispositivos de almacenamiento y la memoria de los procesadores. Es decir los datos se van analizando conforme se van generando[89].

  Si el cuello de botella se encuentra en el acceso a los datos, el modelo más flexible es el *Data Fabric* o *Data Grid* (entramado de datos), específicamente diseñado para generar un "tejido" de información a la que se puede acceder de forma eficiente con independencia del número de nodos o clientes que estén enviando consultas.
- **Posibilidad de crecimiento**, proporcionando la escalabilidad necesaria para maximizar la vida útil del sistema. En el caso del *Big Data* resulta especialmente apropiado el modelo *Grid Computing*, basado en la computación mediante servidores en red, puesto que permite crear y ampliar sistemas con una gran potencia de cálculo, asignando decenas, cientos o miles de procesadores para el tratamiento en paralelo de la información[90].
- **Garantía de un nivel de servicio** mediante mecanismos de alta disponibilidad, que permitan garantizar una funcionalidad mínima en caso de avería de algún componente.

2.2. Plataformas y Herramientas

El siguiente paso para la construcción de un sistema *Big Data* es la definición de las plataformas y las herramientas que dan soporte al tratamiento de los datos.

Teniendo en cuenta las fuentes de datos y los sistemas de información implicados en el proceso, se deben definir los flujos y métodos de comunicación necesarios, considerando los distintos requisitos para la integración de sistemas, el establecimiento de conexiones lógicas entre ellos y la seguridad en el intercambio de datos. Normalmente se utilizarán plataformas de integración (*middleware*) que realizan las funciones de extracción, transformación, codificación y generación de los datos, optimizándolos para su posterior tratamiento[91]. También deben definirse protocolos de comunicaciones y controles de acceso, crearse redes virtuales, e implementarse mecanismos de encriptado de la información, entre otras medidas[92].

En cuanto a las plataformas de almacenamiento, se deben definir en función del tipo de datos, dimensiones y rendimientos esperados respecto a su carga y la explotación[93]. Se pueden emplear distintos tipos de bases de datos (relacionales, NoSQL, documentales, de imágenes, etc.) en función de las prestaciones que se busquen en cada caso, y también se puede elegir entre plataformas centralizadas, distribuidas o en la nube. En caso necesario, puede valorarse una solución híbrida que combine varias de estas herramientas.

La importancia de esta etapa reside en que en ella comienzan los diferentes procesos de gestión del ciclo de vida de los datos, fijando un marco de trabajo para el despliegue del *Big Data* en el que deben estar contemplados una gestión eficiente de los datos, sus flujos, la gestión de la calidad, y las conexiones entre

[88] Wang et al. (2013).
[89] Reed et al. (2011).
[90] Papadimitriou y Sun (2008).
[91] Zikopoulos et al. (2012).
[92] Silva et al. (2012).
[93] Su et al. (2006).

las herramientas que realizan estos procesos, para así poder estimar las prestaciones de la plataforma tecnológica que se necesita[94].

Finalmente, deben coordinarse las distintas herramientas analíticas y los objetivos específicos de cada una de ellas. Para ello hay que comenzar estableciendo los esquemas y agrupaciones de los datos, con vistas a su posterior tratamiento y análisis, y los requisitos de carga y procesamiento de la información. Las herramientas de generación de informes, visualización dinámica y análisis estadístico son el siguiente paso, puesto que ofrecen el primer resultado del *Big Data*, presentando información resumida y de fácil comprensión, permitiendo al usuario consultar grupos de datos estadísticos y en algunos casos interactuar con el sistema para explorar esta información. Después intervienen las herramientas de minería de datos, que ayudarán al descubrimiento de relaciones y dependencias entre los distintos datos o grupos de datos[95]. Estas herramientas facilitan el análisis automatizado y predictivo, y sirven de base para los sistemas de apoyo a la toma de decisiones, ya sean clínicas o de gestión.

2.3. Soluciones

En los últimos años se ha producido un desarrollo creciente de herramientas y soluciones específicas para el *Big Data*, y al mismo tiempo los productos de software generalistas han evolucionado para adaptarse a este nuevo entorno. Como resultado, existe un gran catálogo de soluciones para el tratamiento masivo de datos. Entre las soluciones de propósito general cabe destacar las siguientes:

- **Bases de datos relacionales:** orientadas a un equilibrio entre su rendimiento y flexibilidad, con alto nivel seguridad, autorización, autenticación e integridad. Ejemplos de productos comerciales son Oracle, MySQL, PostgreSQL, MariaDB, SQLite, etc.
- **Bases de datos no relacionales (NoSQL):** orientadas a la escalabilidad, redundancia, flexibilidad y coste, como MongoDB, Redis, Cassandra, CouchDB, etc.

A continuación se enumeran varias soluciones de procesamiento orientadas al *Big Data*:

- **Hadoop Distributed File System (HDFS):** es un sistema de ficheros orientado al almacenamiento de grandes volúmenes de datos no estructurados, distribuido y escalable en lenguaje Java. Se enmarca en las etapas de almacenamiento y explotación de los datos, siendo muy utilizado.
- **MapReduce**: es un marco de software que simplifica el desarrollo y ejecución de aplicaciones altamente paralelizadas. Cuenta con una función "Map" que divide una consulta en múltiples elementos para que sean procesados nodo a nodo, y con una función "Reduce" que agrega los resultados calculados por "Map" para determinar la respuesta planteada en la consulta. Se aplica a las fases de análisis y sus algoritmos son ampliamente aceptados.
- **Hive**: es un marco de *Data Warehouse* basado en Hadoop que permite formular una consulta tipo SQL, definida como HIVEQL, para que pueda ser procesado por MapReduce. Permite la integración y explotación de datos a alto nivel, ya que se pueden hacer consultas complejas.
- **Pig**: es un lenguaje basado en Hadoop orientado al tratamiento de datos en *Big Data*, que permite obviar los límites del SQL. Está orientado a los flujos de datos para programadores.
- **HBase**: es una base de datos no relacional que ofrece un alto rendimiento en búsquedas rápidas sobre Hadoop. Añade funciones de transaccionalidad, permitiendo actualizaciones, inserciones y borrado. Pertenece a Apache Software Foundation y complementa las funcionalidades de Hadoop.
- **Flume**: es un marco de propagación y almacenamiento de datos en Hadoop.

[94] Kuchinke et al. (2016).
[95] Rajaraman and Ullman (2011).

- **Sqoop**: es una herramienta de conectividad que permite la carga de datos de bases de datos relacionales y otros *Data Warehouse* en Hadoop.
- **Mahout**: es una librería de análisis de datos con los algoritmos más frecuentes sobre clustering, regresiones, modelos estadísticos, etc., siguiendo el modelo de MapReduce.
- **ZooKeeper**: es un coordinador de servicios centralizados para el mantenimiento de las configuraciones de la información y su identificación. Permite la agrupación de servicios distribuidos y sincronizados, y realiza las funciones de orquestación.
- **Amazon Web Services (AWS):** proporciona una amplia plataforma de servicios administrados para construir, asegurar y escalar fácilmente aplicaciones de *Big Data* de principio a fin, de forma rápida y sencilla.
- **Cortana Analytics:** es la propuesta de Microsoft como conjunto de herramientas en la gestión de la información, almacenamiento de datos, aprendizaje automático, cuadro de mandos y visualización.
- **IBM Watson Analytics service:** es la solución que propone IBM, basada en la nube, en el procesamiento de lenguaje natural y en modelos de aprendizaje automático, para analizar grandes volúmenes de datos no estructurados.
- **Oracle Big Data Cloud Service** y **Big Data SQL Cloud Service**: son servicios orientados a herramientas en la nube para facilitar su uso y despliegue, manejando tanto datos SQL como NoSQL.
- **Sinequa ES**: es una plataforma de búsqueda y análisis, basada en el procesamiento de lenguaje natural, que realiza análisis estadístico de datos estructurados y análisis semántico y sintáctico de textos.
- **Splunk Enterprise** y **Splunk Cloud**: integra datos desde las diferentes fuentes con un lenguaje de procesamiento de búsquedas, manteniendo una visión global de los datos, tanto históricos como de tiempo real.
- **Tableau**: enfocado al análisis o búsqueda de datos visual, simple y rápido. Puede instalarse en una plataforma local o contratarse en modalidad de pago por uso (SaaS).
- **Trillium Software:** herramienta para la gestión de la calidad de los datos, especializada en varias plataformas, incluyendo Hadoop, orientada al gobierno de la información y a la preparación de los datos para el análisis.

Bibliografía

- Ahmed R., George Karypis. Algorithms for mining the evolution of conserved relational states in dynamic networks, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 603-630.
- Alam et al. 2012, Md. Hijbul Alam, JongWoo Ha, SangKeun Lee, Novel approaches to crawling important pages early, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 707-734.
- Barry W.T., Charles M. Perou, P. Kelly Marcom, Lisa A. Carey, Joseph G. Ibrahim. (2015) The Use of Bayesian Hierarchical Models for Adaptive Randomization in Biomarker-Driven Phase II Studies. Journal of Biopharmaceutical Statistics 25, 66-88.
- Bollen J., H. Mao, and X. Zeng, Twitter Mood Predicts the Stock Market, Journal of Computational Science, 2(1):1-8, 2011.
- Bughin J, M Chui, J Manyika, Clouds, big data, and smart assets: Ten tech-enabled business trends to watch, McKinSey Quarterly, 2010.
- Chang E.Y., Bai H., and Zhu K., Parallel algorithms for mining large-scale rich-media data, In: Proceedings of the 17th ACM International Conference on Multimedia (MM '09), New York, NY, USA, 2009, pp. 917-918.

- Chen R., K. Sivakumar, and H. Kargupta, Collective Mining of Bayesian Networks from Distributed Heterogeneous Data, Knowledge and Information Systems, 6(2):164-187, 2004.
- Chen, H., Chiang, R.H.L. and Storey, V.C. "Business Intelligence and Analytics: From Big Data to Big Impact", MIS Quarterly, 36(4), 2012, pp. 1165-1188.
- Cormode G. and Srivastava D. 2009, Anonymized Data: Generation, Models, Usage, in Proc. of SIGMOD, 2009. pp. 1015-1018.
- Das S., Sismanis Y., Beyer K.S., Gemulla R., Haas P.J., McPherson J., Ricardo: Integrating R and Hadoop, In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD '10), 2010, pp. 987-998.
- Duncan G. 2007, Privacy by design, Science, vol. 317, pp.1178-1179.
- Estape E.A., Mary Helen Mays, Elizabeth A. Sternke. (2016) Translation in Data Mining to Advance Personalized Medicine for Health Equity. Intelligent Information Management 08, 9-16.
- Ghoting A., Pednault E., Hadoop-ML: An infrastructure for the rapid implementation of parallel reusable analytics, In: Proceedinds of the Large-Scale Machine Learning: Parallelism and Massive Datasets Workshop (NIPS-2009).
- Indarte S. y Vero Á. (2014). Sistemas de apoyo a la toma de decisiones clínicas y de gestión en atención primaria de salud. En: Carnicero J., Fernández A. y Rojas D. (coordinadores). Manual de salud electrónica para directivos de servicios y sistemas de salud (II). Aplicaciones de las TIC a la atención primaria de salud. Informes SEIS (10). Comisión Económica para América Latina y el Caribe, Sociedad Española de Informática de la Salud; 2014. 181-202.
- Hitz A., Lea Prevel Katsanis. (2014) A consumer adoption model for personalized medicine: an exploratory study. International Journal of Pharmaceutical and Healthcare Marketing 8, 371-391.
- Hurwitz, J., Nugent, A., Hapler, F. and Kaufman, M., Big Data for Dummies, Hoboken, New Jersey: John Wiley & Sons, 2013.
- Jung KH, Kyung-Han Lee. (2015) Molecular Imaging in the Era of Personalized Medicine. Journal of Pathology and Translational Medicine 49, 5-12
- Keen J., Helen Moore. (2015) The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine. Journal of Personalized Medicine 5, 22-29.
- Kuchinke W., Christian Ohmann, Holger Stenzhorn, Alberto Anguista, Stelios Sfakianakis, Norbert Graf, Jacques Demotes. (2016) Ensuring sustainability of software tools and services by cooperation with a research infrastructure. Personalized Medicine 13, 43-55.
- Labrinidis and Jagadish 2012, A. Labrinidis and H. Jagadish, Challenges and Opportunities with Big Data, In Proc. of the VLDB Endowment, 5(12):2032-2033, 2012.
- Lorch J., B. Parno, J. Mickens, M. Raykova, and J. Schiffman, Shoroud: Ensuring Private Access to Large-Scale Data in the Data Center, In: Proc. of the 11th USENIX Conference on File and Storage Technologies (FAST'13), San Jose, CA, 2013.
- Luo D., Chris Ding, Heng Huang, Parallelization with Multiplicative Algorithms for Big Data Mining, In: Proc. of IEEE 12th International Conference on Data Mining, pp.489-498, 2012.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A.H., Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2011.
- Papadimitriou S., Sun J., Disco: Distributed co-clustering with map-reduce: A case study towards petabyte-scale end-to-end mining. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM '08), 2008, pp. 512-521.
- Raghupathi, W. and Raghupathi, V. Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 2014, 3.
- Rajaraman A. and J. Ullman, Mining of Massive Datasets, Cambridge University Press, 2011.

- Rana A., Sunil Gupta, Dinh Phung, Svetha Venkatesh. (2015) A predictive framework for modeling healthcare data with evolving clinical interventions. Statistical Analysis and Data Mining: The ASA Data Science Journal 8:10.1002/sam.2015.8.issue-3, 162-182.
- Reed C., Thompson D., Majid W., and Wagstaff K. 2011, Real time machine learning to find fast transient radio anomalies: A semi-supervised approach combining detection and RFI excision, Int'l Astronomical Union Sym. on Time Domain Astronomy, UK. Sept. 2011.
- Sariyar M., Irene Schluender, Carol Smee, Stephanie Suhr. (2015) Sharing and Reuse of Sensitive Data and Samples: Supporting Researchers in Identifying Ethical and Legal Requirements. Biopreservation and Biobanking 13, 263-270.
- Schadt E. 2012, The changing privacy landscape in the era of big data, Molecular Systems, 8, Article number 612.
- Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., and Tufano, P., Analytics: The Real-World Use of Big Data: How Innovative Enterprises Extract Value from Uncertain Data, New York: IBM Global Service, 2012.
- Silva A. da, Raja Chiky, Georges Hébrail, A clustering approach for sampling data streams in sensor networks, Knowledge and Information Systems, July 2012, Volume 32, Issue 1, pp 1-23.
- Su K., Huang H., Wu X., and Zhang S., A logical framework for identifying quality knowledge from different data sources, Decision Support Systems, 2006, 42(3): 1673-1683
- Wang Q.; Kui Ren; Wenjing Lou, Privacy-Preserving Public Auditing for Data Storage Security in Could Computing, IEEE Transactions on Computers, 62(2):362-375, 2013.
- Wang, Y., Kung, L., Wang, Y.C., and Cegielski, C. "Developing IT-Enabled Transformation Model: The Case of Big Data in Healthcare", Proceedings of 35th International Conference on Information Systems (ICIS), 2014, Auckland, New Zealand.
- Wu X. and Zhu X. 2008, Mining with Noise Knowledge: Error-Aware Data Mining, IEEE Transactions on Systems, Man and Cybernetics, Part A, vol.38, no.4, pp.917-932.
- Wu X., Yu K., Ding W., Wang H., and Zhu X., Online feature selection with streaming features, IEEE Trans. on Pattern Analysis and Machine Intelligence, 35(5):1178-1192, 2013.
- Ye M., Wu X., Hu X., Hu D., Anonymizing classification data using rough set theory, Knowledge-Based Systems, 43: 82-94, 2013.
- Zikopoulos, P.C., Eaton, C. deRoos, D., Deutsch, T., and Lapis, G., Understanding Big Data: Analytics: Analytics for Enterprise Class Hadoop and Streaming Data. New York: McGraw Hill, 2012.

Los autores

Alberto Andérez González. Licenciado en Derecho por la Universidad de Deusto. Abogado en ejercicio. Letrado de la Administración de la Seguridad Social y Asesor Jurídico del Gobierno de Navarra en excedencia.

Javier Carnicero Giménez de Azcárate. Licenciado en Medicina y Cirugía por la Universidad de Zaragoza. Doctor por la Universidad de Valladolid. Máster en Dirección de Servicios de Salud por la Universidad Pública de Navarra. Jefe del Servicio de Gestión de Prestaciones y Conciertos del Servicio Navarro de Salud. Revisor de Applied Clinical Informatics (ACI), revista electrónica oficial de la Asociación Internacional de Informática Médica (IMIA). Miembro de la Junta Directiva de la Sociedad Española de Informática de la Salud. Coordinador de los Informes SEIS.

Juan Díaz García. Especialista en Medicina Preventiva y Salud Pública. Doctor en Medicina por la Universidad de Granada. Experto en Gestión Sanitaria por la Escuela Andaluza de Salud Pública. Experto en Protección de Datos por la Universidad de Murcia. Cuerpo Superior de Informática de la Junta de Andalucía. Responsable de la Unidad de Gestión de Riesgos Digitales del Servicio Andaluz de Salud. Auditor CISA por ISACA. Miembro de la Junta Directiva de la Sociedad Española de Informática de la Salud. Coordinador del Comité Técnico Asesor de Seguridad de la Información de Salud de la SEIS.

Fernando Escolar Castellón. Doctor en Medicina y Cirugía. Especialista en Medicina Interna y Jefe del Servicio de Medicina Interna del Hospital Reina Sofía de Tudela (Navarra). Secretario de la Sociedad de Medicina Interna de Aragón, Navarra, La Rioja y País Vasco (SOMIVRAN). Autor del modelo de Historia Clínica Informatizada del Gobierno de Navarra.

Pilar León Sanz. Licenciada en Medicina y Cirugía por la Universidad Complutense de Madrid. Doctora por la Universidad de Navarra. Profesora Titular de Historia de la Medicina y Ética Médica en la Facultad de Medicina y miembro del proyecto Cultura Emocional e Identidad en el Instituto de Cultura y Sociedad de la Universidad de Navarra. Research Fellow en el Wellcome Trust Centre for the History of Medicine at UCL (University College London) en 2002 y 2010. Visiting Scholar en el Department of the History of Science, Harvard University (2011). Su investigación se ha orientado al análisis de la profesión y la práctica médica en la España contemporánea.

David Rojas de la Escalera. Ingeniero de telecomunicación (especialidad telemática) por la Universidad de Cantabria. Business Development Senior Consultant en Sistemas Avanzados de Tecnología, S.A. (SATEC). Miembro de la Sociedad Española de Informática de la Salud. Revisor de Applied Clinical Informatics (ACI), revista electrónica oficial de la Asociación Internacional de Informática Médica (IMIA) y de la Asociación de Directores Médicos de Sistemas de Información (AMDIS). Referee del Consejo Editorial de la revista Gestión y Evaluación de Costes Sanitarios de la Fundación Signo.