AUTORES / AUTHORS

Dra. Ernestina Menasalvas
Universidad Politécnica de Madrid, Madrid, España
ernestina.menasalvas@upm.es 0000-0002-5615-6798

Dr. Alejandro Rodríguez-González
Universidad Politécnica de Madrid, Madrid, España
alejandro.rg@upm.es 0000-0001-8801-4762

Dra. María Torrente
Hospital Universitario Puerta de Hierro - Majadahonda, Madrid, España
mtorrente80@gmail.com 0000-0001-8791-7660

Dr. Mariano Provencio
Hospital Universitario Puerta de Hierro - Majadahonda, Madrid, España
mariano.provencio@salud.madrid.org 0000-000-6315-7919

■ Recibido / Received
4 de noviembre de 2019

■ Aceptado / Acepted
10 de noviembre de 2019

■ Páginas / Pages
De la 151 a la 166

■ ISSN: 1885-365X

¿Puede data science ayudarnos a mejorar el pronóstico y tratamiento del paciente oncológico?

Can Data Science help treatment and prognosis of oncological patients?

Resumen

El campo de la informática de la salud está en la cúspide de su período más emocionante hasta la fecha. Las tecnologías de big data, IA y data science están ayudando a tomar decisiones relativas a diagnóstico, tratamiento... La alta implantación de la historia clínica digital es un hecho. El análisis de los datos de la historia clínica permitirá definir nuevas soluciones para todos los integrantes del sistema sanitario. Es necesaria la cobertura de todos los datos posibles para desarrollar nuevos servicios con el objetivo de mejorar el seguimiento y la prevención de enfermedades, y generar valor a partir de ellos. No obstante, el proceso de aplicación de las tecnologías tiene que afrontar todavía retos como el de integración de la información, aplicación de técnicas de lenguaje natural, y elección de las técnicas mas apropiadas dependiendo del problema y de la naturaleza de los datos. En este artículo nos planteamos los retos que tiene la aplicación de estas técnicas en el caso particular del paciente oncológico.

PALABRAS CLAVE: educación, competencias, análisis de datos, evaluación, alumnado universitario, emprendimiento.

Abstract

This work wants to specify preliminary data of the design process of an instrument adapted to a Spanish population based on different questionnaires to evaluate the attributes of entrepreneurial skills of university students, and to contribute a valid and reliable measure that serves as a reference for effective intervention programs in the university environment, and for the development of employability. The instrument provides students with the possibility of discovering their strengths and opportunities related to the sub-competences evaluated: the identification of opportunities, the development of innovative solutions, the ability to learn from failure, and their awareness of their entrepreneurship. An initial content validity study was carried out through the trial of 13 experts, all of them university professors expert of the subject, which

determined the development of the questionnaire that was subsequently tested on a pilot sample of 350 students. It concludes to the suitability and usefulness of the instrument, and discusses the importance of the intervention for the development of entrepreneurial competence in the University.

KEY WORDS: education, skills, data analysis, evaluation, university students, entrepreneurship.

1. Big data en el ámbito sanitario

1.1. Big data. Definición y características

El campo de la informática de la salud está en la cúspide de su período más emocionante hasta la fecha, entrando en una nueva era donde la tecnología está empezando a manejar grandes volúmenes de datos, dando lugar a un potencial ilimitado para el crecimiento de la información. La minería de datos y análisis masivo de datos están ayudando a tomar decisiones relativas a diagnóstico, tratamiento... Y todo finalmente enfocado a una mejor atención al paciente.

El uso de la minería de datos en salud en Estados Unidos puede ahorrar a la industria de la salud hasta 450 mil millones de dólares cada año (Kayyali, Knot y Van Kuiken, 2013). Esto se debe a los volúmenes crecientes de datos generados y de las tecnologías para analizarlos.

El crecimiento explosivo de datos generó, ya en la década de los 80, la aparición de un nuevo campo de investigación que se denominó KDD o *Knowledge Discovery in Databases*. Bajo estas siglas se esconde el proceso de descubrimiento de conocimiento en grandes volúmenes de datos (Fayyad, Piatetsky-Shapiro y Smith, 1996). El proceso de KDD ha servido para unir a investigadores de áreas como la inteligencia artificial, estadística, técnicas de visualización, aprendizaje automático o bases de datos en la búsqueda de técnicas eficientes y eficaces que ayuden a encontrar el potencial conocimiento que se encuentra inmerso en los grandes volúmenes de datos almacenados por las organizaciones diariamente.

Si bien el nombre con el que apareció esta área de investigación fue el de KDD, más adelante se sustituyó por términos como data mining, data analytics, business intelligence y hoy en día Inteligencia Artificial. Si bien es verdad que el énfasis de estos términos es diferente, en lo que están todos de acuerdo es en la extracción de conocimiento de los datos.

Aunque no hay una única definición de data mining, la siguiente es, posiblemente, la más aceptada: «proceso de extracción de información desconocida con anterioridad, válida y potencialmente útil de grandes bases de datos para usarla con posterioridad para tomar decisiones importantes de negocio» (Witten, Frank y Hall, 2011).

El término proceso implica que la extracción de conocimiento es la conjunción de muchos pasos repetidos en múltiples iteraciones. Se dice, por otra parte, que es no trivial, porque se supone que hay que realizar algún tipo de proceso complejo. Los patrones deben ser válidos, con algún grado de certidumbre, y novedosos, por lo menos para el sistema y, preferiblemente, para el usuario, al que deberán aportar alguna clase de beneficio (útil). Por

último, está claro que los patrones deben ser comprensibles, si no de manera inmediata, sí después de ser pre-procesados.

Por su parte, el término Inteligencia Artificial (AI) ha sido definido (BDVA, EU Robotics, 2019) como un término global que cubre la inteligencia tanto digital como física, datos y robótica, y tecnologías inteligentes relacionadas.

Los problemas que se pueden abordar desde la perspectiva de data mining a menudo se agrupan en las siguientes categorías:

- Problemas cuyo objetivo es predecir el valor de un atributo en particular basado en los valores de otros atributos. El atributo que se predice se denomina comúnmente atributo objetivo (o variable dependiente), mientras que los atributos que se utilizan para la predicción son conocidos como atributos explicativos (o variables independientes). Destacan aquí los problemas de clasificación o de estimación de valor y como técnicas podemos destacar los enfoques basados en estadística, regresión, árboles de decisión y redes neuronales.
- Problemas descriptivos cuyo objetivo es derivar patrones (correlaciones, tendencias, agrupaciones o clústeres, trayectorias y anomalías) que resuman las características inherentes a los datos. Dentro de este grupo, cabe destacar el análisis de reglas de asociación para el que el algoritmo "A priori" (Agrawal y Srikant, 1994) es el más conocido, así como los problemas de segmentación o clustering.

Las nuevas características de las tecnologías de la información y las comunicaciones han provocado la aparición de multitud de aplicaciones donde se generan, computan y almacenan data streams (Aguilar-Ruiz y Gama, 2005; Gaber, Krishnaswamy y Zaslavsky, 2005). Estos datos tienen características concretas: flujos de datos continuos en el tiempo, sin límites de tamaño, que aparecen a gran velocidad y cuya distribución evoluciona a lo largo del tiempo. Existen múltiples aplicaciones y ejemplos que generan datos de estas características en el entorno de la salud y en otros entornos: monitores de la UCI, redes de sensores, monitorización de sensores ambientales.

Para diseñar algoritmos eficientes que se adecuen de manera eficaz es necesario establecer qué características identifican a los *data streams*. En concreto, en Aguilar-Ruiz y Gama (2005) y Domingos y Hulten (2000) se identifican las siguientes:

- Cantidad de datos ilimitados.
- Alta velocidad de llegada de datos.
- Búsqueda de modelos a lo largo de un gran período de tiempo.
- El modelo subyacente cambia a lo largo del tiempo (dicho efecto se conoce
- como "evolución del modelo").

Las características propias de los *data streams* provoca que el enfoque clásico utilizado para el análisis de datos no sea aplicable porque la naturaleza de aparición y características en el análisis difiere en ambos casos. De manera general, los algoritmos clásicos de data mining no son capaces de analizar los datos de esta naturaleza puesto que asumen que todos

los datos se encuentran cargados en una base de datos estable y raramente actualizada. Es también importante destacar que el proceso de análisis puede llevar días, semanas o incluso meses, después del cual los resultados son estudiados y, en caso de no ser satisfactorios, dicho análisis se reproduce modificando alguno de las características utilizadas (Domingos y Hulten, 2000).

En el caso de los algoritmos para data streams deben hacer uso limitado de memoria (e incluso de un tamaño fijo) (Aggarwal, Han, Wang y YU, 2003). Además, el hecho de no poder revisar elementos que han aparecido en el pasado produce que estos algoritmos deban ser capaces de generar modelos de una única pasada.

Es importante destacar en este punto que el desarrollo de tecnologías en los últimos 20 años permite contar hoy en día con numerosas soluciones para aplicar dependiendo del tipo de datos, ya sean éstos de índole estática o dinámica (*streams* de datos). El reto, no obstante, radica en entender los problemas y entender cómo integrar, procesar y limpiar los datos, y en aquellos casos en que los datos no están estructurados, como el caso de los textos, estructurarlos.

Como consecuencia de la complejidad del desarrollo de proyectos de minería de datos, a comienzos de los 90 surge el estándar de modelo de proceso denominado CRISP-DM (Wirth, 2000) que divide el proceso en las siguientes fases:

- Comprensión del negocio: se pretende aquí comprender los objetivos del proyecto y sus requerimientos desde la perspectiva del negocio, convirtiendo este conocimiento en un problema de data mining y un plan preliminar para cumplir dichos objetivos.
- Comprensión de los datos: se cuenta en un principio con una colección de datos, se deben identificar los problemas de calidad de los datos, detectar subconjuntos de interés, etc.
- Preparación de los datos: mediante esta fase se construye el conjunto de datos final obtenido de la colección inicial de datos que será proporcionada a las herramientas de modelado.
- Modelado: se seleccionan y aplican varias técnicas de modelado, ajustándolas para obtener valores óptimos.
- Evaluación: una vez construido un modelo se debe evaluar y revisar los pasos ejecutados para construir un modelo que consiga los objetivos de negocio.
- Despliegue: aplicación de los modelos validados para la toma de decisión como parte de algún proceso en la organización.

1.2. Digitalización de los datos en los sistemas sanitarios

Hoy en día la historia clínica digital es un hecho en España. La tecnología ha hecho evolucionar el comportamiento global de usuarios en todos los dominios, investigadores, proveedores de soluciones de tecnologías de la información (TI), profesionales de la salud entre otros. Todos los individuos independientemente de su edad pertenecen a la era digital.

Este nuevo escenario se tiene que tener en cuenta, no sólo para el desafío de almacenamiento de datos y el análisis en el marco del nuevo esquema, sino también para

la construcción de los servicios que se exigen en este período de tiempo siendo el objetivo principal que se persigue, analizar los datos para extraer información que se pueda usar en beneficio de la sociedad. Es importante en primer lugar analizar las fuentes de datos y en este sentido distinguimos:

- Fuentes tradicionales: registros de salud, demografía, etc.
- Fuentes no tradicionales: sensores, redes sociales, genómica, literatura, etc.

El análisis de estos datos permitirá definir nuevas soluciones para todos los integrantes del sistema sanitario: personal sanitario, pacientes, farmacia, investigación, epidemiología, compañías de seguros, administración, servicios de atención. Es necesaria la cobertura de todos los datos posibles para desarrollar nuevos servicios con el objetivo de mejorar el seguimiento y la prevención de enfermedades, y generar valor a partir de ellos.

Gracias a estas soluciones los profesionales de la salud podrán aprender de las recomendaciones que las plataformas inteligentes pueden darles y podrán obtener una mayor capacidad de conocimiento en un tiempo más limitado.

1.3. Retos

Los retos a los que nos enfrentamos son abrumadores, y se pueden agrupar en tecnológicos y no tecnológicos. El primer obstáculo que se nos interpone en el camino es un desafío tecnológico. Gran parte de los datos disponibles del mundo real no están estructurados y están almacenados en miles de hospitales desconectados. La segunda gran barrera es el cambio cultural que se requiere para compartir la información más allá del propio centro, más allá de una región o un país.

Entre los retos tecnológicos a los que nos enfrentamos destacamos:

- El volumen de datos: la genómica, la monitorización (UCI, dispositivos móviles), la ubicuidad, datos sociales. Se requerirán, por una parte, nuevos métodos para el almacenamiento de datos; por otra parte, estos datos requieren nuevas aplicaciones para su integración, consulta y análisis.
- Almacenamiento físico de los datos: los datos requieren de nuevos medios y arquitecturas para su almacenamiento y tratamiento de forma eficiente.
- Problemas de interoperabilidad: diversos hospitales tienen diferentes sistemas de
- almacenamiento. Tiene que haber una capa de interoperabilidad para construir sobre las soluciones de tecnologías de la información.
- Limpieza de datos, integración, análisis, herramientas: cuando se tenga acceso a información de todo tipo: los registros de salud, información de contexto, la genómica, y el resto de datos, serán necesarias nuevas herramientas y servicios para diferenciar el ruido de los datos valiosos.
- Interpretabilidad de los modelos obtenidos con técnicas de inteligencia artificial.
- Impacto de los cambios en los protocolos de registro de datos y en la normativa sobre los datos registrados.

- Acceso a plataformas de open data

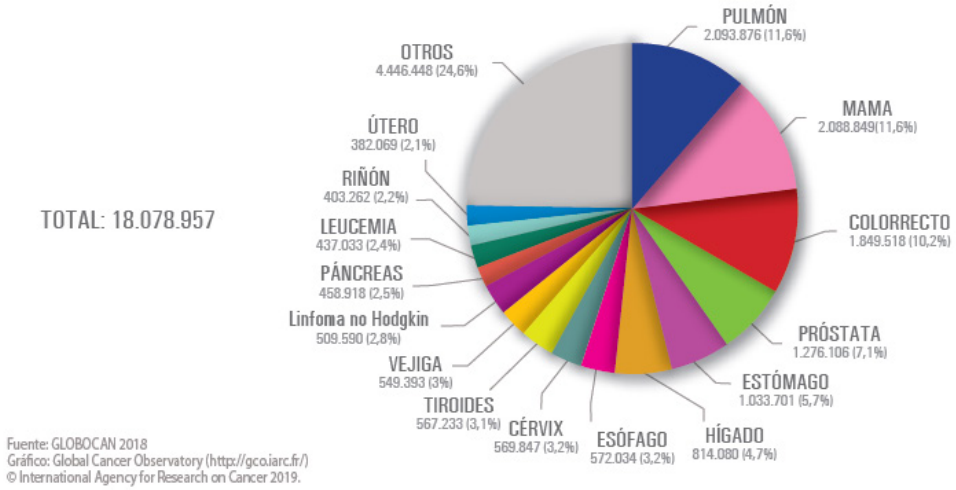
Pero también nos enfrentamos a retos no tecnológicos entre los que destacan las cuestiones legales y de privacidad. Es fundamental tener en cuenta las consecuencias de la entrada en vigor del *Reglamento General de Protección de Datos* en mayo de 2018.

Por su parte los datos se recogen en el lenguaje de cada país/región y consiguientemente nos encontramos con problemas derivados del multilingüismo pues independientemente de cómo se recojan los datos en general, la literatura médica está disponible en inglés. Si bien todos los datos en la historia clínica utilizada en España están en español y, en algunos casos, en las lenguas co-oficiales, la mayoría de las herramientas que se han desarrollado son para inglés.

2. Cáncer y problemática del cáncer

El cáncer sigue siendo uno de los principales problemas de salud pública en todo el mundo. Sigue constituyendo una de las principales causas de morbi-mortalidad del mundo. De acuerdo con los últimos datos disponibles estimados dentro del proyecto GLOBOCAN, el número de tumores continúa creciendo, habiendo aumentado desde los 14 millones de casos en el mundo estimados en el año 2012 a los 18,1 millones en 2018 (Siegel, Miller, Jemal, 2018). Las estimaciones poblacionales indican que el número de casos nuevos aumentará en las dos próximas décadas, alcanzando los 29,5 millones en 2040. Pese a las elevadas cifras de mortalidad, la supervivencia aumenta de forma continua y en España es similar a la del resto de países de nuestro entorno, siendo del 53% a los cinco años.

Figura 1. Tumores más frecuentemente diagnosticados en el mundo. Estimación para 2018 para ambos sexos



A pesar de que Europa comprende sólo una octava parte de la población mundial total, tiene alrededor de una cuarta parte del total de casos de cáncer en el mundo: unos 3,2 millones de pacientes nuevos por año. Los tipos de cáncer más prevalentes en Europa son el cáncer de mama (464.000 casos), seguidos de colorrectal (447.000), próstata (417.000) y pulmón (410.000). Estos cuatro representan la mitad de la carga general del cáncer en Europa. Las causas más comunes de muerte por cáncer son el cáncer de pulmón (353.000 muertes), colorrectal (215.000), de mama (131.000) y de estómago (107.000) (*NSCLC Meta-analysis Collaborative Group*, 2014).

El cáncer es, también en España, una de las principales causas de morbilidad. El número de tumores diagnosticados en España en el año 2019 alcanzará los 277.234, según las estimaciones de REDECAN, en comparación con los 247.771 casos diagnosticados en 2015: 148.827 en varones y 98.944 en mujeres. Los cánceres más frecuentes diagnosticados en España en 2019 serán los de colon y recto (44.937 nuevos casos), próstata (34.394), mama (32.536), pulmón (29.503) y vejiga urinaria (23.819). A mucha distancia, los siguientes cánceres más frecuentes serán los linfomas no Hodgkin, y los cánceres de cavidad oral y faringe, páncreas y estómago.

Es importante destacar que el cáncer de pulmón pasará de ser el cuarto tumor más diagnosticado en mujeres en las estimaciones para el año 2015 al tercero con mayor incidencia para el año 2019, en probable relación con el aumento del consumo de tabaco en mujeres (más tardío). Así, la incidencia de cáncer de pulmón en mujeres ha continuado aumentando, mientras que la tasa de incidencia en varones continúa reduciéndose. Pese a todo, el consumo de tabaco continúa siendo más frecuente en hombres que en mujeres de acuerdo con los datos de Eurostat para el año 2014: 26,2% de fumadores entre los varones frente a un 18,5% de las mujeres. El cáncer de pulmón fue el tumor responsable del mayor número de muertes, 22.896 casos en 2018 (reducción del 0,3% con respecto al año anterior),

y el cáncer colorrectal fue el segundo (reducción del 2,4%) (Sculier, Berghmans y Meert, 2014).

2.1. Problemática actual

Los enfoques actuales para el diagnóstico, tratamiento y seguimiento del cáncer requieren mejoras y avances urgentes. La medicina personalizada, entendida como el desarrollo unitario e independiente de marcadores moleculares más o menos específicos, no cambiará la situación de la enfermedad si no se abordan otros aspectos específicos de la práctica clínica habitual.

En oncología, la evaluación de la respuesta al tratamiento y las predicciones precisas de supervivencia son factores clave para el control efectivo de la enfermedad, así como el diseño de nuevos esquemas de tratamiento y el desarrollo de futuros tratamientos. Ninguno de los métodos de diagnóstico actuales o factores de pronóstico clínico ha identificado completamente la presencia de micrometástasis, que en estadíos iniciales contribuyen a la recaída después de cirugías agresivas, o no identifican la resistencia a los tratamientos precozmente para evitar toxicidades, debido a la heterogeneidad intrínseca de cada tumor y cada paciente en particular (Chansky et al., 2009). El nuevo enfoque para combatir estas enfermedades pasa por añadir información genómica en el momento del diagnóstico, que permita predecir cómo de bien un paciente responderá a los tratamientos. Con la ayuda de sofisticadas pruebas diagnósticas como NGS (*Next Generation Sequencing*), se pueden detectar estos defectos genéticos y podemos aplicar tratamientos dirigidos.

Sin embargo, a pesar de vivir en la era del gran desarrollo de la genómica, los fundamentos clínicos permanecen anclados en un estricto seguimiento y control, basados en el tamaño y la morfología del tumor. Ninguno de los descubrimientos genómicos se ha aplicado en este campo con resultados concluyentes. En esta línea de pensamiento no podemos evitar preguntarnos, ¿qué diferencia implicarían estos avances si se aplicaran en las partes más traslacionales de la oncología y pudiéramos cambiar los paradigmas establecidos hace 50 años?

La situación ideal para desarrollar la prueba de concepto sería el estudio de un tumor suficientemente prevalente, para el cual tenemos la histología en el diagnóstico, un tratamiento homogéneo y una evaluación estandarizada y, desafortunadamente, un mal pronóstico que permitiría una verificación rápida de los indicadores pronósticos. Todos estos factores se pueden encontrar, por ejemplo, en pacientes con cáncer de pulmón de célula no pequeña (NSCLC).

- Diagnóstico: dentro del mismo estadío de la enfermedad, existe una gran variabilidad entre pacientes; en el cáncer de pulmón, por ejemplo, mientras que los pacientes en estadíos iniciales son potencialmente curables, solo el 55% está vivo después de 5 años (Massuti et al., 2012). Como los pacientes fallecen principalmente por metástasis, esto significa que el 45% de los pacientes tenían enfermedad micrometastásica oculta en el momento del diagnóstico, y que no era detectable mediante las pruebas convencionales. Esta situación es similar en otros estadíos y

en todo el campo de la histología. Como resultado, la gran mayoría de los pacientes se someten a tratamientos agresivos y sólo el tiempo dirá si el riesgo-beneficio fue positivo o no (¿los tratamientos fueron apropiados, demasiado tóxicos o demasiado caros?).

- Seguimiento de la enfermedad: el seguimiento actual de la progresión y la remisión de la enfermedad se basa en biopsias, imágenes radiológicas y patología. Estas pruebas se realizan algún tiempo después de los tratamientos y representan sólo una instantánea; actualmente no existe ningún método efectivo de evaluación de la enfermedad en tiempo real. Además, las pruebas basadas en biopsias no tienen en cuenta la variación intratumoral, por lo que la progresión de todo el tumor no se puede evaluar con precisión a partir de un pequeño número de biopsias que representan una pequeña parte del tumor.
- Terapia y respuesta: la situación es similar cuando consideramos los tratamientos ya que la duración óptima de la quimioterapia no está establecida en la mayoría de los casos. Un caso frecuente es que un paciente puede recibir tratamiento de quimioterapia, pero no estamos seguros de cuántos ciclos se requieren si la reevaluación por imagen muestra enfermedad estable. Esto lleva a que se mantengan muchos tratamientos hasta la progresión de la enfermedad o hasta una toxicidad intolerable. Tratamos a muchos pacientes que no obtienen ningún beneficio real de los tratamientos; otros pacientes reciben tratamiento mientras desarrollan resistencia al mismo, y no tenemos ninguna herramienta para determinar esas situaciones de manera precoz y precisa. La importancia de esta ambigüedad no sólo es clínica, sino también económica: la relación coste-beneficio es mayor debido a la variabilidad en la eficacia de los tratamientos.
- Efectos adversos: más del 5% de todos los ingresos hospitalarios se deben a efectos adversos de los tratamientos (Provencio et al., 2012). El número de muertes en todo el mundo debido a esto es de aproximadamente 100.,000 pacientes por año, amenudo debido a medicamentos ineficientes o innecesarios, como se demuestra en las curvas de supervivencia de todos los estudios. A pesar de que el coste farmacéutico es mayor cada año, el coste de las herramientas genómicas para la secuenciación masiva ha disminuido drásticamente: en 2007 costó $ 10 millones secuenciar el genoma humano, hoy cuesta menos de $ 1000.

La gestión subóptima del paciente con cáncer durante el tratamiento y en el seguimiento posterior impide conseguir mejores resultados para el paciente, especialmente en lo relativo a la calidad de vida, y es la responsable de una gran parte de los costes evitables generados. Y es que el cáncer supone una carga económica sustancial a la sociedad. Los mayores costes sanitarios están asociados con su prevención y gestión. A esto hay que sumar el componente psico-social, ya que después del tratamiento, muchos pacientes no pueden seguir trabajando, y muchos dependen de amigos y familiares para recibir apoyo durante el tratamiento o en las últimas fases de la enfermedad. Es por ello que la cuantificación de la carga económica del cáncer en la UE necesita no sólo una estimación del coste del cáncer para los sistemas de salud, sino también una estimación de los ingresos perdidos asociados con la incapacidad para trabajar, bajas laborales y costes sociales relacionados (Provencio,

Isla, Sánchez y Cantos, 2011).

En el caso del envejecimiento y la calidad de vida, la evidencia muestra que la incidencia de la mayoría de los tipos de cáncer depende de la edad. El envejecimiento progresivo de la sociedad está aumentando el número de personas mayores que necesitan tratamiento oncológico (Earle et al., 2003). Los pacientes de edad avanzada presentan características particulares que dificultan la elección de un tratamiento adecuado, también debido a limitaciones metodológicas (mayor frecuencia de analfabetismo, peor cumplimiento de los cuestionarios, enfermedades concomitantes, uso de metodología no validada en población anciana). Además, los pacientes de edad avanzada están poco representados en los ensayos clínicos, lo que dificulta aún más su tratamiento. Recientemente, la calidad de vida relacionada con la salud (QoL) comenzó a considerarse como uno de los puntos más cruciales, pero más complejos en la investigación clínica del anciano con cáncer (Garrido et al., 2007).

Por último, es necesario abordar la problemática existente debida a los costes generados por el uso deficiente de recursos en la etapa final de la vida del paciente oncológico. Hay tres conceptos principales sobre la mala calidad de la atención del cáncer al final de la vida que deben examinarse utilizando los datos disponibles actualmente: el uso de nuevas terapias oncológicas o la continuación de los tratamientos en el último mes de vida; un elevado número de visitas a urgencias hospitalarias, ingresos hospitalarios frecuentes en los últimos meses de vida, ingreso en unidades de cuidados intensivos los últimos días de vida; y una alta proporción de pacientes que no fueron trasladados a tiempo a hospitales de media-larga estancia de cuidados paliativos, solo ingresaron en los últimos días de vida o murieron en un entorno de cuidados agudos (Iadeluca, Mardekian, Chander, Hopps y Makinson, 2017). Conceptos como el acceso a servicios psicosociales y otros servicios multidisciplinares y el control del dolor y los síntomas son importantes y pueden ser posibles, pero actualmente no se pueden aplicar en la mayoría de los sistemas de datos. Los indicadores basados en la limitación del uso de tratamientos con baja probabilidad de beneficio o los indicadores basados en la eficiencia económica no son aceptados por pacientes, familiares o médicos. Se han identificado varios indicadores de calidad prometedores que, si se consideran válidos y fiables dentro de los sistemas de datos, podrían ser útiles para identificar los sistemas sanitarios con necesidades de mejora de cuidado e identificación de pacientes en la última etapa de la enfermedad (Leach et al., 2015).

2.2. El papel de *BigData analytics* para abordar la gestión del paciente con cáncer

2.2.1. Por qué se necesita un enfoque con big data aplicado a la práctica clínica

Gracias a las técnicas de big data integraremos múltiples fuentes de datos de forma que podremos analizar datos no estructurados, notas clínicas y literatura; extraer patrones para predecir toxicidad, o efectos adversos e interacciones de medicamentos a gran escala. Todo esto servirá como herramienta de apoyo en el manejo de los pacientes, para intentar reducir

sobretratamientos, bajas por enfermedad, visitas no programadas a consulta, número de visitas a Urgencias, tiempo dedicado a buscar casos relacionados en la bibliografía o eventos adversos debido a comorbilidades.

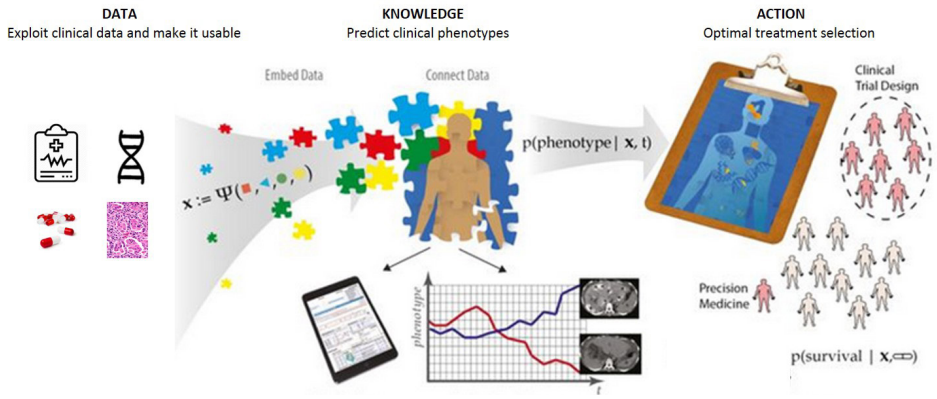
Si bien big data no es la solución única a toda esta problemática, distintos problemas en todo el proceso asistencial del cáncer requieren de las capacidades de *BigData analytics* para ser abordados de una forma efectiva.

- Prevención: identificación de factores relacionados con el cáncer, tanto ocupacional, como geográfico, dietético o familiar.
- Detección: screening no invasivo, sin efectos secundarios, eficiente y con alta sensibilidad/especificidad capaz de identificar pacientes sospechosos de estar desarrollando la enfermedad. Detectar poblaciones especialmente sensibles aún no identificadas.
- Diagnóstico: determinación de una medida de la gravedad de la enfermedad (estadiaje) que correlacione bien con resultados esperados del tratamiento y el estado de salud final del paciente.
- Tratamiento:
    - Predicción de resultados del tratamiento: respuesta, progresión, recaída, supervivencia.
    - Toxicidad y detección temprana de estos eventos.
    - Medicina de precisión: soporte en la toma de decisiones para la elección del mejor.
    - Tratamiento en cada momento para cada paciente.
    - Identificación y actuación proactiva sobre pacientes en alto riesgo de consumo de recursos asistenciales con el fin de reducir procesos hospitalarios innecesarios (ej. visitas no programadas, urgencias hospitalarias, estancia media).
    - Desarrollo de nuevas terapias más efectivas y con menos efectos secundarios en el mundo real (no sólo dentro de ensayo clínico).
- Seguimiento:
    - Detección temprana de recaída, de toxicidad tardía y complicaciones.
    - Determinación del seguimiento clínico óptimo, tanto en resultados en salud como en comodidad y seguridad para el paciente como en eficiencia en costes.

La gran cantidad de datos derivados de las notas clínicas (el Hospital Universitario Puerta de Hierro de Madrid genera alrededor de 2.5 millones de notas clínicas / año), las pruebas y sus resultados, junto con las imágenes digitales y todos los demás datos demográficos, pueden contener información y conocimiento que no es aparente sin asistencia de sistemas informáticos. Actualmente, los sistemas de servicios de salud contienen petabytes de datos de pacientes que, sin duda, contienen información valiosa que podría servir tanto para mejorar

el diagnóstico y el tratamiento de casos específicos, como para llevar a cabo investigaciones que permitan trasladar el conocimiento generado a la práctica clínica. Independientemente de que los datos estén estructurados o no, las técnicas de procesamiento de datos masivos, así como la minería de datos y el aprendizaje automático, permiten descubrir asociaciones entre los valores de variables (problemas de asociación), grupos de sujetos que se comportan de manera similar (segmentación o agrupación) o incluso con base en datos históricos, y son capaces de construir un modelo que permita predicciones o estimaciones del futuro (clasificación). Las técnicas y tecnologías de análisis de big data, como la gestión de bases de datos de alto rendimiento, estadística, minería de datos y aprendizaje automático, incluida la agrupación (K-means, algoritmos jerárquicos y estadísticos y mapas de Kohonen) y algoritmos de asociación (algoritmos a priori y sus variantes), clasificación, regresión y técnicas de aprendizaje supervisado, así como métodos de aprendizaje por refuerzo (RLM), redes neuronales artificiales (ANN) y máquinas de vectores de soporte (SVM), entre otras. Si bien estas técnicas tienen un uso muy extendido en algunos sectores industriales, se han aplicado en un grado mucho menor en el sector de la salud por muchas razones (Guergana, et al., 2019), destacando entre ellas la falta de informatización en los procesos de atención médica y el hecho de que la mayoría de los datos clínicos no están estructurados (es texto libre) y difícil de explotar.

Figura 2. De los datos al conocimiento y a la intervención. La medicina de precisión integra muchas fuentes de datos utilizando algoritmos de aprendizaje automático para ayudar a los médicos a predecir lo que le sucederá al paciente y decidir el mejor tratamiento.



Las nuevas tecnologías transformadoras, tanto de laboratorio como computacionales, permiten recopilar datos clínicos y biológicos de alta resolución, extraer información significativa y luego ofrecer a los pacientes un tipo de asesoramiento médico personalizado. Y eso significa que los pacientes con cáncer pueden esperar un diagnóstico más precoz y, por lo tanto, mejores resultados gracias a los avances en el perfil genómico y molecular y en el modelado computacional y la IA.

2.2.2. Oportunidades para la aplicación de Inteligencia Artificial (IA)

A pesar de las dificultades metodológicas, los algoritmos de IA ya están comenzando a tener un efecto en la investigación del cáncer y la atención clínica, como el diagnóstico y la prevención tempranos, el descubrimiento de fármacos, la inclusión adecuada de pacientes en los ensayos clínicos y las decisiones de tratamiento.

Muchos desarrolladores de medicamentos ya están incorporando de manera rutinaria este tipo de modelos de patología computacional en sus programas de investigación clínica, confiando en el aprendizaje automático para, por ejemplo, cuantificar los niveles de biomarcadores que pueden ayudar a explicar por qué sólo algunos pacientes responden a un determinado tratamiento.

Por ejemplo, las técnicas computacionales avanzadas ya están ayudando a estratificar a los pacientes tratados con inmunoterapia. En una línea similar investigadores de la *Universidad Johns Hopkins*, están aplicando técnicas de *Deep Learning* en la llamada sinapsis inmune, para predecir mejor las respuestas a los medicamentos y ayudar en la toma de decisiones terapéuticas (Polymeri et al., 2019).

Y es que el valor que aportan este tipo de tecnologías va mucho más allá del propio conocimiento, permitiendo crear nuevos modelos asistenciales basados en redes colaborativas y haciendo más eficiente todo el proceso desde la sospecha diagnóstica hasta el tratamiento y optimizando los recursos del sistema sanitario (Wiljer y Hakim, 2019). Con las técnicas de big data tenemos la oportunidad de aprender de la experiencia de cada paciente. En definitiva, permiten minimizar la variabilidad de la práctica clínica e incrementar la calidad asistencial. Tanto las técnicas de big data como de IA representan una oportunidad para romper barreras y habilitar nuevos modelos de atención colaborativos e inteligentes centrados en el paciente, y una pieza clave para acelerar la lucha contra el cáncer.

3. Conclusiones

La disponibilidad de grandes volúmenes de datos ya es una realidad en la atención sanitaria y representa una oportunidad para que los profesionales sanitarios mejoren la atención del cáncer. Sin embargo, los problemas técnicos y socioculturales limitan su uso en la práctica. El desafío es encontrar una manera de procesar todas las variables que proporcionen respuestas simples y útiles. Otro desafío es comprender si una herramienta informática puede adaptarse a las diferencias geográficas, la disponibilidad de medicamentos, etc. Es importante destacar que se necesita un esfuerzo conjunto de todas las partes interesadas (profesionales de la salud, programadores, proveedores de IA, etc.) para discutir y acordar sus ideas, actitudes y objetivos para aplicar la informática en oncología. Esto es vital para garantizar el compromiso mutuo con el desarrollo e integración de herramientas clínicamente útiles y lograr los mejores resultados para los pacientes.

4. Bibliografía

AGGARWAL, Charu C., HAN, Jiawei, WANG, Jianyong y YU, Philip S. (2003). A framework for clustering evolving data streams. En: *Proceedings of the 29th International Conference on Very Large Data Bases* Berlin, Germany: VLDB Endowment, p. 81–92. Recuperado de http://dl.acm.org/citation.cfm?id=1315451.1315460

AGRAWAL, Rakesh y SRIKANT, Ramakrishnan (1994). Fast algorithms for mining association rules in large databases. En: *Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., p. 487–499. (VLDB'94). Recuperado de http://dl.acm.org/citation.cfm?id=645920.672836

AGUILAR-RUIZ, Jesús S. y GAMA, João (2005). Data Streams. *Journal of universal computer science*, 11(8),1349–1352. Recuperado de http://www.jucs.org/jucs_11_8/data_streams/abstract.html

BDVA, EU ROBOTICS (2019). *Strategic research, innovation and deployment agenda for an AI PPP*. Consultation Release.

CHANSKY, Kary, SCULIER, Jean-Paul, CROWLEY, John, J., GIROUX, Dori, VAN MEERBEECK, Jan y GOLDSTRAW, Peter (2009). The International Association for the Study of Lung Cancer Staging Project: Prognostic factors and pathologic TNM stage in surgically managed non-small cell lung cancer. *Journal of thoracic oncology*, 4(7), 792-801. doi: 10.1097/JTO.0b013e3181a7716e

DOMINGOS, Pedro y HULTEN, Geoff (2000). Mining high-speed data Streams. En: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, p. 71–80. doi: 10.1145/347090.347107

EARLE, Craig C., PARK, Elyse R., LAI, Bonnie, WEEKS, Jane C., AYANIAN, John Z., BLOCK, Susan (2003). Identifying potential indicators of the quality of end-of-life cancer care from administrative data. *Journal of clinical oncology*, 21(6), 1133-1138. doi: 10.1200/JCO.2003.03.059

FAYYAD, Usama M., PIATETSKY-SHAPIRO, Gregory y SMYTH, Padhraic (1996). From data mining to knowledge discovery: an overview. En Usama M. Fayyad, Gregory Piatetsky-Shapiro y Padhraic Smyth (Eds). *Advances in knowledge discovery and data mining*. Menlo Park, CA, USA: American Association for Artificial Intelligence. Recuperado de http://dl.acm.org/citation.cfm?id=257938.257942

GABER, Mohamed M., KRISHNASWAMY, Shonali y ZASLAVSKY, Arkady (2005). On-board mining of data streams in sensor networks. En *Advanced methods for knowledge discovery from complex data*. London: Springer. doi: 10.1007/1-84628-284-5_12

GARRIDO, Pilar, GONZÁLEZ-LARRIBA, José Luis, INSA, Amelia, PROVENCIO, Mariano, TORRES, Antonio; ISLA, Dolores, SÁNCHEZ, José Miguel, CARDENAL, Felipe, DOMINE, Manuel,

BARCELO, José Ramón, TARRAZONA, Vicente, VARELA, Andrés, AGUILO, Rafael, ASTUDILLO, Julio, MUGURUZA, Ignacio, ARTAL, Ángel, HERNANDO-TRANCHO, Florentino, MASSUTI, Bartomeu, SÁNCHEZ-RONCO, María, ROSELL, Rafael (2007). Long-term survival associated with complete resection after induction chemotherapy in stage IIIA (N2) and IIIB (T4N0-1) Non small-cell lung cancer patients: The Spanish lung cancer group trial 9901. *Journal of clinical oncology*, 25(30), 4736-42.

IADELUCA, Laura, MARDEKIAN, Jack, CHANDER, Pratibha, HOPPS, Markay, MAKINSON, Geoffrey T. (2018). The burden of selected cancers in the US: health behaviors and health care resource utilization. *Cancer management and research*, 9:721-730. doi: 10.2147/CMAR.S143148

KAYYALI, Basel, KNOT, David, VAN KUIKEN, Steve (2013). The big-data revolution in US health care: Accelerating value and innovation. En: *McKinsey Company. Healthcare systems & services*. McKinsey Company. Recuperado de http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care

LEACH, Corinne R., WEAVER, Kathryn E., AZIZ, Noreen M., ALFANO, Catherine M., BELLIZZI, Keith M., KENT, Erin E., FORSYTHE, Laura P., ROWLAND, Julia H. (2015). The complex health profile of long-term cancer survivors: Prevalence and predictors of comorbid conditions. *Journal of cancer survivorship*, 9(2), 239-51. doi:10.1007/s11764-014-0403-1

MASSUTI, Bartomeu, COBO, Manuel, CAMPS, Carlos, DÓMINE, Manuel, PROVENCIO, Mariano, ALBEROLA, Vicente, VIÑOLAS, Nuria, ROSELL, Rafael, TARÓN, Miguel, GUTIÉRREZ-CADERÓN, Vanesa, LARDELLI, Pilar, ALFARO, Vicente; NIETO, Antonio e ISLA, Dolores (2012). Trabectedin in patients with advanced non-small-cell lung cancer (NSCLC) with XPG and/or ERCC1 overexpression and BRCA1 underexpression and pretreated with platinum. Lung Cancer, 76(3), 354-61. doi: 10.1016/j.lungcan.2011.12.002

NSCLC Meta-analysis Collaborative Group (2014). Preoperative chemotherapy for non-small-cell lung cancer: a systematic review and meta-analysis of individual participant data. *Lancet*, 383(9928), 1561-71. doi: 10.1016/S0140-6736(13)62159-5

POLYMERI, Eirini, SADIK, May, KABOTEH, Reza, BORRELLI, Pablo, ENQVIST, Olof, ULÉN, Johannes, OHLSSON, Mattias, TRÄGARDH, Elin, POULSEN, Mads H., SIMONSEN, Jane A., FLEMMING HOILUND-CARLSEN, Poul, JOHNSSON, Ase A., EDENBRANDT, Lars (2019). Deep learning-based quantification of PET/CT prostate gland uptake: Association with overall survival. *Clinical physiology and functional imaging*, December 3. doi: 10.1111/cpf.12611

PROVENCIO, Mariano, CAMPS, Carlos, COBO, Manuel, DE LAS PEÑAS, R., MASSUTI, Bartomeu, BLANCO, R., ALBEROLA, V., JIMÉNEZ, U., DELGADO, J. R., CARDENAL, F., TARÓN, Miguel, RAMÍREZ, J. L., SÁNCHEZ, Antonio, ROSELL, Rafael (2012). Prospective assessment of XRCC3, XPD and Aurora kinase A single-nucleotide polymorphisms in advanced lung cancer. Cancer chemotherapy and pharmacology, 70(6), 883-90. doi: 10.1007/s00280-012-1985-9

PROVENCIO, Mariano, ISLA, Dolores, SÁNCHEZ, Antonio, CANTOS, Blanca (2011). Inoperable stage III non-small cell lung cancer: Current treatment and role of vinorelbine. *Journal of thoracic disease*; 3(3), 197-204. doi: 10.3978/j.issn.2072-1439.2011.01.02

SAVOVA, Guergana K., DANCIU, Ioana, ALAMUDUN, Folami, MILLER, Timothy, LIN, Chen, BITTERMAN, Danielle S., TOURASSI, Georgia, WARNER, Jeremy L. (2019). Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. *Cancer research*, August 8. doi: 10.1158/0008-5472.CAN-19-0579

SCULIER, Jean-Paul, BERGHMANS, Thierry; MEERT, Anne-Pascale (2014). TNM classification and clinicopathological factors: What is helpful for adjuvant chemotherapy decision after lung cancer resection? *Journal of thoracic oncology*, 9(3), 266-270. doi: 10.1097/JTO.0000000000000110

SIEGEL, Rebecca L., MILLER, Kimberly D. y JEMAL, Ahmedin (2018). Cancer statistics. *CA: A cancer journal for clinicians*, 68(1), 7-30. doi: 10.3322/caac.21442

WILJER, David y HAKIM, Zaki (2019). Developing an artificial intelligence-enabled health care practice: Rewiring health care professions for better care. *Journal of medical imaging and radiation sciences*, 50(4), 8-14. doi: 10.1016/j.jmir.2019.09.010

WIRTH, Rüdiger y HIPP, Jochen (2000). CRISP-DM: Towards a standard process model for data mining. En *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, p. 29–39. Recuperado de https://pdfs.semanticscholar.org/48b9/293cfd4297f855867ca278f7069abc6a9c24.pdf

WITTEN, Ian H., FRANK, Eibe y HALL, Mark A. (2011). *Data mining: Practical machine learning tools and techniques*. Third edition. Burlington, MA: Morgan Kaufmann, 664 p.