

# Big Data en la Salud

## Detección Digital de Enfermedades

Lía L. Aguirre<sup>1</sup>, Juan P. Venditti<sup>1</sup>

<sup>1</sup>Ingeniería en sistemas de información – Universidad Tecnológica Nacional (UTN FRCU)

aguirrel2011, jpvenditti96@gmail.com

**Resumen** *El presente trabajo está orientado a destacar los beneficios del uso de herramientas de big data para procesar grandes volúmenes de datos, con el propósito de detectar tendencias o patrones que comparten las personas frente al padecimiento de distintas enfermedades o síndromes. Esto se logró a través del estudio y análisis de casos donde se ha utilizado la Detección Digital de Enfermedades, en los cuales se demostró que permite asistir a los profesionales de la salud a detectar de forma temprana padecimientos y tomar decisiones al respecto, de esta manera, mejorar sustancialmente la calidad de la atención médica.*

### 1. Introducción

Al utilizar el término big data, se hace referencia a grandes volúmenes de datos digitales, tanto estructurados como no estructurados que, por su elevado volumen, diversidad y complejidad, no pueden ser almacenados ni visualizados con las herramientas tradicionales, como así también, hace referencia a las herramientas necesarias para obtener valor de estos grandes volúmenes de datos. Estos datos son generados día a día desde diversas fuentes, principalmente por *las personas* que realizan *transacciones* online, *marketing electrónico* y *web*, *interacciones máquina a máquina* y *datos biométricos* recolectados [TRC 2013].

Los rápidos avances de las tecnologías de alto rendimiento y la amplia adopción de registros electrónicos de salud (EHRs - Electronic Health Records) han conducido a una rápida acumulación de datos relacionados con la salud. Estos voluminosos datos complejos contienen abundante información para la medicina de precisión, y el análisis de big data puede extraer ese conocimiento para mejorar la calidad de la atención médica [Wu et al. 2017].

Este gran volumen de datos complejos, específicamente aquellos que son generados por servicios de salud, permiten identificar tendencias o patrones que comparten los seres humanos frente al padecimiento de distintas enfermedades o síndromes. A partir de la identificación de estas tendencias o patrones, se pueden determinar qué acciones se deben llevar a cabo, o qué recursos utilizar, para atender dichos padecimientos.

El potencial de la investigación médica impulsada por big data es enorme [Ramos-Casals et al. 2015], principalmente en áreas como la epidemiología, la genética y la terapéutica, y abarca desde análisis relativamente básicos, como determinar incidencias de enfermedades en una población determinada, hasta estudios más complejos, como la eficacia de nuevos fármacos en un entorno del mundo real.

Este artículo se basa, principalmente, en los datos generados por las personas, con el fin de detectar patrones de enfermedades a partir del uso de herramientas de big data para el procesamiento de dichos datos.

## 2. Detección Digital de Enfermedades

Así como existe la epidemiología tradicional, también surgió la epidemiología digital o también conocida como detección digital de enfermedades (DDD - Digital Disease Detection) que, a diferencia de la epidemiología tradicional, se encarga de obtener los datos a través de fuentes que poseen datos electrónicos que surgen al utilizar tecnologías de información. La detección digital de enfermedades se basa en desarrollos tales como la disponibilidad generalizada de acceso a Internet, el crecimiento exponencial de dispositivos móviles y plataformas de intercambio en línea, que generan constantemente grandes cantidades de datos que contienen información relacionada con la salud, aun cuando el objetivo principal por el que se recogen dichos datos no es la salud pública. Además, este enfoque novedoso se basa en la idea de que la información relevante para la salud pública en la actualidad se genera cada vez más directamente por la población a través del uso de servicios en línea, sin que necesariamente se hayan comprometido con el sistema de salud.

En [Vayena et al. 2015] se menciona que *al utilizar datos globales en tiempo real, la DDD promete la detección acelerada de brotes de enfermedades, y ya se han publicado ejemplos de esta mejora en la puntualidad en la detección en la literatura*. Más adelante se van a desarrollar en este artículo algunos casos reales donde se ha utilizado la DDD para la detección temprana de enfermedades.

Este es un campo reciente que se está desarrollando rápidamente [Velasco et al. 2014], que podría generar grandes beneficios para la sociedad. En este trabajo se presentan algunos de sus beneficios, pero no hay que perder de vista que también es necesario reconocer que existen riesgos potenciales y consecuencias imprevistas, relacionados principalmente con cuestiones éticas [Vayena et al. 2015].

### 2.1. Enfermedades coronarias

En el caso del artículo [Letian et al. 2016] el objetivo de estudio fue analizar los datos sobre las enfermedades coronarias en una base de datos existente. Se construyó un modelo matemático a partir de indicadores clínicos y tipos de enfermedades coronarias para realizar la predicción de tipos de enfermedades coronarias.

Este estudio se realizó sobre una muestra de datos clínicos de 5000 pacientes con enfermedades coronarias, que fueron extraídos de una base de datos existente. Algunos de los datos de los pacientes incluían diabetes, presión arterial, edad, sexo, antecedentes de tabaquismo o de enfermedades hereditarias, posición y duración del dolor, tipos de enfermedades, etc. Se utilizó Hadoop [Hadoop 2018] para construir una plataforma de procesamiento de big data clínico para analizar los datos. Se diseñó un experimento de contraste para probar el clasificador propuesto. En el grupo experimental, se usaron la red neuronal de BP y el algoritmo Naive Bayes para analizar el tipo de enfermedad coronaria. El resultado final de la clasificación se obtuvo a través de un Boolean Decision Maker, basado en la probabilidad de qué entradas provenían de la red neuronal de BP y del clasificador Naive Bayes. En el grupo de control, se utilizó el algoritmo de regresión logística para calcular los resultados de la predicción.

Dentro de las 5000 muestras, se tomó un 75 % para entrenar al nuevo clasificador, y el otro 25 % se definió como muestras de experimentos (que se tomaron como entrada tanto para un grupo experimental como para un grupo de control).

Luego de realizar el experimento se pudo concluir que se puede mejorar la predicción utilizando el clasificador combinado de la red neuronal de BP y el algoritmo Naives Bayes con la ayuda de la plataforma de big data. Este nuevo sistema puede ayudar eficazmente a los médicos a tomar decisiones.

## 2.2. Geoepidemiología Autoinmune

Las enfermedades autoinmunes pueden ser localizadas o sistémicas, es decir que afectan específicamente a un órgano o pueden afectar a un conjunto de órganos y sistemas. A pesar de ser un caso especial con respecto a todas las enfermedades de gran variedad que existen en todo el mundo, las enfermedades autoinmunes sistémicas (SADs - Systemic Autoimmune Diseases) suelen ser uno de los principales factores de morbilidad y mortalidad.

Según [Ramos-Casals et al. 2015] *las enfermedades autoinmunes se caracterizan por un amplio espectro de patrones demográficos con respecto a la edad al momento del diagnóstico, la distribución de género y las diferencias étnicas*, es por ello que comenzar a utilizar big data permite el estudio de las distintas regiones geográficas donde se propagan este tipo de enfermedades y de esta manera aportar a la detección de determinados patrones que hacen que ciertas enfermedades ocurran en características genéticas y ambientales específicas. Constantemente se utilizan herramientas que provee big data, y gracias a ello se han logrado detectar este tipo de enfermedades localizadas en ciertas zonas geográficas.

En el caso del artículo [Ramos-Casals et al. 2015], se analizaron 394,827 pacientes con enfermedades autoinmunes sistémicas. Se exploró el potencial del motor de búsqueda de Google para recopilar y combinar grandes series (más de 1000 pacientes) de SAD informadas en la biblioteca Pubmed [NCBI 2018], con el objetivo de obtener una imagen geoepidemiológica de alta definición de cada enfermedad. Las variables recolectadas fueron: enfermedad autoinmune sistémica, número de pacientes, país, fuentes de datos, tipo de base de datos, criterios de inclusión y exclusión, distribución de género, edad media de inicio, diagnóstico y datos de etnia. Finalmente, se incluyeron 85 estudios, de los cuales se recolectaron los datos de 394,827 pacientes con SAD antes mencionados.

El análisis de los casi 400,000 pacientes con SAD ha confirmado la heterogeneidad epidemiológica de estas enfermedades con respecto a la *edad* del diagnóstico, la *influencia de los orígenes étnicos* y la brecha de *género* principalmente en adultos, con un 75 % de pacientes mujeres.

La geoepidemiología puede arrojar pistas importantes sobre los mecanismos desencadenantes genéticos y ambientales de la autoinmunidad [Ramos-Casals et al. 2015]. En este artículo se destaca el uso del motor de búsqueda de Google, debido a que los motores de búsqueda se han convertido en una parte esencial de la vida de las personas, incluyendo la salud, la búsqueda de Google hace que sea realmente fácil encontrar la información correcta.

El uso de un enfoque de big data permitió que surgieran conexiones hasta ahora invisibles en las SAD, y es de gran utilidad para la toma de decisiones médicas en la vida real. En la Figura 1 se aprecia un sesgo significativo en la distribución geográfica de los estudios de big data para cada SAD.

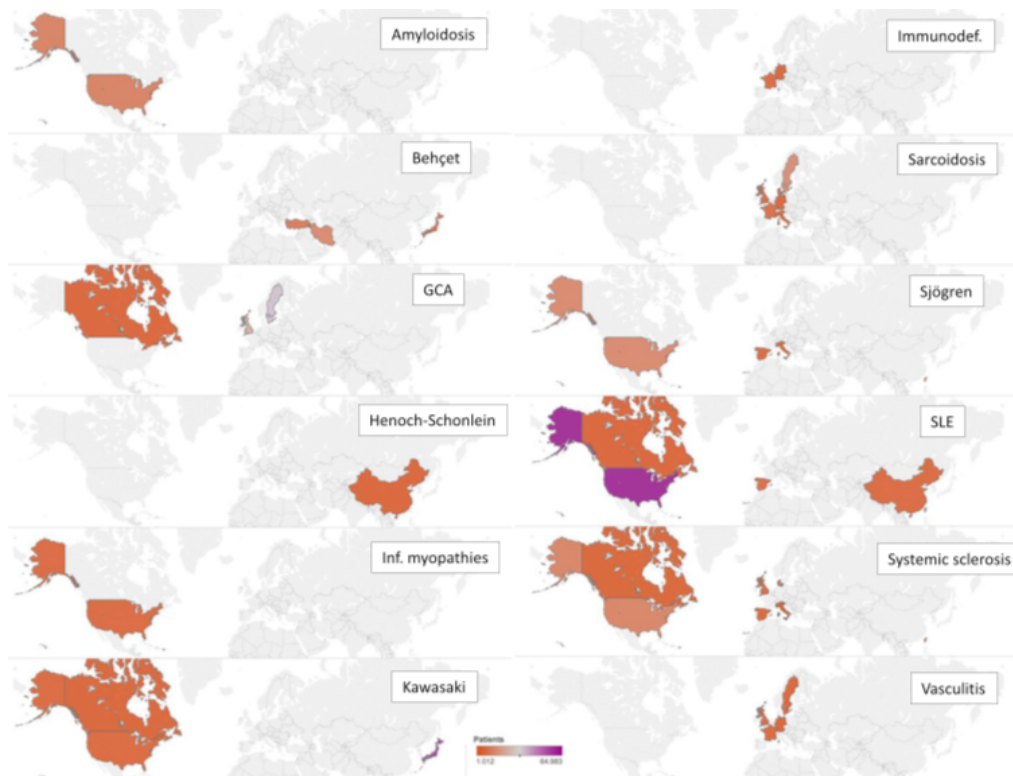


Figura 1. Pacientes incluidos por país de publicación.

### 3. Discusión

Aunque en el presente trabajo se han tratado casos específicos en los que se han aplicado herramientas de big data para la detección digital de enfermedades, los grandes volúmenes de datos que se generan día a día pueden ser tratados para la extracción de patrones o tendencias que se manifiestan alrededor del mundo en que vivimos, y de esta manera, poder comprender y tratar eficientemente la propagación y el tratamiento de ciertas enfermedades.

### 4. Conclusión

Luego de realizar una investigación sobre el uso de herramientas de big data en el ámbito de la salud y de estudiar casos reales donde se han aplicado las mismas, se concluye que la utilización de dichas herramientas brinda facilidades a los profesionales de la salud, debido a que les proporciona información para tomar decisiones y reducir el margen de error que puede existir al realizar un diagnóstico. Esto resulta beneficioso, tanto para los profesionales de la salud como para los pacientes, ya que los primeros mencionados podrían detectar de forma temprana patologías, con menor margen de error y determinar el mejor tratamiento, en consecuencia, se lograría disminuir la tasa de mortalidad de los pacientes y agilizar el proceso de elaboración de diagnósticos, haciendo este proceso más eficiente y eficaz.

### Referencias

Hadoop (2018). [Web; accedido el 23-08-2018].

- Letian, W., Han, L., Zhang, L., and Guo, S. (2016). Gw27-e0397 an analysis and diagnosis system of coronary heart disease based on big data platform. *Journal of the American College of Cardiology*, 68(16 Supplement):C82.
- NCBI (2018). [Web; accedido el 23-08-2018].
- Ramos-Casals, M., Brito-Zerón, P., Kostov, B., Sisó-Almirall, A., Bosch, X., Buss, D., Trilla, A., Stone, J. H., Khamashta, M. A., and Shoenfeld, Y. (2015). Google-driven search for big data in autoimmune geoepidemiology: Analysis of 394,827 patients with systemic autoimmune diseases. *Autoimmunity Reviews*, 14(8):670 – 679.
- TRC, I. (2013). Conceptos básicos de big data.
- Vayena, E., Salathé, M., Madoff, L. C., and Brownstein, J. S. (2015). Ethical challenges of big data in public health. *PLOS Computational Biology*, 11(2):1–7.
- Velasco, E., Aghenza, T., Denecke, K., Kirchner, G., and Eckmanns, T. (2014). Social media and internet-based data in global systems for public health surveillance: A systematic review. *The Milbank Quarterly*, 92(1):7–33.
- Wu, P., Cheng, C., Kaddi, C. D., Venugopalan, J., Hoffman, R., and Wang, M. D. (2017). –omic and electronic health record big data analytics for precision medicine. *IEEE Transactions on Biomedical Engineering*, 64(2):263–273.