Cuba Salud
IV Convención
Internacional de Salud
17-21 de octubre, 2022

Big data y aprendizaje automático para mejorar los procesos en los Ensayos Clínicos: mapeo sistemático de la literatura

Ing. Andrys Adonis Santos Domínguez[1]

[1] Centro de Informática Médica, Universidad de las Ciencias Informáticas, La Habana, Cuba, contacto: aasantos@uci.cu

Resumen: En el siglo XXI la cantidad de datos que se generan a diario es enorme, se han convertido en una mina de la que se puede extraer información muy valiosa. Utilizar estos datos en beneficio de la salud tiene que ser prioritario en nuestro país. Actualmente los ensayos clínicos en Cuba no hacen uso de los datos almacenados previamente para mejorar el diseño de los mismos. Mejorar y optimizar su diseño tiene que ser premisa de los investigadores y para ello el análisis de la literatura constituye un elemento indispensable. Como metodología se aplica la propuesta de Petersen et. al. actualizada en el 2015 para la realización del mapeo. Se realizo la búsqueda entre los años 2017 y 2022 en las revistas y bases de datos más destacadas en el ámbito de la informática médica**.** Se encontraron 309 trabajos de los cuales fueron seleccionados 3. Se observa como el uso de técnicas de minería de datos y aprendizaje automático ayudan a optimizar el diseño de ensayos clínicos lo que puede significar ahorros millonarios en la industria biofarmacéutica.

Palabras clave: big data, machine learning, ensayos clínicos, salud

I. INTRODUCCIÓN

La transformación digital en salud constituye un claro ejemplo de la importancia e impacto de las Tecnologías de la Información y las Comunicaciones (TICs) en la sociedad. Esto ha permitido al sector de la salud, contar con métodos novedosos, sencillos y eficaces de gestión administrativa en consultas, hospitales y centros de investigación biomédica. La Universidad de las Ciencias Informáticas (UCI), más específicamente el Centro de Informática Médica (CESIM) tiene entre sus productos el Sistema para el manejo de datos de ensayos clínicos (SIDEC) el cual agiliza el diseño y conducción de los ensayos clínicos (EC). El mismo impacta de manera positiva en la industria biotecnológica y farmacéutica nacional agilizando el diseño y conducción de los EC dando evidencias de los beneficios de la transformación digital de la industria y de la vinculación Universidad-Empresa (1).

Cuba intensifica la introducción de las TICs en la sociedad mediante el desarrollo y el despliegue de sistemas de información como el anterior. De conjunto al proceso de informatización, se realizan importantes esfuerzos para la creación y el mantenimiento de la infraestructura tecnológica que soporte el flujo masivo de datos en la red nacional y su interconexión con la red internacional. El crecimiento en el volumen de datos generados por diferentes sistemas y actividades cotidianas en la sociedad ha forjado la necesidad de modificar, optimizar y generar métodos y modelos de almacenamiento y tratamiento de datos que suplan las falencias que presentan las bases de datos y los sistemas de gestión de datos tradicionales. Respondiendo a esto aparece el término *Big Data* en su sigla del idioma inglés como término que incluye diferentes tecnologías asociadas a la administración de grandes volúmenes de datos provenientes de diferentes fuentes y que se generan con rapidez (2,3). Uno de los recursos más populares en el campo de la ciencia de datos es el *Machine Learning* (ML) o aprendizaje automático.

El ML es un término amplio que agrupa varias estrategias analíticas cuyo propósito es el desarrollo de algoritmos para extraer información de los datos ya sea para explicación, clasificación o predicción. Desde 2016 forma parte de los términos incluidos en el *Medical Index Subheadings* (MeSH, por sus siglas en inglés) de *Pubmed*. Pese a que se suele considerar como sinónimo de inteligencia artificial, es importante precisar que la inteligencia artificial es una clasificación aún más amplia que incluye tanto técnicas para el análisis de datos estructurados como el aprendizaje automático y datos no estructurados como procesamiento de lenguaje natural (4). En el ámbito de la medicina ha sido utilizado para aumentar la precisión diagnóstica, hacer predicciones de mortalidad hospitalaria o predecir la necesidad de ciertas terapias, dejando de lado su uso en la optimización del diseño de los EC.

Un Ensayo Clínico, es una evaluación experimental de un producto, sustancia, medicamento, técnica diagnóstica o terapéutica que, en su aplicación a seres humanos, pretende valorar su eficacia y seguridad. Los primeros EC que realizó la humanidad fueron en un grupo muy pequeño de personas y algunos de ellos no seguían ciertos estándares de protección a los sujetos que participaban. Como resultado de la evolución cognitiva del ser humano, los EC se han vuelto más complejos cada día. (5)

El uso del *big data* de conjunto con el aprendizaje automático provee de innumerables ventajas que no están siendo explotadas. El presente trabajo tiene como objetivo realizar un mapeo sistemático de la literatura para analizar de forma simplificada como se puede mejorar y/u optimizar cualquiera de las etapas que componen un EC.

II. MÉTODO

Los estudios de mapeo sistemático o estudios de alcance están diseñados para dar una visión general de un área de investigación a través de la clasificación y contabilizar las contribuciones en relación con las categorías de esa clasificación (6). En el presente trabajo se siguen los pasos descritos por Petersen et al. (6) en el año 2015 que incluye las siguientes actividades:

1. Planear el mapeo
    a) Determinar la necesidad y su alcance: incluye la definición de las preguntas de la investigación que deberá de tener en mente para realizar la búsqueda.
    b) Identificar los trabajos: se selecciona y ejecuta la estrategia de búsqueda, se evalúan los resultados obtenidos, se aplican los criterios de inclusión y exclusión. Si se considera necesario, se realizan pruebas a la calidad de los resultados.
    c) Extraer y clasificar los datos: se realiza el proceso de extracción y clasificación, se aplican clasificaciones independientes del tema o específicas de la temática de la investigación.
    d) Visualizar los datos: se emplean gráficos de pasteles, barras u otros, que faciliten las diferentes clasificaciones obtenidas.
    e) Identificar las amenazas de validez: el sesgo entre la publicación de resultados sólo positivos, la escasez de información recopila, la calidad de los estudios seleccionados, entre otros aspectos, deben ser tomados en cuenta según la temática que se investiga.
2. Realizar el mapeo
3. Informar el mapeo

También se especifican las actividades 4) Evaluar el proceso de mapeo y 5) Publicación.

Para identificar las revistas más destacadas en el ámbito de la salud se consultó el sitio web *Scimago Journal & Country Rank*[1] que es un portal disponible públicamente que incluye las revistas y los indicadores científicos de países desarrollados organizados mediante el indicador *SCImago Journal Rank*, que determina el impacto y visibilidad de las instituciones y revistas electrónicas de las bases de datos más importantes basándose en la transferencia de prestigio de una revista a otra una. Se analizaron los resultados obtenidos para el año 2021 (más resiente), con la temática *Health Informatics* (informática médica o informática en salud) y que estuvieran indizadas en la web de la ciencia (WoS).

De las revistas resultantes se tuvieron en cuenta su alcance, popularidad y disponibilidad de sus artículos. Finalmente fueron seleccionadas: *International Journal of Medical Informatics* (IJMI) y *Journal of Biomedical Informatics* (JBI). También se analizaron las principales bases de datos médicas que existen: *PubMed*[2], *IEEE Xplore*[3] y *ACM*[4]. Para la gestión bibliográfica se utilizó el software Zotero.

[1] Disponible en: https://www.scimagojr.com/journalrank.php
[2] Disponible en: https://pubmed.ncbi.nlm.nih.gov/
[3] Disponible en: https://ieeexplore.ieee.org/Xplore/home.jsp

III. RESULTADOS

A. Preguntas de investigación

Para orientar el mapeo, se definieron las siguientes preguntas de investigación: **Q1)** ¿Qué trabajos existen sobre el uso de minería de datos y aprendizaje automático en EC? **Q2)** ¿Dónde y cuándo se publicaron estos trabajos? **Q4)** ¿En qué etapas se enfocan más los trabajos?

B. Identificación de los trabajos

Se realizó la una búsqueda en las bases de datos y revistas usando la siguiente expresión de búsqueda: *("data mining" or "big data" or "machine learning") and ("clinical trials" or "clinical trial")*. Se escogieron esos términos porque se tuvo en cuenta el *Medical Subject Heading* (MeSH) que es el vocabulario controlado que emplea en varias bases de datos biomédicas para procesar la información que se introduce en cada una de ellas. Se analizó las características de los términos para obtener resultados más precisos, reduciendo el número de trabajos irrelevantes a los intereses de los usuarios y eliminando las inoportunas sinonimias, responsables de muchas ausencias en la recuperación de información. (7)

Se tuvieron en cuenta los siguientes criterios de inclusión:

- Trabajos publicados entre 2017-2022.
- Trabajos de acceso público.
- Trabajos en los que el tema central no sea el uso de minería de datos para mejorar los procesos en los EC.

Se excluyeron los trabajos:

- Solo hacen mención a la minería de datos, pero no la aplican.
- Artículos de pago.
- Artículos con metadatos incompletos (el título, resumen o palabras claves no están disponibles).
- Artículos en idiomas distintos de español o inglés.

Los resultados de aplicar la consulta a las bases de datos (ver tabla 1) y a las revistas (ver tabla 2) se pueden apreciar la cantidad trabajos encontrados y seleccionados.

[4] Disponible en: https://dl.acm.org/

Tabla 1 Cantidad de trabajos identificados y seleccionados en las bases de datos consultadas. Fuente: Elaboracion propia

Bases de datos	Trabajos Encontrados	Trabajos Seleccionados
NIH	72	3
IEEE Explore	4	0
ACM	1	0
Total	77	3

Tabla 2 Cantidad de trabajos encontrados y seleccionados en revistas. Fuente: Elaboración propia

Revista	Trabajos Encontrados	Trabajos Seleccionados
IJMI	66	0
JBI	166	0
Total	232	0

Se observa en los resultados que, a pesar de encontrarse varios trabajos, muy pocos cumplen los criterios de inclusión antes mencionados.

C. Extracción y clasificación de los datos

La extracción y clasificación de los datos, consiste en registrar los elementos fundamentales cada trabajo seleccionado. A continuación, se expone un breve resumen de los artículos seleccionados.

Mayo et al. (8) proponen ayudar con el diseño de los ECs haciendo uso de los novedosos sistemas de recursos de análisis de *big data* (BDARSs, por sus siglas en inglés) permitiendo un mejor diseño, reducción de costos, mejorar los presupuestos de proyección para el ensayo, garantizar que los pacientes inscritos en el ensayo reflejen la población prevista, podrían utilizarse para estimar con mayor precisión el tamaño de la muestra, entre otras. Son utilizados fundamentalmente en ensayos controlados aleatorizados, agregan datos clínicos de varios sistemas que incluyen registros electrónicos de salud, sistemas de información de oncología radioterápica, sistemas de planificación de tratamiento y otros. Los mismos se encuentran en una ubicación común diseñada para respaldar el análisis de estos datos para mejorar la atención al paciente. Este enfoque utiliza datos reales en lugar de utilizar proyecciones hipotéticas y evita tener que ajustar el ensayo después de que este se inicia. Los BDARS también podrían usarse para estimar con precisión el número de pacientes elegibles.

Cai et al. (9) proponen el uso de aprendizaje automático para potenciar y ayudar a los investigadores a acelerar la identificación de pacientes y su reclutamiento, lo que permite avanzar hacia EC más eficientes. Normalmente la selección de sujetos requiere revisiones laboriosas y costosas de historias clínicas de los pacientes. Mediante los datos de registro de salud electrónico y el uso del aprendizaje automático que incorpora códigos de facturación y datos de notas clínicas procesadas mediante técnicas de procesamiento natural del lenguaje se puede mejorar la eficiencia de la selección de elegibilidad. La aplicación del algoritmo redujo los pacientes no elegibles de la revisión de historias clínicas en un 40,5 % en el centro de atención terciaria y en un 57,0 % en el hospital comunitario.

Vázquez et al. (10) utilizan técnicas de aprendizaje automático para obtener una comprensión más profunda de los datos al descubrir patrones y tendencias que no son aparentes para descubrir las caracte-

rísticas de los individuos que tienen más probabilidades de mostrar interés en participar en ECs. Luego de probar distintos algoritmos llegaron a predecir la probabilidad de que un individuo exprese su interés en participar en un ensayo con una precisión del 75%. La precisión indica que 75 de cada 100 veces, el algoritmo es probable que prediga correctamente si un individuo mostraría interés en un estudio.

D. Visualización

Las revistas y bases de datos escogidas destacan por ser de las más importantes en el sector de la informática médica, tienen una alta demanda y publican varios números por años. Destaca *PubMed*, con más de 34 millones de artículos, la búsqueda arrojó 73 artículos, de los cuales se seleccionados 3 (ver Figura 1). Se evidencia un creciente interés en el tema ya que se observa un aumento en la cantidad de publicaciones por año (ver Figura 2). El idioma inglés corresponde al 100% de las publicaciones seleccionadas.

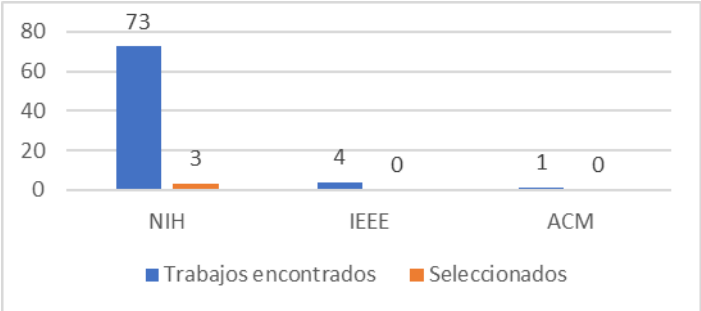

Fig. 1 Trabajos encontrados y seleccionados de las bases de datos. Fuente: Elaboración propia

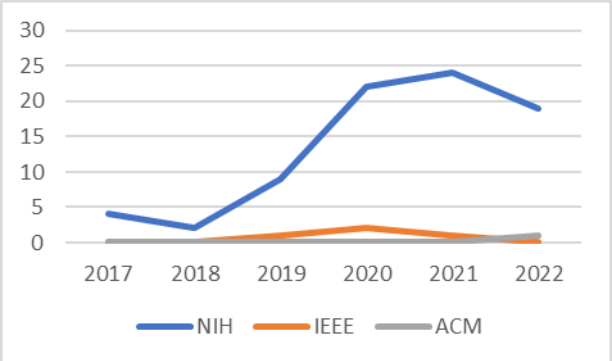

Fig. 2 Trabajos identificados por años. Fuente: Elaboración propia

La búsqueda en las revistas no arroja resultados satisfactorios no encontrándose ningún artículo que cumpla con los criterios de inclusión previamente mencionados.

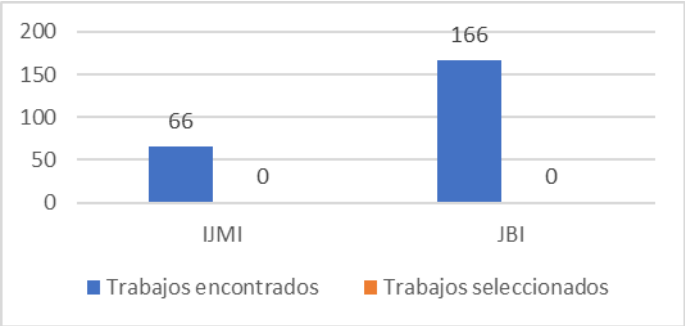

Fig. 3 Trabajos identificados en las revistas seleccionadas. Fuente: Elaboración propia

E. Discusión

Existen pocos trabajos publicados sobre este tema, a pesar de que poco a poco se va mostrando más interés. La etapa que más se aborda en la literatura es el diseño, enfocándose a un diseño óptimo, y, mediante el uso de las nuevas tecnologías descubrir patrones que no están siendo analizados por los investigadores. Existe una tendencia al incremento en la cantidad de artículos publicados, siendo estos aun insuficientes.

IV. CONCLUSIONES

Los EC se han vuelto más costosos, impulsados por un número creciente de partes interesadas que requieren más criterios de valoración, poblaciones de pacientes más diversas y un entorno normativo estricto. El uso del *big data* y el aprendizaje automático ayuda a realizar un diseño más óptimo lo que permite ahorros millonarios en la industria.

Su creciente aplicación es evidencia de que esta tecnología está cambiando la forma en la que diseñadores de ensayos toman sus decisiones aumentando el éxito del reclutamiento de pacientes y la reducción de la duración del ensayo.

REFERENCIAS

1. Vega Izaguirre L, Quintana Díaz VM, Tamayo Peña R, Domínguez Izquierdo YD, Molina Hernández Y, Vega Izaguirre L, et al. Sistema para el manejo de datos de Ensayos Clínicos XAVIA SIDEC. Rev Cuba Informática Médica [Internet]. junio de 2021 [citado 1 de julio de 2022];13(1). Disponible en: http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S1684-18592021000100005&lng=es&nrm=iso&tlng=es

2. Hernández-Leal EJ, Duque-Méndez ND, Moreno-Cadavid J. Big Data: una exploración de investigaciones, tecnologías y casos de aplicación. TecnoLógicas. 2 de mayo de 2017;20(39):15-38.

3. Vitón-Castillo AA, Linares-Cánovas LP, Vitón-Castillo AA, Linares-Cánovas LP. Big data en el contexto de la salud cubana. Rev Cuba Salud Pública [Internet]. septiembre de 2019 [citado 13 de

julio de 2022];45(3). Disponible en: http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S0864-34662019000300017&lng=es&nrm=iso&tlng=es

4. Pedrero V, Reynaldos-Grandón K, Ureta-Achurra J, Cortez-Pinto E, Pedrero V, Reynaldos-Grandón K, et al. Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. Rev Médica Chile. febrero de 2021;149(2):248-54.

5. Martínez Nieto C. ENSAYOS CLÍNICOS: Actualización en ética, normativa, metodología y nuevas tecnologías.

6. Petersen K, Vakkalanka S, Kuzniarz L. Guidelines for conducting systematic mapping studies in software engineering: An update. Inf Softw Technol. agosto de 2015;64:1-18.

7. Pinillo León AL, Cañedo Andalia R. El MeSH: una herramienta clave para la búsqueda de información en la base de datos Medline. ACIMED. abril de 2005;13(2):1-1.

8. Mayo CS, Matuszak MM, Schipper MJ, Jolly S, Hayman JA, Ten Haken RK. Big Data in Designing Clinical Trials: Opportunities and Challenges. Front Oncol. 31 de agosto de 2017;7:187.

9. Cai T, Cai F, Dahal KP, Cremone G, Lam E, Golnik C, et al. Improving the Efficiency of Clinical Trial Recruitment Using an Ensemble Machine Learning to Assist With Eligibility Screening. ACR Open Rheumatol. septiembre de 2021;3(9):593-600.

10. Vazquez J, Abdelrahman S, Byrne LM, Russell M, Harris P, Facelli JC. Using supervised machine learning classifiers to estimate likelihood of participating in clinical trials of a de-identified version of ResearchMatch. J Clin Transl Sci. 2021;5(1):e42.