

Alfredo Madrid-García¹, Beatriz Merino-Barbancho², Luis Rodríguez-Rodríguez¹

¹Grupo de Patología Musculoesquelética. Hospital Clínico San Carlos. Instituto de Investigación Sanitaria San Carlos (IdISSC). Prof. Martin Lagos s/n, Madrid, 28040, Spain

²Escuela Técnica Superior de Ingenieros de Telecomunicación. Universidad Politécnica de Madrid, Avenida Complutense, 30, Madrid, 28040, Spain

INTRODUCCIÓN

El modelado de temas (TM) es una técnica de minería de textos que permite descubrir la estructura semántica latente, los temas, presentes en una colección de documentos de forma no supervisada, así como la clasificación de dichos documentos en los distintos temas. Esta técnica se puede emplear también para estudiar tendencias bibliométricas. El objetivo de este estudio es: a) aplicar TM para descubrir los temas presentes en los artículos en los que se ha aplicado inteligencia artificial (IA) en el ámbito de las enfermedades musculoesqueléticas (RMDs), b) estudiar el conjunto de palabras que caracterizan cada tema, c) analizar qué temas predominan en las publicaciones de las revistas especializadas de reumatología.

MATERIAL Y MÉTODOS

Se extraen los *abstracts* de los artículos científicos publicados en *PubMed* entre el año 2000 hasta el 3 de octubre de 2023, mediante una *query* combinando los siguientes términos: *Artificial Intelligence, Big Data, Data Mining, Supervised Learning, Unsupervised Learning, Deep Learning con Rheumatology, Rheumatic, Musculoskeletal*.

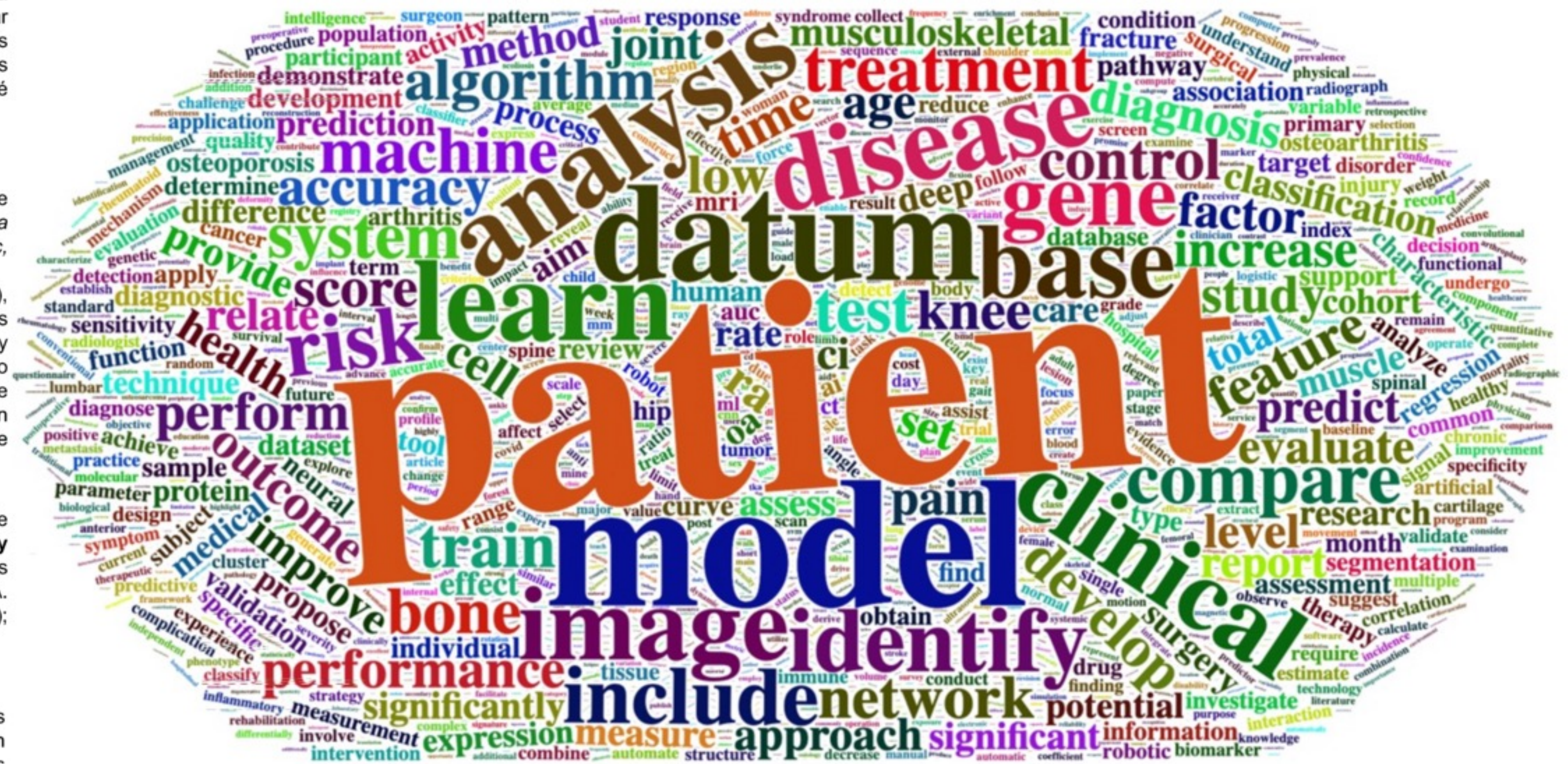
Se pre-procesa el texto de cada *abstract* mediante la eliminación de las frases vacías (e.g., copyright © elsevier), de *stopwords*, de símbolos de puntuación, de dígitos, se aplica lematización y nos quedamos con aquellas palabras que aparecen al menos entre el 1% - 50% de los documentos. Se crea una matriz término-documento y se calcula el número óptimo de temas mediante el *coherence score*, probando de 1 a 25. Se aplica el modelo bayesiano *Latent Dirichlet Allocation* para clasificar cada *abstract*, en función de la probabilidad *gamma*, y se hace una representación mediante nube de palabras. Finalmente, seleccionamos los *abstracts* publicados en revistas del área de la reumatología, según el índice JCR, para estudiar los temas representativos en este subconjunto.

RESULTADOS

Se recuperan 7,358 artículos, de los cuales 7,085 tienen *abstract* registrado en PubMed. El número óptimo de temas es 20. Los cinco temas principales son: **expresión génica** (10.16%); **aprendizaje profundo, imagen y segmentación** (8.64%); **práctica clínica** (7.52%), **marcha y rehabilitación** (7.01%); y modelos predictivos (5.97%). Por último, en 31 de 35 revistas de reumatología, identificadas por JCR, se identificaron *abstracts* de IA. Los cinco temas principales de este subconjunto son: **artritis reumatoide** (17.05%); **práctica clínica** (11.05%); **factores de riesgo** (8.75%); **revisión bibliográfica** (8.45%), y **expresión génica** (8.29%).

CONCLUSIONES

El MT permite estudiar la estructura semántica de un conjunto de documentos. En este caso, nos hemos centrado en literatura científica que combinan las RMDs y la IA. La mayor parte de artículos que combinan ambos mundos guardan relación con la expresión génica. Sin embargo, si particularizamos a las revistas más especializadas, el tema central sobre el que versan los *abstracts* y presumiblemente los artículos asociados, es la **artritis reumatoide**.



Palabras más prevalentes

Artritis reumatoide	Metodología	Respuesta terapéutica	Aprendizaje profundo	Expresión génica	Marcha y rehabilitación	Miscláneo	Miscláneo	Revisión inteligencia artificial	Factores de riesgo
Artrosis de rodilla	Identificación génica	Dolor	Algoritmos AI	Patología ósea	Carga rodilla	Práctica clínica	Modelos predictivos	Historia clínica electrónica	Columna vertebral

Palabras más representativas de cada tema