

Mapping two decades of research in rheumatology-specific journals: a topic modeling analysis with BERTopic

Alfredo Madrid-García , Dalifer Freites-Núñez , Beatriz Merino-Barbancho ,
Inés Pérez Sancristobal and Luis Rodríguez-Rodríguez 

Ther Adv Musculoskelet Dis

2024, Vol. 16: 1–17

DOI: 10.1177/
1759720X241308037

© The Author(s), 2024.
Article reuse guidelines:
sagepub.com/journals-
permissions

Abstract

Background: Rheumatology has experienced notable changes in the last decades. New drugs, including biologic agents and Janus kinase (JAK) inhibitors, have blossomed. Concepts such as window of opportunity, arthralgia suspicious for progression, or difficult-to-treat rheumatoid arthritis (RA) have appeared; and new management approaches and strategies such as treat-to-target have become popular. Statistical learning methods, gene therapy, telemedicine, or precision medicine are other advancements that have gained relevance in the field. To better characterize the research landscape and advances in rheumatology, automatic and efficient approaches based on natural language processing (NLP) should be used.

Objectives: The objective of this study is to use topic modeling (TM) techniques to uncover key topics and trends in rheumatology research conducted in the last 23 years.

Design: Retrospective study.

Methods: This study analyzed 96,004 abstracts published between 2000 and December 31, 2023, drawn from 34 specialized rheumatology journals obtained from PubMed. BERTopic, a novel TM approach that considers semantic relationships among words and their context, was used to uncover topics. Up to 30 different models were trained. Based on the number of topics, outliers, and topic coherence score, two of them were finally selected, and the topics were manually labeled by two rheumatologists. Word clouds and hierarchical clustering visualizations were computed. Finally, hot and cold trends were identified using linear regression models.

Results: Abstracts were classified into 45 and 47 topics. The most frequent topics were RA, systemic lupus erythematosus, and osteoarthritis. Expected topics such as COVID-19 or JAK inhibitors were identified after conducting dynamic TM. Topics such as spinal surgery or bone fractures have gained relevance in recent years; however, antiphospholipid syndrome or septic arthritis have lost momentum.

Conclusion: Our study utilized advanced NLP techniques to analyze the rheumatology research landscape and identify key themes and emerging trends. The results highlight the dynamic and varied nature of rheumatology research, illustrating how interest in certain topics has shifted over time.

Keywords: artificial intelligence, BERTopic, natural language processing, PubMed, topic modeling, transformers, trend analysis

Received: 22 July 2024; revised manuscript accepted: 3 December 2024.

Introduction

Over the past decades, the volume of academic literature has experienced significant growth.^{1,2}

The field of rheumatic and musculoskeletal diseases (RMDs) has not been immune to this growth (Supplemental Figure 1). Moreover,

Correspondence to:
Alfredo Madrid-García
Grupo de Patología
Musculoesquelética,
Hospital Clínico San
Carlos, Instituto de
Investigación Sanitaria San
Carlos, Prof. Martín Lagos
s/n, Madrid 28040, Spain
alfredo.madrid@salud.madrid.org

Dalifer Freites-Núñez
Inés Pérez Sancristobal
Luis Rodríguez-Rodríguez
Grupo de Patología
Musculoesquelética,
Hospital Clínico San
Carlos, Instituto de
Investigación Sanitaria San
Carlos, Madrid, Spain

Beatriz Merino-Barbancho
Escuela Técnica
Superior de Ingenieros
de Telecomunicación,
Universidad Politécnica de
Madrid, Madrid, Spain

RMDs have undergone an unprecedented change in recent years. To begin with, a drug development revolution took place in the early 2000s—which is still active today—with the arrival of promising drugs such as biologic agents or Janus kinase (JAK) inhibitors.^{3–5} Furthermore, the adoption of therapeutic strategies, such as treat-to-target,⁶ the earlier initiation of disease-modifying treatments, or the paradigm shift in how diseases are analyzed, not only by their mortality rate but also by their disability, propitiated a new scenario for RMDs.^{7,8} Concepts such as the window of opportunity,⁹ arthralgia suspicious for progression,¹⁰ erosive disease,¹¹ or difficult-to-treat rheumatoid arthritis (RA)¹² have gained momentum.

In this context of continuous change, we hypothesize that the study of trends in scientific publications could be beneficial to better understand the historical research priorities in rheumatology and the evolving landscape of RMD management and treatment. However, with almost 100,000 original articles published in the last 23 years, the process of comprehending and identifying the main trends is becoming increasingly challenging.

Conventional review methods can be labor-intensive, overwhelming or unfeasible, and non-exhaustive. Hence, we propose the use of modern natural language processing (NLP) techniques to characterize the evolution of the research topics addressed over time in rheumatology scientific publications. Topic modeling (TM) techniques are ideally suited for this, as they can model the evolution of topics over time. Briefly, TM is a suite of unsupervised learning algorithms (i.e., no tags/labels are provided with the input data), within the field of machine learning, designed to identify prevalent topics within a corpus of documents, usually through probabilistic methods.^{13,14} In that collection, the documents are observed while the topic structure (i.e., the topics, per-document topic distributions, and the per-document per-word topic assignments) is hidden.^{15,16} The outcome of a typical TM algorithm is clusters of related words. These techniques operate under the assumption that each topic is defined by a distinct collection of words and that a document consists of a blend of multiple topics in varying proportions. One of the most widely used TM techniques is Latent Dirichlet Allocation (LDA), a generative probabilistic model. However, with the recent advances in NLP and the introduction of the transformer's architecture, new TM

techniques that consider semantic relationships among words and their context have arisen (i.e., BERTopic).

Consequently, this study aims to apply BERTopic, a state-of-the-art TM technique, to analyze 20 years of rheumatology research within specialty journals. By mapping thematic trends, we aim to reveal both long-term research priorities and shifts in focus areas, providing insights into evolving themes and pinpointing research strengths and potential gaps.

Related work

TM has been used in a multitude of fields, including social networks, software engineering, crime science, political science, geography, medicine, and linguistics.¹⁷ In addition, it has proven effective in analyzing historical documents such as newspapers and humanistic texts,¹⁸ as well as in educational research,¹⁹ and the study of organizational phenomena.²⁰

TM has been widely applied in rheumatology research. Tedeschi et al.²¹ employed a TM approach, sureLDA, followed by penalized regression, to predict pseudogout probability in large datasets. TM was also applied to characterize the temporal evolution of ANCA-associated vasculitis (AAV).²² Temporal trends, in more than 113,000 clinical notes, before and after the treatment initiation date for a diagnosis of AAV, were modeled with LDA, finding 90 different topics that included diagnosis (e.g., granulomatosis with polyangiitis), treatments (e.g., AAV specific-treatment), and comorbidities and complications of AAV (e.g., glomerulonephritis, infections, skin lesions).

A prior study conducted by Dzibur et al.²³ explored the application of TM to understand the concerns and perceptions of patients with ankylosing spondylitis regarding biological therapies. The researchers analyzed over 25,000 social media posts using LDA and identified 112 topics. Medication uncertainty, lack of trust in physician's decisions, patient worries, and seeking alternative treatments highlighted were those most prevalent.

On its behalf, Li and Yacyshyn²⁴ analyzed the posts published over a year in the Reddit subforum "r/Behcet" to investigate the perspectives and experiences of people affected by Behcet's

disease. The authors identified 6 themes and 16 subthemes, including *finding connectedness through shared experiences, the struggles of the diagnostic odyssey, and sharing or inquiring about symptoms*.

Tang *et al.*²⁵ pursue to uncover the themes present in the electronic health record (EHR) of patients with RA prior to the start of targeted treatments and to explore their relationship with the subsequent course of treatment. On the other hand, Flurie *et al.*²⁶ evaluated two social media communities, a Facebook group, and a public subreddit (i.e., r/gout), identified 30 topics, and conducted sentiment analysis.

Moreover, Eaneff *et al.*²⁷ characterized systemic lupus erythematosus (SLE) patients' experiences in an online health community by applying LDA in free-text data extracted from the *PatientsLikeMe* community.

Eventually, Sperl *et al.*²⁸ applied LDA to analyze responses to open-ended questions from an online survey designed to assess motivations among health professionals for participating in postgraduate rheumatology education and to identify barriers and facilitators for participation in current EULAR educational offerings.

Supplemental Table 1 shows the most relevant characteristics of each study discussed above.

Materials and methods

Materials

Data from the *RheumaLpack* corpus,²⁹ which includes 96,004 rheumatology-related abstracts along with associated metadata, up to 19 variables including *title*, *PMID/DOI*, *abstract*, *publication year*, *journal*, *keywords*, or *volume*, were extracted. The criteria that were applied to select the articles used in the present work were as follows:

- Indexed in MEDLINE PubMed,
- A publication date between January 1, 2000 and December 31, 2023, both included,
- Belonging to journals classified by the 2023 Journal Citation Reports (Supplemental Table 2) as “RHEUMATOLOGY—SCIE,” and
- With an available abstract.

Briefly, the process that was followed to obtain the articles included the following steps: (a) we performed manual queries to retrieve PMIDs from PubMed, with the name of each journal followed by “[Journal]” (e.g., “Annals of the rheumatic diseases” [Journal]; Supplemental Table 3 shows the search strategy for each journal), saved using the PubMed save settings, and merged into a single document (122,426 PMIDs were recovered this way); (b) next, we used R's *rentrez* library to collect the following information from each retrieved PMIDs: DOI, MeSH keywords, volume, issue, pages, abstract, has abstract, publication type, language, PubMed central papers citation, sort first author, and affiliation. These data were gathered in three batches: the first batch contained the abstract and other publication details such as Medical Subject Headings (MeSH) keywords; the second batch containing the language, the publication type, and the PubMed central papers citation; and the third one containing the affiliation data. The remaining variables (i.e., title, authors, citation, journal/book, publication year, create date, PMCID, and NIHMS ID) were directly retrieved from the PubMed webpage, during the previous step; (c) Finally, not all the selected articles had an abstract since this information is only collected for a certain type of articles (e.g., original research articles, reviews); therefore, we excluded those without this information.

BERTopic was used for TM.³⁰ This technique generates topic representations through three steps. First, each abstract is converted into a vector representation (i.e., embeddings), using language models (e.g., *all-mpnet-base-v2*). Abstracts with similar meanings will have closely related vector representations, making them more likely to be grouped into the same topic. Second, since the generated vectors are of high dimensionality, they are transformed to reduce dimensionality and make clustering less computationally intensive and more efficient. This is achieved with a dimensionality reduction algorithm (e.g., UMAP). Following this, the reduced embeddings are grouped using clustering algorithms (e.g., HDBSCAN) to form distinct topics. Third, the significance of each word within a topic is calculated using a weighting scheme (i.e., c-TF-IDF). For more details, see the “Supplemental BERTopic: topic representation generation” section and Supplemental Figure 2.

Methodology

The *abstract*, *title*, *publication year*, and *journal* information for the 96,004 original articles were retrieved from the *RheumaLpack* corpus. The number of tokens per abstract was computed to guide the selection of the embedding model. This is crucial because texts that exceed the model's maximum length limit are truncated during the embedding process, leading to a loss of information. Depending on the median token size, two options were considered: (a) to concatenate the title and the abstract, so only a complete and single text for each article is studied and (b) to focus the study solely on abstract information.

Data pre-processing was omitted to preserve the original text structure, which is relevant for transformer-based models to effectively comprehend the context. Hence, stopwords were not omitted. From here onward, the modular approach of BERTopic was applied, with considerations made for each step. For a more detailed explanation of how BERTopic's modularity features were applied, please refer to the "Supplemental BERTopic: methodology followed and application" section.

The number of words extracted per topic was set to 20 (i.e., *top_n_words*), as the optimal number of words in a topic is between 10 and 20. Beyond this range, topics tend to lose coherence. We explored all potential combinations involving two embedding models (i.e., *all-mpnet-base-v2* and *S-PubMedBert-MS-MARCO*), three different dimensionality reduction UMAP initialization states (i.e., seeds 42, 52, and 62), and five cluster minimum size values (i.e., 50, 100, 150, 200, and 250). A total of $2 \times 3 \times 5 = 30$ models were explored.

Two final models were selected for further analysis: one using *all-mpnet-base-v2* and the other using *S-PubMedBert-MS-MARCO*. This selection was based on several criteria, including the number of outliers, the number of topics, and the topic coherence score (i.e., *u_mass*). The chosen models were required to contain fewer than one-third of the total documents classified as outliers ($n < 32,000$), support more than 40 topics, and minimize the *u_mass* score. This score is an intrinsic evaluation method (i.e., measures the quality of the topic model itself without considering any specific external task) that evaluates the quality of a topic based on co-occurrences of word pairs,³¹ which was introduced in the study

conducted by Mimno et al.³² Other coherence measures were calculated (i.e., *c_v*, *c_nmpi*, and *c_uci*) but the final decision was guided by *u_mass*. Afterward, outliers were excluded from the analyses.

After analyzing the keywords and the different topic representations, the topics were labeled through a mutual agreement among DF-N and LR-R authors. (B) Tag was used to identify basic science topics, and (C) tag was used to identify clinical science topics. Word clouds were generated to show the keywords linked to the topics and the topics' distribution. The size of each word is proportional to its relevance to the topic. Hierarchical clustering representations were generated to show how topic embeddings can be combined at various cosine distances. Dynamic TM was employed to explore the evolution of topics over time, using the two selected models.

Eventually, we applied the same methodology described in Karabacak and Margetis³³ to model trends. The publication year and the topic probabilities (i.e., the probability of an abstract being classified under a particular topic based on its content) were retrieved. The mean topic probability per publication year and per topic was computed. Bivariate linear regression models were developed for each topic, with the mean topic probability serving as the dependent variable, and the publication year as the independent variable. By examining the slopes of these regression lines, topics were categorized as hot if they had positive slopes and cold if they had negative slopes.

All models were trained in Google Colab, with a T4 GPU and a high-RAM runtime, using Python.

The reporting of this study conforms to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist³⁴ (Supplemental File PRISMA Checklist). Although this article is not a scoping review per se, we found this checklist to be the most suitable for reporting our work.

Results

The number of articles retrieved per year and journal can be found in Supplemental Table 4. The median number of tokens per abstract was 375 (*Q1*: 287, *Q3*: 442). When combining both the abstract and title, the median was 401 (*Q1*: 310, *Q3*: 471); therefore, we chose to analyze

only the abstract. The number of topics identified by the models ranged from 42 to 296, while the number of initial outliers ranged from 19,075 to 35,332. In Supplemental Table 5, the results of the 30 trained models are shown, including the minimum cluster size, the seed, the number of topics and outliers, and the coherence score values. As the number of topics decreases (and the number of the minimum cluster size increases), the topic coherence scores are better. In Supplemental Excel File Models Output, the topic number, the count, the default topic name, the different topic representations, and the three abstracts that best encapsulate the thematic content of each topic are shown. Supplemental Excel File Top 5 Topics shows the five topics with the highest number of documents for all models.

The model that exhibited the lowest u_{mass} coherence score utilized a minimum cluster size of 250, with seed values of 52 for the *all-mpnet-base-v2* model (-0.279) and 42 for the *S-PubMedBert-MS-MARCO* model (-0.288). A total of 73,736 and 69,316 abstracts were classified into 47 topics and 45 topics for the *all-mpnet-base-v2* and the *S-PubMedBert-MS-MARCO* models, respectively. The remaining documents were classified as outliers and discarded. Tables 1 and 2 present a detailed overview of the topics, outlined by a unique set of keywords that capture their essential themes.

Hierarchical clustering plots and word clouds for the top 10 topics are shown in Supplemental Figures 3 and 4, and 5 and 6, respectively.

Regarding the dynamic modeling of topics, for each model, we studied the themes in batches of 10. Figures 1 and 2, and Supplemental Figures 7 and 8 show the results. Moreover, a bar chart of the hot and cold topics for the two models is displayed in Figures 3 and 4. Finally, a comparison of the topics of the two final models is presented in Supplemental Table 6.

Trends in rheumatology

When comparing the top 10 topics identified in the two models, *all-mpnet-base-v2* and *S-PubMedBert-MS-MARCO*, there is considerable overlap between them. This overlap could lend credibility to the findings. For instance, 8 of the 10 primary topics were consistent across the models, with (C) Knee osteoarthritis, and (C) Rheumatoid arthritis being the most studied

topics. The relevance of (C) Spondyloarthritis, (C) Psoriatic arthritis, (B) Systemic lupus erythematosus, and (C) Osteoporosis topics differ between both models. However, when combining all the topics related to RA and SLE, the number of documents is 13,927 and 5950 for the *all-mpnet-base-v2* model, and 13,297 and 7149 for the *S-PubMedBert-MS-MARCO*. Therefore, globally, the three most studied topics are RA, SLE, and OA.

Some of the topics expected to be found (e.g., (C) COVID-19 and (C) JAK inhibitors) were present after applying dynamic TM, which further strengthens the reliability of the results. Conversely, other unexpected topics such as (C) Spinal surgery or (C) Bone fractures have gained relevance in recent years. As shown in Figures 3 and 4; (C) Gout, (C) Spondyloarthritis, and (C) Psoriatic arthritis are nowadays *hot topics*, whereas (C) Antiphospholipid syndrome, (C) Septic arthritis, or (C) Reactive arthritis are *cold topics*.

As the final number of topics is relatively low, no specific topics related to artificial intelligence (AI) or new statistical learning techniques that became popular a few years ago, such as trajectory analysis, were identified. However, when analyzing models with a higher number of topics such as *all-mpnet-base-v2* (minimum number of cluster: 50, seed: 42), we found the following topics: [learning, machine, algorithms, machine learning, algorithm, ai, deep learning, artificial intelligence, artificial, intelligence]. Something similar occurs with social media data topic [websites, internet, information, social media, readability, search, media, social, google, online], with telemedicine [app, apps, mobile, smartphone, digital, application, care, health, mhealth, patient], and with wearables: [app, apps, mobile, smartphone, digital, application, care, health, mhealth, patient]. A similar situation appears when considering topics related to patients' education, a fundamental issue in rheumatology. Hence, the use of models with a larger number of topics could be useful to identify new emerging trends. See Supplemental Excel File Models Output.

Discussion

We applied BERTopic, an NLP-based TM technique, to explore trends and research themes within rheumatology journals over the past 23 years. Our study uncovered 45 and 47 distinct topics when using the models with the greatest

Table 1. Summary of the topics for the *all-mpnet-base-v2* model.

Topic (custom label)	Count	Keywords
(C) Knee osteoarthritis	7805	knee, oa, osteoarthritis, pain, hip, knee oa, study, joint, patients, cartilage
(C) Rheumatoid arthritis	7409	ra, patients, disease, arthritis, mtx, treatment, rheumatoid, disease activity, rheumatoid arthritis, activity
(B) Osteoarthritis	4936	cartilage, oa, chondrocytes, expression, articular, osteoarthritis, knee, articular cartilage, collagen, joint
(C) Systemic sclerosis	4888	ssc, systemic sclerosis, sclerosis, ssc patients, patients, skin, scleroderma, systemic, sclerosis ssc, patients ssc
(C) Spondyloarthritis	3103	axspa, spondylitis, ankylosing, ankylosing spondylitis, spa, patients, axial, spondyloarthritis, disease, basdai
(C) Vasculitis	2531	vasculitis, anca, aav, gpa, patients, polyangiitis, granulomatosis, anca associated, mpa, associated
(B) Rheumatoid arthritis	2515	il, cells, ra, synovial, arthritis, expression, cell, rheumatoid, mice, induced
(C) Gout	2483	gout, urate, crystals, allopurinol, crystal, uric, uric acid, hyperuricemia, msu, acid
(C) Psoriatic arthritis	2438	psa, psoriatic, psoriasis, psoriatic arthritis, arthritis, patients, arthritis psa, disease, patients psa, psa patients
(C) Polymyositis and dermatomyositis	2208	dm, muscle, myositis, iim, dermatomyositis, anti, pm, patients, ild, jdm
(C) Sjögren's syndrome	1868	pss, ss, sjögren, sjögren syndrome, salivary, syndrome, patients, primary, pss patients, gland
(C) Systemic lupus erythematosus	1802	sle, lupus, damage, disease, systemic lupus, patients, erythematosus, lupus erythematosus, systemic, activity
(C) Osteoporosis	1744	osteoporosis, bone, bmd, fracture, fractures, women, density, mineral, bone mineral, risk
(C) Fibromyalgia	1669	fm, fibromyalgia, pain, fms, patients, fm patients, sleep, symptoms, fiq, fibromyalgia fm
(C) Giant cell arteritis, polymyalgia rheumatica, and Takayasu's arteritis	1664	gca, pmr, arteritis, ta, tak, patients, giant cell, giant, cell arteritis, Takayasu
(C) Juvenile idiopathic arthritis	1646	jia, children, juvenile, idiopathic arthritis, juvenile idiopathic, arthritis, idiopathic, disease, arthritis jia, patients
(C) Antiphospholipid syndrome	1375	aps, apl, antiphospholipid, thrombosis, antiphospholipid syndrome, antibodies, syndrome, thrombotic, patients, syndrome aps
(C) Musculoskeletal pain spinal	1330	pain, lbp, low pain, neck, low, work, musculoskeletal, disability, chronic, study
(C) Joint imaging	1308	mri, joints, synovitis, ra, joint, ultrasound, erosions, imaging, arthritis, doppler
(C) Rheumatoid arthritis cardiovascular risk	1269	ra, risk, cardiovascular, cvd, ra patients, patients, cv, disease, patients ra, rheumatoid
(C) COVID-19	1244	covid, covid 19, 19, vaccination, sars, vaccine, cov, sars cov, patients, pandemic
(C) Spinal surgery	1217	lumbar, spinal, fusion, disc, surgery, group, spine, cervical, scoliosis, screw
(C) Autoinflammatory diseases	1177	fmf, fever, mutations, fmf patients, colchicine, mefv, mediterranean, mediterranean fever, patients, familial mediterranean
(C) Behcet's disease	1157	bd, behçet, behçet disease, bd patients, disease, patients, disease bd, bs, patients bd, involvement
(B) Rheumatoid arthritis genetics	1087	ra, hla, association, polymorphism, drb1, allele, susceptibility, polymorphisms, genetic, genotype

(Continued)

Table 1. (Continued)

Topic (custom label)	Count	Keywords
(C) Systemic lupus erythematosus pregnancy	1025	pregnancy, women, pregnancies, maternal, sle, birth, fetal, lupus, pregnant, outcomes
(B) Systemic lupus erythematosus	815	sle, cells, lupus, cell, sle patients, ifn, expression, mice, il, cd4
(B) Rheumatoid arthritis autoantibodies	786	anti, ccp, anti ccp, ra, rf, citrullinated, aca, antibodies, rheumatoid, positive
(B) Spondyloarthritis genetics	723	hla, b27, hla b27, spa, ibd, spondylitis, ankylosing, ankylosing spondylitis, gut, patients
(C) Musculoskeletal pain shoulder	719	shoulder, cuff, rotator, rotator cuff, pain, shoulder pain, tendon, group, patients, study
(C) Hepatic virus	694	hcv, hbv, hepatitis, infection, hiv, virus, patients, ebv, hepatitis virus, reactivation
(C) Systemic lupus erythematosus neurolupus	636	npsle, sle, neuropsychiatric, cognitive, lupus, np, brain, systemic lupus, patients, erythematosus
(C) Systemic lupus erythematosus kidney	618	renal, ln, nephritis, lupus nephritis, lupus, mmf, proteinuria, patients, class, biopsy
(C) Systemic lupus erythematosus cardiovascular risk	609	sle, sle patients, atherosclerosis, risk, lupus, patients, cardiovascular, patients sle, systemic lupus, factors
(C) Rheumatoid arthritis body mass index	604	ra, bmi, glucocorticoid, glucocorticoids, patients, gc, obesity, rheumatoid, body, cortisol
(C) IgG4-related disease	526	igg4, igg4 rd, rd, igg4 related, related disease, related, serum igg4, disease, disease igg4, patients igg4
(C) Septic arthritis	495	septic, septic arthritis, arthritis, infection, rea, patients, joint, bacterial, cases, chlamydia
(B) Systemic lupus erythematosus genetics	445	sle, allele, polymorphisms, association, susceptibility, polymorphism, gene, lupus, sle patients, controls
(B) miRNA	445	mir, mirnas, expression, mirna, 5p, 3p, oa, 146a, micrnas, mir 146a
(C) Adult-onset Still's disease	374	aosd, adult onset, onset disease, aosd patients, disease aosd, adult, disease, onset, patients aosd, ferritin
(C) Vitamin D	319	vitamin, 25 oh, oh, 25, deficiency, vitamin deficiency, levels, vitamin levels, vdr, serum
(C) Bone tumors	310	tumor, bone, recurrence, tumors, surgical, pvns, cases, diagnosis, case, resection
(C) Bone fractures	288	fractures, fracture, fixation, distal, humeral, humerus, plate, proximal, radius, distal radius
(C) Kawasaki disease	275	kd, ivig, kawasaki, kawasaki disease, coronary, disease kd, coronary artery, children, kd patients, artery
(C) Juvenile idiopathic arthritis-associated uveitis	275	uveitis, jia, ocular, eye, children, associated uveitis, patients, idiopathic, juvenile, arthritis
(C) Biologic drug- associated infections	265	infections, infection, tnf, risk, ra, patients, anti tnf, 95, anti, incidence
(C) Rheumatoid arthritis-associated interstitial lung disease	257	ild, ra ild, ra, lung, lung disease, interstitial, hrct, patients, pulmonary, interstitial lung

topic coherence. These topics represent a wide array of clinical and basic research areas, demonstrating the heterogeneous nature of RMDs. Shifts in emerging and declining areas of focus were also assessed. The three most populated topics, common to both models, were those related to the

clinical aspects of knee OA (kOA), RA, and systemic sclerosis (SSc). The frequent focus on kOA and RA aligns with their significant prevalence and burden, which drives research aimed at improving disease outcomes.^{35–37} SSc, a complex multisystem disorder, continues to garner attention due to

Table 2. Summary of the topics for the *S-PubMedBert-MS-MARCO* model.

Topic (custom label)	Count	Keywords
(C) Knee osteoarthritis	6178	knee, oa, pain, osteoarthritis, hip, knee oa, study, joint, patients, years
(C) Rheumatoid arthritis	5988	ra, patients, disease, mtx, activity, disease activity, arthritis, treatment, remission, rheumatoid
(C) Systemic sclerosis	4866	ssc, patients, systemic sclerosis, sclerosis, ssc patients, ild, skin, systemic, scleroderma, pulmonary
(B) Osteoarthritis	4680	cartilage, oa, chondrocytes, expression, articular, osteoarthritis, articular cartilage, collagen, knee, matrix
(B) Rheumatoid arthritis	4458	ra, il, cells, synovial, expression, arthritis, cell, levels, rheumatoid, mice
(B) Systemic lupus erythematosus	2894	sle, lupus, cells, sle patients, patients, systemic lupus, ln, erythematosus, lupus erythematosus, levels
(C) Vasculitis	2756	vasculitis, anca, aav, gpa, patients, polyangiitis, granulomatosis, associated, anca associated, cytoplasmic
(C) Gout	2383	gout, urate, crystals, allopurinol, uric, uric acid, crystal, hyperuricemia, msu, acid
(C) Osteoporosis	2283	bone, osteoporosis, bmd, fracture, fractures, density, mineral, bone mineral, women, risk
(C) Polymyositis and dermatomyositis	2196	dm, myositis, muscle, iim, anti, dermatomyositis, pm, patients, jdm, ild
(C) Spondyloarthritis	1951	axspa, spa, spondylitis, ankylosing, ankylosing spondylitis, axial, spondyloarthritis, mri, patients, disease
(C) Giant cell arteritis, polymyalgia rheumatica, and Takayasu's arteritis	1786	gca, pmr, arteritis, ta, patients, tak, giant, giant cell, cell arteritis, takayasu
(C) Sjögren's syndrome	1767	pss, ss, sjögren, sjögren syndrome, salivary, syndrome, primary, patients, pss patients, gland
(C) Systemic lupus erythematosus	1598	sle, lupus, damage, systemic lupus, disease, patients, erythematosus, lupus erythematosus, systemic, activity
(C) Fibromyalgia	1531	fm, fibromyalgia, pain, fms, patients, fm patients, sleep, symptoms, fiq, fibromyalgia fm
(C) Musculoskeletal pain spinal	1481	pain, lbp, low pain, musculoskeletal, neck, low, work, disability, chronic, health
(C) Joint imaging	1416	mri, synovitis, joints, joint, ultrasound, ra, imaging, arthritis, doppler, clinical
(C) Antiphospholipid syndrome	1386	aps, apl, antiphospholipid, thrombosis, antibodies, antiphospholipid syndrome, syndrome, patients, acl, igg
(C) Juvenile idiopathic arthritis	1199	jia, children, juvenile, idiopathic arthritis, juvenile idiopathic, arthritis, disease, idiopathic, uveitis, arthritis jia
(C) Autoinflammatory diseases	1175	fmf, fever, mutations, fmf patients, colchicine, mefv, mediterranean, patients, mediterranean fever, mutation

(Continued)

Table 2. (Continued)

Topic (custom label)	Count	Keywords
(C) Rheumatoid arthritis cardiovascular risk	1158	ra, risk, cardiovascular, cvd, ra patients, patients, cv, disease, patients ra, rheumatoid
(C) Behcet's disease	1139	bd, behçet, behçet disease, bd patients, disease, patients, disease bd, bs, patients bd, involvement
(C) Spinal surgery	1030	lumbar, fusion, spinal, surgery, group, disc, cervical, screw, patients, postoperative
(C) Knee osteoarthritis treatment	1008	pain, knee, nsaid, placebo, oa, injection, nsaid, mg, osteoarthritis, celecoxib
(C) COVID-19	956	covid, covid 19, 19, sars, cov, sars cov, vaccination, pandemic, patients, vaccine
(C) Systemic lupus erythematosus kidney	898	ln, renal, lupus, nephritis, lupus nephritis, mmf, patients, sle, proteinuria, treatment
(C) Bone tumors	891	bone, case, diagnosis, tumor, rare, cases, old, year old, patient, report
(C) Psoriatic arthritis	884	psa, psoriasis, psoriatic, psoriatic arthritis, arthritis, arthritis psa, disease, patients, patients psa, psa patients
(B) Rheumatoid arthritis autoantibodies	852	anti, ccp, anti ccp, acpa, ra, citrullinated, rf, antibodies, rheumatoid, positive
(B) Rheumatoid arthritis genetics	841	ra, polymorphism, allele, association, polymorphisms, genotype, hla, susceptibility, gene, drb1
(C) Systemic lupus erythematosus neurolypus	507	npsle, sle, neuropsychiatric, cognitive, np, brain, lupus, patients, systemic lupus, erythematosus
(C) Biologic drug-associated infection	481	tnf, anti tnf, infections, anti, cancer, risk, infliximab, alpha, etanercept, necrosis factor
(C) JAK inhibitors	469	tofacitinib, baricitinib, mg, jak, placebo, mtx, ra, filgotinib, safety, patients
(C) Adult-onset Still's disease	462	aosd, mas, adult onset, onset disease, disease aosd, sjia, aosd patients, onset, adult, ferritin
(C) IgG4-related disease	441	igg4, igg4 rd, rd, igg4 related, related disease, related, serum igg4, disease igg4, patients igg4, disease
(C) Systemic lupus erythematosus cardiovascular risk	420	sle, atherosclerosis, sle patients, risk, lupus, carotid, patients, cardiovascular, patients sle, factors
(C) Septic arthritis	416	septic, septic arthritis, arthritis, infection, cases, tuberculosis, joint, diagnosis, patients, case
(B) Spondyloarthritis genetics	385	hla, b27, hla b27, spa, spondylitis, ankylosing, ankylosing spondylitis, controls, hla 27, gut
(C) Reactive arthritis	356	rea, arthritis, reactive arthritis, lyme, reactive, infection, chlamydia, trachomatis, lyme arthritis, patients
(C) Pregnancy	338	pregnancy, women, pregnancies, birth, ra, maternal, disease, trimester, pregnant, postpartum

(Continued)

Table 2. (Continued)

Topic (custom label)	Count	Keywords
(C) Kawasaki disease	309	kd, ivig, kawasaki, kawasaki disease, coronary, disease kd, coronary artery, children, kd patients, artery
(C) Systemic lupus erythematosus pregnancy	297	pregnancy, sle, women, pregnancies, lupus, maternal, fetal, birth, women sle, preterm
(C) Systemic lupus erythematosus gastrointestinal	277	lupus, sle, erythematosus, lupus erythematosus, systemic lupus, systemic, abdominal, case, erythematosus sle, abdominal pain
(C) Influenza and vaccination	271	vaccination, vaccine, influenza, pneumococcal, vaccines, patients, antibody, response, h1n1, vaccinated
(C) Systemic lupus erythematosus antimalarials	258	hcq, hydroxychloroquine, hydroxychloroquine hcq, sle, patients, lupus, dose, use, adherence, retinopathy
JAK, Janus kinase.		

Topics over Time mpnet_52_250_0-9

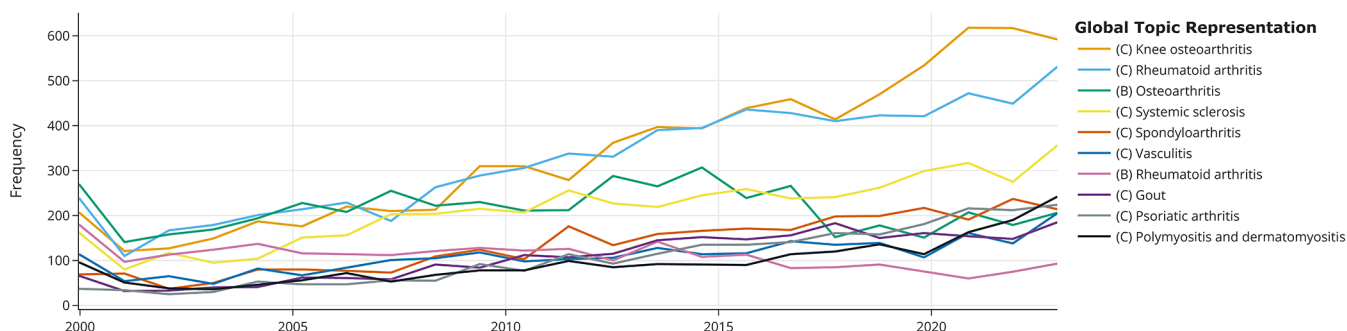


Figure 1. Dynamic topic modeling of the best *all-mpnet-base-v2* model.

recent advances in diagnosis and therapeutic approaches, including the development of targeted treatment strategies.³⁸

Our analysis of research trends reveals a significant shift in publication priorities. The “hottest” topics (clinical research in kOA, COVID-19, and spinal surgery) offer further insight into current clinical priorities. The increasing attention on kOA may be related to a growing interest in the study of risk factors, and other clinical aspects of this condition, considering the aging population and the rising incidence of degenerative joint diseases.³⁹ COVID-19, as expected, emerged as a

significant focus, reflecting its impact on RMDs, including how the pandemic has influenced the management of immunosuppressive therapies in rheumatology.⁴⁰

The rise in “spinal surgery”-related research corresponds to the significant burden of low back pain (LBP), the leading cause of disability.⁴¹ Degenerative spine conditions, such as disc degeneration, lumbar stenosis, and spondylolisthesis, are among the main causes for LBP⁴² and spinal surgeries,⁴³ despite limited evidence supporting its use for degenerative LBP, except in cases involving radiculopathy, neurogenic

Topics over Time pubmed_42_250_0-9

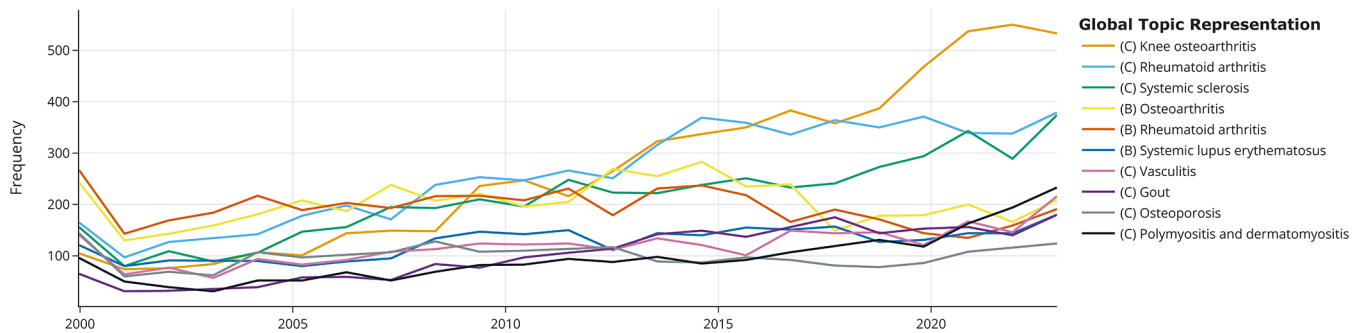


Figure 2. Dynamic topic modeling of the best *S-PubMedBert-MS-MARCO* model.

claudication, cancer, or infection.⁴⁴ Nonetheless, studies have shown an increase in surgical interventions for LBP.^{44,45}

Conversely, certain topics have cooled off, such as antiphospholipid syndrome (APS), and OA and RA basic research. This decline could signal that the foundational understanding of these areas has stabilized, or that fewer groundbreaking developments have emerged recently. For example, APS treatments have not seen significant breakthroughs in recent years.⁴⁶

Regarding RA and OA basic research, the substantial progress made in understanding their pathophysiology may have shifted the research focus toward translating this knowledge into clinical applications, such as the development of new therapies for RA.⁴⁷ For OA, research emphasis may have moved toward prevention strategies, non-pharmacological management, and personalized medicine approaches.^{48–50}

The use of TM techniques on PubMed abstracts is not new. These methods have been used in different medical fields for trend analysis and for uncovering hidden topics over the past few years. For example, Sperandeo *et al.*⁵¹ evaluated the usage of “personality” and “mental health” terms within the titles and abstracts of articles published in PubMed from 2012 to 2017. The researchers employed LDA on more than 7500 abstracts and found 30 topics organized in 8 hierarchical clusters, concluding that personality is linked to a broad spectrum of conditions. The suitable number of clusters was determined using a five-fold cross-validation approach.

Tighe *et al.*⁵² applied TM on a corpus of more than 200,000 abstracts related to pain. The abstracts collected, retrieved through searches using the “pain” [MeSH] term, corresponded to articles published between 1949 and 2017. On this occasion, both LDA and latent semantic indexing techniques were employed. After following a topic coherence strategy, the researchers identified an optimal topic count of 40. One of the conclusions of this research was that TM can help identify critical research avenues by evaluating the gaps in the literature concerning a specific topic.

On their behalf, Abba *et al.*⁵³ focused on the use of TM techniques to uncover hidden topics from 100 years of peer-reviewed hypertension publications (*i.e.*, 1900–2018). LDA was applied to more than 580,000 abstracts. Most of the identified topics, $n=20$, fell into four distinct categories: preclinical, epidemiology, complications, and treatment-related studies. Topic trends were evaluated by calculating the annual proportion of abstracts for each topic relative to the cumulative total of articles associated with that topic.

Shi *et al.*⁵⁴ examined AI-related studies published in PubMed, from 2000 to 2022, to highlight the current situation of medical AI research and to provide insights into its future developments. With that aim, scholars downloaded metadata from 307,000 articles (*e.g.*, title, abstract, journals, authors) and applied LDA to titles and abstracts. They divided the data into intervals of 5 years, performing unique TM for each period. The authors presented the five main topics in eight different domains of AI. These domains

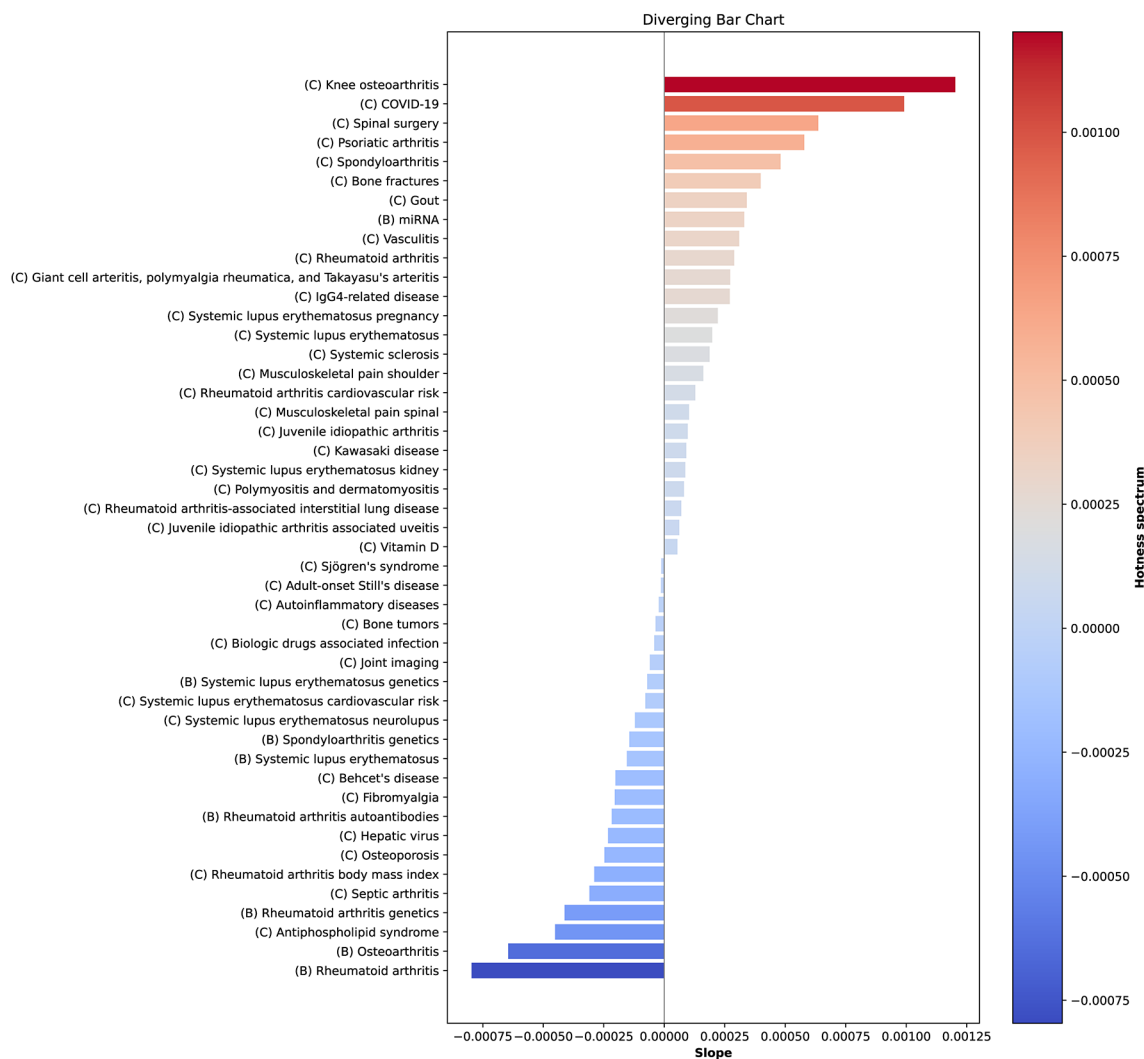


Figure 3. Bar chart of hot and cold topics. *All-mpnet-base-v2* model.

were described by the European Commission Joint Research Centre.

Depression, anxiety, and burnout in academia were studied using BERTopic.⁵⁵ The authors extracted 2846 abstracts from PubMed ranging from 1975 to 2023 using a complex query that did not include MeSH terms. Afterward, the authors compared BERTopic models with different sets of parameters, each of them being run three times. The best model was chosen based on different criteria (i.e., proportion of outliers, topic interpretability, topic coherence, and diversity); this model comprised 27 topics. After studying their evolution, the authors showed, among

others, how the COVID-19 pandemic influenced the burnout of medical professionals.

Eventually, Grubbs et al.⁵⁶ studied the topics present in a specific academic journal—*Gynecologic Oncology*—over a 30-year period (i.e., 1990–2020), as well, as the interest in them over time. With that aim, they used LDA on 11,200 abstracts and determined the number of topics using the coherence score. The best model contained 26 topics, and 3 of them were merged after manual assessment by 3 reviewers. Thanks to the experiments carried out, researchers could hypothesize the evolution of some topics related to oncology gynecology for the next

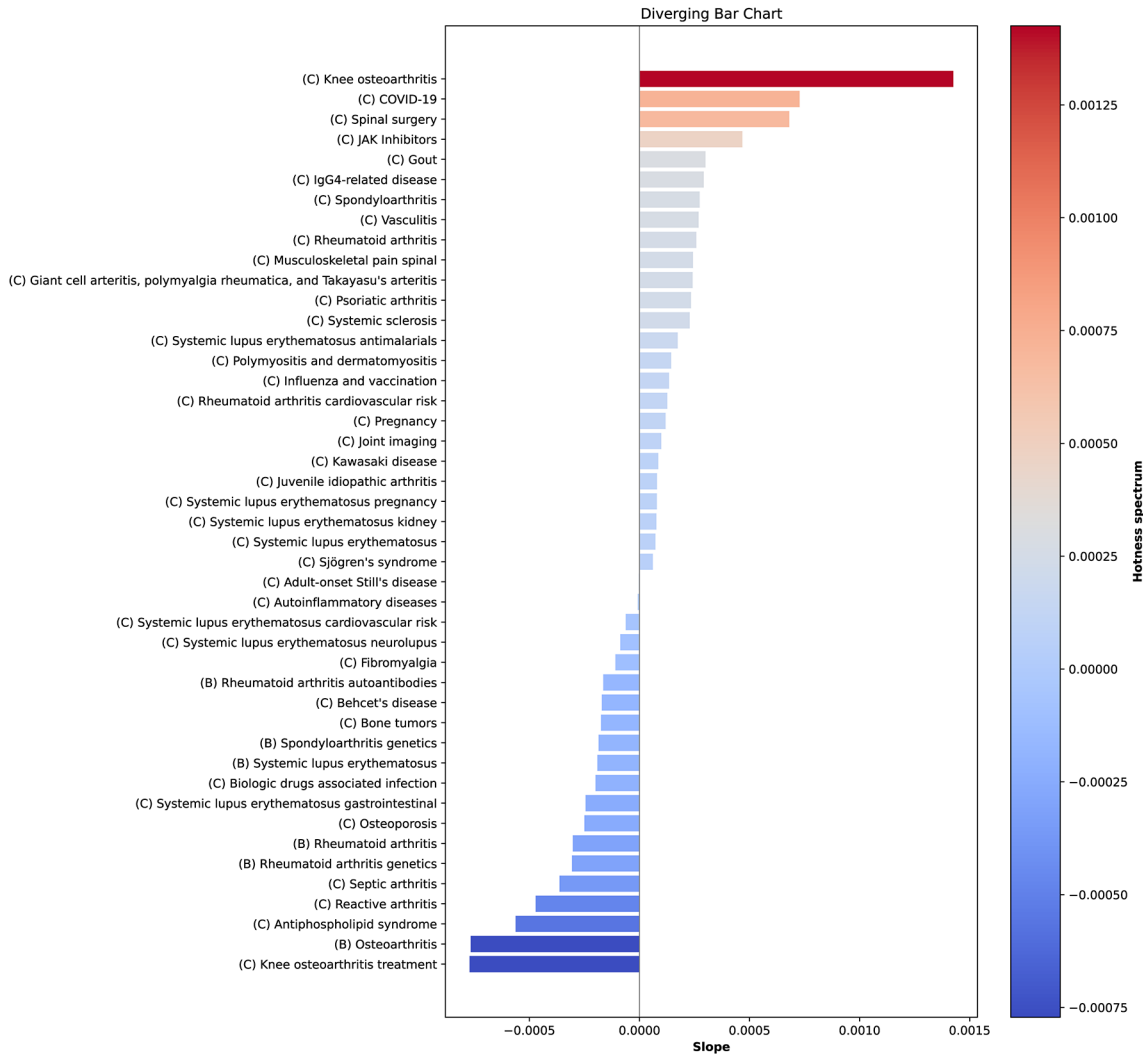


Figure 4. Bar chart of hot and cold topics. *S-PubMedBert-MS-MARCO* model.

years, such as an increase in surgical topics and epidemiological and health outcomes research topics; and a decrease in chemotherapy and radiation.

As can be seen from the above studies, there is a real interest in uncovering latent topics in medical documentation. In this study, we have demonstrated how dynamic TM can be applied to abstracts indexed in PubMed, and published in *Rheumatology* journals from 2000 to 2023.

To the best of our knowledge, the BERTopic approach has not been previously applied to examine trends within this medical field. A potentially more intriguing application of dynamic TM would involve its use with EHR data, to

characterize the natural history of diseases. This approach was taken a few years ago, but applying LDA over AAV histories.²²

Furthermore, each clinical note could be categorized into specific topics. Should there be a requirement for a manual review of the record contents, pre-classifying them by topic could assist physicians in assembling patient cohorts for targeted studies.

Finally, these models could be used as recommendation systems to direct unpublished scientific articles to the journal that maximizes their likelihood of publication based on the latent topics contained in the abstract and other structured data (e.g., year, affiliation of the first author).

Strengths and limitations

The study's findings can offer practical insights for healthcare and policy. By identifying evolving research trends, this analysis allows the assessment of whether rheumatology research aligns with global health priorities, such as those in the Global Burden of Disease study. This alignment could help optimize resource allocation and support efficient policymaking for prevalent, high-cost diseases. The "hot" and "cold" topic analysis can further guide research funding toward areas with significant clinical potential. In addition, understanding these trends enables healthcare systems to anticipate future demands, particularly for conditions like OA, informing public health strategies and ultimately enhancing patient outcomes and healthcare efficiency. However, this study has limitations. Our analysis window begins in 2000, missing the early evolution of biological agents introduced in 1999. TM carries subjectivity, and BERTopic assumes each document covers only one topic, potentially overlooking multifaceted articles. Limiting our dataset to rheumatology journals ensures focused results but may exclude interdisciplinary perspectives. While a broader journal scope could offer a fuller view, varying editorial standards may also complicate trend analysis. Research trends are not correlated with other metrics, such as patents or clinical trials. In addition, patient education was not a predominant theme despite its complementary role in disease management, especially for conditions like fibromyalgia, where programs promoting patient-centered care and self-management (e.g., "Amigos de Fibro"⁵⁷) can significantly improve quality of life. Finally, by limiting the corpus to rheumatology-specific journals, our thematic analysis may primarily reflect traditional or well-established themes within rheumatology, potentially overlooking emerging interdisciplinary perspectives that appear in broader medical or life sciences journals. However, considering that a substantial number of papers within each topic is critical for creating well-defined themes in TM, including isolated or infrequent papers from non-specialized journals would likely have resulted in outliers rather than cohesive topics, thus compromising the robustness of the analysis. Furthermore, research with significant interdisciplinary impact, such as COVID-19 studies, often migrates into rheumatology journals as its relevance to the field becomes apparent. Therefore, we believe this approach provides a balanced view of core research trends while capturing key interdisciplinary developments.

Conclusion

To our knowledge, this is the first study that uses BERTopic, and dynamic TM to identify the key topics in rheumatology research using a set of abstracts extracted from PubMed. The two-sentence embedding models employed provided similar results, highlighting the dynamic and varied nature of rheumatology research and illustrating how interest in certain topics has shifted over time. As the number of scientific publications increases, the use of NLP techniques will be necessary to efficiently analyze and synthesize information, helping to identify trends, gaps, and emerging areas of interest across various medical fields.

Declarations

Ethics approval and consent to participate

Our study did not require approval from an ethics board because it used information that is freely accessible in the public domain and not subject to privacy protections.

Consent for publication

Not applicable.

Author contributions

Alfredo Madrid-García: Conceptualization; Data curation; Formal analysis; Methodology; Software; Visualization; Writing – original draft; Writing – review & editing.

Dalifer Freitas-Núñez: Investigation; Writing – review & editing.

Beatriz Merino-Barbancho: Conceptualization; Software; Writing – original draft.

Inés Pérez Sancristobal: Investigation; Writing – review & editing.

Luis Rodríguez-Rodríguez: Investigation; Supervision; Writing – review & editing.

Acknowledgements

The authors would like to thank professors Anselmo Peñas and Alejandro Rodríguez González for their feedback and guidance.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Instituto de Salud Carlos III, Ministry of Health, Madrid, Spain (RD21/002/0001). The sponsor or funding organization

had no role in the design or conduct of this research.





Competing interests

The authors declare that there is no conflict of interest.

Availability of data and materials

All data used in this manuscript are available online at <https://pubmed.ncbi.nlm.nih.gov/>. Data processing is described in Madrid-García *et al.* Further inquiries can be directed to the corresponding author.

ORCID iDs

Alfredo Madrid-García  <https://orcid.org/0000-0002-1591-0467>
 Dalifer Freites-Núñez  <https://orcid.org/0000-0002-0966-2778>
 Beatriz Merino-Barbancho  <https://orcid.org/0000-0001-5070-4178>
 Luis Rodríguez-Rodríguez  <https://orcid.org/0000-0002-2869-7861>

Supplemental material

Supplemental material for this article is available online.

References

- Thelwall M and Sud P. Scopus 1900–2020: growth in articles, abstracts, countries, fields, and journals. *Quant Sci Stud* 2022; 3: 37–50.
- Bornmann L, Haunschild R and Mutz R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanit Soc Sci Commun* 2021; 8: 1–15.
- Olsen NJ and Stein CM. New drugs for rheumatoid arthritis. *N Engl J Med* 2004; 350: 2167–2179.
- Smolen JS. Insights into the treatment of rheumatoid arthritis: a paradigm in medicine. *J Autoimmun* 2020; 110: 102425.
- Kerrigan SA and McInnes IB. Reflections on “older” drugs: learning new lessons in rheumatology. *Nat Rev Rheumatol* 2020; 16: 179–183.
- van Vollenhoven R. Treat-to-target in rheumatoid arthritis—are we there yet? *Nat Rev Rheumatol* 2019; 15: 180–186.
- Kyu HH, Abate D, Abate KH, *et al.* Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018; 392: 1859–1922.
- James SL, Abate D, Abate KH, *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018; 392: 1789–1858.
- Burgers LE, Raza K, Van Der Helm-Van AH, *et al.* Window of opportunity in rheumatoid arthritis—definitions and supporting evidence: from old to new perspectives. *RMD Open* 2019; 5: e000870.
- van Steenbergen HW, Aletaha D, de Voorde LJJ, *et al.* EULAR definition of arthralgia suspicious for progression to rheumatoid arthritis. *Ann Rheum Dis* 2017; 76: 491–496.
- Van Der Heijde D, Van Der Helm-Van AHM, Aletaha D, *et al.* EULAR definition of erosive disease in light of the 2010 ACR/EULAR rheumatoid arthritis classification criteria. *Ann Rheum Dis* 2013; 72: 479–481.
- Nagy G, Roodenrijs NMT, Welsing PMJ, *et al.* EULAR definition of difficult-to-treat rheumatoid arthritis. *Ann Rheum Dis* 2021; 80: 31–35.
- Churchill R and Singh L. The evolution of topic modeling. *ACM Comput Surv* 2022; 54: 1–35.
- Abdelrazek A, Eid Y, Gawish E, *et al.* Topic modeling algorithms and applications: a survey. *Inf Syst* 2023; 112: 102131.
- Blei DM. Introduction to probabilistic topic models. *Commun ACM* 2011; 55: 77–84.
- Blei DM. Probabilistic topic models. *Commun ACM* 2012; 55: 77–84.
- Jelodar H, Wang Y, Yuan C, *et al.* Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl* 2019; 78: 15169–15211.
- Boyd-Graber J, Hu Y, Mimno D, *et al.* Applications of topic models. *Found Trends Inf Retr* 2017; 11: 143–296.
- Mulunda CK, Wagacha PW and Muchemi L. Review of trends in topic modeling techniques, tools, inference algorithms and applications. In: *2018 5th international conference on soft computing*

- & machine intelligence (ISCM), Nairobi, Kenya, 2018, pp. 28–37. New York, NY: IEEE.
20. Valeri M. Organizational phenomenon. In: Valeri M (ed.) *Organizational studies: implications for the strategic management*. Cham: Springer International Publishing, pp. 1–17.
 21. Tedeschi SK, Cai T, He Z, et al. Classifying pseudogout using machine learning approaches with electronic health record data. *Arthritis Care Res (Hoboken)* 2021; 73: 442–448.
 22. Wang L, Miloslavsky E, Stone JH, et al. Topic modeling to characterize the natural history of ANCA-associated vasculitis from clinical notes: a proof of concept study. *Semin Arthritis Rheum* 2021; 51: 150–157.
 23. Dzubur E, Khalil C, Almario CV, et al. Patient concerns and perceptions regarding biologic therapies in ankylosing spondylitis: insights from a large-scale survey of social media platforms. *Arthritis Care Res (Hoboken)* 2019; 71: 323–330.
 24. Li JX and Yacyshyn E. Thoughts and experiences of Behçet disease from participants on a Reddit Subforum: qualitative online community analysis. *JMIR Form Res* 2023; 7: e49380.
 25. Tang J, Weisenfeld D, Dahal K, et al. The “topics” in the electronic health record of rheumatoid arthritis patients before initiating targeted therapies and association with future treatment course. *Arthritis Rheumatol* 2023; 75(Suppl. 9): 884–886.
 26. Flurie M, Converse M, Parker C, et al. Understanding community perspectives on disease management: a social media analysis of gout care strategies. *Arthritis Rheumatol* 2023; 75(Suppl. 9): 2356–2357.
 27. Eaneff S, Vaughan T, Barut V, et al. How do patients describe their “new normal” in systemic lupus erythematosus? Use of probabilistic topic modelling to characterize patients’ experiences recorded in an online health community. *Arthritis Rheumatol* 2018; 70(Suppl. 9): 2615–2616.
 28. Sperl L, Stamm T, Andrews MR, et al. OP0214-HPR educational needs among health professionals in rheumatology: low awareness of EULAR offerings and unfamiliarity with course content as a major barrier a EULAR funded European survey. *Ann Rheum Dis* 2022; 81: 139–140.
 29. Madrid-García A, Merino-Barbancho B, Freites-Núñez D, et al. From web to rheumalpack: creating a linguistic corpus for exploitation and knowledge discovery in rheumatology. *Comput Biol Med* 2024; 179: 108920.
 30. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure, <https://arxiv.org/abs/2203.05794> (2022, accessed 1 October 2024).
 31. Rosner F, Hinneburg A, Röder M, et al. Evaluating topic coherence measures. arXiv preprint arXiv:14036397, (2014, accessed 1 September 2024).
 32. Mimno D, Wallach H, Talley E, et al. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, Edinburgh, 2011, pp. 262–272. Stroudsburg, PA: Association for Computational Linguistics.
 33. Karabacak M and Margetis K. Natural language processing reveals research trends and topics in the spine journal over two decades: a topic modeling study. *Spine J* 2024; 24: 397–405.
 34. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018; 169: 467–473.
 35. Langworthy M, Dasa V and Spitzer AI. Knee osteoarthritis: disease burden, available treatments, and emerging options. *Ther Adv Musculoskelet Dis* 2024; 16: 1759720X241273009. DOI: 10.1177/1759720X241273009.
 36. Brown P, Pratt AG and Hyrich KL. Therapeutic advances in rheumatoid arthritis. *BMJ* 2024; 384: e070856.
 37. Di Matteo A, Bathon JM and Emery P. Rheumatoid arthritis. *Lancet* 2023; 402: 2019–2033.
 38. Bukiri H and Volkmann ER. Current advances in the treatment of systemic sclerosis. *Curr Opin Pharmacol* 2022; 64: 102211.
 39. Giorgino R, Albano D, Fusco S, et al. Knee osteoarthritis: epidemiology, pathogenesis, and mesenchymal stem cells: what else is new? An update. *Int J Mol Sci* 2023; 24: 6405.
 40. D’Silva KM and Wallace ZS. COVID-19 and disease-modifying anti-rheumatic drugs. *Curr Rheumatol Rep* 2021; 23: 28.
 41. Ferreira ML, de Luca K, Haile LM, et al. Global, regional, and national burden of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *Lancet Rheumatol* 2023; 5: e316–e329.
 42. Knezevic NN, Candido KD, Vlaeyen JWS, et al. Low back pain. *Lancet* 2021; 398: 78–92.

43. Chou R, Baisden J, Carragee EJ, et al. Surgery for low back pain. *Spine (Phila Pa 1976)* 2009; 34: 1094–1109.
44. Evans L, O'Donohoe T, Morokoff A, et al. The role of spinal surgery in the treatment of low back pain. *Med J Aust* 2023; 218: 40–45.
45. Jensen RK, Schiøttz-Christensen B, Skovsgaard CV, et al. Surgery rates for lumbar spinal stenosis in Denmark between 2002 and 2018: a registry-based study of 43,454 patients. *Acta Orthop* 2022; 93: 488–494.
46. Ambati A, Knight JS and Zuo Y. Antiphospholipid syndrome management: a 2023 update and practical algorithm-based approach. *Curr Opin Rheumatol* 2023; 35: 149–160.
47. Lin Y-J, Anzaghe M and Schülke S. Update on the pathomechanism, diagnosis, and treatment options for rheumatoid arthritis. *Cells* 2020; 9: 880.
48. Whittaker JL, Runhaar J, Bierma-Zeinstra S, et al. A lifespan approach to osteoarthritis prevention. *Osteoarthritis Cartilage* 2021; 29: 1638–1653.
49. Gibbs AJ, Gray B, Wallis JA, et al. Recommendations for the management of hip and knee osteoarthritis: a systematic review of clinical practice guidelines. *Osteoarthritis Cartilage* 2023; 31: 1280–1292.
50. Mobasheri A, Thudium CS, Bay-Jensen A-C, et al. Biomarkers for osteoarthritis: current status and future prospects. *Best Pract Res Clin Rheumatol* 2023; 37: 101852.
51. Sperandeo R, Messina G, Iennaco D, et al. What does personality mean in the context of mental health? A topic modeling approach based on abstracts published in PubMed over the last 5 years. *Front Psychiatry* 2020; 10: 449078.
52. Tighe PJ, Sannapaneni B, Fillingim RB, et al. Forty-two million ways to describe pain: topic modeling of 200,000 PubMed pain-related abstracts using natural language processing and deep learning-based text generation. *Pain Med* 2020; 21: 3133–3160.
53. Abba M, Nduka C, Anjorin S, et al. One hundred years of hypertension research: topic modeling study. *JMIR Form Res* 2022; 6: e31292.
54. Shi J, Bendig D, Vollmar HC, et al. Mapping the bibliometrics landscape of AI in medicine: methodological study. *J Med Internet Res* 2023; 25: e45815.
55. Lezhnina O. Depression, anxiety, and burnout in academia: topic modeling of PubMed abstracts. *Front Res Metr Anal* 2023; 8: 1271385.
56. Grubbs AE, Sinha N, Garg R, et al. Use of topic modeling to assess research trends in the journal Gynecologic Oncology. *Gynecol Oncol* 2023; 172: 41–46.
57. Antunes MD, da Rocha Loures FCN, de Souza IMB, et al. A web-based educational therapy intervention associated with physical exercise to promote health in fibromyalgia in Brazil: the Amigos De Fibro (Fibro Friends) study protocol. *Trials* 2023; 24: 655.

Visit Sage journals online
journals.sagepub.com/
home/tab

 Sage journals