
TWO DECADES OF RHEUMATOLOGY RESEARCH (2000-2023): A DYNAMIC TOPIC MODELING PERSPECTIVE

A PREPRINT

 **Alfredo Madrid-García***[†]

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

 **Dalifer Freites-Núñez**

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

 **Luis Rodríguez-Rodríguez**

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

June 6, 2024

ABSTRACT

Background: Rheumatology has experienced notably changes in last decades. New drugs, including biologic agents and janus kinase inhibitors, have blossomed. Concepts such as *window of opportunity*, *arthralgia suspicious for progression*, or *difficult-to-treat rheumatoid arthritis* 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 characterise the research landscape and advances in rheumatology, automatic and efficient approaches based on natural language processing should be used. The objective of this study is to use topic modeling techniques to uncover key topics and trends in the rheumatology research conducted in the last 23 years.

Methods: This study analysed 96,004 abstracts published between 2000 and December 31, 2023, drawn from 34 specialised rheumatology journals obtained from PubMed. BERTopic, a novel topic modeling 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 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 rheumatoid arthritis, systemic lupus erythematosus and osteoarthritis. Expected topics such as COVID-19 or JAK inhibitors were identified after conducting the dynamic topic modeling. Topics such as spinal surgery or bone fractures have gained relevance in last years, however, antiphospholipid syndrome, or septic arthritis have lost momentum.

Conclusions: Our study utilized advanced natural language processing techniques to analyse the rheumatology research landscape, and identify key themes and emerging trends. The results highlight

*First author

†Corresponding author

the dynamic and varied nature of rheumatology research, illustrating how interest in certain topics have shifted over time.

Keywords Artificial intelligence · Natural language processing · PubMed · BERTopic · Topic modeling · Trend analysis · Transformers

1 Introduction

Over the past decades the volume of academic literature has experienced significant growth Thelwall and Sud [2022], Bornmann et al. [2021]. The field of rheumatic and musculoskeletal diseases (RMDs) has not been immune to this growth, Figure 1. Moreover, 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 inhibitors Olsen and Stein [2004], Smolen [2020], Kerrigan and McInnes [2020]. Furthermore, the adoption of therapeutic strategies, such as treat-to-target van Vollenhoven [2019], the earlier initiation of disease-modifying treatments, or the paradigm shift in how diseases are analysed, not only by their mortality rate, but also by their disability, propitiated a new scenario for rheumatic and musculoskeletal conditions Kyu et al. [2018], James et al. [2018]. Concepts such as the window of opportunity Burgers et al. [2019], arthralgia suspicious for progression van Steenberghe et al. [2017], erosive disease Van Der Heijde et al. [2013], or difficult-to-treat rheumatoid arthritis Nagy et al. [2021] have gained momentum.

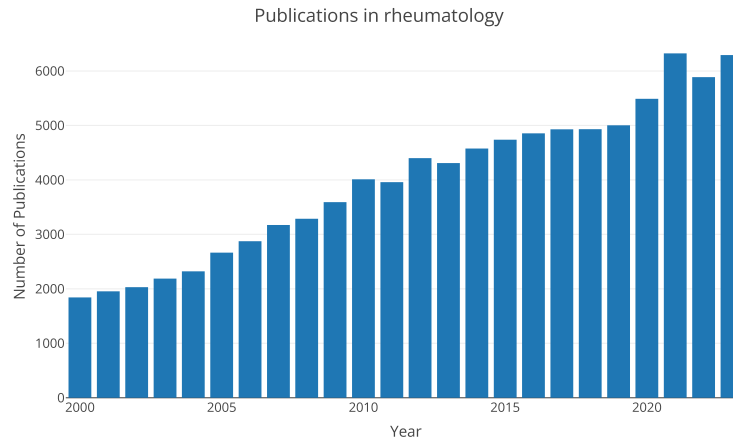


Figure 1: Number of rheumatology-related publications until December 31st 2023

In this context of continuous change, we hypothesise 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 RMDs 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 techniques to characterise 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 Churchill and Singh [2022], Abdelrazek et al. [2023]. 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 structure Blei [2011, 2012]. The outcome of a typical topic modeling 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 arose (i.e. BERTopic).

Consequently, this research study seeks to address the question: *How has rheumatology research evolved in recent years?* To do so, we employ BERTopic to uncover trends related to rheumatology and to explore the publication landscape of rheumatology research within the scientific literature over the past two decades.

2 State of the art

TM has been used in a multitude of fields, including social networks, software engineering, crime science, political science, geography, medicine, and linguistics Jelodar et al. [2019]. Additionally, it has proven effective in analyzing historical documents such as newspapers and humanistic texts Boyd-Graber et al. [2017], as well as in educational research Mulunda et al. [2018], and the study of organizational phenomena Valeri [2021].

TM has been widely applied in rheumatology research.

The authors in Tedeschi et al. [2021] employed a TM approach, sureLDA, followed by penalised regression, to predict pseudogout probability in large datasets. TM was also applied to characterise the temporal evolution of ANCA-associated Vasculitis (AAV) in Wang et al. [2021]. Temporal trends, in more than 113,000 clinical notes, before and after the treatment initiation date for a diagnosis of AAV, were modelled 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, Dzubur et al. [2019], explored the application of TM to understand the concerns and perceptions of patients with ankylosing spondylitis regarding biologic 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, in Li and Yacyshyn [2023], scholars analysed the posts published over the course of 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*.

In noa [2023a], the authors pursue to uncover the themes present in the Electronic Health Record (EHR) of patients with rheumatoid arthritis (RA) prior to the start of targeted treatments, and to explore their relationship with the subsequent course of treatment. On the other hand, in noa [2023b] the authors evaluated two social media communities, a Facebook group, and a public subreddit, ‘r/gout’, identified 30 topics and conduct sentiment analysis.

Moreover, investigators in noa [2018] characterised systemic lupus erythematosus (SLE) patients’ experiences in an online health community by applying LDA in free text data extracted from *PatientsLikeMe* community.

Eventually, in Sperl et al. [2022], LDA was applied to analyze responses to open-ended questions from an online survey designed to assess motivations among health professionals for participating in post-graduate rheumatology education; and to identify barriers and facilitators for participation in current EULAR educational offerings.

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

Table 1: Rheumatology studies in which topic modeling has been employed. EHR: Electronic Health Record. LDA: Latent Dirichlet Allocation

Author	Disease	Data source	Number of data points / patients	Topics	Algorithm
Sara K Tedeschi et al. (2021) Tedeschi et al. [2021]	Pseudogout	EHR	30,089 patients	-	sureLDA
Liqin Wang et al. (2021) Wang et al. [2021]	Vasculitis	EHR	113,000 notes	90	LDA
Eldin Dzubur et al. (2019) Dzubur et al. [2019]	Ankylosing spondylitis	Social media data	27,416 posts from 601 social media sites	112	LDA
Jenny Xiaoyu Li et al. Li and Yacyshyn [2023] (2023)	Behçet	Social media data (Reddit)	196 threads	6	-
Jason Tang et al. (2023) noa [2023a]	Rheumatoid arthritis	EHR	1,102 patients	4	LDA
Maurice Flurie et al. (2023) noa [2023b]	Gout	Social media data (Facebook and Reddit)	175,000 posts comments	30	-
Stephanie Eaneff et al. (2023) Eaneff et al. [2018]	Systemic lupus erythematosus	Online community	138,409 posts	150	LDA
L. Sperl et al. (2022) Sperl et al. [2022]	-	Online survey answers	667 answers	-	-

3 Materials and Methods

3.1 Materials

Data from the *RheumaLpack* corpus Madrid et al. [2024], 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. These abstracts were compiled from original articles indexed in PubMed from January 1, 2000, to December 31, 2023; and came from 34 rheumatology-specific journals, as identified by the Journal Citation Reports (JCR), see Supplementary Tables 1 and 2. R's *rentrez* library was used to collect the data.

BERTopic was used for topic modeling Grootendorst [2022]. This technique generates topic representations through three steps:

- **Document embeddings:** Unlike LDA, a probabilistic topic modeling approach, BERTopic utilizes pre-trained language models to create representations that can be compared semantically. Therefore, clusters of semantically similar documents, abstracts, are created.
- **Document clustering:** to overcome the *curse of dimensionality*, the dimensionality of document embeddings generated in previous step is reduced. Uniform Manifold Approximation and Projection (UMAP) algorithm is commonly employed for that purpose. After that, the reduced embeddings are clustered using HDBSCAN algorithm. This is a soft-clustering approach that prevent the merging of dissimilar topics, this is, the algorithm strategically generates outliers (i.e., documents that do not fall within any of the created topics) to handle the noise. In BERTopic, these outliers are tagged as *topic "-1"*.
- **Topic representation:** each cluster is assigned to a topic. To measure the relevance of each term (i.e., word) in a topic, the class based TF-IDF (c-TF-IDF) approach was used. This is a modification of TF-IDF, that models the importance of words in clusters instead of in documents. With c-TF-IDF, it is also possible to model how topics evolved over time following a *dynamic topic modeling* approach.

3.2 Methodology

The *abstract*, *title*, *publication year*, and *journal* information for the 96,004 original articles was 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; 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 onwards, the modular approach of BERTopic was applied, with considerations made for each step:

- Embeddings were calculated to feed the BERTopic model. Two models were considered:
 - *all-mpnet-base-v2*: sentence-transformer model that maps sentences and paragraphs to a 768 dimensional dense vector space. This model was trained on a 1B sentence pair, comes from the pre-trained MPnet model Song et al. [2020], and was fine-tuned using a contrastive objective. It was the best positioned model in the sentence transformers rank by March 2024 sbe [2024]. By default, input text longer than 384 word pieces is truncated. This model has been applied in Ramamoorthy et al. [2024], Ng et al. [2023], Guizzardi et al. [2023], Meaney et al. [2022].
 - *S-PubMedBert-MS-MARCO*: a sentence-transformer model specially optimised for medical texts Deka et al. [2022]. This model max sequence length is 350. Input text longer than this size is truncated. This embedding model has been used in the past for similar tasks Karabacak and Margetis [2023], Karabacak et al. [2024a,b,c], Ozkara et al. [2023, 2024].
- The embeddings were resized using the UMAP dimensionality reduction algorithm. The algorithm's parameters were set to default, except for the *random_state* parameter, which controls the algorithm's stochastic behavior by fixing a seed. To assess the consistency of the generated topics, three different seeds (i.e., *random_state*) were applied to each tested model. This approach facilitated a comparison across various initializations, using stability as an intrinsic evaluation metric for evaluating performance.
- HDBSCAN was used as the default clustering algorithm. The cluster minimum size (i.e., *min_cluster_size*) was set to 50, 100, 150, 200 and 250 (i.e., minimum number of documents per topic). As this number increases, the number of microclusters decreases, resulting in fewer topics.

- The default vectorizer model, *CountVectorizer*, was chosen to preprocess the topic representations after the documents were assigned to topics. Stopwords and infrequent words were removed in this step. The n-gram range considered was 1-2, meaning that topic representations made up of one or two words were allowed. Other representations were also explored such as *KeyBERTInspired*, *MaximalMarginalRelevance* (i.e., pursues to maximize the diversity of keywords) and *PartOfSpeech* (i.e., extract keywords based on their Part-of-Speech).

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 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 $2 * 3 * 5 = 30$ models were trained.

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 minimise 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 Rosner et al. [2014], which was introduced in Mimno et al. [2011]. Other coherence measures were calculated (i.e., *c_v*, *c_nmpi* and *c_uci*) but the final decision was guided by *c_umass*. Afterwards, outliers were excluded from the analyses.

After analyzing the keywords and the different topics representations, the topics were labelled through a mutual agreement among D.F.N and L.R.R authors. 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 in the topic. Hierarchical clustering representations were generated to show how topic embeddings can be combined at various cosine distances. Dynamic topic modeling 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 Margetis [2023] to model trends. The publication year, and the topics 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.

4 Results

The median number of tokens per abstract was 375 (Q1: 287, Q3: 442). When combining both 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 Supplementary Table 3, 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 *Supplementary 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. *Supplementary 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 2 and 3 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 ten topics are shown in the Supplementary Figures 1 and 2, and 3 and 4, respectively.

Regarding the dynamic modeling of topics, for each model we studied the themes in batches of 10. Figures 2 and 3 show the results. Moreover, a bar chart of the hot and cold topics for the two models is displayed in Figures 4 and 5. Finally a comparison of the topics of the two final models is presented in Supplementary Table 4.

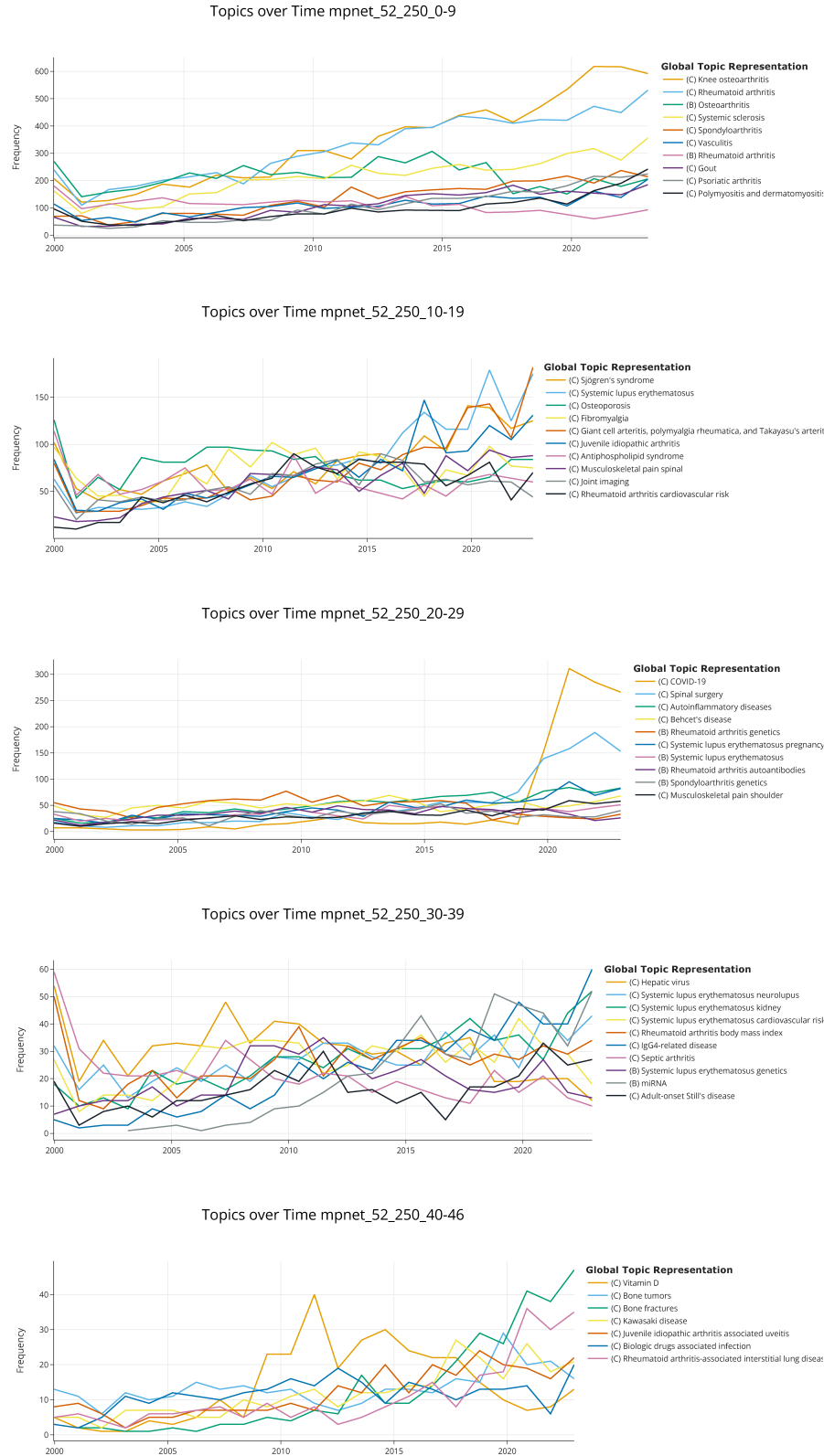


Figure 2: Dynamic topic modeling of the best *all-mpnet-base-v2* model

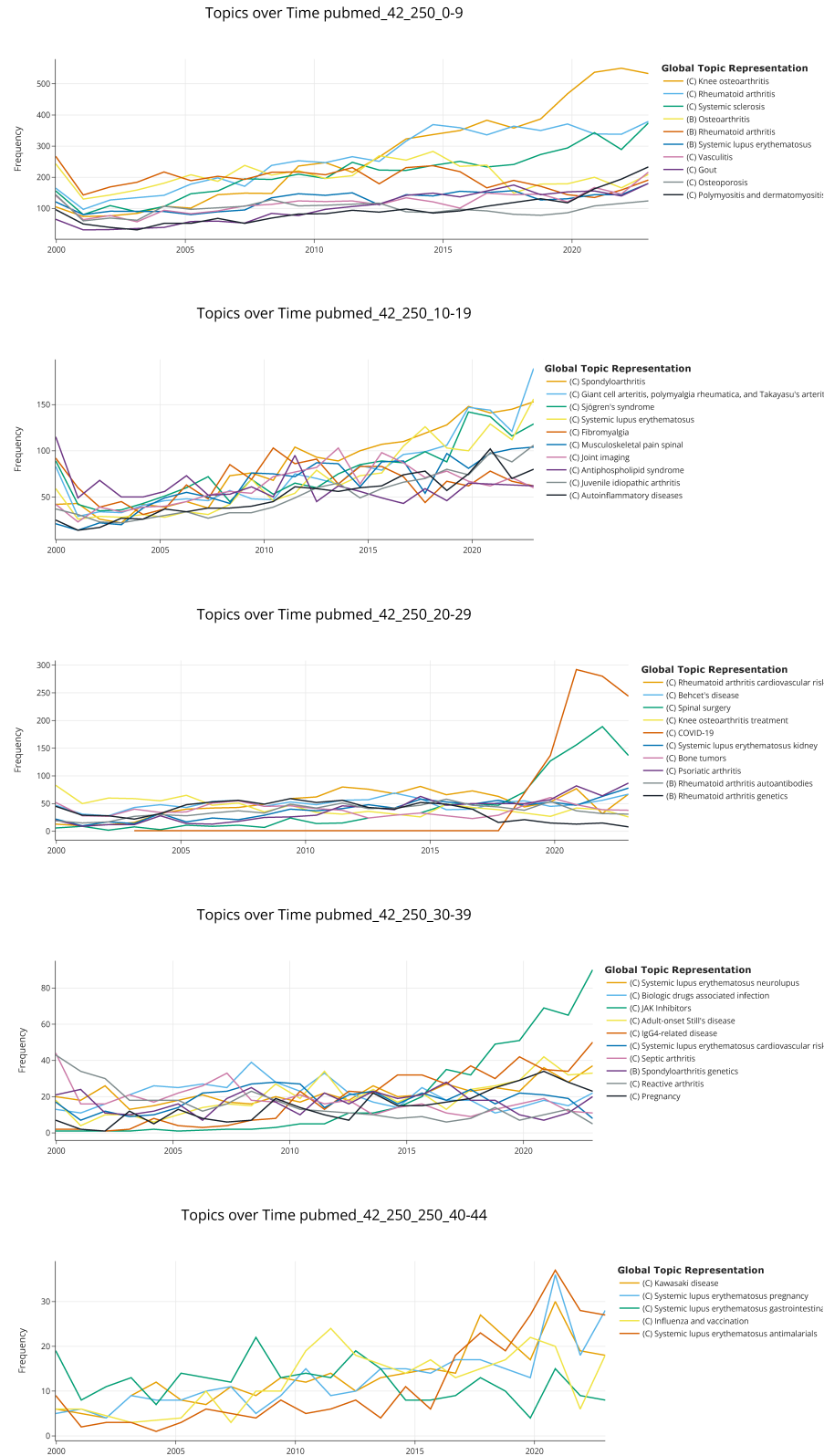


Figure 3: Dynamic topic modeling of the best *S-PubMedBert-MS-MARCO* model

Table 2: 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
(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, acpa, 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) miRNA	445	mir, mirnas, expression, mirna, 5p, 3p, oa, 146a, micromnas, mir 146a
(B) Systemic lupus erythematosus genetics	445	sle, allele, polymorphisms, association, susceptibility, polymorphism, gene, lupus, sle patients, controls
(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) Juvenile idiopathic arthritis associated uveitis	275	uveitis, jia, ocular, eye, children, associated uveitis, patients, idiopathic, juvenile, arthritis
(C) Kawasaki disease	275	kd, ivig, kawasaki, kawasaki disease, coronary, disease kd, coronary artery, children, kd patients, artery
(C) Biologic drugs associated infection	265	infections, infection, tmf, risk, ra, patients, anti tmf, 95, anti, incidence
(C) Rheumatoid arthritis-associated interstitial lung disease	257	ild, ra ild, ra, lung, lung disease, interstitial, hrct, patients, pulmonary, interstitial lung

5 Discussion

5.1 Trends in rheumatology research

When comparing the top ten 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, eight of the ten 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 5,950 for the *all-mpnet-base-v2* model, and 13,297 and 7,149 for the *S-PubMedBert-MS-MARCO*. Therefore, globally, the three most studied topics are: RA, SLE, and osteoarthritis.

Some of the topics expected to be found (e.g., (C) COVID-19 and (C) JAK inhibitors) were present after applying dynamic topic modeling, 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 5 and

Table 3: 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, polyalgia 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
(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 neuropathology	507	np, sle, neuropsychiatric, cognitive, np, brain, lupus, patients, systemic lupus, erythematosus
(C) Biologic drugs 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
(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

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 or new statistical learning techniques that became popular a few years ago, such as trajectory analysis, were identified. However, when analysing 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]. Hence, the use of models with a larger number of topics could be useful to identify new emerging trends. See *Supplementary Excel File Models Output*.

5.2 Topic modeling in PubMed abstracts

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, the authors in Sperandeo et al. [2020] 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 7,500 abstracts and found 30 topics organised in eight hierarchical clusters, concluding that personality is linked to a broad spectrum of conditions. The suitable number of clusters was determined using a 5-fold cross-validation approach.

The authors in Tighe et al. [2020] applied TM on a corpus of more than 200,000 abstracts related to pain. The abstracts collected, retrieved through searches using "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

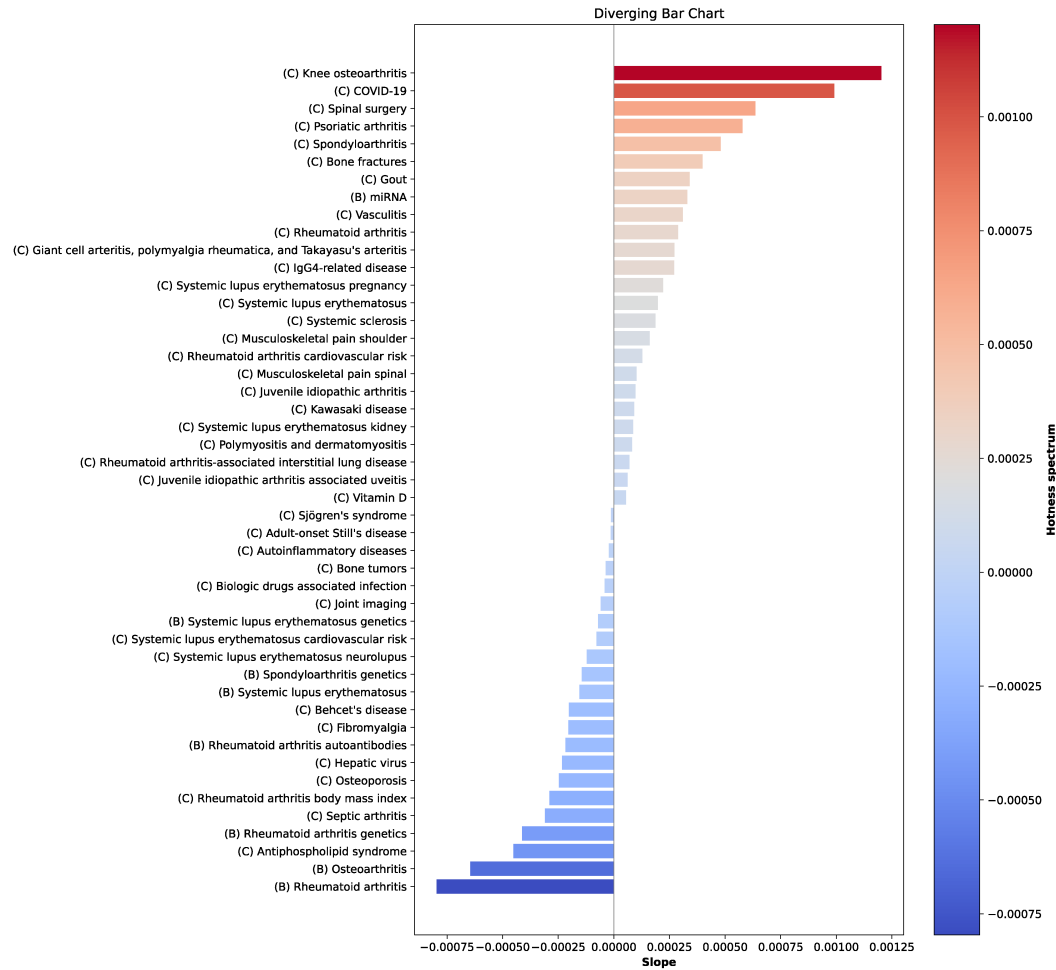


Figure 4: Bar chart of hot and cold topics. *all-mpnet-base-v2* model

coherence strategy, the researchers identified an optimal topic count of 40. One of the conclusions of this research was that TM can be helpful in identifying critical research avenues by evaluating the gaps in the literature concerning a specific topic.

On their behalf, researchers in Abba et al. [2022], 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.

Researchers in Shi et al. [2023] examined artificial intelligence (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 five years, performing unique TM for each period. The authors presented the five main topics in eight different domains of AI. These domains were described by the European Commission Joint Research Centre.

Depression, anxiety, and burnout in academia were studied through BERTopic in Lezhnina [2023]. The authors extracted 2,846 abstracts from PubMed ranging from 1975 to 2023 using a complex query that did not include MeSH terms. Afterwards, 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.

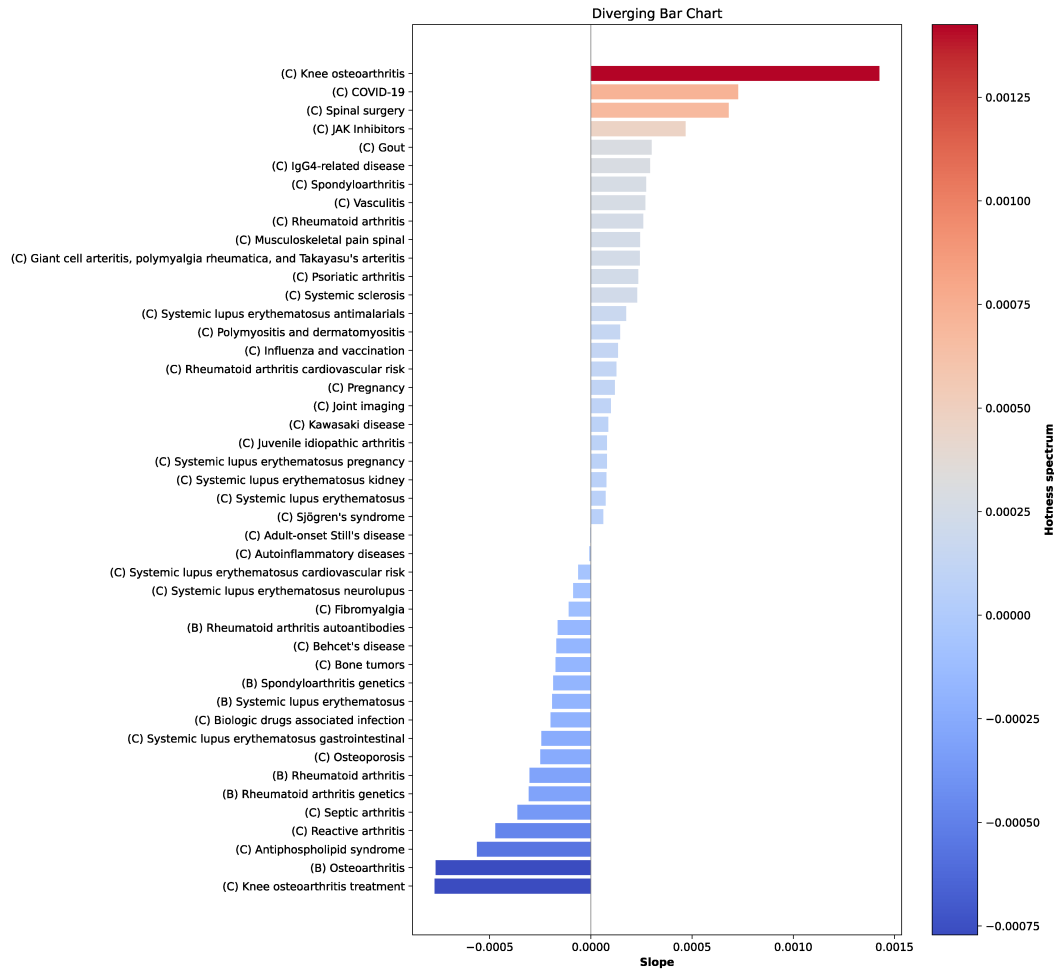


Figure 5: Bar chart of hot and cold topics. *S-PubMedBert-MS-MARCO* model

Eventually, in Grubbs et al. [2023] the researchers studied the topics present in a specific academic journal -Gynecologic Oncology- over a thirty-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 three of them were merged after manual assessment by three reviewers. Thanks to the experiments carried out, researchers could hypothesise the evolution of some topics related to oncology gynecology for the next years, such as an increase in surgical topics and in 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 topic modeling 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 topic modeling 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 Wang et al. [2021].

Furthermore, each clinical note could be categorized into a 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 maximises 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).

5.3 Limitations

- Biologic agents were introduced in the market in 1999. As our study window begins in 2000 we missed the evolution of this topic, from early experiments and clinical trials to their commercial release.
- Topic modeling involves a degree of subjectivity. The results we showcased suggest that topic modeling can be used to discover and understand research trends, rather than assessing the performance of BERTopic as a topic model.
- BERTopic has some noteworthy limitations, as documented by Grootendorst in 2022 Grootendorst [2022]. One significant limitation is the assumption that each document pertains to only one topic, which overlooks the likelihood of documents covering multiple topics.
- Analyzing multiple journals might offer a more comprehensive view, but it also brings variability from the distinct scopes and editorial standards of each journal. This variability may complicate the analysis of research topics and trends. However, both methods—analyzing a single journal Ozkara et al. [2024] and examining multiple journals Karabacak and Margetis [2023]—have been utilized in previous research.
- We have not associated the research trends with other indicators, such as the number of patented products or the volume of clinical trials.

6 Conclusion

To our knowledge, this is the first study that uses BERTopic, and dynamic topic modeling to identify the key topics in rheumatology research using a set of abstracts extracted from PubMed. The two sentence embeddings 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 natural language processing techniques will be necessary to efficiently analyze and synthesize information, helping to identify trends, gaps, and emerging areas of interest across various medical fields.

Data availability statement

All data used in this manuscript is available online at <https://pubmed.ncbi.nlm.nih.gov/>. Data processing is described in Madrid-García, A., Merino-Barbancho, B., Freites-Núñez, D., Rodríguez-Rodríguez, L., Menasalvas-Ruíz, E., Rodríguez-González, A., & Peñas, A. (2024). From Web to RheumaLpack: Creating a Linguistic Corpus for Exploitation and Knowledge Discovery in Rheumatology. medRxiv, 2024-04.Madrid et al. [2024].

Further inquiries can be directed to the corresponding author.

Funding statement

This study did not receive any funding

CRediT author statement

Alfredo Madrid-García: Conceptualization of this study, methodology, coding, review, writing (original draft preparation). **Dalifer Freites Núñez:** Methodology **Luis Rodríguez Rodríguez:** Methodology, review

Supplementary material files

- Supplementary Excel File Models Output: topics identified in the 30 models trained
- Supplementary Excel File Top 5 Predominant Topics: predominant topics in the 30 models trained

Acknowledgement

The authors would like to thank: Inés Pérez San Cristobal, Anselmo Peñas, and Alejandro Rodríguez González

Conflicts of interest

None declared

References

- Mike Thelwall and Pardeep Sud. Scopus 1900–2020: Growth in articles, abstracts, countries, fields, and journals. *Quantitative Science Studies*, 3(1):37–50, 2022.
- Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15, 2021.
- Nancy J Olsen and C Michael Stein. New drugs for rheumatoid arthritis. *New England Journal of Medicine*, 350(21):2167–2179, 2004.
- Josef S Smolen. Insights into the treatment of rheumatoid arthritis: a paradigm in medicine. *Journal of autoimmunity*, 110:102425, 2020.
- SA Kerrigan and IB McInnes. Reflections on ‘older’ drugs: learning new lessons in rheumatology. *Nature Reviews Rheumatology*, 16(3):179–183, 2020.
- Ronald van Vollenhoven. Treat-to-target in rheumatoid arthritis—are we there yet? *Nature Reviews Rheumatology*, 15(3):180–186, 2019.
- Hmwe Hmwe Kyu, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, Ibrahim Abdollahpour, Rizwan Suliankatchi Abdulkader, Molla Abebe, Zegeye Abebe, Olifan Zewdie Abil, and Victor Aboyans. 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. *The Lancet*, 392:1859–1922, 11 2018. ISSN 01406736. doi:10.1016/S0140-6736(18)32335-3.
- Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, 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. *The Lancet*, 392(10159):1789–1858, 2018.
- Leonie E Burgers, Karim Raza, Annette H Van Der Helm-Van, et al. Window of opportunity in rheumatoid arthritis—definitions and supporting evidence: from old to new perspectives. *RMD open*, 5(1):e000870, 2019.
- Hanna W van Steenberg, Daniel Aletaha, Liesbeth JJ Beart-van de Voorde, Elisabeth Brouwer, Catalin Codreanu, Bernard Combe, João E Fonseca, Merete L Hetland, Frances Humby, Tore K Kvien, et al. Eular definition of arthralgia suspicious for progression to rheumatoid arthritis. *Annals of the rheumatic diseases*, 76(3):491–496, 2017.
- Désirée Van Der Heijde, Annette HM Van Der Helm-Van, Daniel Aletaha, Clifton O Bingham, Gerd R Burmester, Maxime Dougados, Paul Emery, David Felson, Rachel Knevel, Tore K Kvien, et al. Eular definition of erosive disease in light of the 2010 acr/eular rheumatoid arthritis classification criteria. *Annals of the rheumatic diseases*, 72(4):479–481, 2013.
- György Nagy, Nadia MT Roodenrijs, Paco MJ Welsing, Melinda Kedves, Attila Hamar, Marlies C Van Der Goes, Alison Kent, Margot Bakkers, Etienne Blaas, Ladislav Senolt, et al. Eular definition of difficult-to-treat rheumatoid arthritis. *Annals of the rheumatic diseases*, 80(1):31–35, 2021.
- Rob Churchill and Lisa Singh. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35, 2022.
- Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131, 2023.
- David M Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2011.
- David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
- Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78:15169–15211, 2019.
- Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296, 2017.
- Christine K. Mulunda, Peter W. Wagacha, and Lawrence Muchemi. Review of trends in topic modeling techniques, tools, inference algorithms and applications. In *2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMCI)*, pages 28–37, 2018. doi:10.1109/ISCMCI.2018.8703231.
- Marco Valeri. *Organizational Phenomenon*, pages 1–17. Springer International Publishing, Cham, 2021. ISBN 978-3-030-87148-2. doi:10.1007/978-3-030-87148-2_1. URL https://doi.org/10.1007/978-3-030-87148-2_1.

Sara K Tedeschi, Tianrun Cai, Zeling He, Yuri Ahuja, Chuan Hong, Katherine A Yates, Kumar Dahal, Chang Xu, Houchen Lyu, Kazuki Yoshida, et al. Classifying pseudogout using machine learning approaches with electronic health record data. *Arthritis care & research*, 73(3):442–448, 2021.

Liqin Wang, Eli Miloslavsky, John H Stone, Hyon K Choi, Li Zhou, and Zachary S Wallace. Topic modeling to characterize the natural history of anca-associated vasculitis from clinical notes: A proof of concept study. In *Seminars in arthritis and rheumatism*, volume 51, pages 150–157. Elsevier, 2021.

Eldin Dzubur, Carine Khalil, Christopher V Almario, Benjamin Noah, Deeba Minhas, Mariko Ishimori, Corey Arnold, Yujin Park, Jonathan Kay, Michael H Weisman, et al. Patient concerns and perceptions regarding biologic therapies in ankylosing spondylitis: insights from a large-scale survey of social media platforms. *Arthritis care & research*, 71(2):323–330, 2019.

Jenny Xiaoyu Li and Elaine Yacyshyn. Thoughts and experiences of behçet disease from participants on a reddit subforum: Qualitative online community analysis. *JMIR Form Res*, 7:e49380, 12 2023. ISSN 2561-326X. doi:10.2196/49380. URL <https://doi.org/10.2196/49380>.

The “topics” in the electronic health record of rheumatoid arthritis patients before initiating targeted therapies and association with future treatment course. <https://acrabstracts.org/abstract/the-topics-in-the-electronic-health-record-of-rheumatoid-arthritis-patients-before-initiating-targeted-therapies-and-association-with-future-treatment-course>. ., 09 2023a. Accessed: 2024-1-25.

Understanding community perspectives on disease management: A social media analysis of gout care strategies. <https://acrabstracts.org/abstract/understanding-community-perspectives-on-disease-management-a-social-media-analysis-of-gout-care-strategies>. ., 09 2023b. Accessed: 2024-1-25.

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. <https://acrabstracts.org/abstract/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>. ., 08 2018. Accessed: 2024-1-25.

L. Sperl, T. Stamm, M. R. Andrews, M. Bjork, C. Boström, J. Cappon, J. de la Torre-Aboki, A. de Thurah, A. Domjan, R. Dragoi, F. Estevez-Lopez, R. J. O. Ferreira, G. E. Fragoulis, J. Grygielska, K. Korve, M. L. Kukkurainen, C. Madelaine-Bonjour, A. Marques, J. Meesters, R. H. Moe, E. Moholt, E. Mosor, C. Naimer-Stach, M. Ndosi, P. Pchelnikova, J. Primdahl, P. Putrik, A. K. Rausch Osthoff, H. Smucrova, S. Stefanac, M. Testa, L. van Bodegom-Vos, W. Peter, H. A. Zangi, O. Zimba, T. P. M. Vliet Vlieland, and V. Ritschl. 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. *Annals of the Rheumatic Diseases*, 81(Suppl 1):139–140, 2022. ISSN 0003-4967. doi:10.1136/annrhumdis-2022-eular.4304. URL https://ard.bmj.com/content/81/Suppl_1/139.1.

Stephanie Eaneff, Timothy Vaughan, Volkan Baruta, Jesper Havsol, Brad Nohe, and Cathy Emmas. 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. <https://acrabstracts.org/abstract/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>. ., 08 2018. Accessed: 2024-1-25.

Alfredo Madrid, Beatriz Merino Barbancho, Dalifer Dayanira Freites Nuñez, Luis Rodriguez Rodriguez, Ernestina Menasalvas Ruiz, Alejandro Rodriguez Gonzalez, and Anselmo Peñas. From web to rheumalpack: Creating a linguistic corpus for artificial intelligence exploitation and knowledge discovery in rheumatology. *medRxiv*, 2024. doi:10.1101/2024.04.26.24306269. URL <https://www.medrxiv.org/content/early/2024/04/27/2024.04.26.24306269>.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MpNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.

Pretrained models — sentence transformers documentation, 2024. URL https://www.sbert.net/docs/pretrained_models.html. Accessed: 2024-06-01.

Thilagavathi Ramamoorthy, Vaitheeswaran Kulothungan, and Bagavandas Mappillairaju. Topic modeling and social network analysis approach to explore diabetes discourse on twitter in india. *Frontiers in Artificial Intelligence*, 7: 1329185, 2024.

Qin Xiang Ng, Dawn Yi Xin Lee, Chun En Yau, Yu Liang Lim, Clara Xinyi Ng, and Tau Ming Liew. Examining the public messaging on ‘loneliness’ over social media: An unsupervised machine learning analysis of twitter posts over the past decade. In *Healthcare*, volume 11, page 1485. MDPI, 2023.

- Stefano Guizzardi, Maria Teresa Colangelo, Prisco Mirandola, and Carlo Galli. Modeling new trends in bone regeneration, using the bertopic approach. *Regenerative Medicine*, 18(9):719–734, 2023.
- Christopher Meaney, Michael Escobar, Therese A Stukel, Peter C Austin, Liisa Jaakkimainen, et al. Comparison of methods for estimating temporal topic models from primary care clinical text data: Retrospective closed cohort study. *JMIR Medical Informatics*, 10(12):e40102, 2022.
- Pritam Deka, Anna Jurek-Loughrey, and P Deepak. Improved methods to aid unsupervised evidence-based fact checking for online health news. *Journal of Data Intelligence*, 3(4):474–504, 2022.
- Mert Karabacak and Konstantinos Margetis. Natural language processing reveals research trends and topics in the spine journal over two decades: A topic modeling study. *The Spine Journal*, 2023.
- Mert Karabacak, Pemla Jagtiani, Ankita Jain, Fedor Panov, and Konstantinos Margetis. Tracing topics and trends in drug-resistant epilepsy research using a natural language processing–based topic modeling approach. *Epilepsia*, 2024a.
- Mert Karabacak, Pemla Jagtiani, Carl Moritz Zipser, Lindsay Tetreault, Benjamin Davies, and Konstantinos Margetis. Mapping the degenerative cervical myelopathy research landscape: Topic modeling of the literature. *Global Spine Journal*, page 21925682241256949, 2024b.
- Mert Karabacak, Ankita Jain, Pemla Jagtiani, Zachary L Hickman, Kristen Dams-O’Connor, and Konstantinos Margetis. Exploiting natural language processing to unveil topics and trends of traumatic brain injury research. *Neurotrauma Reports*, 5(1):203–214, 2024c.
- Burak B Ozkara, Mert Karabacak, Konstantinos Margetis, Vivek S Yedavalli, Max Wintermark, and Sotirios Bisdas. Assessment of computed tomography perfusion research landscape: A topic modeling study. *Tomography*, 9(6): 2016–2028, 2023.
- Burak Berksu Ozkara, Mert Karabacak, Konstantinos Margetis, Wade Smith, Max Wintermark, and Vivek Srikar Yedavalli. Trends in stroke-related journals: Examination of publication patterns using topic modeling. *Journal of Stroke and Cerebrovascular Diseases*, page 107665, 2 2024. ISSN 10523057. doi:10.1016/j.jstrokecerebrovasdis.2024.107665.
- Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettleing, and Andreas Both. Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*, 2014.
- David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272, 2011.
- Raffaele Sperandeo, Giovanni Messina, Daniela Iennaco, Francesco Sessa, Vincenzo Russo, Vincenzo Monda, Marcellino Monda, Antonietta Messina, Silvia Dell’Orco, Enrico Moretto, 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. *Frontiers in psychiatry*, 10:449078, 2020.
- Patrick J Tighe, Bharadwaj Sannapaneni, Roger B Fillingim, Charlie Doyle, Michael Kent, Ben Shickel, and Parisa Rashidi. 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 Medicine*, 21(11):3133–3160, 2020.
- Mustapha Abba, Chidozie Nduka, Seun Anjorin, Shukri Mohamed, Emmanuel Agogo, Olalekan Uthman, et al. One hundred years of hypertension research: Topic modeling study. *JMIR Formative Research*, 6(5):e31292, 2022.
- Jin Shi, David Bendig, Horst Christian Vollmar, and Peter Rasche. Mapping the bibliometrics landscape of ai in medicine: methodological study. *Journal of Medical Internet Research*, 25:e45815, 2023.
- Olga Lezhnina. Depression, anxiety, and burnout in academia: topic modeling of pubmed abstracts. *Frontiers in Research Metrics and Analytics*, 8:1271385, 2023.
- Allison E Grubbs, Nikita Sinha, Ravi Garg, and Emma L Barber. Use of topic modeling to assess research trends in the journal gynecologic oncology. *Gynecologic oncology*, 172:41–46, 2023.

Supplementary Material

Supplementary Tables

Supplementary Table 1: Rheumatology journals classified by the Journal Citation Report index as "RHEUMATOLOGY - SCIE". The journal name is written as appears in JCR webpage. Aktuelle Rheumatologie was excluded from this list

Journal name	
Nature Reviews Rheumatology	Rheumatology and Therapy
ANNALS OF THE RHEUMATIC DISEASES	CLINICAL AND EXPERIMENTAL RHEUMATOLOGY
Lancet Rheumatology	CLINICAL RHEUMATOLOGY
Arthritis & Rheumatology	JCR-JOURNAL OF CLINICAL RHEUMATOLOGY
Osteoarthritis and Cartilage	LUPUS
RMD Open	Pediatric Rheumatology
RHEUMATOLOGY	International Journal of Rheumatic Diseases
BEST PRACTICE & RESEARCH IN CLINICAL RHEUMATOLOGY	BMC MUSCULOSKELETAL DISORDERS
CURRENT OPINION IN RHEUMATOLOGY	Advances in Rheumatology
Current Rheumatology Reports	RHEUMATIC DISEASE CLINICS OF NORTH AMERICA
SEMINARS IN ARTHRITIS AND RHEUMATISM	Modern Rheumatology
ARTHRITIS RESEARCH & THERAPY	SCANDINAVIAN JOURNAL OF RHEUMATOLOGY
ARTHRITIS CARE & RESEARCH	Archives of Rheumatology
JOINT BONE SPINE	ZEITSCHRIFT FUR RHEUMATOLOGIE
Therapeutic Advances in Musculoskeletal Disease	Acta Reumatologica Portuguesa
RHEUMATOLOGY INTERNATIONAL	ARP Rheumatology
JOURNAL OF RHEUMATOLOGY	
Lupus Science & Medicine	

Supplementary Table 2: Number of articles with abstract published by year, considering the 34 JCR journals with the category "RHEUMATOLOGY - SCIE". Although 2024 appears, it must be noted that the time interval studied is 2000-2023. This inconsistency is due to the difference in creation and indexing in PubMed and the date of publication.

Journal	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Acta Reumatol Port	0	0	0	0	0	0	27	38	46	74	60	45	42	34	48
Adv Rheumatol	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ann Rheum Dis	158	199	205	219	278	352	302	303	299	295	386	361	320	306	303
Arch Rheumatol	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARP Rheumatol	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Arthritis Care Res (Hoboken)	0	0	0	0	0	0	0	0	0	0	203	188	222	238	226
Arthritis Res Ther	0	0	0	88	108	201	232	181	211	292	308	305	334	284	292
Arthritis Rheumatol	0	0	0	0	0	0	0	0	0	0	0	0	0	0	332
Best Pract Res Clin Rheumatol	0	51	54	60	55	65	71	65	70	61	68	63	58	57	59
BMC Musculoskelet Disord	1	10	25	28	49	61	103	129	171	167	287	289	265	365	450
Clin Exp Rheumatol	150	178	178	190	176	176	167	207	229	248	241	227	252	247	278
Clin Rheumatol	107	102	109	104	117	133	198	457	301	251	220	237	249	276	258
Curr Opin Rheumatol	82	76	94	98	95	90	86	84	95	88	95	88	91	96	98
Curr Rheumatol Rep	62	65	63	57	54	60	59	60	59	58	63	69	80	90	79
Int J Rheum Dis	0	0	0	0	0	0	0	0	0	61	77	59	79	95	110
J Clin Rheumatol	57	63	54	53	65	66	60	67	66	76	78	85	68	71	70
J Rheumatol	427	402	367	402	346	368	371	331	327	365	325	348	304	233	290
Joint Bone Spine	83	64	90	81	103	89	121	108	129	122	120	113	101	106	79
Lancet Rheumatol	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lupus	122	137	117	153	157	167	143	143	139	195	218	192	237	201	208
Lupus Sci Med	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26
Mod Rheumatol	49	65	65	66	91	84	78	100	105	105	105	116	138	193	167
Nat Rev Rheumatol	0	0	0	0	0	0	0	0	0	72	83	84	71	82	82
Osteoarthritis Cartilage	73	106	109	96	118	128	168	173	206	216	221	170	194	238	230
Pediatr Rheumatol Online J	0	0	0	0	0	0	0	22	20	21	30	35	39	46	52
Rheum Dis Clin North Am	56	48	50	46	49	46	44	39	57	55	43	40	49	44	47
Rheumatol Int	40	70	86	70	78	160	204	182	222	257	240	268	642	458	232
Rheumatol Ther	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Rheumatology (Oxford)	181	174	182	214	226	226	240	292	342	290	289	286	298	276	278
RMD Open	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Scand J Rheumatol	74	67	71	64	73	80	80	69	72	73	75	70	69	78	73
Semin Arthritis Rheum	41	43	43	31	37	56	43	44	43	47	50	93	73	84	102
Ther Adv Musculoskelet Dis	0	0	0	0	0	0	0	0	0	9	28	27	33	22	17
Z Rheumatol	77	34	66	66	45	55	75	77	77	93	96	100	91	88	86

Journal	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	Total
Acta Reumatol Port	50	48	40	38	43	41	39	0	0	0	713
Adv Rheumatol	0	0	0	40	61	47	71	45	56	0	320
Ann Rheum Dis	311	295	275	225	192	182	171	187	186	0	6,310
Arch Rheumatol	18	52	77	49	59	73	77	70	22	0	497
ARP Rheumatol	0	0	0	0	0	0	0	46	38	0	84
Arthritis Care Res (Hoboken)	201	221	225	222	179	200	209	229	320	27	3,110
Arthritis Res Ther	372	288	283	276	295	282	290	269	235	0	5,426
Arthritis Rheumatol	318	283	209	193	197	194	224	181	236	25	2,392
Best Pract Res Clin Rheumatol	59	63	63	65	57	57	39	45	56	0	1,361
BMC Musculoskelet Disord	386	497	549	455	638	843	1037	1130	962	0	8,897
Clin Exp Rheumatol	285	276	267	265	270	290	283	300	326	21	5,727
Clin Rheumatol	299	411	368	440	436	443	528	383	336	79	6,842
Curr Opin Rheumatol	83	91	91	89	91	78	74	50	54	20	2,077
Curr Rheumatol Rep	77	75	81	86	76	90	76	41	31	5	1,616
Int J Rheum Dis	105	159	224	269	274	186	158	155	277	58	2,346
J Clin Rheumatol	60	50	51	49	61	68	130	163	83	11	1,725
J Rheumatol	301	264	241	188	195	200	228	166	232	0	7,221
Joint Bone Spine	75	96	92	87	84	70	86	79	92	8	2,278
Lancet Rheumatol	0	0	0	0	23	63	65	64	57	5	277
Lupus	196	197	202	281	212	210	262	200	178	19	4,486
Lupus Sci Med	29	35	27	45	39	46	73	89	70	0	479
Mod Rheumatol	162	161	167	151	145	144	159	156	265	0	3,037
Nat Rev Rheumatol	82	63	62	53	51	54	53	55	58	4	1,009
Osteoarthritis Cartilage	255	239	241	188	194	153	166	144	164	20	4,210
Pediatr Rheumatol Online J	61	67	80	82	83	90	162	109	138	0	1,137
Rheum Dis Clin North Am	44	46	44	45	43	51	48	56	54	13	1,157
Rheumatol Int	254	207	245	310	244	234	237	235	270	39	5,484
Rheumatol Ther	15	25	37	44	46	66	127	97	110	11	582
Rheumatology (Oxford)	267	262	264	282	261	445	666	518	686	0	7,445
RMD Open	80	86	112	113	107	128	147	205	270	0	1,248
Scand J Rheumatol	72	74	62	58	59	58	60	56	79	9	1,675
Semin Arthritis Rheum	103	107	116	125	155	204	165	150	171	27	2,153
Ther Adv Musculoskelet Dis	22	18	27	22	31	81	107	107	48	0	599
Z Rheumatol	94	97	105	95	101	116	105	106	132	7	2,084

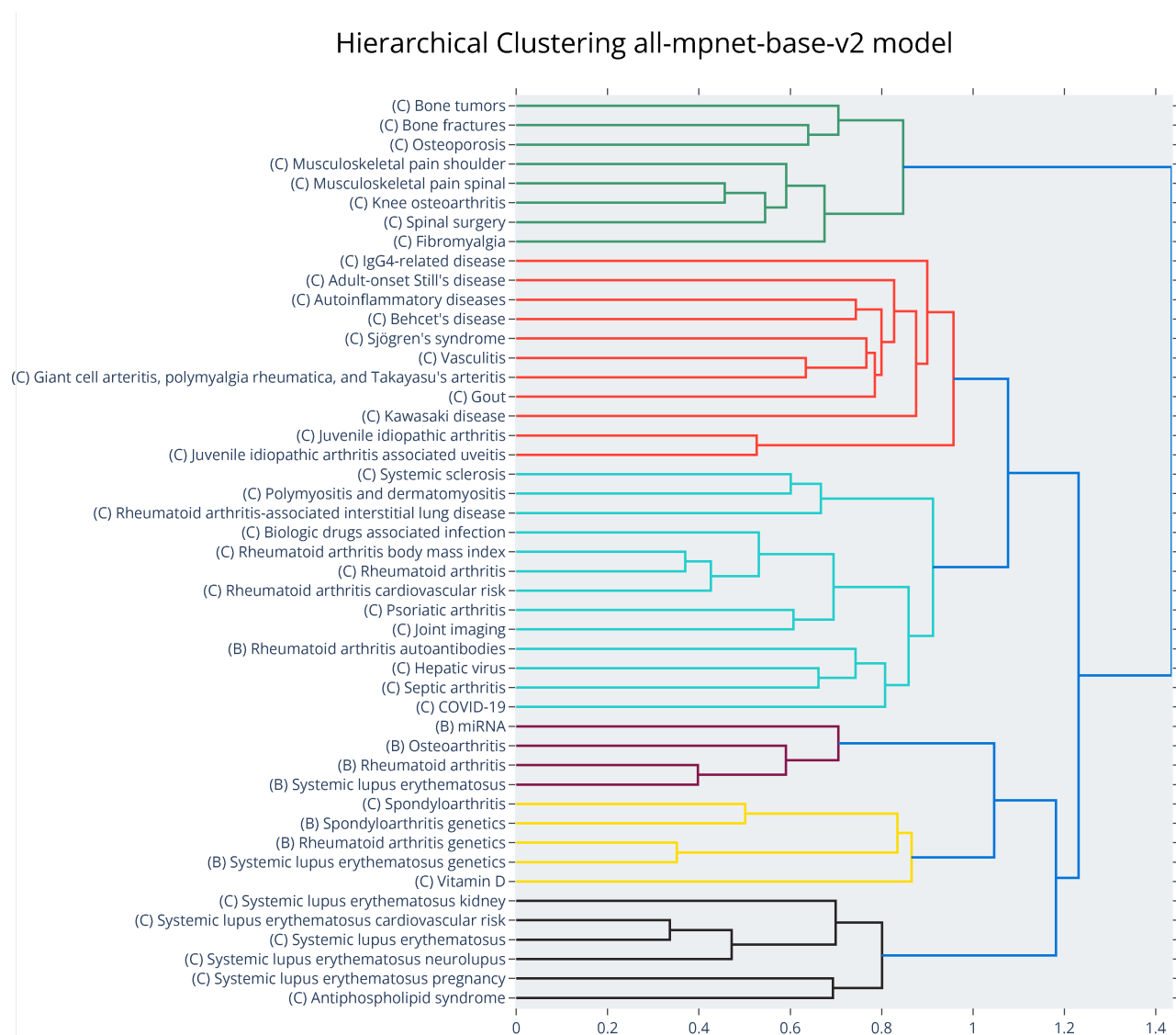
Supplementary Table 3: Results of the models

Model	Min clusters size	Seed	Topics	Outliers	Coherence score			
					u_mass	c_v	c_nmpi	c_uci
all-mpnet-base-v2	50	42	296	31045	-0.933	0.419	-0.006	-2.689
all-mpnet-base-v2	50	52	288	30739	-0.937	0.418	-0.006	-2.703
all-mpnet-base-v2	50	62	294	29168	-0.923	0.419	-0.006	-2.713
S-PubMedBert	50	42	238	33087	-0.819	0.408	-0.012	-2.885
S-PubMedBert	50	52	246	33676	-0.816	0.410	-0.013	-2.917
S-PubMedBert	50	62	246	34309	-0.823	0.409	-0.011	-2.875
all-mpnet-base-v2	100	42	151	29265	-0.689	0.492	0.008	-2.449
all-mpnet-base-v2	100	52	148	30212	-0.680	0.492	0.009	-2.439
all-mpnet-base-v2	100	62	154	29951	-0.689	0.489	0.009	-2.454
S-PubMedBert	100	42	129	35332	-0.610	0.486	-0.005	-2.762
S-PubMedBert	100	52	126	33162	-0.589	0.479	-0.008	-2.838
S-PubMedBert	100	62	129	35102	-0.606	0.492	0.002	-2.601
all-mpnet-base-v2	150	42	87	27896	-0.480	0.535	0.009	-2.438
all-mpnet-base-v2	150	52	85	26068	-0.477	0.533	0.012	-2.351
all-mpnet-base-v2	150	62	71	21245	-0.417	0.521	-0.003	-2.692
S-PubMedBert	150	42	83	35288	-0.467	0.532	0.004	-2.566
S-PubMedBert	150	52	84	35021	-0.444	0.537	0.000	-2.636
S-PubMedBert	150	62	81	33608	-0.461	0.537	0.005	-2.537
all-mpnet-base-v2	200	42	63	26786	-0.385	0.549	0.006	-2.506
all-mpnet-base-v2	200	52	55	21398	-0.347	0.541	-0.002	-2.680
all-mpnet-base-v2	200	62	61	25178	-0.361	0.547	0.005	-2.548
S-PubMedBert	200	42	49	23908	-0.320	0.574	0.015	-2.241
S-PubMedBert	200	52	56	26246	-0.336	0.546	-0.001	-2.653
S-PubMedBert	200	62	54	27647	-0.330	0.562	0.004	-2.537
all-mpnet-base-v2	250	42	44	22334	-0.291	0.556	0.001	-2.621
all-mpnet-base-v2	250	52	48	22628	-0.279	0.561	-0.002	-2.664
all-mpnet-base-v2	250	62	42	19075	-0.281	0.565	0.004	-2.472
S-PubMedBert	250	42	46	26688	-0.288	0.558	-0.004	-2.686
S-PubMedBert	250	52	48	28045	-0.291	0.551	-0.006	-2.771
S-PubMedBert	250	62	48	28343	-0.301	0.551	-0.007	-2.764

Supplementary Table 4: Unique themes of the two selected models

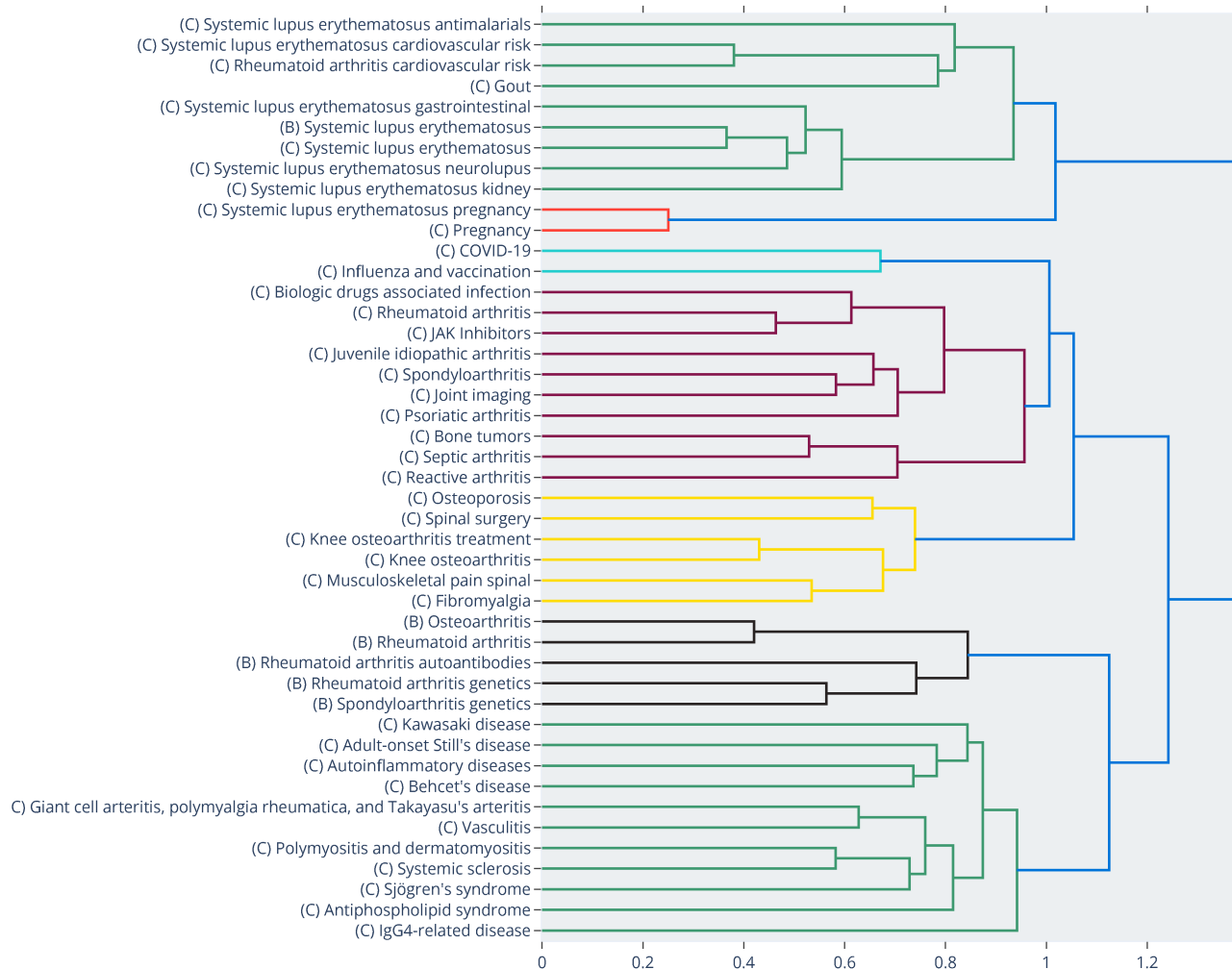
all-mpnet-base-v2	S-PubMedBert-MS-MARCO
(B) miRNA	(C) JAK Inhibitors
(B) Systemic lupus erythematosus genetics	(C) Knee osteoarthritis treatment
(C) Bone fractures	(C) Pregnancy
(C) Hepatic virus	(C) Reactive arthritis
(C) Rheumatoid arthritis-associated interstitial lung disease	(C) Systemic lupus erythematosus antimalarials
(C) Juvenile idiopathic arthritis associated uveitis	(C) Systemic lupus erythematosus kidney
(C) Musculoskeletal pain shoulder	(C) Influenza and vaccination
(C) Rheumatoid arthritis	
(C) Rheumatoid arthritis body mass index	
(C) Vitamin D	

Supplementary Figures

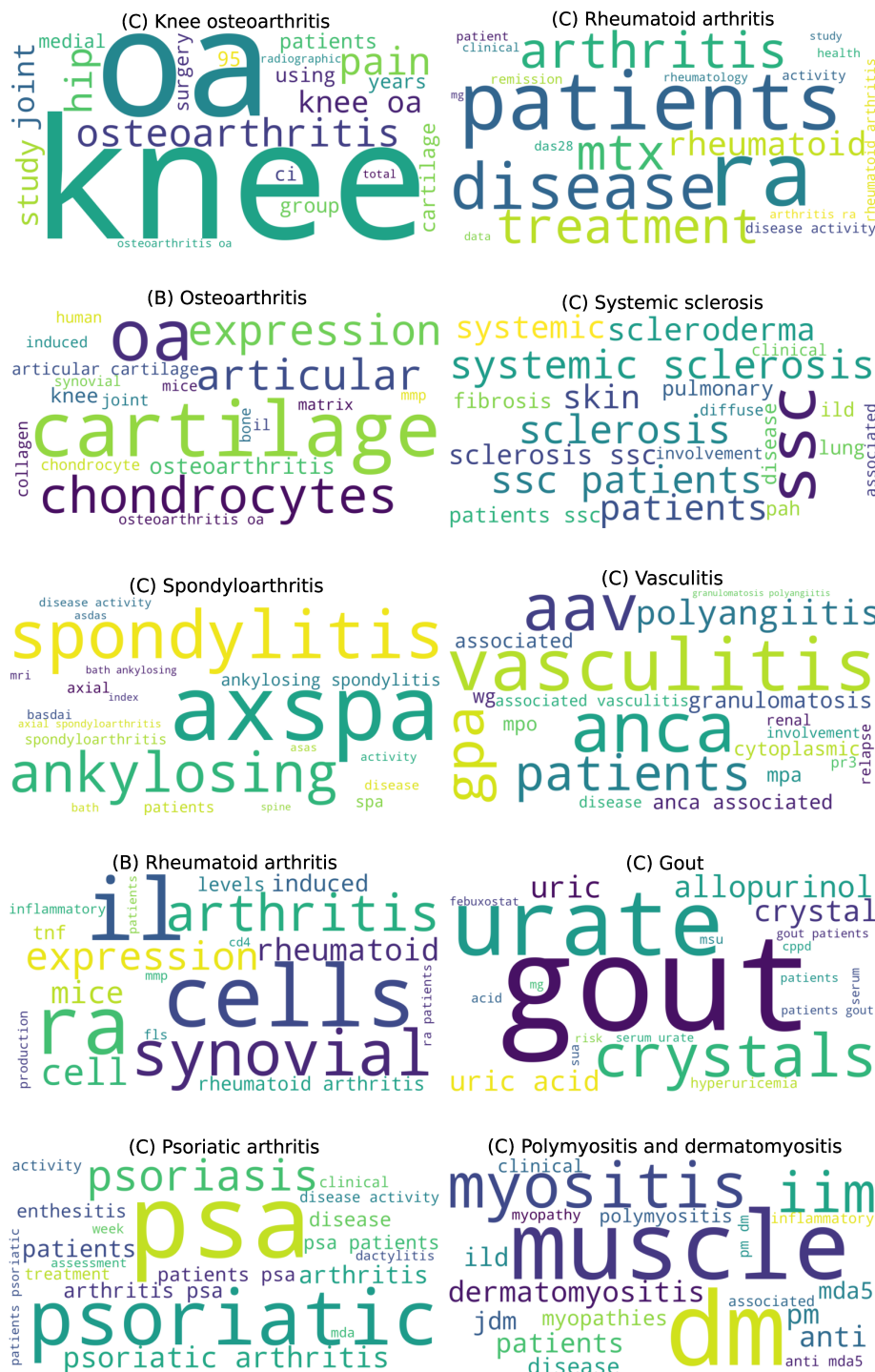


Supplementary Figure 1: Hierarchical structure of the topics labeled with the agreed label. Best *all-mpnet-base-v2* model

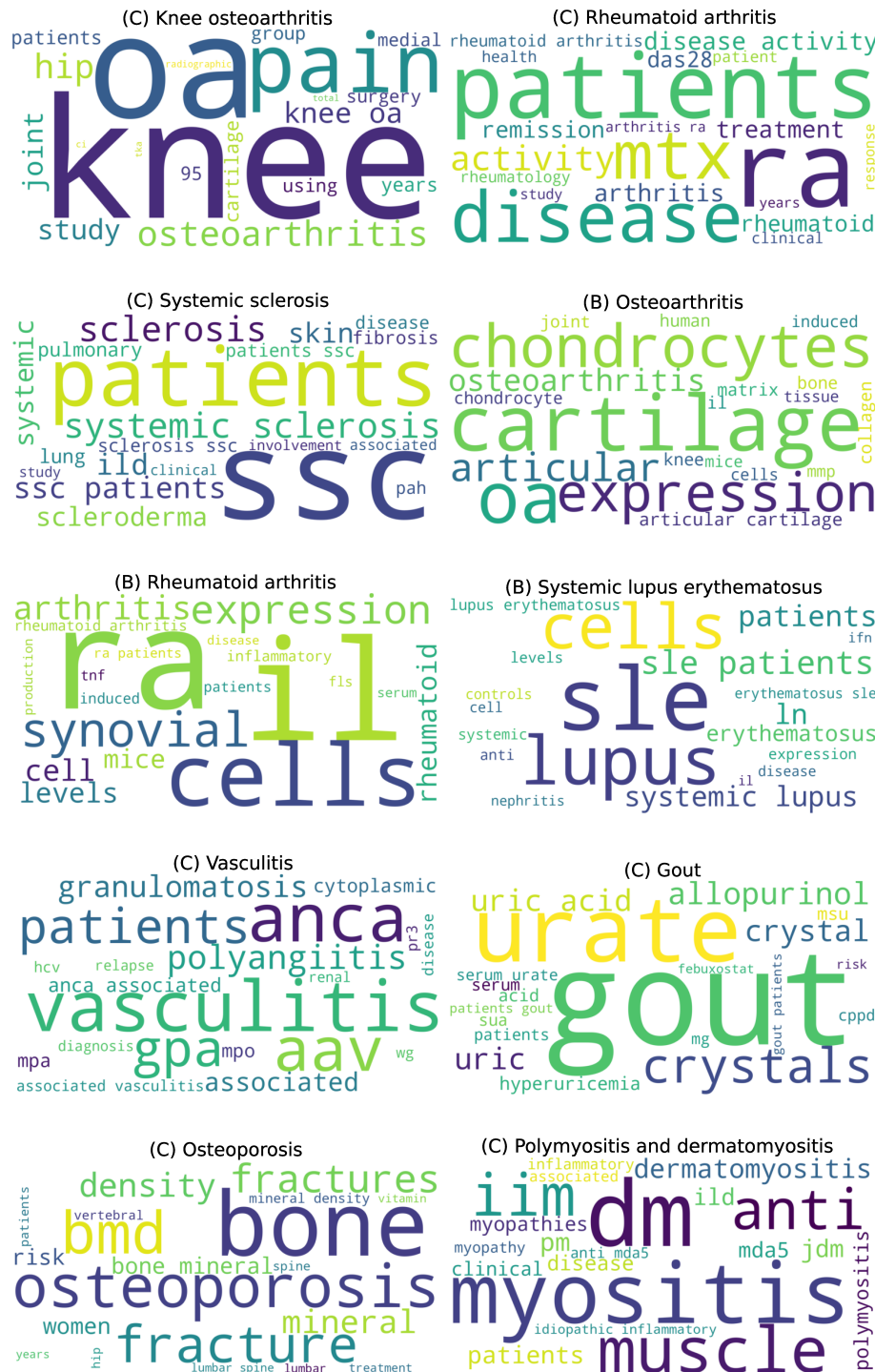
Hierarchical Clustering S-PubMedBert-MS-MARCO



Supplementary Figure 2: Hierarchical structure of the topics labeled with the agreed label. Best *S-PubMedBert-MS-MARCO* model



Supplementary Figure 3: Wordclouds of the top 10 topics of the best *all-mpnet-base-v2* model



Supplementary Figure 4: Wordclouds of the top 10 topics of the best *S-PubMedBert-MS-MARCO* model