RMD Open
Rheumatic & Musculoskeletal Diseases

REVIEW

Artificial intelligence in rheumatology research: what is it good for?

José Miguel Sequí-Sabater ,[1,2,3] Diego Benavent [4]

To cite: Sequí-Sabater JM, Benavent D. Artificial intelligence in rheumatology research: what is it good for?. *RMD Open* 2025;**11**:e004309. doi:10.1136/rmdopen-2024-004309

Received 31 August 2024
Accepted 8 December 2024

Check for updates

[1]Rheumatology Department, La Ribera University Hospital, Alzira, Spain
[2]Rheumatology Deparment, La Fe University and Polytechnic Hospital, Valencia, Spain
[3]Division of Rheumatology, Department of Medicine Solna, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden
[4]Rheumatology Department, Hospital Universitari de Bellvitge, L'Hospitalet de Llobregat, Barcelona, Spain

Correspondence to
Dr Diego Benavent;
d_benavent@hotmail.com

ABSTRACT

Artificial intelligence (AI) is transforming rheumatology research, with a myriad of studies aiming to improve diagnosis, prognosis and treatment prediction, while also showing potential capability to optimise the research workflow, improve drug discovery and clinical trials. Machine learning, a key element of discriminative AI, has demonstrated the ability of accurately classifying rheumatic diseases and predicting therapeutic outcomes by using diverse data types, including structured databases, imaging and text. In parallel, generative AI, driven by large language models, is becoming a powerful tool for optimising the research workflow by supporting with content generation, literature review automation and clinical decision support. This review explores the current applications and future potential of both discriminative and generative AI in rheumatology. It also highlights the challenges posed by these technologies, such as ethical concerns and the need for rigorous validation and regulatory oversight. The integration of AI in rheumatology promises substantial advancements but requires a balanced approach to optimise benefits and minimise potential possible downsides.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Discriminative artificial intelligence (AI) is advancing rheumatology with machine learning models that enhance disease diagnosis and prediction by analysing structured data, imaging data and text.

WHAT THIS STUDY ADDS

⇒ Generative AI, using large language models, may significantly support research by assisting the process and refining study development via general and specialised chatbots, although its application in rheumatology is still in early development.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ To fully harness AI's potential in rheumatology research, it is crucial to balance innovation with responsibility, ensuring robust methodologies and the preservation of research integrity.

INTRODUCTION

Artificial intelligence (AI) has emerged as a transformative technology in medicine, providing rheumatology with innovative tools for research. AI, known as the capability of computational systems to perform tasks that typically require human intelligence, include learning patterns from prior data, understanding natural language, perception, reasoning, problem-solving.[1] The impact of this technology in health sciences research is increasingly evident, with multiple applications gradually being integrated into the field of rheumatology.[1 2] Indeed, different algorithms have led to the development of models for the diagnosis, evaluation, prognosis and prediction of disease.[3] As AI has evolved, it has become increasingly important to differentiate between discriminative AI, widely used for studies on disease classification and prediction, and the more recently emergent generative AI, which holds promise for novel applications in research like hypothesis generation, clinical trial design, drug development, literature synthesis and writing support. Discriminative and generative AI differ in how they process data and apply their learning algorithms. While discriminative models focus on finding decision limits to predict labels, generative models analyse the underlying data distribution aiming to generate new data. Figure 1 summarises the main models used by these technologies and applications for research that we will explore in this review.

Discriminative AI includes a wide range of capabilities, such as distinguishing data to make classifications or predictions, as well as performing tasks like outlier detection and clustering. Radiology exemplifies the significant impact of discriminative AI.[4] As a matter of fact, 723 of the 950 (76%) AI/ML (machine learning)-enabled medical devices approved by the Food and Drug Administration (FDA) as of August 2024 are related to this specialty.[5] Other notable examples can be found in ophthalmology, where algorithms have shown the ability not only to diagnose ocular pathologies with greater accuracy than expert

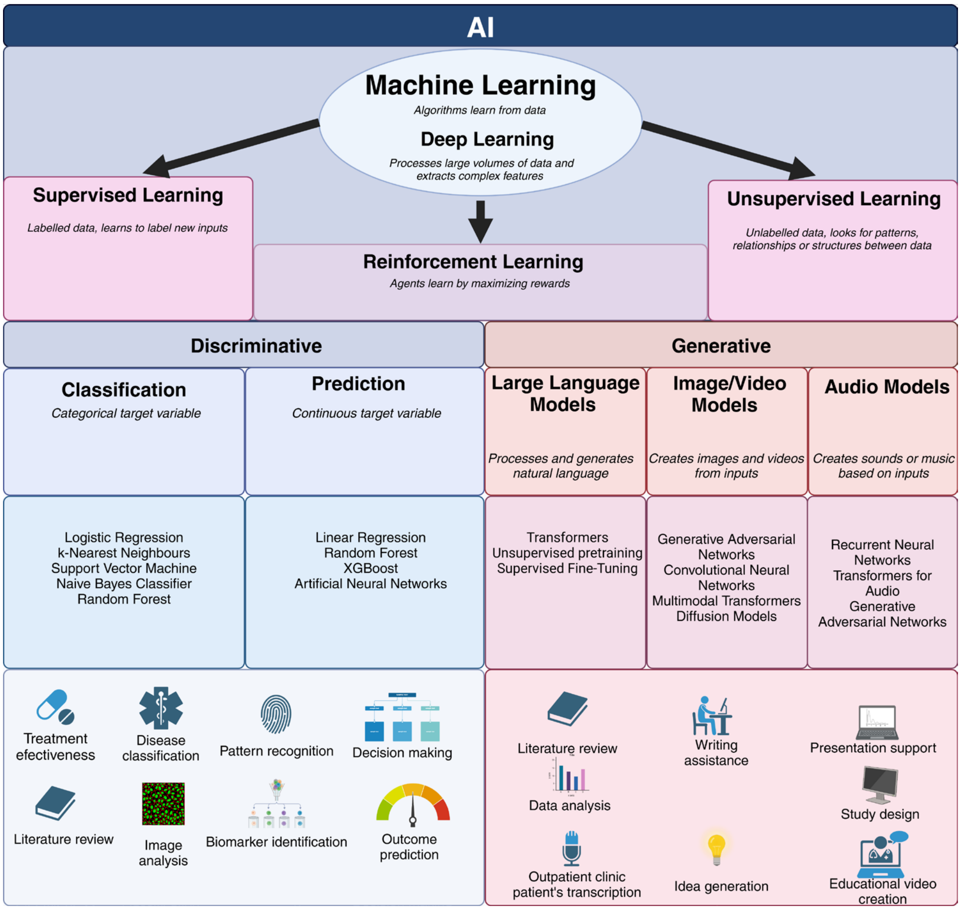

Figure 1 Main artificial intelligence (AI) models for rheumatology research.

ophthalmologists but also to predict cardiovascular risk factors undetectable by fundus examination.[6] Discriminative AI models have proven effective beyond image analysis, extending to fields like linguistics and other data types. For instance, a model used voice recordings and demographic data to predict dementia onset in patients with mild cognitive impairment, achieving around 80% accuracy.[7] Although no FDA-approved AI/ML applications currently exist in rheumatology,[5] numerous promising studies within the field of discriminative AI will be discussed in further detail.

Generative AI has recently transformed the AI landscape, particularly following the release of the chatbot ChatGPT in 2022, which has made AI more accessible to the general public.[8] Generative AI can create new content based on existing data from various sources, including text generation, image or video creation. The tools based on this technology have shown promising applications in medicine, including demonstrating clinical knowledge by successfully achieving high accuracy in standardised examinations.[9] Beyond knowledge-based tasks, ChatGPT's responses to patient questions were shown to be often preferred to those by doctors for their quality and empathy.[10] These results are remarkable given that the ChatGPT model is general-purpose and not specifically designed for medicine. The accuracy and reasoning skills of large language models (LLMs) have also been demonstrated in clinical examinations. Research has led to the development of specialised medical models like Med-Pathways Language Model (PaLM2), which achieved an accuracy similar to clinician answers (both exceeding 90%) in answering medical questions after being trained on six medical question-answering datasets.[11]

Rheumatology is a rapidly evolving specialty, thanks to the advent of advanced therapies and new technologies. Focusing on the management of chronic diseases with potential systemic involvement, it is a field rich in data and complex decision-making. Therefore, the use of AI tools holds the promise to transform clinical practice, leading to more informed decision making. Beyond the diagnostic level, there are promising predictive capabilities. In this regard, AI can assess disease activity, predict flares, determine optimal treatment dosages and anticipate patient responses based on clinical and serological biomarkers.[12] Moreover, generative AI can function as a clinical decision support system, assist with administrative tasks, and enhance the quality of patient information and education.

The aim of this review is to highlight the current and future applications of AI in rheumatology, examine the mechanisms of AI, analyse state-of-the-art investigations and explore its integration into daily research practice. To achieve this, we conducted a narrative review including an electronic search in Medline and

Embase for English-language sources from inception to September 2024. We employed a range of free-text terms including, but not limited to: “Artificial intelligence AND rheumatology”, “Machine learning AND rheumatology”, “Deep learning AND rheumatology”, “(Machine learning OR Deep learning) AND (rheumatoid arthritis OR spondyloarthritis OR psoriatic arthritis OR osteoarthritis OR lupus OR Sjogren)”, “Large language models”, “Natural language processing”, “(Predictive modeling OR Electronic medical records OR Risk stratification) AND rheumatology”, “(Large language models OR Natural language processing) AND rheumatology”, “ChatGPT AND rheumatology”. Furthermore, we conducted a manual search by examining the references cited in the included studies and technical computer science books. Priority was given to seminal references or those published within the last 2 years.

KEY AI CONCEPTS FOR RHEUMATOLOGY RESEARCH

The integration of AI into rheumatology research is becoming increasingly relevant, as the availability of complex datasets and advanced computation redefine how we approach and conduct scientific investigations.[13] Given its capacity for use in research, AI-related concepts can help rheumatologists to effectively use these technologies in their work. Table 1 summarises the core principles in the most widely used AI algorithms, as well as their application in rheumatology.

An *AI algorithm* is a computational model designed to perform tasks by learning from data and identifying patterns, rather than relying solely on a predefined set of rules or instructions. These algorithms may improve their performance over time through experience, which is gained through an iterative process. These algorithms are typically used for classification or predictive purposes, such as diagnostic or prescriptive applications in medicine. This can be achieved by analysing large datasets, identifying relevant features and applying learnt patterns to new, unseen data.[13]

ML is a branch of AI that operates by feeding an algorithm with input data that reflects past observations, enabling it to construct a model to assess new, previously unseen observations. ML algorithms can be classified into four main types according to their training: supervised, unsupervised, self-supervised and reinforcement learning.[13] Supervised algorithms are trained on a dataset where the output results are known and are used to label the outcomes. These have been the most used for clinical research. Unsupervised algorithms work with unlabelled data to identify patterns or clusters within datasets, making them useful for exploratory data analysis. There are two main types: clustering, which groups similar data points (eg, K-means, DBSCAN, hierarchical clustering), and dimensionality reduction, which simplifies data by reducing the number of features while preserving essential information (eg, PCA, t-SNE). These methods help uncover hidden structures without the need for labelled examples.[14] Self-supervised learning creates internal labels within an unlabelled dataset, allowing models to learn without external annotation and guidance.[15] Reinforcement learning adapts dynamically using reward-based feedback to maximise the performance of the algorithm.[13]

Deep learning (DL) is a subtype of ML that involves neural networks. A neural network is a particular ML algorithm based on successive layers of data transformation, inspired by the neural connections in the human brain. Neural networks are particularly effective with large volumes of data and demand significant processing power, which can be provided by processing units working in parallel. DL uses a high number of neuron layers, allowing for multiple levels of abstraction and has achieved noteworthy results in various applications, including text and image recognition.[16] Transfer learning enables the adaptation of a DL model to specific imaging tasks (eg, rheumatological imaging classification) by leveraging pre-existing knowledge from extensive, non-specialised image datasets, enhancing model performance and reducing the need for large, specialised training datasets. Applications of DL include image recognition and natural language processing (NLP), which use images and text as input data, respectively.[17] Indeed, one type of deep neural network algorithm gave birth to transformer technology.[11 18] Transformers have revolutionised NLP with the so-called self-attention mechanism, which allows for capturing relations between words, allowing for efficient and accurate text generation. The seminal paper on this technology has garnered 140 000 citations by November 2024, reflecting the significant impact of language models on society.[18]

LLMs are advanced neural networks based on the transformer architecture.[18] They are pretrained on vast amounts of unlabelled text data, typically sourced from the web, using self-supervised learning. This self-supervised learning involves predicting the next word in a sentence given the previous words (context), for which the model uses the surrounding context as the signal to learn and improve.[17] These models are fine-tuned for specific tasks like question-answering and named entity recognition, showing their versatility and effectiveness in language understanding and generation. When models are able to process and integrate multiple types of data such as text, images and audio, this is known as multimodality.

Validation is a process that ensures a model’s generalisability and reliability by assessing its performance on unseen data before it is deployed in real-world applications.[19] Evaluating discriminative AI models involves metrics familiar to rheumatologists, such as sensitivity (also known as recall in the field of ML) and specificity, which assess the ability to correctly identify true positives and true negatives.[19] Precision, similar to positive predictive value (PPV), measures the proportion of true positives among all positive predictions, while the F1 score combines precision and recall into a single measure.

Table 1 Applications of artificial intelligence models in rheumatology

AI model	Description	Examples of use in research	Studies in the field of rheumatology
Logistic regression	Uses a logistic function to model binary dependent variables	▶ Disease diagnosis ▶ Outcome prediction	Prediction of relapses in RA[45] Diagnosis of SpA[51] Diagnosis of systemic autoimmune diseases[58] Prediction of hospitalisations in SLE[25] Prediction of mortality in systemic sclerosis[34]
Linear regression	Analyses the relationship between a dependent variable and one or more independent continuous variables	▶ Outcome prediction ▶ Decision support ▶ Risk factor analysis	Prediction of response to methotrexate in RA[30] Prediction of response to bDMARDs in RA[32]
Support vector machine	Analyses data for classification and regression analysis by finding the hyperplane that best divides a dataset into classes	▶ Treatment optimisation ▶ Outcome prediction ▶ Image segmentation and anomalies detection	Fatigue prediction in RA through brain MRI[42] Prediction of complications during pregnancy prediction in SLE[26]
Decision tree	Employs a tree-structured approach for decision-making, representing decisions and their possible outcomes, including chance events	▶ Decision support ▶ Outcome prediction	Diagnose and prediction of D2T RA[23] Prediction of response to bDMARDs in RA[32] Prediction of complications during pregnancy in SLE[26]
Random forest	Implements an ensemble learning method for classification, regression and other tasks, using multiple decision trees to improve predictive accuracy	▶ Disease prediction ▶ Risk factor identification ▶ Outcome prediction ▶ Data imputation	Prediction of response to DMARDs in RA[31] Prediction of hospital readmission in SLE[33] Prediction of mortality in systemic sclerosis[34] Prediction of response to methotrexate in RA[30]
Naive Bayes	Applies probabilistic classification based on Bayes' theorem, assuming independence between features	▶ Disease diagnosis ▶ Patient stratification ▶ Outcome prediction	Systemic sclerosis mortality prediction[34] Predict hospitalisations in SLE patients[25] Diagnosis of RA[50]
K-nearest neighbour	Uses a non-parametric method for classification and regression, basing predictions on the k closest examples in the feature space	▶ Classification/clustering ▶ Pattern recognition	Prediction of complications during pregnancy in SLE[26] Diagnose and prediction of OA with MRI[44] Diagnosis of RA with thermography[49]
K-means	Clusters unsupervised data based on partitioning a dataset into a specified number (K) of distinct clusters based on the similarity of data points	▶ Risk stratification ▶ Treatment patterns ▶ Image analysis ▶ Clinical trial design	Classification of clusters in SpA[24]
XGBoost	Gradient boosting algorithm that combines sequential decision trees to improve accuracy, optimised for efficient classification and regression on large datasets	▶ Disease diagnosis ▶ Outcome prediction ▶ Biomarker identification	Prediction of relapses in RA[45] Prediction of response to bDMARDs in RA[32] Diagnosis and prediction of D2T RA[23]
Recurrent neural networks	Processes sequential data by maintaining a temporal memory of past inputs. They use recurrent connections to propagate information from previous time steps, allowing them to capture dependencies in sequences	▶ Text analysis ▶ Time series analysis	Assessment of prevalence and disease management of RA-ILD[54] EHR Data analysis in SpA[57] EHR diagnosis in PsA[52]
Convolutional neural networks	Designed for processing high-dimensional data such as images by using convolutional layers to hierarchically extract spatial features from input data	▶ Medical image analysis ▶ Disease diagnosis ▶ Segmentation	Diagnosis of RA[50] Detection of SpA-related lesions via MRI[43] Diagnosis of GCA through US[46]
Transformers	Use an attention mechanism to process entire sequences in parallel, efficiently capturing long-range dependencies	▶ Clinical text analysis ▶ Medical report summarisation ▶ Drug discovery	Summarising information, aiding in the composition of clinical notes[64] Disease diagnosis support[65 68] Efficiency in drug design[27]

ANCA, anti-neutrophil cytoplasmic antibody; bDMARDs, biological disease-modifying antirheumatic drugs; DMARDs, disease-modifying antirheumatic drugs; EHR, electronic health record; GCA, giant cell arteritis; ILD, interstitial lung disease; OA, osteoarthritis; PsA, psoriatic arthritis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; SpA, spondyloarthritis; US, ultrasound.

Accuracy, which represents the overall correctness of the model's predictions, reflects the proportion of true results (both true positives and true negatives) among the total cases. The area under the receiver operating characteristic curve (AUC-ROC) offers a visual summary of the model's performance across thresholds, with the area reflecting the ability of the model to distinguish between classes.[19]

In evaluating generative AI models, additional metrics provide an understanding of model performance besides accuracy.[20] Perplexity measures the model's ability to predict the next word in a sequence, with lower scores indicating more precise predictions. Bilingual Evaluation Understudy assesses the similarity between the model-generated text and a human reference by comparing overlapping word sequences. Recall-Oriented Understudy for Gisting Evaluation measures the degree of n-gram overlap between generated and reference texts, emphasising recall. BERTScore further enhances evaluation by using Bidirectional Encoder Representations from Transformers (BERT) embeddings to compare the semantic similarity between generated and reference texts, capturing context-sensitive alignment. These complementary metrics allow for a comprehensive assessment of generative AI in clinical applications.[21]

DISCRIMINATIVE AI IN RHEUMATOLOGY

Discriminative AI algorithms are focused on two main objectives. Classification analysis aims to evaluate previously described phenomena, attempting to describe features and ideally associations between risk factors (independent or predictor variables) and outcomes (dependent variables or events). On the other hand, predictive analysis, aims to forecast future events. Traditionally, this has been achieved using regression methodologies, including linear, logistic or Cox regression.[19] Recently, AI has been employed to classify diseases and predict their progression using ML algorithms. These approaches may use various data types, including structured data, images and free-text information. Interpreting performances across studies is complex due to variations in datasets, patient cohorts and study outcomes, making direct comparison challenging. While metrics such as AUC and F1 metrics offer insights into model performance, their practical value depends on whether these AI advances lead to real-world clinical benefits. Validations against conventional models and interventions remain essential to establish AI's utility and ensure its impact on patient care.

The application of AI in analysing structured data is advancing diagnostic accuracy, risk prediction and patient management in rheumatic and RMDs. In rheumatoid arthritis (RA), for example, a neural network model trained on demographic and laboratory data (including age, sex, rheumatoid factor, anti-citrullinated cyclic peptide and anti-carbamylated protein) achieved an F1 score of 0.92 in diagnosing RA,[22] demonstrating accuracy comparable to, or exceeding, conventional diagnostic approaches. In predicting difficult-to-treat (D2T) RA, an extreme gradient boosting (XGBoost) model combined structured and unstructured data from 1873 patients, achieving an AUC-ROC of 0.88 for D2T identification and 0.73 for future D2T development prediction.[23] In combination with structured data, integrating unstructured data—such as clinical notes and imaging—further enhances AI models' ability to predict complex outcomes, as demonstrated in the previous study identifying RA subsets like D2T RA. Collectively, these applications underscore AI's potential to exceed or complement standard statistical approaches by improving accuracy, sensitivity and specificity across RMD diagnostic and prognostic tasks.

For prognostic applications, structured data analyses have provided insights across various RMDs. In spondyloarthritis (SpA), K-means clustering applied to a longitudinal dataset identified two distinct disease activity trajectories—one with persistently high activity and another evolving to low activity—highlighting potential therapeutic approaches based on trajectory patterns.[24] Similarly, in systemic lupus erythematosus (SLE), a random forest (RF) model predicted hospitalisations with an AUC-ROC of 0.75, using clinical markers such as dsDNA positivity, C3 levels, blood cell counts, inflammatory markers and albumin.[25] Additionally, in pregnancy outcomes for women with SLE, a pre-pregnancy RF model achieved an AUC-ROC of 0.92, demonstrating high sensitivity (0.89) and specificity (0.94) in identifying adverse outcomes, a notable improvement compared with traditional models.[26]

The use of structured data and their analysis through AI has become an important axis in drug development and molecule generation, mainly based on ML and DL algorithms.[27] Some of the use cases aim to identify drug targets and binding sites as well as to predict chemical properties (affinity, ability, lipophilicity, solubility, toxicity) of a compound. ML and DL algorithms may be used for efficacy evaluations of drugs through big data modelling and analysis.[28] A relevant advancement in this regard has been conducted by AlphaFold, developed by DeepMind, in predicting structures of proteins. This enabled researchers to understand molecular targets more precisely, therefore supporting in identifying binding sites, refining drug designs and predicting protein interactions.[29] Additionally, AI may assist clinical trial design and implementation by supporting selection of promising lead molecules based on patient-specific profiles, identifying suitable patient profiles and improving recruitment for clinical trials.

Predicting the suitability of treatments is crucial for improving research and clinical practice. One study used an ML model to predict methotrexate (MTX) response in RA patients using clinical data. A *Least Absolute Shrinkage and Selection Operator* algorithm, a method for fitting linear models, was employed in this project, achieving better performance than RF, with an AUC-ROC=0.79 (vs 0.68 in RF); this effective categorisation of patients into good and poor responders was achieved with baseline Disease Activity Score 28 (DAS-28), anti-citrullinated protein antibody and Health Assessment Questionnaire as top predictors.[30] Combining clinical data with genomic biomarkers (single-nucleotide polymorphisms) and baseline DAS-28 has also shown promise in predicting MTX response in early RA; metrics of different supervised ML methods showed an AUC-ROC=0.84 in the training cohort, and a validation cohort accuracy of 0.76.[31] Similarly, different ML models (linear regression, random forest, XGBoost and CatBoost) were evaluated for their ability to predict the probability of therapeutic response for bDMARDs in RA in the ESPOIR cohort, predicting response to tumorous necrosis factor inhibitors with an AUC-ROC of 0.72 (0.68 to 0.73), and yielding key predictors such as DAS28, lymphocytes, aspartate aminotransferase, neutrophils, age, weight and smoking status.[32]

Models aiming to predict the readmission risk of patients with RMDs after discharge, or the evolution of a given disease, have also been developed. By analysing data from electronic health records (EHRs), RF-based models aiming to predict patient's return to the clinic achieved an AUC-ROC of 0.65, a sensitivity of 0.38 and a specificity of 0.79; follow-up duration, the prescription of DMARDs, corticosteroids, diagnosis of chronic polyarthritis, quality of life and patient occupation were identified as key variables.[33] In the context of life-threatening illnesses such as systemic sclerosis, predictive modelling has been employed to estimate mortality rates drawing on clinical, demographic and spirometric data.[34] The Naïve Bayes Classifier, a supervised ML algorithm, achieved an AUC-ROC=0.76 to predict 5-year mortality rates after internal cross-validation, which demonstrated superior predictive capability as compared with other algorithms, including RF (AUC-ROC=0.73), logistic regression (AUC-ROC=0.75) and Cox regression (AUC-ROC=0.724).[34]

Concerning *imaging*, studies have used various techniques from simple to complex. Using X-rays, ML models achieved up to 90.7% accuracy in distinguishing RA and OA from normal hand radiographs, though accuracy decreased (80.6%) when classifying all three classes together.[35] Moving on to osteoarthritis (OA), a DL model was trained on knee radiographs to identify patients with and without pain progression, as measured by the Western Ontario and McMaster Universities Arthritis Index (WOMAC) pain score.[36] The DL model achieved an AUC-ROC=0.80 in predicting pain progression, significantly higher (p<0.001) than a traditional model trained on demographic, clinical and radiographic risk factors. In axial imaging, a neural network based on 1553 pelvis X-rays evaluating the presence or absence of definite radiographic sacroiliitis as agreed in a central reading session, identified definite sacroiliitis with an AUC-ROC=0.94, a sensitivity of 0.92 and a specificity of 0.81 for the test dataset.[37] CT has also benefited from AI, where neural networks trained on CT-derived 3D joint shapes distinguished hand joint patterns in RA with AUC-ROC=75%, psoriatic arthritis (PsA) with AUC-ROC=68% and healthy controls with AUC-ROC=82%. These models additionally identified disease-specific regions prone to erosions and bony spurs, contributing to classifying undifferentiated arthritis.[38] Convolutional neural networks (CNNs) trained on sacroiliac joint images detected structural lesions such as erosion and ankylosis, achieving sensitivities of 0.95 and 0.82 and specificities of 0.85 and 0.97.[39] Additionally, in Sjögren's syndrome, a DL model using 500 CT images detected salivary gland damage in parotid glands with 96% accuracy, comparable to diagnosis of experienced radiologists.[40]

Regarding MRI, CNNs have also demonstrated the ability to differentiate between patients with RA and PsA based on patterns from hand MRIs, achieving AUC-ROC=0.75 for seropositive RA versus PsA, 0.74 for seronegative RA versus PsA and 0.67 for seropositive versus seronegative RA. Interestingly, adding demographic or clinical data to the networks did not provide improve classification.[41] Non-articular MRI applications, such as brain MRI, have been used to evaluate fatigue in RA, showing that brain structural metrics were superior to clinical measures, with the highest prediction accuracy reaching 0.67.[42] In SpA, MRI models have been developed to detect sacroiliac joint active damage. In fact, a deep neural network developed to detect MRI changes in sacroiliac joints indicative of axial SpA (axSpA) achieved a sensitivity of 0.88 and specificity of 0.71 for detecting inflammatory changes, and a sensitivity of 0.85 and specificity of 0.78 for structural changes in external validation.[43] A multi-purpose MRI-based model using compound image transformations analysed knee cartilage in T2-weighted images to predict progression to symptomatic OA with an accuracy of 0.75, as defined by the WOMAC score 3 years post-baseline.[44]

Other imaging modalities, such as ultrasound, have demonstrated potential in predicting RA relapses and assessing joint conditions. A study comparing three ML classifiers found XGBoost to be the best-performing model (AUC-ROC=0.75), identifying 10 key features, including superb microvascular imaging scores of wrist and metatarsophalangeal joints.[45] On vasculitis ultrasound, a study assessed the use of a CNN for detecting the halo sign in colour Doppler images for diagnosing giant cell arteritis, achieving an AUC-ROC=0.84 on the test set, with a 0.95 specificity and 0.60 sensitivity.[46] For Sjögren's syndrome, DL models used transfer learning to improve the automated segmentation of salivary gland ultrasonography, achieving a higher Intersection-over-Union (0.85) compared with both inter-observer agreement (0.76) and intra-observer agreement (0.84), indicating superior accuracy and consistency.[47] Thermography, combined with AI, can detect RA activity by analysing temperature changes in hand joints. An ML-based method, ThermoJIS, for detecting joint inflammation in RA using hand thermography, correlated moderately with ultrasound scores and demonstrated with good diagnostic performance (AUC-ROC=0.78).[48] Building on this, the study developed and validated two composite disease activity indices, ThermoDAI and ThermoDAI-CRP, which showed stronger correlations with ultrasound-determined synovitis (GS=0.52–0.58; PD=0.56–0.61) compared with patient global assessment (PGA) and PGA+CRP, and strong correlations with clinical indices (ρ>0.81).[49]

In the context of *text analysis*, discriminative AI using NLP has aided the analysis of vast amounts of EHRs, including tasks such as disease identification and clinical characteristics assessments. Several studies highlight NLP's utility in rheumatology for disease detection. For instance, a validated ML pipeline identified RA patients with high performance, with support vector machines (AUC-ROC=0.98, F1 score 0.83) and gradient boosting (AUC-ROC=0.94, F1 score 0.82) outperforming simpler word-matching methods.[50] Other study demonstrated the capability to identify axSpA through an unsupervised algorithm, incorporating both the NLP concept and ICD codes, with a sensitivity of 0.78, a specificity of 0.94 and an AUC-ROC of 0.93.[51] This has also been explored in PsA, in which a sensitivity of 0.79 and a PPV of 0.93 were achieved when NLP was combined with billing codes.[52] Further, a tool combining text mining with NLP-based exclusion accurately identified ANCA-associated vasculitis cases, achieving a PPV of 0.86 and outperforming traditional ICD-10 coding.[53] This growing body of evidence supports the adoption of NLP technologies in accurately identifying RMDs.

Additional studies have focused on extracting clinical information beyond diagnoses from EHRs. For example, a recent study that included a dataset with around 64 million EHRs focused on the demographic and clinical characteristics of RA patients with interstitial lung disease (RA-ILD), yielding relevant information on prevalence, comorbidities and drug use in real life, with a high precision (F1 score over 0.7) for most of the assessed variables.[54] Another algorithm extracted forced vital capacity from EHRs, strongly correlating (r=0.94) with pulmonary function test values.[55] In RA, a study identified MTX-induced liver toxicity using NLP with a string-matching algorithm, achieving a PPV of 0.76.[56] In another study, the analysis of structured and free-text EHR data from three hospitals showed limited disease activity evaluations in axSpA and PsA patients.[57] For systemic autoimmune rheumatic diseases, an ML model predicted autoantibody testing needs and specialist referrals in systemic autoimmune diseases with AUC-ROC values from 0.91 to 0.94, enabling early detection up to 5 years before diagnosis.[58] Another example illustrating the potential of AI in using large-scale real-world data is EPIC Cosmos, a vast inter-hospital database aggregating de-identified EHR from millions of patients across multiple health systems.[59] EPIC Cosmos has enabled studies in different fields including rheumatology, such as recent work on SLE where researchers used Cosmos to enhance disease phenotyping and diagnosis. This study applied ICD codes to identify SLE patients and validated data quality against EULAR/ACR classification criteria, highlighting the need for integrating clinical notes to improve data completeness beyond structured EHR fields. While the study primarily relied on structured ICD codes for SLE phenotyping, the authors acknowledge plans to develop an NLP pipeline to analyse clinical notes, aiming to improve data completeness.[60]

GENERATIVE AI IN RHEUMATOLOGY

Generative AI, the latest advancement in AI, is an emerging technology capable of creating new content in audio, image, video and text formats. It is based on foundation models—large-scale AI systems that acquire emergent capabilities across domains such as language, vision, robotics, reasoning and interaction. Their versatility allows them to adapt to diverse tasks, from NLP to computer vision and robotic control, by leveraging unlabelled data and self-supervised learning techniques. Among these, text-oriented applications have shown the most potentialities for research. At the core of generative AI are LLMs, which use transformer architecture to generate human-like responses based on input data.[18] LLMs process and analyse input to generate outputs that mimic human reasoning based on statistical correlations, a capability that distinguishes them from discriminative AI, whose models produce a label or category based on the input, requiring explicit interpretation of the results.[61] An example of interacting with these models is through widely recognised chatbots such as ChatGPT by OpenAI or Gemini by Google.[62]

Clinical workflow and decision-making

Generative AI has yielded some results in research on clinical practice use, though its applications are still in the early stages. Current LLMs fine-tuned on medical data such as Med-PalM or Meditron show promise, nearing expert human performance in answering medical questions, which could serve as a decision support; nonetheless, they may fall short when addressing individual patient circumstances.[11 63] Moreover, LLMs can significantly reduce administrative burdens by summarising and rephrasing information, aiding in the composition of clinical notes and discharge reports with real-time suggestions.[64] Future developments will likely see major software companies integrating LLMs into administrative workflows, serving as clinical decision support systems and automating tasks such as documenting information from consultations, video calls and emails.

Concerning disease diagnosis, LLMs have demonstrated significant results. One study conducted in early 2023 with ChatGPT-3.5 highlighted its strong performance across various clinical tasks, achieving an overall accuracy of 76.9% in making final diagnoses.[65] The multimodal ChatGPT-4 has shown diagnostic capabilities in musculoskeletal radiology, performing at a level comparable to radiology residents when inputting the medical history and imaging findings (accuracy rates of 43% vs 41%) but not matching board-certified radiologists (53%).[66] Interestingly, its text-based diagnostic performance surpassed that of its vision-based counterpart (*Vision* ChatGPT4 version) when processing radiology findings rather than images.[66]

LLM performance has also been assessed in comparison to physicians for differentiating inflammatory rheumatic diseases from non-inflammatory conditions, highlighting its capacity to generate diagnostic insights through pattern recognition in language. ChatGPT-4 correctly identified the most likely diagnosis in 35% of cases, closely matching rheumatologists' 39% (p=0.30).[67] In cases of inflammatory rheumatic disease, ChatGPT-4 performed better, with 71% accuracy versus 62% for rheumatologists. However, it was less accurate in non-inflammatory rheumatic cases.[67] Other LLM-based applications, such as DxGPT, have shown relatively high accuracy in diagnosing rare diseases.[68] This decision support tool revealed that models like Claude 3 Opus achieved 55% strict accuracy and 70% top-5 accuracy using real-world datasets of rare diseases. While these findings highlight the capacity of AI to assist rheumatologists in diagnosing non-prevalent conditions, further validation in clinical settings is essential.[68]

Indeed, as a decision support tool, ChatGPT has proven useful and reliable for answering questions about some RMDs. In a study evaluating LLMs on MTX information for RA, GPT-4 achieved 100% accuracy and completeness, with all 23 MTX-related responses correct and complete as evaluated by two reviewers. In contrast, BARD (now Gemini) scored 73.9% correct answers.[69] In the ChatSLE study, ChatGPT-4 was evaluated against leading rheumatology experts, providing answers to 100 patient-related questions from Lupus100.org.[70] ChatGPT-4's responses were rated as high quality, with a mean quality score of 4.55 (95% CI 4.48 to 4.62) compared with 4.31 (95% CI 4.23 to 4.39) for expert responses ($p<0.0001$). Both sources showed similar empathy scores, but ChatGPT-4 was preferred in 57% of cases ($p=0.01$). Additionally, ChatGPT-4 provided relatively accurate patient information, with a mean score of 8.4±0.7 on a 0–10 scale.[70] Further studies have evaluated ChatGPT's reliability and utility in providing information on common RMDs. For instance, an assessment of ChatGPT's responses regarding conditions such as RA, AS and OA on a 7-point Likert scale, found that ChatGPT achieved the highest reliability score for OA (mean±SD 5.62±1.17), indicating that while the model is a promising tool, clinicians should remain vigilant of its probability to provide misleading information.[71]

Some studies have compared the performance of models in rheumatology. The recent Rheum2Guide study compared treatment plans generated by GPT-4 and GPT-3.5 with those created by a clinical rheumatology board using 20 fictional patient vignettes.[72] GPT-4's plans were selected more frequently than GPT-3.5's for first-line treatments, indicating GPT-4's closer alignment with clinical expectations. Although GPT-4 and GPT-3.5 generated safe and high-quality treatment plans, the rheumatology board's plans were preferred in 68.8% of cases due to higher ratings in guideline adherence, medical appropriateness, completeness and overall quality.[72] Another study evaluated the diagnostic capabilities of ChatGPT-4 and other LLMs like Claude 1.3, Claude 2 and Bard using standardised prompts in The Lancet's Picture Quiz Gallery focused on rheumatic diseases—including the text and not images as part of the input. ChatGPT-4 and Claude 2 both achieved 81% accuracy, outperforming Claude 1.3 (72%) and Bard (66%). However, all models, except Claude 2, struggled with cases involving uncommon infectious diseases, where ChatGPT-4's accuracy dropped to 57%.[73]

The accuracy and reasoning skills of LLMs have also been demonstrated in challenging clinical examinations, in which could be used to aid medical education. For instance, ChatGPT-4 has repeatedly demonstrated proficiency in standardised tests like the US Medical Licensing Examination, where it provided coherent and intuitive responses, surpassing the performance of previous earlier AI systems.[64] Additionally, it successfully answered 93.7% of all rheumatology-related questions from the Spanish Medical Training Examination (MIR) within the years 2009–2023, with a median clinical reasoning score of 4.67 on a 5-point Likert scale, outperforming earlier LLM versions.[74]Although currently the use of AI in day-to-day clinical practice may represent a complement rather than a stand-alone solution, it has been shown that the clinician with AI support versus traditional methods does not improve the situation, but AI alone has shown better results than previous groups, so the potential of this technology will be derived based on the clinician's learning to use it.[75]

Drug development, clinical trials and digital twins

Drug development has leveraged generative AI for molecule generation and molecular property prediction. For instance, BERT—a transformer-based model—has been adapted to learn molecular representations, supporting drug discovery tasks. Similarly, other language models have been fine-tuned for molecule generation and annotation, significantly enhancing efficiency and accuracy in drug design.[27]

In clinical trials, generative models can create synthetic data that closely mirrors real-world data, as illustrated with digital twins (DTs). Pretrained on patient vitals, clinical trajectories, lab results and diagnoses, DTs simulate patient evolution over time based on treatment decisions. Indeed, DT may facilitate the creation of synthetic control arms, which can replicate patient groups for comparative analyses without recruiting additional participants. External controlled arms based clinical trials have been supported by both the FDA and EMA; in rheumatology, these approaches have been applied to research in RA.[76 77]

Optimising the research process

AI, particularly through LLMs, has the opportunity to transform research by offering advanced tools that can support every stage of the process. Central to adopting these capabilities is the concept of *prompting*—the process of giving instructions to AI systems. Prompts can range from simple, direct queries to complex, structured inputs designed to elicit detailed responses.[78] In research, prompting can be executed as 'zero-shot' learning, where the AI is given a task without any prior example or training; more refined prompts can guide AI to produce more focused and relevant information based on

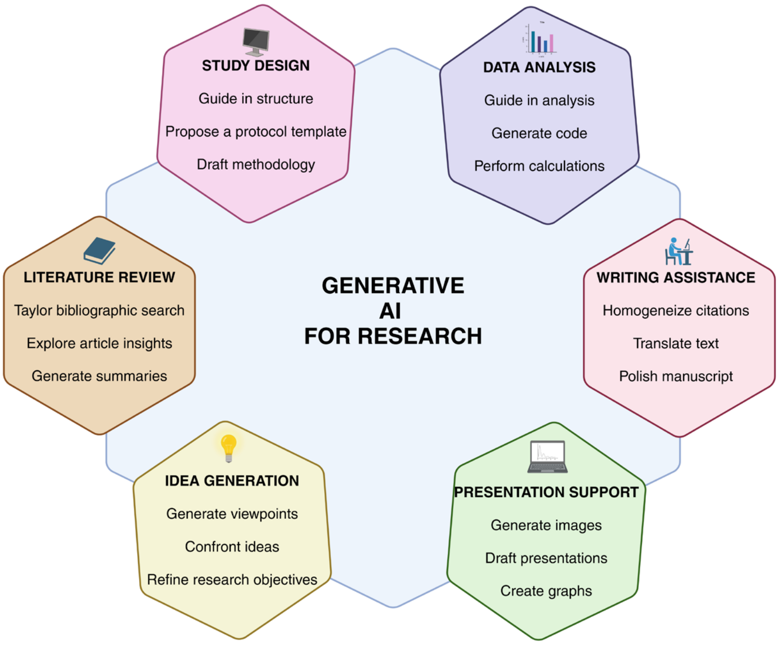

Figure 2 Generative applications of artificial intelligence (AI) in research.

examples, such as few-shot prompting (giving examples) or chain-of-thought (providing step-by-step) prompting.

The role of AI in all aspects of research goes from idea generation and literature review to data analysis and manuscript preparation (figure 2). In the early stages of research, AI can significantly help via brainstorming sessions, in which a wide range of ideas and hypotheses can be explored.[79] LLMs such as ChatGPT-4, Gemini, Perplexity and Claude are capable of generating diverse perspectives on a given research question, helping to promote the creativity that helps refine research objectives.[62] These tools allow researchers to quickly iterate their ideas, explore conceivable investigative angles and develop a clear roadmap for studies. A recent study found that an AI model generated research ideas rated as more original and exciting than those of human scientists, though with slightly lower feasibility. Using the Claude 3.5 model, researchers produced 4000 ideas across several topics, and reviewers assessed these ideas without knowing their source. Despite AI's high novelty scores, only about 200 ideas were genuinely unique, with creativity diminishing over time.[80] The overwhelming abundance of options produced by AI challenges conventional creative processes, pushing researchers to shift from seeking single insights to generating numerous ideas for refinement. Rather than asking for one idea, AI enables researchers to request many altogether, allowing to sift through diverse suggestions and strategies. This surplus demands the skill of curating and discerning the best quality, which highlights a core value of AI in augmenting intellectual creativity and decision-making in research. As the research project progresses, AI tools can assist in refining the design and structure of the study. Beyond generating ideas, these systems can suggest detailed article structures, and help in drafting sections of a manuscript.

Conducting a literature review can also benefit from AI support. Tools such as Elicit or Research Rabbit provide curated bibliographies by searching with NLP.[81 82] In addition, they can offer insights into the state-of-the-art of research concepts and visualise the relationships between key studies, potentially uncovering connections between research articles that may not be immediately apparent. They can also generate summaries from articles, extract precise data and even highlight emerging trends in the field, enabling researchers to stay ahead in their field.

AI tools can also take on a more active role during the data analysis phase. LLMs can assist in data analysis by guiding researchers through their analysis, performing statistical tests and generating insights from datasets. For instance, ChatGPT-4, can provide AI-generated code for statistical tools like SPSS, R or Python that facilitates the execution of complex analyses. Moreover, it can directly perform advanced computational calculations and data queries by inputting the prompt and the dataset to the system. As an additional support, it can also provide preliminary interpretations of data and give insights on the results. A recent study of 187 489 software developers using GitHub Copilot demonstrated how AI tools can reshape work by shifting focus from non-core management tasks to primary tasks, such as coding. This shift allowed developers to work more autonomously, explore new methods, and potentially reduce hierarchical dependencies.[83]

Finally, the writing phase—often the most time-consuming—can be facilitated using AI. Besides

ChatGPT or Gemini, other tools such as Jenni AI can suggest article structures, and help draft coherent and well-organised manuscripts.[84] These platforms can assist with translation, paraphrasing and ensuring that the text adheres to publication standards. As for the presentation of the results, image models can support on creating graphs, images or presentations. For example, platforms like Microsoft Copilot can create a presentation of the research.[85]

Challenges ahead

The integration of AI into the field of rheumatology research is a double-edged sword, offering significant chances alongside profound ethical and practical challenges.

AI models have demonstrated high accuracy in diagnosing and predicting outcomes of rheumatic diseases, sometimes even surpassing traditional methods. Predictive analytics can identify patients at higher risk of disease progression, facilitating proactive management. Nonetheless, accuracy and reliability of these models in clinical practice is yet to be explored. While there are some studies including external validation of the algorithms, clinical trials assessing the efficacy of these models in randomised controlled trials are lacking in rheumatology.

Another primary concern is the rapid pace at which AI is being adopted, particularly as health systems deploy AI-driven support tools with minimal clinician training. Without structured guidance, clinicians may struggle to use these tools effectively, which could limit their impact on diagnostic accuracy. A recent randomised trial demonstrated that access to an LLM alone did not improve physicians' diagnostic reasoning in challenging cases, even though the LLM performed well when operating independently. Unexpectedly, the LLM alone significantly outperformed physicians in diagnostic reasoning for complex cases.[75] This finding suggests that simply having access to AI tools does not inherently enhance clinical reasoning skills and that effective use of these tools requires comprehensive training.

Besides, flawed training data may possibly lead to algorithmic bias; AI models trained on non-representative data might produce skewed outcomes, disadvantaging certain patient groups.[83] As an example in SLE, AI models trained predominantly on data from non-Hispanic white populations may produce less accurate predictions for under-represented groups, such as black, Hispanic or Asian patients, due to differing symptom patterns and disease progression, potentially leading to skewed outcomes in diagnosis and treatment.

Concerning generative AI, some researchers have raised concerns about the readiness of LLMs for medical application. For example, there have been instances where the unethical use of LLMs, such as generating fraudulent research or using undisclosed AI assistance in manuscript writing, has led to the retraction of scientific papers. Additionally, LLMs are prone to 'hallucinations', where they generate plausible-sounding but incorrect information, which is a matter of debate for clinical practice, where accuracy and evidence-based knowledge are paramount. In addition, the black-box nature of many AI algorithms also raises transparency issues, making it difficult for practitioners to understand and trust AI-generated insights.

To address these concerns, new reporting guidelines have emerged for both discriminative and generative models, such as TRIPOD-AI for validating AI interventions, CONSORT-AI for clinical trials, DECIDE-AI for decision support systems and CLAIM-AI for imaging technologies.[86–89] Additionally, guidelines like CANGARU have been developed specifically for generative AI models, reflecting the growing need for transparency and accountability in AI-driven research.[90] In addition, regulatory frameworks, such as the European Union's AI Act, set to be enforced in 2026, are now being formulated.[91] As we stand at the crossroads of innovation, regulation and ethics, the responsible evolution of AI in rheumatology requires a collective commitment from researchers to thoughtfully use these technologies, ensuring they enhance both research and patient care.

With an ageing population and a projected increase in RMDs, AI can assist in managing the growing demand on healthcare systems.[92] In this regard, AI may enhance collaboration between general practitioners and rheumatologists. Decision-support systems can aid in the early detection of RMDs at the primary care level, improving the accuracy and timeliness of referrals. Enhanced communication platforms can lead to a more integrated approach to patient care. Moving forward, it is crucial to balance the promising capabilities of AI with a mindful consideration of its limitations. Training healthcare professionals in AI technologies will facilitate their effective integration into clinical practice. Ongoing research is necessary to enhance the robustness of AI models and adapt them to the evolving needs of rheumatology.

CONCLUSION

The combined strengths of discriminative and generative AI are revolutionising rheumatology research. Discriminative AI's precise classification and prediction capabilities, paired with generative AI's ability to synthesise and create content, may provide rheumatology researchers with powerful tools to enhance their work. These advancements can accelerate the research process and therefore contribute to the development of efficient processes in rheumatology. However, as we integrate these technologies into our research, we must proceed with caution, balancing innovation with responsibility to maximise their prospective impact on the field. The future trajectory of AI in rheumatology is within our hands, with its ultimate impact determined by our collective efforts and thoughtful application.

X José Miguel Sequí-Sabater @drjsequirheum

Acknowledgements The authors thank the Spanish Foundation of Rheumatology for providing medical writing/editorial assistance during the preparation of the manuscript.

Contributors All authors have contributed to this work and approved the final version. DB as guarantor. In the preparation of this manuscript, artificial intelligence tools, specifically OpenAI's language model, was employed to assist in refining the language and structure of the text.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests JMS-S: speakers bureau: Abbvie, Galápagos, Theramex, UCB, Nordicpharma, Grunenthal, Novartis; Congress support: Abbvie, Galápagos, Nordicpharma, UCB, Novartis, Janssen; Advisory Boards: Abbvie. DB: speakers bureau: AbbVie, BMS, Galapagos, Janssen, Lilly, MSD. Research grants: Novartis. Consultancy: Sandoz, UCB. Part-time work in Savana Research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iDs

José Miguel Sequí-Sabater http://orcid.org/0000-0001-8437-1623
Diego Benavent http://orcid.org/0000-0001-9119-5330

REFERENCES

1 Kothari S, Gionfrida L, Bharath AA, *et al*. Artificial Intelligence (AI) and rheumatology: a potential partnership. *Rheumatol (Oxford)* 2019;58:1894–5.
2 Venerito V, Bilgin E, Iannone F, *et al*. AI am a rheumatologist: a practical primer to large language models for rheumatologists. *Rheumatol (Oxford)* 2023;62:3256–60.
3 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
4 López-Úbeda P, Martín-Noguerol T, Luna A. Radiology, explicability and AI: closing the gap. *Eur Radiol* 2023;33:9466–8.
5 FDA Food and Drug Administration. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. 2024. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
6 Poplin R, Varadarajan AV, Blumer K, *et al*. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018;2:158–64.
7 Amini S, Hao B, Yang J, *et al*. Prediction of Alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models. *Alzheimers Dement* 2024;20:5262–70.
8 Naveed H, Khan AU, Qiu S, *et al*. A comprehensive overview of large language models. *arXiv* [Preprint] 2023.
9 Garabet R, Mackey BP, Cross J, *et al*. ChatGPT-4 Performance on USMLE Step 1 Style Questions and Its Implications for Medical Education: A Comparative Study Across Systems and Disciplines. *Med Sci Educ* 2024;34:145–52.
10 Ayers JW, Poliak A, Dredze M, *et al*. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* 2023;183:589–96.
11 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature* 2023;620:172–80.
12 Hügle T. Advancing Rheumatology Care Through Machine Learning. *Pharmaceut Med* 2024;38:87–96.
13 Hügle M, Omoumi P, van Laar JM, *et al*. Applied machine learning and artificial intelligence in rheumatology. *Rheumatol Adv Pract* 2020;4.
14 IBM. What is unsupervised learning? 2024. Available: https://www.ibm.com/topics/unsupervised-learning
15 Jamaludin A, Kadir T, Zisserman A. Self-supervised learning for spinal mris. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. . 2017: 10553. 294–302. Available: https://neptune.ai/blog/self-supervised-learning
16 Piccialli F, Somma VD, Giampaolo F, *et al*. A survey on deep learning in medicine: Why, how and when? *Inf Fus* 2021;66:111–37.
17 Jurafsky D, Martin JH. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with language models*. 3rd edn.2024. Available: https://web.stanford.edu/~jurafsky/slp3/
18 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *arXiv* [Preprint] 2017.
19 Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep* 2024;14:6086.
20 Bedi S, Liu Y, Orr-Ewing L, *et al*. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA* 2024;0:e2421700.
21 Medium. LLM evaluation metrics explained. ROUGE score, BLEU, perplexity, MRR… | by Mehul Gupta | data science in your pocket. 2024. Available: https://medium.com/data-science-in-your-pocket/llm-evaluation-metrics-explained-af14f26536d2
22 Bai L, Zhang Y, Wang P, *et al*. Improved diagnosis of rheumatoid arthritis using an artificial neural network. *Sci Rep* 2022;12:9810.
23 Messelink MA, Roodenrijs NMT, van Es B, *et al*. Identification and prediction of difficult-to-treat rheumatoid arthritis patients in structured and unstructured routine care data: results from a hackathon. *Arthritis Res Ther* 2021;23:184.
24 De Craemer A-S, Renson T, Deroo L, *et al*. Peripheral manifestations are major determinants of disease phenotype and outcome in new onset spondyloarthritis. *Rheumatology (Sunnyvale)* 2022;61:3279–88.
25 Jorge AM, Smith D, Wu Z, *et al*. Exploration of machine learning methods to predict systemic lupus erythematosus hospitalizations. *Lupus (Los Angel)* 2022;31:1296–305.
26 Hao X, Zheng D, Khan M, *et al*. Machine Learning Models for Predicting Adverse Pregnancy Outcomes in Pregnant Women with Systemic Lupus Erythematosus. *Diagnostics (Basel)* 2023;13:612.
27 Zhang Y, Luo M, Wu P, *et al*. Application of Computational Biology and Artificial Intelligence in Drug Design. *IJMS* 2022;23:13568.
28 Paul D, Sanap G, Shenoy S, *et al*. Artificial intelligence in drug discovery and development. *Drug Discov Today* 2021;26:80–93.
29 Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nat New Biol* 2021;596:583–9.
30 Duong SQ, Crowson CS, Athreya A, *et al*. Clinical predictors of response to methotrexate in patients with rheumatoid arthritis: a machine learning approach using clinical trial data. *Arthritis Res Ther* 2022;24.
31 Myasoedova E, Athreya AP, Crowson CS, *et al*. Toward Individualized Prediction of Response to Methotrexate in Early Rheumatoid Arthritis: A Pharmacogenomics-Driven Machine Learning Approach. *Arthritis Care Res (Hoboken)* 2022;74:879–88.
32 Bouget V, Duquesne J, Hassler S, *et al*. Machine learning predicts response to TNF inhibitors in rheumatoid arthritis: results on the ESPOIR and ABIRISK cohorts. *RMD Open* 2022;8.
33 Madrid-García A, Font-Urgelles J, Vega-Barbas M, *et al*. Outpatient Readmission in Rheumatology: A Machine Learning Predictive Model of Patient's Return to the Clinic. *J Clin Med* 2019;8:1156.
34 Beretta L, Santaniello A, Cappiello F, *et al*. Development of a five-year mortality model in systemic sclerosis patients by different analytical approaches. *Clin Exp Rheumatol* 2010;28:S18–27.
35 Üreten K, Maraş HH. Automated Classification of Rheumatoid Arthritis, Osteoarthritis, and Normal Hand Radiographs with Deep Learning Methods. *J Digit Imaging* 2022;35:193–9.
36 Guan B, Liu F, Mizaian AH, *et al*. Deep learning approach to predict pain progression in knee osteoarthritis. *Skeletal Radiol* 2022;51:363–73.
37 Bressem KK, Vahldiek JL, Adams L, *et al*. Deep learning for detection of radiographic sacroiliitis: achieving expert-level performance. *Arthritis Res Ther* 2021;23:106.
38 Folle L, Simon D, Tascilar K, *et al*. Deep Learning-Based Classification of Inflammatory Arthritis by Identification of Joint Shape Patterns-How Neural Networks Can Tell Us Where to "Deep Dive" Clinically. *Front Med (Lausanne)* 2022;9:850552.
39 Van Den Berghe T, Babin D, Chen M, *et al*. Neural network algorithm for detection of erosions and ankylosis on CT of the sacroiliac joints: multicentre development and validation of diagnostic accuracy. *Eur Radiol* 2023;33:8310–23.
40 Kise Y, Ikeda H, Fujii T, *et al*. Preliminary study on the application of deep learning system to diagnosis of Sjögren's syndrome on CT images. *Dentomaxillofac Radiol* 2019;48:20190019.
41 Folle L, Bayat S, Kleyer A, *et al*. Advanced neural networks for classification of MRI in psoriatic arthritis, seronegative, and seropositive rheumatoid arthritis. *Rheumatology (Sunnyvale)* 2022;61:4945–51.

42 Goñi M, Basu N, Murray AD, *et al*. Brain predictors of fatigue in rheumatoid arthritis: A machine learning study. *PLoS One* 2022;17:e0269952.

43 Bressem KK, Adams LC, Proft F, *et al*. Deep Learning Detects Changes Indicative of Axial Spondyloarthritis at MRI of Sacroiliac Joints. *Radiology* 2022;305:655–65.

44 Ashinsky BG, Bouhrara M, Coletta CE, *et al*. Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative. *J Orthop Res* 2017;35:2243–50.

45 Matsuo H, Kamada M, Imamura A, *et al*. Machine learning-based prediction of relapse in rheumatoid arthritis patients using data on ultrasound examination and blood test. *Sci Rep* 2022;12:7224.

46 Roncato C, Gautier G, Ploton G, *et al*. Colour Doppler ultrasound of temporal arteries for the diagnosis of giant cell arteritis: a multicentre deep learning study. *Clin Exp Rheumatol* 2020;38:120–5.

47 Vukicevic AM, Radovic M, Zabotti A, *et al*. Deep learning segmentation of Primary Sjögren's syndrome affected salivary glands from ultrasonography images. *Comput Biol Med* 2021;129:104154.

48 Morales-Ivorra I, Narváez J, Gómez-Vaquero C, *et al*. Assessment of inflammation in patients with rheumatoid arthritis using thermography and machine learning: a fast and automated technique. *RMD Open* 2022;8:e002458.

49 Morales-Ivorra I, Narváez J, Gómez-Vaquero C, *et al*. A Thermographic Disease Activity Index for remote assessment of rheumatoid arthritis. *RMD Open* 2022;8:e002615.

50 Maarseveen TD, Meinderink T, Reinders MJT, *et al*. Machine Learning Electronic Health Record Identification of Patients with Rheumatoid Arthritis: Algorithm Pipeline Development and Validation Study. *JMIR Med Inform* 2020;8:e23930.

51 Zhao SS, Hong C, Cai T, *et al*. Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records. *Rheumatology (Sunnyvale)* 2020;59:1059–65.

52 Love TJ, Cai T, Karlson EW. Validation of psoriatic arthritis diagnoses in electronic medical records using natural language processing. *Semin Arthritis Rheum* 2011;40:413–20.

53 van Leeuwen JR, Penne EL, Rabelink T, *et al*. Using an artificial intelligence tool incorporating natural language processing to identify patients with a diagnosis of ANCA-associated vasculitis in electronic health records. *Comput Biol Med* 2024;168:107757.

54 Román Ivorra JA, Trallero-Araguas E, Lopez Lasanta M, *et al*. Prevalence and clinical characteristics of patients with rheumatoid arthritis with interstitial lung disease using unstructured healthcare data and machine learning. *RMD Open* 2024;10:e003353.

55 England BR, Roul P, Yang Y, *et al*. Extracting forced vital capacity from the electronic health record through natural language processing in rheumatoid arthritis-associated interstitial lung disease. *Pharmacoepidemiol Drug Saf* 2024;33:e5744.

56 Lin C, Karlson EW, Dligach D, *et al*. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *J Am Med Inform Assoc* 2015;22:e151–61.

57 Benavent D, Muñoz-Fernández S, De la Morena I, *et al*. Using natural language processing to explore characteristics and management of patients with axial spondyloarthritis and psoriatic arthritis treated under real-world conditions in Spain: SpAINET study. *Ther Adv Musculoskelet Dis* 2023;15:1759720X231220818.

58 Forrest IS, Petrazzini BO, Duffy Á, *et al*. A machine learning model identifies patients in need of autoimmune disease testing using electronic health records. *Nat Commun* 2023;14.

59 Epic cosmos. 2024. Available: https://cosmos.epic.com/

60 Patel J, Yao L, Vina E, *et al*. Phenotype Systemic Lupus Erythematosus Patients from EPIC Cosmos. *Stud Health Technol Inform* 2024;310:159–63.

61 Databricks. A compact guide to large language models. 2023.

62 Kowalewski KF, Rodler S. Large language models in science. *Urologie* 2024;63:860–6.

63 Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. *Nat Med* 2023;29:1930–40.

64 Wójcik S, Rulkiewicz A, Pruszczyk P, *et al*. Beyond ChatGPT: What does GPT-4 add to healthcare? The dawn of a new era. *Cardiol J* 2023;30:1018–25.

65 Rao A, Pang M, Kim J, *et al*. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *J Med Internet Res* 2023;25:e48659.

66 Horiuchi D, Tatekawa H, Oura T, *et al*. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *Eur Radiol* 2025;35:506–16.

67 Krusche M, Callhoff J, Knitza J, *et al*. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. *Rheumatol Int* 2024;44:303–6.

68 do Olmo J, Logroño J, Mascías C, *et al*. Assessing dxgpt: diagnosing rare diseases with various large language models (pre print). *MedRxiv* 2024.

69 Coskun BN, Yagiz B, Ocakoglu G, *et al*. Assessing the accuracy and completeness of artificial intelligence language models in providing information on methotrexate use. *Rheumatol Int* 2024;44:509–15.

70 Haase I, Xiong T, Rissmann A, *et al*. ChatSLE: consulting ChatGPT-4 for 100 frequently asked lupus questions. *Lancet Rheumatol* 2024;6:e196–9.

71 Uz C, Umay E. 'Dr ChatGPT': Is it a reliable and useful source for common rheumatic diseases? *Int J Rheum Dis* 2023;26:1343–9.

72 Hannah L, Lea-Kristin N, Martin K, *et al*. Vignette-based comparative analysis of chatgpt and specialist treatment decisions for rheumatic patients: results of the rheum2guide study. *Rheumatol Int* 2024;44:2043–53.

73 Venerito V, Puttaswamy D, Iannone F, *et al*. Large language models and rheumatology: a comparative evaluation. *Lancet Rheumatol* 2023;5:e574–8.

74 Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, *et al*. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. *Sci Rep* 2023;13:22129.

75 Goh E, Gallo R, Hom J, *et al*. Large Language Model Influence on Diagnostic Reasoning. *JAMA Netw Open* 2024;7:e2440969.

76 Bordukova M, Makarov N, Rodriguez-Esteban R, *et al*. Generative artificial intelligence empowers digital twins in drug discovery and clinical trials. *Expert Opin Drug Discov* 2024;19:33–42.

77 Thorlund K, Dron L, Park JJH, *et al*. Synthetic and External Controls in Clinical Trials - A Primer for Researchers. *Clin Epidemiol* 2020;12:457–67.

78 Meskó B. Prompt Engineering Is An Emerging Essential Skill For Medical Professionals: A Tutorial. *J Med Internet Res* 2023;25:e50638.

79 Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care* 2023;27:75.

80 Conroy G. Do AI models produce more original ideas than researchers? *Nature* 2024.

81 Research Rabbit. Available: https://www.researchrabbit.ai/

82 Elicit. The ai research assistant. 2024 Available: https://elicit.com/

83 Hoffmann M, Boysel S, Nagle F, *et al*. Generative ai and the nature of work. *SSRN* [Preprint] 2024.

84 Jenni AI. Available: https://jenni.ai/?via=direct&gad_source=1&gclid=EAIaIQobChMI7PHloc2XiAMVcZpoCR1Opht-EAAYASAAEgJwsfD_BwE

85 Copilot. 2024 Available: https://copilot.microsoft.com/

86 Vasey B, Nagendran M, Campbell B, *et al*. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022;377:e070904.

87 Liu X, Rivera SC, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* 2020;26:m3164.

88 Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell* 2020;2:e200029.

89 Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378.

90 Cacciamani GE, Eppler MB, Ganjavi C, *et al*. Development of the chatgpt, generative artificial intelligence and natural large language models for accountable reporting and use (cangaru) guidelines. *arXiv* [Preprint] 2023.

91 Gibney E. What the EU's tough AI law means for research and ChatGPT. *Nature* 2024;626:938–9.

92 van Onna M, Boonen A. Challenges in the management of older patients with inflammatory rheumatic diseases. *Nat Rev Rheumatol* 2022;18:326–34.