Narrative review of machine learning in rheumatic and musculoskeletal diseases for clinicians and researchers: biases, goals, and future directions

The source(s) of support in the form of grants or industrial support: National Institutes of Health/National Institute of Arthritis and Musculoskeletal Diseases (NIH/NIAMS) P30AR072580 and R21AR074685

Conflict of interest: none

Acknowledgment: We would like to thank Dr. Yvonne Golightly for her insightful comments on the draft manuscript.

Amanda E. Nelson, MD MSCR (corresponding)

ORCID: 0000-0002-9344-7877

Associate Professor

University of North Carolina at Chapel Hill

Thurston Arthritis Research Center

Department of Medicine, Division of Rheumatology, Allergy, and Immunology

3300 Thurston Building, Campus Box #7280

Chapel Hill, NC 27599-7280, aenelson@med.unc.edu

Liubov Arbeeva, MS

ORCID: 0000-0001-5004-419X

University of North Carolina at Chapel Hill

Thurston Arthritis Research Center
Abstract

There has been rapid growth in the use of artificial intelligence analytics in medicine in recent years, including in rheumatic and musculoskeletal diseases (RMDs). Such methods represent a challenge to clinicians, patients, and researchers given the “black box” nature of most algorithms and the unfamiliarity of the terms and lack of awareness of potential issues around these analyses. Therefore, this review aims to introduce this area in a way that is relevant and meaningful to clinicians and researchers. We hope to provide some insights into relevant strengths and limitations, reporting guidelines, as well as recent examples of such analyses in key areas with a focus on lessons learned and future directions in diagnosis, phenotyping, prognosis, and precision medicine in RMDs.

Box 1. Definitions and Abbreviations for Key Terms in this Review

Rheumatic and Musculoskeletal Diseases (RMDs): A diverse group of more than 200 diseases that often affect joints but can affect any organ, often caused by problems in the immune system, inflammation, or infection, that can result in significant disability1.

Artificial Intelligence (AI): The development of computer systems that can perform tasks normally requiring human intelligence

Machine Learning (ML): A discipline within AI where computer algorithms are developed to learn and make decisions based on data

Background

Artificial intelligence (AI, Box 1) and its subcategory of machine learning (ML) have rapidly gained traction as analytic methods in a variety of conditions including RMDs. These terms have seemingly taken over the medical literature in recent years, but often in a way that is not readily accessible to most clinicians or researchers. Beam et al. provided a very useful perspective on AI/ML in 2018 as part of a spectrum from fully human-guided analysis and decision making to fully automated network-based algorithms2. They sagely note that AI/ML provides “…no guarantees of fairness, equitability, or even veracity2.”
The European League Against Rheumatism (EULAR) in 2020 endorsed principles relating to the use of big data (defined as large, complex, and/or multidimensional, from heterogeneous sources) in RMDs. These include an imperative to consider ethical issues and an overarching goal to use big data to improve the lives of patients with RMDs. Key points focused on the need for harmonized standards and the FAIR principle (Findable, Accessible, Interoperable, and Reusable), open data platforms with privacy considerations and interdisciplinary collaboration, use of explicit reporting of methods, benchmarking of computational methods, and independent validation, along with interdisciplinary training in big data for clinicians and scientists from various backgrounds.

A variety of recent reviews have focused on the use of ML in RMDs, including an overview of definitions and performance characteristics of ML and a set of representative clinical studies through early 2021, and a more technical overview of definitions, methods, classification procedures, prediction models and algorithms; noting that most datasets are not purpose-built and thus lack necessary sample size (SS) as well as novel features. Additionally, there are several reviews focused specifically on the role of ML in imaging, including in RMDs, so this large topic will not be reviewed here. Therefore, rather than providing a systematic or technical review, we refer the reader to these publications, and instead provide a narrative overview of recent work. The goal of this review is to serve as an introduction to the area of AI/ML for clinicians and researchers in RMDs who are new to this field (see also Hugle et al for a nice introduction to types of AI/ML in rheumatology), to improve understanding around how incorporating these methods might benefit their work, which data types might be useful in AI/ML analyses (Box 2), how they might work with collaborators, and to provide examples of work in key areas, including 1) diagnosis; 2) phenotyping; 3) prognosis; 4) precision medicine; 5) limitations and biases; and 6) future directions.
Box 2. Examples of data sources and types in AI/ML approaches

DATA SOURCES
Individual patients
Observational cohorts
Clinical cohorts
Clinical trials
Electronic health records
National reimbursement databases
Registries

DATA TYPES (standard)
Clinical/demographic
Patient reported outcomes
Laboratory data
Tissue analysis
Imaging
Medications
Clinical notes
Disease activity

DATA TYPES (specialized)
Genetics
Proteomics, other ‘omics
Nutrition
Wearables

1. AI/ML For Diagnosis: identifying the condition of interest in a patient or cohort. AI/ML methods can assist with a variety of diagnostic challenges, including in the clinical setting based on available lab and clinical data, identification of affected patients in the Electronic Health Record (EHR), or optimal selection of clinical study participants.

One of the main interests for clinicians and their patients is the definition of the disease state. As many RMDs are rare diseases (RD), by definition affecting fewer than 1 in 2000 individuals, this can be particularly challenging, but ML holds specific promise to improve strategies and develop new drugs for the treatment of RD. Data from genomic and multiomic approaches have provided new insights, as have other big data like gait assessments and imaging\(^2\). In a recent scoping review, most studies of RD employing ML were focused on diagnosis or prognosis and many suffered from small SS and lack of external validation\(^3\). Registries can increase SS for
even the rarest of conditions, although harmonization is needed for these to be useful; open data sources have similar limitations and may have poorer reliability than carefully curated data sources. Possible solutions include: enhancing SS with incorporation of unlabeled non-case samples outside the RD of interest, artificially built samples, and transfer learning.

Systemic lupus erythematosus (SLE) is an example of a RD and RMD that has been investigated using ML methods given the challenges of making this diagnosis in clinical practice. The SLE Risk Probability Index (SLERPI) has been proposed to assist with diagnosis in the clinical setting improving time to diagnosis and treatment in SLE. Clinical guidance was used to create 20 feature panels each of which were submitted to random forest (RF) and LASSO penalized regression, resulting in 40 models trained on data from two SLE registries. The model with highest accuracy was evaluated in a validation cohort and converted into a scoring system, using a threshold of 7 to separate SLE versus other RMD in adults, adjusted to 8 in a follow-up study in a pediatric SLE cohort. Although using a relatively small SS and retrospective design, this work exemplifies the importance of internal and external validation in ML-based diagnostic algorithms. Another study utilized both structured and narrative data to identify SLE patients in EHR data. They selected definite cases, probable cases, and definite non-cases by chart review to determine the positive predictive value (PPV) of the algorithms and features internally and in external cohorts. The ML algorithm had 92% PPV for define/probable SLE in the internal cohort, and 94% in an external cohort, comparing favorably to PPV of using one or two ICD codes, which was cited as around 50%. The EHR phenotyping protocol is published and available for use in clinical and translational studies. The performance characteristics of previously published algorithms were also tested, demonstrating the importance of adjustment for portability (i.e., their application in other systems), and identified several challenges, such as different medical billing practices, medication prescribing and reporting, and disease prevalence.
2. AI/ML for Phenotyping: defining important subtypes of disease. Individuals with RMDs have variable courses, including rates of progression, transition to other conditions, and response to treatments.

Molecular phenotyping is an area of growing interest given substantial overlap in clinical features and the lack of specific diagnostic studies for many RMDs. Myositis is a good example, given the multiple antibodies and clinical phenotypes within inflammatory myositis and our growing understanding of their impact on outcomes. Muscle biopsies were collected from 119 patients enrolled in several key myositis cohorts, including those with myositis-specific autoantibodies, anti-synthetase syndrome, necrotizing myopathy, or inclusion body myositis, and 20 normal controls\textsuperscript{18}. Ten different ML algorithms, including decision trees and RF, among others, were trained using transcriptomic data to determine disease-specific gene expression patterns which allowed accurate identification of subgroups in over 90% of muscle biopsies using the linear support vector machine model\textsuperscript{18}. Although a small SS, the use of biopsy data is a strength of this work, which also demonstrated the usefulness of objective transcriptomics in the interpretation of tissue biopsies. These markers may be useful to tailor therapies to a specific molecular diagnosis in the future.

Juvenile onset SLE is a rare RMD that is challenging to study and often reliant on small cohorts, particularly for subgrouping. The choice of ML algorithms that can potentially address difficulties associated with RD, as well as cross-validation (of value in settings without a readily available validation cohort) is particularly important. Robinson et al. applied supervised ML approaches for classification (i.e., discrimination of 67 SLE patients from 39 healthy controls) and selection of important variables, including immune cell profiles, which were further used in in an unsupervised k-means clustering that identified four potentially important subgroups among SLE patients\textsuperscript{19}. Limitations include the small SS, low number of Black patients, and imperfect
outcome measures. However, such immune-based phenotyping may improve patient stratification for future clinical studies and may eventually inform clinical practice.

3. AI/ML for Prognosis: defining disease course for targeted intervention. It is important to know which individuals are most likely to worsen rapidly and which may experience improvement or resolution of their condition to best identify those at risk and take appropriate action.

While clinical disease activity measures are available and validated in most RMDs, their application and use in clinical practice is inconsistent. ML can be used to estimate these values from available clinical information to identify patients with active disease in clinical datasets. An earlier study in this area utilized only structured data (e.g., lab values, existing Clinical Disease Activity Index [CDAI] scores, and medications) from the EHR of two distinct clinical settings to build a deep learning model to predict CDAI in patients with RA\(^20\). More than 20 variables were significantly important for accuracy of the predictions. This paper demonstrates both strong statistical design and a detailed discussion of limitations associated with such data, including missing values, subsequent biases, and of usefulness of these models in clinical practice.

Another approach is to use unstructured data, such as clinical notes, to predict other quantifiable outcomes. Alves et al. developed natural language processing algorithms, followed by validation to estimate SLEDAI categories in SLE from unstructured notes\(^21\). This approach was validated to estimate CDAI scores in RA\(^22\). Both groups were able to estimate disease activity with Area Under the Curve (AUC)~0.9 and correlation with true clinical scores around 0.7. The ability to estimate disease activity in the absence of clinician-entered scores would dramatically increase the data available for research use, providing large numbers of patients for outcomes research, or for meta-analyses, and potentially reducing disparities by provider or clinic. Such algorithms would still require sufficient input for estimation and may not be applicable to all settings (e.g., handwritten notes, international/low resource settings, etc.).
combination of structured and unstructured data is likely optimal for prediction of prognosis in clinical datasets.

It is important to identify patients who are most likely to progress or even to develop disease. One such circumstance is patients with undifferentiated arthritis, some of whom go on to develop rheumatoid arthritis (RA) and some of whom do not. A small study assessed the DNA methylome of patients with undifferentiated arthritis (n=72), where about half remained stable and half developed RA after a year, and identified differential methylation between groups. Both supervised and unsupervised methods were used along with internal and external validation. Distinct methylation patterns were seen among those who did and did not develop RA as well as a separate group with RA at baseline, demonstrating the potential of methylation markers to sense “early disease determinants” in these patients. Despite the small SS, this work highlights the possibility of incorporating basic and clinical data for clinically relevant risk assessment.

Data from randomized clinical trials (RCTs) can be used to identify predictors of prognosis, as such studies include well-phenotyped individuals, balanced across treatment groups at baseline, who are followed over time. Pooling individual data across different trials, while appropriately addressing heterogeneity, can increase SS, but this data source provides a limited number of variables. One such study pooled data from nearly 1900 patients with psoriatic arthritis enrolled in four RCTs to determine subgroups of response trajectory to secukinumab therapy over 52 weeks. They applied model-based clustering methods to identify seven clusters of participants, where patients within a cluster had a common distribution of 206 baseline measures; this procedure was repeated on 200 different subsets to access cluster stability. The clusters, characterized according to longitudinal responses, were clinically interpretable, with features such as higher polyarticular disease burden, greater foot symptoms and dactylitis, or more nail and skin involvement. The overall population was skewed due to
the RCT design, with a high proportion of active polyarticular disease compared to other subtypes (such as oligoarticular involvement). However, this type of work could be used to inform trial selection for specific therapeutics or dosing regimens in future RCTs. In another study, data from several RCTs were examined and a remission prediction score for RA patients treated with tocilizumab was developed and validated\textsuperscript{26}. Importantly, this prediction rule was subsequently tested in registry real world data with an extended set of variables in a follow-up paper\textsuperscript{27}, finding that the RCT model could similarly predict discrimination in the registry data (with AUC~0.7 to 0.8). Both studies are excellent examples of robust design and rigorous statistical analysis.

There has been a great deal of interest in predicting progression in osteoarthritis (OA), one of the most common RMDs. An objective endpoint of total joint replacement (TJR) can be used to overcome some of the challenges of subjective pain outcomes and discordance with imaging in OA, although this is complicated by issues of preference, practice variability, and access to care. Jamshidi et al. utilized baseline data from the publicly available Osteoarthritis Initiative (OAI) dataset to predict TJR at 96 months\textsuperscript{28}. Utilizing a LASSO method to select features followed by multiple ML models, they could predict time to TJR with high accuracy (AUC 0.9). Given the nature of the OAI cohort, which included only people with or at risk for OA, these results are not generalizable to the general population, and only features known to be associated with OA were included in the dataset, so there was no opportunity to discover novel features, a challenge for many existing cohorts.

The IMI-APPROACH (Applied Public-Private Research enabling OsteoArthritis Clinical Headway) study utilized a novel selection method to identify individuals most likely to progress\textsuperscript{29}. This group developed algorithms in two large OA cohort studies to best classify progressors (and avoid selection of likely non-progressors), potentially improving efficiency for future clinical studies. This procedure was subsequently utilized to pre-select likely progressors
(high likelihood of joint space loss or pain) from existing OA cohorts\textsuperscript{30}, resulting in 297 participants to be followed for two years. Their inclusion was decided using RF and other supervised ML models, that provide the probability of progression based on structure and/or pain within the lifetime of the study\textsuperscript{31}. To improve the performance of RF, a single model was trained to assign “pain progression” and “structure progression” labels independently (multi-label classification), while duo classifier was used two independent models, each trained to predict a single label (pain or structure)\textsuperscript{31}. As a purpose-built cohort designed for the application of ML methods, this work is an important step forward in OA.

4. AI/ML for Precision medicine: using data to guide therapy and avoid adverse events.

Several recent publications in RMDs reflect the goals of precision medicine, which can be understood as the provision of the right treatment\textsuperscript{32}, at the right dose\textsuperscript{33}, to the right person, at the right time\textsuperscript{34}, while minimizing unnecessary testing, side effects and overuse issues, including opioid use and abuse\textsuperscript{35-37}, specifically opioid use around TJR\textsuperscript{38-40}, and to explore issues of inequity in classification\textsuperscript{41}.

Prediction of clinical response among RMD patients, and thus ability to make an informed decision about optimal treatment recommendations, has long been a goal of clinicians and researchers. Utilizing 275 baseline variables from a dataset described above\textsuperscript{26}, a separate analysis employed Bayesian elastic net, which is useful for a large number of potentially correlated patient characteristics, to determine predictors of 16-week outcome based on starting dose of secukinumab in psoriatic arthritis\textsuperscript{33}. While still limited by RCT data and need for validation, this work provides insight relevant to precision medicine in RMDs. Another study, in a small cohort of 39 women with RA starting anti-TNF therapy, researchers assessed differences in multiomics from peripheral blood mononuclear cells (PBMCs) among EULAR responders and non-responders at 3 months\textsuperscript{42}, although ML methods were not fully integrated into this analysis.
A preliminary study aiming to predict the 6-month clinical response to adalimumab and etanercept was undertaken in 80 RA patients enrolled in an observational cohort in the Netherlands as they started biologic treatment\textsuperscript{32}. The investigators obtained PBMCs prior to biologic therapy and performed genome-wide expression and DNA methylation assays, which demonstrated different signatures in those who eventually responded to therapy. RF models utilizing these multiomics data had greater than 80% accuracy for prediction of response\textsuperscript{32}. Several internal cross-validation techniques were used, although the validation and training sets were from the same sample\textsuperscript{43}. Key for this study is the incorporation of true multiomics data and integrated data analysis in the prediction models. Future work will benefit from larger samples with robust outcomes and truly independent external validation sets to avoid overfitting and to mitigate feature instability (often challenging in RDs). Another example is work utilizing consortium data to develop an algorithm to predict methotrexate response in patients with early RA (n=643)\textsuperscript{34}. A RF model was trained on all UK patients (n=336) and externally validated on independent, non-UK patients (n=307, Sweden, Netherlands). Overfitting and class imbalance were directly addressed; the sample included only White Europeans, so generalizability remains limited. The incorporation of genetic data in the prediction algorithm substantially improved prediction accuracy, supporting the feasibility of pharmacogenomic markers for precision medicine, although the overall response rate remained low\textsuperscript{34}.

We utilized 24 ML algorithms to select the optimal model and to develop individualized treatment rules based on RCT data from the Intensive Diet and Exercise for Arthritis (IDEA) trial\textsuperscript{44}. IDEA randomized overweight or obese individuals with symptomatic knee OA to 3 groups: exercise alone, diet alone, or a combination of diet plus exercise\textsuperscript{45}. Using data from 343 participants and multiple outcome RF and list-based models, subgroups of participants were identified who would have improved outcomes for weight loss and for IL-6 (an inflammatory cytokine) if they had been assigned according to the decision rule rather than to...
the diet plus exercise intervention using value functions\textsuperscript{44}. This work highlights the use of RCT data from a non-pharmacologic trial, exploration of multiple features and outcomes, and multiple model evaluation which could improve the design of future studies.

5. Limitations and biases in AI/ML. Here we discuss several key issues including: a) Bioethics; b) Missing data; c) Model bias; and d) Translation.

a. Bioethics. A recent excellent piece on bioethics in big data and RMD research identified four main areas of potential concern: privacy, informed consent, impact on the medical profession, and justice\textsuperscript{46}. First, privacy and confidentiality are a challenge when large datasets are linked, as the detailed information that results could increase the risk of re-identification even when the datasets themselves are “de-identified” or even fully public. These may not even be considered “human subjects” data but can still be used to extract sensitive information. The authors astutely recommend the use of an “honest broker” to maintain and distribute data, avoiding providing full access to any potentially interested entity (e.g., private funders, industry). Secondly, the nature of these big data analytics means that future developments, potential uses, and consequences are not known at the time of data collection, making fully informed consent a challenge to participants\textsuperscript{47} and investigators, as well as institutional review boards and ethics committees. Third is the potential impact on the medical profession: if an algorithm makes a mistake that causes harm, who is responsible? ML analytics carry the potential to undermine the physician-patient relationship with negative consequences. This leads to the fourth area, justice, reflected in the potential for these technologies to worsen the existing digital divide and local and global health disparities. The risk of security breaches and hacking are higher in areas with lower health literacy, greater corruption, or rapid technology expansion without appropriate oversight, further placing underserved populations at risk\textsuperscript{46}.

b. Missing data. Considerations around health equity in relation to ML and big data have recently gained more attention, including in the study of RMDs. It is essential to consider who is
in the dataset, who is not, and why not, as well as the impact these “missing” data may have on results from a ML analysis. For example, missing data could represent inconsistent care, an issue that more often affects individuals of low socioeconomic status, those with mental health issues, or immigrant populations. The existence of multiple care instances in a single EHR is often required for diagnostic algorithms and may thus exclude these individuals. Such missing data are not random, leading to potentially erroneous inferences from models that assume random missingness\textsuperscript{48}. Individuals of lower socioeconomic status may already receive suboptimal care; failure to recognize this could result in an algorithm that preferentially directs these patients to inadequate care\textsuperscript{48}. A lack of health care is not equivalent to lower disease burden but could be interpreted as such by an ML algorithm lacking appropriate context. Use of proxies for health, such as mortality, readmission, or cost can introduce biases due to unequal access to care, resulting in underestimated illness burden and potentially further inequities in access\textsuperscript{49}. Over-the-counter medications are often missing or incompletely reported in EHRs and national reimbursement databases, and more accurate prescription dispensation data may require linkage to pharmacy or other databases to get a more complete picture of what patients are taking\textsuperscript{50}. EHRs often lack data on social determinants that might improve the ability of the ML algorithm to identify such equity issues. Similarly, race/ethnicity and preferred language may be missing or incorrect, leading to misclassification\textsuperscript{48}. Specific analyses focused on addressing these issues, including subgroup analysis, stratification, and validation in a representative cohort, can be considered\textsuperscript{60}. Importantly, while such attention to fairness can avoid potential harms, when fully appreciated it can also identify areas of greatest need and lead to improved equity\textsuperscript{51}.

c. Model Bias. Given the subjectivity of the model selection process, there is an obvious need for both clinical/provider and patient input in making these decisions\textsuperscript{52}. The inclusion of patient collaborators in RMD research, including when using big data/ML applications, is important and
may help address some of these issues\textsuperscript{52}. This is of great importance as, in this author’s experience, most papers using big data or ML methods state, without evidence, that it is somehow not possible or not reasonable to involve patient collaborators due to the nature of the work.

There are of course many other potential sources of bias in ML models\textsuperscript{53}. A systematic review of prediction models utilizing supervised ML methods found that the vast majority of the \textasciitilde150 studies reviewed were at high risk of bias for a few key reasons, including an inadequate number of events per predictor, and overfitting, issues which did not improve in the literature over time\textsuperscript{54}. Another study, focused on biases in observational clinical studies in secondary databases, identified confounding, selection bias, and measurement bias as the most reported and provided a detailed summary table\textsuperscript{55} as well as guidance regarding potential ways to address these issues. That ML algorithms can pick up on non-informative features and incorrectly interpret them is well-described, such as prioritizing studies marked “urgent” or “stat” or recognizing features indicative of portable versus departmental imaging\textsuperscript{50}. Investigators may be concerned regarding SS, resulting in lack of consideration of potentially important subgroups in the data that are smaller in number, thus affecting prediction for underrepresented or minority groups\textsuperscript{50}.

Temporal data drift is an uncommonly discussed limitation to generalizability of ML algorithms but can have substantial implications\textsuperscript{50,56}. A systematic review focused on approaches to mitigate the effects of temporal shift found only 15 papers explicitly covering this topic in clinical areas\textsuperscript{57}, although this phenomenon is better studied and appreciated in non-clinical work\textsuperscript{58}. Temporal shifts can occur at the patient (demographics, referrals, new diseases); practice (trial or guideline results, practice patterns, drug/test availability, reimbursement policies), or administration (EHR modification, vendor, coding system and practices) level and can impact performance and reproducibility. Strategies to address this issue, as well as those to be
developed in the future, would benefit from a rigorous benchmarking procedure to best characterize impact and solutions\textsuperscript{57}.

“All models are wrong, but some are useful”. This aphorism is often used to emphasize the importance of acknowledgment of limitations, assumptions, and potential biases relevant to the analyses being used, whether ML or more traditional statistical methods. For the researcher new to the area, awareness of potential bias and limitations is important. Use of reporting guidelines such as TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis)\textsuperscript{59,60}, or critical appraisal tools like PROBAST (Prediction model Risk Of Bias Assessment Tool)\textsuperscript{61} may be useful to avoid, address, and identify such potential biases, and are required by many peer-review journals. In recognition of the specific issues around prediction models utilizing AI and ML, extensions of these tools, TRIPOD-ML, TRIPOD-AI (reporting guideline) and PROBAST-AI (critical appraisal tool) are currently under development\textsuperscript{60}. Other tools developed for fields such as cardiology\textsuperscript{62}, orthopaedics\textsuperscript{63} or for clinical trials\textsuperscript{64,65} are also available.

\textbf{d. Translation}. There are a variety of challenges with translation of AI/ML to clinical practice, many of which are directly related to the challenges above\textsuperscript{50}. Selection of reliable outcomes is essential but can be challenging. In addition, the frequent (and understandable) use of retrospective studies to develop algorithms will result in better performance metrics compared to application to prospective, real-world data, making their implementation difficult and potentially unreliable\textsuperscript{50}. It is essential that studies with the goal of eventual clinical adoption be rigorously performed, appropriately reported and peer-reviewed; many such studies are published only as pre-prints\textsuperscript{50}. The development of understandable and clinically relevant assessments of model performance, reflecting its practical importance is a key in the clinical realm. As noted above and throughout this review, it is difficult to compare algorithms due to methodologies, populations, sample distributions and characteristics, and differing performance metrics, again
highlighting the need for independent test sets and large open datasets for validation and benchmarking.

6. Future directions for AI/ML in RMDs

A variety of other clinical uses for AI/ML are also in various stages of development, but space precludes extensive discussion, including digital health, smart technology, wearables, care algorithms, and monitoring of adherence. Wearables are of particular interest given the potential for continuous monitoring. This type of data could allow for automated alerts to patients or their physicians, direct patient feedback, and/or algorithm-based automatic intervention, while providing an opportunity to increase access and potentially improve monitoring and outcomes. An obvious limitation, in addition to cost and use of the technology itself, is the need for enhanced health and digital literacy of both patients and their care providers to allow optimal use of such tools. So-called “explainable ML” has been another hot topic of late, implying that ML algorithms, not always straightforwardly interpretable to humans as regression coefficients or heatmaps, which undermines the creditability of these models. Therefore, explanation techniques are needed to make these so-called black box approaches explainable and trusted, particularly in the health care setting. Unfortunately, the currently available methods (e.g., regression with understandable coefficients or heat maps for imaging applications) do not imply accurate performance and may give false assurances, and thus are better understood as tools for developers. Any tool to be utilized in clinical care must undergo “robust assessments of the efficacy, affordability, and scalability of AI in the context of digital health for rare connective tissue diseases…to avoid the detrimental waste of scarce resources.”

AI/ML techniques can inform all stages of drug development and repurposing, including identification of potential targets, validation of those targets, identification of biomarkers and optimization of clinical trial endpoints. These methods can harness a variety of datatypes,
incorporating information from images, text, wearables, assays, and complex ‘omics data, which can be used in concert to objectively inform some of the previously trial-and-error steps in this complex process\textsuperscript{71}.

Additional AI/ML applications have been developed in other fields that will likely appear soon in RMD research. For example, epigenetic biomarkers of aging have been studied in cardiovascular disease, Alzheimer’s disease, and various cancers\textsuperscript{72}, but not yet in RMDs. Epigenetic clocks, reflecting one’s biological age, were developed to study age-related diseases and excess mortality. Clocks based on age-related inflammation have been created using ML but have not yet been studied in RMDs, which is obviously a promising direction\textsuperscript{73}. Other such clocks have been developed using a variety of ‘omics data, although frequently in isolation\textsuperscript{74-76}.

In contrast, the simultaneous incorporation of multi-modal data, e.g., genetic, ‘omics, images, psychosocial, and/or clinical data, which holds substantial promise, is challenging due to the need to integrate multiple data types, potentially from different studies and cohorts. To date, most studies with such data are relatively small and primarily focused on the multi-omics aspect rather than integration across all data types\textsuperscript{42}. It will be important in the future to collect these types of multi-omics data on larger and more representative samples and fully integrate the different data types into ML models. A few studies have incorporated such multi-modal data\textsuperscript{34}, and others are collecting it\textsuperscript{30,77}, but additional rigorously designed longitudinal studies will be needed to establish this knowledge base and allow discovery and validation using existing and newly developed methodologies capable of handling this type of multi-modal information.

**Summary**

The promise of ML for advances in RMD research and clinical care is enormous, although not yet fully realized. As exemplified by papers discussed in this review (summarized in Table 1), development and implementation of ML algorithms requires collaborative efforts from a variety
of experts including those with analytic, programming, and subject area expertise working together to achieve robust results. Examples discussed in this review include a range of RMDs, data types, data sources, approaches and outcomes reflecting the breadth of AI’s potential while also considering its limitations. We mention examples of future directions, although these are nearly limitless as technologies evolve. RMD research stands to benefit greatly from such technologies given the challenge of studying these rare diseases with traditional methodologies, but care must be taken to mitigate rather than amplify potential disparities and other potential biases.

Table 1 here
Bibliography

1. van der Heijde D, Daikh DI, Betteridge N, et al. Common language description of the term rheumatic and musculoskeletal diseases (RMDs) for use in communication with the lay public, healthcare providers and other stakeholders endorsed by the European League Against Rheumatism (EULAR) and the American College of Rheumatology (ACR). Ann Rheum Dis 2018;77:829-32.


<table>
<thead>
<tr>
<th>Aim/Research question (Ref)</th>
<th>Data Types</th>
<th>Methods Used</th>
<th>What does this study add?</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To develop an algorithm that can aid SLE diagnosis (14)</td>
<td>Clinical/demographic</td>
<td>RF, LASSO-Logistic Regression</td>
<td>Developed and validated an accurate, clinician-friendly algorithm based on classical disease features for early SLE diagnosis and treatment to improve patient outcomes</td>
<td>Small SS; retrospective design; further validation in prospective studies needed</td>
</tr>
<tr>
<td>To generate ML based algorithms to accurately identify SLE patients from EHR records (16)</td>
<td>EHRs (structured and narrative data)</td>
<td>Natural Language Processing; penalized logistic regression</td>
<td>Generated algorithms were internally and externally validated and compared to existing rule-based algorithms. The EHR phenotyping protocol is published and available for use in clinical and translational studies (17)</td>
<td>The authors emphasize the importance of adjustment for portability</td>
</tr>
<tr>
<td><strong>Phenotyping</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To define unique gene expression profiles in muscle biopsies from patients with different types of myositis (18)</td>
<td>Transcriptomics data</td>
<td>Decision Trees, RF, Linear SVM model, and other classification algorithms</td>
<td>Demonstrated the usefulness of objective transcriptomics to tailor therapies to a specific molecular diagnosis in the future</td>
<td>Small SS</td>
</tr>
<tr>
<td>Task</td>
<td>Data Sources</td>
<td>Methodology</td>
<td>Findings</td>
<td>Limitations</td>
</tr>
<tr>
<td>---------------------------------------------------------------------</td>
<td>-----------------------------------</td>
<td>------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>To characterize the immune cell profile of patients with juvenile-onset SLE and investigate links with the disease trajectory over time (19)</td>
<td>Laboratory data</td>
<td>Supervised ML approaches for classification (balanced RF, sPLS-DA) and feature selection; used in in an unsupervised k-means clustering</td>
<td>Identified four potentially important subgroups among SLE patients; may eventually inform clinical practice and clinical studies</td>
<td>Small SS; low number of Black patients, and imperfect outcome measures</td>
</tr>
<tr>
<td>Prognosis</td>
<td></td>
<td>------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>To assess the ability of an artificial intelligence system to predict disease activity of patients with RA at their next clinical visit (20)</td>
<td>Medications; Disease activity; Clinical/demographic; Laboratory data</td>
<td>Recurrent Neural Networks (longitudinal deep learning model)</td>
<td>The findings suggest that building accurate models to forecast complex disease outcomes using EHRs is possible</td>
<td>Missing data; usefulness of these models in clinical practice is not clear</td>
</tr>
<tr>
<td>To validate a ML model to estimate SLEDAI score categories from clinical notes and to apply the model to a large, real-world dataset to generate estimated score categories (21)</td>
<td>Clinical notes</td>
<td>Natural Language Processing</td>
<td>Developed natural language processing algorithms, followed by validation to estimate SLEDAI categories from unstructured notes</td>
<td>Notes without sufficient text were excluded; the model will need to be modified for different data sources</td>
</tr>
<tr>
<td>To validate a ML model developed in (21) to estimate CDAI scores in RA using clinical notes (22)</td>
<td>Clinical notes</td>
<td>Natural Language Processing</td>
<td>External validation of previously developed algorithm</td>
<td>The authors emphasize the importance of validation in non-U.S. practices</td>
</tr>
<tr>
<td>To investigate whether alterations in the DNA methylation profiles of immune cells can discriminate between disease subtypes (23)</td>
<td>Clinical/demographic; DNA methylome</td>
<td>Logistic Regression; RF; SVM</td>
<td>Demonstrated the potential of methylation markers as early disease determinants</td>
<td>Small SS</td>
</tr>
<tr>
<td>Identify distinct clusters of psoriatic arthritis patients based on their baseline features and explore clinical and therapeutic implications (25)</td>
<td>Clinical/demographic</td>
<td>Model-based Clustering Methods</td>
<td>Data were pooled from 4 RCTs (increased SS); identified seven distinct clusters; some showed greater improvements if using higher dose of secukinumab</td>
<td>RCT population</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>To identify patients with RA most likely to achieve remission with tocilizumab monotherapy by developing and validating a prediction model and associated remission score (26)</td>
<td>Clinical/demographic; Disease activity; Laboratory data</td>
<td>Logistic regression</td>
<td>Data were pooled from 4 RCTs (increased SS); derived a remission prediction score; robust design and rigorous statistical analysis</td>
<td>Inconsistency in variables between trials; validation using real-world data (RWD) is needed; low disease activity was not considered</td>
</tr>
<tr>
<td>External validation of the prediction score using “real world data” (27)</td>
<td>Clinical/demographic; Disease activity; Laboratory data</td>
<td>Logistic regression, RF</td>
<td>Remission prediction scores derived in RCTs discriminated patients in RWD about as well as in RCTs. Discrimination was further improved by retraining models on RWD.</td>
<td>Missing data; patients were from North America only; Other DMARDs) will be considered in the future work</td>
</tr>
<tr>
<td>To build a model to predict risk and time to total knee replacement (TKR) in osteoarthritis (28)</td>
<td>Clinical/demographic; Patient reported outcomes; Imaging; Medications</td>
<td>Cox regression, DeepSurv, RF, linear/kernel SVM, and linear/neural multi-task logistic regression models</td>
<td>Developed a model using the OAI cohort to predict with high accuracy if and when a given knee would require TKR</td>
<td>Not generalizable to the general population, and only features known to be associated with OA were included in the dataset</td>
</tr>
<tr>
<td>To develop and utilize a novel selection method to identify likely progressors (within the lifetime)</td>
<td>Clinical/demographic; Patient reported outcomes; Imaging; Medications</td>
<td>Logistic Regression; Multinomial Logistic Regression; kNN classifier;</td>
<td>The procedure was developed and subsequently utilized to pre-select likely progressors (within the lifetime)</td>
<td>Weak preprocessing strategy; difficult to implement in clinical practice</td>
</tr>
<tr>
<td>To determine predictors of 16-week outcome based on starting dose of secukinumab in psoriatic arthritis (33)</td>
<td>Clinical/demographic</td>
<td>Bayesian elastic net</td>
<td>Determined predictors of 16-week outcome based on starting dose of secukinumab in psoriatic arthritis; provides insight relevant to precision medicine</td>
<td>RCT data, need for validation</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>----------------------</td>
<td>---------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>To detect biomarkers and expression signatures of treatment response to TNF inhibition (42)</td>
<td>Multiomics from peripheral blood mononuclear cells</td>
<td>Linear Models with Regularization; RF, SVM with Radial Basis Function kernel</td>
<td>Suggested new predictive models of anti-TNF treatment in RA patients</td>
<td>ML methods were not fully integrated into this analysis; small SS</td>
</tr>
<tr>
<td>To predict response to anti-TNF prior to treatment in patients with RA, and understand mechanisms of response (32)</td>
<td>multi-omics</td>
<td>RF</td>
<td>Incorporation of true multiomics data and integrated data analysis in the prediction models</td>
<td>Future work will benefit from larger SS, robust outcomes and independent external validation</td>
</tr>
<tr>
<td>To test the ability of ML approaches with clinical and genomic biomarkers to predict methotrexate treatment response in patients with early RA (34)</td>
<td>Clinical/demographic; Genetics</td>
<td>RF</td>
<td>Integration of clinical and genomic data for individualized prediction of response. The incorporation of genetic data in the prediction algorithm substantially improved prediction accuracy.</td>
<td>The sample included only White Europeans; limited generalizability</td>
</tr>
</tbody>
</table>
To apply a precision medicine approach to maximize expected outcomes in a knee OA RCT (44) Clinical/demographic; Patient reported outcomes; Laboratory data; Medications

24 ML algorithms; optimal model was selected to develop individualized treatment rules

Highlights the use of RCT data from a non-pharmacologic trial, exploration of multiple features and outcomes, and multiple model evaluation which could improve the design of future studies

Participants with missing outcomes and covariates with a large proportion of missing data were excluded; data were from a single RCT