Original Paper

Predicting Rheumatoid Arthritis Flares Using Wearable Device Data: Deep Learning Study

Pragya Sharma[1], PhD; Matteo Danieletto[1,2], PhD; Jessica K Whang[3], MS; Kyle Landell[1,2], BS; Drew Helmus[3], MPH, MA; Mayte Suárez-Fariñas Suárez-Fariñas[4], PhD; Percio S Gulko[5], MD; Robert P Hirten[3], MD

[1]Windreich Department of AI and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[2]The Hasso Plattner Institute for Digital Health at the Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[3]The Dr. Henry D. Janowitz Division of Gastroenterology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[4]Department of Population Health Science and Policy, Center for Biostatistics, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[5]Division of Rheumatology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:

Robert P Hirten, MD
The Dr. Henry D. Janowitz Division of Gastroenterology, Department of Medicine
Icahn School of Medicine at Mount Sinai
Annenberg Building RM 5-12, 1468 Madison Avenue
New York, NY 10029
United States
Phone: 1 212-241- 0150
Fax: 1 646-537-8647
Email: robert.hirten@mountsinai.org

Abstract

Background: Rheumatoid arthritis (RA) flares are common and unpredictable. Current monitoring relies on episodic clinical visits and laboratory testing, thereby limiting the ability to anticipate disease worsening. Wearable devices may support continuous remote monitoring, but models must account for missing real-world data, heterogeneous consumer devices, and the distinction between symptomatic disease activity and systemic inflammation.

Objective: The study aimed to develop and evaluate deep learning (DL) time-series models using wearable-derived physiological data to predict symptomatic and inflammatory RA flares up to 28 days in advance.

Methods: In a prospective cohort study, adults with RA used a smartwatch (Apple Watch; Apple Inc), a wearable fitness tracker (Fitbit; Google LLC), or a smart ring (Oura Ring; Oura Health Oy). Daily wearable-derived heart rate, heart rate variability, and step-count were harmonized across devices and linked with smartphone-based patient-reported outcomes (PROs) and available C-reactive protein results. Symptomatic flares were defined using longitudinal RAPID3 (Routine Assessment of Patient Index Data 3) scores, and inflammatory flares were defined using a prespecified C-reactive protein threshold. Classical machine learning and DL models were trained to predict daily flare vs remission status at 0-, 7-, 14-, 21-, and 28-day horizons. Model evaluation used a temporally ordered 70/30 within-participant train-test split. Performance was assessed using receiver operating characteristic area under the curve (ROCAUC), area under the precision-recall curve (AUPRC), and participant-level sensitivity. This prediction-model analysis was not prospectively registered.

Results: Forty-nine participants contributed 5624 days of overlapping wearable and clinical data. Participants used Apple Watch (n=31), Fitbit (n=15), or Oura Ring (n=3). DL models outperformed machine learning models using imputed wearable data, with average ROCAUC improvements of 5.3% for symptomatic flares and 18.4% for inflammatory flares, and AUPRC improvements of 9.3% and 44.1%, respectively. For 7-day-ahead prediction, transformer models incorporating wearable-derived physiology and demographic features achieved ROCAUCs of 0.95 for both symptomatic and inflammatory flares, with AUPRCs of 0.92 and 0.91, respectively. Adding daily PRO improved participant-level sensitivity for inflammatory flare prediction from 0.67 to 0.87. Performance remained stable across prediction horizons for up to 4 weeks.

Conclusions: Wearable-derived physiological signals analyzed with DL may support personalized forecasting of symptomatic and inflammatory RA flare states weeks in advance. By integrating passive digital biomarkers with PROs, this approach offers

a scalable framework for proactive RA monitoring that could help identify disease exacerbations earlier than episodic clinical assessment alone. Further validation in larger, more diverse cohorts and prospective studies is needed to determine whether wearable-guided forecasting improves clinical decision-making, timely intervention, and long-term outcomes.

J Med Internet Res 2026;28:e95089; doi: 10.2196/95089

Keywords: rheumatoid arthritis; disease flare prediction; wearable devices; digital biomarkers; remote patient monitoring; smartphone; deep learning; real-world data; longitudinal data; AI

Introduction

Rheumatoid arthritis (RA) is a chronic immune-mediated condition characterized by synovial inflammation. It affects approximately 0.5% to 1% of adults in the United States and is responsible for significant disability [1,2]. Disease flares, or exacerbations, are characterized by joint pain, swelling, morning stiffness, and functional impairment, with effects on physical function and quality of life that can persist for months [3,4]. Current treatment guidelines emphasize inflammation control to prevent progressive joint damage [5-7]. However, flares remain common, occurring in at least 35% of patients annually and are poorly predicted [8,9].

Traditional biomarkers such as C-reactive protein (CRP) are routinely used to assess inflammation in RA, but CRP may be normal in a substantial proportion of patients with active disease or histologic inflammation [10]. Conversely, symptoms such as pain, stiffness, and fatigue may be influenced by central sensitization, psychosocial factors, pain catastrophizing, comorbidities, and individual resilience, and therefore do not always correlate directly with peripheral inflammation [11-13]. Early detection and prediction of flares could enable prompt therapeutic adjustment and improve long-term outcomes [5,14-16]. However, current monitoring relies on in-person physician visits and periodic laboratory assessments, resulting in significant uncertainty for patients, many of whom resort to self-management of flares [5,17].

Recent advances in wearable sensors and smartphone technologies have introduced new opportunities for continuous remote monitoring in RA, allowing passive, noninvasive, and longitudinal evaluation of disease activity [18]. Activity trackers and actigraphy-based measures have been used in RA studies to compare physical activity patterns between patients with RA and healthy controls and to identify flare-associated reductions in activity [18-20]. However, these devices provide limited insight into underlying physiology. In a study by Creagh et al [21] evaluating 30 individuals with moderate-to-severe RA and 30 healthy controls, a machine learning (ML) framework combining app-based symptoms, actigraphy, and heart sensor data distinguished RA status and severity. However, this work did not evaluate flare prediction.

Building on this work, we recently demonstrated that physiological signals from commercially available wearables, including heart rate (HR) and heart rate variability (HRV), change significantly during both symptomatic and inflammatory RA activity and begin to diverge from baseline up to 4 weeks before flare onset [22]. These findings suggest that continuous physiological monitoring could enable earlier flare detection and, if prospectively validated, support remote monitoring tools for patients and clinicians to better assess evolving disease activity. Therefore, the aim of this study was to develop and evaluate deep learning (DL) time-series models that integrate wearable-derived physiological features, demographic variables, and patient-reported outcomes (PROs) to forecast symptomatic and inflammatory RA activity up to 4 weeks in advance. We hypothesized that changes in longitudinal HR, HRV, and activity patterns captured by consumer wearables would predict RA flare states, that DL models applied to these data would outperform classical ML approaches in the setting of incomplete real-world wearable data, and that integration of PROs would improve participant-level flare prediction.

Methods

Dataset

Data Collection

The RA Forecast study enrolled 53 individuals with RA between 2023 and 2024 from throughout the United States, collecting physiological data through commercially available wearable devices, including the Apple Watch, Fitbit, and Oura Ring [22]. Baseline demographics, laboratory values, and PROs were collected using our custom ehive smartphone application [23]. Participants were eligible if they had a confirmed diagnosis of RA, were aged 18 years or older, and were receiving medication for RA management. Exclusion criteria included pregnancy and other chronic conditions, a pacemaker or defibrillator, and the use of medications known to affect HR or HRV (eg, beta-blockers, benzodiazepines). To enroll, participants needed to be willing to use a wearable device (Apple Watch, Fitbit, or Oura Ring), either their own or one provided by the study team. For participants within the Mount Sinai Health System who did not provide demographic information via the app, data were supplemented from the electronic health record (EHR). Missing values for age (n=5) and sex (n=1) were imputed using multivariate imputation methods, incorporating the complete set of available demographic and wearable-derived physiological features across participants. Participants completed all study activities remotely from their homes.

Sensor Data Description and Physiological Features Extraction

Participants were recommended to use their devices at least 4 days per week and for a minimum of 8 hours per day. Physiological data from commercial wearables are typically preprocessed by device manufacturers. Because signal quality and available data streams vary by device, all physiological

data from the 3 wearables (Apple Watch, Fitbit, and Oura Ring) were cleaned prior to feature extraction. Outlier values outside physiological ranges were excluded, including HR <30 or >200 beats per minute and resting heart rate (RHR) <30 or >150 beats per minute. Daily step counts were retained up to 60,000 steps, while values <50 steps were removed as they likely reflected inadequate wear time. Step counts are particularly sensitive to missing wear periods, as it represents a cumulative daily measure.

To enable daily flare prediction, sensor signals were processed to derive daily features used as model inputs. The raw HR data were aggregated into daily features: mean daily HR (HR), minimum daily HR (HR_{min}), mean daytime HR (HR_{day}), and mean nighttime HR (HR_{night}). Fitbit and Oura reported HRV as the root-mean-square of successive RR interval differences (RMSSD) during detected sleep sessions. These were similarly aggregated into daily (HRV_{RMSSD}), nighttime ($HRV_{night_{RMSSD}}$), and daytime ($HRV_{day_{RMSSD}}$) values. The Apple Watch reported HRV using the SD of NN intervals and interbeat intervals; interbeat intervals were converted to HRV_{RMSSD} to harmonize HRV estimates across devices. Daily step counts and RHR were extracted from all 3 devices.

PROs and Laboratories

In addition to wearable sensor data, participants completed smartphone-based surveys at scheduled intervals. Psychological resilience was assessed at enrollment using the 10-item Connor-Davidson Resilience Scale (CDRISC-10) [24]. Biweekly surveys measured stress using the Perceived Stress Scale 4 (PSS4) [25], and functional status and quality of life using the Multidimensional Health Assessment Questionnaire (MDHAQ) [26]. Two daily surveys included the Routine Assessment of Patient Index Data 3 (RAPID3) [27]—a validated subset of the MDHAQ capturing disease activity through physical function, pain, and a patient global assessment—and a 10-point fatigue rating. Participants also self-reported CRP levels through the study app; for those enrolled at Mount Sinai, additional laboratory results were extracted from their EHRs.

RA Flares: Inflammatory and Symptomatic Definitions

RA is characterized by intermittent exacerbation of disease activity, or flares. Although validated measures of RA activity exist [27,28], symptom expression varies considerably among individuals, and patient-reported symptoms may diverge from clinical assessments [12,29]. Clinical evaluation, combining physician scores and PROs, remains the gold standard but occurs only during infrequent clinical visits. To enable continuous flare prediction, 2 operational flare definitions were applied: symptomatic and inflammatory. Symptomatic flares were defined using daily RAPID3 scores. It has been validated and strongly correlates with the scores obtained by doctors and clinical investigators using the Disease Activity Score 28 (DAS28) and Clinical Disease Activity Index (CDAI), and therefore is a widely acceptable tool to monitor disease activity [27]. Scores range from 0 to 30 and are used to categorize disease activity as near-remission (≤3), low (3-6), moderate (6-12), and high (>12). Symptomatic activity was defined as a RAPID3 >3. To minimize the influence of day-to-day variability and missing responses, a 7-day rolling window (±3 d) was applied. A given day was labeled "flare" if at least 4 days within that window had RAPID3 scores, and at least 3 of those scores were consistent with a flare. Inflammatory flares were defined as CRP values >5 mg/L, independent of symptom status. CRP was selected because it is a routinely obtained objective marker of systemic inflammation in RA care and is commonly used alongside symptoms, physical examination, and composite disease activity measures to inform disease monitoring and treatment decisions. Because CRP is nonspecific and may be influenced by comorbid inflammatory conditions, infection, obesity, and medication exposure, this outcome was considered an operational measure of systemic inflammation rather than a definitive RA-specific flare adjudication. CRP measured at discrete time points was imputed over a 15-day window (±7 days around the test date).

Flare Prediction

Prediction Overview

All models incorporated daily physiological features derived from wearable sensors and were developed to predict daily symptomatic and inflammatory flares up to N_{pred}=0, 7, 14, 21, and 28 days prior to flare onset (N_{pred}=0 denotes same-day prediction).

Missing Data Handling

Wearable data collected in real-world settings are prone to missingness, arising from 2 main sources: (1) nonwear days, when participants did not wear the device (with or without completing surveys), and (2) low-quality intervals, when sensor signals were unreliable due to poor skin contact or excessive motion and were automatically filtered out by the device. To ensure comparability, we applied uniform processing steps to all 3 devices, including outlier removal, daily aggregation, and imputation. Since the effects of imputation of missing data are not well established for ML models [30], we evaluated two alternative workflows: (1) Complete-case analysis excluded calendar days with missing values for any input feature, yielding a smaller but higher-quality dataset. (2) Participant-specific statistical imputation was applied to each participant and physiological variable to address intermittently missing data, resulting in a fully populated feature matrix.

Model performance was compared across the 2 workflows to assess the impact of missing-data handling strategies on flare-prediction accuracy.

ML Model

Several classical ML algorithms were first benchmarked for flare prediction. Two temporal resolutions were evaluated: (1) single-day inputs (N_{win}=1) and (2) 7-day aggregated inputs (N_{win}=7). The candidate feature set (for N_{win}=1) included HR and HRV variables, with the addition of step count. For N_{win}=1, statistical features of each of these physiological

signals were calculated. Feature selection procedures were applied within the training set to reduce overfitting.

Model development used a 70/30 train-test split applied longitudinally per participant, with the first 70% of days used for training. Within the training set, 3-fold cross-validation was used for hyperparameter optimization and model selection. The random forest classifier model with standardized features and balanced class weights demonstrated the best performance for both symptomatic and inflammatory flares. The classification thresholds were selected by maximizing the training-set F_1-score.

Time-Series–Based DL Models

In contrast to classical ML models, which depend on statistical features computed over predefined windows, time-series–based DL methods are known to improve results without requiring additional feature-engineering steps. We evaluated three architectures: (1) long short-term memory (LSTM), (2) transformer, and (3) an LSTM+transformer fusion model with an LSTM front-end. LSTMs address the vanishing-gradient problem, allowing gradual physiological drift to be modeled over time. Transformer models use self-attention mechanisms that allow each time point to relate to all others. Two transformer variants were tested, with and without sinusoidal positional encoding (PE), to assess the impact of explicit temporal order.

Daily physiological variables were arranged into 7-day sliding windows (N_{win}=7). Model inputs included daily HR, HRV, and activity-related features harmonized across devices. All days in N_{win} were continuous, with missing values imputed as described above. Separate models were trained independently for each prediction horizon (N_{pred}=0, 7, 14, 21, and 28 d) using consistent architecture, early stopping, and hyperparameter tuning.

Performance Evaluation

Model performance was assessed with nonoverlapping, out-of-sample, and out-of-time test sets. For each participant, the first 70% of chronologically ordered study days were assigned to the training set, and the remaining 30% to the test set. Model training incorporated early stopping based on training loss. Architecture-specific hyperparameters were optimized using a random search strategy, with model selection based on predictive performance while limiting overfitting. Reported metrics include receiver operating characteristic area under curve (ROCAUC), area under precision-recall curve (AUPRC), precision, sensitivity, specificity, and accuracy, with thresholds determined by the probability value that maximized the F_1-score in the training set. Pairwise differences in ROCAUC were evaluated using DeLong nonparametric method. In addition to aggregated metrics, we computed sensitivity on a per-subject basis. This analysis determined whether predictive performance gains were driven by a small number of individuals with longer observation periods or were distributed across the entire cohort, independent of the proportion of flare vs remission days. Because per-participant variability was more pronounced for inflammatory flare prediction, given the smaller dataset, these analyses were focused on that subgroup. We further explored methods to reduce variation in per-participant sensitivity for inflammatory flare prediction by augmenting the physiological baseline model with PROs. In addition to daily RAPID3 and fatigue scores, biweekly RAPID3 scores were estimated from MDHAQ surveys, reflecting participant status over the preceding 2 weeks. Both biweekly survey scores (RAPID3 biweekly and PSS4) were imputed over the prior 2 weeks. To identify demographic, behavioral, and comorbidity factors associated with flare-state persistence, continuous variables (eg, age and BMI) were compared between participant groups using the Mann-Whitney U test, and categorical variables (eg, sex and race) were compared using Fisher exact test.

Ethical Considerations

This study was approved by the Institutional Review Board of the Icahn School of Medicine at Mount Sinai (STUDY-22-01589). All participants provided informed consent prior to participation. Study procedures were conducted in accordance with the ethical standards of the Mount Sinai Institutional Review Board and the principles of the Declaration of Helsinki. Participant privacy and confidentiality were maintained throughout the study. Data collected through the smartphone application, wearable devices, laboratory results, and EHR review were stored using study-specific identifiers and analyzed in a deidentified manner. Access to identifiable participant information was restricted to authorized study personnel. No compensation was provided to participants.

A protocol specific to this deep learning analysis was not prepared or made publicly available. This analysis was not registered in a public study or trial registry.

Results

Study Cohort Characteristics and Data Availability

We analyzed a dataset consisting of overlapping daily physiological measurements and corresponding flare-outcome days. In total, 49 out of 53 enrolled participants contributed wearable and patient-reported data, with 47 included in the symptomatic-flare prediction cohort and 37 in the inflammatory-flare prediction cohort. The mean age was 51.4 (SD 14.6) years, and mean BMI was 28.9 (SD 7.9) kg/m^2. Out of the 49 participants, 44 (89.8%) were female and 29 self-identified as White. Sixteen participants reported a single comorbidity, 11 reported 2 or more, and 22 reported none. Table S1 and Figure S1 in Multimedia Appendix 1 summarize the demographic characteristics of the study cohort.

Thirty-one participants used an Apple Watch, 15 used a Fitbit, and 3 used an Oura Ring as the primary device, defined as the device worn for the majority of the duration if multiple devices were used. Across all participants, 5624 total days of data were collected, averaging 115 (SD 100) days of overlapping wearable and flare status data per participant. In the symptomatic-flare cohort, each participant contributed an average of 111 (SD 104) days, including a mean of 39 (SD

66) flare days. In the inflammatory-flare cohort, participants contributed an average of 26 (SD 19) days, including 7 (SD 12) days during flare. Additional information about raw data completeness and missing wearable data during flare periods is provided in Table S2 in Multimedia Appendix 1.

Daily RAPID3 assessments and CRP measurements (interpolated over ±7 d) were simultaneously available over 262 participant-days. Elevated RAPID3 (>3) was observed on 51.5% (n=135) of days, whereas elevated CRP (>5 mg/L) was observed on 16.8% (n=44) of days. Both measures were concurrently elevated on only 9.5% (n=25) of participant-days, highlighting the limited overlap between symptomatic and inflammatory disease activity in our cohort.

Characterization of Flare State Dynamics

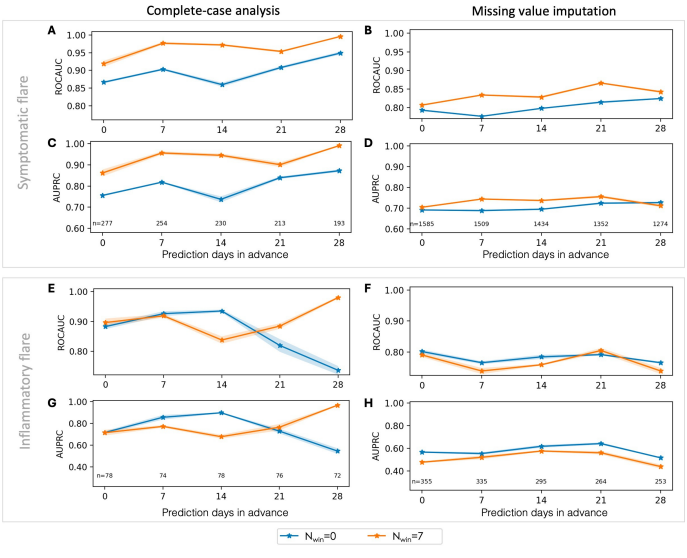
Within the inflammatory flare cohort, 86% (32/37) of the participants remained persistently in either flare (n=9) or remission (n=23) throughout the study, compared with 64% (30/47) in the symptomatic flare cohort (including 18 and 12 participants in persistent flare and remission states, respectively). Among participants with at least one flare-to-remission or remission-to-flare state transition, the mean number of transitions was 5.9 (SD 5.8) for symptomatic flares and 1.4 (SD 0.6) for inflammatory flares. Detailed transition counts are summarized in Table S3 in Multimedia Appendix 1, and participant-level flare trajectories are shown in Figure S2 in Multimedia Appendix 1.

To investigate factors associated with persistent flare states, we compared demographic, behavioral, and comorbidity characteristics among participants with persistent flare, persistent remission, and at least one state transition. Variables examined included age, sex, BMI, race/ethnicity, smoking history, and comorbidity burden (any comorbidity and the 3 most common: asthma, osteoporosis, and hypertension). BMI differed significantly between the persistent flare and persistent remission groups (median symptomatic: 32.1, IQR 26.5-39.8 vs 23.1, IQR 22.3-24.4 kg/m^2; median inflammatory: 38.9, IQR 29.5-42.4 vs 22.7, IQR 21.5-27.9 kg/m^2), in both the symptomatic (P=.001) and inflammatory (P=.002) cohorts. BMI was also significantly higher in the persistent flare group than in the transition group for the symptomatic cohort (median 32.1, IQR 26.5-39.8 vs 25.6, IQR 21.3-32.5 kg/m^2; P=.03). In the symptomatic cohort, the presence of asthma differed significantly between the persistent flare and remission groups (7/18, 38.9% vs 0/12, 0.0%; P=.02), as well as between the persistent flare group and the remaining participants (7/18, 38.9% vs 3/29, 10.3%; P=.03). Given the limited number of participants in each subgroup, particularly within the inflammatory cohort, these findings should be interpreted with caution.

ML-Based Flare Prediction

Using a 7-day window (N_{win}=7) with feature imputation, the random forest classifier achieved a mean ROCAUC/AUPRC of 0.83/0.74 for symptomatic flares and 0.77/0.55 for inflammatory flares when predicting outcomes 7 days in advance (N_{pred}=7). When observations containing any missing values were excluded through complete-case analysis, performance improved with the ROCAUC increasing by 10.0% to 15.3% and the AUPRC by 27.4% to 29.7% across both flare types (Figure 1). However, this improvement came at the cost of a substantial (~6-fold) reduction in sample size. Thus, ML models face a 2-fold challenge: (1) sensitivity to naïve imputation strategies or (2) the inability to make predictions on days with incomplete data, both of which limit their applicability in real-world settings.

Figure 1. Performance of random forest models for rheumatoid arthritis (RA) flare prediction across varying prediction horizons (N_{pred}=0, 7, 14, 21, 28 days, x-axis) on test data. Results are shown for 2 missing data handling strategies: complete-case analysis (left panels) and when missing values are imputed (right panels), with n showing total samples. Panels A to D illustrate symptomatic-flare prediction where 7-day features aggregation (N_{win}=7) consistently yields the highest discriminative performance. Panels E to F depict inflammatory-flare prediction. Shaded regions represent ±1 SD around the mean performance curves. Panels C, D, G, and H display the corresponding number of test-day samples. AUPRC: area under precision-recall curve; ROCAUC: receiver operating characteristic area under curve.



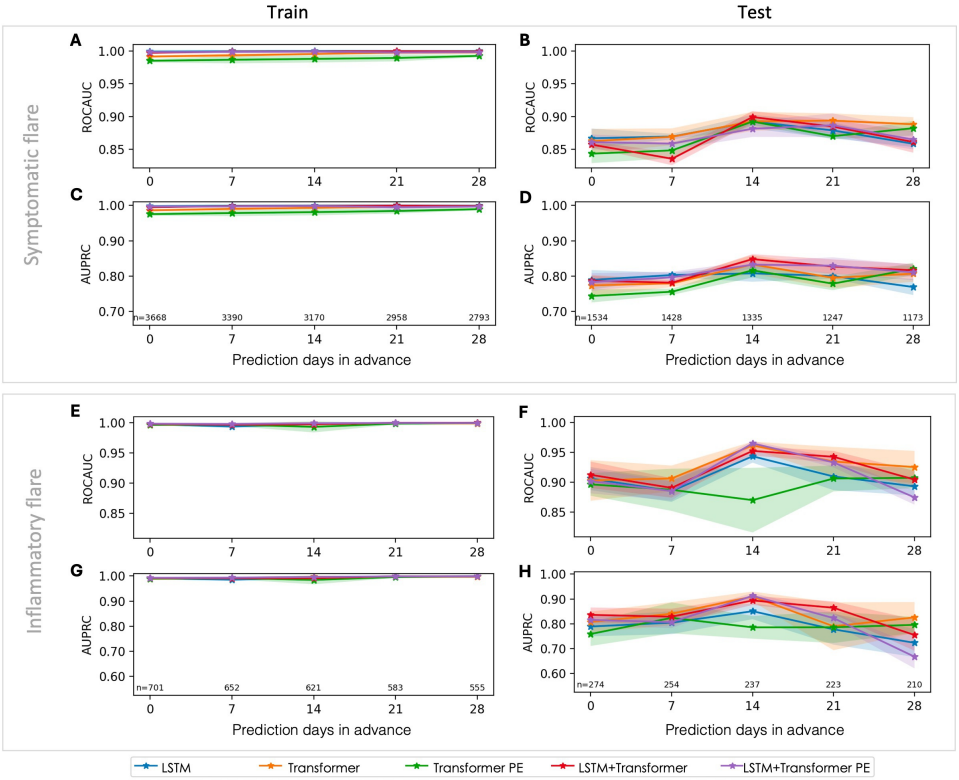
Seven-day feature aggregation (N_{win}=7) outperformed single-day inputs (N_{win}=1) for symptomatic flare prediction across nearly all prediction horizons (N_{pred}), with an increase of 6.9% and 7.6% in ROCAUC and AUPRC, respectively, for N_{pred}=7. In contrast, no clear performance gain was observed for inflammatory flare prediction, likely reflecting the smaller sample size within that subgroup. Model performance remained stable across prediction horizons up to 28 days, aligning with the strong temporal persistence of flare and remission states in this cohort.

DL-Based Flare Prediction

Model Performance and Feature Importance

To overcome the sensitivity of traditional ML models to missing features and incomplete wear-day data, DL models were designed to handle sparsity and enhance robustness to missing features, thereby supporting development in real-world scenarios. Five DL architectures were evaluated following hyperparameter optimization. Overall, model performance was comparable across architectures (Figure 2). The temporal Transformer model achieved the highest mean ROCAUC across all prediction horizons (N_{pred}), while the LSTM+Transformer fusion model yielded the highest average AUPRC for both flare types (Table S4 in Multimedia Appendix 1). Given the comparable performance across horizons, subsequent analyses focused on the more parsimonious transformer model. Relative to the ML random forest model with imputation, the DL-based transformer model improved ROCAUC by an average of 5.3% for symptomatic flares and 18.4% for inflammatory flares and improved AUPRC by 9.3% and 44.1%, respectively. Detailed sample counts and test-set metrics for each prediction horizon (N_{pred}) are provided in Tables S5 (symptomatic) and S6 (inflammatory) in Multimedia Appendix 1, with thresholds selected to maximize the training F_1-score.

Figure 2. Receiver operating characteristic area under curve (ROCAUC) and area under precision-recall curve (AUPRC) performance of 5 deep learning architectures for flare prediction across multiple prediction horizons (N_{pred}=0,7,14,21 and 28 days). Models evaluated include (1) long short-term memory (LSTM), (2) transformer, (3) transformer with positional encoding (PE), (4) LSTM+transformer, and (5) LSTM+transformer PE. Left panels display training performance, and the right panels show test performance. Panels A to D depict symptomatic-flare prediction, with test-set ROCAUC ranging from 0.84 to 0.90 and AUPRC from 0.74 to 0.85. Panels E to H present inflammatory-flare prediction, with test-set ROCAUC between 0.87 and 0.96 and AUPRC between 0.67 and 0.91. Across models, training performance exceeded test performance, consistent with mild overfitting.



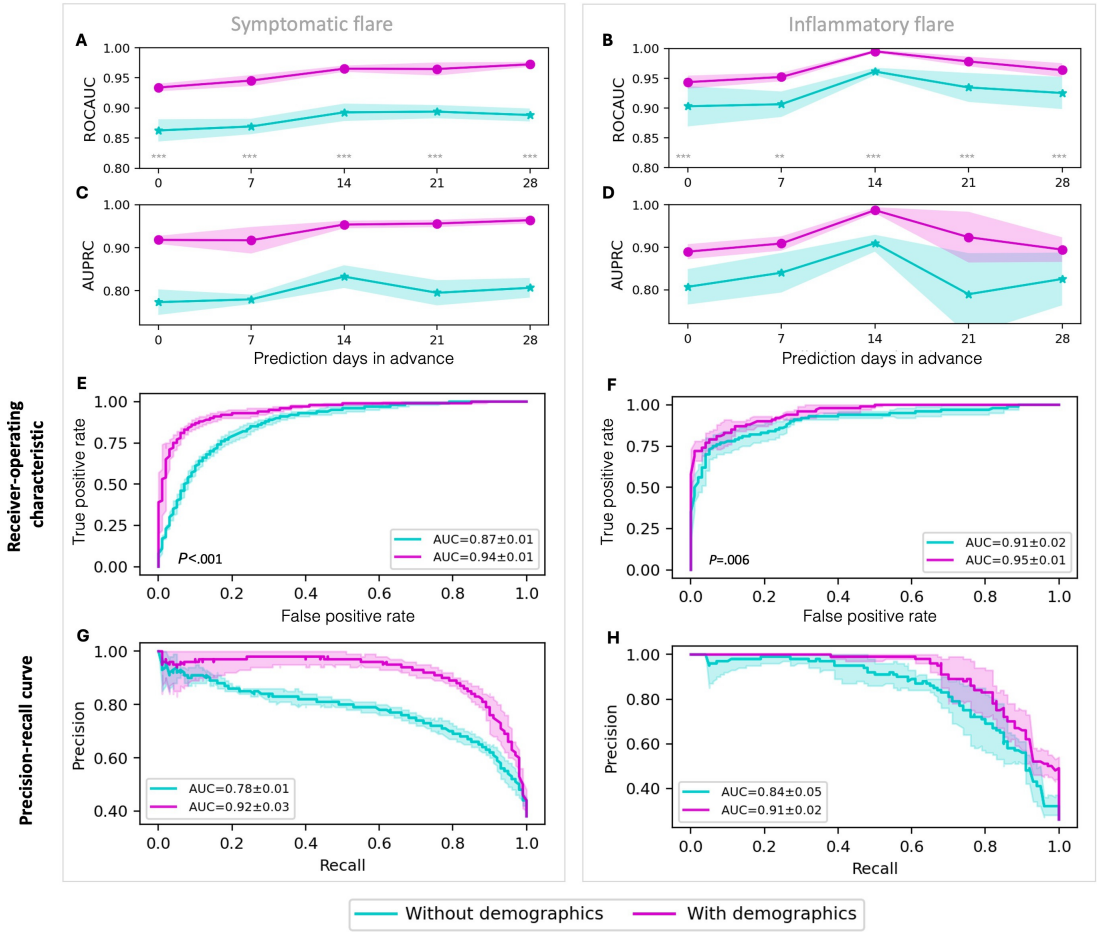
Performance remained consistent across Apple Watch, Fitbit, and Oura Ring devices for both flare types (Table S7 in Multimedia Appendix 1), with a slight reduction in performance for inflammatory flare prediction using Fitbit data (n=12). Shapley additive explanations (SHAP) analysis showed that the DL models relied on clinically meaningful physiological features, with HR- and RHR-based measures consistently among the most influential predictors across flare types, regardless of whether demographic features were included in the models or not.

Impact of Demographic Features

We next evaluated whether incorporating demographic variables (age, sex, and BMI) improved predictive performance. Figure 3 compares Transformer models trained using only wearable-derived physiological features with those augmented by demographic information. Across all prediction

horizons and flare categories, the inclusion of demographic features improved predictive performance. For 7-day-ahead prediction (N_{pred}=7), ROCAUC significantly increased from 0.87 to 0.95 (9.2% improvement, DeLong test P<.001) for symptomatic flares and from 0.91 to 0.95 (4.4% improvement, P=.006) for inflammatory flares. AUPRC improved from 0.78 to 0.92 (18.0%) and from 0.84 to 0.91 (8.3%), respectively. Figure 3 shows corresponding improvement in precision, sensitivity, specificity, and accuracy plots with the addition of demographic features. Comprehensive performance metrics for symptomatic and inflammatory flare prediction with the addition of demographic features are provided in Tables S8 and S9 in Multimedia Appendix 1.

Figure 3. Effect of demographic features on Transformer-based flare prediction. Left panels depict symptomatic-flare prediction and right panels depict inflammatory-flare prediction. Cyan curves represent models trained without demographic variables, while magenta curves represent models including demographics; shaded areas denote ±1 SD across runs. Panels A to D show mean test receiver operating characteristic area under curve (ROCAUC) and area under precision-recall curve (AUPRC) values across prediction horizons (N_{pred}=0,7,14,21, and 28 days). DeLong test confirms area under the curve (AUC) differences between models with and without demographics are statistically significant, with *** and ** denoting P<.001 and P<.01, respectively. Panels E to H illustrate ROC (top row) and precision-recall (PR) curves (bottom row) for N_{pred}=7 days, demonstrating statistically significantly different AUCs for both prediction outcomes. Incorporating demographic variables consistently enhanced ROCAUCs and AUPRCs across all prediction horizons and flare types.



The observed improvement with demographic features may reflect participant-specific disease characteristics and the persistence of flare or remission states within individuals. To evaluate whether performance gains were driven primarily by participants with persistent states, we performed a sensitivity analysis restricted to participants exhibiting at least one flare-to-remission or remission-to-flare state transition. In the symptomatic cohort (n=17), incorporating demographic features continued to improve performance, increasing ROCAUC from 0.77±0.06 to 0.83±0.06 and AUPRC from 0.55±0.13 to 0.70±0.14. In contrast, no consistent improvement was observed for inflammatory flares, likely reflecting the limited sample size (n=5). Complete performance metrics for this subset are provided in Table S10 in Multimedia Appendix 1.

With the addition of demographic variables, HR-derived features remained among the top predictors, while BMI emerged as an important predictor for both flare types. Age was found to be an important predictor of symptomatic flares. The relative ranking of wearable-derived physiological features also shifted, particularly for inflammatory flare prediction. These findings suggest that demographic variables provide additional predictive information beyond wearable-derived signals.

Impact of PROs for Inflammatory Flare Prediction

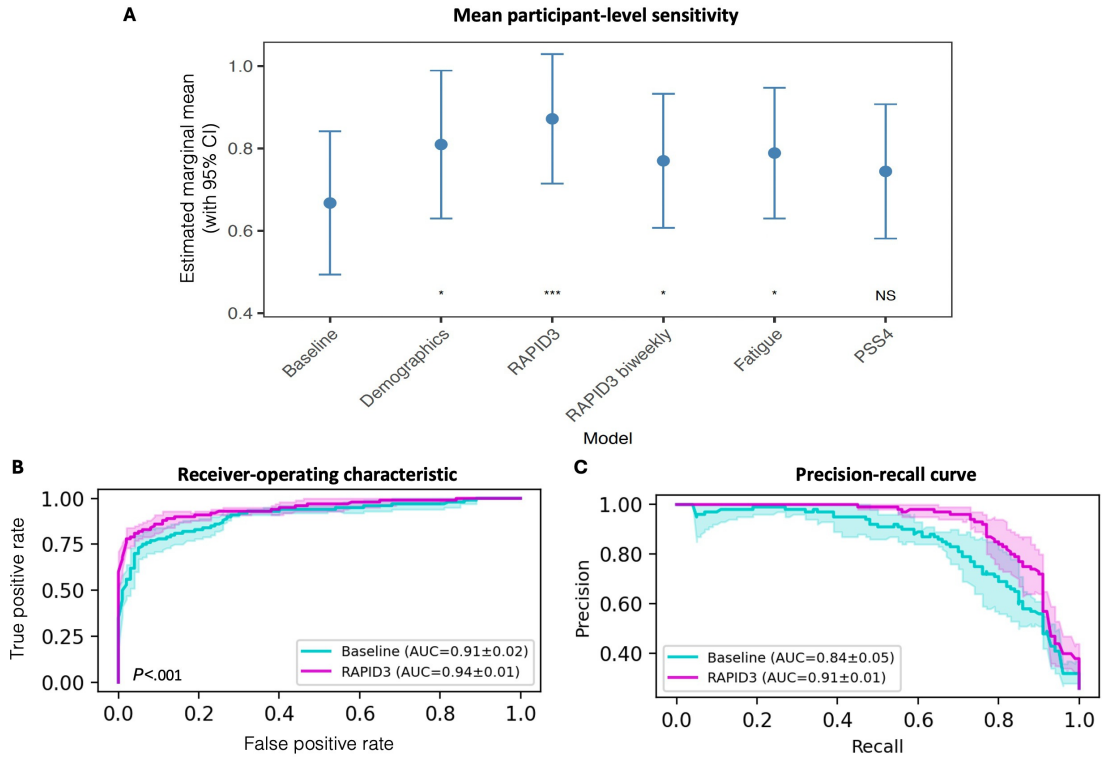
Because participants exhibited wide variability in flare and remission duration, we evaluated performance not only at the aggregate level but also across individual participants. Sensitivity was chosen as the primary patient-level metric. The distribution of per-participant sensitivities was modeled using a linear mixed-effects framework, with participants treated as a random-effect. The resulting expected marginal mean (EMM) was used as a summary performance estimate. For 7-day-ahead prediction (N_{pred}=7), the baseline model using only physiological features achieved an EMM sensitivity of 0.67 (95% CI 0.49-0.84). Incorporating demographic variables increased EMM sensitivity to 0.81, while adding daily RAPID3 survey scores further improved it to 0.87, with narrower CI. Table 1 summarizes EMM and ROCAUC metrics across all feature sets: baseline (physiological features only), baseline+demographics, and baseline+PROs. PROs included the daily and biweekly RAPID3 scores, daily fatigue ratings, and biweekly PSS4. Figure 4 presents EMM sensitivities with corresponding CIs across feature sets, while the lower panels display ROC and precision-recall curves comparing the RAPID3-augmented model with the baseline. DeLong test confirmed significant AUC improvements with PRO augmentation (P<.001). Figure S4 in Multimedia Appendix 1 illustrates the narrower interparticipant variability in sensitivity achieved with PRO integration. Overlaid ROCAUC and AUPRC curves further demonstrate consistent performance gains relative to the baseline.

Table 1. Participant-level sensitivity by feature set, using estimated marginal means (EMMs), for 7-day inflammatory flare prediction (N_{pred}=7).

Feature set	Baseline	Demographics	RAPID3[a]	RAPID3[a] biweekly	Fatigue	PSS4[b]
Participant-level sensitivity, EMM (95% CI)	0.67 (0.49-0.84)	0.81 (0.63-0.99)	0.87 (0.72-1.03)	0.77 (0.61-0.93)	0.79 (0.63-0.95)	0.74 (0.58-0.91)
ROCAUC[c]	0.91	0.95	0.94	0.94	0.95	0.92

[a]RAPID3: Routine Assessment of Patient Index Data 3.
[b]PSS4: Perceived Stress Scale 4.
[c]ROCAUC: receiver operating characteristic area under curve.

Figure 4. Effect of incorporating patient-reported outcomes (PROs) into the transformer model with physiological features for inflammatory-flare prediction (N_{pred}=7). The top panel displays participant-level sensitivity via estimated marginal means (EMMs) derived from a linear mixed-effects model. Each point represents the mean participant-level sensitivity, and vertical bars indicate 95% CIs. Models along the x-axis include the baseline (physiological features only) and sequential addition of demographic covariates, daily Routine Assessment of Patient Index Data 3 (RAPID3), biweekly RAPID3, daily fatigue, and biweekly Perceived Stress Scale 4 (PSS4). Asterisks denote statistical significance of each model's contrast relative to baseline (*P<.05; ***P<.001). The bottom panels show receiver operating characteristic (ROC) and PR curves comparing the baseline model with the daily RAPID3-augmented model.



Discussion

Principal Findings

The aim of this study was to develop and evaluate DL time-series models that integrate wearable-derived physiological features, demographic variables, and PROs to predict symptomatic and inflammatory RA flares up to 4 weeks in advance. In this prospective longitudinal cohort, physiological signals from commercially available wearable devices predicted future RA disease activity. DL models outperformed classical ML approaches and maintained strong discriminative performance across multiple prediction horizons. Transformer-based models incorporating wearable-derived physiological features and demographic variables achieved high performance for both symptomatic and inflammatory flare prediction. In addition, the incorporation of daily PROs, particularly RAPID3 scores, improved participant-level sensitivity for inflammatory flare prediction. These findings support the hypothesis that longitudinal wearable-derived physiological patterns, when analyzed with time-series DL methods, may enable personalized forecasting of RA flares.

Interpretation and Comparison With Prior Work

The current paradigm of RA monitoring relies on episodic in-person assessments and intermittent laboratory testing, capturing only single time point evaluations of disease activity. Commonly used disease activity indices, such as the DAS28, CDAI, and Simple Disease Activity Index (SDAI), require clinician assessment and laboratory data, limiting their use outside of the clinical setting [31]. Although these measures are central to RA management, they provide discrete assessments of a continuous disease process that fluctuates over time and may fail to capture changes between visits. Given the dynamic nature of RA, up to 10% of patients experience clinically meaningful flares between office visits even when seeing a physician at 3-month intervals [32]. Moreover, symptoms and inflammation can occur discordantly, creating uncertainty for both patients and clinicians when attempting to distinguish inflammatory disease activity from pain, fatigue, and functional limitations that contribute to symptom burden [11]. Currently, no validated approach is routinely used in clinical care to anticipate RA flare onset before symptoms or objective inflammation is recognized. Our study addresses this critical gap by developing models capable of identifying and predicting both symptomatic and inflammatory flares in advance, enabling earlier identification of disease worsening and potentially timelier intervention.

Prior studies using wearable devices in RA have primarily focused on physical activity and actigraphy-based measures, demonstrating associations between symptomatic flares and reduced activity levels [18-20]. Active RA has also been associated with autonomic dysregulation, which may be captured indirectly through wearable-derived measures such as HR, RHR, and HRV [33]. Our prior work extended these observations by demonstrating that wearable-derived physiological signals, including HR and HRV, change during symptomatic and inflammatory RA activity and prior to flare onset [22]. The present study extends this work by evaluating whether longitudinal physiological data can be used to forecast future symptomatic and inflammatory flare states across clinically relevant prediction horizons.

The physiological signals evaluated in this study should be interpreted as markers of broader physiologic perturbation associated with RA activity rather than as disease-specific biomarkers in isolation. HR, RHR, and HRV are indirect measures of autonomic function and may be influenced by inflammatory, infectious, behavioral, and psychological stimuli. Similarly, activity-derived features may reflect joint pain, stiffness, and functional impairment but are also affected by device wear time, gait limitations, and contextual factors unrelated to RA. However, the temporal pattern of these perturbations may provide important disease-contextual information. In our prior RA work [22], wearable-derived physiologic changes persisted for weeks before symptomatic and inflammatory flare activity, and similar sustained preflare alterations have been observed in our inflammatory bowel disease studies [34]. This contrasts with our COVID-19 work, in which physiologic changes emerged closer to the infection diagnosis and normalized over a shorter period after diagnosis [35]. These differences suggest that although wearable-derived HR, RHR, HRV, and activity measures are not disease-specific on their own, their timing, duration, and integration with disease-relevant PROs and inflammatory biomarkers may help distinguish chronic inflammatory disease activity from acute physiologic stressors.

The distinction between symptomatic and inflammatory flare prediction is clinically important. The RAPID3, which is a validated assessment of RA activity, may also be influenced by other musculoskeletal conditions such as osteoarthritis, fibromyalgia, and other comorbidities [36]. CRP provides an objective inflammatory measure, but it is nonspecific and may be affected by other inflammatory conditions [37]. By evaluating both symptomatic and inflammatory flare definitions, this study provides an initial framework for distinguishing patient-experienced disease activity from systemic inflammation. At the same time, these definitions should be viewed as operational labels for model development rather than definitive measures of RA flare biology.

DL models outperformed classical ML approaches, particularly when missing wearable data were imputed. This is important because missingness is inherent to real-world wearable monitoring, reflecting variable device wear and sensor quality. DL models were able to learn temporal patterns from partially observed longitudinal data, supporting their potential utility in real-world remote monitoring settings. Transformer architectures may be especially useful in this context because they can capture nonlinear relationships and temporal dependencies across physiological signals without relying solely on manually engineered summary features.

The improvement observed after adding demographic variables highlights the importance of individualized context when interpreting wearable-derived physiology. Age, sex,

and BMI may influence physiologic metrics and therefore provide important context for personalizing model predictions. Because many participants remained consistently in either flare or remission, demographic variables may have helped the model account for stable differences between individuals, in addition to any dynamic physiological changes that occurred before flare transitions. These findings support the potential value of demographic context for model personalization and highlight the need for validation in larger, more diverse cohorts with greater numbers of flare-remission state transitions.

PROs provided complementary information and improved participant-level sensitivity for inflammatory flare prediction. Wearable-derived features may capture autonomic, behavioral, and functional changes, whereas RAPID3 and fatigue scores reflect patient-experienced symptoms and function. Together, these data streams may offer a more complete representation of RA activity than either source alone. Clinically, this supports a future monitoring framework in which passive digital biomarkers function as an early warning layer that can prompt targeted symptom assessment, laboratory testing, or clinical review, while remaining adjunctive to standard disease activity measures and physician judgment.

In this proposed monitoring framework, model evaluation should reflect the intended clinical use case. Thus, the temporally ordered train-test strategy was designed to approximate personalized forecasting, in which an initial observation period is used to calibrate predictions for future days within the same individual. This approach is more aligned with within-person remote monitoring than random observation-level splitting. However, it does not establish performance in previously unseen patients or the number of days required for personalization. Future studies should therefore evaluate subject-level validation, external cohorts, and prospective real-time prediction before clinical implementation.

Limitations

Several limitations should be considered. First, the sample size was modest, and the inflammatory flare cohort was smaller because CRP testing was performed as part of routine clinical care rather than at fixed study intervals. This limited temporal precision around inflammatory flare onset and resolution, and may have introduced ascertainment bias, as testing may have been more likely during periods of suspected disease activity. In addition, CRP-defined flares should be interpreted as an operational inflammatory outcome rather than a definitive RA-specific flare definition. Although CRP is routinely used in RA care as an objective marker of inflammatory burden, it is nonspecific and may be influenced by obesity, infection, comorbid inflammatory conditions, and medications such as corticosteroids, interleukin-6 inhibitors, or Janus kinase inhibitors. Second, symptomatic flares were defined using RAPID3, which is feasible for remote monitoring but may be influenced by other musculoskeletal conditions such as noninflammatory pain, osteoarthritis, and fibromyalgia. This may be particularly relevant for participants with persistently elevated RAPID3 scores, in whom models may partly reflect stable symptom burden rather than discrete flare onset. Despite these limitations, symptoms, physical function, and patient global assessment remain central domains in RA monitoring and treatment decisions, supporting the clinical relevance of modeling RAPID3-defined symptomatic flare states in remote monitoring frameworks. Third, many participants remained persistently in either flare or remission, particularly in the inflammatory flare cohort (Table S3 in Multimedia Appendix 1), limiting the number of flare-remission state transitions available for model learning. Performance analyses restricted to participants with at least one state transition continued to show improved prediction with the addition of demographic features in the symptomatic cohort, suggesting that the observed performance gains cannot be explained solely by persistence of disease state. Nevertheless, because relatively few transition events were available, particularly for inflammatory flares, some of the reported predictive performance likely reflects forecasting of persistent disease state in addition to physiological changes preceding flare transitions. Accordingly, the current analysis should be interpreted as forecasting future daily flare/remission state, rather than as definitive event-level prediction of new flare onset. Larger cohorts with more frequent objective inflammatory assessments and greater numbers of flare transitions are needed to better evaluate flare onset, persistence, and resolution. Fourth, the within-participant 70/30 train-test split approximates personalized forecasting after an individual calibration period, but it does not establish performance in previously unseen patients or with smaller calibration duration. Given the limited cohort size, a separate validation set was not retained; consequently, the temporally held-out test set was used for hyperparameter selection, while early stopping relied on training loss. A larger dataset would enable dedicated training, validation, and test sets for more rigorous hyperparameter optimization, early stopping, and unbiased performance estimation. External validation will be necessary before broader generalization. Fifth, sleep metrics were available but were not incorporated into the present models. Given the complex and potentially bidirectional relationship between sleep and RA disease activity, as well as differences in sleep measurement across commercial wearable platforms, sleep features are being evaluated in a dedicated analysis. Future work should determine whether sleep metrics provide incremental predictive value beyond HR, HRV, activity, demographic features, and PROs. Additionally, the modest sample size and demographic composition of the cohort also limited evaluation of model fairness across subgroups, including sex and race/ethnicity. Larger, more diverse cohorts will be needed to assess whether predictive performance is consistent across demographic groups and to identify potential disparities before clinical implementation. Finally, data were collected from 3 commercial wearable platforms with different sensors, sampling approaches, proprietary processing, and available data streams. The cohort was also weighted toward Apple Watch users, limiting device-specific comparisons. Step counts and other activity features may be affected by wear time, gait limitations, device

placement, and proprietary algorithms. Wearable missingness may also be nonrandom, potentially biasing complete-case and imputed analyses if nonwear or signal quality varies with health status or participant behavior. In addition, requiring wearable use and excluding participants with conditions or medications affecting HR or HRV may limit generalizability. Future studies should evaluate device-specific performance, missingness patterns, sleep metrics, and model interpretability to better define which physiological and behavioral features drive prediction.

Conclusions

This study suggests that consumer wearable-derived physiological data analyzed with DL may support personalized forecasting of symptomatic and inflammatory RA flare states weeks in advance. Rather than replacing clinical assessment, laboratory testing, or clinician judgment, wearable-enabled forecasting may serve as an adjunctive monitoring strategy to identify periods of increased risk, prompt earlier objective evaluation, and support more proactive disease management. More broadly, these findings illustrate the potential for continuous real-world physiological data to shift chronic inflammatory disease care from episodic assessment toward dynamic, personalized monitoring. Larger, more diverse cohorts with standardized inflammatory assessments, external validation, improved interpretability, and prospective implementation studies are needed to determine whether this approach improves treatment decisions and patient outcomes.

Generative AI Use

The authors acknowledge the assistance of ChatGPT (version 5.5; OpenAI, 2026) for copyediting, language polishing, and grammar refinement of this manuscript. The authors have reviewed and edited all AI-generated suggestions for scientific accuracy and take full accountability for the final content.

Funding

Support for this study was provided by K23DK129835 (RPH). The funder had no role in study design, data collection, analysis, or interpretation of the data, preparation of the manuscript, or the decision to submit the manuscript for publication.

Data Availability

The datasets analyzed during the current study may be available from the corresponding author upon reasonable request. The code developed during the current study is proprietary and part of ongoing work and is therefore not publicly available but may be made available by the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: RPH
Data curation: PS, MD
Formal analysis: PS
Funding acquisition: RPH
Methodology: MS-F, RPH
Project administration: MD, JKW, DH
Software: MD, KL
Supervision: MS-F, PSG, RPH
Validation: MS-F, PSG
Visualization: PS
Writing – original draft: PS, RPH
Writing – review & editing: PS, MD, KL, JKW, DH, MS-F, PSG, RPH

Conflicts of Interest

RPH reports consulting from Bristol Meyers Squibb, Abbvie. Stock options from Salvo Health. Speaking for Sanofi. Research support from Janssen, Intralytix, Enlisense.

Multimedia Appendix 1

Participant characteristics, data availability, and deep-learning flare prediction.
[DOCX File (Microsoft Word File), 1440 KB-Multimedia Appendix 1]

References

1. Cai Y, Zhang J, Liang J, et al. The burden of rheumatoid arthritis: findings from the 2019 global burden of diseases study and forecasts for 2030 by Bayesian age-period-cohort analysis. J Clin Med. Feb 6, 2023;12(4):1291. [doi: 10.3390/jcm12041291] [Medline: 36835827]
2. Allaire S, Wolfe F, Niu J, Lavalley MP. Contemporary prevalence and incidence of work disability associated with rheumatoid arthritis in the US. Arthritis Rheum. Apr 15, 2008;59(4):474-480. [doi: 10.1002/art.23538] [Medline: 18383413]

3. van Mulligen E, Weel AEAM, Kuijper TM, Hazes JMW, van der Helm-van Mil AHM, de Jong PHP. The impact of a disease flare during tapering of DMARDs on the lives of rheumatoid arthritis patients. Semin Arthritis Rheum. Jun 2020;50(3):423-431. [doi: 10.1016/j.semarthrit.2020.02.011] [Medline: 32224045]
4. Di Matteo A, Bathon JM, Emery P. Rheumatoid arthritis. The Lancet. Nov 2023;402(10416):2019-2033. [doi: 10.1016/S0140-6736(23)01525-8] [Medline: 38240831]
5. Radu AF, Bungau SG. Management of rheumatoid arthritis: an overview. Cells. Oct 23, 2021;10(11):2857. [doi: 10.3390/cells10112857] [Medline: 34831081]
6. Smolen JS, Kerschbaumer A, Aletaha D, Robinson WH. Rheumatoid arthritis in adults: a review. JAMA. Oct 13, 2026;336(14):1257-1271. [doi: 10.1001/jama.2026.15455] [Medline: 42720931]
7. Fraenkel L, Bathon JM, England BR, et al. 2021 American College of Rheumatology guideline for the treatment of rheumatoid arthritis. Arthritis Rheumatol. Jul 2021;73(7):1108-1123. [doi: 10.1002/art.41752] [Medline: 34101376]
8. Perniola S, Alivernini S, Gremese E, et al. A multiparametric risk table for loss of clinical remission status in patients with rheumatoid arthritis: a STARTER study post-hoc analysis. Rheumatology (Oxford). Feb 1, 2025;64(2):526-532. [doi: 10.1093/rheumatology/keae094] [Medline: 38364299]
9. Sparks JA. Rheumatoid arthritis. Ann Intern Med. Jan 1, 2019;170(1):ITC1-ITC16. [doi: 10.7326/AITC201901010] [Medline: 30596879]
10. Orr CK, Najm A, Young F, et al. The utility and limitations of CRP, ESR and DAS28-CRP in appraising disease activity in rheumatoid arthritis. Front Med (Lausanne). 2018;5:185. [doi: 10.3389/fmed.2018.00185] [Medline: 30123796]
11. Kay J, Morgacheva O, Messing SP, et al. Clinical disease activity and acute phase reactant levels are discordant among patients with active rheumatoid arthritis: acute phase reactant levels contribute separately to predicting outcome at one year. Arthritis Res Ther. Feb 3, 2014;16(1):R40. [doi: 10.1186/ar4469] [Medline: 24485007]
12. Lee YC, Frits ML, Iannaccone CK, et al. Subgrouping of patients with rheumatoid arthritis based on pain, fatigue, inflammation, and psychosocial factors. Arthritis Rheumatol. Aug 2014;66(8):2006-2014. [doi: 10.1002/art.38682] [Medline: 24782222]
13. Georgopoulos V, Smith SL, Walsh DA, McWilliams DF. Central pain sensitivity is associated with changes in fatigue in RA: data from the CAP-RA study. Rheumatology (Oxford). Apr 6, 2026;65(4):keag130. [doi: 10.1093/rheumatology/keag130] [Medline: 41863329]
14. van Nies JAB, Krabben A, Schoones JW, Huizinga TWJ, Kloppenburg M, van der Helm-van Mil AHM. What is the evidence for the presence of a therapeutic window of opportunity in rheumatoid arthritis? A systematic literature review. Ann Rheum Dis. May 2014;73(5):861-870. [doi: 10.1136/annrheumdis-2012-203130] [Medline: 23572339]
15. Bozzalla-Cassione E, Grignaschi S, Xoxi B, et al. Insights into the concept of rheumatoid arthritis flare. Front Med (Lausanne). Mar 17, 2022;9:852220. [doi: 10.3389/fmed.2022.852220] [Medline: 35372374]
16. Bykerk VP, Shadick N, Frits M, et al. Flares in rheumatoid arthritis: frequency and management. A report from the BRASS registry. J Rheumatol. Feb 2014;41(2):227-234. [doi: 10.3899/jrheum.121521] [Medline: 24334643]
17. Myasoedova E, Crowson CS, Giblon RE, et al. Optimization of flare management in patients with rheumatoid arthritis: results of a randomized controlled trial. Clin Rheumatol. Nov 2019;38(11):3025-3032. [doi: 10.1007/s10067-019-04664-5] [Medline: 31264032]
18. Davergne T, Rakotozafiarison A, Servy H, Gossec L. Wearable activity trackers in the management of rheumatic diseases: where are we in 2020? Sensors (Basel). Aug 25, 2020;20(17):4797. [doi: 10.3390/s20174797] [Medline: 32854412]
19. Gossec L, Guyard F, Leroy D, et al. Detection of flares by decrease in physical activity, collected using wearable activity trackers in rheumatoid arthritis or axial spondyloarthritis: an application of machine learning analyses in rheumatology. Arthritis Care Res (Hoboken). Oct 2019;71(10):1336-1343. [doi: 10.1002/acr.23768] [Medline: 30242992]
20. Hashimoto T, Yoshiuchi K, Inada S, et al. Physical activity of elderly patients with rheumatoid arthritis and healthy individuals: an actigraphy study. BioPsychoSocial Med. Dec 2015;9(1). [doi: 10.1186/s13030-015-0046-0] [Medline: 26442128]
21. Creagh AP, Hamy V, Yuan H, et al. Digital health technologies and machine learning augment patient reported outcomes to remotely characterise rheumatoid arthritis. NPJ Digit Med. Feb 12, 2024;7(1):33. [doi: 10.1038/s41746-024-01013-y] [Medline: 38347090]
22. Sharma P, Danieletto M, Whang JK, et al. Wearable devices detect physiological changes that precede and are associated with symptomatic and inflammatory rheumatoid arthritis flares. Sci Rep. Nov 30, 2025;16(1). [doi: 10.1038/s41598-025-29748-y] [Medline: 41318620]
23. Hirten RP, Danieletto M, Landell K, et al. Development of the ehive digital health app: protocol for a centralized research platform. JMIR Res Protoc. Nov 16, 2023;12:e49204. [doi: 10.2196/49204] [Medline: 37971801]

24. Connor KM, Davidson JRT. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety. 2003;18(2):76-82. [doi: 10.1002/da.10113] [Medline: 12964174]
25. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav. Dec 1983;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
26. Pincus T, Swearingen C, Wolfe F. Toward a Multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis Rheum. Oct 1999;42(10):2220-2230. [doi: 10.1002/1529-0131(199910)42:10<2220::AID-ANR26>3.0.CO;2-5] [Medline: 10524697]
27. Pincus T, Yazici Y, Bergman MJ. RAPID3, an index to assess and monitor patients with rheumatoid arthritis, without formal joint counts: similar results to DAS28 and CDAI in clinical trials and clinical care. Rheum Dis Clin North Am. Nov 2009;35(4):773-778. [doi: 10.1016/j.rdc.2009.10.008] [Medline: 19962621]
28. van Riel PLCM, Renskers L. The Disease Activity Score (DAS) and the Disease Activity Score using 28 joint counts (DAS28) in the management of rheumatoid arthritis. Clin Exp Rheumatol. 2016;34(5 Suppl 101):S40-S44. [Medline: 27762189]
29. Kvrgic Z, Asiedu GB, Crowson CS, Ridgeway JL, Davis 3rd JM. "Like No One Is Listening to Me": a qualitative study of patient–provider discordance between global assessments of disease activity in rheumatoid arthritis. Arthritis Care Res (Hoboken). Oct 2018;70(10):1439-1447. [doi: 10.1002/acr.23501] [Medline: 29266857]
30. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Methodol. Jul 25, 2020;20(1):199. [doi: 10.1186/s12874-020-01080-1] [Medline: 32711455]
31. Konzett V, Kerschbaumer A, Smolen JS, et al. Definition of rheumatoid arthritis flare based on SDAI and CDAI. Ann Rheum Dis. Jan 11, 2024;83(2):169-176. [doi: 10.1136/ard-2023-224742] [Medline: 37890977]
32. Walker UA, Mueller RB, Jaeger VK, et al. Disease activity dynamics in rheumatoid arthritis: patients' self-assessment of disease activity via WebApp. Rheumatology (Oxford). Oct 1, 2017;56(10):1707-1712. [doi: 10.1093/rheumatology/kex229] [Medline: 28957553]
33. Ingegnoli F, Buoli M, Antonucci F, Coletto LA, Esposito CM, Caporali R. The link between autonomic nervous system and rheumatoid arthritis: from bench to bedside. Front Med (Lausanne). ;7:589079. [doi: 10.3389/fmed.2020.589079] [Medline: 33365319]
34. Hirten RP, Danieletto M, Sanchez-Mayor M, et al. Physiological data collected from wearable devices identify and predict inflammatory bowel disease flares. Gastroenterology. May 2025;168(5):939-951. [doi: 10.1053/j.gastro.2024.12.024] [Medline: 39826619]
35. Hirten RP, Danieletto M, Tomalin L, et al. Use of physiological data from a wearable device to identify SARS-CoV-2 infection and symptoms and predict COVID-19 diagnosis: observational study. J Med Internet Res. Feb 22, 2021;23(2):e26107. [doi: 10.2196/26107] [Medline: 33529156]
36. Chua JR, Jamal S, Riad M, et al. Disease burden in osteoarthritis is similar to that of rheumatoid arthritis at initial rheumatology visit and significantly greater six months later. Arthritis Rheumatol. Aug 2019;71(8):1276-1284. [doi: 10.1002/art.40869] [Medline: 30891933]
37. Foster ZJ, Day AL, Miller J. Polyarticular joint pain in adults: evaluation and differential diagnosis. Am Fam Physician. Jan 2023;107(1):42-51. [Medline: 36689970]

Abbreviations

AUPRC: area under precision-recall curve
CDAI: Clinical Disease Activity Index
CDRISC-10: 10-item Connor-Davidson Resilience Scale
CRP: C-reactive protein
DAS28: Disease Activity Score 28
DL: deep learning
EHR: electronic health record
EMM: expected marginal mean
HR: heart rate
HRV: heart rate variability
LSTM: long short-term memory
MDHAQ: Multidimensional Health Assessment Questionnaire
ML: machine learning
PE: positional encoding
PRO: patient-reported outcome
PSS4: Perceived Stress Scale 4

RA: rheumatoid arthritis
RAPID3: Routine Assessment of Patient Index Data 3
RHR: resting heart rate
RMSSD: root-mean-square of successive RR interval differences
ROCAUC: receiver operating characteristic area under curve
SDAI: Simple Disease Activity Index
SHAP: Shapley additive explanations

Edited by Ivan Steenstra; peer-reviewed by Carlos Alberto Sanches, Kenneth Baker; submitted 10.Mar.2026; initial comments to author 02.Jun.2026; final revised version received 29.Sep.2026; accepted 30.Sep.2026; published 25.Nov.2026

Please cite as:
Sharma P, Danieletto M, Whang JK, Landell K, Helmus D, Suárez-Fariñas MSuáRFariñA, Gulko PS, Hirten RP
Predicting Rheumatoid Arthritis Flares Using Wearable Device Data: Deep Learning Study
J Med Internet Res 2026;28:e95089
URL: https://www.jmir.org/2026/1/e95089
doi: 10.2196/95089

© Pragya Sharma, Matteo Danieletto, Jessica K Whang, Kyle Landell, Drew Helmus, Mayte Suárez-Fariñas, Percio S Gulko, Robert P Hirten. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 25.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.