

Classifying Self-reported Rheumatoid Arthritis Flares Using Daily Patient-generated Data from a Smartphone App: An Exploratory Analysis Applying Machine Learning Approaches

Julie Gandrup, David A Selby, William G Dixon

Submitted to: JMIR Formative Research
on: July 09, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4

Supplementary Files..... 21

 Figures 22

 Figure 1..... 23

 Figure 2..... 24

 Multimedia Appendixes 25

 Multimedia Appendix 1..... 26

Classifying Self-reported Rheumatoid Arthritis Flares Using Daily Patient-generated Data from a Smartphone App: An Exploratory Analysis Applying Machine Learning Approaches

Julie Gandrup¹ MD, PhD; David A Selby^{1, 2}; William G Dixon^{1, 3}

¹Centre for Epidemiology Versus Arthritis University of Manchester Manchester GB

²Department of Computer Science Technische Universität Kaiserslautern Kaiserslautern DE

³Department of Rheumatology Northern Care Alliance NHS Foundation Trust Salford GB

Corresponding Author:

Julie Gandrup MD, PhD

Centre for Epidemiology Versus Arthritis

University of Manchester

Oxford Rd

Manchester

GB

Abstract

Background: The ability to predict rheumatoid arthritis (RA) flares between clinic visits based on real-time, longitudinal patient-generated data could potentially allow for timely interventions to avoid disease worsening.

Objective: In this exploratory study, we aimed to investigate the feasibility of using machine learning methods to classify self-reported RA flares based on a small dataset of daily symptom data collected on a smartphone app.

Methods: We used daily symptoms and weekly flares reported on the REMORA smartphone app from 20 RA patients over three months. Predictors were several summary features of the daily symptom scores collected in the week leading up to the flare question. We fitted three binary classifiers: logistic regression +/- elastic net regularization, a random forest and naïve Bayes. Performance was evaluated according to the area under the curve (AUC) of the receiver operating characteristic curve.

Results: The data comprised an average of 60.6 daily reports and 10.5 weekly reports per participant. AUCs were broadly similar between models, but logistic regression with elastic net had the highest AUC of 0.82. At a cut-off requiring specificity to be 0.80, the corresponding sensitivity to detect flares was 0.60 for this model. The positive predictive value in this population was 53%, and negative predictive value 85%.

Conclusions: Predicting self-reported flares based on daily symptom scorings in the preceding week using machine learning methods was feasible. The observed predictive accuracy might improve as we obtain more data, and these exploratory results should be validated in an external cohort. In the future, analysis of frequently collected patient-generated data may allow us to predict flares before they unfold, opening opportunities for just-in-time interventions.

(JMIR Preprints 09/07/2023:50679)

DOI: <https://doi.org/10.2196/preprints.50679>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a peer-reviewed journal, I will be able to remove the preprint label and make the full manuscript available to all users.

Original Manuscript

Classifying Self-reported Rheumatoid Arthritis Flares Using Daily Patient-generated Data from a Smartphone App: An Exploratory Analysis Applying Machine Learning Approaches

Julie Gandrup¹, David A Selby^{1,2}, William G Dixon^{1,3}

¹Centre for Epidemiology Versus Arthritis, University of Manchester, Manchester, UK

²Department of Computer Science, Technische Universität Kaiserslautern, Kaiserslautern, Germany

³Department of Rheumatology, Northern Care Alliance NHS Foundation Trust, Salford, UK

Corresponding author

Dr Julie Gandrup

Centre for Epidemiology Versus Arthritis

Division of Musculoskeletal and Dermatological Sciences, School of Biological Sciences

The University of Manchester

2nd floor, Stopford Building

Oxford Road

Manchester M13 9PT, UK

E-mail: Julie.gandrup@gmail.com

Telephone: 0741 501 1601

<https://orcid.org/0000-0002-4109-4889>

ABSTRACT

Background: The ability to predict rheumatoid arthritis (RA) flares between clinic visits based on real-time, longitudinal patient-generated data could potentially allow for timely interventions to avoid disease worsening.

Objective: In this exploratory study, we aimed to investigate the feasibility of using machine learning methods to classify self-reported RA flares based on a small dataset of daily symptom data collected on a smartphone app.

Methods: We used daily symptoms and weekly flares reported on the REMORA smartphone app from 20 RA patients over three months. Predictors were several summary features of the daily symptom scores (e.g. pain and fatigue) collected in the week leading up to the flare question. We fitted three binary classifiers: logistic regression with and without elastic net regularization, a random forest and naïve Bayes. Performance was evaluated according to the area under the curve (AUC) of the receiver operating characteristic curve. For the best performing model, we considered sensitivity and specificity for different thresholds in order to illustrate different ways in which the predictive model could behave in a clinical setting.

Results: The data comprised an average of 60.6 daily reports and 10.5 weekly reports per participant. Participants reported a median of two flares (IQR 0.75-4.25) each over an average follow-up time of 81 days (IQR 79–82). AUCs were broadly similar between models, but logistic regression with elastic net regularization had the highest AUC of 0.82. At a cut-off requiring specificity to be 0.80, the corresponding sensitivity to detect flares was 0.60 for this model. The positive predictive value (PPV) in this population was 53%, and negative predictive value (NPV) 85%. Given the prevalence of flares, the best PPV achieved meant only around two of every three positive predictions were correct (PPV 0.65). By prioritising a higher NPV, the model correctly predicted over nine in every ten non-flare weeks, but the accuracy of predicted flares fell to only one in two being correct (NPV and PPV 0.92 and 0.51, respectively).

Conclusion: Predicting self-reported flares based on daily symptom scorings in the preceding week using machine learning methods was feasible. The observed predictive accuracy might improve as we obtain more data, and these exploratory results need to be validated in an external cohort. In the

future, analysis of frequently collected patient-generated data may allow us to predict flares before they unfold, opening opportunities for just-in-time interventions. Depending on the nature and implication of an intervention, different cut-off values for an intervention decision need to be considered as well as the level of predictive certainty required.

KEYWORDS

Rheumatoid arthritis, flare, patient-generated health data, smartphone, mHealth, machine learning

INTRODUCTION

Rheumatoid arthritis (RA) is characterized by fluctuations in disease severity over time, with periods of worsening referred to as “flares”. Flares represent a significant burden on patients, including uncontrollable symptoms and compromised ability to perform everyday tasks [1], and are associated with negative outcomes such as loss of functional ability and structural damage [2,3]. To minimise the impact of significant flares on the patient, it is important that a flare is identified early, so necessary interventions can be initiated. However, changes in disease severity often occur between scheduled visits to a clinician (usually every 6–12 months) which might hamper optimal disease management. In the early stages of a flare, patients self-manage, then progress to seeking medical help when they feel they are losing control [4]. Understanding when a flare is happening—or about to happen—could remove some of the barriers to seeking help.

Patient-generated health data, including patient-reported symptoms, could play an increasingly important role in clinical decision-making [5]. Smartphones, tablets and wearable devices can facilitate collection of self-reported symptom data between scheduled clinical appointments and at a much higher frequency, e.g. daily or weekly. This would allow us to “listen in” on the short-term patterns of RA disease severity and identify flares earlier or even predict flares before they unfold. The ability to identify or predict flares between clinical appointments based on patient-generated data would potentially allow for timely interventions. These might include self-management advice, medication adjustment, triggering a remote consultation or bringing forward a planned visit. Just-in-time adaptive interventions are an emerging area of research which, until now, have primarily been deployed in mental health and behaviour-change treatments [6,7]. Before using predictive algorithms, however, it is important to understand how well the prediction performs and whether such performance would be acceptable in a clinical setting.

Due to the potentially high-dimensional and non-linear nature of intensively-collected patient-generated data, modern machine learning methods could offer benefits over traditional tools, such as logistic regression, for accurate prediction. Machine learning is increasingly being employed in rheumatology: see for example Hügler et al. (2020) [8]. However, the literature on predicting distant outcomes such as flares through longitudinal patient-generated health data is still in its infancy and currently limited by heterogeneity in predictors, flare definitions, frequency of data collection and classification methods [9,10].

In previous work, we investigated the association between patient-reported flares and daily symptom scores [11]. The purpose of the current analysis was to build on this work by investigating the feasibility of using machine learning methods to classify self-reported RA flares based on a small dataset of daily symptom data collected through a smartphone app. Specifically, the objectives of this exploratory study were 1) to fit three binary classifiers and consider their performance, 2) to illustrate the initial implications of different cutoff values for predicting a flare, and finally 3) frame an agenda for future work supporting ways to meaningfully leverage digital patient-generated health data to predict flares and improve patient outcomes.

METHODS

Data

This study was a post-hoc analysis of data from the first phase of the Remote Monitoring of Rheumatoid Arthritis (REMORA) study [12], which involved 20 RA patients using a smartphone application to track their daily symptoms over three months.

Participants received prompts every evening to report several symptoms on a 0–10 numerical rating scale based on the RAID scale adapted for daily use [13]: pain, function ('difficulty in doing daily activities'), fatigue (attributed to RA), sleep quality, overall physical and emotional wellbeing and 'ability to cope'. Users reported the duration of morning stiffness daily using one of seven time intervals. Weekly questionnaires asked patients about self-assessed tender and swollen joint counts and the binary flare question: 'Have you experienced a flare in the last week?'. These questions were prompted by a notification every seven days to complete the weekly question set. Eligibility criteria were: 1) clinician-verified RA, 2) treated at a specific outpatient clinic, 3) willingness to participate and 4) able to provide written consent. The app and its content was co-designed with patients, clinicians and researchers. For further details of the REMORA study, see

Austin et al. (2019) [12].

Definition of outcomes and explanatory variables

We treated each weekly flare report as a binary outcome. It was left up to the patient answering the question to decide what classified as a flare. Weeks with an unanswered (missing) flare question were not included in this analysis.

To fit a binary classification model, it is necessary to extract a 'feature vector' or list of predictors from the sequence of daily symptom data that are mapped to each weekly flare report. The seven days up to and including each flare report were treated as the exposure period. For each exposure period, we calculated the following five symptom summary features for each of the eight daily symptoms: Minimum, maximum, mean score, standard deviation (SD) and slope. Isolated daily reports (those not followed by a flare report in the next 6 days) were discarded, so every remaining exposure period contained at least two daily data points. Although not prompted, participants were able to answer the weekly flare question at any time during the week outside of the 7-day schedule, resulting in some partially overlapping exposure periods. In that case, we allowed the intersecting daily symptom reports to correspond to multiple outcomes. Where the same participant responded more than once on the same date, we assumed later-recorded responses superseded earlier ones.

The patient-reported symptom scores were collected using integer numerical rating scales from 0-10 (morning stiffness on a 7-point ordinal scale). For this exploratory analysis, all symptoms were treated as continuous variables. This approach was chosen because it allows ease of comparison with other work in intraindividual pain variability[14]. Additionally, distributions of pain scores, both during 'flare' and 'no-flare' weeks, was not noticeably skewed, suggesting minimal influence of ceiling or floor effects. One alternative to our approach would be incorporating monotonic constraints in the ML model, however there were no readily available packages for monotonic random forests in either Python or R at the time of writing. However, imposition of monotonic constraints, while improving interpretability of the model, is as likely as not to decrease predictive performance.

Statistical analysis

Machine learning classification concerns the task of recognizing objects and being able to separate them into categories. With our analysis, we aimed to classify each week as either a flare week or non-flare week based on the symptom summary features. While the most popular binary classification models are simpler ones, like logistic regression, the seemingly high-dimensional and nonlinear nature of disease activity motivates more complex 'black box' machine learning approaches, including random forest classifiers. We fitted three distinct classes of binary

classification models to the data: logistic regression with and without elastic net regularization, a random forest and naïve Bayes. Random forest models use decision trees as building blocks. Decision trees use features to divide the observations into subgroups (or classes) that are as different from each other as possible. Many decision trees operate as an ensemble and the class selected by most trees will become the final output. Under a Naïve Bayes classifier, continuous predictors may be assumed to follow independent univariate normal distributions with a separate mean and variance estimated for each class. Given a feature vector (list of predictor values), the predicted class is then inferred probabilistically via Bayes rule [15]. These three methods were chosen based on previous work that aimed to predict flares in RA [16,17] and in addition they have the benefit of being straightforward to fit on a generic binary classification problem as presented here.

When evaluating the performance of classifiers, a training dataset is required for fitting them and another distinct dataset is needed for the subsequent evaluation and test of those classifiers. We trained our models using the R package *mlr3* [18]. Fine-tuning of the models beyond the default settings was not performed for this exploratory study and no imputation was performed on missing values. 10-fold cross-validation was performed, with 18 (90%) participants comprising the training sets and the remaining two (10%) the test sets. The validation was repeated ten times, each time reserving two different participants for testing. In the case of longitudinal data collected from individuals, the training-test data splits should fall between participants, so that data associated with a particular patient fall entirely in a training set or a test set, so testing and training is not performed within the same patient timeline. In other words, the models were tested on different patients to those on which they were trained [16]. We then evaluated the performance of each of the models against patient-reported flares as the gold standard according to the area under the curve (AUC) of the receiver operating characteristics (ROC) curve. The model with the highest AUC in the test dataset was considered as the best final model.

We considered sensitivity and specificity for ten different thresholds in order to illustrate different ways in which the predictive model could behave in a clinical setting. Sensitivity is the proportion of those with a flare who have a positive prediction, while the specificity is the proportion of those without a flare that have correctly been predicted to have no flare. We did this by setting the sensitivity from 0.5 up to 0.9 in 0.1 unit increments, then doing the same for specificity (see Table 1). Corresponding positive predictive values, i.e. the probability that those with a predicted flare indeed go on to have a flare, and negative predictive values, i.e. the probability that those with a predicted non-flare indeed do not experience a flare, were also considered for

these different thresholds to illustrate their potential impact and clinical utility.

Ethical considerations

The original study was reviewed and approved by the Greater Manchester Central Research Ethics Committee (ref. 15/NW/0172). All participants completed informed consent forms which included consent for secondary use of de-identified data for research purposes. Participants did not receive any financial compensation for their participation.

RESULTS

The collected dataset comprised 20 unique participants completing a total of 1325 daily and 213 weekly questionnaires. Each participant reported an average of 61 daily reports and 11 weekly reports over an average follow-up time of 81 days (IQR 79–82). Of the participants, 60% were female, all except one were of white British ethnicity, and mean age was 57 ± 11 years. Patterns of daily and weekly responses for each app user are shown in Figure 1.

Participants reported a median of two flares (IQR 0.75-4.25) each over the course of the study resulting in 57 flares in total. The largest number of flares reported by a single participant was nine, while five participants reported no flares at all.

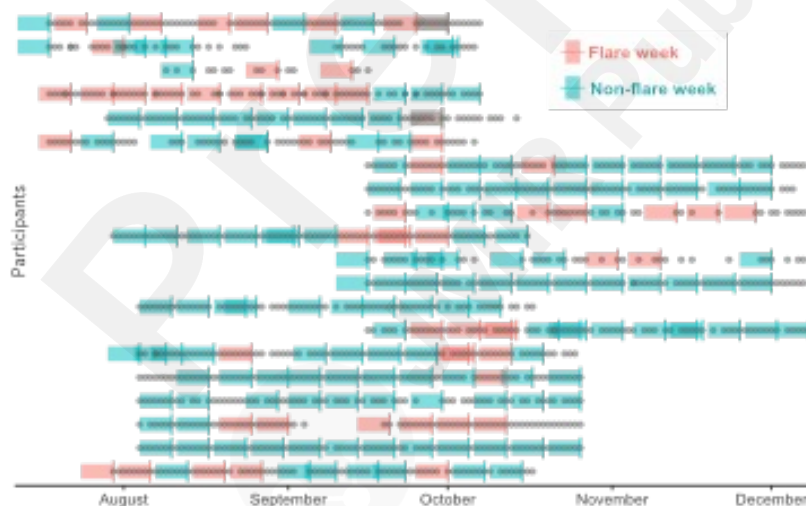


Figure 1 Patterns of daily and weekly data entry. Each row is a different participant. Vertical lines denote weekly responses, point daily responses. The shaded bands represent the week preceding each weekly response and the two colors denote whether patients reported a flare or no flare in that week.

Classifier performances are visualized in Figure 2. AUCs were broadly similar for all models, but the model with the highest AUC was the logistic regression with elastic net regularization with an AUC of 0.82. This

was followed by naïve Bayes and random forest with AUCs of 0.77 and 0.75, respectively. Un-regularized logistic regression, as expected, had the lowest AUC of 0.71. Supplementary Figure S1 shows precision-recall curves, and Figure S2 shows the average relative importance of each predictor for the random forest model and logistic regression (data not shown).

Figure 2 Classifier performance for each of the four models.

Table 1 shows sensitivity, specificity, and positive and negative predictive values for a range of different thresholds for the model with the highest AUC. At a cutoff requiring specificity to be 0.80, the corresponding sensitivity to detect flares was 0.60 for the regularized logistic regression model, meaning that the prediction model correctly identified three in every five self-reported flares, and correctly identified four in every five non-flares. At this cutoff, and given the prevalence of flares within our dataset, the positive predictive value was 0.53 and the negative predictive value was 0.85, meaning there was (only) a 53% chance that the patient actually had a flare after the algorithm predicted a flare, but an 85% chance the patient did not have a flare, if the algorithm predicted a non-flare.

For that same model, we also considered a different threshold that favored identifying true positives, i.e. ability to correctly identify those reporting a flare. At a cutoff requiring *sensitivity* to be 0.80, the corresponding specificity was 0.72. The positive predictive value was 0.51 and negative predictive value 0.90 for this threshold. Of all the sensitivity and specificity options, ranging from 0.5-0.9, the greatest positive predictive value was 0.65 (with an associated negative predictive value of 0.83) and the highest negative predictive value was 0.92 (where the best corresponding positive predictive value was 0.51).

Table 1 Sensitivity, specificity, positive and negative predictive values, and implications at different cut-offs. Shown for logistic regression with elastic net regularization.

	Psychometric properties			
	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Cut-off 1	0.50	0.90	0.65	0.83
Cut-off 2	0.60	0.80	0.53	0.85
Cut-off 3	0.70	0.74	0.49	0.87
Cut-off 4	0.80	0.72	0.51	0.90
Cut-off 5	0.90	0.43	0.37	0.92
Cut-off 6	0.88	0.50	0.39	0.92
Cut-off 7	0.87	0.60	0.44	0.92
Cut-off 8	0.83	0.70	0.51	0.92
Cut-off 9	0.60	0.80	0.53	0.85

Cut-off	0.50	0.90	0.65	0.83
10				

DISCUSSION

With this exploratory study, we showed that it is feasible to use robust machine learning methods to classify patient-reported flares based on daily symptom scorings in the preceding week with decent accuracy. Of the three classifiers fitted, logistic regression with elastic net regularization had the highest overall AUC of 0.82, but across the different models AUCs were broadly similar. Random forest classifiers tend to overfit especially for high-dimensional data. Therefore, any advantage of random forest for our dataset is most likely due to overfitting on the small number of observations, which might explain the seemingly better performance of the regularized logistic regression model. For the model with the highest AUC, at a cut point requiring specificity to be 0.80, sensitivity to detect flare was 0.60 resulting in accurate prediction of three out of five flares from the prior week's daily symptom data. Given the prevalence of flares in this cohort, the best PPV we could achieve meant only around two of every three positive predictions were correct (PPV 0.65). If we instead prioritised a higher NPV, we were able to correctly predict over nine in every ten non-flare weeks, although this meant the accuracy of predicted flare weeks fell to only one in two being correct (NPV and PPV 0.92 and 0.51, respectively). In the future, it will be necessary to find the optimal balance between identifying true flares (or, in other words, not missing flares) without overburdening the health service by identifying flares incorrectly. Models were fitted to a relatively small dataset of 20 highly selected RA patients with three months of daily symptoms, so interpretations should be cautious. Nonetheless, our study serves as an early indicative example of how classification of flares based on daily patient-generated data is ambitious but feasible.

Other examples of predicting RA flares using longitudinal patient-generated health data (in contrast to using longitudinal clinical routine data) are sparse. Haynes et al. attempted to classify weekly-reported flares from a combination of daily RA symptom scorings and weekly flare questionnaires collected on a smartphone. Similar to our results, their best performing logistic regression classification model had an AUC of 0.81 and, at a cutoff requiring specificity to be ≥ 0.80 , sensitivity to detect flare was 0.62 [9]. As an alternative to patients actively entering the data, Gossec et al. predicted weekly patient-reported flares based on passively-collected step counts from fitness trackers in 155 patients with RA and axial spondyloarthritis. Using a naïve Bayes classification model, they found that patient-reported flares were strongly associated to physical activity and proposed that processing of patient-level physical activity data using machine learning can be used to

accurately detect flares [10]. Similarly, Rao et al demonstrated the ability of physical activity tracker data to classify health status over time (not specifically flares) in patients with rheumatoid arthritis [19].]. Creagh et al. observed that augmenting standard patient-reported outcomes with objective sensor-based data improved the estimation of RA severity levels [20]. Combined, these results raise the possibility for passive surveillance that might, in the future, lead to just-in-time interventions without the need for continuous active symptom tracking.

Limitations

The methodology of our study has several limitations. First, as already mentioned, the dataset is limited in size, which makes interpretation of results more challenging and additionally limits the possibility of meaningful interpretation of the importance of different predictors for classifying a flare. Additionally, most patients were of white British ethnicity, which limits the generalisability of our results to other populations. Second, laboratory data, such as CRP or ESR, or traditional disease activity measures, were not available. This means we cannot correlate patient-reported data with clinician-reported information and limits our understanding of the generalisability of the population. (That said, it would not be plausible to have a clinician assessment whenever a patient experiences a flare). We fitted several different models, but the lack of an external validation dataset also limits the generalisability of our results. There is a need to externally validate our findings in a larger, more diverse dataset. Third, for modelling purposes, we treated the original ordinal features as continuous. This preserves the information in the ordering but requires the assumption that the numerical distance between each category is approximately equal. We assumed that this was reasonable for our analysis, but other more complex methods could be utilized to account for ordinal data [21]. Fourth, the feature vectors also do not account for temporal dependence (or autocorrelation) within or between patient-weeks, i.e. the fact that pain today may depend on pain yesterday, or that likelihood of reporting a flare this week is affected by reports in previous weeks. Fifth, isolated daily scores — those not within seven days of a subsequent flare report—were discarded. However, in a different analysis approach, these could be treated as censored observations. We included variables from the week prior to the self-reported flare and not any data from preceding weeks. We therefore did not assess how far in advance it was possible to predict a flare: we would want better performance from the more proximal data before extending the time window further. Sixth, the period of symptom tracking was limited to three months. While we know patients can sustain symptom tracking for up to six months [22], we do not yet know how much longer they would continue, nor whether the predictive algorithm is stable over time. Finally, our definition of flare was a non-validated, pragmatic, patient-centred one, which left it to the patient to decide when a flare occurred, and

therefore it could be interpreted differently by different patients. Multiple definitions of RA flares have been suggested [23,24], but to date no reference standard has been agreed. This might consequently make it harder to predict a 'flare' if each patient's interpretation of a flare is different. There is a need to develop a validated, accepted and easy to use (digital) flare definition in RA which can be used prospectively.

While sensitivity and specificity of a test are stable, positive and negative predictive values are influenced by the prevalence of the disease in the population. When prevalence decreases, the positive predictive value decreases too. In contrast, the negative predictive value will increase. The prevalence of patient-reported flares in our cohort therefore influences predictive values, and its broader usability is dependent on our cohort's representativeness of the broader RA population.

Clinical implications

Our results point to a future where real-time analysis of frequently collected patient-generated data from symptom tracking may allow us to predict imminent flares before they unfold. This in turn opens opportunities for just-in-time adaptive interventions (JITAI). JITAI "leverage mobile technology to deliver the right type of support, at the right time based on ongoing information about the individual's internal state and context" [25]. Until now, they have primarily been deployed in supporting health behaviour change [7,26], but they hold enormous potential for fluctuating diseases like RA where timely intervention for increase in disease activity is beneficial. Depending on the nature and implication of a JITAI, different cut-off values for an intervention decision need to be considered. Because of cost and other implications, different interventions will require different levels of predictive certainty before an action is triggered. In RA, we could imagine, say, two different scenarios in response to a predicted flare: One where a self-management advice is newly offered or promoted via a notification within the app, and a second where a scheduled clinical consultation is brought forward based on the data entered by the patient. Striking the right balance between missing true flares and flagging up false positives is crucial. We might tolerate to serve up an automated written self-management advice for more false positives, because the implications are relatively few. This would also mean we rarely miss the opportunity of providing the advice to someone with a true flare that the predictive test has failed to identify. On the other hand, we need more caution when offering a clinical consultation. Here, tolerance for false positives should be low because of the high implications – scheduling an expensive consultation in an already busy and overworked clinic because a flare is predicted, but where that consultation is wasted as there is no true flare. In this instance, a high positive predictive value of the algorithm is essential. If we apply these considerations to our results, we could foresee that self-management could be usefully

delivered in response to the predicted flares. Whether incorrectly promoting self-management advice to one in every two people who might have a flare (PPV 0.51, NPV 0.92) would need formal evaluation to see if this is indeed acceptable. Conversely, given the current model performance and prevalence of flares, we would be unlikely to use the predictive model to trigger a time- and resource-intensive clinical intervention because, at best, only two in three of these predicted flares would be correct (PPV 0.65, NPV 0.83). In addition to careful assessment of quality, cost-benefit analyses of various interventions depending on model performance will be a key piece to the assessment of predictive models in clinical practice. For example, evaluating whether the costs of additional telephone consultations and/or clinic visits in response to automated flare predictions are outweighed by the benefits of earlier interventions including potential shorter recovery, earlier return to work or quality of life, and better long-term outcomes.

Conclusions

Classifying self-reported flares based on daily symptom scorings in the preceding week using machine learning methods was feasible in this exploratory study, with regularized logistic regression seeming to outperform the other machine learning methods in this small dataset. It is possible that the observed predictive accuracy will improve as we obtain more data and external validation in larger datasets is an important next step. As we begin to understand how we can use regular symptom tracking data to predict imminent flares in RA before they unfold, we in turn open opportunities for just-in-time adaptive interventions. This is now a tangible future, but more data and more research are needed to realise the goal of using machine learning to offer a personalised care approach that ultimately improves patient outcomes.

AVAILABILITY OF DATA AND MATERIALS

The data that support the findings of this study are available on request from the corresponding author (JG). The data are not publicly available due to them containing information that could compromise research participant consent.

ACKNOWLEDGEMENTS

Not applicable.

CONFLICTS OF INTEREST

JG is currently employed at UCB, but this work was completed before joining UCB. WGD has received consultancy fees from Google, unrelated to this work. DAS declares no conflicts of interest.

FUNDING

This work was supported by the Centre for Epidemiology Versus Arthritis (grant number 21755).

AUTHORS' CONTRIBUTIONS

All authors contributed to the conception of the idea for the study and research questions. DAS and JG performed the analyses and drafted the manuscript. All authors were involved in interpreting the findings and revising the manuscript critically. All authors approved the final version to be submitted for publication. All authors had full access to all the data in the study.

ADDITIONAL FILES

“Additional File 1” is of format .docx and contains “Supplementary Figure S1”, which depicts the importance of different predictors in the respective models. “Supplementary Figure S2” depicts importance of different predictors in the best model. Data is not shown in the main text.

REFERENCES

1. Hewlett S, Sanderson T, May J, Alten R, Bingham CO III, Cross M, March L, Pohl C, Woodworth T, Bartlett SJ. ‘I’m hurting, I want to kill myself’: rheumatoid arthritis flare is more than a high joint count—an international patient perspective on flare where medical help is sought. *Rheumatology* 2012 Jan 1;51(1):69–76. doi: 10.1093/rheumatology/keq455
2. Markuse IM, Dirven L, Gerards AH, van Groenendael JHLM, Roday HK, Kerstens PJSM, Lems WF, Huizinga TWJ, Allaart CF. Disease flares in rheumatoid arthritis are associated with joint damage progression and disability: 10-year results from the BeSt study. *Arthritis Res Ther* 2015;17(1). PMID:26321751
3. Kuettel D, Primdahl J, Christensen R, Ørnbjerg L, Hørslev-Petersen K. Impact of patient-reported flares on radiographic progression and functional impairment in patients with rheumatoid arthritis: a cohort study based on the AMBRA trial. *Scandinavian Journal of Rheumatology* 2018 Mar 4;47(2):87–94. PMID:28841103
4. Flurey CA, Morris M, Richards P, Hughes R, Hewlett S. It’s like a juggling act: rheumatoid

arthritis patient perspectives on daily life and flare while on current treatment regimes. *Rheumatology* 2014 Apr 1;53(4):696–703. doi: 10.1093/rheumatology/ket416

5. Fautrel B, Alten R, Kirkham B, de la Torre I, Durand F, Barry J, Holzkaemper T, Fakhouri W, Taylor PC. Call for action: how to improve use of patient-reported outcomes to guide clinical decision making in rheumatoid arthritis. *Rheumatol Int* 2018;38(6):935–947. PMID:29564549
6. Balaskas A, Schueller SM, Cox AL, Doherty G. Ecological momentary interventions for mental health: A scoping review. *PLoS One* 2021 Mar 11;16(3):e0248152. PMID:33705457
7. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAs) to promote physical activity. *Int J Behav Nutr Phys Act* 2019 Apr 3;16:31. PMID:30943983
8. Hügler M, Omoumi P, van Laar JM, Boedecker J, Hügler T. Applied machine learning and artificial intelligence in rheumatology. *Rheumatology Advances in Practice* 2020 Jan 1;4(1):rkaa005. doi: 10.1093/rap/rkaa005
9. Haynes VS, Curtis J, Xie F, Zhao H, Kannowski CL, Poon JL, Gavigan K, Curtis D, Nolot SK, Nowell WB. Using self-reported outcomes to detect new-onset flare in a real-world study of participants with rheumatoid arthritis - Interim results from the digital tracking of arthritis longitudinally (DIGITAL) study. *Ann Rheum Dis* 2020;79:576.
10. Gossec L, Guyard F, Leroy D, Lafargue T, Seiler M, Jacquemin C, Molto A, Sellam J, Foltz V, Gandjbakhch F, Hudry C, Mitrovic S, Fautrel B, Servy H. 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)* 2019;71(10):1336–1343. PMID:30242992
11. Gandrup J, Selby D, van der Veer SN, McBeth J, Dixon WG. Using patient-reported data from a smartphone app to capture and characterise real-time patient-reported flares in rheumatoid arthritis. *Rheumatology Advances in Practice* 2022 Mar 16;rkac021. doi: 10.1093/rap/rkac021
12. Austin L, Sharp CA, van der Veer SN, Machin M, Humphreys J, Mellor P, McCarthy J, Ainsworth J, Sanders C, Dixon WG. Providing 'the bigger picture': benefits and feasibility of integrating remote monitoring from smartphones into the electronic health record: Findings from the Remote Monitoring of Rheumatoid Arthritis (REMORA) study. *Rheumatology* 2020 Feb 1;59(2):367–378. doi: 10.1093/rheumatology/kez207
13. Gossec L, Paternotte S, Aanerud GJ, Balanescu A, Boumpas DT, Carmona L, de Wit M, Dijkmans B a. C, Dougados M, Englbrecht M, Gogus F, Heiberg T, Hernandez C, Kirwan JR, Mola EM, Cerinic MM, Otsa K, Schett G, Scholte-Voshaar M, Sokka T, von Krause G, Wells GA, Kvien TK. Finalisation and validation of the rheumatoid arthritis impact of disease score, a patient-derived composite measure of impact of rheumatoid arthritis: a EULAR initiative. *Ann Rheum Dis* 2011 Jun;70(6):935–942. PMID:21540201
14. Mun CJ, Suk HW, Davis MC, Karoly P, Finan P, Tennen H, Jensen MP. Investigating intraindividual pain variability: methods, applications, issues, and directions. *Pain* 2019 Nov;160(11):2415–2429. PMID:31145212

15. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition. Springer; 2017.
16. Vodenčarević A, Goes MC van der, Medina OAG, Groot MCH de, Haitjema S, Solinge WW van, Hoefer IE, Peelen LM, Laar JM van, Zimmermann-Rittereiser M, Hamans BC, Welsing PMJ. Predicting Flare Probability in Rheumatoid Arthritis using Machine Learning Methods: Proceedings of the 7th International Conference on Data Science, Technology and Applications Porto, Portugal: SCITEPRESS - Science and Technology Publications; 2018. p. 187–192. doi: 10.5220/0006930501870192
17. Vodencarevic A, Tascilar K, Hartmann F, Reiser M, Hueber AJ, Haschka J, Bayat S, Meinderink T, Knitza J, Mendez L, Hagen M, Krönke G, Rech J, Manger B, Kleyer A, Zimmermann-Rittereiser M, Schett G, Simon D, on behalf of the RETRO study group. Advanced machine learning for predicting individual risk of flares in rheumatoid arthritis patients tapering biologic drugs. *Arthritis Research & Therapy* 2021 Feb 27;23(1):67. doi: 10.1186/s13075-021-02439-5
18. Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, Au Q, Casalicchio G, Kotthoff L, Bischl B. mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software* 2019 Dec 11;4(44):1903. doi: 10.21105/joss.01903
19. Rao K, Speier W, Meng Y, Wang J, Ramesh N, Xie F, Su Y, Nowell WB, Curtis JR, Arnold C. Machine Learning Approaches to Classify Self-Reported Rheumatoid Arthritis Health Scores Using Activity Tracker Data: Longitudinal Observational Study. *JMIR Formative Research* 2023 Jun 26;7(1):e43107. doi: 10.2196/43107
20. Creagh AP, Hamy V, Yuan H, Mertes G, Tomlinson R, Chen W-H, Williams R, Llop C, Yee C, Duh MS, Doherty A, Garcia-Gancedo L, Clifton DA. Digital health technologies and machine learning augment patient reported outcomes to remotely characterise rheumatoid arthritis. *npj Digit Med* Nature Publishing Group; 2024 Feb 12;7(1):1–12. doi: 10.1038/s41746-024-01013-y
21. Bürkner P-C, Charpentier E. Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology* 2020;73(3):420–451. doi: 10.1111/bmsp.12195
22. Dixon WG, Beukenhorst AL, Yimer BB, Cook L, Gasparrini A, El-Hay T, Hellman B, James B, Vicedo-Cabrera AM, Maclure M, Silva R, Ainsworth J, Pisaniello HL, House T, Lunt M, Gamble C, Sanders C, Schultz DM, Sergeant JC, McBeth J. How the weather affects the pain of citizen scientists using a smartphone app. *npj Digit Med* 2019 Oct 24;2(1):1–9. doi: 10.1038/s41746-019-0180-3
23. Bykerk VP, Lie E, Bartlett SJ, Alten R, Boonen A, Christensen R, Furst DE, Hewlett S, Leong AL, Lyddiatt A, March L, May JE, Montie P, Orbai A-M, Pohl C, Voshaar MS, Woodworth T, Bingham CO, Choy EH. Establishing a Core Domain Set to Measure Rheumatoid Arthritis Flares: Report of the OMERACT 11 RA Flare Workshop. *J Rheumatol* 2014 Apr;41(4):799–809. PMID:24584927
24. Berthelot J-M, De Bandt M, Morel J, Benatig F, Constantin A, Gaudin P, Le Loet X, Maillefert J-F, Meyer O, Pham T, Saraux A, Solau-Gervais E, Spitz E, Wendling D, Fautrel B, Guillemin F, STPR

group of French Society of Rheumatology. A tool to identify recent or present rheumatoid arthritis flare from both patient and physician perspectives: the “FLARE” instrument. *Ann Rheum Dis* 2012 Jul;71(7):1110–1116. PMID:22072015

25. Carpenter SM, Menictas M, Nahum-Shani I, Wetter DW, Murphy SA. Developments in Mobile Health Just-in-Time Adaptive Interventions for Addiction Science. *Curr Addict Rep* 2020 Sep;7(3):280–290. PMID:33747711
26. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAs) to reduce harmful substance use: a systematic review. *Addiction* n/a(n/a). doi: 10.1111/add.15687

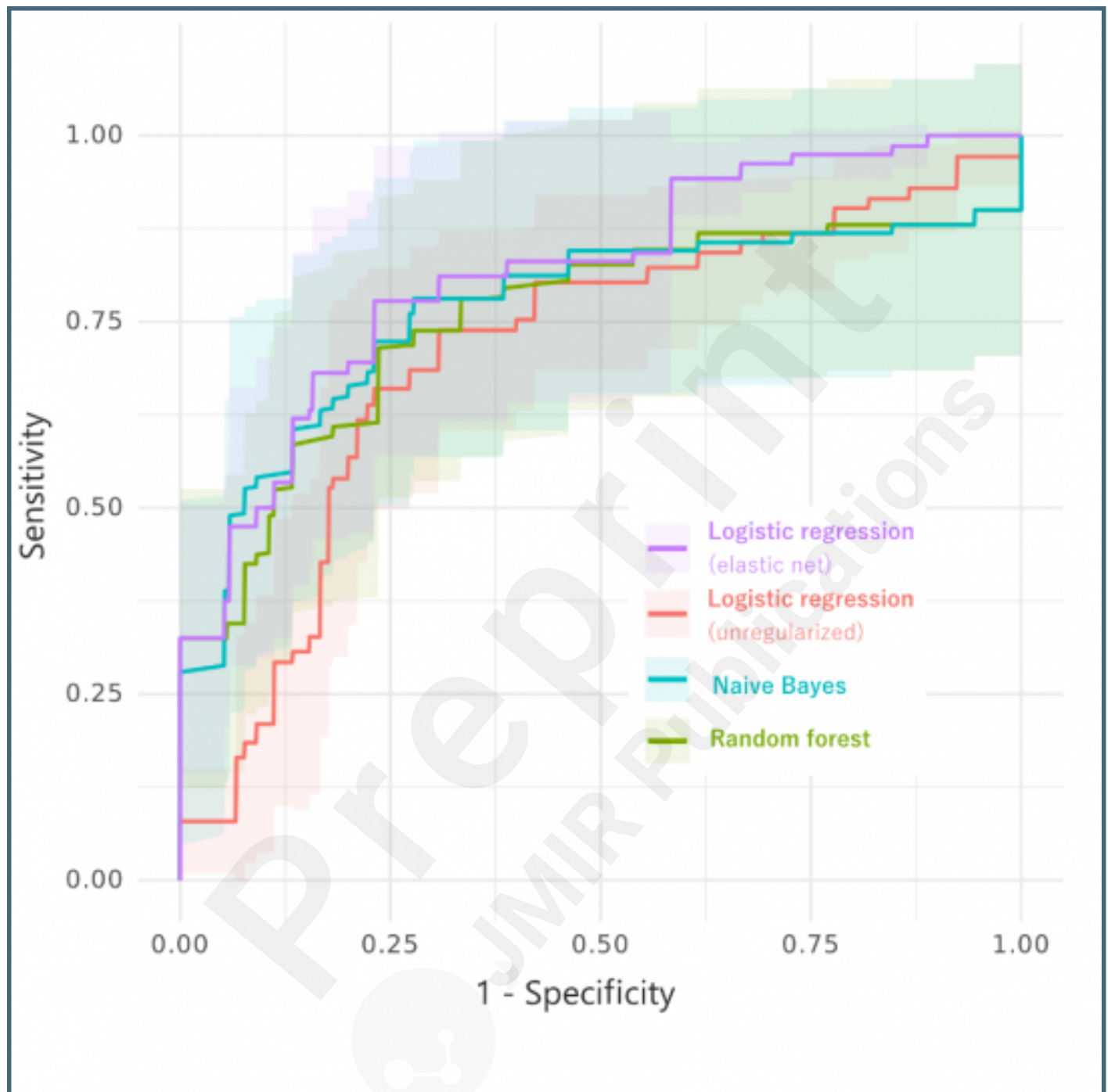
Supplementary Files

Figures

Patterns of daily and weekly data entry. Each row is a different participant. Vertical lines denote weekly responses, point daily responses. The shaded bands represent the week preceding each weekly response and the two colors denote whether patients reported a flare or no flare in that week.



Classifier performance for each of the four models.



Multimedia Appendixes

Supplementary figures.

URL: <http://asset.jmir.pub/assets/69a8328fec2017e93dd35df368ac85d6.docx>

