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

Laure Gossec, Frédéric Guyard, Didier Leroy, Thomas Lafargue, Michel Seiler, Charlotte Jacquemin, Anna Moltó, Jérémie Sellam, Violaine Foltz, Frederique Gandjbakhch, et al.

Running Head: Machine-learning detection of flares

Title: 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

Laure Gossec MD, PhD[1], Frédéric Guyard PhD[2], Didier Leroy MSc[2], Thomas Lafargue MSc[2], Michel Seiler MSc[3], Charlotte Jacquemin, MD, MSc[1], Anna Molto MD, PhD[4,5], Jérémie Sellam MD, PhD[6], Violaine Foltz MD, MSc[1], Frédérique Gandjbakhch MD, MSc[1], Christophe Hudry MD, MSc[4], Stéphane Mitrovic, MD, MSc[1], Bruno Fautrel MD, PhD[1], Hervé Servy MSc[7]

1. Sorbonne Université, Paris France; Pitié Salpêtrière hospital, APHP, Rheumatology department, Paris, France.

2. Orange IMT, Nice, France

3. Orange Healthcare, Paris, France

4. Rheumatology B Department, Cochin Hospital, AP-HP, Paris, France.

5. INSERM (U1153), Clinical Epidemiology and Biostatistics, PRES Sorbonne Paris-Cité, Paris Descartes University, Paris, France

6. Sorbonne Université, INSERM UMRS_938, Paris France; St-Antoine Hospital, AP-HP, DHU i2B, Rheumatology department, Paris, France

7. Sanoïa, e-Health services, Gémenos, France

Corresponding author

Pr Laure GOSSEC

Hôpital Pitié- Salpêtrière, Service de Rhumatologie

47-83 bd de l'hôpital, 75013 PARIS FRANCE

Email : laure.gossec@gmail.com

Tel=+33 142178421

Word count: 3607words, 47 references, 3 tables, 3 figures, 4 online tables

Funding:

The ActConnect initial study was funded by unrestricted academic grants from Lilly France, BMS France and Pfizer France. Orange IMT provided for free the technical infrastructure, the software used to perform Machine Learning and the experts to perform data management and analysis, during 10 weeks.

Disclosures:

Laure Gossec, Charlotte Jacquemin, Anna Molto, Jérémie Sellam, Violaine Foltz, Frédérique Gandjbakhch, Christophe Hudry, Stéphane Mitrovic, Bruno Fautrel: no relevant disclosures for this study.

Hervé Servy is the main shareholder of the SANOIA platform operating company: e-Health Services which acted as CRO for this study.

Frédéric Guyard,Didier Leroy, Thomas Lafargue are employees of Orange IMT and Michel Seiler is an employee of Orange Healthcare.

ABSTRACT (N=249 words)

Background

Flares in rheumatoid arthritis (RA) and axial spondyloarthritis (axSpA) may influence physical activity. The objective was to assess longitudinally the association between patient-reported flares and activity-tracker-provided steps per minute, using machine-learning.

Methods

This prospective observational study (ActConnect) included patients with definite RA or axSpA. During 3 months, physical activity was assessed continuously by number of steps/minute, using a consumer grade activity tracker, and flares were self-assessed weekly. Machine-learning techniques were applied to the dataset. After intra-patient normalization of the physical activity data, using multiclass Bayesian methods, sensitivities, specificities and predictive values of the machine-generated models of physical activity to predict patient-reported flares were calculated.

Results

In all, 155 patients (1339 weekly flare assessments and 224,952 hours of physical activity assessment) were analyzed: for RA (N=82) and axSpA (N=73) patients respectively, mean age was 48.9±12.6 and 41.2±10.3 years; mean disease duration was 10.5±8.8 and 10.8±9.1 years; 14 (17.1%) and 41 (56.2%) were males. Disease was well-controlled (mean DAS28: 2.2±1.2; mean BASDAI: 3.1±2.0) but flares were frequent (22.7% of all weekly assessments). The model generated by machine-learning performed well against patient-reported flares (mean sensitivity: 96% [95% confidence interval 94-97%], mean specificity: 97% [96-97%], mean positive and negative predictive value: 91% [88-96and 99% [98-100%]). Sensitivity analyses were confirmatory.

Conclusion

Although these pilot findings will have to be confirmed, the correct detection of flares by a machine learning processing of activity trackers data opens the way for future studies of remote-control monitoring of disease activity, with great precision and minimal patient burden.

Significance and Innovation

- Patient-reported flares were associated to less physical activity, measured by activity trackers, in rheumatoid arthritis (RA) and axial spondyloarthritis (axSpA), confirming the objective consequences of patient-reported flares.
- Using machine-learning, changes in physical activity patterns were found to be associated to patient-reported flares based on physical activity data with a sensitivity of 96% and a specificity of 97%.
- Given the relatively small sample size and the lack of a separate validation population, these findings should be further confirmed.
- Connected activity trackers with machine-learning processing may be an opportunity for continuous indication of disease activity in RA and axSpA.

ARTICLE TEXT

The evolution of rheumatoid arthritis (RA) and axial spondyloarthritis (axSpA) is marked by alternated periods of flares and stable disease activity.[1-7] Flares are important for patients since they contribute to the unpredictability of the disease.[8,9] Furthermore, due to the link between inflammation and structural degradation, flares are important to assess for disease management.[10-12] There is growing interest in both RA and axSpA, to characterize the reality behind the concept of patient-reported flares.[2,4,13-16] Flares appear to have objective consequences on daily life and in particular on physical activity.[17,18] Physical activity including everyday walking as well as aerobic exercise may be objectively and longitudinally assessed using connected activity trackers. These devices allow an interactive feedback of physical activity and the visualization of activity patterns, according to duration, intensity and frequency of physical activity.[19]

The ActConnect study was a 3-months longitudinal study of patients with either RA or axSpA where patient-reported flares were assessed weekly and physical activity was collected continuously using a connected activity tracker.[20, 21] The data were analysed using standard statistics and we found flares were related to a moderate decrease in physical activity, since during weeks with flares, there was a relative decrease in physical activity of 12-21%, i.e., an absolute decrease of 836-1462 steps/day.[21] This study thus confirmed objectively the functional impact of patient-reported flares. However at the group level and on amalgamated data, the link between flares and physical activity was weak and it was not possible to determine modifications in physical activity patterns which could adequately reflect patient-reported flares.[21]

Machine-learning allow multiple analyses of large datasets and make the best use of the available data, with minimal data amalgamation.[22] Although machine-learning methods have been little used in rheumatology to date [23], their usefulness in other medical fields has been clearly shown.[24-29] The specificity of such analyses is that the data is fed into a machine-learning operations tool, which will build – by itself - classification models, generated most often using an "averaging" of numerous naive Bayesian classifications.

The objective of this reanalysis of the ActConnect dataset was to assess longitudinally the association between patient-reported flares and activity-tracker-provided continuous flows of steps per minute, using machine-learning.

PATIENTS AND METHODS

Study design and patients

As previously reported, the ActConnect study was a prospective, multicenter, pragmatic, longitudinal observational study in France in 2016.[20,21] Briefly, patients had definite clinician-confirmed RA or axSpA and owned a smartphone or tablet which was compatible with the connected activity tracker and had an Internet access. There were no inclusion criteria related to disease activity or to physical activity. Ethical approval was obtained from the institutional review board (CPP Ile de France VI) and the human research ethics committee (CCTIRS, number 16.057bis).

Data collection

General and patient-reported data

Patient demographics and disease characteristics were collected at baseline including ongoing pharmacological treatment. Where available, in RA patients, the status for rheumatoid factor (RF) and for anti-cyclic citrullinated peptide (anti-CCP), the presence of radiographic erosions and the Disease Activity Score 28 (DAS28) at inclusion were recorded.[30] In axSpA patients, the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI, range 0 to 10), [31,32] the HLA B27 status, history of peripheral and of extra-articular symptoms, and the presence of sacroiliitis according to the medical file, on X-rays and/or on Magnetic Resonance Imaging (MRI) were recorded. Physical function was assessed by the modified Health Assessment Questionnaire (mHAQ).[33] Comorbidities were collected using the Functional Comorbidity Index, which ranges from 0 (0=no comorbidity, however the minimal score was 1 in the present study because of the rheumatic disease) to 18.[33bis]

Flares were assessed from the patient perspective by the question: « has your disease flared up since the last assessment ? », which has already been used in previous studies.[2] The categorical response was: no flare, flare lasting 1 to 3 days (short flare) or flare lasting more than 3 days (persistent flare). Flares were completed online from home each week during 3 months, following a text message reminder.[21]

Physical activity data

Each participant received an activity tracker (the Withings® Activité Pop watch[34]) and was instructed to wear it every day for 3 months. The Withings® tracker records the number of steps per minute. Ninety consecutive days from the first Monday following activation of the device were collected. No instruction about physical activity was given to the participants, however patients could visualize their physical activity on their smartphones.[21]

Statistical analysis

Patients were analyzed if they had at least one complete time point (i.e., physical activity data over the 7 days preceding a flare assessment).

Data preparation

Weeks containing more than 12 consecutive hours of missing or blank data, were removed from the data set (leading to 10559 days of a potential total 13950 which could be analysed). The remaining short-duration missing activity data (mostly due to non-wear periods and mostly at night) were not imputed (and were not analysed in the algorithm). For each patient, the mean and variance of number of steps for each aggregation interval with no flare were bootstrapped. The data for each patient were then normalized using these values, leading to a distribution of steps during a given aggregation interval with no flare with a mean of 0 and a variance of 1. Data preparation was performed on R.[35] The normalization was performed several times, since data were normalized for each aggregation interval. *Physical activity data aggregation*

The ActConnect study collected physical activity information (steps) at the minute level, during 3 months, leading to 13.5 million information points. Although very limited data aggregation is necessary for machine-learning software using Bayesian analyses, several levels of aggregation (24, 12, 4, and 1 hour) were tested, resulting in 4 distinct models.

Longitudinal relationship between physical activity and disease activity

The goal of the analysis was to classify each week as flare/no flare, based on the weekly physical activity data. The models were built using only the normalized steps and the patient-reported flare. Steps were analyzed both for deviation with respect to the reference week and for the importance of the time intervals with deviations. No other covariate was used. Of note, the models were developed at the population level not at the patient level (but patient data were individually normalized).For all the analyses, multiclass selective naïve Bayesian methods were performed using Khiops© (Orange Labs).[36-38] Naïve Bayes classifiers are

among the standard classification methods used in machine-learning and are based on a direct application of Bayes theorem.[26, 39-41]. Models corresponding to the 4 distinct aggregation intervals were built for 10 training/validation sets, and analyses were initially performed for the 3 levels of flare (no flare, flare ≤ 3 days, flare > 3 days) but this did not perform well (**online supplementary Table 1 and 2**). Thus, the binary variable (flare/no flare) was used. Then, the performance of the models was evaluated using patient-reported flares (assessed weekly) as gold standard and sensitivity and specificity, as well as positive and negative predictive values were assessed. Furthermore, to assess agreement, Cohen's kappa was calculated.[42]

Training and validation sets

In order to evaluate the performance of a classifier (and of any machine-learning model in general), the classifier is designed using a set of data (the training set) and its performance is evaluated using the classification of a distinct set of data (the validation set). To select a model (here the aggregation interval) and to take into account its mean performance but also the variation of the performance (i.e., the bias-variance trade-off [39]), 10 different training/validation sets were built. The generation of training/validation set was set at the weekly observation level On each of the 10 sets, the analyses were then performed for each aggregation interval. Each training/validation set was constructed using a random stratified 70% of the total weekly dataset (i.e. 936 weeks) as the training set, and the 30% remaining data as validation set (403 weeks). Each training/validation set used all of the available data (and the sets were not at all mutually exclusive). Thus, the datasets overlapped and on average, a given week was counted in 3 different validation sets. Data were stratified on flare/no flare. Performances were calculated for each set and on the pooled validation sets. Of note, in the pooled analysis, each week's data were used several times since the training/validation sets overlapped –thus these results should be considered as indicative only.

Illustration of results

The variations of the main statistical characteristics of the classifications in relation with the aggregation interval were reported for two training/validation examples. To illustrate changes for a single patient, a patient correctly classified as flare/no flare was chosen based on having age, number of flares and overall physical activity close to the population means. For this patient, mean physical activity for weeks without flare and for weeks with flare was graphically presented.

In the results of the modeling phase, Khiops provides an evaluation of the importance (the weight) of each explanatory variable in the model. Using these weights, a time-line map of "significant" moments of activity during the week was created. This time line shows the weight of the moments of the days used by the algorithm to perform the score calculation for the classification. The aggregation being hourly, these weights characterize the importance of each hour of each day of the week in the resulting flare classification.

Sensitivity analyses

All analyses were performed twice, from the data preparation to the model building, on Khiops®, independently by 2 statisticians. The analyses were also performed again using another machine-learning method (random forests classifiers [43]) on R and the code for this is given in **online supplementary appendix**.

RESULTS

Patients

Among the 170 patients included in the study, 155 (82 RA and 73 axSpA patients) were analyzed. This corresponds to 1339 weekly flare assessments and 224,952 hours physical activity assessment timeframes. Physical activity being provided at the minute, the dataset contained close to 13.5 million activity points.

For the 155 patients, mean age was for RA (n=82) 48.9±12.6 years, mean disease duration 10.5±8.8 years; 14 (17.1%) were males and for axSpA (n=73) 41.2±10.3 years, mean disease duration 10.8±9.1 years; 41 (56.2%) were males. Disease was well-controlled (mean DAS28: 2.2±1.2; mean BASDAI: 3.1±2.0) (**Table 1**). RA patients had a mean DAS28 at baseline of 2.2 (±1.2); 48/82 (58.5%) had radiographic erosions, and 63/79 (79.7%) had positive rheumatoid factor and/or anti-CCP. Among axSpA patients, 44/73 (60.3%) had experienced extra-articular symptoms, 42/70 (60.0%) had past or present peripheral symptoms; 50/65 (76.9%) carried HLA B27, and 54/64 (84.4%) had radiographic and/or MRI sacroiliitis. At baseline the mean BASDAI was 3.1 (±2.0). Overall, 81/155 (52.3%) patients were receiving biologics and 106/155 (68.4%) were stable in terms of treatment over the 3 months prior to inclusion (Table 1).

Among the 155 patients, 112 (72.2%) patients reported at least one flare over the 3 months follow up. Patients reported having experienced a flare on average at 22.7% of the questionnaires.

Over all the assessments, the mean number of steps was 6838 (±4033) steps/day with a median of 6265 (interquartile range [quartile 1 3843; quartile 3 9144]; range [minimum 0,maximum 38212] steps per day..

Detection of flares

The Khiops© program detected correctly both flares and absence of flare (mean sensitivity 95.7% [95% confidence interval: 94.4-97.0], mean specificity 96.7% [95% confidence interval: 96.0-97.3]) with high predictive values as well (**Table 2 and 3**).

Performances increased as the aggregation interval decreased: the best performance in terms of proportion of correctly/incorrectly classified instances was evidenced for 1-hour intervals (**Table 2**). The increase in the agreement between flares and predicted flares was also reflected in the substantial increase of the Kappa coefficient when the size of the aggregation intervals decreased (**Figure 1**).

The variations of the main statistical characteristics of the classifications in relation with the aggregation interval are reported for two training/validation examples in **online supplementary Table 3.**

Illustration of results

Figure 2 shows mean physical activity over weeks with versus without flares, for a random RA patient. As seen, there were considerable fluctuations overall and variations in patterns.

The model generated by the machine established “significant” moments of activity during the week that were more strongly related to flares (**Figure 3)**. It appeared working day mornings were not highly “significant” while the ends of the afternoons as well as Saturday afternoons appeared strongly associated with flare detection. These might be moments when patients can rest more, when in flare. In other words: when physical activity is different (from previous weeks) at a "significant" time point (eg Saturday afternoons) this is a flare state change

indicator for the machine. Reversely, a physical activity change in a "non significant" time point in the week, would be less contributive to flare state detection.

Sensitivity analyses

The second round of analyses and the analyses using another statistical technique on R, were confirmatory with >95% sensitivities and specificities (**online supplementary Table 4).**

DISCUSSION

This study demonstrated that patient-reported flares were strongly linked to physical activity and that machine-learning processing of patient-level physical activity can be used to detect flares with great accuracy. Furthermore, this study also demonstrated the usefulness of machine-learning applied to large rheumatology datasets.

This study has strengths and weaknesses. Firstly, the sample size in the present study was only moderate, and the relatively low number of flares may lead to power issues. However, the dataset for physical activity time points was very large. Furthermore, patients had either RA or axSpA, and as the analyses were pooled, this study does not allow to interpret possible differences between these 2 diseases. Secondly, this was a French, Paris-based study thus extrapolation to other cultures and social habits merits discussion, given the role of cultural background in perception and expression of patient-reported outcomes [44]. The strong link found between modifications in physical activity and patient-reported flares should not be directly interpreted as demonstrating causality. Confounding factors may have intervened. Indeed, other factors than flares that may impact physical activity (e.g., illness, mood, weather) were not collected, and it is currently uncertain how machine learning methods solely based on physical activity can distinguish between activity variation caused by disease flare and by other causes, although the model performances were remarkable here. The analyses were run several times on several subsets of the data which is a strength. However, the subsets were selected for frequency of flares; which could introduce a bias if the obtained models were used on another dataset since the proportions of week with flares may be different. Night movements and thus impact of flares on sleep could not be analysed here. Another point refers to the definition of flares – here, flares were classified as 'short' or 'long' but these definitions have not been validated. Finally, the obtained classifications are not easy to interpret since the machine chooses its criteria to build models, without human guidance.

Flares significantly impacted physical activity. These results confirm that patient-reported flares have an observable functional impact. The reality of the concept of flares has been much discussed.[1-7,16] Here, changes in physical activity patterns allowed to accurately detect patient-reported flares – in particular short flares, rather than longer (more than 3 day) flares – probably because shorter flares were much more frequent, although their clinical relevance is not well-established. The classification model generated by the machine allowed accurate detection of both periods of flare and periods of absence of flare. It is interesting to note that although flares were reported weekly (thus with moderate granularity), shorter intervals of aggregation for the physical activity data led to better prediction of flares than longer ones. Reasons for this include that flares may be more represented by the "way" the patients move during the day, implementing their own copying strategy, than the total number of steps during the day. Indeed, the model generated by the machine, established "significant" moments of activity during the week, which were markers of flares. These were patient-dependent, but we observed that they were often –but not only- the "less stressful" moments where patients can slow down or post-pone their physical activity (e.g. in the evenings or on Saturdays). In contrast, weekday mornings seemed less "significant" to detect

flares. This indicates flares of moderate severity may not lead to huge changes in physical activity such as being bed-ridden, but rather to the patient self-pacing his/her efforts.[6,16] We hypothesize patients may force themselves to deal with every day activities and work, even when in flare; and slow down mainly when this is not too disruptive to their lives; in particular in cases of moderate flares not necessitating medical intervention, as was the case here. It would be interesting to explore the long-term consequences of such moderate flares.[12]

The present reanalysis was centered on machine-learning. In this study, flares were collected from the patient regularly, once a week, and cross-tabulated with objective measurements of physical activity using a connected activity tracker. The dataset was first analyzed using traditional statistical methods for longitudinal datasets, and a significant but moderate decrease in physical activity was noted concomitantly to patient-reported flares. Furthermore, it was not possible to determine specific cutoffs of decrease in activity which would allow to predict a flare (for example, a decrease of 20% of steps or of 500 steps per day).[21] The interesting finding in the present reanalysis is that when applying different, innovative statistics with machine-learning, it was possible to find strong associations with excellent predictive capacities, between steps performed and patient-reported flares. This may be in part due to the capacity of the machine to compare the patient to himself.

Machine-learning statistics are complex procedures. Typically, the machine will use a dataset to develop a model, then a validation phase is necessary.[45] In the present study, this 2-phase approach was applied; but the lack of a separate validation group is a weakness. Indeed, the present findings should be seen as 'pilot findings' and it will be important to reproduce these findings in another cohort. Several techniques have been proposed for machine-learning. The random forest method performs well on datasets of reasonable size.[43] However, it is a lengthy process and is hard to industrialise with very large datasets. The other main method relies on Bayesian statistics and was applied here. Both methods appear to perform similarly.[46, 47] Although Bayesian modeling performed extremely well here, as in all machine-learning processes, some part of mystery remains since the exact decision mechanism of the machine when predicting flares is internal and implicit rather than explicit, making the interpretation difficult. Khiops© is based on sophisticated Naïve Bayes method (using both features selection and models averaging). It was initially developed as an easy to use and efficient tool in marketing.[36-38] Khiops© is used in many domains where classification or clustering is the subject and where massive data need to be analyzed. This study is a pilot for the use of Khiops© on healthcare data.

Connected devices and Internet of Things bring continuous flows of data that cannot be handled with traditional statistical tools without important complexity reduction and data aggregation. Using machine-learning, here the data could be analysed with minimal data aggregation. However, the comparison of models for different time aggregations was necessary. The fact that smaller timeframes performed better probably reflects the fact that, for flare characterization, the way patients are moving during the day is more indicative than their total activity over one day.

Data preparation was an essential (and time-consuming) step in the present analyses. Preliminary analyses confirmed that patterns of physical activity of two distinct patients may be quite different: a given physical activity for a patient during a week with flare may be similar to the physical activity of another patient during a week with no flare (data not shown). The normalization of steps led to all patients becoming comparable during weeks without flare and, from there, classification models were possible. It is probable that such normalization would allow analyses in different datasets with different characteristics, but this remains to be proven by further studies. Of note, measurement error (variability) in the device was not taken into account since trends over time were studied here.

Despite the conception choices made to limit take-off cases and so non-wear periods, including choosing a device that does not require to be plugged to a power source for

months or removed to wash, non-wear period were detected. This means we lost some information due to non-wear.

Furthermore, the use of overlapping subsets of patients and the lack of an independent testing set is an important issue which means that overfitting is a possibility. Other studies are needed. Overall, machine-learning technologies are still a growing field and require high-level technology but also relevant human expertise. These constraints need to be taken into account when planning future studies.

The correct detection of flares by the activity tracker and adapted statistics is of great interest. Indeed, activity trackers have great accuracy and lead to minimal patient burden, compared to online questionnaires or in-person visits. These results open perspectives to integrate in the future of connected devices in the monitoring of patients with chronic arthritis, in clinical research as well as in clinical practice. It is possible to imagine mixed-methods monitoring with continuous data collection via activity trackers, and physical assessments in person in case of frequent flares, for example. Of course, the cost of the wearable device needs to be taken into account. In a context of treating to a target, such continuous assessments (passively for the patient) may be of capital importance as the healthcare organization could benefit from more targeted outpatient visits (i.e., in case of flares). Finally, machine-learning methods may contribute to a more precise quantification of existing links or to the identification of new links in rheumatologic datasets.

In conclusion, this pilot application of machine-learning to physical activity assessment will open the way to future studies. The design of operational monitoring systems based on machine-learning models would however require careful validations on much larger datasets and the present analyses should be considered as a proof of concept of such an approach.

REFERENCES

1. Stone MA, Pomeroy E, Keat A, Sengupta R, Hickey S, Dieppe P, et al. Assessment of the impact of flares in ankylosing spondylitis disease activity using the Flare Illustration. Rheumatol. Oxf. Engl. 2008;47:1213–8.

2. Bykerk VP, Bingham CO, Choy EH, Lin D, Alten R, Christensen R, et al. Identifying flares in rheumatoid arthritis: reliability and construct validation of the OMERACT RA Flare Core Domain Set. RMD Open 2016;2:e000225.

3. Fautrel B, Morel J, Berthelot J-M, Constantin A, De Bandt M, Gaudin P, et al. Validation of flare-ra, a self-administered tool to detect recent or current rheumatoid arthritis flare. Arthritis Rheumatol. 2017;69(2):309-319.

4. Godfrin-Valnet M, Prati C, Puyraveau M, Toussirot E, Letho-Gyselink H, Wendling D. Evaluation of spondylarthritis activity by patients and physicians: ASDAS, BASDAI, PASS, and flares in 200 patients. Jt. Bone Spine Rev. Rhum. 2013;80:393–8.

5. Cooksey R, Brophy S, Gravenor MB, Brooks CJ, Burrows CL, Siebert S. Frequency and characteristics of disease flares in ankylosing spondylitis. Rheumatol. Oxf. Engl. 2010;49:929–32.

6. Bykerk VP, Shadick N, Frits M, Bingham CO, Jeffery I, Iannaccone C, et al. Flares in rheumatoid arthritis: frequency and management. A report from the BRASS registry. J. Rheumatol. 2014;41:227–34.

7. Bartlett SJ, Bykerk VP, Cooksey R, Choy EH, Alten R, Christensen R, et al. Feasibility and Domain Validation of Rheumatoid Arthritis (RA) Flare Core Domain Set: Report of the OMERACT 2014 RA Flare Group Plenary. J. Rheumatol. 2015;42:2185–9.

8. Gossec L, Paternotte S, Aanerud GJ, Balanescu A, Boumpas DT, Carmona L, et al. 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;70:935–42.

9. Kiltz U, van der Heijde D, Boonen A, Cieza A, Stucki G, Khan MA, et al. Development of a health index in patients with ankylosing spondylitis (ASAS HI): final result of a global initiative based on the ICF guided by ASAS. Ann. Rheum. Dis. 2015;74:830–5.

10. Aletaha D, Alasti F, Smolen JS. Rheumatoid arthritis near remission: clinical rather than laboratory inflammation is associated with radiographic progression. Ann. Rheum. Dis. 2011;70:1975–80.

11. Cooksey R, Brophy S, Dennis M, Davies H, Atkinson M, Irvine E, et al. Severe flare as a predictor of poor outcome in ankylosing spondylitis: a cohort study using questionnaire and routine data linkage. Rheumatol. Oxf. Engl. 2015;54:1563–72.

12. Raheel S, Matteson EL, Crowson CS, Myasoedova E. Improved flare and remission pattern in rheumatoid arthritis over recent decades: a population-based study. Rheumatology (Oxford). 2017 Dec 1;56(12):2154-2161.

13. Gossec L, Portier A, Landewé R, Etcheto A, Navarro-Compán V, Kroon F, et al. Preliminary definitions of “flare” in axial spondyloarthritis, based on pain, BASDAI and ASDAS-CRP: an ASAS initiative. Ann. Rheum. Dis. 2016;75:991–6.

14. Kirwan JR, Hewlett SE, Heiberg T, Hughes RA, Carr M, Hehir M, et al. Incorporating the patient perspective into outcome assessment in rheumatoid arthritis--progress at OMERACT 7. J. Rheumatol. 2005;32:2250–6.

15. Kwoh CK, Ibrahim SA. Rheumatology patient and physician concordance with respect to important health and symptom status outcomes. Arthritis Rheum. 2001;45:372–7.

16. Hewlett S, Sanderson T, May J, Alten R, Bingham CO, Cross M, et al. “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;51:69–76.

17. Hernández-Hernández V, Ferraz-Amaro I, Díaz-González F. Influence of disease activity on the physical activity of rheumatoid arthritis patients. Rheumatol. Oxf. Engl. 2014;53:722–31.

18. Brophy S, Cooksey R, Davies H, Dennis MS, Zhou S-M, Siebert S. The effect of physical activity and motivation on function in ankylosing spondylitis: a cohort study. Semin. Arthritis Rheum. 2013;42:619–26.

19. Van Genderen S, Boonen A, van der Heijde D, Heuft L, Luime J, Spoorenberg A, et al. Accelerometer Quantification of Physical Activity and Activity Patterns in Patients with Ankylosing Spondylitis and Population Controls. J. Rheumatol. 2015;42(12):2369-75.

20. Jacquemin C, Servy H, Molto A, Sellam J, Foltz V, Gandjbakhch F, et al. Physical Activity Assessment Using an Activity Tracker in Patients with Rheumatoid Arthritis and Axial Spondyloarthritis: Prospective Observational Study. JMIR Mhealth Uhealth. 2018 Jan 2;6(1):e1. doi: 10.2196/mhealth.7948.

21. Jacquemin C, Molto A, Servy H, Sellam J, Foltz V, Gandjbakhch F, et al. Flares assessed weekly in patients with rheumatoid arthritis or axial spondyloarthritis and relationship with physical activity measured using a connected activity tracker: a 3-month study. RMD Open. 2017 Jun 29;3(1):e000434.

22. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning, Second Edition, Springer Series in Statistics, 2009.

23. González FA. Machine learning models in rheumatology. Rev Colomb Reumatol. 2015 ;22:77-8.

24. Quellec G, Lamard M, Erginay A, Chabouis A, Massin P, Cochener B, et al. Automatic detection of referral patients due to retinal pathologies through data mining. Med Image Anal. 2016;29:47-64.

25. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist- level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.

26. Tomar D. A survey on Data Mining approaches for Healthcare. Int J Bio-Science and Bio-Technology. 2013;5(5):241–66.

27. Nandy, Jay & Hsu, Wynne & Lee, Mong. An Incremental Feature Extraction Framework for Referable Diabetic Retinopathy Detection. In 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI). 2016;908–12.

28. Fergus P, Hussain A, Hignett D, Al-Jumeily D, Abdel-Aziz K, Hamdan H. A machine learning system for automated whole-brain seizure detection. Applied Computing and Informatics. 2016;12:70–89.

29. Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI. Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. Comput Struct Biotechnol J. 2017;15:26–47.

30. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum. 1995;38:44–8.

31. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. J. Rheumatol. 1994;21:2286–91.

32. Claudepierre P, Sibilia J, Goupille P, Flipo RM, Wendling D, Eulry F, et al. Evaluation of a French version of the Bath Ankylosing Spondylitis Disease Activity Index in patients with spondyloarthropathy. J. Rheumatol. 1997;24:1954–8.

33. Pincus T, Summey JA, Soraci SA, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum. 1983;26:1346–53.

33bis. Groll DL, To T, Bombardier C, Wright JG. The development of a comorbidity index with physical function as the outcome. J. Clin. Epidemiol. 2005;58:595–602.

34. «Withings Activity Pop Watch». Nokia. Accessed 20th December 2017, https://support.health.nokia.com/hc/en-us/categories/200208646

35. « R: The R Project for Statistical Computing ». The R Foundation. Accessed 20th December 2017. https://www.r-project.org/.

36. Boullé M. Compression-Based Averaging of Selective Naive Bayes Classifiers. J Mach Learn Res. 2007;8:1659–85.

37. Boullé M. Khiops: outil de préparation et modélisation des données pour la fouille des grandes bases de données. In Extraction et gestion des connaissances (EGC'2008). 2008;229–30.

38. «Khiops software for data mining». PredicSis S.A.S. Accessed may 2017 : https://khiops.predicsis.com .

39. Hastie, T, Tibshirani, R, Friedman, J. Overview of Supervised Learning. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY. 2009;9–41.

40. Rish I. An empirical study of the naïve Bayes classifier. IBM Research Report, RC 22230, November 2, 2001.

41. Boullé M. MODL: A Bayes optimal discretization method for continuous attributes, Mach. Learn. 2006;65:131–165.

42. Agresti A. Categorical Data Analysis, Second Edition, Wiley-Interscience, John Wiley & Sons, 2002.

43. Douglas PK, Harris S, Yuille A, Cohen MS. Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. NeuroImage. 2011;56(2):544–53.

44. Putrik P, Ramiro S, Hifinger M, et al. In wealthier countries, patients perceive worse impact of the disease although they have lower objectively assessed disease activity: results from the cross-sectional COMORA study. Ann Rheum Dis. 2016;75:715–20.

45. "Machine Learning Wars: Amazon vs Google vs BigML vs PredicSis" ; kdnuggets.com ; Accessed 20th December 2017 : http://www.kdnuggets.com/2015/05/machine-learning-wars-amazon-google-bigml-predicsis.html

46. Nabi M, Kumar P, Wahid A. Performance Analysis of Classification Algorithms in Predicting Diabetes. International Journal of Advanced Research in Computer Science. 2017;8(3):456–61.

47. Kukreja M, Stephen AJ, Stafford P. Comparative study of classification algorithms for immunosignaturing data. BMC Bioinformatics. 2012;13:139.

Table 1: Characteristics of 82 RA and 73 axSpA patients

	RA (N=82)	**axSpA (N=73)**
Sex, N (%), males	14 (17.1)	41 (56.2)
Age, mean (SD), years	48,9 (12.6)	41.2 (10.3)
BMI, mean (SD), kg/m²	24.7 (4.5)	24.6 (4.6)
Disease duration, mean (SD), years	10.5 (8.8)	10.8 (9.1)
Work status, N (%), employed	61 (74.4)	61 (83.6)
Of whom,		
Manual work	3 (4.9)	2 (3.3)
Intellectual work	58 (95.1)	62 (96.7)
Studies > high school, N (%)	69 (84.1)	66 (90.4)
Functional comorbidity Index (range, 1-18), mean (SD)	1.6 (0.9)	1.4 (0.9)
mHAQ (0-3), mean (SD)	0.23 (0.39)	0.30 (0.33)
Ongoing treatment		
NSAIDs, N (%)	17 (20.7)	44 (60.3)
Glucocorticoids, N (%)	19 (23.2)	1 (1.4)
Conventional synthetic DMARDs, N (%)	76 (92.7)	17 (23.3)
Of whom, methotrexate, N (%)	66 (86.8)	13 (76.5)
Biological therapy, N (%)	37 (45.1)	44 (60.3)
Of whom, antiTNF, N (%)	23 (62.2)	44 (100)
No change in arthritis drugs over the 3 months prior to inclusion, N (%)	59 (72.0)	47 (64.4)

Percentages are calculated on all complete data; RA= rheumatoid arthritis, SpA= spondyloarthritis, SD= standard deviation, BMI= body mass index, mHAQ: modified Health Assessment Questionnaire,[27] NSAIDs= nonsteroidal anti-inflammatory drugs, DMARDs= disease-modifying antirheumatic drugs.

Table 2: Association between physical activity and self-reported flares.

validation set	Kappa (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
1	0.88 (0.83-0.94)	0.96 (0.91-0.99)	0.96 (0.94-0.98)	0.87 (0.81-0.94)	0.99 (0.98-1.0)
2	0.89 (0.84-0.94)	0.98 (0.95-1.00)	0.95 (0.93-0.98)	0.87 (0.80-0.93)	0.99 (0.98-1.0)
3	0.91 (0.86-0.96)	0.95 (0.90-0.99)	0.97 (0.95-0.99)	0.91 (0.86-0.98)	0.98 (0.97-1.0)
4	0.88 (0.82-0.93)	0.93 (0.88-0.98)	0.96 (0.94-0.98)	0.88 (0.82-0.95)	0.98 (0.97-1.0)
5	0.92 (0.87-0.96)	0.98 (0.95-1.00)	0.97 (0.95-0.99)	0.90 (0.85-0.96)	0.99 (0.98-1.0)
6	0.88 (0.82-0.94)	0.90 (0.84-0.96)	0.97 (0.96-0.99)	0.91 (0.86-0.97)	0.97 (0.96-0.99)
7	0.92 (0.88-0.96)	0.97 (0.93-1.00)	0.97 (0.96-0.99)	0.92 (0.87-0.98)	0.99 (0.98-1.0)
8	0.90 (0.85-0.95)	0.96 (0.91-0.99)	0.97 (0.95-0.99)	0.89 (0.84-0.96)	0.99 (0.97-1.0)
9	0.92 (0.88-0.97)	0.96 (0.90-0.99)	0.98 (0.96-0.99)	0.93 (0.87-0.98)	0.99 (0.98-1.0)
10	0.89 (0.85-0.95)	0.99 (0.96-1.00)	0.95 (0.93-0.98)	0.89 (0.88-0.91)	0.99 (0.98-0.99)
Pooled results	0.90 (0.89-0.92)	0.96 (0.94-0.97)	0.97 (0.96-0.97)	0.89 (0.88-0.91)	0.99 (0.98-1.00)

Kappa statistics, sensitivity and specificity, positive and negative predictive values (PPV and NPV) of the model against self-reported flares as gold-standard, reported here for the 10 validation sets (403 weekly data) and for the aggregation of 1 hour.

(95% CI): 95% confidence interval

Table 3: Detection of patient-reported flares by physical activity: pooled results:

N (% of 4030 weeks)	No patient-reported flare	Patient-reported flare
No flare according to Khiops	3006 (74.6%)	40 (1.0%)
Flare according to Khiops	104 (2.6%)	880 (21.8%)

Footnote. Results presented are the sum over the 10 training/validation sets of the confusion matrices with 1 hour aggregation (4030 weeks containing 3110 weeks patient-reported as no-flare and 920 reported as flare). Each week's data is used several times since the training/validation sets overlapped –thus these results should be considered indicative only.

Online supplementary Table 1: Patient-reported flares as 3 categories (no flare, short flare, long flare) versus machine-learning predicted flares

validation set	kappa	sensitivity (No flare)	sensitivity (short flare)	sensitivity (long flare)	specificity (no flare)	specificity (short flare)	specificity long flare)
1	0.76	0.96	0.91	0.0	0.93	0.91	0.99
2	0.77	0.96	0.96	0.0	0.98	0.91	0.99
3	0.78	0.97	0.95	0.0	0.95	0.92	1.0
4	0.75	0.97	0.69	0.0	0.91	0.92	1.0
5	0.81	0.98	0.95	0.0	0.96	0.93	1.0
6	0.78	0.97	0.95	0.0	0.88	0.94	1.0
7	0.79	0.97	0.96	0.0	0.96	0.92	1.0
8	0.81	0.98	0.95	0.0	0.92	0.94	1.0
9	0.81	0.98	0.96	0.0	0.95	0.93	1.0
10	0.79	0.96	0.97	0.0	0.98	0.91	1.0

3-levels classification: Kappa statistics, sensitivity and specificity for the 10 validation sets (403 weekly data) and for the aggregation of 1 hour.

Online supplementary Table 2. Patient-reported flares in 3 categories (no flare, short flare, long flare) versus machine-learning predicted flares overall and corresponding bootstrapped indicators

3-levels classification	No patient-reported flare	Patient-reported flare ≤ 3 days	Patient-reported flare > 3 days
No flare according to Khiops©	3016	30	24
Flare ≤ 3 days according to Khiops©	94	693	166
Flare >= 3 days according to Khiops©	0	7	0

Sum over all training/validation sets of the confusion matrices with 1 hour aggregation (4030 weeks: ie, patient-reported as no-flare: 3110, flare ≤ 3 days: 730, flare> 3 days: 190) for the 3-levels classification

kappa	**sensitivity (No flare)**	**sensitivity (short flare)**	**sensitivity (long flare)**	**specificity (no flare)**	**specificity (short flare)**	**specificity long flare)**
0.78 [0.76,0.80]	**0.97 [0.96,0.98]**	**0.95 [0.93,0.96]**	**0.0**	**0.94 [0.92,0.96]**	**0.92 [0.91,0.93]**	**1.0**

Comment: The model failed to precisely make the distinction between short and long flares, but was sensitive to the distinction flare vs. no flare.

Online supplementary Table 3: Variations of agreements between patient-reported flares and flares predicted by physical activity (steps) using multiclass Bayesian classification for two examples of training/validation sets.

Validation set 4 (403 weeks)	Se	Sp	PPV	NPV	Kappa	a	b	c	d
24h	0.49	0.94	0.71	0.86	0.4851	45 (11%)	18 (4%)	47 (12%)	293 (73%)
12h	0.70	0.93	0.76	0.91	0.6540	65 (16%)	21 (5%)	27 (7%)	290 (72%)
4h	0.87	0.94	0.80	0.96	0.7874	80 (20%)	19 (5%)	12 (3%)	292 (72%)
1h	0.93	0.96	0.88	0.98	0.8761	86 (21%)	12 (3%)	6 (2%)	299 (74%)
Validation set 9 (403 weeks)									
24h	0.45	0.98	0.89	0.86	0.5214	41 (10%)	5 (1%)	51 (13%)	306 (76%)
12h	0.57	0.98	0.91	0.88	0.6341	52 (13%)	5 (1%)	40 (10%)	306 (76%)
4h	0.84	0.96	0.87	0.95	0.8076	77 (19%)	12 (3%)	15 (4%)	299 (74%)
1h	0.9-	0.98	0.93	0.99	0.9234	88 (22%)	7 (2%)	4 (1%)	304 (75%)

Results are presented for 2 validation sets, one with high kappa (validation set 9 in Table 2) and one with low kappa (validation set 4 in Table 2).

Se:sensitivity and Sp: specificity, against the gold standard (patient reported flares)

PPV, NPV: positive and negative predictive values; CI: confidence interval

a, b, c, d: corresponds to the 2x2 tables, specifically, a=flare correctly predicted, b=no flare according to the patient but flare predicted, c= flare according to the patient, not predicted; d= absence of flare correctly predicted (each number is n (% of total which is 403 in all cases)).

Online supplementary Table 4: Sensitivity analyses using random forest methods. Patient-reported flares versus random forest predicted flares overall

Aggregation interval (h)	Kappa (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
1	0.88 (0.86-0.90)	0.84 (0.82-0.86)	0.99 (0.99-0.99)
4	0.83 (0.81-0.85)	0.79 (0.76-0.82)	0.99 (0.99-0.99)
12	0.65 (0.62-0.68)	0.63 (0.60-0.66)	0.96 (0.96-0.97)
24	0.54 (0.51-0.58)	0.52 (0.49-0.55)	0.96 (0.95-0.97)

Figure 1 Title Agreement between patient-reported flares and predicted flares for different time-aggregation intervals: kappa statistics.

Figure 1 Footnote Footnote: box plots for the kappas observed on the different validation sets are presented. Scale for kappa: very bad: <0, weak:]0,0.2], decent:]0.2,0.4], moderate:]0.4,0.6], substantial:]0.6,0.8], almost perfect:]0.8,1[, perfect: 1.

Figure 2 Title. One example of activity for weeks with versus without flare, for a randomly chosen patient over all weeks of activity (mean values are presented for weeks with or without flares)

Figure 2 Footnote.

X-axis: hours over one week (starting on Sunday 1 AM), Y-axis: physical activity in steps per hour (mean values for weeks with or without flare).

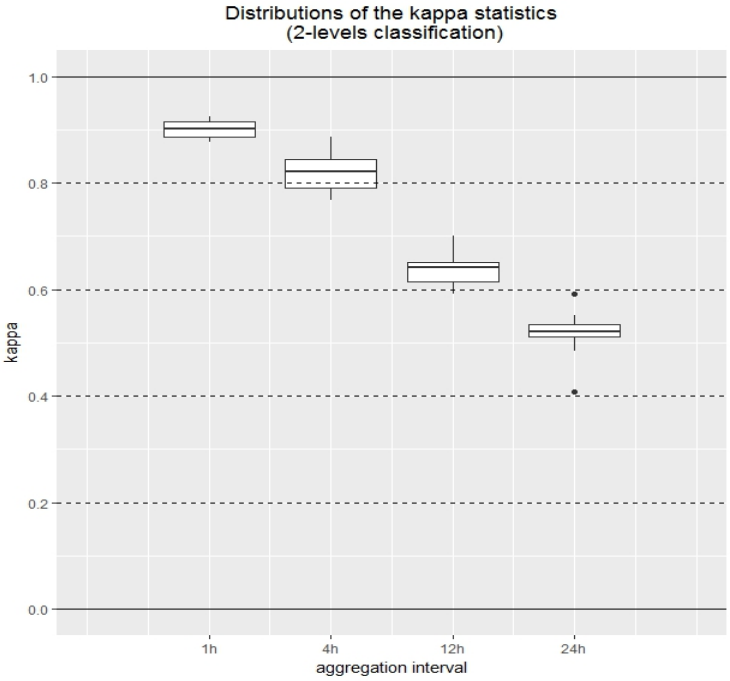
The dotted line represents mean steps per hour in weeks with flare, and the full line mean steps per hour in weeks without flare.

Figure 3 Title. Illustration of importance (weight) of the various time intervals over a week to predict flares.

Figure 3 Footnote

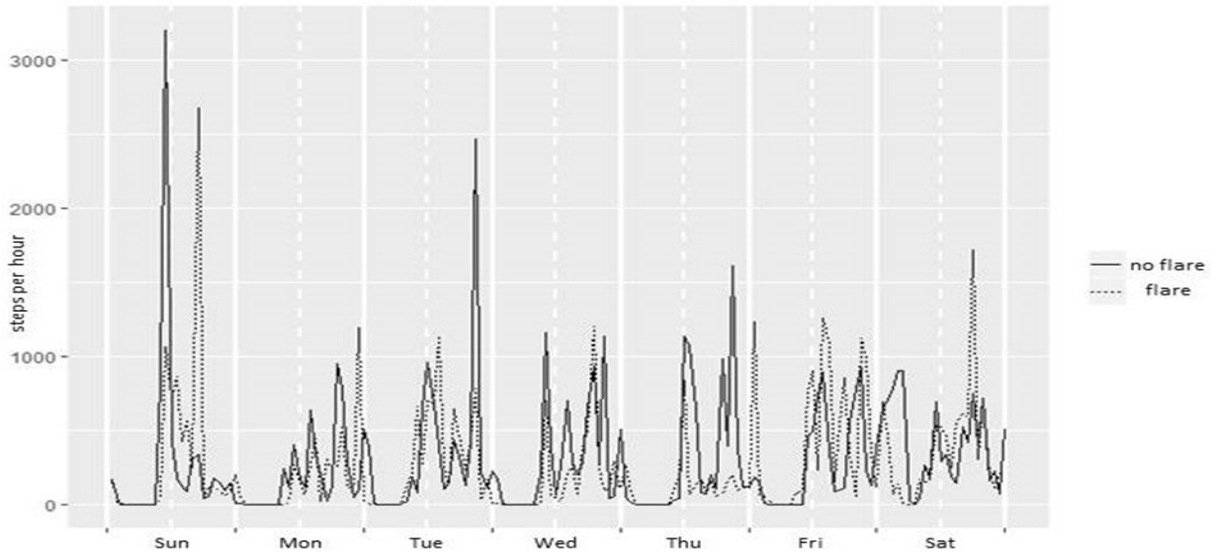
This figure shows how the moments of the week (and the day) are weighted by the algorithm to perform the classification, in one example (ie, one instance of training/validation sets). The X axis present the days of the week, split in one-hour intervals (the dotted line divides AM and PM for each day). The darker the colour, the more important is a time interval in the classification.

Figure 1 Agreement between patient-reported flares and predicted flares for different time-aggregation intervals: kappa statistics.



Footnote: box plots for the kappas observed on the different validation sets are presented. Scale for kappa: very bad: <0, weak:]0,0.2], decent:]0.2,0.4], moderate:]0.4,0.6], substantial:]0.6,0.8], almost perfect:]0.8,1[, perfect: 1.

Figure 2. One example of activity for weeks with versus without flare, for a randomly chosen patient over all weeks of activity (mean values are presented for weeks with or without flares)

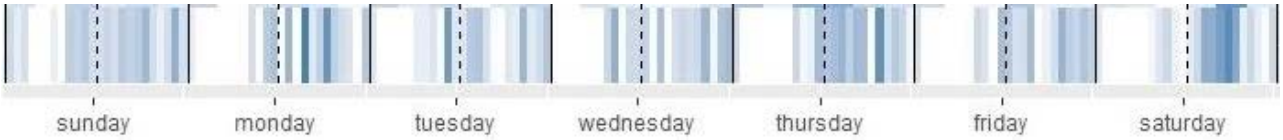


Footnote.

X-axis: hours over one week (starting on Sunday 1 AM), Y-axis: physical activity in steps per hour (mean values for weeks with or without flare).

The dotted line represents mean steps per hour in weeks with flare, and the full line mean steps per hour in weeks without flare.

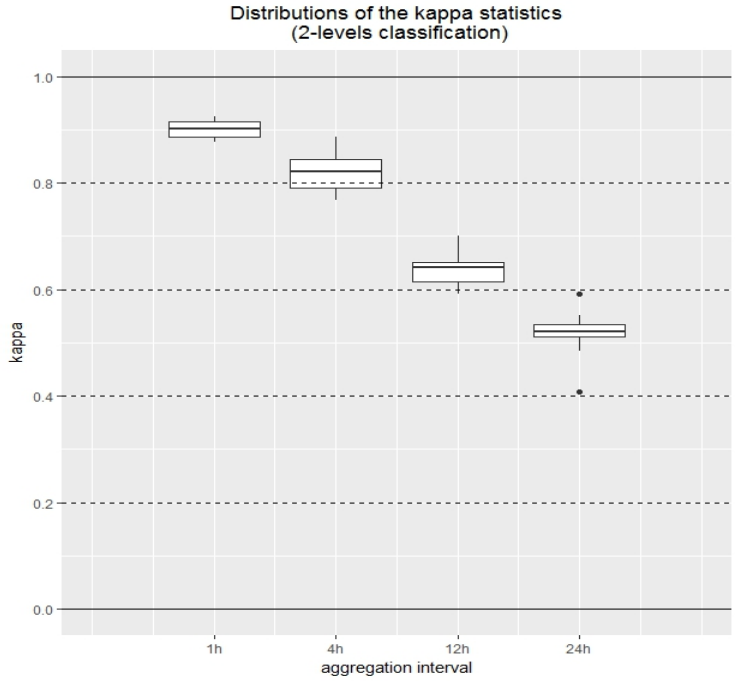
Figure 3. Illustration of importance (weight) of the various time intervals over a week to predict flares.



Footnote

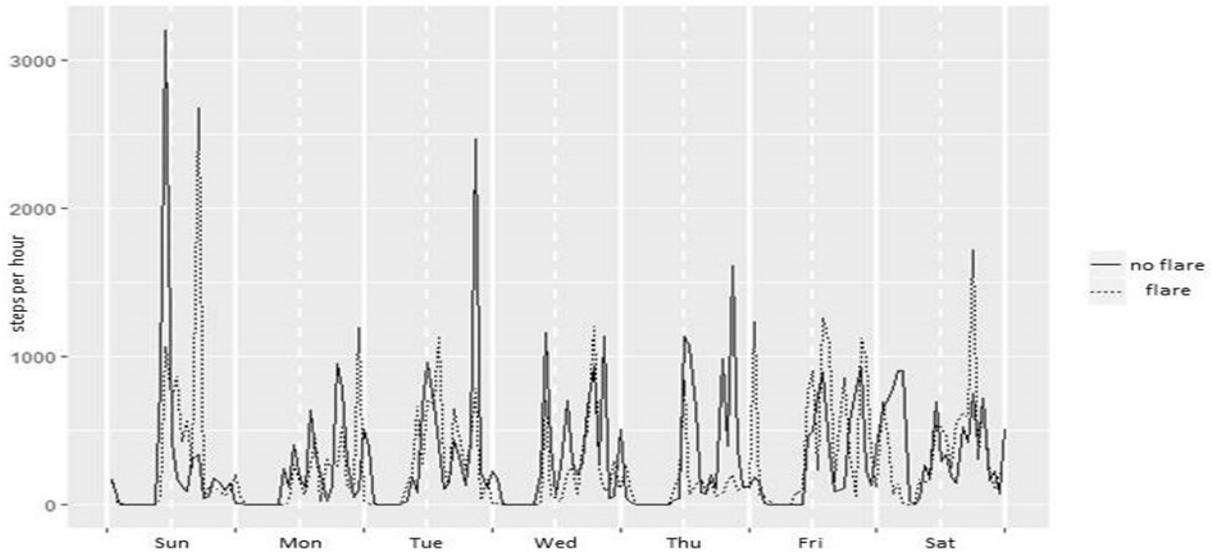
This figure shows how the moments of the week (and the day) are weighted by the algorithm to perform the classification, in one example (ie, one instance of training/validation sets). The X axis present the days of the week, split in one-hour intervals (the dotted line divides AM and PM for each day). The darker the colour, the more important is a time interval in the classification.

Figure 1 Agreement between patient-reported flares and predicted flares for different time-aggregation intervals: kappa statistics.



Footnote: box plots for the kappas observed on the different validation sets are presented. Scale for kappa: very bad: <0, weak:]0,0.2], decent:]0.2,0.4], moderate:]0.4,0.6], substantial:]0.6,0.8], almost perfect:]0.8,1[, perfect: 1.

Figure 2. One example of activity for weeks with versus without flare, for a randomly chosen patient over all weeks of activity (mean values are presented for weeks with or without flares)

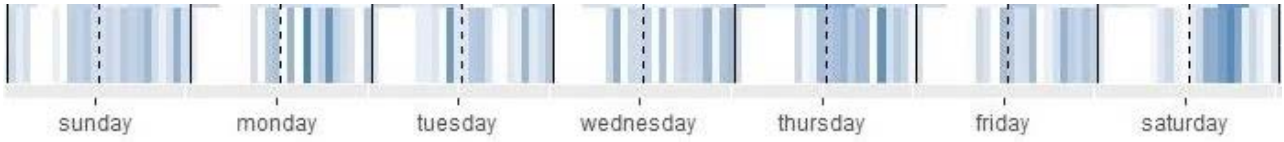


Footnote.

X-axis: hours over one week (starting on Sunday 1 AM), Y-axis: physical activity in steps per hour (mean values for weeks with or without flare).

The dotted line represents mean steps per hour in weeks with flare, and the full line mean steps per hour in weeks without flare.

Figure 3. Illustration of importance (weight) of the various time intervals over a week to predict flares.



Footnote

This figure shows how the moments of the week (and the day) are weighted by the algorithm to perform the classification, in one example (ie, one instance of training/validation sets). The X axis present the days of the week, split in one-hour intervals (the dotted line divides AM and PM for each day). The darker the colour, the more important is a time interval in the classification.