npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

Development of digital measures for nighttime scratch and sleep using wrist-worn wearable devices

Nikhil Mahadevan[1,5✉], Yiorgos Christakis[1,5], Junrui Di[1], Jonathan Bruno[1], Yao Zhang[1], E. Ray Dorsey[2], Wilfred R. Pigeon[2,3], Lisa A. Beck[2], Kevin Thomas[4], Yaqi Liu[4], Madisen Wicker[4], Chris Brooks[4], Nina Shaafi Kabiri[4], Jaspreet Bhangu[4], Carrie Northcott[1] and Shyamal Patel[1]

Patients with atopic dermatitis experience increased nocturnal pruritus which leads to scratching and sleep disturbances that significantly contribute to poor quality of life. Objective measurements of nighttime scratching and sleep quantity can help assess the efficacy of an intervention. Wearable sensors can provide novel, objective measures of nighttime scratching and sleep; however, many current approaches were not designed for passive, unsupervised monitoring during daily life. In this work, we present the development and analytical validation of a method that sequentially processes epochs of sample-level accelerometer data from a wrist-worn device to provide continuous digital measures of nighttime scratching and sleep quantity. This approach uses heuristic and machine learning algorithms in a hierarchical paradigm by first determining when the patient intends to sleep, then detecting sleep–wake states along with scratching episodes, and lastly deriving objective measures of both sleep and scratch. Leveraging reference data collected in a sleep laboratory (NCT ID: NCT03490877), results show that sensor-derived measures of total sleep opportunity (TSO; time when patient intends to sleep) and total sleep time (TST) correlate well with reference polysomnography data (TSO: $r = 0.72$, $p < 0.001$; TST: $r = 0.76$, $p < 0.001$; $N = 32$). Log transformed sensor derived measures of total scratching duration achieve strong agreement with reference annotated video recordings ($r = 0.82$, $p < 0.001$; $N = 25$). These results support the use of wearable sensors for objective, continuous measurement of nighttime scratching and sleep during daily life.

npj Digital Medicine (2021)4:42 ; https://doi.org/10.1038/s41746-021-00402-x

INTRODUCTION

Pruritus (itch) is a primary symptom seen in numerous chronic eczematous conditions, especially prevalent in patients with atopic dermatitis (AD)[1]. A common reaction to the pruritus sensation is to scratch the affected area[2,3], which results in additional inflammation/lesion formation thus exacerbating the pruritus and perpetuating the itch–scratch cycle[4]. Furthermore, pruritus often occurs during the evening and at night and disrupts patients' sleep[5]. The itch–scratch cycle compounded with sleep disturbances reduces the quality of life of patients as well as caregivers[6,7].

Traditional assessments of pruritus and sleep are primarily based on clinical outcome assessments (COAs) and patient reported outcome assessments (PROs). COAs are aimed at assessing total body surface area (BSA) of the lesion[8] as well as lesion severity (redness, induration, excoriations, etc.)[9] but these are physician-derived measurements and provide limited insight into the fluctuations of symptoms experienced outside the clinic. In contrast, while PROs provide insight into the perceived condition from the patient's perspective, they are subjective and can be affected by mood or suggestion, lack compliance, and are qualitative in nature[10]. Therefore, there is a need for more objective measures that accurately reflect the impact of AD on a patient's daily life. These types of measurements not only have the potential to provide more reliable indicators of intervention efficacy, but may also help improve management of the disease.

Advances in wearable sensor technology have already led to more objective measures of health, both within and outside of healthcare settings. Measurement of sleep/wake cycles using wrist-worn accelerometers has continued to evolve since their introduction[11–13]. With improved algorithms, the hope is that actigraphy would provide a more practical and valid option to longitudinally monitor sleep compared to polysomnography (PSG) (the gold standard for sleep assessment)[14]. While PSG provides rich information beyond distinguishing between sleep and wake states (e.g. identifying sleep stages, gross body movements, and respiration patterns), the technically demanding, in-clinic, overnight requirements and cost make it a poor choice for long-term monitoring in situations when these additional parameters are not needed. By leveraging wrist-worn accelerometers, high-resolution measurements can be collected for weeks to months at a time with minimal disturbance to a patient's daily life. Although methods that rely on accelerometer data are unable to reliably detect sleep stages, they have been used to effectively detect long-term changes in circadian rhythms and sleep quantity[12–14].

More recently, there have been efforts to leverage data captured using wrist-worn accelerometers in combination with machine learning (ML) techniques to measure nighttime scratching[15–18]. Feuerstein et al.[16] utilized four signal features derived from accelerometer data with a k-means clustering technique to segment simulated scratching movements (scratching performed on command in a clinic setting) from walking and restless movements during sleep. Petersen et al.[17] built on this approach by leveraging the same four signal features with logistic regression to also classify simulated scratching movements from walking and restless movements during sleep. However, while both methods achieved high sensitivity in predicting scratch movements (0.90 and 0.96, respectively), because both methods rely on simulated scratching movements and were not designed

[1]Pfizer, Inc., Cambridge, MA, USA. [2]University of Rochester Medical Center, Rochester, NY, USA. [3]Department of Veterans Affairs, Canandaigua, NY, USA. [4]Boston University School of Medicine, Boston, MA, USA. [5]These authors contributed equally: Nikhil Mahadevan, Yiorgos Christakis. ✉email: nmdevan816@gmail.com

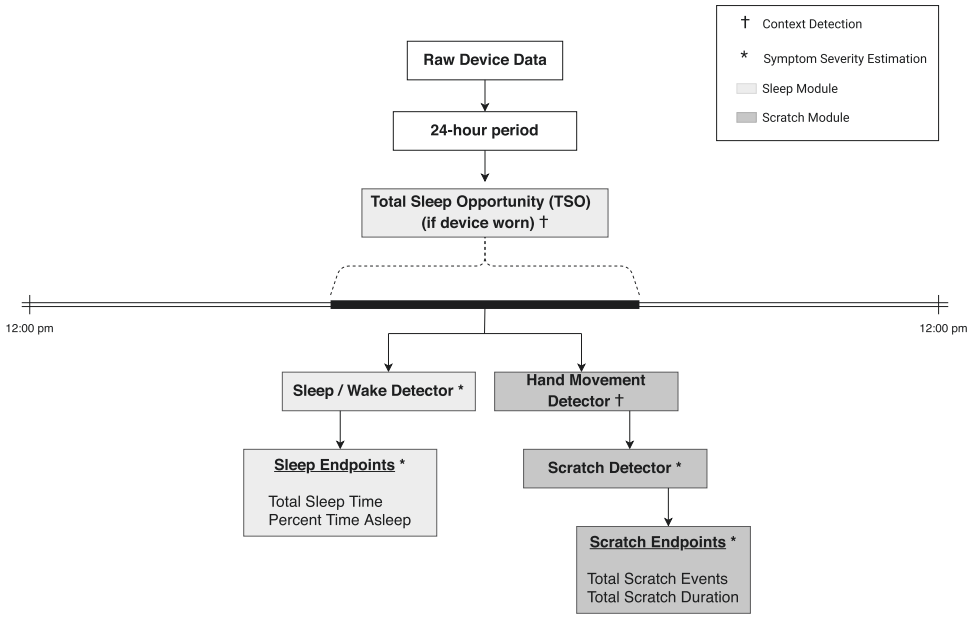

Fig. 3 Flowchart highlighting hierarchical approach for detection and assessment of sleep and nighttime scratch using accelerometer data from a wrist-worn device. Raw accelerometer data are first sliced to a 24-h period (12:00 PM–12:00 PM), then segmented to the total sleep opportunity (TSO) window, and finally measures of sleep and scratch are computed during the TSO window.

restless movement criteria used to define and annotate scratch behaviors (developed by Boston University School of Medicine Laboratory for Human Neurobiology) are in Supplementary Note 1. In addition to videography, during the second overnight clinic visit, participants' sleep was monitored via limited PSG. Following the in-clinic visit, participants were monitored for two nights at home. All procedures in this study had approval from the University of Rochester Medical Center Institutional Review Board. All participants in the study gave written informed consent prior to enrollment. Data from both in-clinic visits were used for scratch algorithm development and data from the second in-clinic visit were used for sleep algorithm development.

Instrumentation

Participants wore two devices (GeneActiv Original; Activinsights, Kimbolton, UK), one on each wrist, during both in-clinic and at-home visits. These devices have a watch-like form factor (although no watch face) and are designed for continuous, multi-day recordings in both free-living and clinical environments. Sample-level sensor data (triaxial acceleration, near-body temperature, and ambient light) is logged on the device and can be downloaded at the end of the monitoring period. In this study, data from a triaxial accelerometer (sampling rate: 100 Hz, unit: g), ambient light sensor (sampling rate: 100 Hz, unit: Lux), and near-body temperature sensor (sampling rate: 0.334 Hz, unit: Celsius) were collected. Participants were instructed to wear the devices at least 3 h prior to the first overnight clinic visit and leave them on throughout the evaluation period (i.e. not remove during the day). During the second overnight clinic visit, participants underwent PSG, which was used as the ground truth measurement of sleep. PSG recordings consisted of external electrodes (three electroencephalography (EEG) sites (C3, C4, and Occipital), two electrooculography (EOG) sites, two facial EMG sites, reference electrodes, and ground) that were placed on the head and face (one on either side of each eye, one behind each ear, and two on the chin/jawline). PSG recordings did not include measurement of respiration, limb movement, or oximetry, and were scored in 30-s epochs per revised American Academy of Sleep Medicine scoring guidelines[31]. Both in-clinic visits included thermal videography recorded in 72,000 frame batches at 60 Hz (equating to 20 min per video) for the duration of the subject's overnight visit.

Analytical approach

As illustrated in Fig. 3, the proposed method for assessment of nighttime scratch and sleep follows a hierarchical paradigm by first performing context detection and then estimating symptom severity. Context detection consists of wear detection (on-body vs. off-body), detection of the subject's TSO window (defined as the largest period in a 24-h window during which sleep is the intended behavior, or more simply the time from when the participant laid down to go to bed to the time when they rise in the morning), and detection of hand movement during the predicted TSO window. Symptom severity estimation consists of detection and assessment of sleep quantity and nighttime scratch. In order to evaluate sleep and scratch independently, we separate the method into two modules: (1) sleep module, consisting of the wear detector, TSO detector, sleep/wake detector, and sleep assessment, and (2) scratch module, consisting of the hand movement detector, scratch detector, and scratch assessment. The development procedure for each module of the method is explained in detail below.

Sleep module

The sleep module incorporates several previously published algorithms[11,24,32] in a modular framework to provide measures of sleep quantity[33]. An overview of the processing pipeline can be seen in Fig. 4.

Accelerometer data obtained from the wrists were first down sampled from 100 Hz to 20 Hz. Data were then separated into 24-h segments (12:00 PM to 12:00 PM the next day). Any 24-h periods with less than 6 h of recording time were discarded. This was done to exclude data recorded before and after the official visit period, incomplete data, or data from a misconfigured device (each recording was expected to be approximately 48 h). Periods of non-wear were determined by applying a heuristic rule to the near-body temperature data recorded by the GeneActiv Original device. The temperature data were first processed similarly to the sample-level accelerometer data (5-s rolling median, consecutive 5-s average, rolling 5-min median) so that the wear/non-wear periods would be aligned with the candidate TSO periods. Any candidate period with a temperature value less than 25 °C was considered non-wear. The 25 °C threshold was empirically derived from known wear data during sleep (Supplementary Fig. 7).

Candidate TSO periods for each 24-h segment were then determined using a heuristic approach based on change in arm angle calculated using accelerometer data from the wrist[24]. Any candidate TSO period that was classified as non-wear was excluded. Of the remaining candidate TSO periods in a given 24-h segment, the longest one was chosen as the TSO window. Once the TSO window was identified, predictions of sleep and wake were generated for each 1-min epoch using a heuristic approach[11]. The previously published sleep–wake classification algorithm that was implemented relies on a proprietary method to derive activity counts,

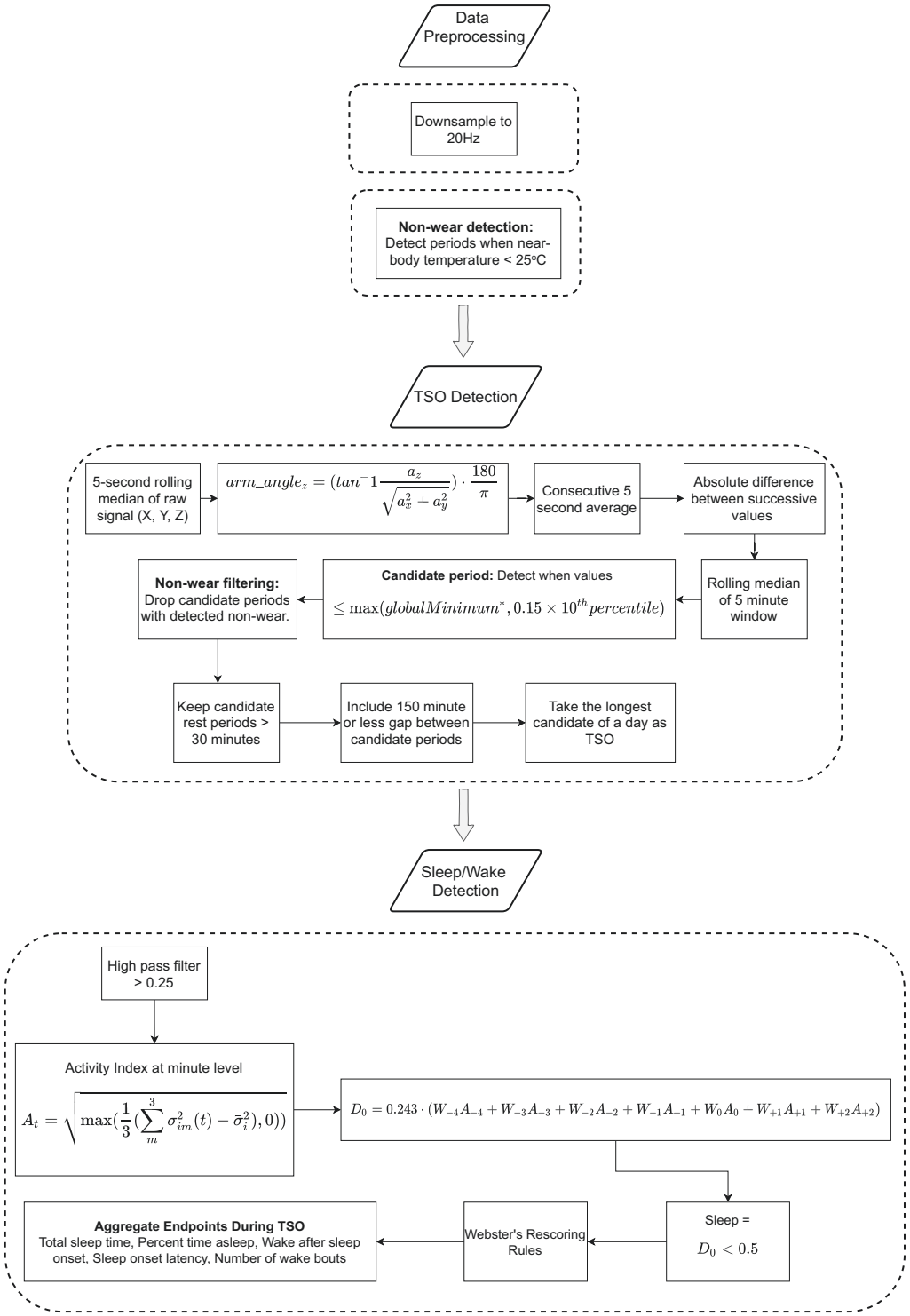

Fig. 4 Overview of the sleep module processing pipeline. The pipeline consists of data preprocessing, total sleep opportunity (TSO) and wear detection, sleep/wake classification, and sleep assessment. *Global minimum is set to 0.1 based on the 25th percentile value of all valid (on-body) data (see Supplementary Fig. 7). Block with a dashed outline provides a detailed illustration of steps in the preceding block with a solid outline.

which were not available for the wearable device used in this study. Therefore, an open-source activity index metric[32] was used as a proxy for activity counts in our implementation. Webster's rescoring rules were applied to the binary sleep–wake predictions to improve specificity[11]. Digital measures of sleep were derived for each 24-h segment by processing the sleep–wake predictions during the determined TSO window as seen in Table 4.

Scratch module

Predictions of nighttime scratch were generated via a two-tiered approach (Fig. 5b). First, the presence of hand movement was determined, and then those periods of hand movement were classified as either scratch or non-scratch events. Sample-level accelerometer data were segmented into 3-s non-overlapping windows within the selected TSO window for a given 24-h period. After testing multiple window lengths (1, 2, and 3 s), we found

Table 4. Description of digital measures output from the analytics solution.

Digital measure	Type	Units	Description
Total sleep opportunity (TSO)	Sleep	Minutes	Largest window of time where sleep is the intended behavior.
Total sleep time (TST)	Sleep	Minutes	Total time spent asleep during the total sleep opportunity window.
Percent time asleep (PTA)	Sleep	Percentage	Percentage of the total sleep opportunity window spent in the sleep state.
Total scratch events	Scratch	Counts	Total scratch bouts during the total sleep opportunity window.
Total scratch duration	Scratch	Minutes	Total time scratching during the total sleep opportunity window.

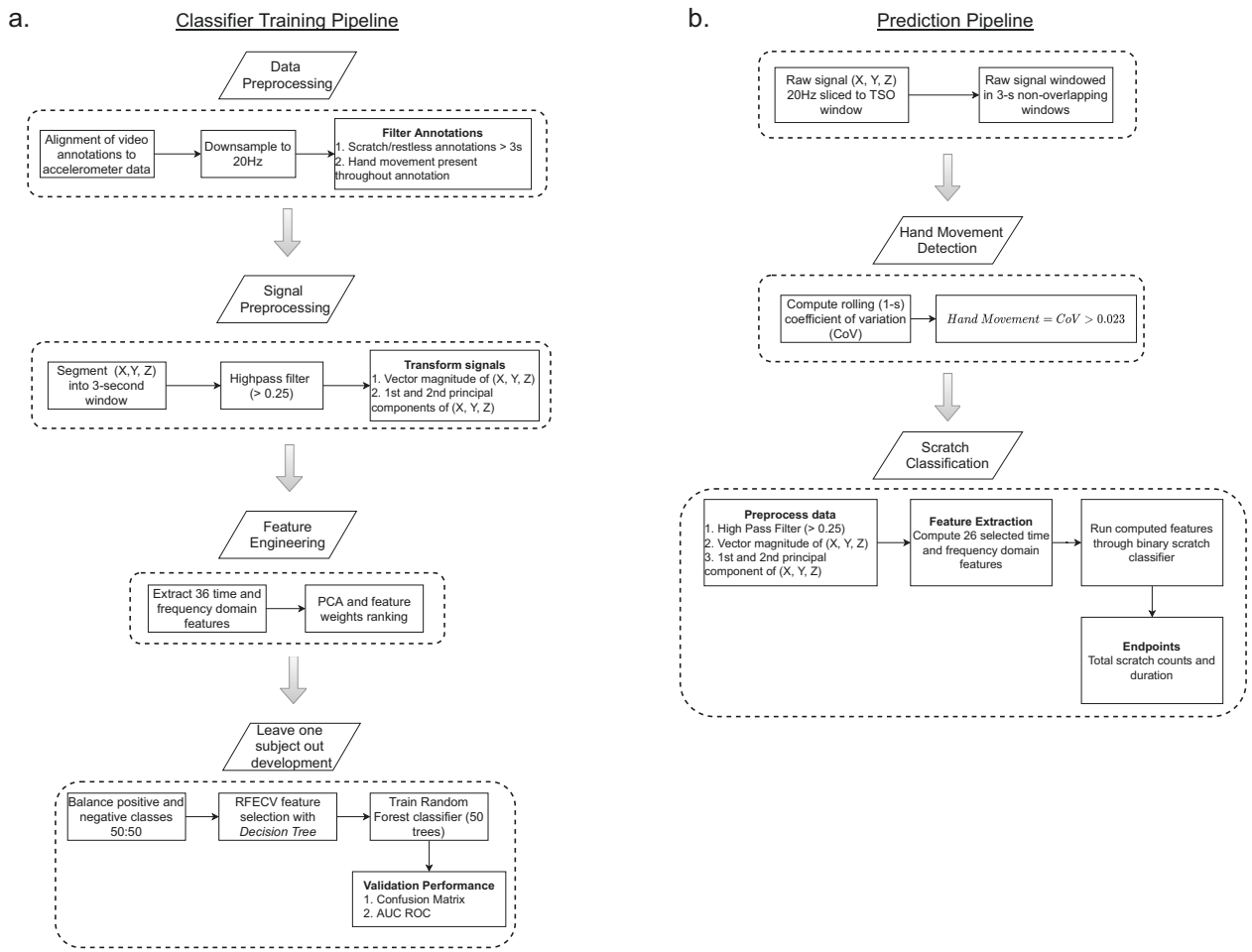

Fig. 5 Overview of the prediction and classifier training pipeline for the scratch module. a Scratch classifier training pipeline, consisting of data preprocessing, signal preprocessing, feature engineering, and leave-one-subject-out validation. **b** Scratch module prediction pipeline, consisting of data preprocessing, hand movement detection, scratch classification, and scratch assessment.

that a 3-s window achieved a good tradeoff between temporal resolution and detection performance. This choice is in agreement with prior work on scratch classification[16,17] and human activity recognition[34]. Each 3-s window was passed through a heuristic hand movement detection algorithm[35] to determine the presence of hand movement. The primary parameter of the hand movement algorithm (a threshold applied to rolling coefficient of variation) was tuned empirically based on our dataset. After testing several threshold values, we selected the 25th percentile of the distribution of coefficient of variation values (0.023) based on our dataset. Scratch classification was performed on a 3-s window only if hand movement was present for the entirety of the window.

We trained a binary ML classifier to detect the presence of scratch. For training, we used all available 3-s windows across both in-clinic visits. To generate labels for training the classifier, instances of nighttime scratch and restless (non-scratch) movements observed via thermal videos of each in-clinic participant visit were annotated by human raters using criteria outlined in Supplementary Note 1. Annotations were performed by two annotators and reviewed by an arbitrator if there were disagreements with regard to timing or behavior classification. Each annotation included metadata about which hand was moving (right, left, or both), the affected body location, as well as severity of scratching (mild, moderate, severe; see Supplementary Note 1 for definitions). Annotations of 3 s or longer were used for training the binary classifier. If an annotation was greater than 3 s, it was segmented into 3 s windows with 50% overlap prior to training to maximize data availability. To ensure that the ground truth was reliable, all annotations were manually time-aligned with the accelerometer data based on a prescribed clap event (participant instructed to clap in front of camera while wearing accelerometer devices) during each in-clinic visit.

The pipeline for training the binary scratch classifier included steps for preprocessing, feature extraction, feature selection, model training, and

model evaluation (Fig. 5a). The preprocessing step generated three processed signals by applying filtering and dimensionality reduction to the sample-level accelerometer data. First, the data were filtered using a first-order Butterworth infinite impulse response (IIR) high-pass filter with a cutoff frequency of 0.25 to remove acceleration due to gravity. Next, to reduce dependence on device orientation, the signal vector magnitude (SVM) ($\sqrt{x^2+y^2+z^2}$) as well as the first (PC1) and second (PC2) principal components of the filtered signal were computed.

A total of 36 time and frequency domain features were then calculated from the preprocessed signals for each window (Supplementary Table 5). Observations were then randomly sampled to balance the positive and negative classes prior to feature selection. We performed feature selection using recursive feature elimination with cross-validation with a decision tree as the estimator[36], which resulted in a total of 26 selected features. A random forest classifier with 50 estimators was then trained based on the selected features. Performance of the binary model was assessed using a leave-one-subject-out validation. We evaluated multiple settings for number of estimators in the random forest classifier (25, 50, 75, and 100) and saw no significant improvement in model performance as we increased the number of estimators past 50.

Digital measures of nighttime scratch were derived by processing the binary scratch predictions during the predicted TSO window for each 24-h segment (Table 4). Total scratch counts were computed by taking the sum of contiguous 3-s bouts of predicted scratch detected from both wrists. Total scratch duration was computed by taking the sum of the duration of all predicted scratch bouts from both wrists.

Statistical methods to measure agreement between sensor-derived measures and reference data

Performance of both the sleep and scratch algorithms were evaluated at the epoch (30 s and 3 s, respectively) and summary endpoint (summary metrics seen in Table 4) levels. With the aim of transitioning to a single device setup in the future, epoch predictions of sleep were derived from both the left and right wrists and were assessed independently against PSG during the second in-clinic visit. A leave-one-subject-out validation was used to assess scratch performance at the epoch level. Annotated scratch and restless movements from the left and right wrists during both in-clinic visits were pooled together to train a single scratch classifier. Conventional classification performance metrics (accuracy, sensitivity, specificity, F1 score, and AUC of the ROC) were used to summarize epoch level performance for both sleep and scratch modules. SHAP[21,22], a game theoretic approach, was used to analyze the importance of the selected features used in the ML scratch classifier.

Epoch predictions of sleep and scratch (with scratch based on leave-one-subject-out model predictions) are summarized for each participant night to obtain endpoint level predictions. Summary statistics of scratch algorithm performance are calculated for different ISGA severities (taken at screening) and sex. Pearson correlation coefficients (along with their p-values), Bland-Altman plots, and limits of agreement were used to assess agreement with reference data on endpoint level metrics throughout. The Bland-Altman limits of agreement describe the 95% confidence interval between the measurements being compared. The agreement between sensor-derived sleep endpoints from the left and right wrists was also assessed. Subsequently, aggregate endpoints of sleep (derived by taking the average between the left and right wrists) were assessed against PSG-derived endpoints of sleep during the second in-clinic visit. Aggregate endpoints of scratch (derived by taking the sum of left and right wrist outputs) during the TSO window predicted by the sleep module were assessed against video annotation derived endpoints of scratch during PSG determined TSO window. Since the distributions of sensor-derived scratch endpoints were right-skewed, log transformation was applied (specifically, $\log(x+1)$ to include possible zero values).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Upon request, and subject to review, Pfizer will provide the data that support the findings of this study. Subject to certain criteria, conditions, and exceptions, Pfizer may also provide access to the related individual anonymized participant data. See https://www.pfizer.com/science/clinical-trials/trial-data-and-results for more information.

CODE AVAILABILITY

The sleep module outlined in this publication is available on GitHub: https://github.com/elyiorgos/sleeppy.

Received: 11 September 2020; Accepted: 15 January 2021;
Published online: 03 March 2021

REFERENCES

1. Kapur, S., Watson, W. & Carr, S. Atopic dermatitis. *Allergy Asthma Clin. Immunol.* **14**, 52 (2018).
2. Ebata, T., Aizawa, H. & Kamide, R. An infrared video camera system to observe nocturnal scratching in atopic dermatitis patients. *J. Dermatol.* **23**, 153–155 (1996).
3. Endo, K., Sano, H., Fukuzumi, T., Adachi, J. & Aoki, T. Objective scratch monitor evaluation of the effect of an antihistamine on nocturnal scratching in atopic dermatitis. *J. Dermatol. Sci.* **22**, 54–61 (1999).
4. Camfferman, D., Kennedy, J. D., Gold, M., Martin, A. J. & Lushington, K. Eczema and sleep and its relationship to daytime functioning in children. *Sleep Med. Rev.* **14**, 359–369 (2010).
5. Oliveira, C. & Torres, T. More than skin deep: the systemic nature of atopic dermatitis. *Eur. J. Dermatol.* **29**, 250–258 (2019).
6. Hon, K.-L. E. et al. Assessing itch in children with atopic dermatitis treated with tacrolimus: objective versus subjective assessment. *Adv. Ther.* **24**, 23–28 (2007).
7. Thorburn, P. T. & Riha, R. L. Skin disorders and sleep in adults: where is the evidence? *Sleep Med. Rev.* **14**, 351–358 (2010).
8. Hanifin, J. M. et al. The eczema area and severity index (EASI): assessment of reliability in atopic dermatitis. *Exp. Dermatol.* **10**, 11–18 (2001).
9. Futamura, M. et al. A systematic review of Investigator Global Assessment (IGA) in atopic dermatitis (AD) trials: many options, no standards. *J. Am. Acad. Dermatol.* **74**, 288–294 (2016).
10. Murray, C. & Rees, J. Are subjective accounts of itch to be relied on? The lack of relation between visual analogue itch scores and actigraphic measures of scratch. *Acta Derm. Venereol.* **91**, 18–23 (2011).
11. Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J. & Gillin, J. C. Automatic sleep/wake identification from wrist activity. *Sleep* **15**, 461–469 (1992).
12. Ancoli-Israel, S. et al. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep* **26**, 342–392 (2003).
13. Van De Water, A. T. M., Holmes, A. & Hurley, D. A. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography - a systematic review. *J. Sleep Res.* **20**, 183–200 (2011).
14. Ancoli-Israel, S. et al. The SBSM guide to actigraphy monitoring: clinical and research applications. *Behav. Sleep Med.* **13**, S4–S38 (2015).
15. Ebata, T., Iwasaki, S., Kamide, R. & Niimura, M. Use of a wrist activity monitor for the measurement of nocturnal scratching in patients with atopic dermatitis. *Br. J. Dermatol.* **144**, 305–309 (2001).
16. Feuerstein, J., Austin, D., Sack, R. & Hayes, T. L. Wrist actigraphy for scratch detection in the presence of confounding activities. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 3652–3655 (IEEE, 2011).
17. Petersen, J., Austin, D., Sack, R. & Hayes, T. L. Actigraphy-based scratch detection using logistic regression. *IEEE J. Biomed. Health Inform.* **17**, 277–283 (2013).
18. Moreau, A. et al. Detection of nocturnal scratching movements in patients with atopic dermatitis using accelerometers and recurrent neural networks. *IEEE J. Biomed. Health Inform.* **22**, 1011–1018 (2018).
19. Lipton, Z. C., Berkowitz, J. & Elkan, C. A critical review of recurrent neural networks for sequence learning. Preprint at https://arxiv.org/abs/1506.00019 (2015).
20. Ikoma, A. et al. Measurement of nocturnal scratching in patients with pruritus using a smartwatch: initial clinical studies with the itch tracker app. *Acta Derm. Venereol.* **99**, 268–273 (2019).
21. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
22. Lundberg, S. M. & Lee, S. -I. A unified approach to interpreting model predictions. In *31st Conference on Neural Infromation Processing Systems* (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
23. Fekedulegn, D. et al. Actigraphy-based assessment of sleep parameters. *Ann. Work Expo. Health* **64**, 350–367 (2020).
24. van Hees, V. T. et al. Estimating sleep parameters using an accelerometer without sleep diary. *Sci. Rep.* **8**, 12975 (2018).
25. Smith, M. P. et al. Emerging methods to objectively assess pruritus in atopic dermatitis. *Dermatol. Ther. (Heidelb)* **9**, 407–420 (2019).