iPROLEPSIS

D3.2 / First report on digital phenotyping for PsA

Editor	Contractual delivery date	Actual delivery date
Apostolidis Georgios (AUTH)	June 2025	October 2025
Deliverable type	**Dissemination level**	**Version - date**
R - Document, report	PU – Public	1.0 - 09/10/2025

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency. Neither the European Union nor the European Health and Digital Executive Agency can be held responsible for them.

Deliverable ID

Project acronym	iPROLEPSIS
Project full title	Psoriatic arthritis inflammation explained through multi-source data analysis guiding a novel personalised digital care ecosystem
Grant Agreement ID	101095697
Deliverable number	D3.2
Deliverable title	First report on digital phenotyping for PsA
Work package	WP3 - Research on PsA inflammation drivers and monitoring
Deliverable type	R - Document, report
Dissemination level	PU – Public
Version - date	1.0 - 09/10/2025
Contractual delivery date	June 2025
Actual delivery date	October 2025
Lead partner	AUTH
Editor	Georgios Apostolidis (AUTH)
Contributors	Vasilis Charisis (AUTH); Eleni Vasileiou (AUTH); Nikos Melanitis (AIN); Dimitrios Konstantinidis (CERTH); Kosmas Dimitropoulos (CERTH); Vasiliki Georgali (CERTH); Katerina Chatziaugeri (CERTH); Vasilis Stamatis (CERTH); Sílvia Reis (PLUX); Bárbara Raposo (PLUX); Rodrigo Braga (PLUX)
Reviewed by	Angleos Karlas (TUM-Med) Jolanda Luime (EMC)
Approved by	Leontios Hadjileontiadis (AUTH, Project Coordinator)
Keywords	Artificial intelligence in rheumatology, Digital biomarkers, Digital phenotyping, Psoriatic arthritis (PsA), Remote disease monitoring, Smart devices-based assessment, Trustworthy AI

Document history

Version	Date	Contributors	Action / status
0.1	15/10/2024	AUTH	Document structure (table of contents) ready
0.2	19/03/2025	AUTH, CERTH, PLUX	Content about the followed approach (Section 2). Preliminary work on joint range of motion analysis (Section 4). Preliminary work on typing dynamics (Section 7). Intestinal motility analysis was finalised by PLUX (Section 10).
0.3	25/07/2025	AUTH, CERTH	Content about the overall dataset (Section 3). The joint range of motion analysis was finalised by CERTH (Section 4). The swollen joint analysis was finalised by AUTH (Section 5). The typing events analysis was finalised by AUTH (Section 7).
0.4	18/08/2025	AUTH, CERTH, AIN	The nail psoriasis analysis was finalised by CERTH (Section 6) The Physical activity level and gait characteristics based on both smartphone (Section 8) and smartwatch (Section 9) IMU data analyses were finalised. The HRV analysis was finalised by AUTH (Section11). The Garmin all-day metrics analysis was finalised by AIN (Section 12).
0.5	09/09/2025	AUTH	Ready for internal review
0.6	19/9/2025	EMC, TUM-Med	Reviewed by Jolanda Luime and Angelos Karlas
0.7	8/10/2025	AUTH, CERTH, AIN	Document revised
1.0	9/10/2025	AUTH	Approved by the Project Coordinator. Submitted to the EC by the Project Coordinator

Contents

List of figures

List of tables

List of abbreviations

AUROC	Area Under Receiver Operating Characteristic curve
BASDAI	Bath Ankylosing Spondylitis Activity Index
BBI	Beat to Beat Intervals
BMI	Body Mass Index
CASPAR	ClASsification for Psoriatic ARthritis
CI	Confidence Interval
CNN	Convolutional Neural Network
COI	Concept Of Interest
COPD	Chronic Obstructive Pulmonary Disease
CRP	C-Reactive Protein
CTTI	Clinical Trials Transformation Initiative
CV	Cross Validation
CWT	Continuous Wavelet Transform
DAPSA	Disease Activity index for Psoriatic Arthritis
dBM	Digital Biomarker
DHT	Digital Health Technologies
DIP	Distal Interphalangeal
DOA	Description Of Action
EGG	Electrogastrograph
ENMO	Euclidean Norm Minus One
FDR	False Discovery Rate
GRC	Global Rate of Change
GSS	Garmin's Stress Score
HAQ	Health Assessment Questionnaire
HRV	Heart Rate Variability
IMU	Inertial Measurement Unit
IQR	Interquartile Range

KPI	Key Performance Indicator
LOSO	Leave-One-Subject-Out
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MAH	Meaningful Aspect of Health
MAPE	Mean Absolute Percentage Error
MCP	Metacarpophalangeal
MDA	Minimal Disease Activity
MMA	Multi-manifold Multi-head Attention
MVPA	Moderate to Vigorous Physical Activity
PEST	Psoriasis Epidemiology Screening Tool
PIP	Proximal Interphalangeal
PoC	Proof of Concept
PPG	Photoplethysmogram
PROs	Patient-Reported Outcomes
PsA	Psoriatic Arthritis
PsAID	Psoriatic Arthritis Impact of Disease
PsO	Psoriasis
QC	Quality Control
RF	Random Forest
RFECV	Recursive Feature Elimination with Cross-Validation
RMSE	Root Mean Square Error
SJC	Swollen Joints Count
TAI	Trustworthy Artificial Intelligence
TJC	Tender Joint Count
URD	Unobtrusive Remote Disease
VAS	Visual Analogue Scale

Executive summary

This deliverable (D3.2) presents the first report on digital phenotyping for psoriatic arthritis (PsA) within the iPROLEPSIS project. The work focuses on developing and validating digital biomarkers (dBMs) derived from smartphone, smartwatch, and smartbelt sensors, using data from the PDPID study cohort.

Traditional clinical and patient-reported assessments of PsA are limited by episodic measurement, subjectivity, and survey fatigue. By contrast, digital phenotyping offers the possibility of continuous, objective, and unobtrusive monitoring of disease activity in real-world settings. The aim of Task 3.2 is to establish and validate dBMs that capture key PsA-related parameters such as joint range of motion, swelling, fine motor function, physical activity, gait, sleep quality, gastrointestinal activity, cardiovascular response to disease exacerbation, and psychological stress.

The report documents:

- The digital health technology development framework adopted, aligned with iPROLEPSIS's Trustworthy AI principles and CTTI guidelines.
- The structure and characteristics of the first PDPID dataset, combining digital and clinical data.
- The development and preliminary validation of multiple dBMs, including joint mobility measures, nail psoriasis detection, typing dynamics, gait and activity patterns based on smartphone and smartwatch accelerometer data, heart rate variability, gastrointestinal motility, and all-day metrics.
- Early findings indicating correlations between selected digital features and clinical or patient-reported outcomes, though limited statistical power due to the current small sample size and single baseline visit restricts definitive conclusions at this stage. These findings provide a foundation for more robust results as larger sample size from the PDPID study become available.

This deliverable establishes the methodological foundation for future iterations (D3.6, D3.8), where larger and longitudinal datasets will enable stronger validation, feature refinement, and predictive modelling. The ultimate goal is to integrate validated dBMs into the miPROLEPSIS ecosystem, supporting personalised care and earlier detection of disease flares.

1 Introduction

Developing and integrating digital biomarkers (dBMs) into clinical pathways of psoriatic arthritis (PsA) holds significant promise for transforming disease monitoring and management. Traditional clinical assessments, although considered the gold standard, are inherently limited by their episodic nature (i.e. occur in irregular and rare visits), subjectivity, and dependence on the availability and expertise of healthcare providers. In contrast, dBMs, which are derived from smartphones, wearables, and other smart devices, offer a novel, patient-centred approach to disease evaluation. These technologies can passively and actively capture continuous, objective data on movement patterns, joint function, pain levels, fatigue, and morning stiffness, thereby providing a granular and valid view of disease activity in real-world settings.

The ability to track these parameters longitudinally enables early detection of subtle changes in inflammatory activity that may precede clinical flares or structural damage. Moreover, dBMs facilitate remote monitoring, which is particularly valuable in the context of telemedicine and decentralised clinical trials. This could lead to a paradigm shift from reactive, episodic care toward a more proactive and personalised disease management model.

Also, they can enable predictive modelling of disease trajectories, stratification of patients based on dynamic symptom profiles. However, realising this potential requires robust validation studies, standardisation of data collection protocols, and careful attention to data privacy and patient engagement.

1.1 Document scope

Following data collection in the PDPID study (T5.2), task T3.2 is responsible for implementing data processing pipelines that generate dBMs. These dBMs capture disease-related parameters such as disease activity status, pain, joint range of motion, and fatigue. Each type of dBM targets specific parameters depending on the digital sensors that produce the corresponding raw data.

In the PDPID study, data are collected from multiple sources:

- **Smartphone sensors:** videos, photos, typing events, accelerometer/gyroscope data
- **Smartwatch sensors:** accelerometer, beat-to-beat intervals
- **Smart belt:** electrogastrography, bowel sounds

The purpose of this report is to document the methodology and validation of each dBM category, organised by sensor type. This is the first of three planned reports (followed by D3.6 and D3.8). It focuses on data collected from the first cohort of patients recruited during the initial six months of the study. The dBMs developed in T3.2 will support disease modelling (T3.5) and will be integrated into the miPROLEPSIS patient and HCP app (T4.3).

According to Description of Action the activities of T3.2 are associated with objective *O4*

“Set up a system of digital biomarkers for objective, privacy-aware tracking of key PsA risk and progression drivers in daily living (Addressed in WP2, WP3, WP5).”

hence, the corresponding KPIs are

- Development, implementation and verification **of ≥5 new dBM**s for motor, physical, emotional, nutritional, gastrointestinal activity, sleep quality and functional level tracking

- Correlation of dBMs with relevant clinical scores **>0.70 (*p*<0.05)**

1.2 Document structure

The document is structured as follows:

- **Section 2** presents the adopted framework for technology development, including the concepts, principles, and approaches considered in the design and implementation of dBMs.
- **Section 3** provides a brief description of the dataset, covering both digital and clinical data.
- **Remaining sections (Section 4-12)** describe each processing pipeline and its corresponding analysis. Each section follows the same structure: *Background and Objectives*, *Methods*, *Results*, *Discussion*, and a summary of next steps leading to the submission of the following deliverable.
- **Section 13** concludes the report.

2 Digital health technology development framework

2.1 Introduction

This section delineates the framework underpinning the design and development of digital measures within Task T3.2. It begins with a brief overview of the PDPID study rationale and objectives to establish context. Subsequently, it outlines the relevant requirements derived from the iPROLEPSIS Trustworthy AI (TAI) framework. In addition, it introduces key recommendations from the Clinical Trials Transformation Initiative (CTTI), which serve as foundational principles for the definition of digital measures and the specification of associated assessment steps. The final subsection describes the implementation of these recommendations, resulting in the formal specification of the digital measures.

2.2 PDPID study rationale and aim

In PsA, both disease activity and patients' perceptions significantly influence treatment decisions. Disease activity is currently assessed using clinical measures and patient-reported outcomes (PROs), which require regular outpatient visits. While PROs enable remote monitoring, their long-term use is limited by survey fatigue, and no standardised method exists for unobtrusive remote monitoring.

The rationale of the PDPID study is that the widespread use of smartphones and smartwatches offers a unique opportunity to develop Unobtrusive Remote Disease (URD) activity monitoring using passive behavioural data from device sensors. It is hypothesized that higher disease activity in PsA results in measurable changes in key disease-related characteristics, as captured by these devices. Digital measures may provide insights complimentary to clinical assessments and PROs, while also capturing additional symptoms, such as fatigue and sleep disturbances, enhancing the ability to differentiate between disease states. We also anticipate that patients will view smartphone-based monitoring as a privacy-conscious and acceptable way to gain greater insight into their condition.

The aim of the study is to develop an unobtrusive and affordable digital tool capable of detecting changes in disease activity, including flares. This technology may support applications that help patients better understand their disease status, determine the need for follow-up visits, and identify potential flare triggers ultimately contributing to improved disease control.

2.3 Relation to iPROLEPSIS trustworthy AI (TAI) framework

The iPROLEPSIS TAI framework, detailed in Deliverable D2.3, is based on the EU Ethics Guidelines for Trustworthy AI[1] and tailored to the specific needs of the iPROLEPSIS project. It outlines recommendations across seven key areas of the AI lifecycle:

1. Human Agency and Oversight,
1. Technical Robustness and Safety,
2. Privacy and Data Governance,
3. Transparency,
4. Diversity, Non-discrimination and Fairness,
5. Societal and Environmental Wellbeing, and
6. Accountability.

[1] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Task T3.2 is responsible for developing algorithmic techniques that produce digital measures, i.e., reliable indicators of disease-related characteristics. These measures are foundational to most AI outcomes within the iPROLEPSIS ecosystem, either directly or indirectly. Therefore, T3.2 is primarily aligned with recommendations under "Technical Robustness and Safety", including:

- Test whether specific contexts or conditions need to be taken into account to ensure reproducibility.
- Put in place verification and validation methods and documentation (e.g., logging) to evaluate and ensure different aspects of the system's reliability and reproducibility.
- Clearly document and operationalize processes for the testing and verification of the reliability and reproducibility of the AI system.

To meet these requirements, we follow best practices for data evaluation and model performance assessment. Additionally, all model-based techniques are documented using model cards, an industry-standard approach to support transparency and reproducibility. Our model cards follow a hybrid methodology, combining elements from Mitchell et al. (2019) and the Hugging Face Model Card Guidebook (Ozoani et al., 2022). The completed model cards are included in the Appendix I Model cards.

2.4 Recommendations of Clinical Trials Transformation Initiative

The CTTI, a public-private partnership established in 2007 by the U.S. Food and Drug Administration and Duke University, aims to improve the quality and efficiency of clinical trials. CTTI has issued several recommendations on the use of digital health technologies (DHTs) in clinical research. In Task T3.2, we follow the approach outlined in CTTI's document "**Steps for Novel Endpoint Development, with Suggested Approaches and Considerations**"[2]. This iterative, two-phase process includes:

a. Selecting the outcome assessment and patient population
b. Developing the outcome assessment into a clinical trial endpoint, including technology validation

The key steps in selecting the outcome assessment and patient population are:

a1. Describe the target population, including subpopulations, to ***define the context of use***. Begin with a narrow context, which can be broadened later.
a2. ***Identify a meaningful aspect of health (MAH)*** that matters to patients and is not well-assessed. Patient and caregiver input is essential.
a3. ***Define the concept of interest (COI)***, a measurable element of the MAH. A COI should reflect changes in the MAH. While available DHTs may be considered, technology selection should not occur at this stage. Consensus definitions for DHT-enabled COIs must be established.
a4. ***Select the specific outcome assessment*** that best represents the COI. Evaluate whether a DHT-derived measure offers added value over existing assessments, incorporating input from patients, caregivers, and clinical experts.

Key steps in developing the outcome assessment into a clinical trial endpoint are:

b1. ***Conduct technical verification*** to ensure the DHT meets performance specifications (e.g., accuracy, precision) over time and conditions. Consistency is critical, especially when multiple DHTs are used.

[2] https://ctti-clinicaltrials.org/wp-content/uploads/2021/06/CTTI_Novel_Endpoints_Detailed_Steps.pdf

b2. ***Perform technical validation*** in a representative population, comparing the novel endpoint to established measures. This may involve adding it as an exploratory endpoint in ongoing studies. While high correlation with existing measures is not always expected, reliability and strategies to mitigate compliance-related variability should be demonstrated.

b3. ***Assess usability*** to ensure participant tolerability and acceptability. Usability testing must confirm participants can use the DHT as intended.

Steps a1–a4 are implemented in Section 2.5 to conclude the digital measures and their assessment for this task. Steps b1–b2 are used to structure the reporting of each digital measure category in the subsequent sections. Step b3 will be applied in the final deliverable of T3.2 (D3.9), which will include usability data.

It should be noted that we did not adopt the full set of CTTI recommendations, but only those relevant to the objectives of task T3.2. Additional procedures suggested by CTTI for the design of digital measures are, in the case of iPROLEPSIS, explicitly defined in the DoA.

Additionally, study design (WP5) and user research (WP2) are handled as separate tasks, with outcomes reported in their respective deliverables.

2.5 iPROLEPSIS digital phenotyping measures

This section outlines the key concepts involved in the development of digital technologies for measuring disease-related constructs in the context of digital biomarkers. These concepts (**Table 1**), as defined in the relevant literature (Walton, 2015; Houts, 2020; Goldsack, 2021), are structured in a hierarchical organisation ranging from high-level conceptual definitions to specific digital measures and clinically meaningful endpoints. Such organisation is essential for the systematic analysis of each use case.

Table 1 Key concepts in mapping participant experience to digital measures (Reproduced from (Goldsack, 2021)).

Concept	Definition	Patient input	Consideration
Meaningful Aspect of Health (MAH)	Aspect of a disease that a patient: (1) does not want to become worse, (2) wants to improve, or (3) wants to prevent	What do you wish that you could do, but your condition prevents you from doing it? What part of your life is most frustratingly impacted by your condition?	May be shared across some conditions and diseases
Concept of Interest (COI)	Simplified or narrowed element that can be practically measured	What are the symptoms that most impact your ability to do these activities?	Patients may have different symptoms. Symptoms may vary over time. Symptom relevance may vary over time.
Measure	Specific measurable characteristics	Do these measures make sense to you?	Measures may be relevant to multiple symptoms. Assess technical specifications of sensor and whether it is suitable for measuring this outcome in this population.

Concept	Definition	Patient input	Consideration
Endpoint	Precisely defined, statistically analysed variables with demonstrated relevance to clinical benefit	How much change do we need to see in this symptom before it really starts to make a positive difference in your life?	Sensors may support multiple measures & endpoints

In psoriatic arthritis (PsA), the primary MAH is ***the ability to perform ambulatory activities*** (McGagh, 2024).

Key COIs in PsA include:

1. Joint pain, stiffness, and swelling
2. Physical functioning and activity levels
3. Fatigue
4. Sleep disturbance

A secondary MAH concerns gastrointestinal health, particularly ***dietary restrictions and reduced confidence in social settings***, with gastrointestinal activity levels identified as the corresponding COI.

To represent these COIs, novel digital measures are developed by processing data from diverse sources, including:

- Smartphone camera, keyboard, and inertial measurement units (IMUs), such as accelerometers and gyroscopes.
- Smartwatch data, including raw beat-to-beat intervals and accelerometer signals, as well as all-day metrics from Garmin's proprietary algorithms (e.g., body battery, stress level, sleep, and wellness).

The effectiveness of these digital measures is validated by assessing their associative ability with target clinical variables. **Table 2** summarizes the digital measures, input data sources, and corresponding clinical targets.

Table 2 Summary of the developed digital measures along with the respective data sources and targeted clinical and patient-reported outcomes.

Digital measure(s)	Data source	Target(s)
Joint range of motion scores	Videos captured from smartphone's camera	• Joint stiffness and pain • Disease activity scores (MDA, PASDAS, DAPSA) • Disease change (GRCQ)
Joint swelling score	Photographs captured from smartphone's camera	• Swollen joint count (SJC66)
Nail psoriasis detection	Photographs captured from smartphone's camera	• CASPAR-Nail dystrophy

Digital measure(s)	Data source	Target(s)
Typing dynamics	Typing events from smartphone's keyboard	• Disease activity scores (DAPSA, PASDAS) • Pain (PsAID-1) • Fatigue (PsAID-2) • Hand function (66/68JC)
Physical activity level (sedentary/light/MVPA) and gait characteristics (Coarse type)	Data from smartphone's IMU sensors	• Disease activity scores (MDA, DAPSA, PASDAS) • Disease impact scores (PsAID) • Pain (PsAID-1) • Lower-extremity pain (66/68JC)
Physical activity level (sedentary/light/MVPA) and gait characteristics (Fine type)	Data from smartwatch's accelerometer sensors	• Disease activity scores (MDA, DAPSA, PASDAS) • Disease impact scores (PsAID) • Hand function (66/68JC) • Pain (PsAID-1) • Sleep (PsAID-7) • Stiffness (BASDAI-5)
Intestinal motility scores	EGG and bowel sound signals acquired from PLUX's smart belt	• Disease activity (MDA) • Gastrointestinal health status (Crohn's Disease Activity Index, Mayo score)
Digital inflammation response index	BBI from smartwatch's PPG sensor(s)	• DAPSA-CRP • Disease activity scores (MDA, PASDAS, DAPSA) • Clinician-assessed flare • Patient-perceived flare
Garmin's all-day metrics: • **Body battery** • **Stress level** • **Sleep measures** • **Wellness measures**	Produced by the proprietary algorithms of Garmin processing the sensors of Garmin's smartwatches.	• Disease activity scores (MDA, DAPSA, PASDAS) • Pain (PsAID-1) • Fatigue (PsAID-2) • Sleep (PsAID-7) • Stiffness (BASDAI-5)

3 PDPID preliminary dataset

The PDPID preliminary dataset consists of 111 adults with PsA who were recruited between September 2024 (the initiation of the study) and April 2025. Demographic and clinical characteristics from participants' first clinic visit (T0) are presented in **Table 3** and **Table 4**, respectively. Values are reported as counts, medians with interquartile ranges (IQR) or percentages, depending on the variable type. The percentage of missing values is also provided.

Table 3 Demographic characteristics for the first PDPID study cohort at T0 (PDPID study cohort v1).

Characteristic	Value (Count or Median [IQR] or (%))	Missing (%)
No. Participants	111	-
NL	29 (26%)	-
UK	32 (29%)	-
GR	50 (45%)	-
Age	55 [45, 61]	0%
Sex (male)	57 (51%)	0%
Disease duration (years)	8 [4, 16]	9%
PsO History	69 (79%)	21%
BMI	28 [25, 34]	5%
BMI ≥ 30	40 (38%)	5%
Smoking	35 (32%)	1%
Education (years)	13 [12, 16]	1%
Employed (yes)	42 (51%)	26%
Handedness	110	1%
Right-handed	94 (85%)	-
Left-handed	14 (13%)	-
Ambidextrousness	2 (2%)	-
Typing tools usage (yes)*	29 (26%)	1%
**Typing tools usage (yes): Participant reported using a handheld input tool—e.g., a stylus/digital pen—to type on the smartphone's virtual keyboard (vs. finger-only typing).*		

Table 4 Clinical characteristics and PROs for the first PDPID study at T0.

Characteristic	Value (Count or Median [IQR] or (%))	Missing (%)
Clinical assessment		
Disease phenotype	78	30%
Polyarticular	42 (54%)	-
Oligoarticular	18 (23%)	-
Monoarticular	5 (6%)	-

Axial	9 (12%)	-
Dactylitis	2 (3%)	-
Entheseal	2 (3%)	-
SJC	0 [0,1]	0%
Upper body (>0)*	29 (26%)	-
Lower body (>0)*	16 (14%)	-
Hands (>0)*	28 (25%)	-
TJC	1 [0,7]	0%
Upper body (>0)*	54 (49%)	-
Lower body (>0)*	50 (45%)	-
Hands (>0)*	44 (40%)	-
CRP (mg/dL)	0.4 [0.2, 1]	23%
CRP (>3 mg/dL)*	9 (11%)	23%
DAPSA	41 [21, 101]	49%
PASDAS	3.4 [2.3, 4]	43%
MDA (yes)	35 (52%)	39%
Flare (physician)	20 (19%)	4%
Nail dystrophy*	49 (49%)	9%
PROs		
Flare (patient)	12 (15%)	27%
HAQ	0.6 [0.1, 1.1]	25%
PsAID12	2 [1, 4]	26%
Pain (PsAID12)	2 [1, 5]	26%
Fatigue (PsAID12)	2 [1, 4]	26%
Sleep (PsAID12)	2 [0,4]	26%
VAS Pain	22 [9, 50]	25%
Stiffness	3 [1, 6]	1%

**Upper body: sternoclavicular, acromioclavicular, glenohumeral (shoulders), elbows, wrists, and hand joints (MCP, PIP, DIP).*
**Lower body: hips (tender only), knees, ankles, tarsus/midfoot, metatarsophalangeal (MTP), and toe interphalangeal/PIP joints.*
**Hands: wrists and hand joints (MCP, PIP, DIP).*
**CRP (>3mg/dL): Inflammatory flare.*
**Nail dystrophy (CASPAR): Typical psoriatic nail dystrophy (onycholysis, pitting, and hyperkeratosis on exam).*

Regarding the digital data, they are collected using miPROLEPSIS app, which supports the continuous passive recording from smartphone IMU (accelerometer and gyroscope), Garmin smartwatch accelerometers and beat-to-beat intervals, and smartphone keyboard typing events. It also supports two types of active tests: (1) photo and (2) video tests.

For this analysis, only data from the first two weeks after the T0 visit are used, as clinical characteristics during this period are considered stable. Data availability varies significantly across participants. In the following sections, where the analysis of each modality is presented, more details about the respective data are provided.

4 Joint range-of-motion scores

4.1 Background and objectives

Detecting fine motor skills is an important aspect of the assessment of the severity of PsA in patients. Joints are commonly affected by PsA, leading to tenderness or swelling that can significantly reduce the range-of-motion in patients, posing challenges to their everyday lives. Therefore, simple movement tests can be designed and used as a useful tool for the estimation of patients' range-of-motion and consequently for the assessment of the disease status and progression.

In the framework of the iPROLEPSIS project, a set of 4 hand and 2 body exercises were designed, namely small and full fist formation, wrist extension and flexion, and upper and lower body movement. In this way, a tool was developed that employs the above 6 simple exercises to evaluate the physical functioning of a large number of joints in the human body, ranging from fingers and wrists to shoulders, ankles and knees. The tool allows a user to perform these exercises using the smartphone's front camera. Then, the recorded videos are processed using a well-known skeletal extraction algorithm, the angles between joints are computed to assess the range-of-motion and a set of scores that indicate the difficulty of the user in performing the exercises (which can be attributed to the severity of psoriatic arthritis) is provided. A key advantage of the proposed range-of-motion assessment approach is its non-intrusive nature, allowing patients to perform exercises naturally and comfortably in their own environment. Unlike wearable sensors, which can be cumbersome, the proposed approach ensures seamless movement and enhances patient compliance, while still providing accurate analysis results.

The objective of the proposed approach is to collect novel digital biomarkers that are related with joint range-of-motion and perform a correlation with clinical measurements and ePROs. Specifically, swollen or tender joints recorded during clinical visits can be used as clinical targets and search for associations with the computed digital biomarkers. Additionally, composite scores, such as MDA, PASDAS, DAPSA, PEST and PSAID can also be employed as clinical targets. Moreover, questionnaire responses, such as those in GRCQ can also serve as target outputs to measure changes in physical functioning.

4.2 Methods

4.2.1 Dataset(s)

There are no publicly available datasets of video exercises or skeletal data related to PsA or other arthritic diseases. Through the video capturing and analysis procedure of the proposed approach, the coordinates of the skeletal joints are collected and saved to form a new dataset of hand and body joint motion data. The new dataset is stored on the Cloud Data Management System of the iPROLEPSIS project, and it aims to be one of the few large datasets collected from real PsA patients in various countries to be used for the assessment of the range-of-motion in PsA patients and to foster the identification of correlations with clinical measurements and the progression of the disease. The format of the newly formed dataset is presented below and it slightly differs based on whether the test is related with the movement of hands (**Table 5**) or body (**Table 6**).

Table 5 Hand joint motion data stored in the dataset.

Variable	Description
patientid	Pseudonymized ID of the patient
start_timestamp	The time the specific exercise started in ms (unix time)
end_timestamp	The time the specific exercise ended in ms (unix time)
assessment_type	The type of exercise performed, being: RoM_Hands_Small_Fist, RoM_Hands_Large_Fist, RoM_Hands_Wrist_Flexion or RoM_Hands_Wrist_Extension
completed	Boolean flag of whether the exercise was successfully performed
hand_joints	Nested list of hand joint coordinates with size (L, N, 3), where L is the temporal dimension, N = 21 hand joints and 3 are the coordinates in space (x,y,z)
handedness	Nested list of confidence scores of the skeletal extraction algorithm per hand with size (L, H), where L is the temporal dimension and H can be either an empty list (no hands were detected) or contain 1 or 2 elements that represent the left and right hand, as well as a confidence score for each hand
joint_indices	List of names of the 21 hand joints

Table 6 Body joint motion data stored in the dataset.

Variable	Description
patientid	Pseudonymized ID of the patient
start_timestamp	The time the specific exercise started in ms (unix time)
end_timestamp	The time the specific exercise ended in ms (unix time)
assessment_type	The type of exercise performed, being RoM_Body_Lower or RoM_Body_Upper
completed	Boolean flag of whether the exercise was successfully performed
body_joints	Nested list of body joint coordinates with size (L, N, 3), where L is the temporal dimension, N = 33 body joints and 3 are the coordinates in space (x,y,z)
joint_indices	List of names of the 33 body joints

4.2.2 Development

The pipeline that was followed for the assessment of the joint range-of-motion from the video exercises is presented in **Figure 1**.

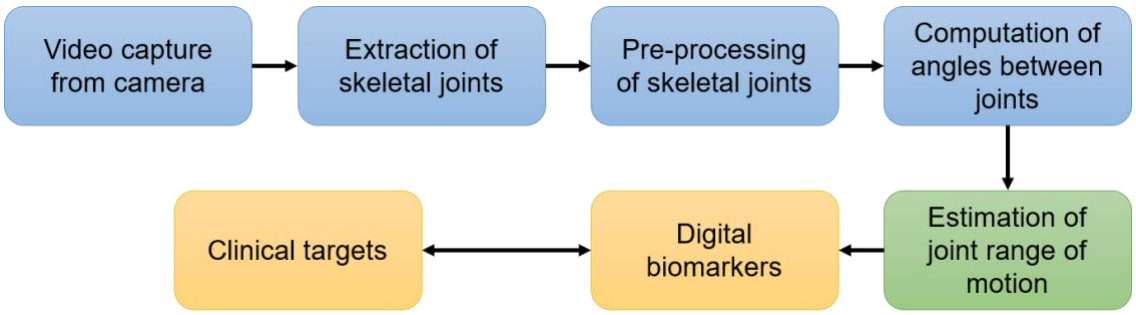


Figure 1 Pipeline of the proposed approach for the assessment of joint range-of-motion through video exercises.

Before presenting the main steps that comprise the pipeline of the proposed approach, a preliminary procedure was required for the definition of the movement exercises and the type and the number of joints for which the range-of-motion should be assessed. Physicians from the project's consortium were employed to propose candidate movement exercises that were then revised by the clinicians of the project's consortium. From this procedure, a set of 6 hand and body movement exercises was designed and executed by FMH, and snapshots from these exercises are presented in **Figure 2**.

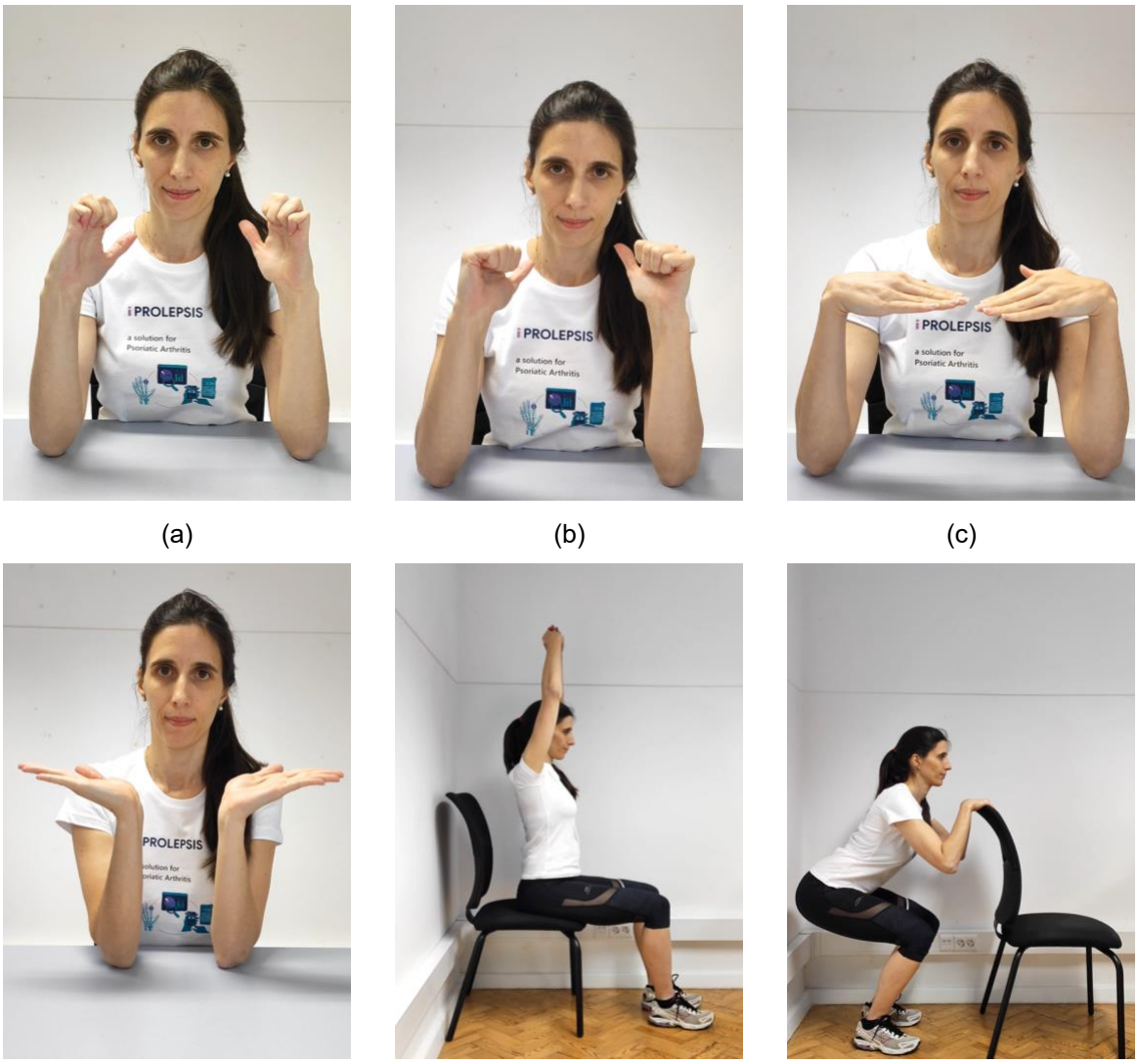


Figure 2 Snapshots from (a) small fist, (b) full fist, (c) wrist flexion, (d) wrist extension, (e) upper body and (f) lower body movement exercises.

The joints whose range-of-motion are assessed for each movement exercise are presented in **Table 7**.

Table 7 Joints for assessment for each movement exercise.

Movement test	Joints for assessment
Small fist	Distal interphalangeal (DIP) and proximal interphalangeal (PIP) joints of the 4 fingers (except for thumb) of each hand
Large fist	DIP, PIP and metacarpophalangeal (MCP) joints of the 4 fingers (except for thumb) of each hand
Wrist flexion	Wrist joint of each hand
Wrist extension	Wrist joint of each hand
Upper body movement	Shoulder joint of each hand
Lower body movement	Knee and ankle joints of each leg

The steps of the processing pipeline are described below:

Video capture from camera: Given the definition of the video movement exercises, the first step of the proposed approach involves the capturing of the videos through a dedicated mobile application utilizing a smartphone camera. Patients are given detailed video and text instructions on how to perform each exercise, as well as the distance from the smartphone camera that should be maintained to ensure optimal video recording results.

Extraction of skeletal joints: The captured videos that represent the movement exercises are processed using Google's AI models for hand landmark detection (Google AI, 2024a) and pose landmark detection (Google AI, 2024b) that are components of the Mediapipe software. The hand landmark detection model is employed for the assessment of the hand-related movement exercises (i.e., small/large fist and wrist flexion/extension), while the pose landmark detection model is employed for the assessment of the body-related movement exercises (i.e., upper/lower body movements). Using these two models, the position (coordinates in the 3D space) of the joints of the hands and body parts are computed for each frame of the video sequence and are sent to the Cloud Data Management System of the iPROLEPSIS project.

Pre-processing of skeletal joints: The aim of the skeletal joint pre-processing procedure is to prepare the data for accurate angle computations by removing outliers and missing values. Initially, each temporal joint sequence is analysed to identify joints that are not detected in one or more temporal frames. When such a sequence is found, a cubic spline interpolation is performed on the joint sequence to fill in the missing values. This step ensures that there are no gaps in the temporal sequences of the joints that can lead to errors in the angle computations. Subsequently, a moving average processing step with a sliding window of 5 frames is applied to each temporal joint sequence in order to smooth out potential extreme values or outliers. In this way, sudden spikes on the values of the joint coordinates are smoothed out, ensuring that there are no abrupt changes in the computed angles.

Computation of angles between joints: This step involves the computation of angles in the 3D space for the joints of interest (**Table 7**). The angles are defined as the ones between the lines formed by the specific joints and the previous/next joints in the skeletal sequence. For instance, the angle of the PIP joint of the index finger is defined as the angle between the line formed by the PIP and the DIP joints of the index finger and the line formed by the PIP and

the MCP joints of the index finger (**Figure 3**). On the other hand, the angle of the shoulder joint is defined as the angle between the line formed by the shoulder and the elbow joints and the line formed by the shoulder and the hip joints. The aforementioned procedure is performed for each temporal frame of the video sequence, leading to the creation of a temporal sequence of angles representing the evolution of the motion for each joint. Finally, before proceeding to the estimation of joint range of motion, a moving average operation with a window of 5 frames is applied to smooth the sequences of angles and remove spikes or troughs.

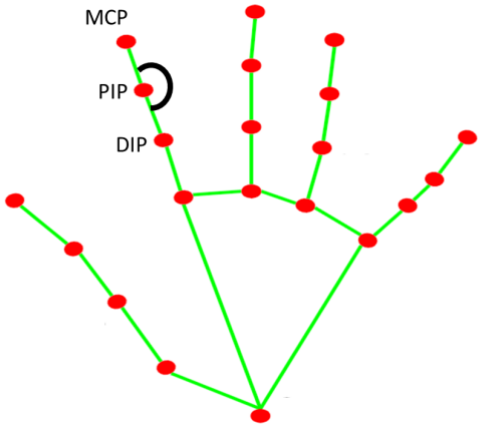


Figure 3 Computation of angle of PIP joint of index finger as the angle between the lines formed by the MCP and PIP joints and the PIP and DIP joints.

Estimation of joint range-of-motion (RoM): To estimate the RoM of each joint, the temporal sequence of angles for the specific joint $\{a_i\}_{i=1}^{T}$ is analyzed and the minimum and maximum angles are identified. The difference between these two angles represents the RoM for the specific joint, as it quantifies how much a joint can be extended and flexed. As a result, the observed RoM of a joint J is equal to:

$$RoM_J^{obs} = \max\left(\{a_i\}_{i=1}^{T}\right) - min\left(\{a_i\}_{i=1}^{T}\right)$$

Digital biomarkers: The developed digital biomarkers that correspond to the joint RoM are defined as scores representing the deviation of the patients' joint RoM from the joint RoM of healthy individuals. As a result, higher scores correspond to larger deviations from the healthy population, which can be attributed to higher severity of psoriatic arthritis. To acquire the RoM for the joints of the healthy population, an extensive literature review was performed and the RoM for all important joints in the form of mean and standard deviation was identified. The joint RoM for hand (Ngeo et al., 2014), wrist (Than et al., 2012), shoulder (Gill et al., 2020), and knee and ankle (Lorenzetti et al., 2018) joints is presented in **Table 8**.

Table 8 Joint RoM of healthy individuals found in the literature in the form of mean ± std.

Name of joint	Joint RoM of healthy individuals (mean ± std) in degrees
Index finger MCP	62.29 ± 14.27
Index finger DIP	72.55 ± 16.87
Index finger PIP	45.51 ± 22.25
Middle finger MCP	69.39 ± 11.61
Middle finger DIP	80.07 ± 16.52

Name of joint	Joint RoM of healthy individuals (mean ± std) in degrees
Middle finger PIP	57.07 ± 22.42
Ring finger MCP	62.51 ± 11.04
Ring finger DIP	88.58 ± 14.21
Ring finger PIP	58.94 ± 19.99
Pinky finger MCP	69.27 ± 6.07
Pinky finger DIP	72.94 ± 14.27
Pinky finger PIP	84.54 ± 12.6
Wrist joint flexion	68.3 ± 10.6
Wrist joint extension	68.2 ± 10.3
Left shoulder	157.1 ± 18.2
Right shoulder	158.5 ± 19.1
Knee	91 ± 12
Ankle dorsiflexion	29.6 ± 8
Ankle plantar flexion	26.5 ± 7.2

Given the observed RoM of a joint RoM_J^{obs} that is computed during a movement exercise and the mean μ_J^{ref} and standard deviation σ_J^{ref} RoM of the same joint from the literature, the RoM score RS_J for the specific joint J is defined as the standard score or z-score (DeVore, 2017) that measures the deviation of the observed RoM from the distribution of joint RoM in the healthy population, as shown below:

$$RS_J = \frac{RoM_J^{obs} - \mu_J^{ref}}{\sigma_J^{ref}}$$

In the case that the patient who performs the movement exercises achieves a better performance than the healthy population (i.e., negative RS_J due to higher observable RoM_J^{obs} for a joint than μ_J^{ref}), then the RS_J for the specific joint is set to zero. As a result, the computed RoM scores / digital biomarkers of the proposed approach can only take positive values, with a score of 0 denoting the optimal movement exercise performance and higher values indicating issues with performing the movement exercises. As the difficulty of performing a movement exercise increases due to joint stiffness or pain, so does the computed RS_J, indicating the severity of a patient's PsA illness. Following the aforementioned procedure, a RoM score is calculated for each joint that needs to be assessed. Additionally, RoM scores are computed for joints that are assessed in more than one movement exercises, as well as for clusters of joints. Joints that belong to the first case are the finger DIP and PIP joints that are assessed during the small and large fist exercises and the wrists that are assessed during the wrist flexion and extension exercises. On the other hand, joints that belong to the second case include the left and right finger joints (i.e., index, middle, ring and pinky), the entire left and right hands and the full body. The additional RoM scores, computed for the

aforementioned two cases of joints, are considered composite scores and are computed using the following weighted average technique based on standard deviation:

$$RS_J = \frac{\sum_{i=1}^{n} w_J^i RS_J^i}{\sum_{i=1}^{n} w_J^i}$$

In the equation above, n is the number of exercises or joints that need to be aggregated to compute the composite RoM scores and $w_J^i = \frac{1}{\left(\sigma_J^{ref}\right)_i^2}$ is the weight of the specific RoM score based on the square of its standard deviation. For instance, given the RoM score for the wrist joint during the wrist flexion exercise RS_{wrist}^{WF} and the RoM score for the same joint during the wrist extension exercise RS_{wrist}^{WE}, the composite RoM score for the wrist joint is defined as:

$$RS_{wrist} = \frac{\frac{1}{\left(\sigma_{wrist}^{ref}\right)_{WF}^2} RS_{wrist}^{WF} + \frac{1}{\left(\sigma_{wrist}^{ref}\right)_{WE}^2} RS_{wrist}^{WE}}{\frac{1}{\left(\sigma_{wrist}^{ref}\right)_{WF}^2} + \frac{1}{\left(\sigma_{wrist}^{ref}\right)_{WE}^2}}$$

where $\left(\sigma_{wrist}^{ref}\right)_{WF}^2$ represents the square of the standard deviation of the wrist RoM for the healthy population during the wrist flexion exercise and $\left(\sigma_{wrist}^{ref}\right)_{WE}^2$ represents the square of the standard deviation of the wrist RoM for the healthy population during the wrist extension exercise. In total, a set of 54 RoM scores are computed for each patient after they finish their prescribed 6 movement exercises, as shown in **Table 9**, and these scores comprise the digital biomarkers that reflect the physical functioning of the PsA patient.

Table 9 The set of computed RoM scores / digital biomarkers and the components used for their computation.

Digital biomarker / RoM score	Components
Left/Right Index MCP	Left/Right index finger MCP joint
Left/Right Index PIP	Left/Right index finger PIP joint during small and large fist exercises
Left/Right Index DIP	Left/Right index finger DIP joint during small and large fist exercises
Left/Right Index Total	Left/Right index finger DIP, PIP and MCP joints
Left/Right Middle MCP	Left/Right middle finger MCP joint
Left/Right Middle PIP	Left/Right middle finger PIP joint during small and large fist exercises
Left/Right Middle DIP	Left/Right middle finger DIP joint during small and large fist exercises
Left/Right Middle Total	Left/Right middle finger DIP, PIP and MCP joints
Left/Right Ring MCP	Left/Right ring finger MCP joint

Digital biomarker / RoM score	Components
Left/Right Ring PIP	Left/Right ring finger PIP joint during small and large fist exercises
Left/Right Ring DIP	Left/Right ring finger DIP joint during small and large fist exercises
Left/Right Ring Total	Left/Right ring finger DIP, PIP and MCP joints
Left/Right Pinky MCP	Left/Right pinky finger MCP joint
Left/Right Pinky PIP	Left/Right ring finger PIP joint during small and large fist exercises
Left/Right Pinky DIP	Left/Right pinky finger DIP joint during small and large fist exercises
Left/Right Pinky Total	Left/Right pinky finger DIP, PIP and MCP joints
Left/Right Hand Total	Left/Right index, middle, ring and pinky fingers
Left/Right Wrist Flexion	Left/Right wrist joint flexion
Left/Right Wrist Extension	Left/Right wrist joint extension
Left/Right Wrist Total	Left/Right wrist joint during wrist flexion and extension movement exercises
Left/Right Shoulder	Left/Right shoulder joint
Left/Right Leg Total	Left/Right knee and ankle (dorsiflexion and plantar flexion) joints
Left/Right Ankle Dorsiflexion	Left/Right ankle joint dorsiflexion
Left/Right Ankle Plantar Flexion	Left/Right ankle joint plantar flexion
Left/Right Ankle Total	Left/Right ankle joint during dorsiflexion and plantar flexion
Left/Right Knee	Left/Right knee joint
Full Body (No Ankles)	Left/Right hands, Left/Right wrists, Left/Right shoulders, and Left/Right knees
Full Body	Left/Right hands, Left/Right wrists, Left/Right shoulders, Left/Right ankles and Left/Right knees

Clinical targets: Several clinical measurements, questionnaire responses and composite scores (e.g., DAPSA, PASDAS) are obtained when patients perform clinical visits to assess the disease progression every 3 months. These clinical measurements, responses and scores comprise potential clinical targets that can be correlated with the digital biomarkers extracted from the previous step of the proposed approach. Identifying positive correlations between the computed digital biomarkers and clinical targets is crucial for the ability of the proposed joint

range of motion assessment approach to accurately estimate disease status and progression through a set of simple movement exercises.

4.2.3 Verification

The proposed approach offers a built-in data quality mechanism within the dedicated mobile application to ensure that the collected data can be used for the accurate estimation of joint RoM values and the corresponding digital biomarkers. Specifically, through the mobile application the proposed approach tracks the number of joints detected in each temporal frame in real-time for each movement exercise. If for a given exercise, the proposed approach cannot reliably detect at least 80% of the joints in at least 80% of the frames, then the mobile application informs patients that the exercise was not successfully performed and that they need to repeat the exercise. This procedure ensures that the patients remain in front of the smartphone camera for the duration of the exercises to reliably capture their movements.

Additionally, an extensive internal evaluation of the computed digital biomarkers was performed at the premises of the consortium partner developing the method, CERTH. This evaluation involved a user performing the exercises both in a correct way and by intentionally not moving specific body parts to verify that the computed digital biomarkers / RoM scores accurately reflect the movement capabilities of the user. The results indicated that the lowest scores were achieved when the user performed the exercises correctly, with higher scores were attributed to the user not moving one or more body parts.

4.2.4 Validation

The proposed approach is validated by identifying correlations between the computed digital biomarkers / RoM scores and questionnaire responses, clinical measurements and composite scores obtained during their baseline visit in the clinic. More specifically, the validation procedure involves the collection of the patients' clinical data during their baseline visit in the clinic and the computed RoM scores from the movement exercises performed in the same day or after at most 14 days in order to identify whether high scores (low physical functioning) are correlated with specific clinical measurements.

4.3 Results

4.3.1 Overall statistics

Starting from September 2024 to April 2025, a total of 91 patients have successfully performed a set of movement exercises during their first clinic visit or at most 14 days afterwards. Out of the 91 patients, 87 of them (~95.6%) performed all 6 movement exercises successfully, while 1 of them (~1.1%) did not perform the small fist movement exercise successfully and the rest 3 (~3.3%) patients did not perform the lower body movement exercise successfully. These results indicate that the hardest movement exercise to perform was the lower body movement exercise that involved the movement of knee and ankle joints.

Figure 4 presents the distribution of some indicative computed RoM scores, i.e., left hand, wrist, shoulder, knee and ankle joints, and full body, for all patients that performed the movement exercises. Similar distributions exist for the right hand, wrist, shoulder, knee and ankle joints but are not presented here for the sake of space economy.

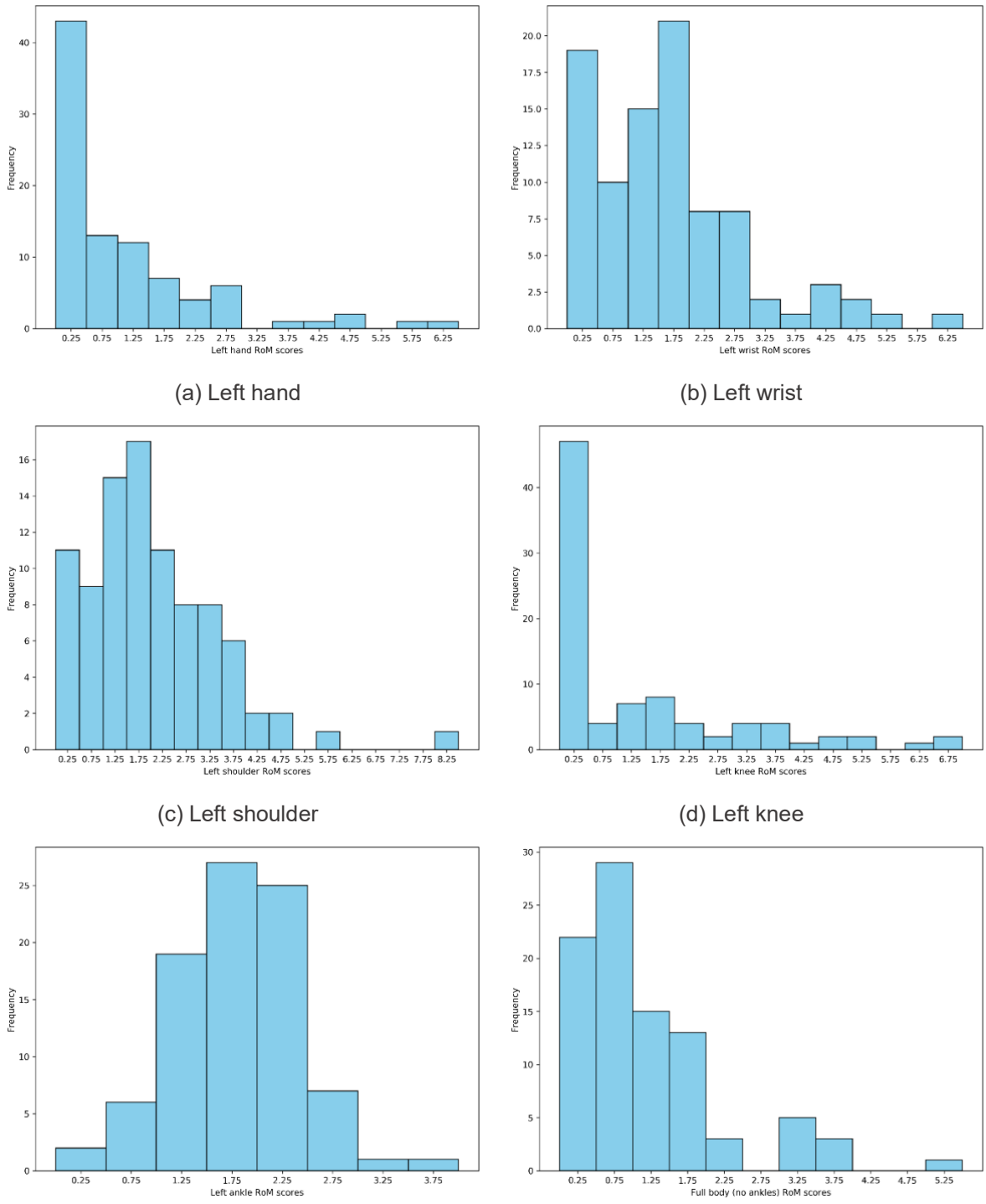


Figure 4 Distribution of RoM scores (i.e., deviation of the observed RoM from the distribution of RoM in the healthy population) for (a) the left hand, (b) the left wrist, (c) the left shoulder, (d) the left knee, (e) the left ankle and (f) the full body (no ankles) for all patients that performed the movement exercises.

Figure 4 reveals that the computed RoM scores for the hand and knee joints are primarily concentrated in the range [0-0.5], suggesting that patients generally exhibit high physical capabilities in flexing their finger and knee joints. In contrast, the RoM scores for the wrist, shoulder and ankle joints are more concentrated in the range [1-2.5], indicating greater difficulty in flexing or extending the wrist, shoulder and ankle joints. By observing the full body

(no ankles) composite RoM scores, it can be seen that most scores remain low in the range [0-1.5]. Nevertheless, there are a few instances of significantly high scores (5-6), corresponding to patients who experience considerable difficulty in performing the movement exercises. The reason for the computation of a composite full body RoM score that excludes ankle joints lies in the observation that the patients exhibit significant difficulty in accurately performing the part of the lower body movement exercise that includes the movement of ankles. As a result, a full body RoM score that excludes ankle motion evaluation can act as a more unbiased estimator of the physical capabilities of patients.

4.3.2 Validation results

For the validation of the proposed approach, a study to identify correlations between the computed digital biomarkers/RoM scores and several clinical measurements (anthropometrics, questionnaire-based and composite scores) for all patients with valid data as recorded during their first/baseline visit in the clinic was performed. Unless otherwise stated, the following validation results consider as full body RoM scores the ones that exclude the ankle motion evaluation (i.e., Full Body (No Ankles) digital biomarker).

Group comparisons were performed using the non-parametric Mann–Whitney U test (for 2 groups) and the non-parametric Kruskal-Wallis H test (for more than 2 groups) to compare the RoM scores between patients of different sex, gender and frequency of physical exercises, as well as between patients with and without tender joints. Correlation analysis was conducted using Spearman's correlation coefficients *r* to examine associations between RoM scores and both clinical targets (TJC, SJC, DAPSA) and patient-reported outcomes (PROs) (HAQ grip, VAS pain, PSAID1, PSAID3, PSAID5).

4.3.2.1 Correlation with anthropometric measurements

A study on the impact of sex and age on the computed full body RoM scores during the movement exercises drew some interesting observations. The male patients tend to perform the exercises slightly better than their female counterparts, as shown in **Table 10**. The Mann-Whitney U test reveals a statistically significant difference between the two groups (p-value of 0.02).

Table 10 Impact of sex on the performance in the movement exercises.

Sex	Count	Full body RoM score (Mean ± std)
Male	48	1.06 ± 1.01
Female	43	1.36 ± 0.94

Regarding age, it can be seen that as expected the younger patients perform the movement exercises with higher accuracy than the older patients (**Table 11**).

Table 11 Impact of age on the performance in the movement exercises.

Age group	Count	Full body RoM score (Mean ± std)
<40	16	0.80 ± 0.52
40s	20	1.02 ± 0.76
50s	25	1.51 ± 1.11
60+	30	1.27 ± 1.13

Specifically, the patients under 40 years old achieve the lowest mean full body RoM scores of 0.8, which steadily increases until the value of 1.51 for the patients at their 50s, before declining slightly to the mean value of 1.27 for the patients of over 60 years old. This demonstrates the higher physical capabilities of the younger population, despite the status of their PsA disease. However, in this case, the Kruskal-Wallis H test reveals a not statistically significant difference among the different age groups (p-value = 0.221).

4.3.2.2 Correlation with questionnaire responses

During their first clinical visit, patients were asked to complete several questionnaires regarding their habits and their disease status. To this end, associations between the patients' responses on physical activity and joint-related questions and the computed RoM scores can be identified. yr

Based on patients' responses on the frequency of exercising and the results presented in **Table 12**, it can be concluded that a high frequency of exercising is positively correlated with high physical capabilities and thus lower full body RoM scores. Specifically, the Kruskal-Wallis H test reveals a statistically significant difference among the different patient groups (p-value = 0.019).

Table 12 Impact of frequency of physical exercising on the performance in the movement exercises.

Frequency of physical exercises	Count	Full body RoM score (Mean ± std)
0 days per week	24	1.6 ± 1.2
1 to 4 days per week	33	1.16 ± 0.82
5 to 7 days per week	26	0.92 ± 0.86

These results indicate that the patients that exercise frequently maintain high physical capabilities and they are capable of performing the movement exercises more accurately (lower full body RoM scores).

On the other hand, by taking into account the physical examination tests performed on patients regarding which joints are swollen or tender and the respective RoM scores / digital biomarkers for those joints, we can investigate whether the reported swollen or tender joints can be linked with difficulties in moving those specific joints. To achieve this, we group joints based on their location in the human body and we first consider hand joints that include all finger joints. A comparison between the reports on swollen or tender finger joints and the performance of the same patients on the small and large fist movement exercises (i.e., average of the Left and Right Hand Total digital biomarkers) is presented in **Figure 5**. Out of the 60 patients with reported swollen/tender joints, 22 patients have "no swollen/tender hand joints" and achieved an average hand RoM score of 0.68 ± 0.91, while for the rest 38 patients one or more swollen/tender finger joints have been reported, leading to a hand RoM score of 1.43 ± 1.32. According to the Mann-Whitney U test, the difference between the two groups is statistically significant (p-value = 0.011).

The same procedure was applied to identify potential associations between the patients' reported swollen/tender wrist joints and the patients' performance on the wrist flexion and extension movement exercises as computed by the average of the Left and Right Wrist Total digital biomarkers. The results, presented in **Figure 6**, indicate that there are 46 patients with no swollen/tender wrist joints, achieving an average wrist RoM score of 1.58 ± 1.07 and 14 patients with reported swollen/tender wrist joints, achieving an average wrist RoM score of

2.61 ± 1.66. According to the Mann-Whitney U test, the difference between the two groups is statistically significant (p-value = 0.026).

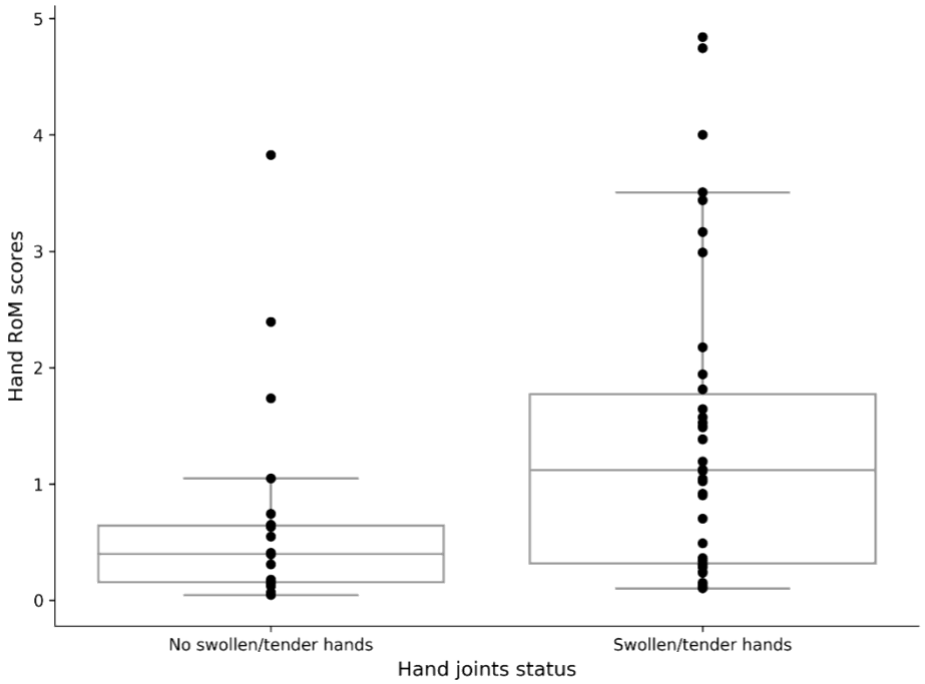


Figure 5 Box plot visualization of the distribution of hand RoM scores (i.e., average of the Left and Right Hand Total digital biomarkers) based on patients reported swollen/tender finger joints.

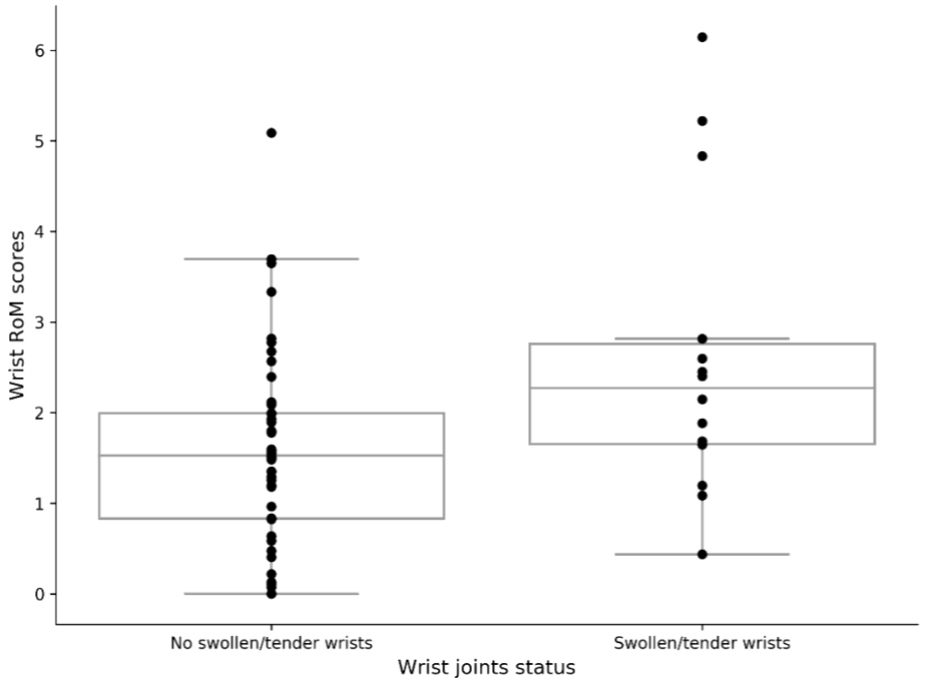


Figure 6 Box plot visualization of the distribution of wrist RoM scores (i.e., average of the Left and Right Wrist Total digital biomarkers) based on patients reported swollen/tender wrist joints.

In the same fashion, potential associations between the patients' reported swollen/tender shoulder joints and the patients' performance on the upper body movement exercises as computed by the average of the Left and Right Shoulder digital biomarkers were searched. The results, presented in **Figure 7**, indicate that there are 37 patients with no swollen/tender shoulder joints, achieving an average shoulder RoM score of 1.91 ± 1.08 and 23 patients with reported swollen/tender shoulder joints, achieving an average shoulder RoM score of 3.11 ± 1.59. According to the Mann-Whitney U test, the difference between the two groups is statistically significant (p-value = 0.002).

Finally, associations between the patients' reported swollen/tender lower body (i.e., knee, spine and/or foot) joints and the patients' performance on the lower body movement exercises as computed by the average of the Left and Right Knee digital biomarkers were searched. The results, presented in **Figure 8**, indicate that there are 13 patients with no swollen/tender lower body joints, achieving an average knee RoM score of 0.82 ± 1.6 and 46 patients with reported swollen/tender lower body joints, achieving an average knee RoM score of 1.57 ± 1.73. According to the Mann-Whitney U test, the difference between the two groups is not statistically significant (p-value = 0.085).

From these results, it can be deduced that there is a positive correlation between patients reported swollen/tender joints and the actual performance of the patients in the movement exercises. As a result, pain or swelling in the joints can affect the execution of the prescribed movement exercises and the performance of the patients in the proposed active tests can be used as an indicator for the existence or not of swelling or pain in the joints. Although the correlation between the reports and the RoM scores is of mediocre strength, in most cases it is of high statistical confidence, indicating the importance of these correlations, which can be further verified and enhanced as additional data from patients are collected.

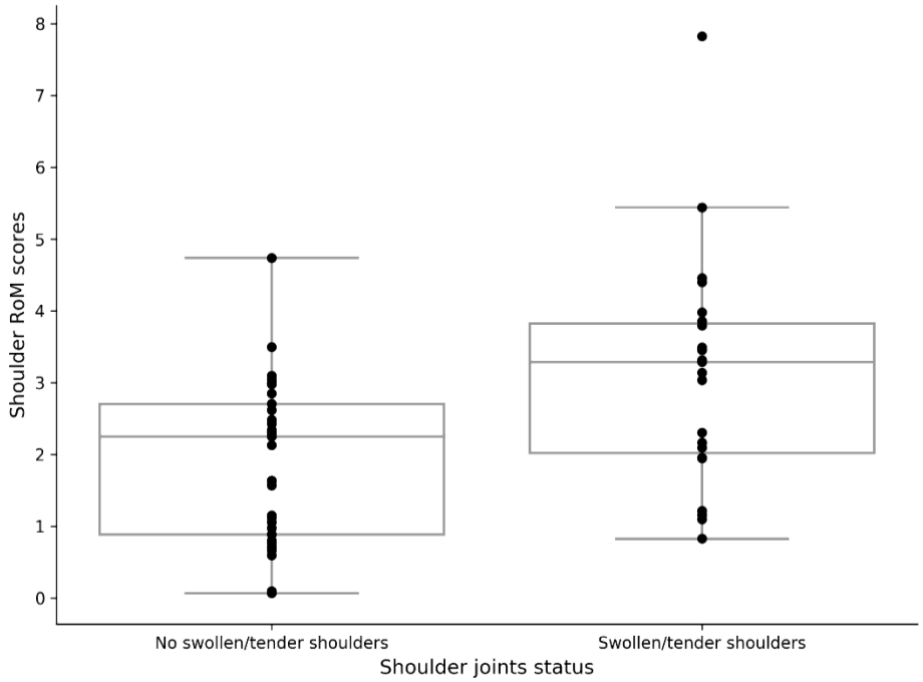


Figure 7 Box plot visualization of the distribution of shoulder RoM scores (i.e., average of the Left and Right Shoulder digital biomarkers) based on patients reported swollen/tender shoulder joints.

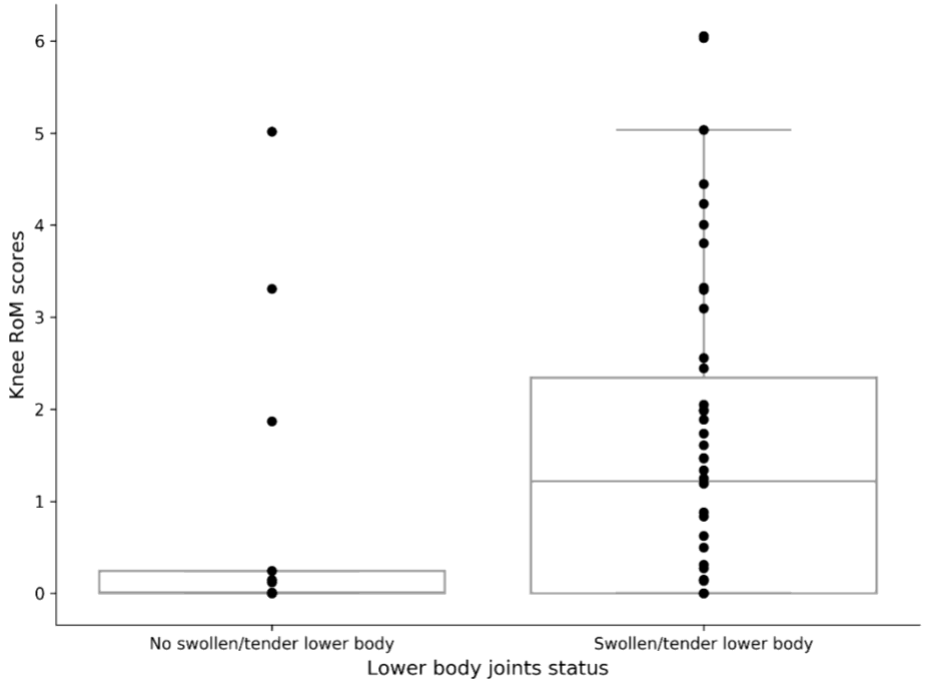


Figure 8 Box plot visualization of the distribution of knee RoM scores (i.e., average of the Left and Right Knee digital biomarkers) based on patients reported swollen/tender lower body joints.

4.3.2.3 Correlation with clinical composite scores

Clinical composite scores, such as MDA, PASDAS, DAPSA, HAQ and PSAID, offer an overall assessment of the disease status of the patient by taking into account physical and mental functioning, psoriasis surface area, pain, fatigue and several other aspects of the daily living of the patient. The goal of this section is to assess whether correlations can be identified among such clinical composite scores and the full body RoM scores that assess the physical capabilities of patients during all 6 movement tests.

The most important correlation was found to be between the overall HAQ score that measures disability and the full body RoM scores (i.e., Full Body (No Ankles) digital biomarker). **Figure 9** demonstrates that higher disability measured by the overall HAQ score is positively correlated with higher full body RoM scores. According to the Kruskal-Wallis H test, the differences among the different groups of patients are statistically significant (p-value = 0.045).

Another important correlation was identified between the number of swollen and tender joints reported in SJC66 and TJC68 clinical variables and the predicted number of swollen and tender joints. To identify swollen or tender joints, a threshold parameter t is defined and any joint with a RoM score above the threshold is considered as swollen or tender. **Figure 10** depicts the Mean Absolute Error (MAE) between the ground truth and the predicted number of swollen and tender joints with respect to the threshold parameter t. It can be seen that the optimal results (MAE equal to 6.5) are achieved for a threshold value of 2.5. For this parameter value, the Spearman correlation between the number of predicted and ground truth swollen/tender joints is positive of mediocre correlation and with high statistical significance (ρ = 0.427, p-value = 0.007).

Finally, positive and negative correlations with other composite scores and PROs were also identified, such as with MDA (ρ = 0.06), PSAID1 (ρ = 0.11), PSAID3 (ρ = 0.16), PSAID5 (ρ = 0.11), DAPSA (ρ = -0.05) and PCS (ρ = -0.08), but these correlations were found to be very weak. These weak correlations can be attributed to the dependence of several composite scores from other aspects, such as mental health and skin issues that are not related to the

physical capabilities needed to perform the movement exercises. The acquisition of additional data from patients can also strengthen some of these correlations, leading to more reliable results.

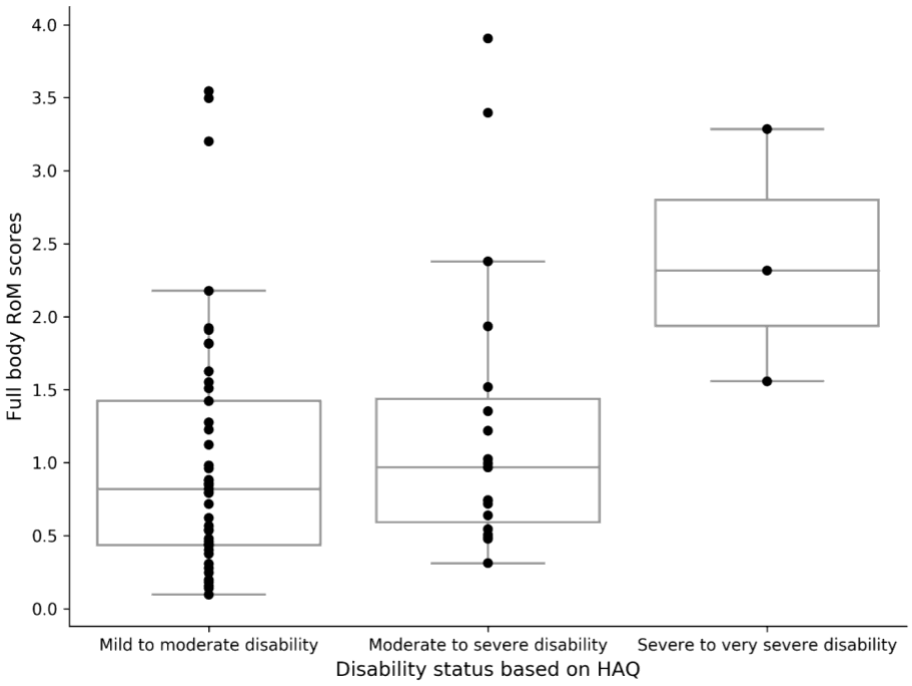


Figure 9 Box plot visualization of the correlation between the patients' disability status measured using the overall HAQ score and the full body RoM scores.

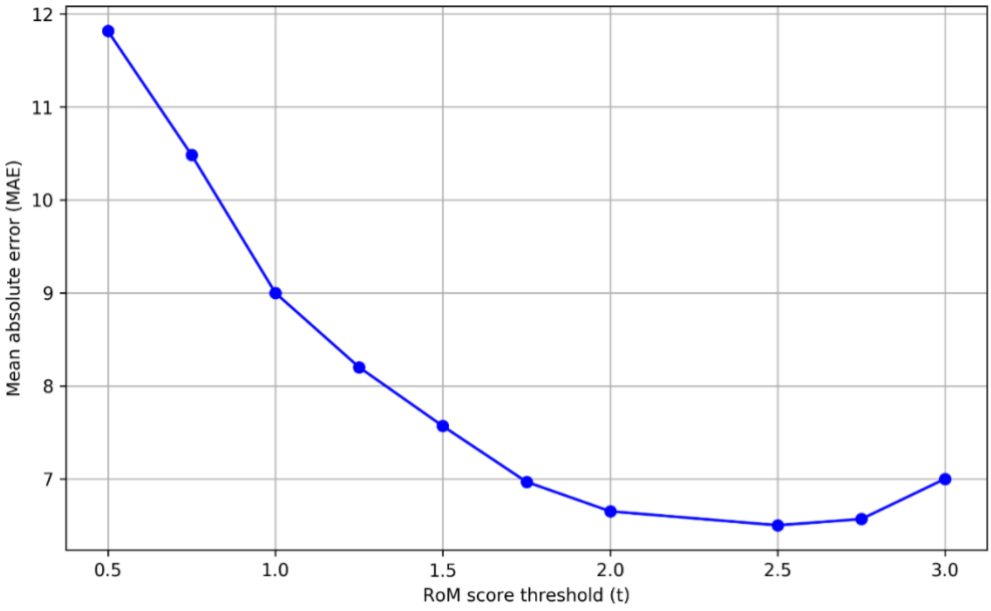


Figure 10 MAE between ground truth and predicted number of swollen/tender joints with respect to RoM score threshold t.

4.3.3 Visualization of RoM evolution

A better understanding of the computed range of motion for a given joint and the estimated corresponding RoM score can be earned by visualizing the RoM evolution of the joint as a patient performs a movement exercise. **Figure 11** illustrates the left shoulder joint RoM

evolution in degrees during the upper body movement exercise for two patients. In the figure, the left shoulder joint RoM appears as a blue line, while a dashed green line represents the average left shoulder joint RoM of the healthy population and the height of the green box equals 2 x standard deviations of the left shoulder joint RoM of the healthy population, as defined in **Table 8**. The upper part of the figure shows an excellent performance, in which the patient extends the left shoulder joint reaching a maximum RoM of around 154°, close to the average RoM of 157.1° in the healthy population, thus leading to a very small RoM score of just 0.12. On the other hand, the lower part of the figure presents a patient who performs the exercise poorly, extending the left shoulder up to a maximum RoM of 70°, a value much lower than the one of the healthy populations. The patient reported swollen shoulder joints and consequently they face significant difficulties in performing the upper body movement exercise, producing a RoM score of 4.82.

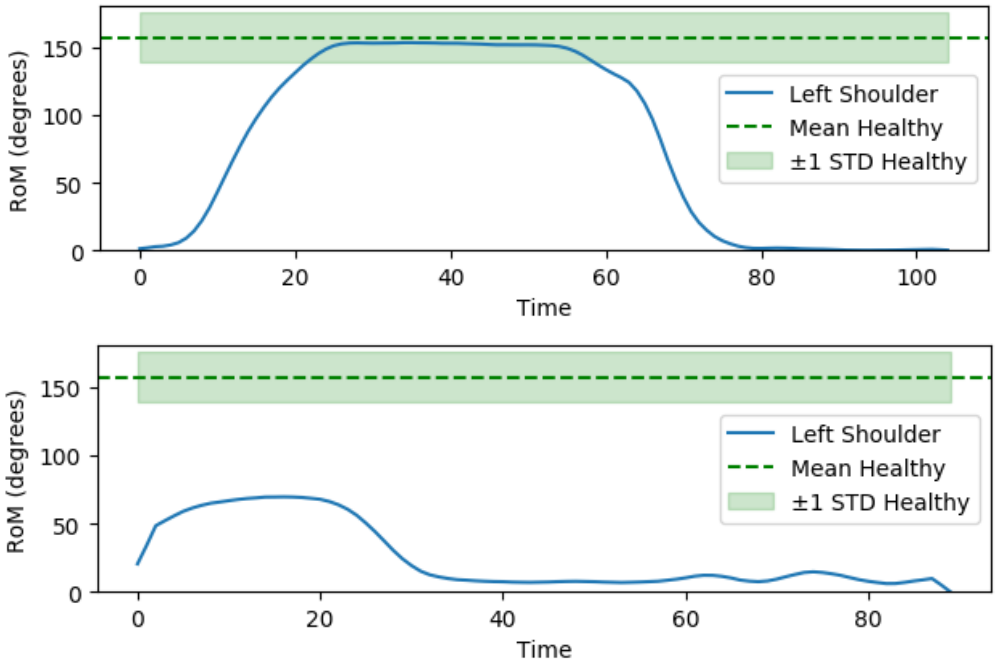


Figure 11 Left shoulder joint RoM evolution during the upper body movement exercise for a patient with high physical capabilities (upper figure) and a patient with low physical capabilities (lower figure).

Similarly, **Figure 12** illustrates the right knee joint RoM evolution in degrees during the lower body movement exercise for two patients. In the figure, the right knee joint RoM appears as a blue line, while a dashed green line represents the average right knee joint RoM of the healthy population and the height of the green box equals 2 standard deviations of the right knee joint RoM of the healthy population, as defined in **Table 8**. The upper part of the figure shows an excellent performance, in which the patient flexes the right knee joint reaching a maximum RoM of around 110°, higher than the average RoM of 91° in the healthy population, thus generating a RoM score of 0. On the other hand, the lower part of the figure shows a patient performing the lower body movement exercise poorly, flexing the right knee by only 65°, a value lower than the one of the healthy populations. The patient reported tender knee joints and consequently they face significant difficulties in performing the lower body movement exercise, producing a RoM score of 3.06.

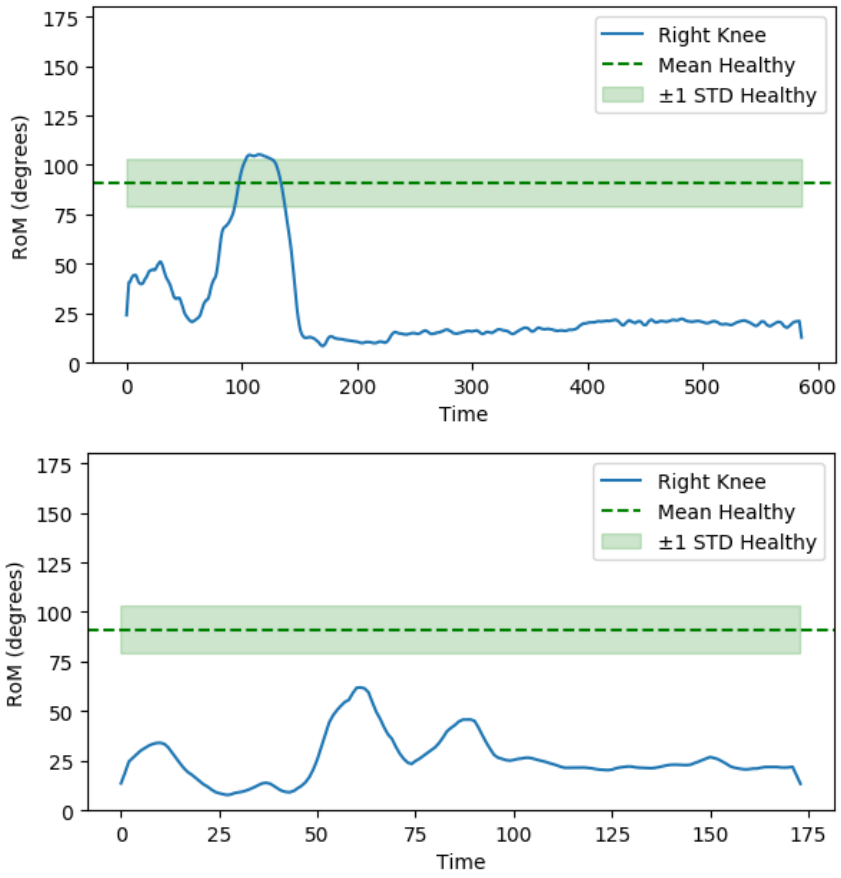


Figure 12 Right knee joint RoM evolution during the lower body movement exercise for a patient with high physical capabilities (upper figure) and a patient with low physical capabilities (lower figure).

4.4 Discussion

4.4.1 Interpretation

Through a set of simple movement exercises that target specific key joints of the human body, the proposed joint RoM score estimation approach can quite accurately assess which joints of patients are swollen or in pain as shown by the validation results. Patients with swollen or tender joints find it challenging to perform movement exercises that require maximum ranges of motion. As a result, such patients can achieve smaller RoM than the healthy population. These smaller RoM correspond to higher RoM scores computed by the proposed approach and they can be used as an indication of joint problems. The analysis of the results reveals that the proposed joint RoM score estimation approach can robustly capture these difficulties with small RoM scores indicating high physical capabilities and high RoM scores indicating poor physical capabilities and problematic joint movements. Therefore, the proposed approach with the dedicated mobile application can be a convenient tool for the automatic assessment of the physical functioning of PsA patients, and consequently a significant indicator of the disease status and progression. The study of potential correlations with questionnaire responses and composite clinical scores reveals that some associations can be found, such as with the frequency of physical activity, the joint-related pain or swelling and the HAQ overall score. However, further investigation for correlations is required due to the fact that many composite clinical scores depend on several factors that cannot be associated with physical functioning.

4.4.2 Limitations

Through the analysis of the validation results, it can be seen that high RoM scores can quite reliably be associated with joint issues, such as swelling and pain. However, there can be other reasons that can affect the physical functioning of patients that can range from inherent biomechanical limitations of joints that do not allow patients to reach the maximum RoM of healthy people to poor physical activity skills or instruction misunderstandings, leading to patients not performing the exercises as expected. There are also cases of patients reporting pain or having swollen/tender joints and still perform the exercises adequately. Moreover, medication can also affect the performance of patients as it can alleviate the pain from the joints. Several of these limitations can be overcome by performing a much larger analysis of patient data and by identifying important correlations with several different clinical measurements, tasks that are scheduled as more patient data are collected. Furthermore, the visualization of skeletal animations can empower our researchers to differentiate erroneous exercise execution from actual physical difficulties in performing the prescribed movement exercises.

4.5 Work until D3.6

One of the goals for the next working period (until M36) is to enhance the validity of the results by increasing the number of patients to perform validation. This will be possible as more patient data become available. An increase in the tested number of patients will allow to derive more reliable results regarding the correlation between clinical measurements and digital biomarkers, empowering the researchers to identify different and additional factors that can affect the physical capabilities of patients while they perform the movement exercises, as well as tackle the limitations of the current validation analysis. Finally, as more clinical visits of the same patients will become available, changes in patients' physical functioning can be calculated, paving the way for the identification of associations between these changes and patients' responses in the GRCQ questionnaire and/or changes in the disease status.

5 Joint swelling score

5.1 Background and objectives

Joint inflammation causes pain, tenderness, and swelling, which may be localised to a single joint or extend across an entire digit. Symptoms can also involve the foot joints, spine, and skin. Early diagnosis and effective monitoring of symptoms can improve quality of life and help prevent permanent joint damage (Dures et al., 2019).

In clinical practice, the assessment of PsA includes evaluation of joints using swollen and tender joint counts. Joint examination involves applying pressure (approximately 4 kg/cm^2) to assess tenderness and swelling, scored as swollen or not swollen. The scores are summed to determine the total number of swollen joints. Common joint counts used for PsA assessment include the 28-joint count, the 76/78-swollen joint counts (SJC76), and the 66/68-swollen joint counts (SJC66) (Ogdie et al., 2020). However, these assessments are subjective, relying on palpation and clinical judgment, and require experienced rheumatologists.

Advances in digital health technologies have the potential to address challenges in rheumatology through hybrid patient pathways that combine in-person and remote care (Knitza et al., 2024). Specifically for hand joint swelling assessment, a few works exist in literature, and they have been reported in the state-of-the-art deliverables, i.e. D2.2[3] and D2.4[4]. In brief, the Psorcast suite of mobile applications has introduced several smartphone-based assessments (Webster et al., 2024). One such tool is the “Hand image processing” module, which generates hand landmarks on the palm and finger joints from images and processes them to estimate joint thickness. Additionally, (Hugle et al., 2022) proposed an AI-based approach to detect and monitor joint swelling in patients with rheumatoid arthritis. Their method involves extracting dorsal finger fold lines and processing them using a convolutional neural network (CNN) to classify joints as swollen or not. Similarly, (Phatak et al, 2023) developed a method for analysing smartphone photographs of hands to detect inflammation in three selected joints, the index and middle PIP joints and the wrist, using a CNN model. The CNN achieved the highest accuracy, sensitivity, and specificity for detecting synovitis in the middle finger PIP joint, followed by the index finger PIP joint and the wrist. However, their method was not assessed for detecting swelling in DIP joints.

The work in this subtask aims to digitally characterise interphalangeal joints as swollen or not using logistic regression modelling applied to smartphone photographs. The logistic regression model uses the effective width of the joint, a metric reflecting joint thickness, as its primary predictor, with age, sex, BMI, and joint type (i.e. PIP or DIP) included as covariates. The effective width plays a central role in the proposed method. Its calculation involves developing a pipeline of computer vision algorithms that segment individual fingers, detect joint landmarks, and compute distances between points of interest.

The proposed model was validated using the dataset described in Section 3. For each participant, a pair of hand photographs and demographic and clinical data from the first clinical visit were used. The validation approach adopted was the k-fold cross-validation, where each validation fold contained exactly one patient with at least one swollen joint. This approach was chosen due to the limited dataset size.

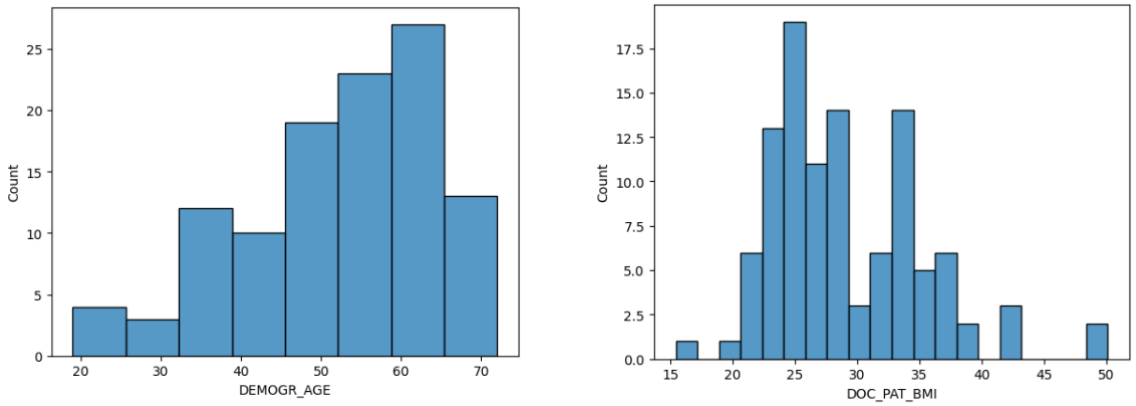
[3] iPROLEPSIS - D2.2 Initial report on the state-of-the-art and datasets
[4] iPROLEPSIS - D2.4 Updated report on the state-of-the-art and datasets

5.2 Methods

5.2.1 Dataset(s)

For digital assessment of swollen joints, patients captured photographs of their hands themselves or with assistance from healthcare professionals in the clinic. Images were taken via the miPROLEPSIS app, a mobile application developed to support the PDPID study. Patients received instructions to ensure high-quality images, including placing the dorsal side of the hand on a surface with high contrast to their skin (ideally white paper) and keeping their fingers separated. Healthcare professionals also recorded clinical assessments, including joint examinations, to provide ground truth data.

The dataset for studying digital assessment of joints consists of 85 patients, and it was evenly distributed by sex (42 male, 43 female), with a mean age of 52.6 years (SD 11.8) and a mean BMI of 28.94 kg/m^2 (SD 6.43). The age and BMI distribution are shown in **Figure 13**. The number of patients used is 85, rather than the 111 reported in Section 3, due to the exclusion of a few incorrectly captured photographs (see Verification subsection for details). Of the 85 participants, ***ten*** had at least one swollen interphalangeal joint, with a total of 40 swollen interphalangeal joints reported in the cohort.



(a) (b)

Figure 13 Distributions of (a) age and (b) BMI in digital assessment of swollen joints.

A panel with example photographs from the dataset is shown in **Figure 14**. These examples were selected to highlight the heterogeneity of image views, despite the instructions provided, illustrating the challenges of analysing smartphone photographs captured in real-life settings.

5.2.2 Development

Swelling of a joint appears as an abnormal increase in size. In interphalangeal joints, one way to measure this enlargement is by comparing the width of the DIP or PIP joint to the distance between them, a metric known as effective width (Webster et al., 2024). This measure can be derived from hand photographs using computer vision techniques. However, effective width alone cannot accurately predict interphalangeal joint swelling, as it is influenced by other factors such as anatomical characteristics. Therefore, we propose a model that incorporates effective width along with covariates such as age, BMI, sex, and joint type (PIP or DIP) to improve swelling detection performance.

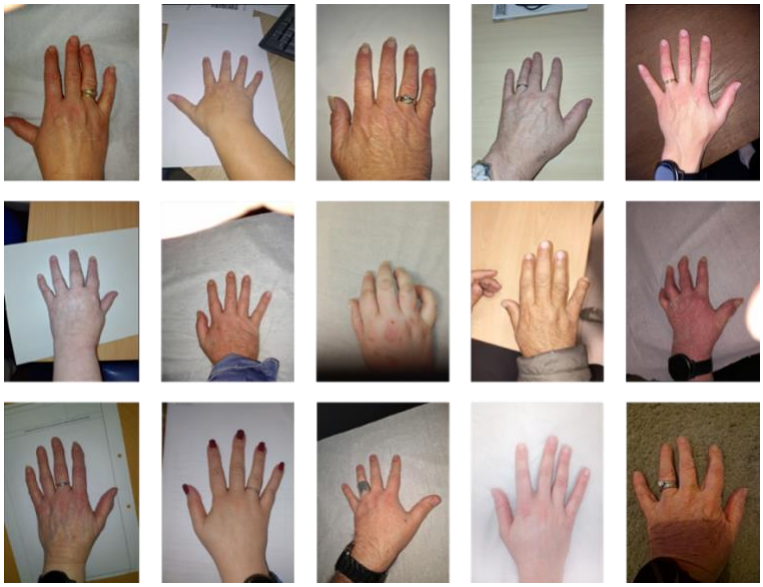


Figure 14 A panel of example hand photographs from the dataset.

Calculating the effective width from smartphone photographs involves several steps. First, hand landmarks are detected using the MediaPipe Hand Landmarker (Lugaresi et al., 2019), which provides coordinates for the hand joints and fingertips. These landmarks are used throughout the effective width calculation process. Next, a neural network, specifically, the U2-Net pretrained for human segmentation (Qin et al., 2020), is applied to the photograph to generate a hand segmentation mask. After segmentation, the contour of the hand is extracted. By combining the hand contour with the finger landmarks, the contours of individual fingers are extracted. The contour points of each finger allow the cropping of the finger into a rectangular bounding box encompassing all contour points. For each finger, the widths of the PIP and DIP joints (W) and the distance between PIP and DIP joints (D) are measured in pixels. The effective width (EFW) of each interphalangeal joint is defined as:

$$\mathrm{D} = \frac{1}{4} \cdot \left(\mathrm{D}_{index} + \mathrm{D}_{middle} + \mathrm{D}_{ring} + \mathrm{D}_{pinky}\right),$$

$$EFW_k^i = \frac{W_k^i}{D},$$

where $i \in \{index, middle, ring, pinky\}$ and $k \in \{PIP, DIP\}$. Here, D represents the average PIP-DIP distance and serves as a common denominator for all effective width calculations in the hand. Note that this definition of effective width does not apply to the thumb. The effective width calculation process is illustrated in **Figure 15**.

The logistic regression model that classifies a joint as swollen or not is formulated as:

$$logit\big(P(Swollen = 1)\big) = \beta_0 + \beta_1 \cdot EFW + \beta_2 \cdot Age + \beta_3 \cdot BMI + \beta_4 \cdot MALE + \beta_5 \cdot DIP,$$

where *EFW*, *Age*, and *BMI* are numerical variables, while *MALE* and *DIP* are binary ones. The regression coefficients β_j are estimated by maximizing the penalized log-likelihood with an L2 penalty term $\lambda\|\beta\|^2$.

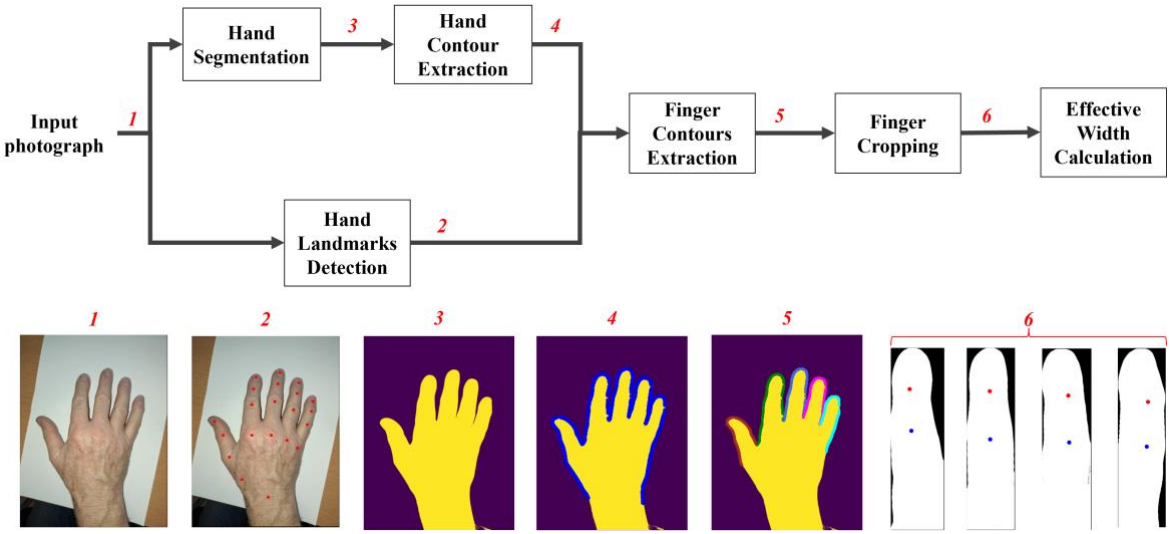


Figure 15 The effective width calculation process. A six-step processing workflow from input photograph to effective width calculation: (1) image input, (2) landmark detection, (3) hand segmentation, (4) contour extraction, (5) finger cropping, and (6) width measurement. Bottom panel shows representative outputs at each processing stage.

5.2.3 Verification

The verification phase involves a series of checks to ensure that the condition and quality of the photographs guarantee the reliable calculation of the effective width measure. The first check is performed by MediaPipe Hand Landmarker. If it fails to extract landmarks, the corresponding photographs are excluded from the analysis (see example in the **Figure 16**).

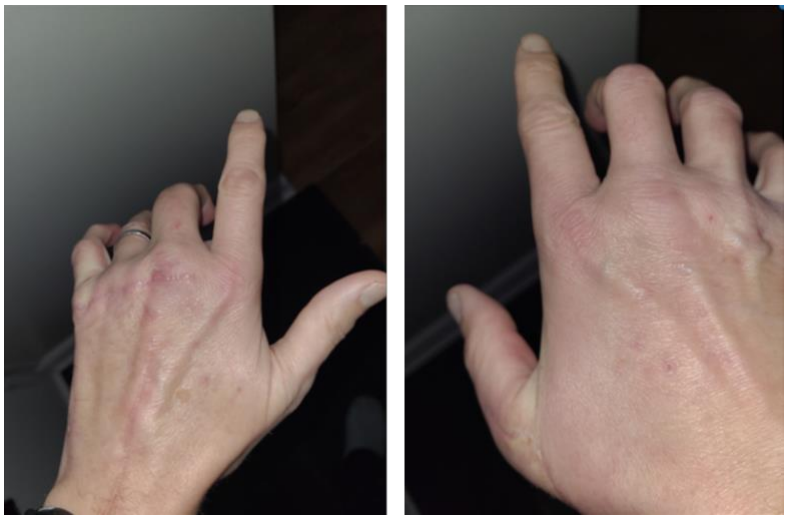


Figure 16 Examples of hand photographs for which the MediaPipe Hand Landmarker does not provide landmarks.

Moreover, in some cases, MediaPipe extracts landmarks, but they are inaccurate. This produces results outside the expected range or of no numerical values. Two examples of such cases are shown in the **Figure 17**. These errors are often caused by background objects behind the fingers (e.g., a pencil) or strange lighting artifacts, especially when the wrist is not visible.

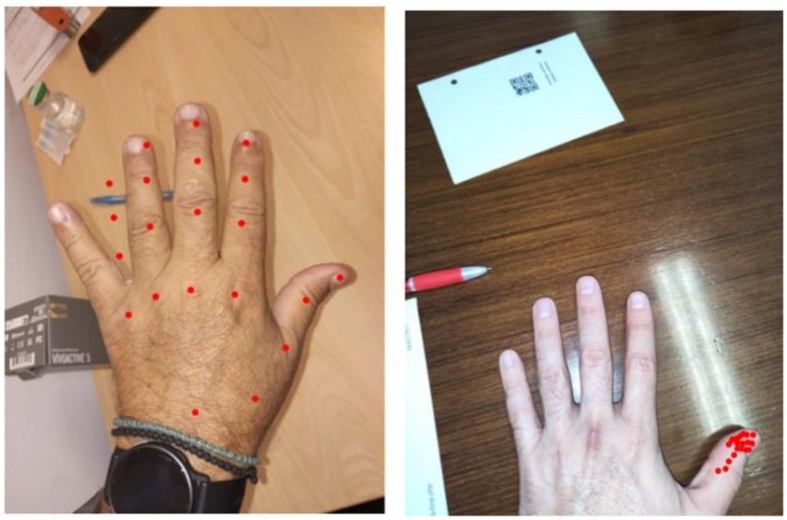
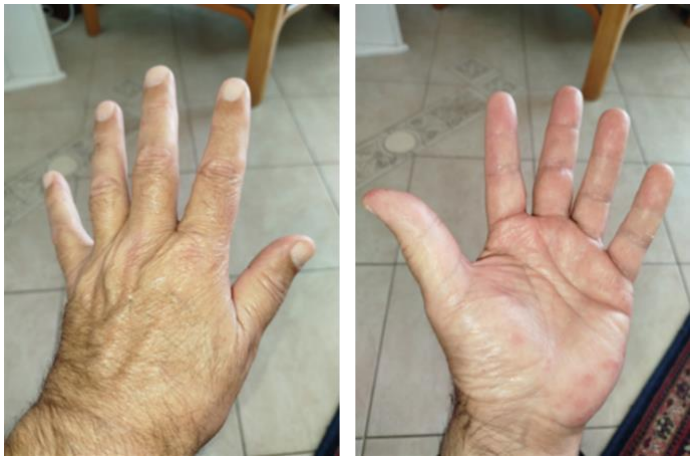
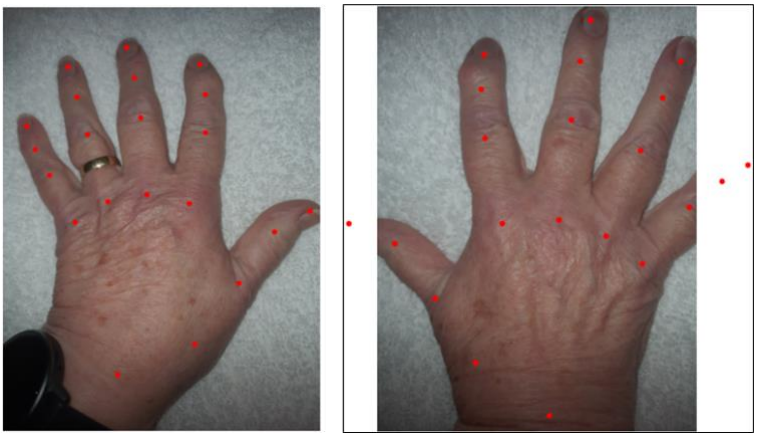


Figure 17 Examples of hand photographs for which the MediaPipe Hand Landmarker provide incorrect landmarks. The red dots indicate the extracted landmarks.

Finally, some photographs are excluded due to incorrect capture, such as two images of the same hand (e.g., two left hands) or hands partially out of frame. These are also excluded from the analysis (see examples in the **Figure 18**).



(a)



(b)

Figure 18 Examples of incorrectly captured hand photograph tests: (a) Frontal side of a hand on the right photograph; (b) Partially captured hand on the right photograph.

5.2.4 Validation

Model validation employed 10~~k~~-fold cross-validation, with each validation fold containing precisely one patient exhibiting at least one swollen joint. Additionally, a nested cross-validation strategy was used to select the inverse regularization parameter $C = 1/\lambda$. Specifically, a grid search was conducted in the inner loop to identify the value of C that maximized the area under the receiver operating characteristic curve (AUROC) on the validation set. The range of C values tested is [0.01, 0.1, 1, 10, 100].

5.3 Results

The true and false positive rates from the best model in each inner loop were retained for cumulative ROC analysis. **Figure 19** shows the average ROC curve along with upper and lower standard deviation bands. Additionally, the median and interquartile range (IQR) of the AUROC metric were calculated to assess overall model performance. The results indicate good performance, although variability across folds is substantial.

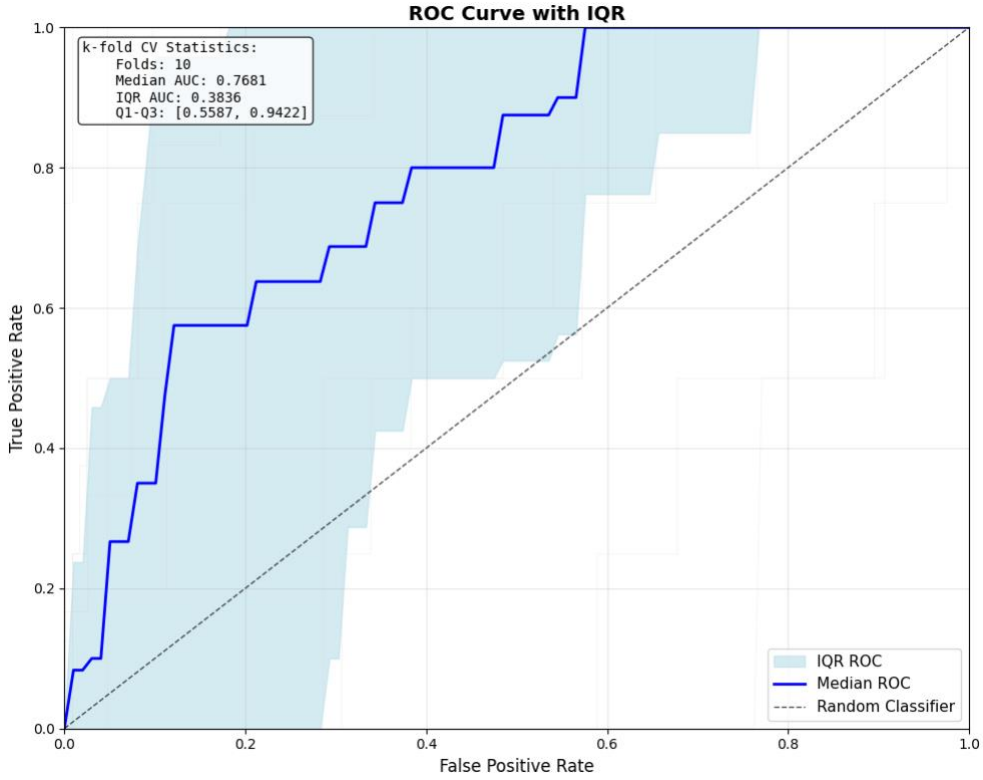


Figure 19 Receiver operating characteristic (ROC) curve showing the median swollen joint classification performance of the proposed logistic regression model, with shaded bands representing the interquartile range (IQR) across cross-validation folds.

The importance of independent variables was evaluated by examining the logistic regression coefficient magnitudes. Higher absolute coefficient values reflect a greater contribution to classification. Since regularization was applied, coefficient magnitudes also indicate the relative importance of each feature compared to the others. **Figure 20** displays the median absolute coefficients across all cross-validation folds, bounded by the first and third quartiles. As expected, the effective width emerged as the most important independent variable. However, the other covariates also showed smaller but notable contributions to swollen joint classification.

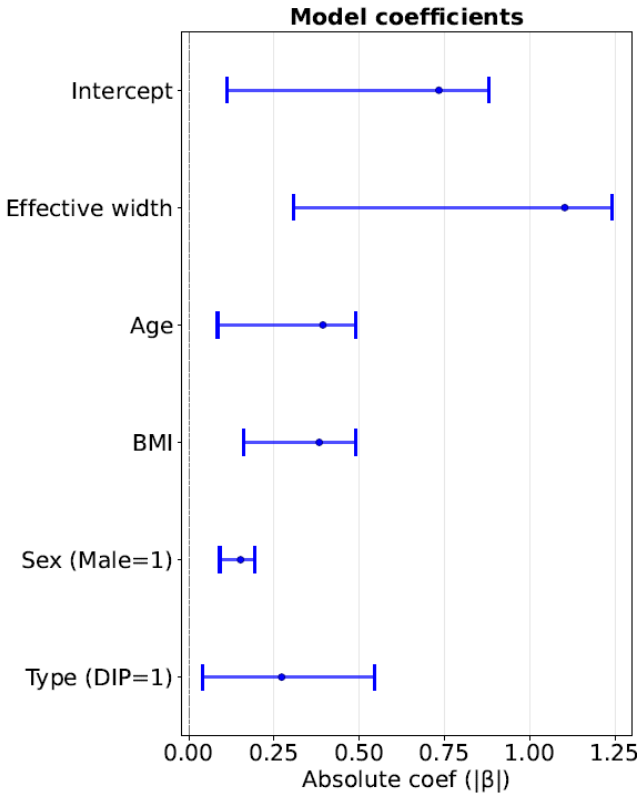


Figure 20 Magnitudes of logistic regression coefficients indicating the importance of independent variables in classifying swollen joints. For each coefficient, the median value across all cross-validation folds is shown, with error bars representing the first and third quartiles.

5.4 Discussion

5.4.1 Interpretation

This work demonstrates that a logistic regression model combining a joint thickness measure, the effective width, with auxiliary variables such as age, BMI, sex, and joint type can serve as a promising tool for detecting interphalangeal joint swelling. The results highlight the value of effective width as an indicator of joint swelling, validated using a real-world dataset of photographs taken under varying conditions. While promising, the model currently exhibits greater performance variability than is ideal, likely due to the limited number of positive (swollen) cases. Further validation on larger datasets is needed. With more data, the model could also be enhanced by incorporating interactions among independent variables to improve classification performance. Finally, the proposed model holds promise for monitoring joint swelling in PsA patients over time, although this within-subject evaluation would require longitudinal photographic data from the same individuals.

5.4.2 Limitations

One limitation of the approach is its inability to assess swelling in other hand joints, such as the MCPs and the wrist, where swelling does not manifest as relative thickness in a specific direction. Additionally, the method cannot be applied to toe joints because MediaPipe does not detect foot joint landmarks.

Another limitation concerns how the photograph is captured. The pipeline may fail if the background is cluttered, lighting artifacts are present, or the hand is partially covered or out of frame. While the effective width pipeline performs well under various challenging conditions, it fails in certain extreme cases.

5.5 Work until D3.6

As more data are collected in the PDPID study, the dataset will grow and be used to augment the proposed model. This augmentation will include further evaluation of the contribution of covariate variables and their interactions. In addition to effective width, new measures will be explored to capture other characteristics of swelling, such as skin folds over affected joints. Finally, an additional model will be developed to represent longitudinal changes in swelling and their association with clinical outcomes.

6 Nail psoriasis detection

6.1 Background and objectives

Nail psoriasis is a common manifestation of psoriasis, a chronic autoimmune skin disorder. It affects the fingernails and toenails, leading to a range of symptoms including pitting, discoloration, thickening (hyperkeratosis), crumbling, and separation of the nail from the nail bed (onycholysis). Nail involvement occurs in up to 50% of people with psoriasis and is even more prevalent in those with psoriatic arthritis. Despite being confined to the nails, nail psoriasis can have a profound impact on a patient's quality of life. It can cause pain, functional impairment, and emotional distress due to the visible and often stigmatizing appearance of affected nails. Moreover, nail psoriasis is considered an important clinical marker, as it may signal more severe disease or serve as an early indicator of psoriatic arthritis. Therefore, recognizing and effectively managing nail psoriasis is crucial not only for improving physical and psychological well-being, but also for potentially mitigating progression to joint involvement. As such, it represents an important focus in the comprehensive care of individuals with psoriasis.

In the framework of the iPROLEPSIS project, a mobile application was developed that allows a user to capture images of their hands and feet using a simple smartphone camera. Then, an automated deep learning-based methodology was developed to process these images, identify nails and classify potential nail diseases. The objective of the proposed approach is to assess the ability of such a methodology to accurately assess nail psoriasis on fingers and toes and correlate these findings with clinical measurements derived from patients' clinical visits.

6.2 Methods

6.2.1 Dataset(s)

6.2.1.1 Nail Detection Dataset

The Nail Detection Dataset (Markus, 2024) contains images of hands with annotated fingernails in YOLO format. In total, there are 5209 nail images, split in training set (3949 images), validation set (840 images) and test set (420 images). Although this dataset was publicly available until March 2025, it is now no longer available for download and use.

6.2.1.2 Nail Disease Detection Dataset

The Nail Disease Detection Dataset (Kaggle, 2021) contains cropped healthy hand nail images or nail images with different diseases. Specifically, the dataset depicts nails with 9 diseases, namely Acral Lentiginous Melanoma, Beaus Line, Blue Finger, Clubbing, Koilonychia, Muehrckes Lines, Onychogryphosis, Pitting and Terry's Nail. In total, the dataset contains 6670 nail images, with 6363 of them belonging to the training set and the rest 307 of them belonging to the test set. The distribution of nail images per disease is presented in **Table 13**.

The dataset was publicly available until March 2025, but it is now private. To overcome this issue, a new dataset (Kaggle, 2024) has been identified that contains in total 3835 hand nail images, split among healthy nails, as well as nails with 5 different diseases, namely Acral Lentiginous Melanoma, Blue Finger, Clubbing, Onychogryphosis and Pitting.

Regarding feet nail detection and disease classification, no datasets were identified, and thus the hand datasets were also employed in this case.

Table 13 Distribution of nail images per disease in the Nail Disease Detection Dataset.

Nail disease	Training images	Testing images
Acral Lentiginous Melanoma	753	36
Beaus Line	456	22
Blue Finger	612	29
Clubbing	783	38
Healthy	645	30
Koilonychia	537	28
Muehrckes Lines	336	16
Onychogryphosis	690	34
Pitting	657	32
Terry's Nail	894	42

6.2.2 Development

The pipeline that was followed for the detection of nails affected by psoriasis is presented in **Figure 21**.

Figure 21 Pipeline for the detection and classification of nail psoriasis.

Specifically, two deep networks were developed and trained, the nail detection network and the nail classification network that are described in detail below.

Nail detection network

The nail detection network is a YOLOv8 deep network, pretrained on ImageNet and finetuned with the Nail Detection dataset. This network accepts as input a hand image and is responsible for detecting and extracting bounding boxes that contain the nail regions from the input image. The network was finetuned for 3 epochs and achieved very satisfactory results in the test set of the Nail Detection dataset as shown in **Table 14**.

Table 14 Accuracy metrics of the Nail detection network in the test set of the Nail Detection dataset.

Recall	Precision	F1-score
0.88	0.92	0.9

Nail classification network

The nail classification network consists of a Uniformer small Transformer network (Li et al., 2023), enhanced with the Multi-manifold Multi-head Attention (MMA) mechanism, proposed in

(Konstantinidis et al., 2023). By projecting the feature space in the Euclidean, Semi-positive definite and Grassmann manifolds through the MMA mechanism, the Transformer is capable of computing more descriptive representations of the initial feature space and improving its performance. The Uniformer small Transformer network is pretrained on ImageNet and finetuned using the Nail Disease Detection Dataset. It takes as input a nail region and classifies the nail as healthy, psoriatic or other. To achieve this, the diseases of Beaus Line, Onychogryphosis and Pitting were grouped together as they are heavily associated with psoriasis according to the international biography (Ji et al., 2021, while the other 6 diseases were also grouped together and were labelled as "Other". The nail classification network was trained for a maximum number of 100 epochs and the training ended when the network achieved the lowest test loss, leading to a very high accuracy of **95.29%** in the test set of the Nail Disease Detection dataset.

6.2.3 Verification

The miPROLEPSIS mobile app offers a built-in data quality mechanism to ensure that the patients actually provide images of their hands and feet. Specifically, during the image capturing procedure, outlines of the hands and feet are drawn so that the patients align their hands and feet with the outlines. This mechanism facilitates the correct capturing procedure, minimizing cases of hands and feet captured partially and/or too distant or too close to the smartphone camera. Unfortunately, building and deploying an algorithm to ensure that the captured images actually contain hands or feet is too demanding for an application that runs in real-time.

6.2.4 Validation

The proposed approach is validated by identifying associations among its classification results and whether the patients exhibit nail dystrophy as recorded during the clinical visits. More specifically, a patient will be considered to exhibit nail dystrophy by the proposed approach if one or more of their hands and/or foot fingernails are classified as psoriatic. These results will be directly associated with the nail dystrophy recorded during the patients' clinical visits.

6.3 Results

Starting from September 2024 to April 2025, a total of 86 patients has successfully uploaded images of their hands and feet during their first clinic visit or at most 14 days afterwards. These images were collected and processed to identify correlations between nail psoriasis detection results from the proposed methodology and the nail dystrophy results recorded during the first clinic visit. Initially, the nail detection network was executed to extract nail regions for further processing. The results of the testing of the nail detection network in the 172 hand images collected from the patients are presented in **Table 15**.

Table 15 Performance of the trained nail detection network in the 172 hand images of patients.

Recall	Precision	F1-score
0.69	0.90	0.78

From these results, it can be seen that the trained nail detection network demonstrates strong performance in the hand images, achieving an F1-score of 0.78. Subsequently, it was decided to use the same network in the 172 images of feet collected from the patients in order to evaluate its ability to accurately extract the nail regions from the foot images. However, its performance on foot images declined significantly, with an F1-score of only 0.33, characterized by high precision (0.9) but very low recall (0.2). These findings indicate that while the network is capable of making accurate predictions when confident, it fails to detect many true nail

regions in foot images, therefore highlighting the need for training with foot images to improve detection performance.

A few results from employing the nail detection network on patients' hand and foot images are presented in **Figure 22**. It can be seen that the network can identify several nail regions, but some of them are not recognized successfully, due to the need for better training, especially in foot images.

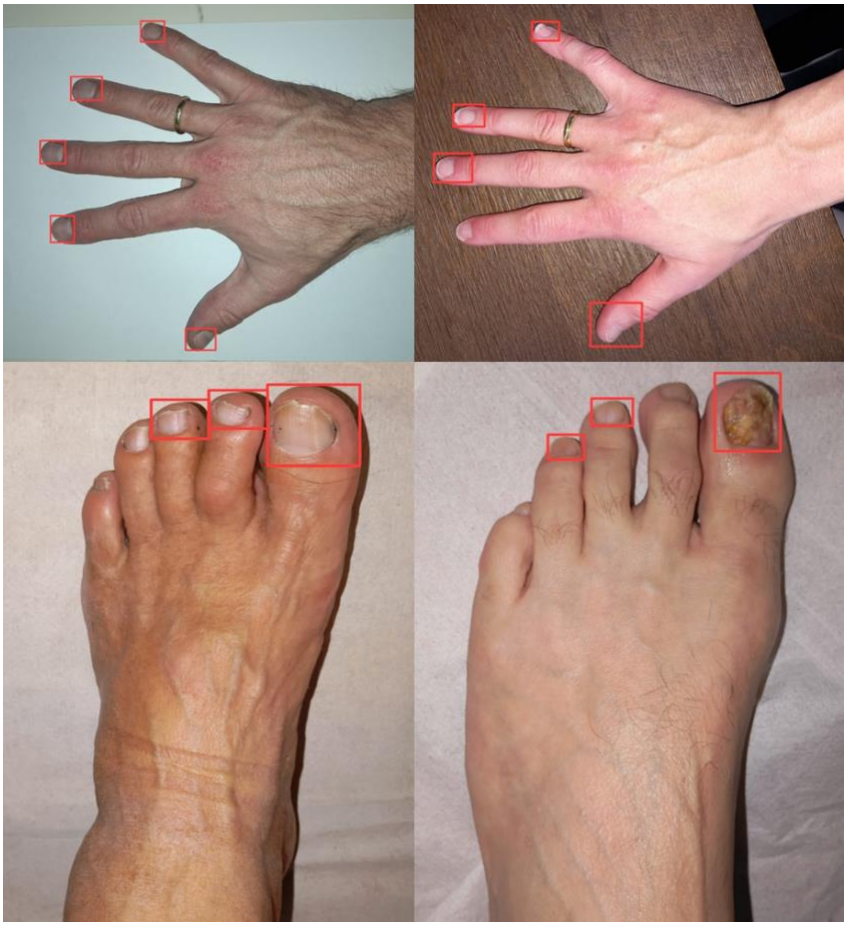


Figure 22 Nail detection results on hand and foot images from the proposed method.

Finally, the nail classification network was evaluated on the nail regions extracted from the hand and foot images of the patients. We consider a patient as having nail psoriasis when the nail classification network identifies a single nail on either hands or feet as psoriatic. Otherwise, the patient is labelled as healthy in terms of nail diseases. These results are then correlated with nail dystrophy, which can be either positive or negative, and it is measured during the clinical visit. The result of this comparison is presented as a confusion matrix in

Table 16.

Table 16 Confusion matrix from the comparison of the nail dystrophy results (ground truth) and the output of the proposed method (prediction).

		Prediction (from proposed method)	
		Nail psoriasis	**No nail psoriasis**
Ground truth	**Nail psoriasis**	**11**	**24**

(from nail psoriasis dystrophy)	**No nail psoriasis**	**12**	**39**

From the confusion matrix, an accuracy of 0.58, a recall of 0.31, a precision of 0.48 and a F1-score of 0.38 can be calculated. These values indicate a mediocre performance, despite the fact that the nail classification network, which is trained on hand images, is applied on foot images. These findings indicate that the proposed nail classification network needs further training, especially on foot images, to improve its performance, since 24 patients with nail psoriasis are not recognized as such, while 12 patients with no nail psoriasis are falsely identified as having nail psoriasis.

6.4 Discussion

6.4.1 Interpretation

From the results presented above, it can be deduced that the proposed method achieves quite accurate performance in the task of hand nail detection but needs improvements in its accuracy for the task of foot nail detection. The lack of publicly available foot nail detection datasets significantly affects the detection results and the use of a network trained on hand images to foot images cannot easily mitigate this issue. As a result, a wider search for foot nail detection datasets should be performed and/or transfer learning approaches should be utilized so that the proposed hand nail detection network can be successfully employed for foot images.

Additionally, regarding nail classification, the performance of the proposed method can be considered mediocre (accuracy of 58%), meaning that there is a need of exploring more robust network architectures, curating foot-specific datasets, and employing advanced data augmentation strategies to enhance classification accuracy.

6.4.2 Limitations

The evaluation of the proposed method in hand and foot images collected from patients shows that the performance is affected by cases of blurred images, painted nails, tattoos, rings and watches. Painted nails tend to disguise the actual disease in the nails, leading to erroneous disease classification results. Moreover, objects and/or patterns present in the hand and foot images, such as tattoos, rings and watches can lead to false positive nail detection results. As a result, there is a need for a more robust nail detection network that can be achieved with larger datasets and better data augmentation approaches.

6.5 Work until D3.6

Future research will focus on further enhancing the performance of the proposed nail psoriasis detection method, especially for the task of foot nail detection. Efforts will be directed toward exploring advanced network architectures and identifying additional publicly available datasets. In parallel, novel synthetic data generation strategies will be developed to complement and expand existing datasets. To improve robustness, diverse data augmentation techniques will also be employed. Finally, whenever feasible, potential associations between the nail psoriasis outcomes and other relevant clinical variables will be investigated.

7 Typing dynamics

7.1 Background and objectives

Impairment of hand function is common and significant in individuals with psoriatic arthritis (PsA) (Liphardt et al., 2020). Clinical assessment typically relies on musculoskeletal exams performed by experienced clinicians, such as palpation to evaluate joint tenderness or swelling. While effective, these methods require in-person visits and provide only a momentary snapshot of disease activity. Hand function is also indirectly evaluated through patient-reported questionnaires, which are accessible but subjective and limited in their ability to capture subtle or fluctuating fine motor deficits.

Remote digital assessments have shown promise in neurological conditions. Keystroke dynamics have been linked to fine motor impairment in Parkinson's disease (Iakovakis et al., 2018) and multiple sclerosis (Lam et al., 2021). More recently, Matias et al. (2024) extended this approach to rheumatic and musculoskeletal diseases, including PsA, using smartphone-based typing features to detect fine motor changes. These findings suggest that natural smartphone typing behaviour may offer a scalable and objective method to monitor upper extremity function in PsA.

The aim of this study is to identify signs of fine motor impairment and behavioural changes in people with PsA through the analysis of touchscreen typing behaviour captured in daily life. Keystroke-derived features—such as key press and release durations, typing speed, and backspace or space usage—are analysed alongside accelerometer-based measures of phone motion and orientation. These digital signals are then examined in relation to clinical outcomes (e.g., joint counts, composite scores) and patient-reported symptoms (fatigue, pain, stiffness) to support the development of remote digital biomarkers of hand function.

7.2 Methods

7.2.1 Dataset(s)

To develop the smartphone-based digital biomarkers (dBMs) of PsA using typing dynamics, we employed data from the first batch of participants recruited in the PDPID study (see section 3 for details). Typing data were collected passively in real-world conditions during everyday smartphone use via the miPROLEPSIS app. This app includes a custom keyboard based on the latest release of the Android Open-Source Project (AOSP) keyboard (Latin IME)[5]. It replaces the native Android keyboard, supporting multiple languages and typing facilitators such as word suggestions and auto-correction. The custom keyboard replaces the device's native keyboard and records timestamped touch events (press and release), inter-key distances, and additional metadata such as language, phone orientation, and screen layout, during each typing session. A typing session is defined as the period between the keyboard being shown (popped up) and subsequently hidden (minimized). To ensure user privacy, the actual characters typed are never recorded.

Each participant contributed multiple typing sessions. In parallel, continuous smartphone accelerometer data were also recorded. Using the start and end times of each typing session, we aligned and paired each session with the corresponding accelerometer signal. The typing dynamics and associated metadata are summarized in **Table 17**.

[5] https://android.googlesource.com/platform/packages/inputmethods/LatinIME/

Table 17 Typing data captured by the custom keyboard of the miPROLEPSIS app.

Variable	Description
startTimestamp	The time the typing session started in ms (unix time)
stopTimestamp	The time the typing session ended
downTime	Timestamps corresponding to the moment alphanumeric keys were tapped down
upTime	Timestamps corresponding to the moment alphanumeric keys were released
downTimeBackspace	Timestamps corresponding to the moment backspace key was tapped down
upTimeBackspace	Timestamps corresponding to the moment backspace key was released
downTimeSpace	Timestamps corresponding to the moment space key were tapped down
upTimeSpace	Timestamps corresponding to the moment space key were released
pressure	Normalized (0-1) maximum pressure applied for each key tap (where available)
distance	Distance between keys tapped (mm)
isLongPress	Flag denoting whether a key was long pressed
keyboardScale	Scale of the keyboard (0: portrait, 1: landscape)
characterCount	Number of alphanumeric keys tapped
numBackspaces	Number of times the backspace/delete key was tapped
numSpaces	Number of times the space key was tapped
isVibrationOn/ isSoundOn/ isShowPopupOn	Type of key feedback, i.e., vibration or sound (flag)
currentLanguage	Keyboard language (ISO 369 alpha-2 locale)

7.2.2 Development

Typing and phone motion data collected via the custom keyboard and smartphone sensors (see **Table 17**) were processed to derive features indicative of fine motor control, cognitive effort, and behavioural patterns relevant to psoriatic arthritis. Each valid typing session (≥50 characters) includes timestamped key events, inter-key distances, backspace/space usage, and applied pressure (when available), along with continuous accelerometer signals captured during the session. From these signals, we extracted temporal and spatial features—such as hold time, flight time, typing speed, and device motion—that serve as candidate digital biomarkers. A visual summary of the processing pipeline is shown in **Figure 23**.

Typing features extraction

For typing sessions containing at least 50 characters, three primary timing sequences were extracted from the raw keyboard data using the recorded key down-time, up-time, and inter-key distances:

- **Hold time**: Duration between key press (down) and release (up).
- **Flight time**: Time between releasing one key and pressing the next.

- **Typing speed**: Inter-key distance divided by the corresponding flight time.

Definitions of these typing events are provided in **Figure 24**.

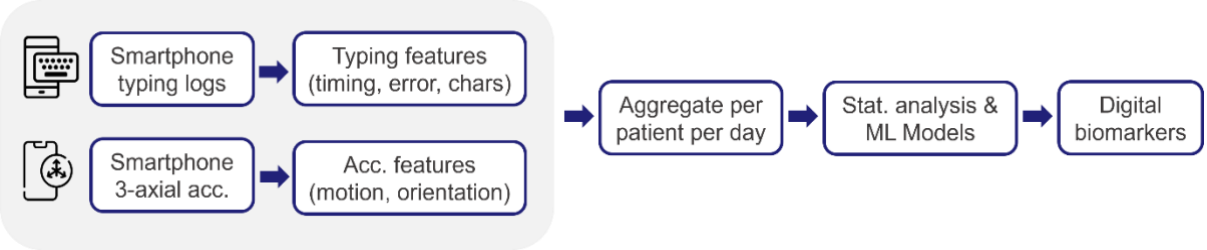


Figure 23 Typing dynamics analysis overview.

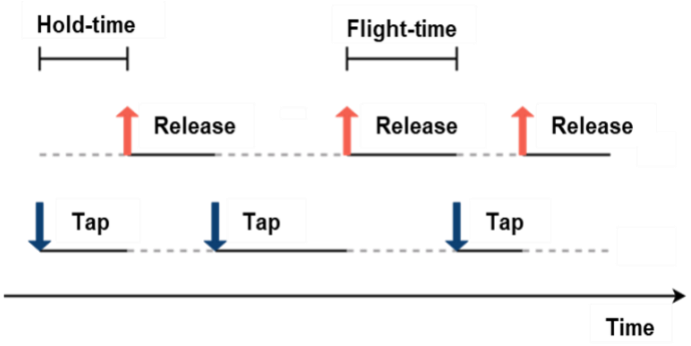


Figure 24 Definitions of the typing events.

To ensure data quality and reduce noise from outliers or non-standard typing behaviour, the following filtering steps were applied:

- **Hold time** values outside the range of 0 to 500 milliseconds were excluded to remove long presses and unintended holds.
- **Flight time** values were retained only if they fell within the 1st to 99th percentile of a real-world typing distribution ([−1270, 1700] ms) (Iakovakis et al., 2018); values below 0 ms (e.g., two-handed typing overlaps) were also excluded.
- **Flight time normalization** was applied within each session by subtracting the session mean, minimizing individual differences in baseline typing skills.

Following filtering, each sequence was processed to compute second- and higher-order statistics per session, including the mean, standard deviation, skewness, kurtosis, and the slope of a linear fit across the sequence. Additionally, two behavioural features were derived: the **error rate**, estimated as the proportion of deleted characters (backspace taps divided by total characters typed), and the **average characters per word**, calculated by dividing the number of characters by the number of space key taps.

Typing sessions were then aggregated on a daily level per participant by computing the **median** of each feature across all sessions in each day. The final set of typing-derived features used for modelling is presented in **Table 18**.

Table 18 List of typing features extracted per typing session.

Feature	Description
Hold Mean	Mean duration of key presses (tap down to release)
Hold Std	Standard deviation of key press durations
Hold Skew	Skewness of hold time distribution (asymmetry)
Hold Kurt	Kurtosis of hold time distribution (peakedness)
Hold Slope	Slope of hold time across the session (fatigue or drift)
Flight Std	Standard deviation of flight times (time between key releases and next press), normalized by session mean
Flight Skew	Skewness of flight time distribution
Flight Kurt	Kurtosis of flight time distribution
Flight Slope	Slope of flight time over session duration
Speed Mean	Mean tap speed, calculated as interkey distance divided by flight time (not normalized)
Speed Std	Standard deviation of tap speed
Speed Skew	Skewness of tap speed distribution
Speed Kurt	Kurtosis of tap speed distribution
Speed Slope	Slope of tap speed over session time
Error Rate	Proportion of deleted characters (backspaces/ total typed)
Chars Per Word	Average characters per word, calculated as character count divided by word count (space key taps per session).

Accelerometer feature extraction

For each valid typing session, the corresponding segment of smartphone accelerometer data was extracted using the session's start and end timestamps. Tri-axial acceleration signals (x, y, z) were recorded continuously by smartphone sensors at a minimum sampling rate of approximately 20 Hz. All sequences were resampled uniformly to 20 Hz using linear interpolation to correct for irregularities.

Preprocessing involved quality control and signal filtering. Sessions were excluded if more than 10% of expected samples were missing within the paired window or if the signal quality was inadequate. After resampling, the acceleration signal was decomposed into gravitational and linear components.

Additional filtering was applied based on the type of features:

- **Posture and device orientation**: Low-pass filters (cutoff = 3 Hz) applied to the raw acceleration signal to retain gravitational component relevant for orientation estimation.
- **Motion intensity and variability**: Low-pass filters (cutoff = 3 Hz) applied to the linear component to extract general motion features.
- **Dynamic movement**: Band-pass filters (3–9 Hz) applied to the linear component to isolate higher-frequency dynamic movement.

Posture and device orientation were estimated using low-frequency raw accelerometer signals. Median acceleration along the x- and z-axes indicated horizontal and vertical positioning, while the tilt angle—defined as the angle between the phone's z-axis and the gravity vector—was used to describe phone orientation. The mean and standard deviation of this angle reflected posture stability and drift over time.

Motion intensity and variability were described using the mean, standard deviation, and skewness of the linear acceleration magnitude. Signal entropy quantified irregularity, while energy and area under the curve (AUC) reflected the total activity during each session.

Dynamic movement was described using the mean and standard deviation of the jerk signal (rate of change in acceleration), as well as the peak-to-peak amplitude and the time between the signal's maximum and minimum values.

All accelerometer-derived features were computed per session and aggregated daily using the **median** value across sessions for each participant. The complete list of features used in the analysis is presented in **Table 19**.

Table 19 List of accelerometer-derived features extracted per typing session.

Feature	Description
AccMag Mean	Mean of linear acceleration magnitude (motion intensity)
AccMag Std	Standard deviation of acceleration magnitude
AccMag Skew	Skewness of acceleration magnitude (asymmetry of motion)
Acc Entropy	Entropy of acceleration magnitude (signal unpredictability)
Acc Energy	Signal energy of linear acceleration magnitude
Acc AUC	Area under the curve of acceleration magnitude over time
Peak2Peak	Peak-to-peak range of acceleration magnitude (max - min)
Peak Time	Time between maximum and minimum acceleration magnitude
Jerk Mean	Mean jerk (rate of change of linear acceleration)
Jerk Std	Standard deviation of jerk
Acc X Median	Median acceleration on x-axis (orientation-related, with gravity)
Acc Z Median	Median acceleration on z-axis (upright/flat posture indicator)
Tilt Mean	Mean angle between the phone's z-axis and the gravity vector (0° = flat)
Tilt Std	Standard deviation of tilt angle across the session

7.2.3 Verification

Data capture was tested on multiple phones and shown to be consistent, especially for typing data, which was verified through device-level checks. The applied filters (e.g., session length, invalid timing values) ensured usable input but are considered part of preprocessing rather than verification. Accelerometer data were more sensitive to issues like sensor quality and temperature, so additional checks for missing data and sampling rate consistency were applied, and low-quality sessions were excluded.

7.2.4 Validation

Statistical analysis

To assess the reliability and stability of each extracted feature across repeated measures, the intraclass correlation coefficient ICC(3, k) was computed over the 14-day period. This two-way mixed-effects model for average measures estimates the consistency of each feature across study days within individuals, where *k* represents the number of repeated daily measurements. ICC values below 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, between 0.75 and 0.9 indicate good reliability, and values above 0.9 indicate excellent reliability (Koo & Li, 2016).

Group comparisons were performed using the non-parametric Mann–Whitney U test to compare typing and accelerometer features between patients with tender hand joints and those without. A significance threshold of $p < 0.05$ was used without correction for multiple comparisons. As this analysis was exploratory, *p*-values are reported for descriptive interpretation.

Correlation analysis was conducted using Spearman's correlation coefficients *r* ($r > 0.75$: good-to-excellent; $r = 0.50$-0.75: moderate-to-good; $r = 0.25$-0.49 fair; $r < 0.25$ no correlation) to examine associations between extracted features and both clinical targets (TJC, SJC, DAPSA, PASDAS) and patient-reported outcomes (PROs) (HAQ grip, fatigue, VAS pain, stiffness).

Finally, a univariate ROC analysis was carried out to evaluate the discriminatory power of individual features in distinguishing between patients with and without tender joints. The area under the ROC curve (AUC) was used to quantify each feature's classification performance.

Modelling

To assess the discriminatory power of typing dynamics, accelerometer features, and their combination, both linear models and tree-based models were employed. The goal was to classify participants into two groups based on the clinician-assessed **TJC68 hand subset** (wrists and MCP/PIP/DIP): **TJC Hand > 0** (any tender hand joint) vs **TJC Hand = 0** (none). Model performance was evaluated using area under the ROC curve (AUCROC), F1 score, and Cohen's kappa (κ). In all models, the following variables were included as covariates to account for potential confounding: sex, age, years of education, handedness, and self-reported typing style, i.e., one- vs. two-handed typing.

For linear models, logistic regression with L1 (LASSO), L2 (RIDGE), and ElasticNet regularization was used to address feature collinearity and control model complexity. For non-linear modelling, Random Forest and XGBoost classifiers were used to capture potential interactions between features and non-linear decision boundaries.

A nested cross-validation setup was used to ensure robust and unbiased model evaluation. The outer loop applied a leave-one-subject-out (LOSO) scheme to estimate generalizability, while the inner loop used a 5-fold subject-level cross-validation for hyperparameter tuning. This subject-wise split prevented information leakage, given the presence of repeated daily measurements per participant. Predictions were generated per day and aggregated at the subject level via majority voting, with final evaluation metrics computed from the aggregated predictions.

Hyperparameters for each model were tuned over standard search ranges. For XGBoost, parameters such as learning rate, number of estimators, tree depth, and regularization strength were optimized. Random Forest models were tuned by adjusting the number of trees, tree depth, and number of features considered at each split. Logistic regression models were explored using L1 (LASSO), L2 (RIDGE), and ElasticNet penalties with varying regularization strengths and solvers. All models included class balancing where applicable.

7.3 Results

For this analysis, only data from the first two weeks after the baseline visit were used, as clinical characteristics during this period are considered stable. Participants using typing tools (e.g., stylus or pen) or lacking sufficient data (fewer than 10 valid typing sessions) were excluded. A total of 46 participants were included in the final analysis. Socio-demographic and clinical characteristics of this cohort are summarized in **Table 20**. Additional details on participant recruitment and consent procedures are available in the Section 3.

Table 20 Demographic and clinical characteristics of the PsA patients (n = 46) at T0.

Variable	Median [IQR] (or Count %)	Missingness %
Demographics		
Age	53 [40, 59]	0%
Sex (male)	27 (59%)	0%
Education (years)	15 [12, 16]	0%
Smartphone usage confidence level	-	46%
Not confident	2 (8%)	-
Slightly confident	1 (4%)	-
Moderately confident	2 (8%)	-
Very confident	3 (12%)	-
Extremely confident	17 (68%)	-
Handedness (right-handed)	42 (91%)	0%
Typing Mode (one-handed)	33 (71%)	0%
Data availability		
Available days per patient	11 [7, 13]	0%
Typing sessions per patient	48 [19, 105]	0%
Clinical assessment		
DAPSA	36 [22, 92]	41%
PASDAS	3 [2, 4]	44%
TJC Hand (>0)	19 (41%)	0%
TJC Upper body (>0)	19 (41%)	0%
SJC Hand (>0)	10 (22%)	0%
SJC Upper body (>0)	11 (24%)	0%
HAQ Grip* (>0)	17 (37%)	33%
Fatigue (PSAID2)	3 [2, 4]	35%

Variable	Median [IQR] (or Count %)	Missingness %
VAS Pain	16 [7, 50]	33%
Stiffness	3 [1, 6]	0%
**HAQ Grip was computed as the maximum score among the three grip-related items.*		

The results of feature stability over time are shown in **Figure 25**. Typing features demonstrated moderate to good reliability across the 14-day period, while accelerometer features showed poor to moderate reliability. ICC values tended to stabilize after approximately 5 days, with higher values observed in the initial days. It is important to note that participants used a custom keyboard, which may differ in feel and behaviour from the standard Android keyboards provided by device manufacturers. This could have influenced early typing consistency.

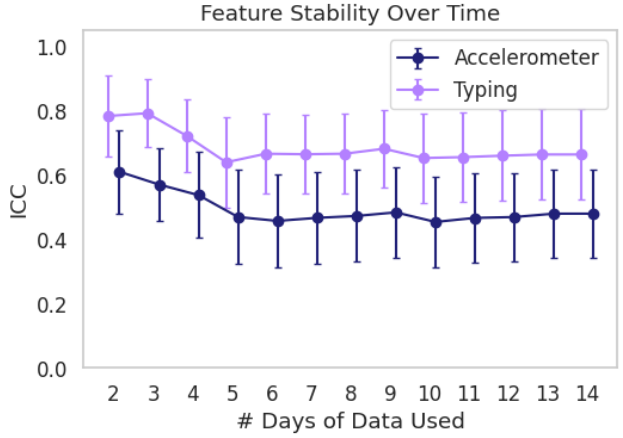


Figure 25 Intraclass correlation coefficient (ICC(3, k)) over increasing numbers of days used in analysis for typing and accelerometer features. Error bars indicate standard deviation across features.

Figure 26 shows the distribution of the most informative features identified by the Mann–Whitney U test ($p < 0.05$), comparing patients with and without tender joints in the hands. Patients with tender joints exhibited lower variability in both hold time and flight time, but higher skewness and kurtosis, indicating more asymmetric and peaked typing patterns. They also tended to type shorter words, as reflected in lower characters-per-word values. Regarding phone handling, patients with tender joints appeared to hold the phone in a more upright position, with a higher median tilt angle (closer to 90°) and lower median acceleration along the z-axis, consistent with a more vertical orientation compared to those without joint tenderness.

Correlation analysis revealed several fair associations between digital typing and accelerometer features and clinical or patient-reported outcomes (PROs), as shown in **Figure 27** and **Table 21**. Among all features, Chars per Word demonstrated the strongest correlations, showing consistent and moderate negative associations with TJC Hand (r = -0.419, p-value = 0.04), TJC Upper (r = -0.434, p-value = 0.003), SJC Hand, and SJC Upper indicating that patients with more joint involvement tend to type shorter words.

For timing-based features, Flight Std showed a fair negative correlation with TJC Hand (r = -0.381, p-value = 0.009), reflecting reduced variability in flight time among patients with tender joints. This was followed by positive correlations for Flight Skew and Flight Kurt, suggesting more asymmetric and peaked timing distributions. Additionally, Hold Std was negatively correlated with TJC Hand, and posture-related features such as Tilt Mean and Acc Z Median

were positively and negatively correlated, respectively, indicating more upright phone orientation in patients with hand tenderness.

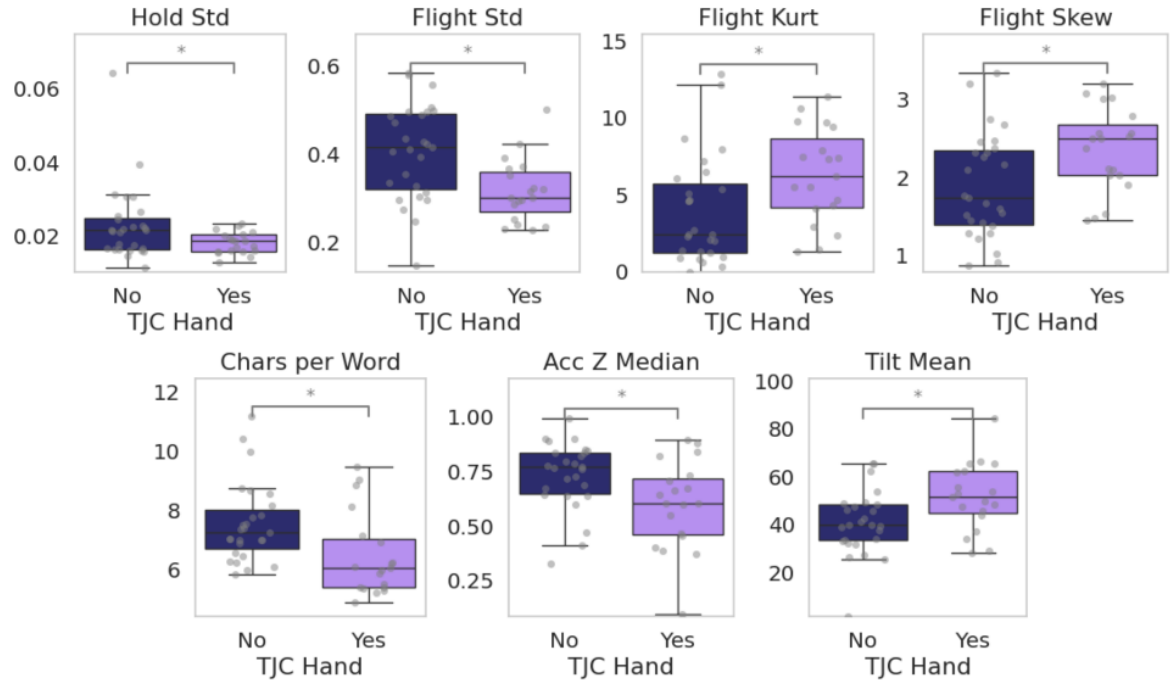


Figure 26 Distribution of selected features that significantly differed between participants with and without tender hand joints (*p < 0.05, Mann–Whitney U test).

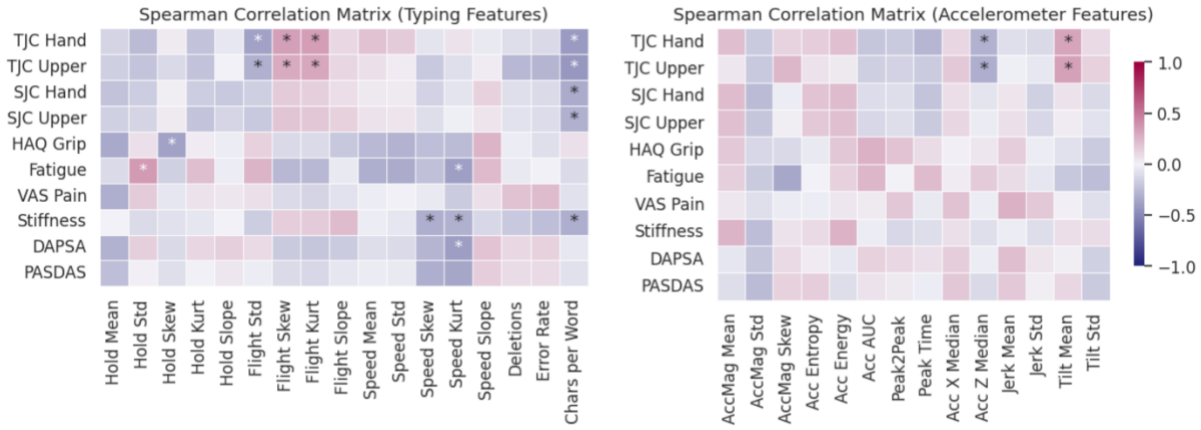


Figure 27 Spearman correlation matrices between digital features (typing on the left, accelerometer on the right) and clinical/PRO outcomes. Asterisks (*) denote p < 0.05.

Table 21 List of statistically significant Spearman correlations (p < 0.05) between selected digital features and clinical or PRO outcomes.

Features	Clinical/ PRO outcomes	Spearman's *r*	p-value
Chars per Word	TJC Hand	-0.419	0.004
Flight Std	TJC Hand	-0.381	0.009
Flight Kurt	TJC Hand	0.346	0.019
Flight Skew	TJC Hand	0.340	0.021
Tilt Mean	TJC Hand	0.317	0.034
Acc Z Median	TJC Hand	-0.305	0.042
Chars per Word	TJC Upper	-0.434	0.003

Features	Clinical/ PRO outcomes	Spearman's *r*	p-value
Tilt Mean	TJC Upper	0.328	0.028
Acc Z Median	TJC Upper	-0.327	0.028
Flight Std	TJC Upper	-0.316	0.033
Flight Kurt	TJC Upper	0.311	0.036
Flight Skew	TJC Upper	0.297	0.045
Chars per Word	SJC Hand	-0.344	0.019
Chars per Word	SJC Upper	-0.320	0.030
Hold Skew	HAQ Grip	-0.379	0.035
Speed Kurt	DAPSA	-0.402	0.037
Speed Kurt	Fatigue	-0.395	0.031
Hold Std	Fatigue	0.363	0.049
Speed Kurt	Stiffness	-0.313	0.034
Speed Skew	Stiffness	-0.324	0.028

For PROs, Speed Kurt showed consistent negative correlations with Fatigue, Stiffness, and DAPSA, while Speed Skew was also negatively associated with Stiffness. Hold Std showed a positive correlation with Fatigue, while Hold Skew was negatively correlated with HAQ Grip. The complete correlation matrices for typing and accelerometer features are presented in the Appendix (**Figure 83**; **Figure 84**).

Univariate ROC analysis results are presented in **Figure 28**, showcasing features with AUC greater than 0.60. The full list of features is available in the Appendix (**Table 60**). Among typing-derived features, flight time skewness and kurtosis demonstrated the strongest discriminatory performance for detecting TJC in the hand, followed by mean and standard deviation of typing speed. For accelerometer-derived features, tilt mean—reflecting phone orientation—achieved the highest AUC, followed by motion-related features such as acceleration energy, entropy, and mean acceleration magnitude.

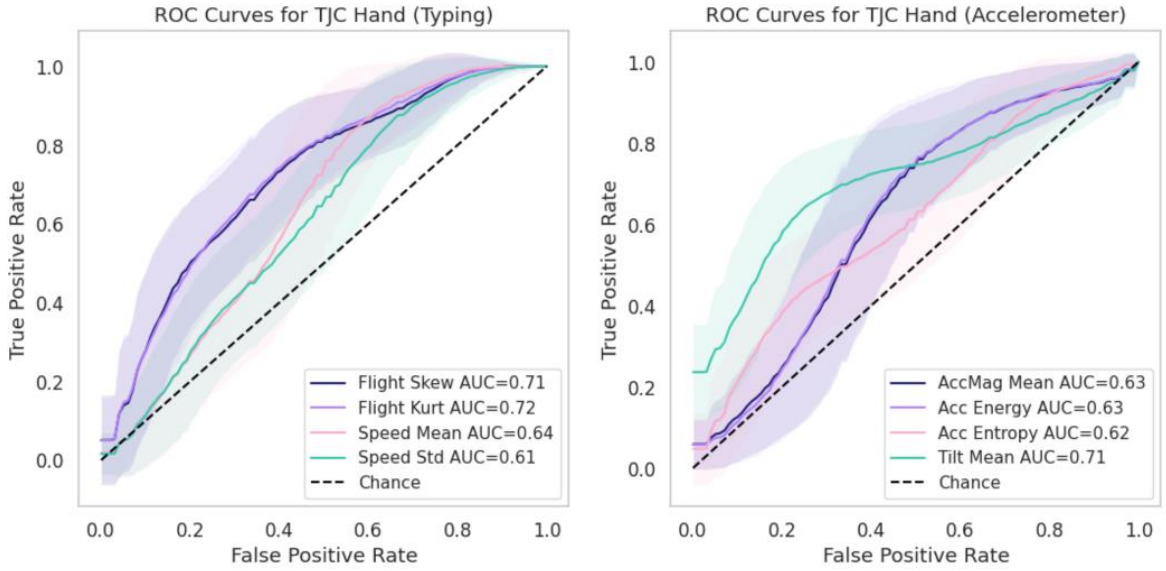


Figure 28 ROC curves for the classification of patients with tender joints in the hand using individual typing (left) and accelerometer (right) features. Only features with AUC > 0.60 are displayed. Shaded areas indicate 95% confidence intervals; dashed line denotes chance level.

Figure 29, **Figure 30**, and **Table 22** to **Table 24** summarize the results of the classification models used to distinguish patients with and without tender joints in the hands. **Figure 29** shows ROC curves for the best-performing logistic regression models trained on combined typing and accelerometer features, typing only, and accelerometer only. The highest performance was observed for the RIDGE model using combined features, with a mean AUC of 0.73. For accelerometer-only models, RIDGE also performed best (AUC = 0.70), while typing-only models reached an AUC of 0.70 with LASSO regularization.

Figure 30 presents the top 10 features ranked by the absolute value of their coefficients for the best-performing models. In the combined model, acceleration magnitude and energy were the most important accelerometer features, both contributing positively, while hold and flight time variability showed strong negative contributions. In the accelerometer-only model, average tilt had the highest positive contribution, while variability in acceleration magnitude contributed negatively. For the typing-only model, flight and hold time variability remained the most influential, with negative coefficients.

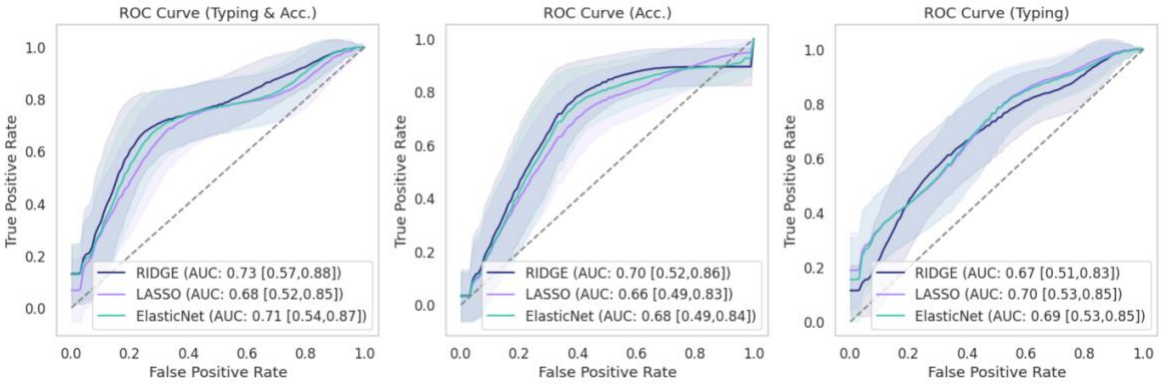


Figure 29 ROC curves for logistic regression models trained on typing and accelerometer features. Left: Combined typing and accelerometer features; Middle: Accelerometer-only; Right: Typing-only. Shaded areas represent 95% confidence intervals.

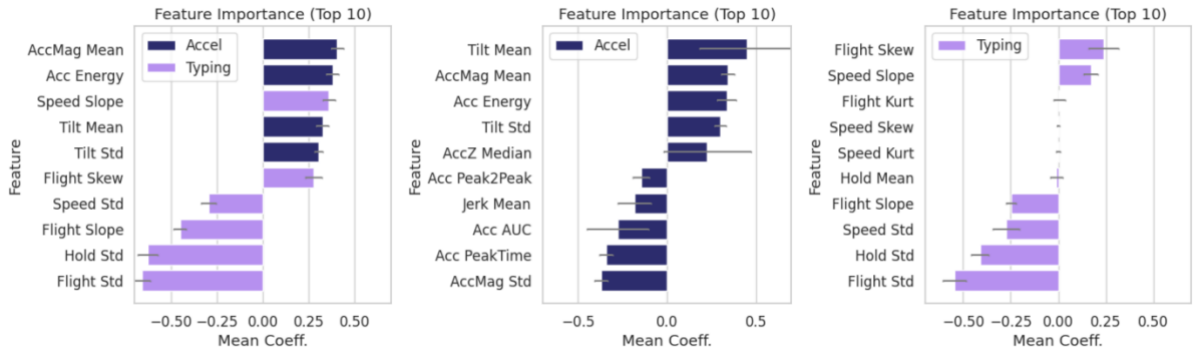


Figure 30 Top 10 most important features based on mean absolute coefficients from the best-performing models. Left: Combined typing and accelerometer model; Middle: Accelerometer-only; Right: Typing-only. Negative values indicate inverse associations with TJC Hand label.

Table 22, **Table 23**, and **Table 24** provide full model metrics, including AUC, Cohen's κ, and F1 score. Overall, logistic regression models outperformed tree-based models. The combined model achieved moderate agreement (κ = 0.52) and strong class balance (F1 = 0.76).

Table 22 Performance metrics (AUCROC, Cohen's κ, and F1 score with 95% CI) for models trained on combined typing and accelerometer features.

Model	AUCROC	κ	F1
RIDGE	**0.733** [0.569, 0.879]	**0.524** [0.261, 0.780]	**0.760** [0.627, 0.890]
LASSO	0.688 [0.520, 0.849]	0.449 [0.192, 0.697]	0.722 [0.589, 0.847]
ElasticNet	0.710 [0.536, 0.867]	0.491 [0.232, 0.737]	0.744 [0.608, 0.869]
RandomForest	0.649 [0.471, 0.831]	0.433 [0.160, 0.684]	0.713 [0.571, 0.842]
XGBoost	0.658 [0.522, 0.804]	0.291 [0.022, 0.561]	0.575 [0.387, 0.743]

Table 23 Performance metrics (AUCROC, Cohen's κ, and F1 score with 95% CI) for models trained on typing features.

Model	AUCROC	κ	F1
RIDGE	0.676 [0.512, 0.833]	0.382 [0.145, 0.617]	0.685 [0.564, 0.808]
LASSO	**0.701** [0.533, 0.851]	0.389 [0.160, 0.627]	0.685 [0.562, 0.808]
ElasticNet	0.697 [0.531, 0.849]	0.390 [0.156, 0.635]	0.687 [0.564, 0.817]
RandomForest	0.652 [0.474, 0.825]	**0.435** [0.155, 0.698]	**0.711** [0.565, 0.848]
XGBoost	0.645 [0.469, 0.815]	0.426 [0.161, 0.695]	0.710 [0.569, 0.846]

Table 24 Performance metrics (AUCROC, Cohen's κ, and F1 score with 95% CI) for models trained on accelerometer features.

Model	AUCROC	κ	F1
RIDGE	**0.698** [0.519, 0.859]	**0.477** [0.217, 0.733]	**0.734** [0.602, 0.866]
LASSO	0.664 [0.490, 0.826]	0.403 [0.155, 0.652]	0.696 [0.564, 0.826]
ElasticNet	0.680 [0.495, 0.845]	0.454 [0.189, 0.697]	0.723 [0.587, 0.848]
RandomForest	0.665 [0.491, 0.821]	0.381 [0.138, 0.609]	0.683 [0.552, 0.804]
XGBoost	0.647 [0.469, 0.813]	0.389 [0.112, 0.650]	0.692 [0.551, 0.825]

7.4 Discussion

7.4.1 Interpretation

Patients with tender hand joints—clinician-assessed TJC68 restricted to wrists and MCP/PIP/DIP—showed reduced variability in hold and flight times, and more asymmetric and peaked distributions, along with shorter words and a more upright phone posture. These differences were reflected in feature correlations with clinical measures, where typing features, particularly characters per word and flight variability, showed fair negative associations with joint counts. Motion and posture features such as tilt angle and vertical acceleration were also relevant.

Several individual features, including flight skewness and kurtosis and average tilt, showed fair discriminatory performance (AUC > 0.70) in univariate ROC analysis. Combining typing and accelerometer features in logistic regression models yielded the best results, with an AUC of 0.73 and moderate agreement (κ = 0.52) for detecting tender joints. Typing variability features (e.g., flight and hold time standard deviation) and motion/posture signals emerged as key contributors. Notably, some features, such as flight time and phone tilt, consistently

appeared across correlation analysis, univariate classification, and model importance rankings, underlining their potential relevance as candidate digital biomarkers.

7.4.2 Limitations

Limitations of the current analysis include the small sample size (n = 46), a single clinical visit (baseline), and a 14-day data collection window. Participants using typing tools (e.g., stylus or pen) were excluded; however, such behaviours may influence results in real-world scenarios and should be explored further. Similarly, one- vs two-handed typing, especially affecting typing speed and phone motion, was not explicitly modelled. While some adjustments were included in the analysis (e.g., self-reported typing style), these behaviours need more focused analysis. Medication or medication changes after baseline were not considered and may have influenced results, future models should account for treatment effects.

7.5 Work until D3.6

Future steps include validating the findings in a larger PDPID dataset as recruitment continues and incorporating longitudinal analysis as more clinical visits become available. The typing pipeline will also be expanded to include users of alternative input methods such as pens. One- vs. two-handed typing will be looked at more closely using unsupervised methods to understand how it affects typing-based digital biomarker signals. Finally, additional classification and regression models will be explored to cover clinical targets and PROs that were not modelled in this deliverable due to the current dataset size. Medication data will be integrated to adjust for treatment-related changes.

8 Physical activity level and gait characteristics from smartphone's IMU sensors

8.1 Background and objectives

Psoriatic arthritis (PsA) varies widely across patients and can limit physical function and quality of life, creating physical, psychological, and social burdens. Standard assessments—in-clinic visits and patient-reported outcomes—provide brief snapshots and are prone to recall bias, so they often miss fluctuations and unpredictable flares. Continuous, out-of-clinic monitoring with smartphones and smartwatches can unobtrusively capture real-world sensor data from which digital biomarkers can be derived, complementing routine care (Knitza et al., 2024). Most prior work in inflammatory arthritis has used wrist-worn smartwatches; however, smartphones have also been leveraged to develop digital biomarkers and to track physical activity (Karas et al., 2024), given their near-universal availability and frequent on-person carry.

This study aims to detect physical activity and activity-specific characteristics in patients with PsA by analysing patterns in smartphone IMU data (tri-axial accelerometer and gyroscope). We analyse these sensor streams to derive features representing activity levels—e.g., walking activity counts—and gait characteristics. We then evaluate these features as candidate digital biomarkers by assessing their associations with clinical measures and patient-reported outcomes that reflect changes in physical function.

8.2 Methods

8.2.1 Dataset(s)

To develop smartphone-based digital biomarkers (dBMs) for PsA, we used data from the first cohort of participants enrolled in the PDPID study (see Section 3). Sensor data were collected passively in real-world conditions during everyday smartphone use via the miPROLEPSIS app, which records tri-axial accelerometer and tri-axial gyroscope streams continuously (24/7) at around 20 Hz. The variables are summarized in **Table 25**.

Table 25 Data captured by the miPROLEPSIS app.

Variable	Description
Accelerometer	
timestamp	Timestamp in ms (Unix time)
x	Acceleration along the x axis (includes gravity) (m/s^2)
y	Acceleration along the y axis (includes gravity) (m/s^2)
z	Acceleration along the z axis (includes gravity) (m/s^2)
Gyroscope	
timestamp	Timestamp in ms (Unix time)
x	Angular velocity of x axis. (rad/s)
y	Angular velocity of y axis (rad/s)
z	Angular velocity of z axis (rad/s)

8.2.2 Development

Tri-axial accelerometer and gyroscope data were captured continuously (24/7) via the miPROLEPSIS smartphone app in free-living conditions, with unspecified on-body phone location. Accelerometer data were used to discriminate walking from non-walking activities; accelerometer and gyroscope signals were further processed to derive gait features (e.g., step count, step/stride regularity, cadence). Walking and non-walking segments were then used to compute walking volume, fragmentation, and intensity features. Features were aggregated at the daily level and summarized weekly for association with clinical scores and patient-reported outcomes. A schematic is shown in **Figure 31** Smartphone IMU data analysis overview.

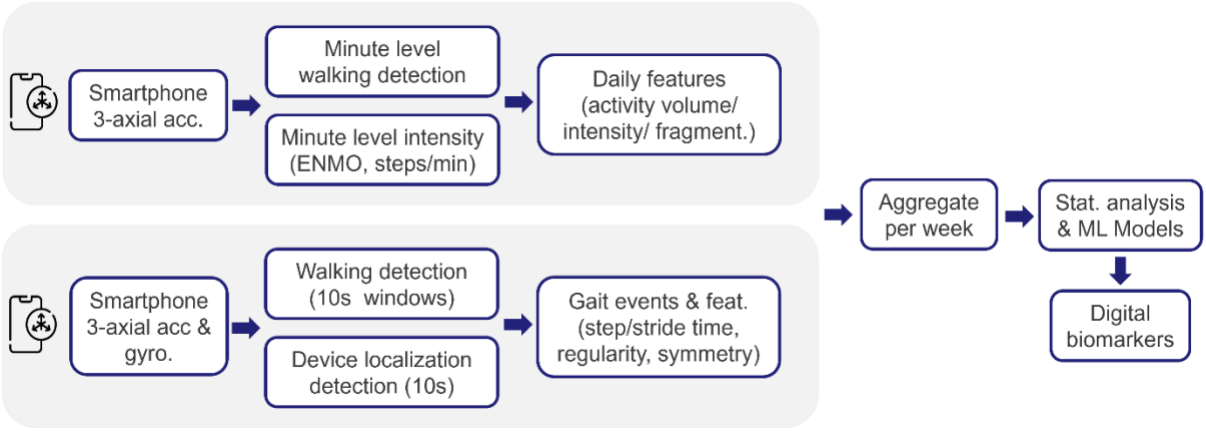


Figure 31 Smartphone IMU data analysis overview.

Physical activity detection

We modeled physical activity by detecting walking segments from raw tri-axial accelerometer data using a published walking-recognition algorithm (Straczkiewicz et al., 2023a). This method classifies walking based on intensity, periodicity, and minimum duration, and was validated on 20 publicly available annotated datasets spanning wrist-worn and phone recordings at multiple body locations. It has also been evaluated in an independent oncology cohort, demonstrating feasibility for free-living monitoring (Straczkiewicz et al., 2023b).

Raw acceleration was converted to gravitational units and resampled via linear interpolation to a uniform 10 Hz across axes. To mitigate effects of changing phone orientation and position, the tri-axial signal was reduced to a vector magnitude and operationalized as ENMO (Euclidean Norm Minus One g) subtracting the static gravitational component and truncating negatives to zero.

Signals were partitioned into 10-s windows. Candidate walking was identified using peak-to-peak amplitude in 1-s sub-windows and a predefined amplitude threshold. A continuous wavelet transform (CWT) then estimated the dominant step frequency and derived gait-harmonic parameters. Segments failing amplitude/frequency criteria or minimum duration were discarded; remaining segments were labelled as walking. Per-second step counts were computed from the estimated step frequency and summed over each walking period.

ENMO was expressed in milligravity (mg) and aggregated to 1-min epochs. Walking labels and per-second step counts were likewise aggregated to 1-min, yielding a per-minute activity vector (walking/non-walking indicator, steps/min, mean ENMO). Minutes were then classified as Inactive (0 steps/min; phone still or sedentary), Light activity (> 0 steps/min and cadence < 100 steps/min), and MVPA (moderate-to-vigorous physical activity; cadence ≥ 100 steps/min, consistent with a cut-point validated in a cross-sectional study of 80 healthy adults aged 41–60 years (Tudor-Locke et al., 2020).

From the per-minute series, we derived daily features capturing activity volume, intensity, and fragmentation, then summarized them weekly using mean values; we also reported weekly minima and maxima for daily step counts and total walking minutes. The full list of daily features is presented in **Table 26**.

Valid weeks require at least 72 h of data covering all 24 clock hours. This threshold is based on simulations in 29,765 UK Biobank wrist-accelerometry participants, where 72 h produced estimates within 10% of a 7-day measure (Doherty et al., 2017). Although our data come from smartphone accelerometers, we adopt this criterion to target general physical activity. Missing minutes were imputed using the mean at the same time of day on other observed days; for multiweek summaries we used, in order: same weekday and time, same day type (weekday/weekend) and time, then the overall time-of-day mean.

Table 26 List of daily physical-activity features extracted from smartphone accelerometer data.

Feature	Description
Activity volume	
Mean ENMO (mg)	Daily mean ENMO magnitude (mg)
Steps	Total step count per day
Walking (min)	Minutes spent walking per day
Inactive (min)	Total minutes classified inactive per day
Light (min)	Total minutes of light activity per day
MVPA (min)	Total minutes of moderate–vigorous activity per day
Activity fragmentation	
Avg Inactive (min)	Average length of inactive bouts (minutes)
Avg Light (min)	Average length of light bouts (minutes)
Avg MVPA (min)	Average length of MVPA bouts (minutes)
Max Inactive (min)	Longest inactive bout (minutes)
Max Light (min)	Longest light bout (minutes)
Max MVPA (min)	Longest MVPA bout (minutes)
Prob Inactive → Light	Probability next epoch is Light given current is Inactive
Prob Light → Inactive	Probability next epoch is Inactive given current is Light
Prob Inactive → MVPA	Probability next epoch is MVPA given current is Inactive
Inactive bouts/hr	Inactive bouts per hour
Mobility	
Walking bouts (≥10min)	Walking bouts ≥10 min (count)
Top 1 Cadence	Mean cadence of the top 1 highest step/min minute(s)
Top 30 Cadence	Mean cadence of the top 30 highest step/min minutes
95th Cadence	95th-percentile steps/min for the day

Gait analysis

To capture gait-specific features from smartphone IMU signals, we first identified walking segments (as above) and determined whether the phone was carried in a trouser pocket during those segments. On-body device localization was performed with a Random Forest classifier trained on the RealWorld dataset (Sztyler & Stuckenschmidt, 2016), which contains walking data with the phone at multiple body locations (chest, forearm, head, shin, thigh, upper arm, waist). From 10-s windows of accelerometer and gyroscope data, we extracted summary statistics (e.g., quantiles), inter-axis correlations, dominant frequencies, and peak counts, and used these features to classify device location. Only windows classified as thigh/waist (i.e., pocket carry) were retained for gait analysis. Classification results are reported in the Appendix (**Table 66**; **Figure 86**).

For gait estimation, accelerometer and gyroscope signals in 10-s windows labeled as walking and thigh/waist were processed with the Mobilise-D algorithm pipeline (Kirk et al., 2024; Micó-Amigo et al., 2023) using its official implementation in the mobgap[6] Python library. The pipeline yields initial contacts (step events), from which we derived step and stride durations, regularity, and gait symmetry (see **Table 27** for definitions). For each valid window, we computed the median step and stride times and their mean absolute deviations (MADs); these window-level metrics were then averaged per day and finally averaged across the 14-day period.

Table 27 List of gait features extracted per 10-s walking segment (smartphone IMU).

Feature	Description
Step Reg (step regularity)	Autocorrelation coefficient at the step lag of the acceleration vector magnitude; higher values indicate more periodic step-to-step pattern (unitless)
Stride Reg (stride regularity)	Autocorrelation coefficient at the stride lag of the acceleration vector magnitude; higher values indicate more periodic stride-to-stride pattern (unitless)
Gait Symm	Absolute stride minus step regularity (symmetry)
Cadence	Walking cadence (steps/min)
Step Med	Median step time (interval between consecutive steps), in seconds
Step MAD	Mean absolute deviation of step time
Stride Med	Median stride time (interval between consecutive strides), in seconds
Stride MAD	Mean absolute deviation of stride time

8.2.3 Verification

We did not conduct formal device-level verification (e.g., testing against a reference-grade accelerometer) because smartphone hardware was unconstrained beyond minimal OS version requirements and participants used their own devices. Instead, verification relied on preprocessing and quality control at the stream level—checking timestamp integrity (monotonicity, duplicates, gaps); screening sampling rate and excluding irregular segments (with interpolation where appropriate); removing clipped/saturated segments; mitigating orientation via vector magnitude/ENMO; and applying device localization to retain thigh/waist in-pocket windows for gait analyses. We also enforced valid-week coverage (≥72 h spanning all 24 clock hours) and used time-of-day imputation for short gaps to keep derived day-level features representative and unbiased.

[6] https://github.com/mobilise-d/mobgap/

Coverage before quality control is shown in **Figure 32** (accelerometer) and **Figure 33** (gyroscope). For the accelerometer data, we collected 1328 days in total (median 14 [13–14] days per participant), i.e., 86% of the expected 14-day window; most participants contributed all 14 days, while a small subset provided none due to hardware/OS constraints. For the gyroscope data, we collected 1045 days (median 14 [1–14]), i.e., 67% of expected days; ~25 participants contributed no gyroscope data despite providing accelerometer data. This lower gyroscope coverage likely reflects Android heterogeneity and power management—gyroscope capture is more resource-intensive and often throttled in the background. Early days occasionally show missing data from misconfiguration or smartphone model-specific restrictions.

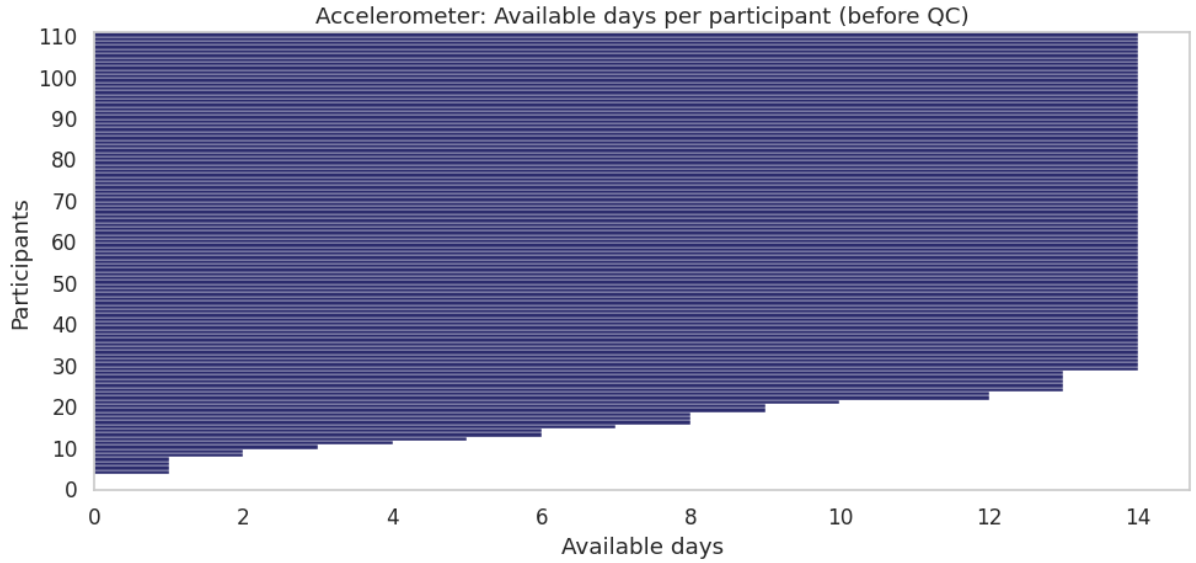


Figure 32 Daily coverage of smartphone accelerometer data during the first two weeks post-recruitment, prior to quality control.

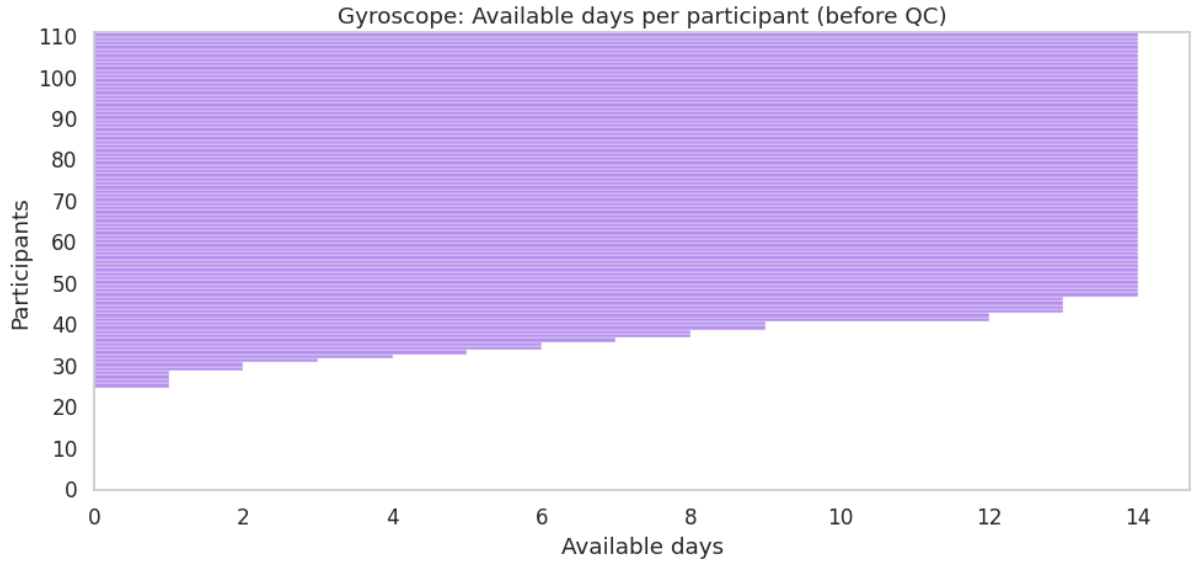


Figure 33 Daily coverage of smartphone gyroscope data during the first two weeks post-recruitment, prior to quality control.

Figure 34 shows post-QC accelerometer coverage for the participants retained for analysis (n=95; 85.6%): 1294 days in total (median 14 [14–14]), corresponding to 97% of the expected

days for the retained cohort and 83% relative to all 111 enrolled. **Figure 35** illustrates same-time, same day-type imputation for steps/min; for example, a ~1.5-hour gap is filled using the participant's weekday/weekend profile at the same clock minutes, preserving 24-h completeness of daily features, whereas leaving such gaps empty would bias day-level summaries.

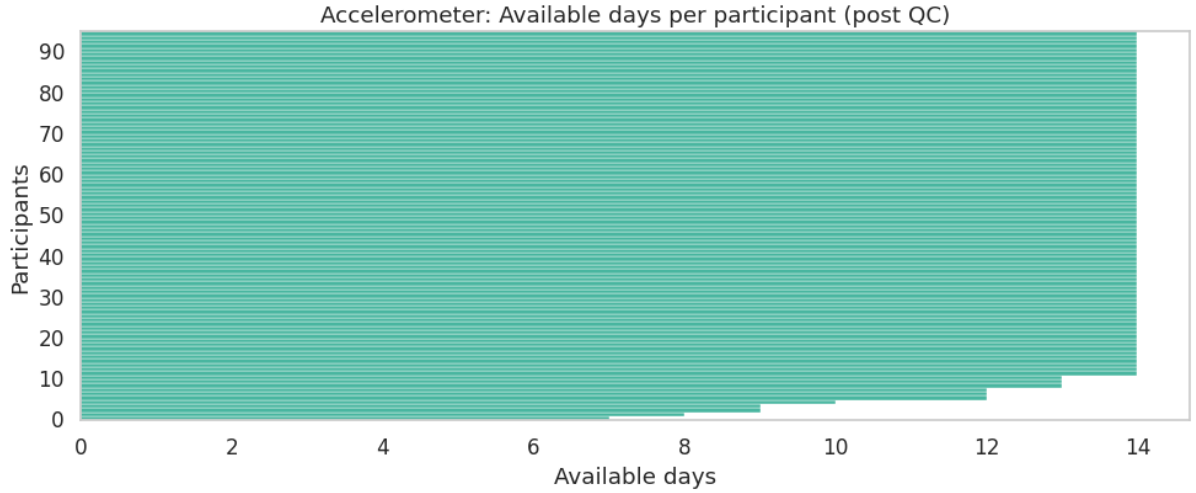


Figure 34 Daily coverage of smartphone accelerometer data in the first two weeks after recruitment, after quality control. Weeks counted as valid if ≥72 h of data covered all 24 h; shown sample reflects those retained for analysis.

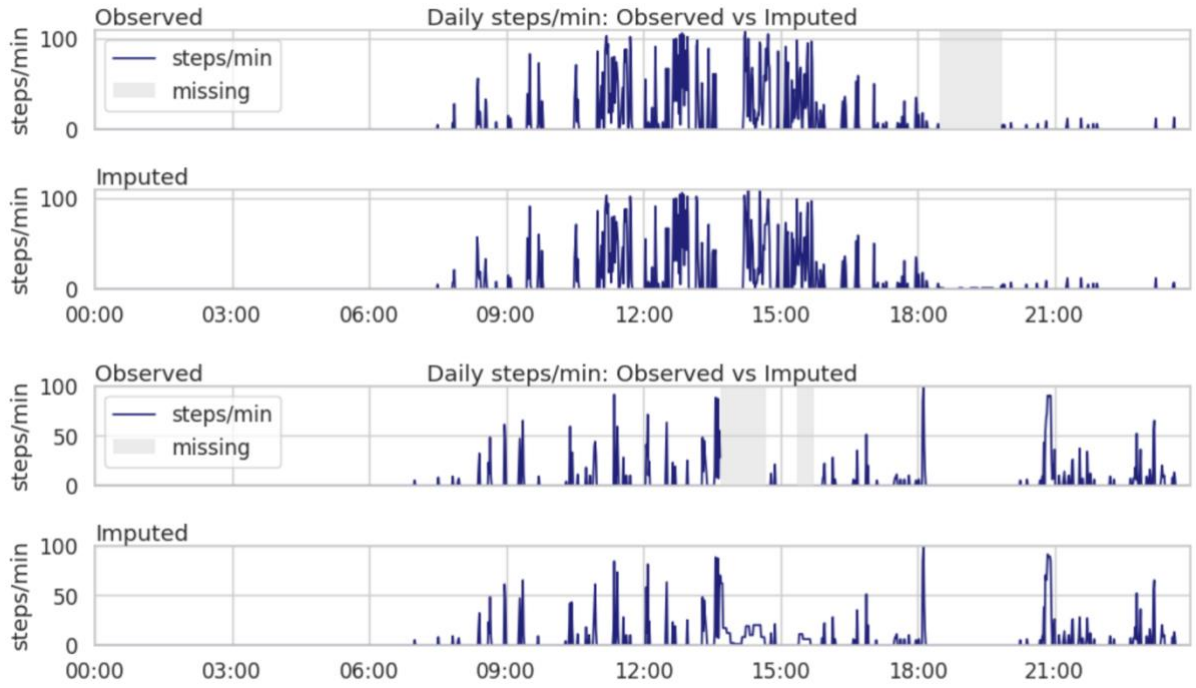


Figure 35 Example of same-time, same day-type imputation. For two days, the top panel shows the observed steps/min with missing intervals (grey); the bottom panel shows the imputed series obtained from the average at the same time of day across other days.

Figure 36 and **Figure 37** present gait-analysis availability—walking segments with both accelerometer and gyroscope present and sufficiently clean to detect gait events. Without on-body restriction (all locations), 52 participants (47%) contributed segments, totalling 5,226

segments (median 34 [7–106]) across 258 participant-days (median 4 [2–8]). With the in-pocket restriction, 44 participants (40%) contributed segments—85% of those with any location—with 2635 segments (median 16 [2–81]) across 169 participant-days (median 3 [1–6]). Requiring ≥10 valid in-pocket walking bouts yielded 29 participants (26% of all; 66% of in-pocket contributors), with 2614 segments (median 47 [18–115]) across 150 participant-days (median 4 [3–7]).

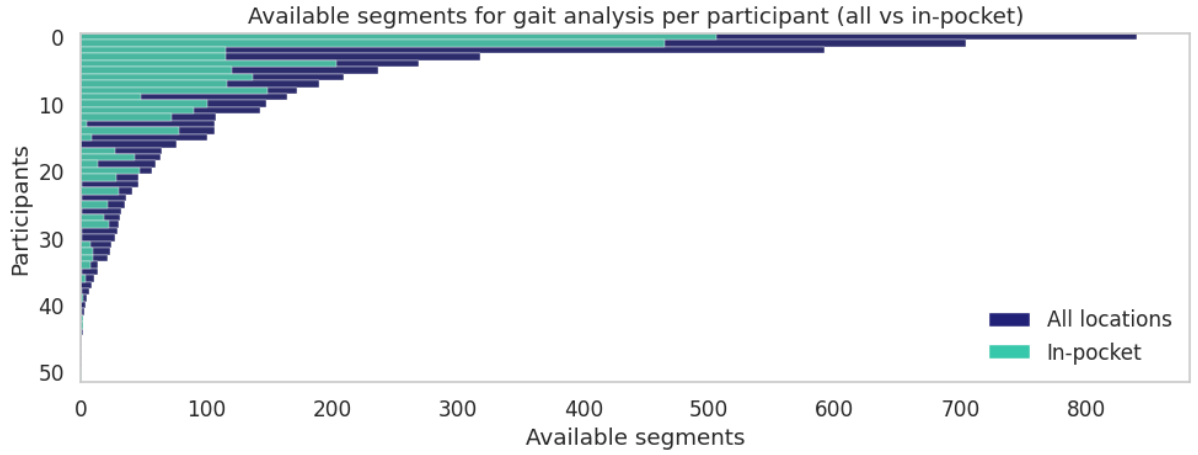


Figure 36 Number of valid walking segments per participant from the smartphone accelerometer and gyroscope, shown without on-body location restriction ("All locations") and with the "in-pocket" restriction.

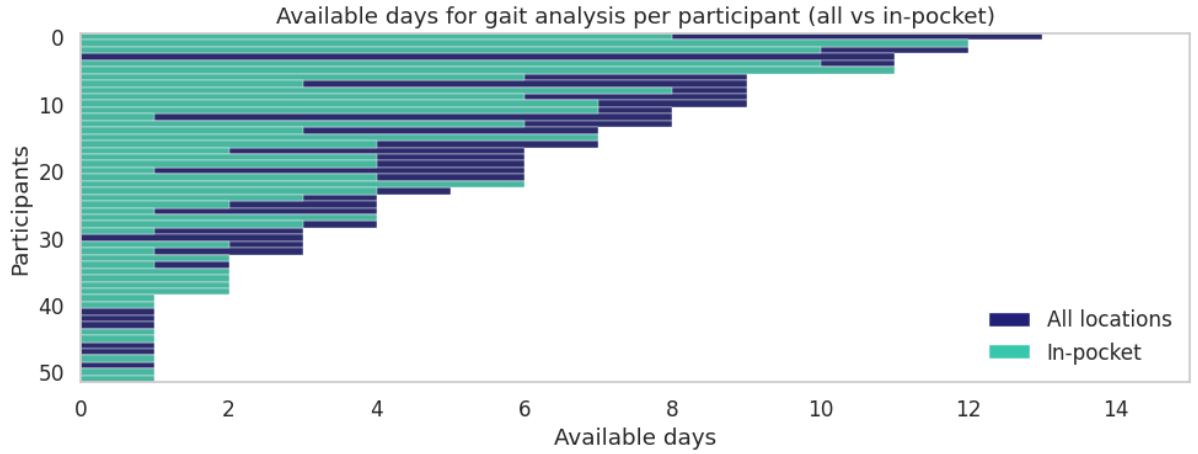


Figure 37 Daily availability of valid walking segments per participant from the smartphone accelerometer and gyroscope, shown without on-body location restriction ("All locations") and with the "in-pocket" restriction.

8.2.4 Validation

Statistical analysis

To assess reliability across repeated measures, we computed the intraclass correlation ICC(3, k) for each feature—over 14 days for physical-activity features and 7 days for gait features. This two-way mixed-effects model for average measures estimates the consistency of each feature across study days within individuals, where *k* represents the number of repeated daily measurements. ICC values below 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, between 0.75 and 0.9 indicate good reliability, and values above 0.9 indicate excellent reliability (Koo & Li, 2016).

Group comparisons were performed using the non-parametric Mann–Whitney U test. Given the exploratory nature of the study, we set $p < 0.05$ without multiplicity correction and report p-values descriptively. Physical-activity features were compared between: (i) flare vs. no flare (both physician-reported and patient-reported), (ii) MDA achieved vs. not achieved, (iii) presence vs. absence of tender lower-extremity joints, and (iv) HAQ walking difficulty vs. none (item: "Are you able to walk outdoors on flat ground?"). For gait features, comparisons focused on lower-extremity tenderness and HAQ walking difficulty.

Associations with clinical status were examined using Spearman's correlation coefficients *r* (r > 0.75 good–excellent; 0.50–0.75 moderate–good; 0.25–0.49 fair; <0.25 none). Correlates included clinical targets (TJC, SJC, DAPSA, PASDAS) and patient-reported outcomes (HAQ, PSAID-12 total and items: fatigue, pain, and stiffness).

Modelling

To evaluate discriminative utility, we built classifiers using both linear and tree-based models. Binary tasks included:

- Lower-extremity tenderness (yes/no),
- MDA (achieved/not achieved),
- Flare status—physician-reported (flare/no flare),
- Flare status—patient-reported (flare/no flare), and
- PRO impact thresholds: PSAID pain ≤ 4 vs. > 4, PSAID fatigue ≤ 4 vs. > 4 (Gossec et al., 2024), and stiffness ≤ 4 vs. > 4 (all on 0–10 scales).

Model performance was evaluated using area under the ROC curve (AUCROC), F1 score, and Cohen's kappa (κ). In all models, the following variables were included as covariates to account for potential confounding: sex, age, BMI, and disease duration.

For linear models, logistic regression with L1 (LASSO), L2 (RIDGE), and ElasticNet regularization was used to address feature collinearity and control model complexity. For non-linear modelling, Random Forest (RF) and XGBoost (XG) classifiers were used to capture potential interactions between features and non-linear decision boundaries.

A nested cross-validation setup was used to ensure robust and unbiased model evaluation. The outer loop applied a leave-one-subject-out (LOSO) scheme to estimate generalizability, while the inner loop used a 5-fold subject-level cross-validation for hyperparameter tuning.

Hyperparameters for each model were tuned over standard search ranges. For XGBoost, parameters such as learning rate, number of estimators, tree depth, and regularization strength were optimized. Random Forest models were tuned by adjusting the number of trees, tree depth, and number of features considered at each split. Logistic regression models were explored using L1 (LASSO), L2 (RIDGE), and ElasticNet penalties with varying regularization strengths and solvers. All models included class balancing where applicable.

8.3 Results

Physical activity results

For this analysis, we used data from the first 14 days after the baseline visit, a window during which clinical characteristics were expected to remain stable. Weeks were deemed valid if they contained ≥72 h of data spanning all 24 clock hours. Missing minutes were imputed using the mean at the same time of day from other observed days in the same period (see Development for details). A total of 95 participants were included in the final analysis. Socio-demographic and clinical characteristics of the cohort are summarized in **Table 28**; recruitment

and consent procedures are described in Section 3. Distributions of indicative feature values for the full cohort are presented in **Table 29**.

Table 28 Demographic and clinical characteristics of the PsA patients (n=95) at T0.

Variable	Median [IQR] (or Count %)	Missingness %
Demographics		
Age	55 [47, 61]	0%
Sex (male)	47 (50%)	0%
Education (years)	14 [12, 16]	1%
BMI	28 [25, 33]	4%
BMI ≥ 30	32 (34%)	4%
Disease duration (years)	8 [4, 15]	10%
Data availability		
Available days per patient	14 [14, 14]	0%
Available days total	1294 (97%)	0%
Clinical assessment		
Disease phenotype		33%
Polyarticular	34 (53%)	
Oligoarticular	14 (22%)	
Monoarticular	4 (6%)	
Axial	9 (14%)	
Dactylitis	2 (3%)	
Entheseal	1 (2%)	
TJC	1 [0, 7]	0%
SJC	0 [0, 1]	0%
TJC Lower body (>0)	45 (47%)	0%
SJC Lower body (>0)	13 (14%)	0%
DAPSA	43 [23, 87]	50%
PASDAS	3.4 [2.3, 4.0]	43%
MDA (achieved)	29 (50%)	39%
Flare (physician) (yes)	20 (22%)	3%
PROs		
Flare (self-reported)	12 (18%)	28%
PSAID12	1.8 [0.9, 3.9]	25%
HAQ	0.6 [0.0, 1.0]	24%
HAQ Walk (>0)	24 (35%)	24%
VAS Pain	23 [9, 50]	37%
Stiffness	3 [1, 6]	0%
Stiffness (>4)	43 (45%)	0%

Variable	Median [IQR] (or Count %)	Missingness %
Fatigue (PSAID12)	2 [1, 5]	25%
Fatigue (>4)	24 (33%)	25%
Pain (PSAID12)	3 [1, 5]	25%
Pain (>4)	23 (32%)	25%
HAQ Walk: "Are you able to walk outdoors on flat ground?" Response scale (0–3): 0 = without any difficulty; 1 = with some difficulty; 2 = with much difficulty; 3 = unable to do. *TJC/SJC Lower body: Tender/Swollen Joint Count limited to lower-extremity joints.* *TJC/SJC Upper body: Tender/Swollen Joint Count limited to upper-extremity joints.*		

Table 29 Distribution of representative features extracted from smartphone accelerometer data for the full cohort (n=95) during the first 14 days post-baseline; features were computed daily and averaged per subject.

Feature	Median [IQR]) (n=95)
Activity volume	
Steps	2926 [1803, 4433]
Walking (min)	103 [68, 154]
Mean ENMO (mg)	11 [8, 14]
Inactive (min)	1337 [1286, 1372]
Light (min)	89 [61, 147]
MVPA (min)	5 [3, 10]
Activity fragmentation	
Max Inactive (min)	404 [353, 454]
Max Light (min)	10 [7, 14]
Max MVPA (min)	2 [1, 3]
Inactive → Light	0.03 [0.02, 0.04]
Light → Inactive	0.48 [0.41, 0.54]
Walk bouts (≥10min)	1 [1, 2]
Mobility	
Top 1 Cadence	100 [88, 110]
Top 30 Cadence	58 [44, 74]

Figure 38 summarizes feature stability over time, showing how reliability changes as additional days of data are included. Most activity-volume, fragmentation, and mobility metrics showed poor reliability over the 14-day window. Exceptions included light-activity minutes, inactive minutes, walking minutes, mean ENMO, and the Light↔Inactive transition probabilities, which reached moderate reliability. ICC trajectories tended to stabilize after ~7 days.

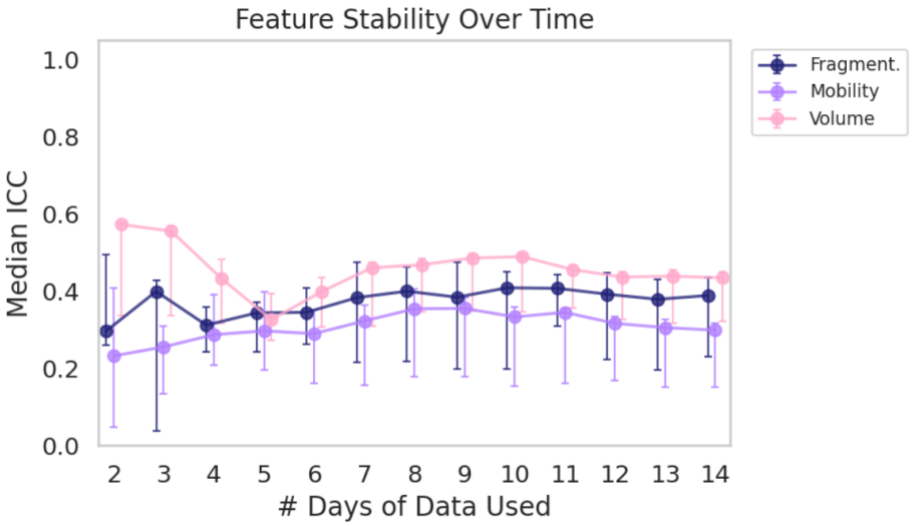


Figure 38 Feature stability over time. Median ICC(3,k) for smartphone-based physical-activity features as a function of days of data included. Curves represent feature groups (Volume, Fragmentation, Mobility); points are the median values across features within each group, and error bars show the standard deviation across features.

Figure 39-**Figure 43** show the distributions of the most informative features identified by the Mann–Whitney U test. Participants reporting walking difficulty spent less time in MVPA, had shorter sustained MVPA bouts, and exhibited lower cadence (Top 30 and 95th-percentile) than those without difficulty. In the lower-extremity pain comparison, the probability of transitioning from Light to Inactive was higher in the pain group ($p < 0.001$), whereas the no-pain group showed more sustained 10-min walking bouts, longer light-activity bouts, higher weekly maximum steps, and higher cadence.

For physician-reported flare, differences were limited; only maximum MVPA-bout duration was higher among those without flare, although distributions suggested a higher Light→Inactive transition during flare. Patient-reported flare showed a clearer pattern: participants without flare spent more time in MVPA, had longer MVPA bouts, lower Light→Inactive transition probability, and higher cadence. Finally, participants achieving MDA accumulated more light-activity time and more sustained 10-min walking bouts.

Overall, time spent in MVPA, the continuity of MVPA bouts, cadence metrics, and the Light→Inactive transition probability were the most informative across groups. Group-wise summaries are provided in the Appendix (**Table 61** to **Table 65**).

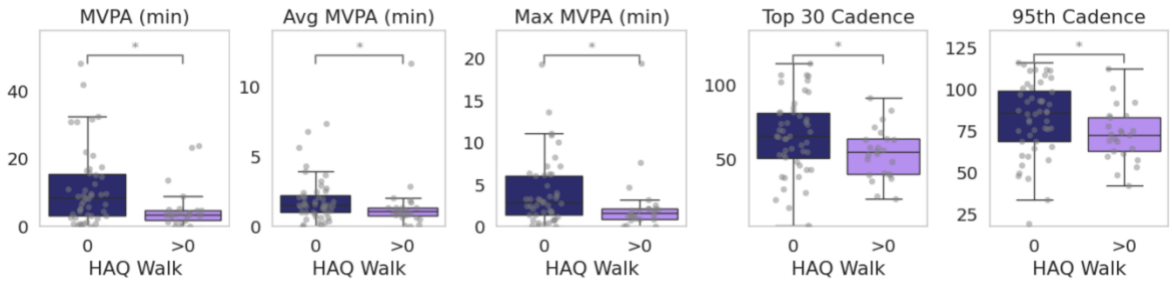


Figure 39 Distributions of selected features showing significant differences between participants with and without walking difficulty (*p < 0.05, Mann–Whitney U test).

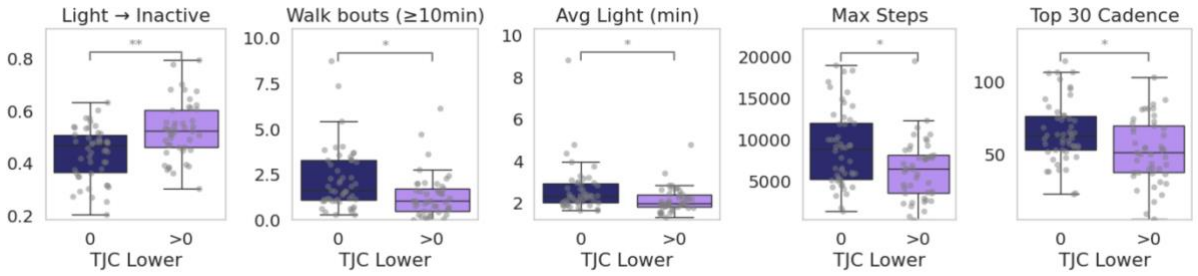


Figure 40 Distributions of selected features showing significant differences between participants with and without lower-extremity tender joints (*p < 0.05, **p < 0.001, Mann–Whitney U test).

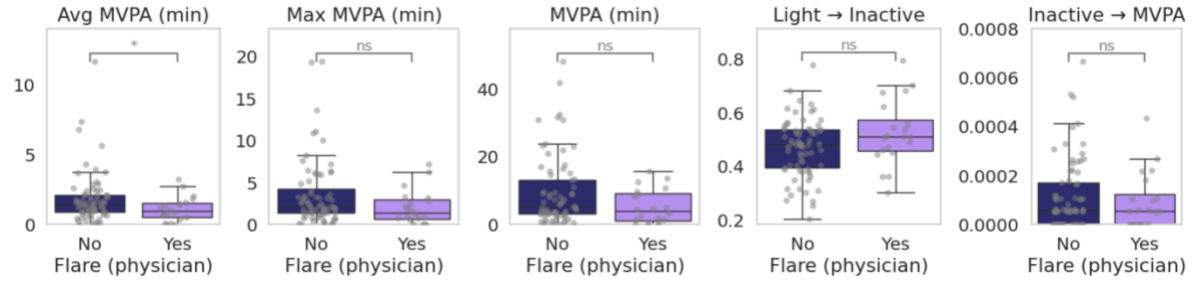


Figure 41 Distributions of selected features for participants with and without physician-reported flare; significant differences, where present, are indicated (*p < 0.05, Mann–Whitney U test).

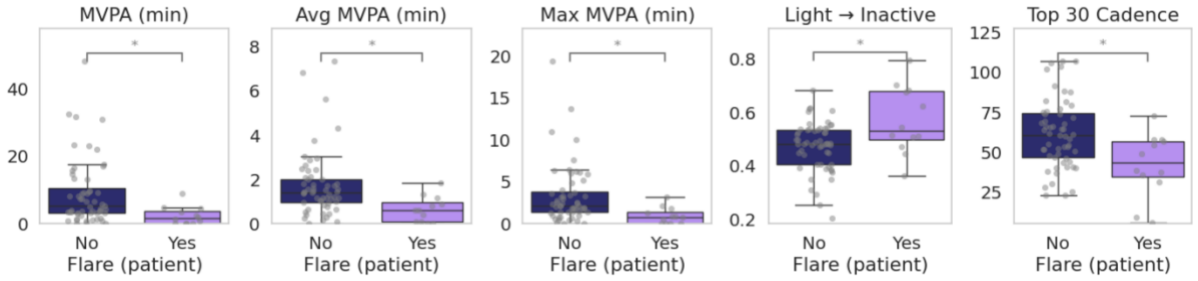


Figure 42 Distributions of selected features showing significant differences between participants with and without patient-reported flare (*p < 0.05, Mann–Whitney U test).

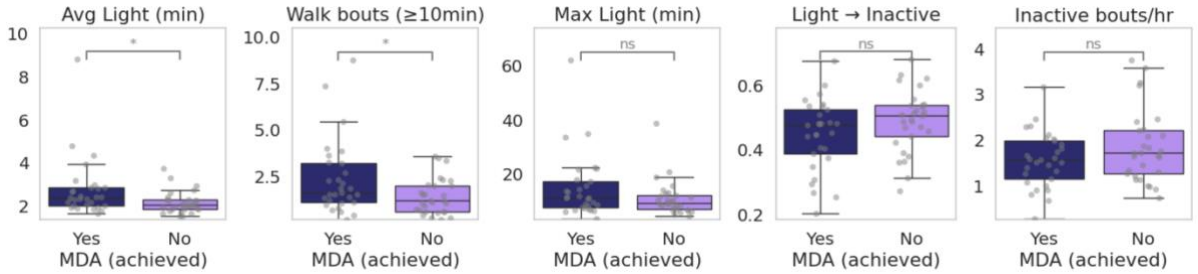


Figure 43 Distributions of selected features for participants achieving vs. not achieving MDA; significant differences, where present, are indicated (*p < 0.05, Mann–Whitney U test).

Results for the Spearman correlation analysis are shown in **Figure 44**, with the 20 strongest associations (by |*r*|) listed in **Table 30**. Higher disease activity and joint counts were consistently associated with fewer sustained 10-min walking bouts, less MVPA time, shorter maximum/average MVPA bouts, lower cadence (Top-30 and 95th percentile), and fewer steps (e.g., walk bouts ≥10 min vs. SJC66 $r = -0.489$; MVPA minutes vs. PASDAS $r = -0.441$; Top-30 cadence vs. SJC66 $r = -0.457$).

Symptom and function measures showed similar patterns: stiffness related to lower cadence and fewer sustained walking bouts, and a higher Light→Inactive transition probability related to more stiffness. Functional impairment (HAQ, HAQ-Walk) correlated negatively with MVPA metrics and cadence. Fragmentation measures were informative: the Light→Inactive transition probability was positively associated with disease activity and tenderness (e.g., PASDAS r = 0.358; DAPSA r = 0.357; TJC68 r = 0.437). Overall, better clinical profiles align with more MVPA, more sustained walking, higher cadence, and fewer transitions to inactivity; worse profiles show the opposite pattern. The full pairwise correlation structure is provided in the Appendix (**Figure 85**).

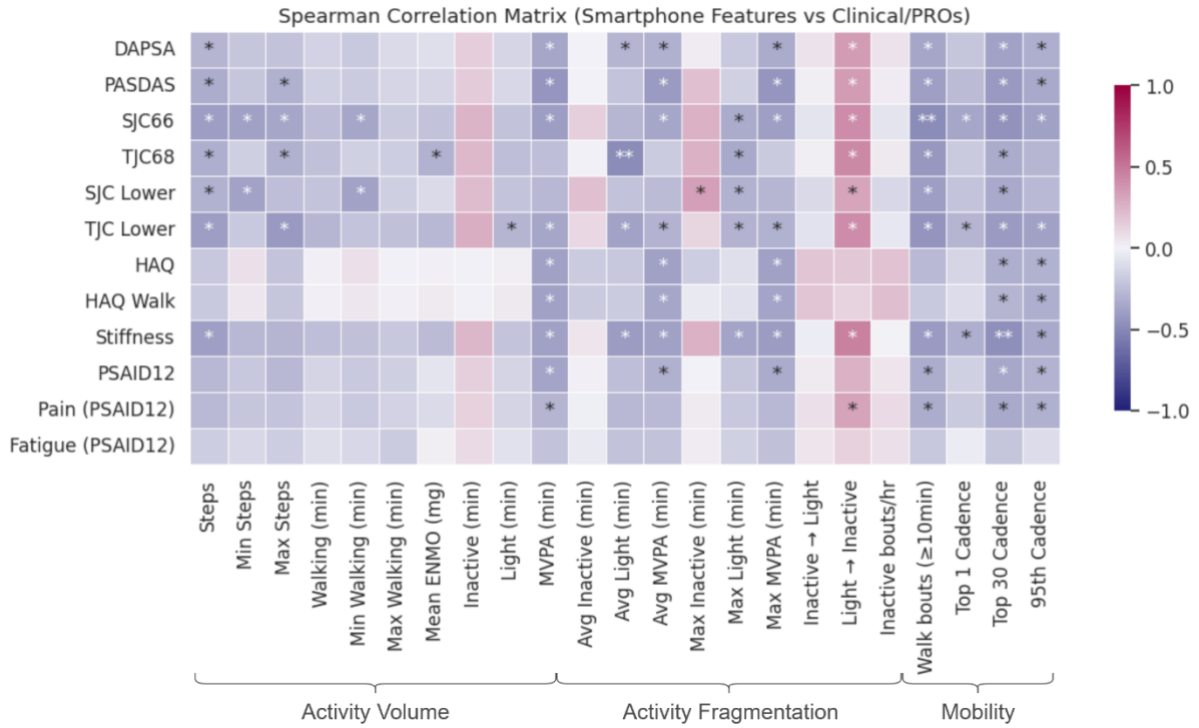


Figure 44 Spearman correlation matrix between smartphone-derived features and clinical/PRO outcomes. Cells show Spearman's r (color scale −1 to 1). Asterisks indicate significance: $p < 0.05$ (*), $p < 0.001$ (**).

Table 30 Top Spearman correlations between smartphone-derived features and clinical/PRO outcomes. Pairs with the 20 largest absolute correlations are reported.

Features	Clinical/ PRO outcomes	Spearman's r	p-value
Avg Light (min)	TJC68	-0.490	< 0.001
Walk bouts (≥10min)	SJC66	-0.489	< 0.001
Top 30 Cadence	Stiffness	-0.473	< 0.001
Light → Inactive	Stiffness	0.464	0.001
Top 30 Cadence	SJC66	-0.457	0.001
Walk bouts (≥10min)	TJC Lower	-0.452	0.002
MVPA (min)	PASDAS	-0.441	0.002
Max MVPA (min)	PASDAS	-0.440	0.002
Light → Inactive	TJC68	0.437	0.002
Walk bouts (≥10min)	TJC68	-0.434	0.003

Features	Clinical/ PRO outcomes	Spearman's r	p-value
Top 30 Cadence	PASDAS	-0.429	0.003
Max Steps	TJC Lower	-0.424	0.003
Light → Inactive	SJC66	0.419	0.004
Light → Inactive	TJC Lower	0.417	0.004
Top 30 Cadence	TJC Lower	-0.417	0.004
Avg MVPA (min)	PASDAS	-0.416	0.004
Walk bouts (≥10min)	Stiffness	-0.415	0.004
Avg Light (min)	Stiffness	-0.412	0.004
Max MVPA (min)	Stiffness	-0.408	0.005
Steps	SJC66	-0.401	0.006

Results for the classification models are summarized in **Table 31**. Performance was best for patient-reported flare (XG AUC-ROC = 0.77 [0.64, 0.88]; κ = 0.37; F1 = 0.53) and MDA (XG AUC-ROC = 0.70 [0.56, 0.83]; κ = 0.52; F1 = 0.76). Stiffness was also moderately predicted with logistic regression model (LASSO) (AUC-ROC = 0.70 [0.59–0.81]; κ = 0.41; F1 = 0.67). Lower-pain and patient-reported flare achieved moderate performance with logistic models (AUC-ROC ≈ 0.65–0.72). Tasks such as fatigue, physician-reported flare, and pain were weaker (AUC-ROC ≤ 0.60 for the best tree models). Across tasks, influential predictors (from logistic coefficients; see **Figure 45** and **Figure 46**) included sustained 10-min walking bouts, MVPA minutes and MVPA bout duration, Light→Inactive transition probability, cadence (Top-30/95th), and step counts, alongside covariates (sex, age, BMI, disease duration).

Table 31 Best model performance by outcome. For each outcome, the best-performing classifier is reported with AUC-ROC, Cohen's *κ*, and F1 score (brackets show 95% CIs).

Outcome	Model	AUCROC	κ	F1
Logistic regression				
Flare (phys.)	ElasticNet	**0.60** [0.46, 0.74]	0.22 [0.04, 0.42]	0.45 [0.29, 0.59]
MDA	ElasticNet	0.59 [0.44, 0.75]	0.33 [0.09, 0.56]	0.71 [0.53, 0.84]
Flare (pat.)	LASSO	0.72 [0.55, 0.87]	0.35 [0.15, 0.60]	0.51 [0.32, 0.70]
Lower pain	LASSO	**0.65** [0.53, 0.76]	0.31 [0.13, 0.49]	0.62 [0.50, 0.74]
Pain	LASSO	**0.58** [0.46, 0.71]	0.23 [0.05, 0.42]	0.54 [0.38, 0.68]
Stiffness	LASSO	**0.70** [0.59, 0.81]	0.41 [0.24, 0.58]	0.67 [0.55, 0.78]
Fatigue	LASSO	**0.59** [0.46, 0.73]	0.28 [0.08, 0.51]	0.54 [0.39, 0.68]
Decision trees				
Flare (phys.)	XG	0.47 [0.33, 0.62]	0.09 [-0.08, 0.29]	0.35 [0.20, 0.50]
MDA	XG	**0.70** [0.56, 0.83]	0.52 [0.31, 0.73]	0.76 [0.63, 0.87]
Flare (pat.)	XG	**0.77** [0.64, 0.88]	0.37 [0.19, 0.59]	0.53 [0.35, 0.70]
Lower pain	RF	0.64 [0.52, 0.76]	0.32 [0.14, 0.51]	0.65 [0.53, 0.76]
Pain	XG	0.54 [0.39, 0.70]	0.18 [-0.03, 0.41]	0.47 [0.33, 0.63]
Stiffness	XG	0.63 [0.52, 0.75]	0.32 [0.13, 0.50]	0.59 [0.47, 0.71]

Outcome	Model	AUCROC	κ	F1
Fatigue	RF	0.47 [0.33, 0.60]	0.16 [-0.03, 0.33]	0.53 [0.37, 0.67]

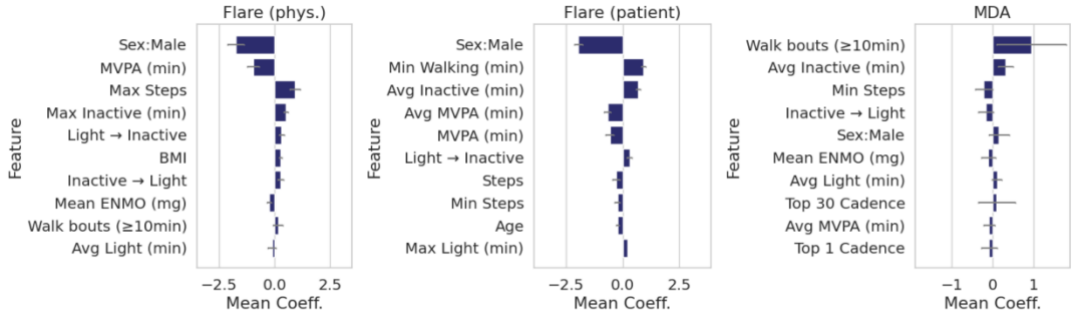


Figure 45 Top coefficients from best logistic models (flare/MDA). Bars show the top 10 coefficients (mean values across inner folds; error bars ±1 SD) for: physician-reported flare, patient-reported flare, and MDA (from left to right). Positive coefficients indicate higher odds of the positive class (flare or MDA achieved).

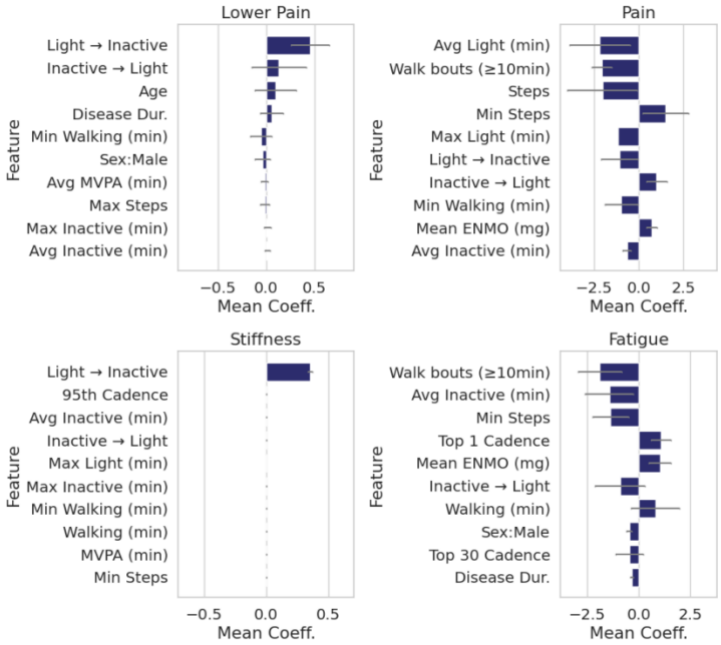


Figure 46 Top coefficients from best logistic models (pain/stiffness/fatigue). Bars show the top 10 coefficients (mean values across inner folds; error bars ±1 SD) for: lower-extremity pain, pain, stiffness, and fatigue. Positive coefficients increase odds of the positive class for each task.

Gait characteristics results

For gait analyses, we used data from the first 14 days after the baseline visit (T0), a period during which clinical characteristics were expected to remain stable. Participants were included if they contributed ≥10 valid walking-in-pocket segments (i.e., walking windows localized to thigh/waist). A total of 29 patients met this criterion and were analysed. Demographic and clinical characteristics at T0 are summarized in **Table 31**; recruitment and consent procedures are described in Section 3.

Table 32 Demographic and clinical characteristics of the PsA patients (n=29) at T0.

Variable	Median [IQR] (or Count %)	Missingness %
Demographics		
Age	50 [41, 61]	0%
Sex (male)	9 (31%)	0%
Education (years)	13 [12, 16]	0%
BMI	28 [23, 34]	0%
BMI ≥ 30	12 (41%)	0%
Disease duration (years)	9 [4, 15]	14%
Data availability		
Available segments per patient	4 [3–7]	0%
Available segments total	2614	0%
Available days per patient	47 [18–115]	0%
Available days total	150 (37%)	0%
Clinical assessment		
Disease phenotype	-	28%
Polyarticular	13 (62%)	-
Oligoarticular	0 (0%)	-
Monoarticular	1 (5%)	-
Axial	5 (24%)	-
Dactylitis	2 (10%)	-
Entheseal	0 (0%)	-
TJC	2 [0, 9]	0%
SJC	0 [0, 1]	0%
TJC Lower body (>0)	13 (45%)	0%
SJC Lower body (>0)	4 (14%)	0%
DAPSA	67 [50, 120]	45%
PASDAS	3.7 [2.9, 4.8]	41%
MDA (achieved)	6 (32%)	35%
Flare (physician) (yes)	6 (21%)	3%
PROs		
Flare (self-reported)	4 (19%)	28%
PSAID12	2.9 [1.6, 5.4]	24%
HAQ	0.9 [0.5, 1.2]	21%
HAQ Walk (>0)	10 (35%)	21%
VAS Pain	37 [20, 63]	31%
Stiffness	5 [2, 6]	0%
Fatigue (PSAID12)	4 [1, 6]	24%

Pain (PSAID12)	3 [2, 7]	24%
**HAQ Walk: "Are you able to walk outdoors on flat ground?" Response scale (0–3): 0 = without any difficulty; 1 = with some difficulty; 2 = with much difficulty; 3 = unable to do.* ***TJC/SJC Lower body: Tender/Swollen Joint Count limited to lower-extremity joints.*		

Gait-feature stability by feature is shown in **Figure 47**. We computed ICC(3, k) on daily aggregates with k = 2–7 consecutive days (rather than 14) because many participants did not provide valid in-pocket walking segments every day across the 14-day window—using 14 days would have excluded a large fraction of the cohort. Across 2–7 days, reliability was generally poor, with a median ICC around 0.30–0.35. Cadence showed the relatively best stability despite a slight decline as more days were added. Step/stride time medians and their MADs followed, while regularity and symmetry were lowest. ICCs plateaued by ~6–7 days, and no gait feature reached the moderate range.

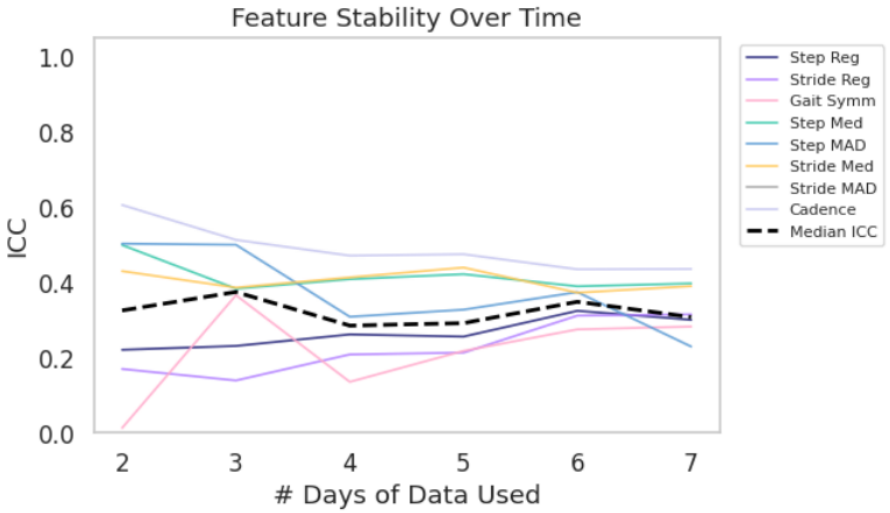


Figure 47 Feature stability for gait characteristics. ICC(3,k) for gait features as a function of days of data (2–7) using thigh/waist "in-pocket" walking segments. Coloured lines denote individual features.

Group comparisons (**Figure 48** and **Figure 49**) revealed limited separation. For walking difficulty (HAQ Walk), no gait feature differed significantly. Participants reporting difficulty walking tended to show lower stride regularity and cadence, and longer step/stride medians and MADs. For lower-extremity tenderness (TJC Lower), step-time MAD was significantly lower in the tender-joint group ($p < 0.05$), while other features showed non-significant trends toward slower timing (higher step/stride medians), higher regularity, and lower cadence. Group-wise summaries are provided in the Appendix (**Table 67** to **Table 68**).

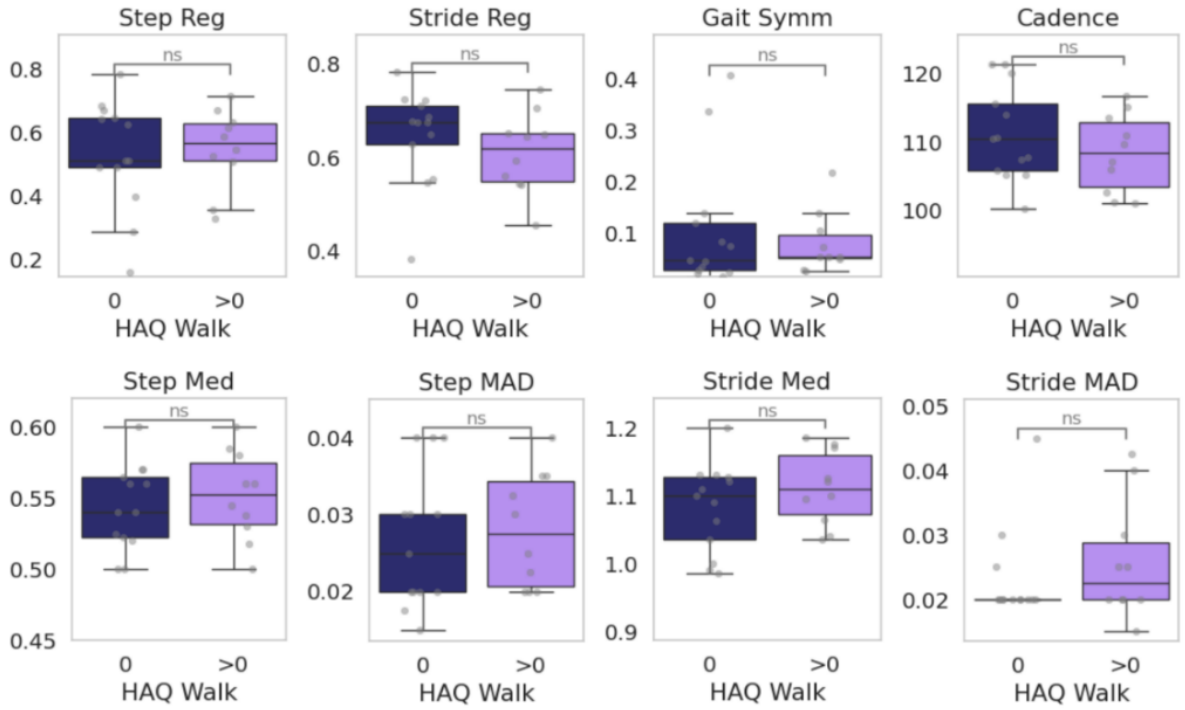


Figure 48 Distributions of gait features for participants with vs. without walking difficulty (HAQ Walk). Significant differences, where present, are indicated (*p < 0.05; Mann–Whitney U).

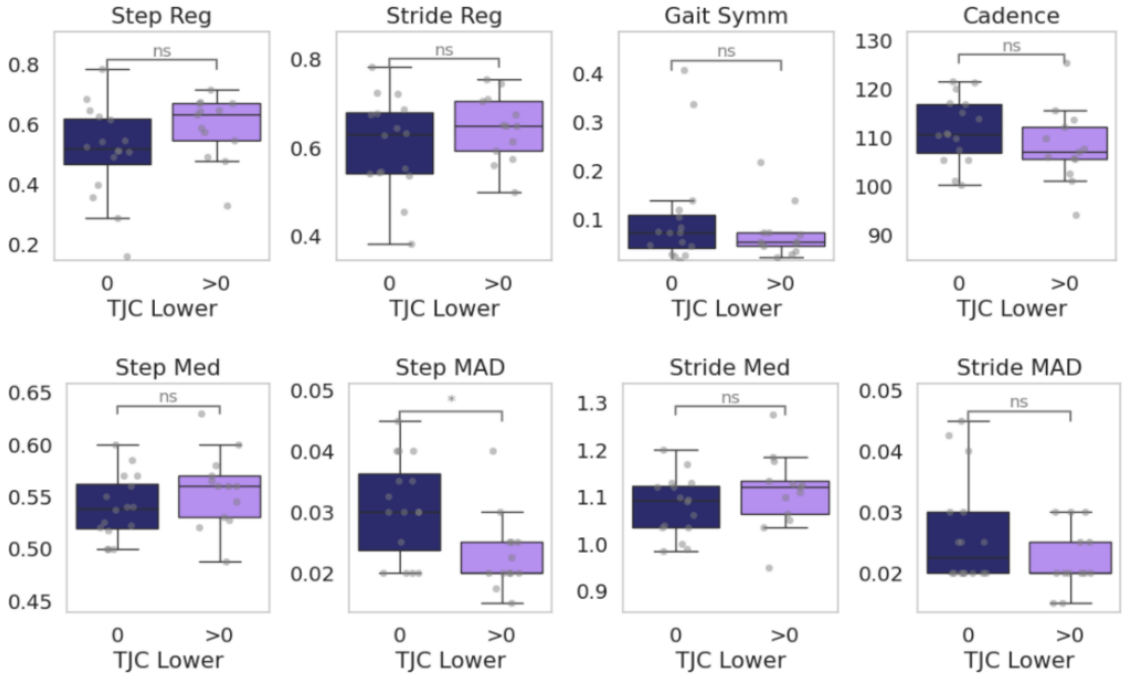


Figure 49 Distributions of gait features for participants with vs. without lower-extremity tenderness (TJC Lower). Significant differences, where present, are indicated (*p < 0.05; Mann–Whitney U).

Spearman correlations (**Figure 50**) likewise yielded few significant associations. The most notable were step-time MAD vs. TJC Lower ($r = -0.53$, *p-value* = 0.044) and stride-time MAD vs. PASDAS ($r = 0.52$, *p*-value = 0.049). Other correlations were small and non-significant; results should be interpreted cautiously given the exploratory analysis. The full pairwise correlation structure is provided in the Appendix (**Figure 87**).

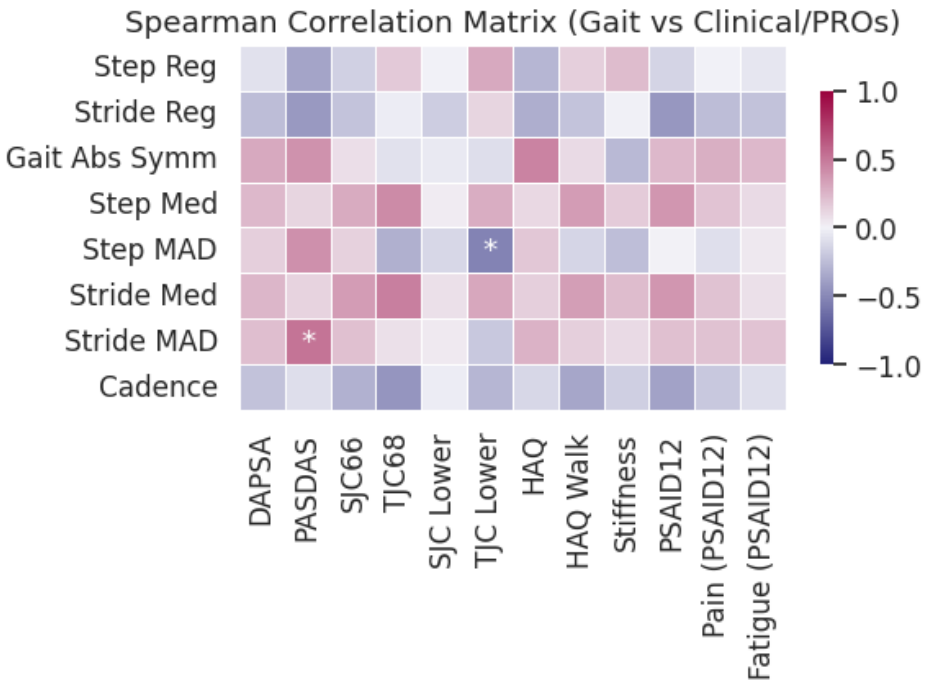


Figure 50 Spearman correlation matrix between smartphone-derived gait features and clinical/PRO outcomes. Cells show Spearman's r (colour scale −1 to 1). Asterisks indicate significance: $p < 0.05$ (*).

8.4 Discussion

8.4.1 Interpretation

Taken together, we observed consistent differences in free-living activity patterns across disease states. Higher inflammation and lower functional capacity were associated with lower daily step volume, less MVPA, and more fragmented activity (shorter sustained bouts) with a higher probability of transitioning to inactivity. Even where individual tests were not significant, the direction of effects aligned across features and analyses. These patterns also aligned with clinical measures (disease activity) and PROs (disability by HAQ; disease impact by PSAID12), supporting that daily movement patterns reflect disease status and burden.

In terms of temporal stability, smartphone-derived features showed modest day-to-day stability that improved with more data, with ICCs plateauing after ~7 days, which is expected because weekly routines repeat, and a full week better captures typical activity patterns. Among these features, light/inactive/walking minutes, mean ENMO, and Light→Inactive transitions achieved moderate reliability.

Group comparisons were directionally consistent. Participants with walking difficulty or lower-extremity pain spent less time in MVPA, had shorter sustained MVPA bouts, lower cadence, and a higher probability of transitioning from light activity to inactivity; patient-reported flare showed a similar pattern, whereas physician-reported flare differences were smaller. These patterns reflect higher inflammatory burden and pain by reducing functional capacity and activity volume and increasing inactivity transitions. Moreover, those achieving MDA accumulated more light activity and more sustained 10-min walking bouts. Correlation analyses mirrored these findings—disease activity, joint counts, disability (HAQ), disease impact (PSAID12), and stiffness were inversely related to sustained walking, MVPA, and cadence, while Light→Inactive transitions were positively related to worse clinical status.

Overall, greater activity volume (e.g., steps), higher-intensity activity, and longer activity bouts were associated with better clinical and functional status.

Models predicting patient-reported flare and MDA performed best, with AUCs around 0.70–0.77; performance was moderate for stiffness and weaker for fatigue, physician-reported flare, and pain, indicating that free-living activity patterns capture functional status and symptom burden. Across tasks, the most informative predictors were the number of sustained 10-minute walking bouts, MVPA minutes and MVPA bout duration, cadence, and step counts—each associated with better status—while a higher probability of Light→Inactive transitions indicated worse status.

For gait, reliability was generally poor over 2–7 days; group separation was limited (no differences for HAQ-Walk; step-time MAD lower with lower-extremity tenderness), and only a few correlations reached significance (e.g., step-time MAD inversely with TJC-Lower; stride-time MAD positively with PASDAS). Cadence, although not significant here as in the all-day analyses, showed consistently negative correlations with disease status and burden; other signals were mixed. These results indicate that passive, free-living (in-pocket) gait capture is feasible but currently less stable and discriminative than all-day activity features, likely requiring more valid walking segments and larger cohorts to improve sensitivity.

8.4.2 Limitations

This analysis is constrained by a small sample—particularly for gait—a single clinical visit (baseline), and a 14-day window, which limits power and generalizability. Medication use or changes after baseline were not modelled and may confound associations. Our in-pocket localization targets segments where the phone is tightly coupled to the body; in everyday use this is often not the case (e.g., loose pockets), so many segments were excluded. Relaxing this constraint would increase yield but likely introduce sensor noise and degrade event detection. Results also depend on phone-carriage behaviour, which varies across individuals, introducing potential measurement bias. Finally, device/sensor heterogeneity and placement variability may reduce comparability across participants.

8.5 Work until D3.6

We will scale to the full PDPID cohort and add multi-visit longitudinal analyses as additional clinical visits become available. We will account for treatment by integrating medication/treatment data as time-varying covariates and conducting sensitivity analyses for therapy changes. To improve gait capture, we will refine on-body localization and evaluate strategies tolerant to lose pocket placement or alternative placements.

9 Physical activity level, sleep, stiffness and gait characteristics from smartwatch's accelerometer sensors

9.1 Background and objectives

Wearable and smartphone accelerometers enable unobtrusive, objective, and reproducible measurement of physical activity, mobility, and gait in free-living settings (Straczkiewicz et al., 2021), offering higher sampling density and less recall bias than questionnaire-based PROs (Knitza et al., 2024). These data can complement clinician assessments and PROs to monitor symptoms and disease severity more efficiently.

Evidence in PsA is still limited. In a hip-worn actigraphy study, Hernández-Hernández et al. (2021) found similar physical activity patterns between PsA and healthy controls but a negative association between physical activity and disease activity over six months. A wrist-worn study (McGagh et al., 2023) monitoring 19 PsA patients for 28 days reported that greater moderate-to-vigorous physical activity (MVPA) was associated with lower disease severity.

The aim of this analysis was to detect physical activity and activity-specific characteristics in PsA patients from smartwatch IMU (tri-axial accelerometer) data. We transformed raw signals into features reflecting daily behaviour (sleep, sedentary, light, MVPA) and walking characteristics (e.g., total steps, cadence). We then evaluated these features as candidate digital biomarkers by testing their associations with clinical measures and patient-reported outcomes that capture changes in physical function.

9.2 Methods

9.2.1 Dataset(s)

To develop smartwatch-based digital biomarkers (dBMs) for PsA, we used data from the first PDPID cohort (see Section 3). Sensor streams were collected passively in free-living conditions. Tri-axial accelerometer data were captured during daily life from a wrist-worn Garmin Vivoactive 5 connected to the miPROLEPSIS app, which recorded tri-axial accelerometer continuously (24/7) at 25 Hz. The variables are summarized in **Table 33**.

Table 33 Smartwatch accelerometer variables captured via miPROLEPSIS.

Variable	Description
Accelerometer	
timestamp	Timestamp in ms
x	Acceleration along the x axis (includes gravity) (mg)
y	Acceleration along the y axis (includes gravity) (mg)
z	Acceleration along the z axis (includes gravity) (mg)

9.2.2 Development

Tri-axial accelerometer data were captured continuously (24/7) from a wrist-worn smartwatch paired with the miPROLEPSIS smartphone app in free-living conditions. Signals were processed in 30-s windows to classify activity levels (sleep, sedentary, light, MVPA) and to detect walking/steps, yielding activity time series.

From this series, we computed daily features describing volume (e.g., steps, minutes per intensity), fragmentation (bout counts/lengths, transition probabilities), intensity (e.g., cadence, ENMO), and sleep. Features were then summarized weekly and related to clinical scores and patient-reported outcomes. A schematic is shown in **Figure 51**.

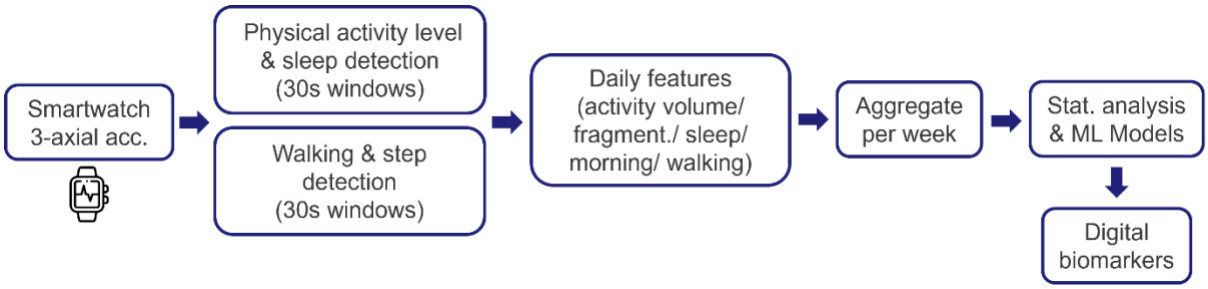


Figure 51 Smartwatch accelerometer data analysis overview.

To passively capture daily physical activity and sleep, wrist-accelerometer data were processed with a deep convolutional neural network (DCNN) HAR pipeline (Yuan et al., 2024). The model was self-supervised pre-trained on UK Biobank data (~100,000 participants; ~700,000 person-days), then fine-tuned for activity recognition using the Capture-24 (Chan et al., 2024) real-world, manually labelled dataset. To incorporate temporal dependence, predictions were post-processed with a Hidden Markov Model (HMM). Preprocessing included data chunking, calibration to local gravity, resampling, and filtering. Signals were segmented into 30-s windows, yielding labels for sedentary, light, MVPA, and sleep. In parallel, using the same backbone and preprocessing, a step-count/walking model fine-tuned on OxWalk dataset (Small et al., 2023) produced 30-s step counts and a walking indicator. ENMO (mg) was computed per 30-s window.

The resulting 30-s time series comprised: activity class, mean ENMO, steps, and walking flag. From this series we derived daily features capturing volume, intensity, fragmentation, sleep, and morning activity, then summarized (mean) weekly. Valid weeks required ≥72 h of data covering all 24 clock hours (threshold from UK Biobank simulations showing 72 h ≈ within 10% of a 7-day measure; Doherty et al., 2017). Missing minutes were imputed using the same time-of-day mean from other observed days; for multi-week summaries we used, in order: same weekday and time, same day-type (weekday/weekend) and time, then the overall time-of-day mean. The full list of daily features is provided in **Table 34**.

Table 34 List of daily features extracted from smartwatch accelerometer data.

Feature	Description
Activity volume	
Mean ENMO (mg)	Average daily Euclidean Norm Minus One (mg)—proxy of movement intensity
Sleep (min)	Total minutes classified as sleep
Sedentary (min)	Total minutes classified as sedentary
Light (min)	Total minutes classified as light activity
MVPA (min)	Total minutes classified as moderate-to-vigorous activity
Steps	Total daily steps
Walking (min)	Daily minutes with walking detected
Activity fragmentation	

Feature	Description
Avg Sleep (min)	Average sleep bout duration
Max Sleep (min)	Longest sleep bout duration
Avg Sedentary (min)	Average sedentary bout duration
Max Sedentary (min)	Longest sedentary bout duration
Avg Light (min)	Average light bout duration
Max Light (min)	Longest light bout duration
Avg MVPA (min)	Average MVPA bout duration
Max MVPA (min)	Longest MVPA bout duration
Pr Sedentary→MVPA	Probability of transitioning from sedentary to MVPA
Pr MVPA→Sedentary	Probability of transitioning from MVPA to sedentary
Pr Sedentary→Light	Probability of transitioning from sedentary to light activity
Pr Light→Sedentary	Probability of transitioning from light to sedentary
Sedentary bouts/hr	Sedentary bout frequency per hour
Morning activity	
Morning ENMO 30min	Mean ENMO in first 30 minutes after wake
Morning ENMO 60min	Mean ENMO in first 60 minutes after wake
Morning ENMO 120min	Mean ENMO in first 120 minutes after wake
Sleep	
Sleep duration (h)	Duration of longest nightly sleep episode
Rest efficiency	Ratio of sleep minutes / time in bed
#Sleep episodes	Number of discrete sleep episodes per night
Rest fragmentation	Frequency of awakenings or interruptions
Sleep movements (#)	Number of movement events during sleep
Sleep movements/hr	Movement events per hour of sleep
Walking	
Walking bouts (≥10min)	Walking bouts ≥10 min (count)
Top 1 cadence (steps/min)	Mean cadence (steps/min) of top 1 min of walking
Top 30 cadence (steps/min)	Mean cadence (steps/min) of top 30 mins of walking
95th cadence (steps/min)	95th percentile of cadence distribution

9.2.3 Verification

We did not perform formal device-level verification for Garmin Vivoactive 5 smartwatch data. Instead, verification relied on preprocessing and quality control: checks for timestamp monotonicity and gaps; calibration to local gravity; screening sampling-rate consistency with exclusion of irregular segments and alignment to 30-s epochs; removal of clipped/saturated or zero-variance windows; plausibility filters for out-of-range ENMO/steps and implausible cadence; alignment of activity labels, ENMO, steps, and walking flags (with HMM smoothing applied to labels); and enforcement of the valid-week rule (≥72 h spanning all 24 clock hours) with time-of-day imputation for short gaps to keep derived day-level features representative

and unbiased. Low-quality segments were excluded to ensure consistent inputs. Additionally, we tested the data pipeline on multiple smartphone vendors during development, as vendor-specific background restrictions can affect transfer from the watch to the miPROLEPSIS app and onward to the cloud.

Figure 52 shows daily coverage before QC: 1237 days in total (median 14 [10–14]), 80% of the expected days. Because participants used their own phones, some models faced transfer issues—especially early on—and in some cases uploads were delayed or blocked by limited or unstable internet access, although initial engagement tended to increase wear time. **Figure 53** shows coverage after QC for the retained cohort (n = 89; 80%): 1087 days (median 14 [12–14]), corresponding to 87% of expected days for those retained for analysis and 70% relative to all participants enrolled.

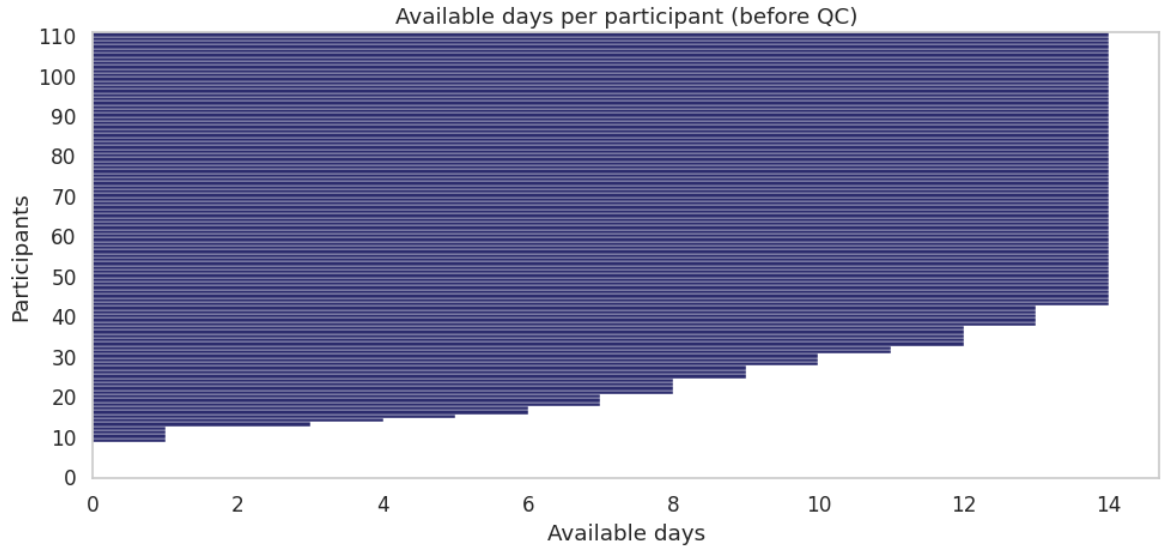


Figure 52 Daily coverage of smartwatch accelerometer data during the first two weeks post-recruitment, prior to quality control.

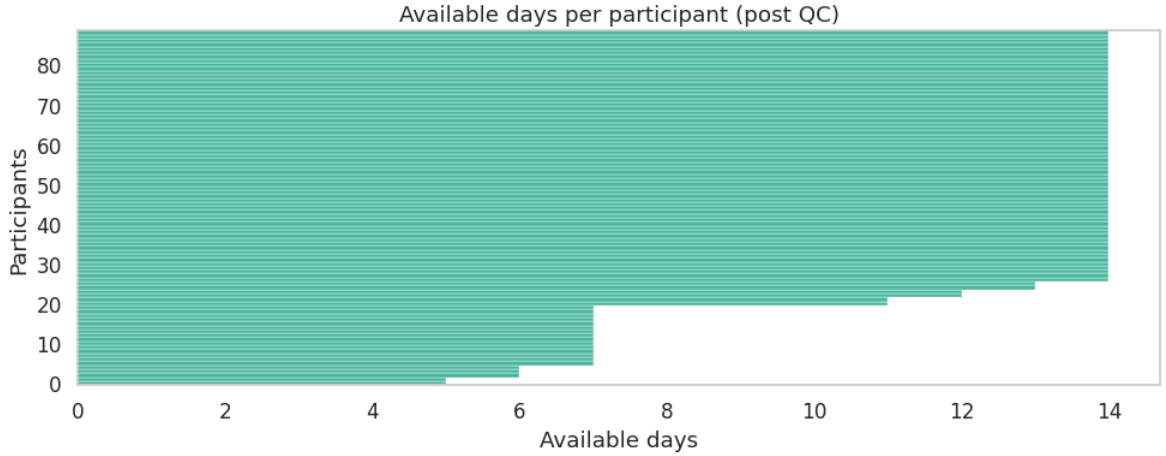


Figure 53 Daily coverage of smartwatch accelerometer recordings in the first two weeks after recruitment, after quality control. Weeks counted as valid if ≥72 h of data covered all 24 h; shown sample reflects those retained for analysis.

Figure 54 illustrates same-time, same day-type imputation presenting the observed activity with missing intervals (grey), and the values imputed from the mean at the same clock minutes

across other weekday/weekend-matched days. For example, a ~2 h gap is filled, preserving 24-h completeness of daily features, whereas leaving such gaps empty would bias day-level summaries.

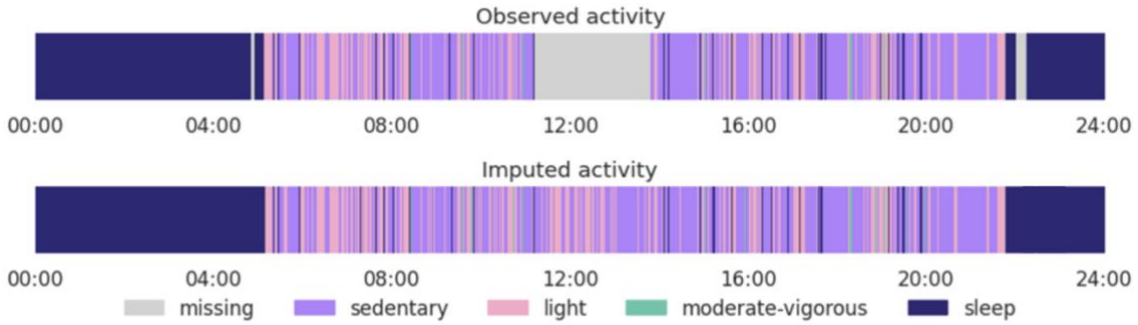


Figure 54 Example of same-time, same day-type imputation. The top panel shows the observed activity with missing intervals (grey); the bottom panel shows the imputed series obtained from the average at the same time of day across other days.

9.2.4 Validation

Statistical analysis

To assess reliability across repeated measures, we computed the intraclass correlation ICC(3, k) for each feature over 14-day period. This two-way mixed-effects model for average measures estimates the consistency of each feature across study days within individuals, where *k* represents the number of repeated daily measurements. ICC values below 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, between 0.75 and 0.9 indicate good reliability, and values above 0.9 indicate excellent reliability (Koo & Li, 2016).

Group comparisons were performed using the non-parametric Mann–Whitney U test. Given the exploratory nature of the study, we set $p < 0.05$ without multiplicity correction and report p-values descriptively. Physical-activity features were compared between: (i) flare vs. no flare (both physician-reported and patient-reported) and (ii) MDA achieved vs. not achieved.

Associations with clinical status were examined using Spearman's correlation coefficients *r* (r > 0.75 good–excellent; 0.50–0.75 moderate–good; 0.25–0.49 fair; <0.25 none). Correlates included clinical targets (TJC, SJC, DAPSA, PASDAS) and patient-reported outcomes (HAQ, PSAID-12 total and items: fatigue, pain, and stiffness).

Modelling

To evaluate discriminative utility, we built classifiers using both linear and tree-based models. Binary tasks included:

- MDA (achieved/not achieved),
- Flare status—physician-reported (flare/no flare),
- Flare status—patient-reported (flare/no flare), and
- PRO impact thresholds: PSAID pain ≤ 4 vs. > 4, PSAID fatigue ≤ 4 vs. > 4 (Gossec et al., 2024), and stiffness ≤ 4 vs. > 4 (all on 0–10 scales).

Model performance was evaluated using area under the ROC curve (AUCROC), F1 score, and Cohen's kappa (κ). In all models, the following variables were included as covariates to account for potential confounding: sex, age, BMI, and disease duration.

For linear models, logistic regression with L1, L2, and ElasticNet regularization was used to address feature collinearity and control model complexity. For non-linear modelling, Random Forest and XGBoost classifiers were used to capture potential interactions between features and non-linear decision boundaries.

A nested cross-validation setup was used to ensure robust and unbiased model evaluation. The outer loop applied a leave-one-subject-out (LOSO) scheme to estimate generalizability, while the inner loop used a 5-fold subject-level cross-validation for hyperparameter tuning.

Hyperparameters for each model were tuned over standard search ranges. For XGBoost, parameters such as learning rate, number of estimators, tree depth, and regularization strength were optimized. Random Forest models were tuned by adjusting the number of trees, tree depth, and number of features considered at each split. Logistic regression models were explored using L1 (LASSO), L2 (RIDGE), and ElasticNet penalties with varying regularization strengths and solvers. All models included class balancing where applicable.

9.3 Results

We analysed data from the first 14 days after the baseline visit, a period during which clinical characteristics were expected to remain stable. Weeks were considered valid if they included ≥72 h of data spanning all 24 clock hours. Missing minutes were imputed using the mean at the same time of day from other observed days in the same period (see Development section for details). A total of 89 participants met these criteria and were included in the analysis. Socio-demographic and clinical characteristics are summarized in **Table 35**; recruitment and consent procedures are described in Section 3. Distributions of indicative feature values for the full cohort are presented in **Table 36**.

Table 35 Demographic and clinical characteristics of the PsA patients (n=89) at T0.

Variable	Median [IQR] (or Count %)	Missingness %
Demographics		
Age	56 [45, 61]	0%
Sex (male)	45 (51%)	0%
Education (years)	14 [12, 16]	1%
BMI	28 [25, 34]	4%
BMI ≥ 30	34 (38%)	4%
Disease duration (years)	9 [4, 18]	9%
Data availability		
Available days per patient	14 [12–14]	0%
Available days total	1087 (87%)	0%
Clinical assessment		
Disease phenotype	-	32%
Polyarticular	33 (54%)	-
Oligoarticular	12 (20%)	-
Monoarticular	4 (7%)	-
Axial	8 (13%)	-
Dactylitis	2 (3%)	-

Variable	Median [IQR] (or Count %)	Missingness %
Entheseal	2 (3%)	-
TJC	1 [0, 7]	0%
SJC	0 [0, 1]	0%
TJC Lower body (>0)	41 (46%)	0%
SJC Lower body (>0)	12 (14%)	0%
TJC Upper body (>0)	45 [51%]	0%
SJC Upper body (>0)	24 [27%]	0%
DAPSA	43 [25, 102]	52%
PASDAS	3.6 [2.7, 4.1]	45%
MDA (achieved)	25 [47%]	40%
Flare (physician) (yes)	16 [19%]	3%
PROs		
Flare (self-reported)	10 [16%]	28%
PSAID12	1.9 [1, 4]	27%
HAQ	0.6 [0, 1]	26%
HAQ Walk (>0)	22 (39%)	26%
VAS Pain	25 [9, 55]	38%
Stiffness	3 [1, 6]	0%
Fatigue (PSAID12)	2 [1, 5]	27%
Pain (PSAID12)	3 [1, 5]	27%
Sleep (PSAID12)	2 [0, 4]	27%

**HAQ Walk: "Are you able to walk outdoors on flat ground?" Response scale (0–3): 0 = without any difficulty; 1 = with some difficulty; 2 = with much difficulty; 3 = unable to do.*

***TJC/SJC Lower body: Tender/Swollen Joint Count limited to lower-extremity joints.*

***TJC/SJC Upper body: Tender/Swollen Joint Count limited to upper-extremity joints.*

Table 36 Distribution of representative features extracted from smartwatch accelerometer data for the full cohort (n=89) during the first 14 days post-baseline; features were computed daily and averaged per subject.

Feature	Median [IQR]
Activity volume	
Mean ENMO (mg)	22 [18, 25]
Sedentary (min)	518 [432, 596]
Light (min)	438 [368, 562]
MVPA (min)	44 [27, 69]
Steps	8132 [6086, 10435]
Walking (min)	351 [278, 405]
Activity fragmentation	

Feature	Median [IQR]
Max Sedentary (min)	25 [18, 32]
Max Light (min)	25 [20, 33]
Max MVPA (min)	4.32 [2.75, 6.57]
Pr Sedentary→MVPA	0.01 [0.01, 0.01]
Pr MVPA→Sedentary	0.13 [0.09, 0.18]
Pr Sedentary→Light	0.09 [0.07, 0.11]
Pr Light→Sedentary	0.11 [0.08, 0.13]
Morning activity	
Morn. ENMO 30min	28 [24, 38]
Morn. ENMO 60min	28 [21, 37]
Morn. ENMO 120min	29 [23, 36]
Sleep	
Sleep duration (h)	8.10 [7.04, 9.94]
Rest efficiency	1.24 [1.04, 1.44]
Rest fragmentation	2.51 [1.49, 5.67]
Sleep movements (#)	19 [9, 36]
Walking	
Walk bouts (≥10min)	9.36 [7.14, 12]
Top 1 cadence (steps/min)	104 [97, 112]
Top 30 cadence (steps/min)	79 [69, 90]

Feature stability results are shown in **Figure 55**. Reliability improved with additional days and plateaued at about one week. By feature group, activity volume and fragmentation showed the highest median stability (moderate); sleep also reached moderate levels; walking and morning features were lowest. At the feature level, rest efficiency was excellent; sleep and light activity were good; and several daily behaviours—sedentary time, walking minutes, ≥10-min walking bouts, transition probabilities, and mean ENMO—were moderate. In contrast, morning-activity metrics, step/cadence measures, and some maximum-type features were poor, indicating greater day-to-day variability.

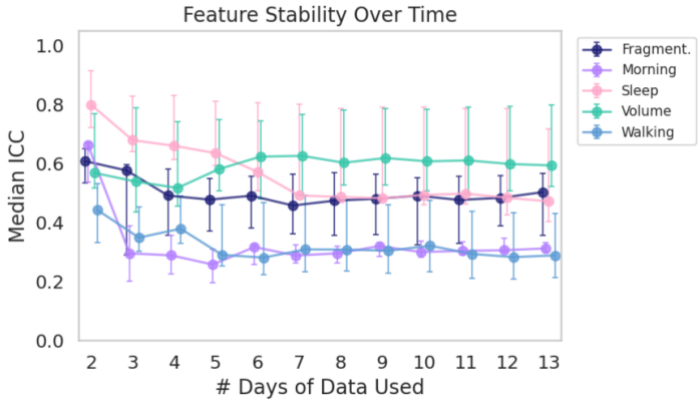


Figure 55 Feature stability over time. Median ICC(3, k) for smartwatch-based physical-activity features as a function of days of data included. Curves represent feature groups (Volume, Fragmentation, Morning activity, Sleep, and Walking); points are the median values across features within each group, and error bars show the standard deviation across features.

Group comparisons are presented in **Figure 56**, **Figure 57**, and **Figure 58**. For physician-reported flare, no features reached significance (Mann–Whitney U), but the flare group tended to have shorter MVPA bouts, fewer sustained ≥10-min walking bouts, less daily walking time and fewer steps, and a higher MVPA→Sedentary transition probability. For patient-reported flare, several differences were significant: MVPA→Sedentary probability was higher in the flare group, while average MVPA bout duration and cadence metrics (Top 1, Top 30, 95th percentile) were lower. For MDA status, participants who did not achieve the target showed higher MVPA→Sedentary and Light→Sedentary transition probabilities, spent more time sedentary, slept less, and had more fragmented sleep. Group-wise summaries are provided in the Appendix (**Table 69** to **Table 71**).

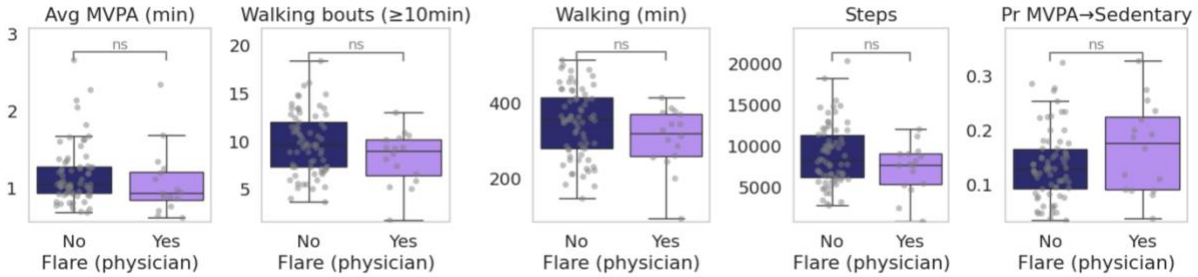


Figure 56 Distributions of selected features for participants with and without physician-reported flare; significant differences, where present, are indicated (*p < 0.05, Mann–Whitney U test).

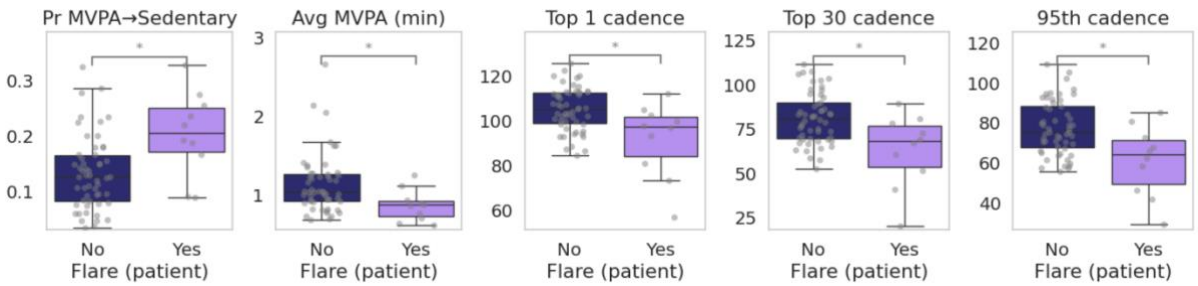


Figure 57 Distributions of selected features for participants with and without patient-reported flare; significant differences are indicated (*p < 0.05, Mann–Whitney U test).

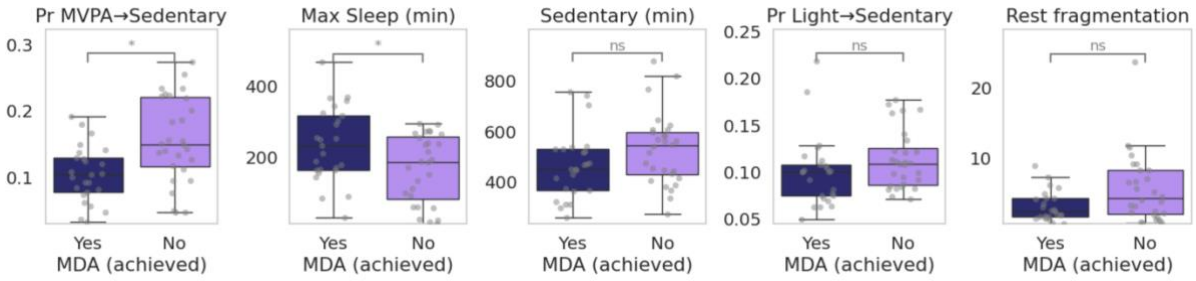


Figure 58 Distributions of selected features for participants achieving vs. not achieving MDA; significant differences, where present, are indicated (*p < 0.05, Mann–Whitney U test).

Spearman correlation analysis results are shown in **Figure 59**, with the top 20 associations listed in **Table 37**; the full pairwise correlation structure is provided in the Appendix (**Figure 88**).

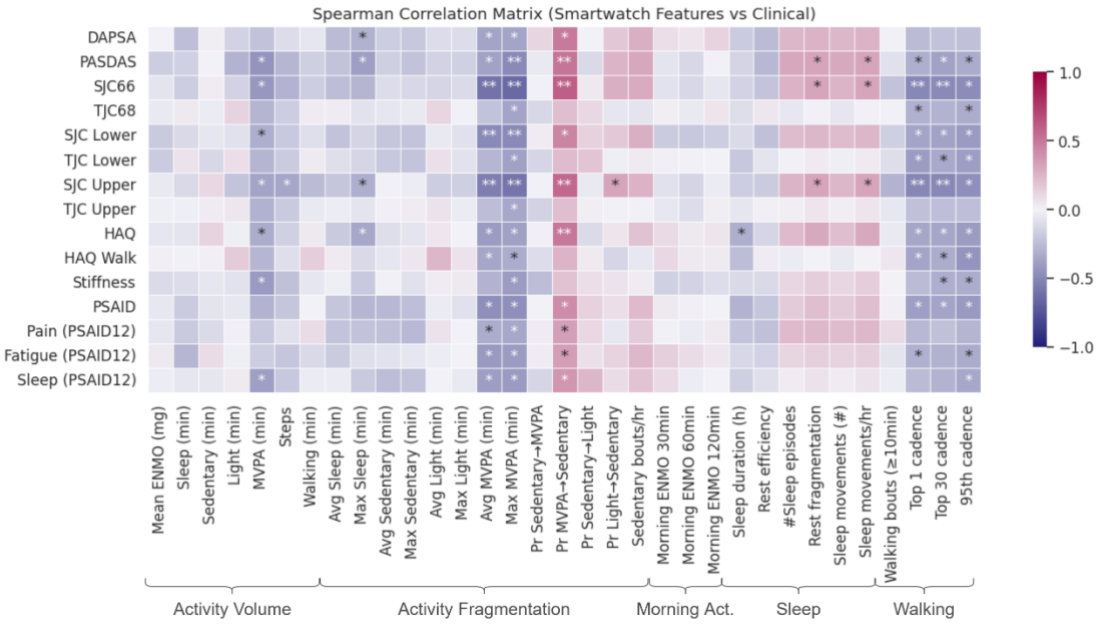


Figure 59 Spearman correlation matrix between smartwatch-derived features and clinical/PRO outcomes. Cells show Spearman's r (colour scale −1 to 1). Asterisks indicate significance: p < 0.05 (*), p < 0.001 (**).

Table 37 Top Spearman correlations between smartwatch-derived features and clinical/PRO outcomes. Pairs with the 20 largest absolute correlations are reported.

Features	Clinical/ PRO outcomes	Spearman's r	p-value
Max MVPA (min)	SJC66	-0.649	< 0.001
Pr MVPA→Sedentary	SJC66	0.609	< 0.001
Avg MVPA (min)	SJC66	-0.608	< 0.001
Max MVPA (min)	SJC Upper	-0.597	< 0.001
Pr MVPA→Sedentary	SJC Upper	0.569	< 0.001
Avg MVPA (min)	SJC Upper	-0.550	< 0.001
Top 1 cadence	SJC Upper	-0.515	< 0.001
Pr MVPA→Sedentary	PASDAS	0.514	< 0.001

Features	Clinical/ PRO outcomes	Spearman's r	p-value
Top 30 cadence	SJC Upper	-0.506	< 0.001
Pr MVPA→Sedentary	HAQ	0.505	< 0.001
Top 1 cadence	SJC66	-0.500	< 0.001
Top 30 cadence	SJC66	-0.500	< 0.001
Max MVPA (min)	SJC Lower	-0.500	< 0.001
Avg MVPA (min)	SJC Lower	-0.499	< 0.001
Max MVPA (min)	PASDAS	-0.497	< 0.001
Pr MVPA→Sedentary	DAPSA	0.490	0.001
95th cadence	SJC66	-0.489	0.001
Max MVPA (min)	PSAID	-0.472	0.002
95th cadence	SJC Upper	-0.471	0.002
Avg MVPA (min)	PSAID	-0.469	0.002

The most consistent pattern was that greater MVPA—both average and maximum minutes—was inversely associated with joint swelling and composite disease activity (SJC66, upper/lower SJC, PASDAS, DAPSA), and with disability and symptom burden (HAQ, PSAID). In contrast, a higher MVPA→Sedentary transition probability was positively associated with these same clinical/PRO measures, indicating a greater tendency to revert to inactivity with worse status. Cadence metrics (Top 1, Top 30, 95th percentile) were also inversely related to swollen joint counts and functional impairment, consistent with reduced walking capacity at higher disease activity.

Sleep features showed complementary trends. Longer maximum nightly sleep bouts were associated with lower disease activity, whereas rest fragmentation and sleep movements were positively related to PASDAS and SJC. Even when not all sleep associations reached significance, features indexing lower sleep volume and poorer rest efficiency tended to relate to worse clinical status and higher disease impact, while fragmentation metrics correlated positively with disease activity.

Overall, higher activity volume and intensity (more steps, more MVPA, longer bouts) correlated with lower disease activity and burden—consistent with better functional capacity—whereas greater activity fragmentation and higher Light/MVPA→Sedentary transition probabilities correlated with worse status. For stiffness, morning activity showed negative correlations, more early movement aligned with less stiffness, while associations with other clinical/PRO outcomes were mixed. Sleep showed a similar pattern, longer maximum sleep bouts related to lower disease activity, whereas fragmentation and sleep movements related to worse status; several correlations did not reach significance but were directionally consistent.

Results for the classification models are summarized in **Table 38**. The best performance was for patient-reported flare with logistic (RIDGE) (AUC 0.82 [0.65–0.94]; κ 0.44; F1 0.55). Physician-reported flare reached ~0.65–0.66 AUC (RIDGE/XG). MDA was moderate (LASSO AUC 0.66; RF AUC 0.64). Pain and fatigue were also moderate with logistic models (~0.64 AUC), while stiffness was weaker. Tree models did not consistently outperform linear ones. Across outcomes, influential predictors (from logistic coefficients; see **Figure 60** and **Figure 61**) included MVPA→Sedentary transition probability, MVPA amount/bout duration, cadence (Top 1/Top 30/95th), walking minutes, sustained ≥10-min bouts, and sleep structure

(episodes, duration, max sleep), alongside covariates (sex, age, BMI, disease duration). We had hypothesized that morning-activity metrics (e.g., morning ENMO 30/60/120 min) would show negative coefficients—indicating better morning activity with better status—but coefficients were mixed across tasks, suggesting heterogeneous behaviour.

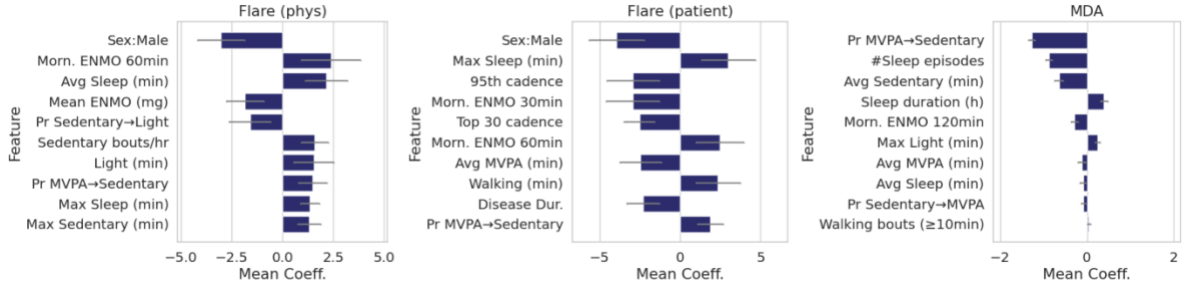


Figure 60 Top coefficients from best logistic models (flare/MDA). Bars show the top 10 coefficients (mean values across inner folds; error bars ±1 SD) for: physician-reported flare, patient-reported flare, and MDA (from left to right). Positive coefficients indicate higher odds of the positive class (flare or MDA achieved).

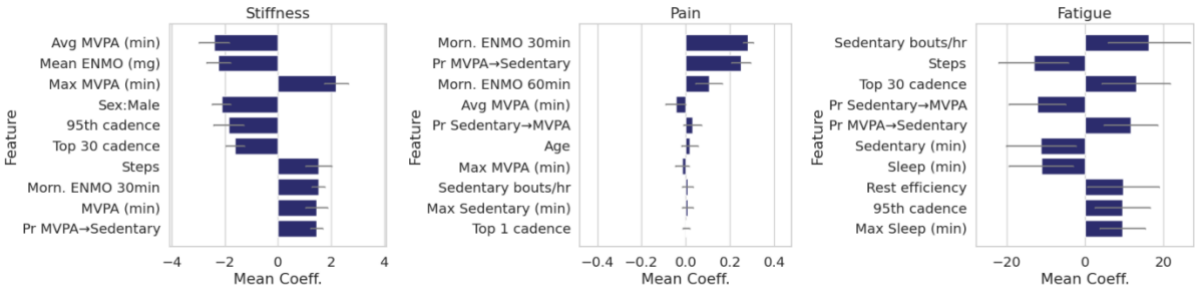


Figure 61 Top coefficients from best logistic models (pain/stiffness/fatigue). Bars show the top 10 coefficients (mean values across inner folds; error bars ±1 SD) for: pain, stiffness, and fatigue. Positive coefficients increase odds of the positive class for each task.

Table 38 Best model performance by outcome. For each outcome, the best-performing classifier is reported with AUC-ROC, Cohen's κ, and F1 score (brackets show 95% CIs).

Outcome	Model	AUCROC	κ	F1
Logistic regression				
Flare (phys.)	RIDGE	0.65 [0.51, 0.78]	0.25 [0.07, 0.45]	0.44 [0.27, 0.59]
MDA	LASSO	**0.66** [0.50, 0.80]	0.38 [0.15, 0.59]	0.69 [0.53, 0.83]
Flare (pat.)	RIDGE	**0.82** [0.65, 0.94]	0.44 [0.18, 0.73]	0.55 [0.31, 0.78]
Pain	ElasticNet	**0.64** [0.48, 0.78]	0.31 [0.08, 0.53]	0.56 [0.39, 0.71]
Stiffness	RIDGE	**0.54** [0.41, 0.66]	0.19 [0.01, 0.39]	0.61 [0.46, 0.73]
Fatigue	LASSO	**0.64** [0.51, 0.77]	0.32 [0.14, 0.50]	0.61 [0.45, 0.75]
Decision trees				
Flare (phys.)	XG	**0.66** [0.51, 0.81]	0.30 [0.07, 0.56]	0.46 [0.28, 0.65]
MDA	RF	0.64 [0.49, 0.80]	0.32 [0.09, 0.58]	0.63 [0.47, 0.78]
Flare (pat.)	RF	0.67 [0.43, 0.89]	0.41 [0.07, 0.69]	0.52 [0.24, 0.75]
Pain	XG	0.54 [0.38, 0.70]	0.20 [-0.02, 0.44]	0.50 [0.32, 0.66]
Stiffness	XG	0.46 [0.34, 0.59]	0.06 [-0.14, 0.28]	0.46 [0.31, 0.61]

Outcome	Model	AUCROC	κ	F1
Fatigue	XG	0.55 [0.41, 0.70]	0.20 [0.00, 0.39]	0.54 [0.37, 0.69]

9.4 Discussion

9.4.1 Interpretation

We observed consistent differences between disease states in PsA and free-living activity/sleep patterns. Higher inflammation and lower functional capacity were associated with lower activity volume and MVPA, shorter active bouts, lower cadence, a greater tendency to transition from MVPA to sedentary, and more fragmented sleep. Despite constraints in sample size and coverage and some tests not reaching significance, the directions were consistent across features and analyses, supporting a coherent pattern.

Smartwatch analyses indicate that passively captured signals carry clinically relevant information, though stability and discrimination vary by feature group. Reliability improved with additional days and plateaued at about one week, suggesting ~5–7 days of wear as a practical minimum. Sleep and activity-volume features were most stable; fragmentation was moderate; walking-intensity and morning-activity metrics were less stable, likely reflecting day-to-day variability.

Group comparisons aligned with clinical status. Patient-reported flare showed the clearest separation: MVPA→Sedentary probability was higher, while MVPA bout duration and upper-range cadence (Top 1/Top 30/95th) were lower. For MDA, not achieving the target was associated with higher MVPA→Sedentary and Light→Sedentary transitions, more sedentary time, shorter sleep, and more fragmented sleep—indicating lower functional capacity and more frequent transitions into inactivity. Physician-reported flare did not yield significant tests, but the same directional patterns were observed, likely reflecting limited power rather than a lack of effect. Overall, worse status was characterized by lower activity volume/intensity, shorter bout continuity, and greater fragmentation.

Correlation patterns were in line with these findings. Greater MVPA (average and maximum minutes) related to lower swollen-joint counts and composite disease activity (SJC66, upper/lower SJC, PASDAS, DAPSA) and to less disability and impact (HAQ, PSAID), whereas higher MVPA→Sedentary probability tracked with higher disease activity and symptom burden. Cadence metrics declined with increasing swelling counts and functional impairment, consistent with reduced walking capacity at higher disease states. Sleep features showed complementary trends: longer maximum sleep bouts related to lower disease activity, while rest fragmentation related to higher disease activity. Not every coefficient was significant, but directions were consistent, strengthening the interpretation.

Models performed best for patient-reported flare and MDA (AUC ≈ 0.66–0.82), were moderate for physician-reported flare, pain, and fatigue, and weaker for stiffness. Predictors mirrored the univariate results: MVPA amount and bout continuity, MVPA→Sedentary transitions, cadence distribution, and aspects of sleep structure. Morning-activity coefficients—expected to be uniformly negative—were mixed across tasks, suggesting heterogeneous behaviour. Overall, smartwatch-derived features—especially MVPA quantity/continuity, cadence, and MVPA→Sedentary transitions—are promising digital correlates of symptom burden and functional status, with sleep fragmentation providing complementary information.

9.4.2 Limitations

This analysis is limited by a single clinical visit (baseline) and a 14-day window, which constrain power and generalizability. Medication use or changes after baseline were not modelled and may confound associations. Results may also reflect wear-pattern and schedule effects (e.g., mornings); although we applied quality control and valid-week rules, residual adherence/sampling variability likely remains. The cohort was modest in size, leading to wide confidence intervals and potential class imbalance—especially for flare models.

9.5 Work until D3.6

We will scale and validate on the expanded PDPID cohort and add longitudinal analyses, including estimation of minimal detectable change, as additional clinical visits become available. Week-level ICCs will be computed to assess stability across weeks (better accounting for schedule variability). We will integrate medication/treatment data as covariates to adjust for treatment effects. Modelling will be extended to regression (for composite scores and continuous outcomes) and broadened classification targets, with improved class-imbalance handling, probability calibration, and subgroup/fairness checks (sex, age, BMI, device). We will also refine feature engineering (e.g., morning metrics, sleep structure).

10 Intestinal motility scores

10.1 Background and objectives

Psoriatic Arthritis (PsA) is a complex inflammatory disease linked with significant comorbidities such as cardiovascular diseases, hypertension, and gastrointestinal (GI) disorders, including inflammatory bowel diseases (IBD). GI involvement is reported in 5–10% of PsA cases, with subclinical manifestations evident in up to two-thirds of patients. These findings suggest an intrinsic association between PsA and GI health, yet the mechanisms remain underexplored. Current clinical approaches for assessing PsA activity primarily depend on patient-reported outcomes (PROs) and episodic clinical evaluations, which may fail to capture the fluctuating nature of GI symptoms. This gap underscores the need for innovative tools capable of providing continuous, objective GI health monitoring.

The study introduces an abdominal wearable device, the SmartBelt, which utilizes electrogastrography (EGG), accelerometry, and acoustic sensors to measure GI myoelectrical activity, motility, and sounds. This device seeks to establish GI motility patterns as potential biomarkers to differentiate between PsA patients and healthy controls. Insights from this research aim to enhance understanding of PsA's impact on GI health and improve management strategies through novel, non-invasive monitoring methods.

Objectives

Primary Objective:

- To perform a comparative analysis between the psoriatic arthritis patient group and the control group to identify potential differences between the two populations.

Secondary Objectives:

- To assess patient acceptance, satisfaction, and compliance with the use of the SmartBelt device for monitoring gastrointestinal health in the context of PsA management.
- To study the link between PsA and GI health.
- To validate the features extracted from GI motility data and assess their use in the development of novel and interpretable machine learning algorithms that can utilise these features to effectively distinguish between different GI motility patterns associated with PsA and healthy states.
- To explore the potential of the SmartBelt device in reducing the need for frequent outpatient clinic visits by providing reliable remote monitoring of GI health in patients with PsA during periods of low disease activity.

This study represents a critical step toward leveraging wearable technologies for precise, non-invasive monitoring of PsA-associated GI dysfunctions, paving the way for future longitudinal research and therapeutic innovations.

10.2 Methods

10.2.1 Dataset(s)

No datasets were identified.

10.2.2 Development

The study is a single-centre, cross-sectional, observational design conducted in Portugal. Participants include individuals diagnosed with Psoriatic Arthritis (PsA) and healthy controls. Recruitment is carried out through patient advocacy groups and online platforms, ensuring eligibility based on predefined inclusion and exclusion criteria.

Study Procedure

1. **Baseline Recording:** Participants undergo a 30-minute recording in a fasting state using the SmartBelt. This is conducted in a quiet, comfortable environment to minimize external disturbances and ensure high-quality data.
2. **Test Meal:** Participants consume a standardized meal tailored to dietary needs (e.g., no-dairy or diabetic-friendly options) to stimulate postprandial GI responses. Meals are designed to provide 250–500 kcal, with consumption completed within 10 minutes.
3. **Postprandial Recording:** Following the meal, a 30-minute recording is performed while participants remain in a resting position to capture post-meal GI activity.
4. **SmartBelt Data Collection:** Participants wear the SmartBelt for a total of 8 hours. The device records electrogastrography (EGG) signals, acoustic data (bowel sounds), and movement (via an accelerometer). Data collection focuses on pre- and post-meal phases, ensuring comprehensive monitoring.
5. **Patient-Reported Data:** Participants complete questionnaires to capture demographic details, GI symptoms, and PsA disease activity. The Inflammatory Bowel Disease Questionnaire is used to assess GI health.

Data Processing and Analysis

The SmartBelt data, including EGG signals, bowel sounds, and movement artifacts, will undergo a multi-stage preprocessing and analysis pipeline to extract meaningful features and develop predictive models. The overall goal is to differentiate between PsA patients and healthy controls based on gastrointestinal motility patterns and to correlate these patterns with clinical and patient-reported data.

The first step of preprocessing is signal quality assessment, where both raw EGG and ACU signals are initially assessed for quality. Segments with excessive noise or artifacts (either due to movement, electrode detachment, or external interference, for example) are identified and either corrected, using techniques like independent component analysis or interpolation, or outright excluded from further analysis, depending on the length and signal-to-noise ratio of the affected signal.

In terms of the filtering applied, for the EGG a band-pass filter is applied to the signals to isolate the frequency range of gastric slow waves (typically 0.01-0.15Hz, which corresponds to 0.6-9 cycles per minute). This is intended to remove high-frequency noise and other biosignals, as well as low-frequency baseline drift. Similarly, for the ACU a band-pass filter will also be applied focused on the frequency range of characteristic bowel sounds (typically 100-1000 Hz), with the main goal of separating bowel sounds from other physiological sounds and environmental noise. The ACU signal will also be segmented into bowel sound windows, with a determined start and finish that contain a detected bowel sound.

At this stage, from the preprocessed EGG and ACU signals, a comprehensive set of features is extracted, including both time-domain and frequency-domain characteristics. These features will be extracted using internal PLUX libraries and are based on established research to capture different aspects of gastric and intestinal motility and are detailed on **Table 39**.

Table 39 List of features extracted from EGG and ACU data.

Feature	Description
Electrogastrography Features	
Dominant Point	The point of the signal that has the highest power or amplitude. (Hz, mV^2/Hz)
Dominant Frequency	The frequency component that has the highest amplitude in the signal spectrum, indicating the predominant rhythm of the gastric activity (Hz)
Dominant Power	The power or amplitude of the dominant point in the signal spectrum. (mV^2/Hz)
Frequency Instability Coefficient	Measures the instability of its dominant frequency component.
Power Instability Coefficient	Measures the instability of its dominant power component.
Percentage of Normal Slow-Waves	Percentage of time that the dominant frequency falls within the normal frequency range. (%)
Power in Bradygastria Range	Power in the bradygastria frequency range.(%)
Power in Normal Range	Power in the normal frequency range.(%)
Power in Tachygastria Range	Power in the tachygastria frequency range. (%)
Total Power in EGG Informative Range	The sum of the power of the electrogastrography signal spectrum in the [0, 0.167] Hz range. (mV^2/Hz)
Crest Factor	Ratio of the signal's peak value to the root mean square value of the EEG signal. (s, a.u.)
Acoustic Features	
Number of Valid Bowel Sounds Windows	Number of valid bowel sounds windows, after pre-processing.
Bowel Sounds per Minute	Number of bowel sounds per minute.
Bowel Sounds windows (start and end indexes)	Start and end indexes of the bowel sound windows.
Bowel Sounds windows (start and end times)	Start and end times of the bowel sound windows.
Peak Frequency	Percentage of daily time spent in walking.
Sound-to-Sound Intervals (SSI)	Time intervals between sounds.
Average SSI	Average of the time intervals between sounds.
Standard Deviation of SSI	Standard deviation of the time intervals between sounds.
Sound Duration (SD)	Duration of each valid sound detected.
Average SD	Average duration of each valid sound detected.
Standard Deviation of SD	Standard Deviation of duration of each valid sound detected.
Sound Index (SI)	Sum of absolute signal amplitudes.
Average SI	Average of the sum of absolute signal amplitudes.
Standard Deviation of SI	Standard deviation of the sum of absolute signal amplitudes.

In addition to these two signals, accelerometer data will also be used to identify and help exclude periods of significant movement that may contaminate the EGG and ACU signals. It acts primarily as a covariate to improve the quality of the GI-specific data. In an initial stage, simple statistics such as mean absolute deviation can be used for this end.

Machine Learning Modelling

The extracted features serve as input to machine learning models. The specific ML algorithms will be selected and optimized based on the characteristics of the dataset (size, feature distribution, among others) and performance during preliminary testing. Candidate models include, but are not necessarily limited to:

- Deep neural networks, particularly convolutional neural networks and recurrent neural networks, which are well-suited for analysing time-series data lie EGG and ACU signals
- Support vector machines, which are effective for classification tasks and are adequate for handling high-dimensional feature spaces
- Random forests, which are ensemble methods that tend to provide robustness and good generalization performance.

Explainable AI techniques will be applied through methods such as SHAP (SHapley Additive exPlanations) to assist in determining the relative importance of each extracted feature in the model's predictions. This is also expected to help identify the biomarkers associated with PsA-related GI dysmotility.

Regarding the training and validation of the models, the dataset will be split into training, validation, and testing sets. The training set is used to train the ML models, the validation set is used to tune hyperparameters and prevent overfitting, and the testing set is used for final, unbiased performance evaluation. Additionally, k-fold cross-validation (e.g., 5-fold or 10-fold) will be employed to ensure robust model evaluation and to mitigate the risk of overfitting to a specific data split. In k-fold cross-validation, the data is divided into k subsets, and the model is trained and evaluated k times, each time using a different subset for testing and the remaining subsets for training.

Finally, model performance will be evaluated using several metrics ranging from accuracy, sensitivity, specificity, F1-score, area under the receiver operating characteristic curve, and Cohen's Kappa in the cases that relevant clinician feedback is obtained.

Main study parameters/endpoints

- GI motility patterns measured via SmartBelt
- Patient-reported GI symptoms measured by a validated GI symptom questionnaire (Inflammatory Bowel Disease Questionnaire)
- Correlation between GI motility patterns and PsA disease:
    - Correlation coefficients analysis (Pearson/Spearman) (including the evaluation of the correlation between GI myoelectrical activity/motility patterns and patient-reported symptoms, such as abdominal pain and bloating)
    - Regression analysis results
- Machine learning algorithm performance:
    - Accuracy, sensitivity and specificity in distinguishing PsA patients from healthy controls
- Feasibility, usability, and patient satisfaction with the SmartBelt

In the Appendix II it is possible to encounter more information on the investigational device.

10.3 Work until D3.6

In the forthcoming period until D3.6, the work will focus on completing patient recruitment and data collection with the SmartBelt device, ensuring high-quality recordings across fasting, postprandial, and daily activity phases. The acquired EGG, acoustic and accelerometer data will be pre-processed, and feature extraction will be finalised. The next steps will include the development and optimisation of the machine learning models for identifying PsA-related gastrointestinal motility patterns, as well as the validation of these models through cross-validation strategies. Patient-reported outcomes and clinical questionnaires will be integrated with the objective sensor data to explore correlations between GI symptoms and PsA disease activity. Finally, emphasis will be placed on evaluating patient acceptance, usability and satisfaction with the SmartBelt, with results expected to inform the preparation of subsequent longitudinal studies and future therapeutic applications.

11 Digital inflammation response index

11.1 Background and objectives

Heart rate variability (HRV) quantifies changes in the time intervals between consecutive heartbeats, known as beat-to-beat intervals (BBI). Because heart rhythm is regulated by the autonomic nervous system (ANS), HRV reflects the activity of both the sympathetic (SNS) and parasympathetic (PNS) branches, as well as their interaction. HRV has been linked to various physiological processes, including respiratory sinus arrhythmia, the baroreceptor reflex, psychological stress, and thermoregulation.

HRV comprises a family of metrics categorized by the method of computation: time-domain, frequency-domain, and non-linear measures. Some metrics capture PNS activity only (vagally mediated HRV), while others reflect both PNS and SNS activity. The time window used to calculate HRV is critical, as different physiological processes occur on different timescales. Most studies use short-term HRV data collected over less than a day in clinical settings, and findings from such analyses may not generalize to long-term, free-living conditions (Hayano & Yuda, 2021). One common application of HRV is the detection of psychological stress. Under psychological stress, HRV typically decreases, reflecting reduced parasympathetic (vagal) nervous system activity and increased sympathetic nervous system activity. This shift aligns with the "fight or flight" stress response, where the sympathetic nervous system becomes more dominant (Immanuel et al., 2023).

In the context of the immune system, the inflammatory reflex (**Figure 62**) that is part of the PNS can suppress immune cell proliferation via the cholinergic anti-inflammatory pathway (Tracey, 2002). Lower inflammation levels are often associated with higher HRV, particularly vagally mediated HRV, although the literature reports mixed findings (Williams et al., 2019; Adam et al., 2023). A recent study most relevant to our work linked HRV to inflammatory bowel disease flares, using smartwatch-derived BBI data collected over several days in free-living conditions. HRV was lower during flares compared to remission phases (Hirten et al., 2025).

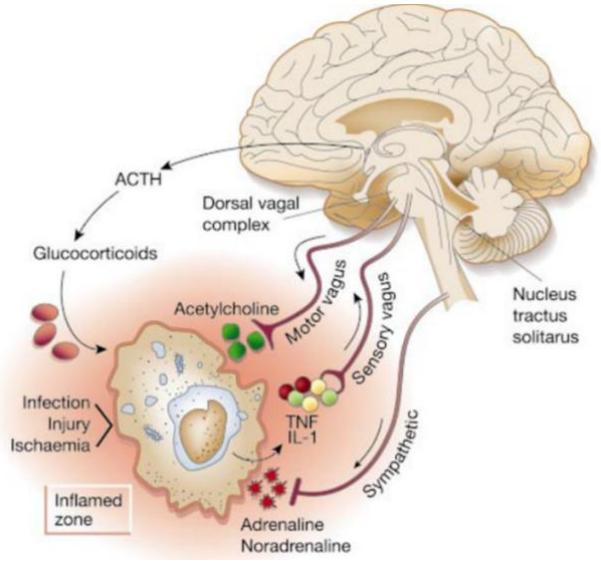


Figure 62 Inflammatory products from damaged tissues trigger afferent signals to the nucleus tractus solitarius. This activates vagus efferents, which inhibit cytokine synthesis via the cholinergic anti-inflammatory pathway. Signals can also reach the hypothalamus and dorsal vagal complex, stimulating ACTH release and activating the humoral anti-inflammatory pathway. Sympathetic activation raises local adrenaline and noradrenaline levels, further suppressing inflammation. (Reproduced from Tracey, 2002).

The objective of this subtask is to investigate the relationship between inflammation in PsA and HRV. Inflammation is quantified using the clinical variables from the PDPID study. HRV will be assessed using a processing pipeline that computes common measures from all HRV categories, incorporating the necessary pre-processing steps and post-processing analyses to explore dynamic changes in HRV over extended periods.

11.2 Methods

11.2.1 Dataset(s)

The dataset contains both clinical and digital data. Clinical data include healthcare professionals' assessments, patient-reported outcome measure (PROM) questionnaires, and laboratory biomarkers collected during the recruitment visit. The clinical variables of interest in this subtask are: flare status as reported by doctors (DOC FLARE) and patients (PAT FLARE), PSAID, DAPSA, PASDAS scores, and CRP levels. Digital data consist of BBI recordings from smartwatches collected over several weeks. For this study, only the first two weeks of recordings after recruitment are used, as clinical status is generally stable during this period.

The total number of patients who are included in this study is 105. It is important to note that: (i) not all patients have complete two-week smartwatch recordings (the verification section below – section 11.2.3 – explains how data coverage is calculated, and which patients are excluded, and (ii) not all clinical variables are available for every patient. **Table 40** shows the distribution of available variables in the HRV study cohort. In addition, CRP values were not always measured on the recruitment date. Only CRP measurements taken within two weeks before or after the baseline visit were retained; others were excluded.

Table 40 Target clinical variables available for analysis and their distribution in HRV study cohort.

	DOC FLARE	PAT FLARE	DAPSA	PASDAS	PSAID	CRP	ALL (excluding CRP)
Number of patients	102	77	54	60	77	34	43

11.2.2 Development

Each BBI sequence is processed to derive HRV metrics. The processing pipeline consists of the following steps:

1. ***Artifact removal***: Artifact intervals are removed using well-established criteria, i.e., the Malik rule (Malik, 1996), the Acar rule (Acar et al., 2000), and exclusion of intervals outside the physiological range of 250–2000 ms.
2. ***Segmentation***: The cleaned BBI sequence is divided into overlapping segments (50% overlap) using a rolling window. The window length is a hyperparameter.
3. ***Missing data filtering***: BBI data from smartwatches in free-living conditions often contain gaps (Bent et al., 2020). Large gaps reduce the reliability of HRV metrics. Following the method of Föll et al. (2021), each segment is accepted if it satisfies the missing data criterion

$$m = N_t \cdot \frac{\overline{BBI}_t}{L} > \ thresh$$

where L is the window length, $\overline{BBI}_t$ is the segment's mean BBI, N_t is the number of BBIs in the segment, and $thresh$ is a user-defined parameter in [0,1]. In this study, we set $thresh = 0.5$ to balance between rejecting incomplete and retaining useful data.

4. ***HRV metric calculation***: For each accepted segment, standard HRV metrics of all categories are computed using the NeuroKit2 Python library (Makowski et al., 2021). The selected metrics are listed in **Table 41**.
5. ***Aggregation*:** The HRV metrics are aggregated across segments (***mean*** and ***standard deviation***) to produce representative values, called **HRV markers**, for the entire recording period (i.e., 2 weeks). Thus, each two-week period is represented by a single HRV marker.

Table 41 The HRV metrics implemented in the study.

	Metric	Unit	Description
Time domain	SDNN	ms	Standard deviation of normal-to-normal (NN) intervals.
	RMSSD	ms	Root mean square of successive RR interval differences.
	HRV triangular index	-	Integral of the density of the RR interval histogram divided by its height.
Frequency domain	ULF power	ms^2	Absolute power of the ultra-low-frequency band ***(≤0.003 Hz)***.
	VLF power	ms^2	Absolute power of the very-low-frequency band ***(0.0033–0.04 Hz)***.
	LF power	ms^2	Absolute power of the low-frequency band ***(0.04–0.15 Hz)***.
	HF power	ms^2	Absolute power of the high-frequency band ***(0.15–0.4 Hz)***.
	LF/HF	%	Ratio of LF-to-HF power.
	Total power	ms^2	Absolute power of the total band
Non-linear	DFA α1	-	Detrended fluctuation analysis, which describes short-term fluctuations.
	DFA α2	-	Detrended fluctuation analysis, which describes long-term fluctuations.

The HRV processing pipeline and an illustrative step-by-step example are shown in **Figure 63**. The resulted HRV markers are calculated for a specific selection of window length. In this study, the HRV pipeline will be executed for three different window lengths, i.e., 1h, 6h, and 12h; hence, the total number of markers is 11 (HRV metrics) x 2 (aggregation statistics) x 3 (window lengths) = 66.

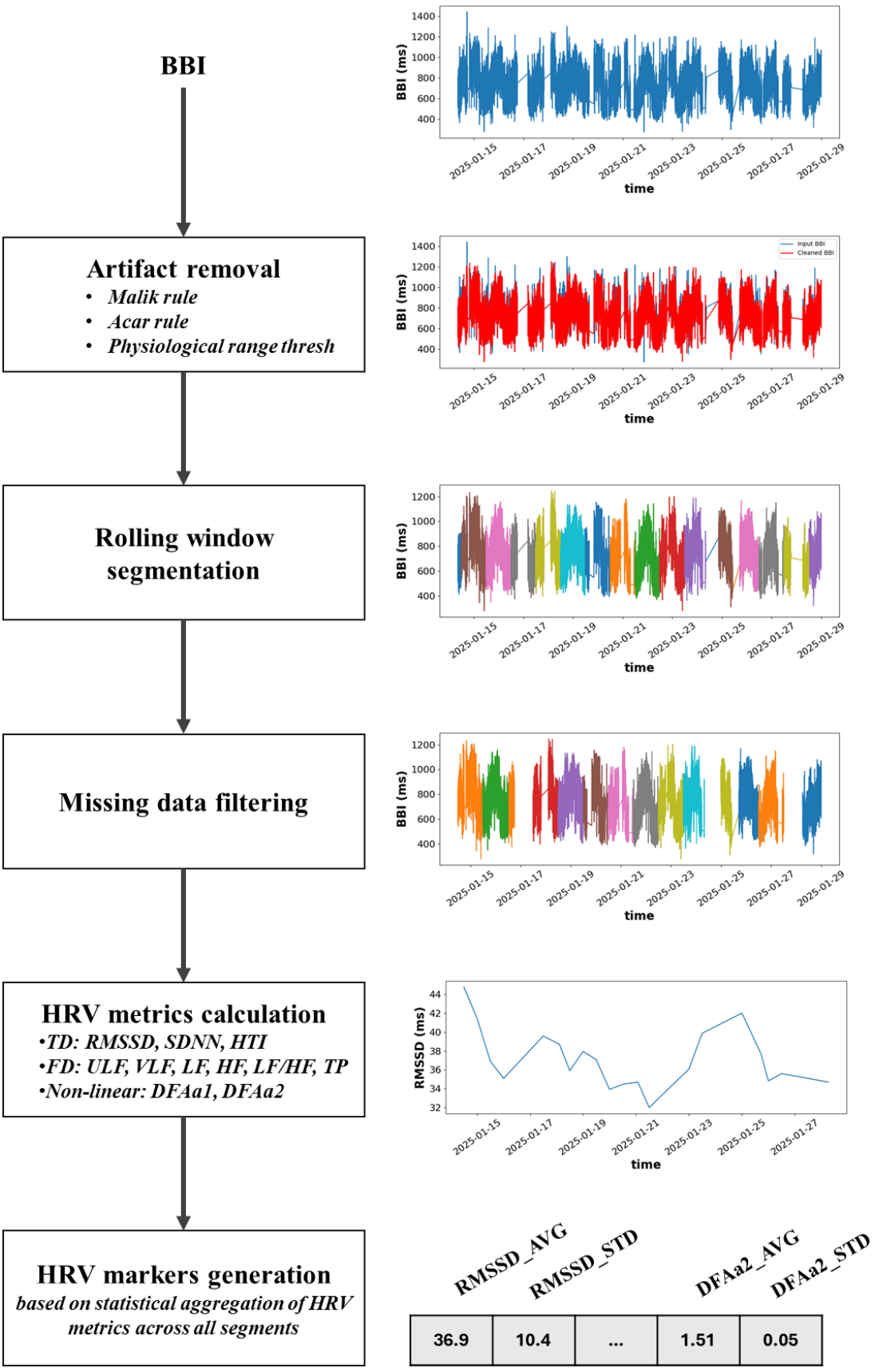


Figure 63 HRV processing pipeline with illustrative step-by-step example.

11.2.3 Verification

Missing or incomplete BBI recordings are common when collected from smartwatches in free-running conditions. Therefore, data completeness must be assessed before applying HRV calculations, as incomplete data can compromise results. This assessment is performed at two levels. In window level, each segment is checked to ensure it contains enough BBI samples. This process was described in the previous section as missing data filtering. An example that illustrates the nature of missing data criterion and supports the decision of choosing the threshold as 0.5 is shown in **Figure 64**.

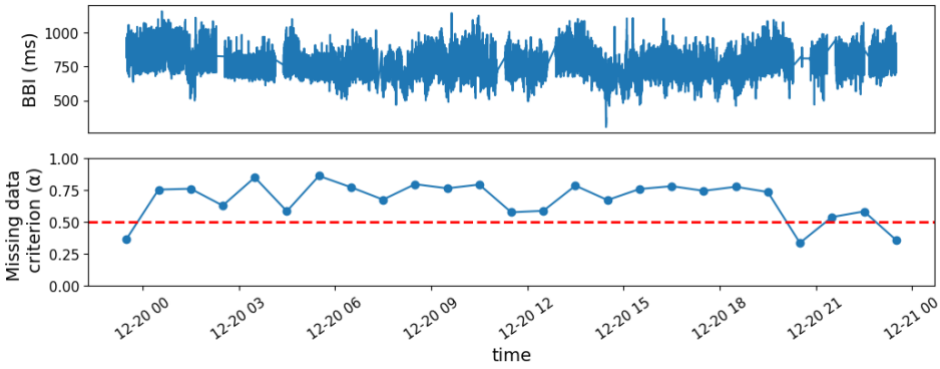


Figure 64 Missing data filtering example. A single-day BBI recording is segmented into non-overlapping hourly windows. The proportion of missing values in each window is evaluated using the missing data criterion.

After removing unacceptable segments, each patient's two-week recording is evaluated for overall data coverage, defined as the ratio of total time with acceptable BBIs to the full two-week period. If coverage is less than seven days, the entire recording is excluded. Low coverage typically occurs when the patient does not wear the smartwatch for several consecutive days. The total number of patients with coverage greater than seven days is 79. **Figure 65** illustrates the coverage during the first two weeks for the 105 patients.

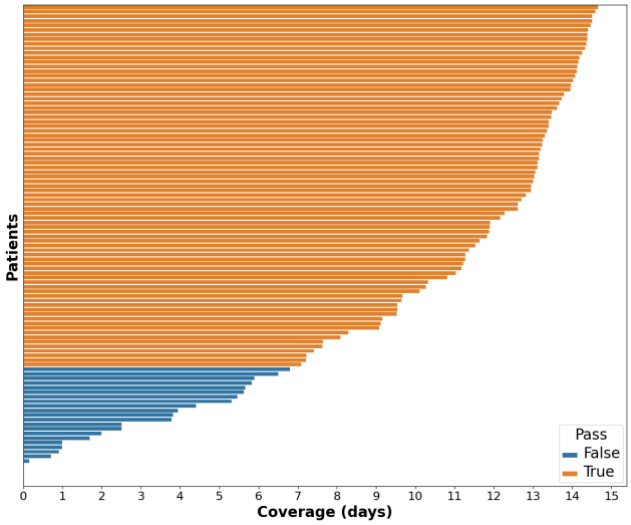


Figure 65 Coverage of BBI data in the first two weeks after recruitment.

11.2.4 Validation

The validation tests the association between calculated HRV markers and inflammation status. Inflammation status is represented by either binary or numerical target variables.

- ***Binary targets***: flare status reported by doctors (DOC FLARE) and patients (PAT FLARE), and inflammatory flare defined as CRP > 3 mg/dL.

- ***Numerical targets***: DAPSA, PASDAS, PSAID scores, and CRP level.

For binary targets, differences between HRV marker groups are assessed using the Mann–Whitney U test. This choice followed Shapiro–Wilk normality tests, which showed that most groups did not meet normality assumptions. For numerical targets, correlations between HRV markers and target variables are assessed using Spearman's rank correlation. Again, normality tests indicated non-normal distributions, justifying the non-parametric approach.

11.3 Results

We analysed data from the first 14 days after the baseline visit, a period during which clinical characteristics were expected to remain stable. Socio-demographic and clinical characteristics of the 79 participants who meet the BBI data coverage criterion described in verification section are summarized in **Table 42**.

Table 42 Demographic and clinical characteristics of the PsA patients (n=79) at T0.

Variable	Median [IQR] (or Count %)	Missingness %
Demographics		
Age	55 [45, 61]	0%
Sex (male)	41 (52%)	0%
BMI	28 [25, 33]	4%
BMI ≥ 30	28 (35%)	4%
Disease duration (years)	8 [4, 19]	10%
Clinical assessment		
Disease phenotype	-	32%
Polyarticular	27 (50%)	-
Oligoarticular	12 (22%)	-
Monoarticular	4 (7%)	-
Axial	8 (15%)	-
Dactylitis	2 (4%)	-
Entheseal	1 (2%)	-
CRP (mg/dL)	0.4 [0.1, 1]	27%
CRP (mg/dL) (14 days)	0.4 [0.2, 1]	57%
DAPSA	48 [24, 101]	53%
PASDAS	3.5 [2.7, 4.2]	47%
MDA (achieved)	21 (45%)	41%
Flare (physician) (yes)	13 (17%)	20%
Flare (inflammatory) (yes)*	3 (9%)	57%
PROs		
Flare (self-reported)	8 (14%)	29%
PSAID12	2.2 [1.1, 4.2]	28%
** Inflammatory flare defined as CRP > 3 mg/dL.*		

11.3.1 Statistical comparison tests

When using doctor-reported flare status (DOC FLARE) as the target variable, the positive group (flare) included 13 patients, while the negative group (no flare) included 63. The distribution of HRV markers as median [IQR], along with Mann–Whitney U test results, are presented in **Tables Table 43**, **Table 44**, and **Table 45** for markers aggregated over 1-hour, 6-hour, and 12-hour windows, respectively. Significant differences are observed for ***time-domain HRV metrics (RMSSD, SDNN, and triangular index)*** when average aggregation is applied. Among the window lengths, the 1-hour window yielded the strongest effects. The boxplots of these HRV markers are shown in **Figure 66**.

Table 43 HRV markers aggregated from 1-hour windows, with their median [IQR] distributions in the study cohort grouped by doctor-reported flare status (Yes/No), along with associated p-values indicating statistical differences. Markers with p-values less than 0.05 are indicated by purple background shading.

HRV marker	DOCTOR FLARE Yes (N=13) Median [IQR]	DOCTOR FLARE No (N=63) Median [IQR]	p value
Time domain markers			
hrv_rmssd_w1h_avg	33.694 [29.579, 36.711]	40.013 [35.185, 47.691]	**0.002**
hrv_rmssd_w1h_std	8.148 [7.161, 10.494]	9.427 [8.105, 11.911]	**0.044**
hrv_sdnn_w1h_avg	64.410 [57.289, 81.768]	79.685 [67.296, 89.072]	**0.004**
hrv_sdnn_w1h_std	20.853 [18.639, 23.792]	23.164 [20.246, 30.822]	**0.036**
hrv_hti_w1h_avg	15.955 [13.715, 18.238]	18.236 [15.867, 21.215]	**0.009**
hrv_hti_w1h_std	4.896 [4.293, 5.498]	5.415 [4.760, 6.373]	**0.048**
Frequency domain markers			
hrv_ulf_w1h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.350
hrv_ulf_w1h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.144
hrv_vlf_w1h_avg	0.002 [0.001, 0.002]	0.002 [0.001, 0.002]	0.267
hrv_vlf_w1h_std	0.002 [0.002, 0.002]	0.002 [0.002, 0.002]	0.434
hrv_lf_w1h_avg	0.001 [0.001, 0.001]	0.001 [0.001, 0.001]	0.272
hrv_lf_w1h_std	0.001 [0.001, 0.002]	0.002 [0.001, 0.002]	0.200
hrv_hf_w1h_avg	0.001 [0.001, 0.001]	0.001 [0.001, 0.001]	0.200
hrv_hf_w1h_std	0.001 [0.001, 0.001]	0.001 [0.001, 0.002]	0.090
hrv_lfhf_w1h_avg	1.560 [1.364, 1.722]	1.640 [1.320, 1.873]	0.407
hrv_lfhf_w1h_std	0.621 [0.484, 0.795]	0.602 [0.413, 1.050]	0.742
hrv_tp_w1h_avg	0.004 [0.003, 0.004]	0.004 [0.003, 0.005]	0.200
hrv_tp_w1h_std	0.004 [0.004, 0.005]	0.005 [0.004, 0.006]	0.123
Non-linear markers			
hrv_dfa1_w1h_avg	1.055 [1.029, 1.124]	1.097 [1.045, 1.127]	0.305
hrv_dfa1_w1h_std	0.115 [0.104, 0.148]	0.110 [0.091, 0.139]	0.900
hrv_dfa2_w1h_avg	1.017 [1.009, 1.050]	1.031 [1.005, 1.051]	0.571
hrv_dfa2_w1h_std	0.155 [0.144, 0.161]	0.152 [0.142, 0.167]	0.324

Table 44 HRV markers aggregated from 6-hour windows, with their median [IQR] distributions in the study cohort grouped by doctor-reported flare status (Yes/No), along with associated p-values indicating statistical differences. Markers with p-values less than 0.05 are indicated by purple background shading.

HRV marker	DOCTOR FLARE Yes (N=13) Median [IQR]	DOCTOR FLARE No (N=63) Median [IQR]	p value
Time domain markers			
hrv_rmssd_w6h_avg	33.964 [29.778, 37.008]	40.274 [35.361, 47.516]	**0.005**
hrv_rmssd_w6h_std	6.002 [4.835, 7.970]	6.848 [5.237, 8.852]	0.230

HRV marker	DOCTOR FLARE Yes (N=13) Median [IQR]	DOCTOR FLARE No (N=63) Median [IQR]	p value
hrv_sdnn_w6h_avg	88.589 [77.254, 102.658]	99.792 [88.823, 122.112]	**0.021**
hrv_sdnn_w6h_std	21.331 [19.820, 24.943]	24.881 [20.091, 31.274]	0.113
hrv_hti_w6h_avg	22.273 [19.271, 27.019]	25.302 [22.198, 29.846]	**0.025**
hrv_hti_w6h_std	6.157 [4.970, 6.686]	6.700 [5.704, 7.947]	0.061
Frequency domain markers			
hrv_ulf_w6h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.499
hrv_ulf_w6h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.720
hrv_vlf_w6h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.679
hrv_vlf_w6h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.720
hrv_lf_w6h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.517
hrv_lf_w6h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	1.000
hrv_hf_w6h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.600
hrv_hf_w6h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.762
hrv_lfhf_w6h_avg	1.525 [1.348, 1.676]	1.612 [1.312, 1.864]	0.956
hrv_lfhf_w6h_std	0.409 [0.274, 0.531]	0.389 [0.261, 0.670]	0.869
hrv_tp_w6h_avg	0.001 [0.000, 0.001]	0.000 [0.000, 0.001]	0.619
hrv_tp_w6h_std	0.000 [0.000, 0.001]	0.000 [0.000, 0.001]	0.934
Non-linear markers			
hrv_dfa1_w6h_avg	1.057 [1.032, 1.127]	1.083 [1.044, 1.133]	0.699
hrv_dfa1_w6h_std	0.064 [0.060, 0.110]	0.070 [0.052, 0.098]	0.195
hrv_dfa2_w6h_avg	1.100 [1.092, 1.125]	1.102 [1.077, 1.132]	0.762
hrv_dfa2_w6h_std	0.088 [0.081, 0.095]	0.079 [0.070, 0.092]	0.172

Table 45 HRV markers aggregated from 12-hour windows, with their median [IQR] distributions in the study cohort grouped by doctor-reported flare status (Yes/No), along with associated p-values indicating statistical differences. Markers with p-values less than 0.05 are indicated by purple background shading.

HRV marker	DOCTOR FLARE Yes (N=13) Median [IQR]	DOCTOR FLARE No (N=63) Median [IQR]	p value
Time domain markers			
hrv_rmssd_w12h_avg	33.422 [30.060, 37.261]	40.314 [35.410, 47.579]	**0.004**
hrv_rmssd_w12h_std	4.477 [3.494, 5.508]	5.176 [4.096, 6.879]	0.133
hrv_sdnn_w12h_avg	106.014 [94.852, 116.398]	114.745 [99.699, 140.105]	0.092
hrv_sdnn_w12h_std	20.735 [17.858, 22.665]	21.312 [18.254, 29.755]	0.200
hrv_hti_w12h_avg	26.245 [24.013, 31.529]	29.071 [25.427, 35.235]	0.053
hrv_hti_w12h_std	5.888 [4.633, 6.715]	6.384 [5.505, 7.784]	0.104
Frequency domain markers			
hrv_ulf_w12h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.448
hrv_ulf_w12h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.408
hrv_vlf_w12h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.581
hrv_vlf_w12h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.945
hrv_lf_w12h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.544
hrv_lf_w12h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.535
hrv_hf_w12h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.619
hrv_hf_w12h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.526
hrv_lfhf_w12h_avg	1.533 [1.352, 1.725]	1.536 [1.275, 1.849]	0.956

HRV marker	DOCTOR FLARE Yes (N=13) Median [IQR]	DOCTOR FLARE No (N=63) Median [IQR]	p value
hrv_lfhf_w12h_std	0.324 [0.213, 0.446]	0.298 [0.223, 0.461]	0.720
hrv_tp_w12h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.473
hrv_tp_w12h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.581
Non-linear markers			
hrv_dfa1_w12h_avg	1.059 [1.043, 1.128]	1.073 [1.042, 1.135]	0.847
hrv_dfa1_w12h_std	0.059 [0.042, 0.076]	0.047 [0.037, 0.070]	0.225
hrv_dfa2_w12h_avg	1.117 [1.111, 1.134]	1.113 [1.090, 1.139]	0.649
hrv_dfa2_w12h_std	0.060 [0.055, 0.062]	0.056 [0.050, 0.065]	0.600

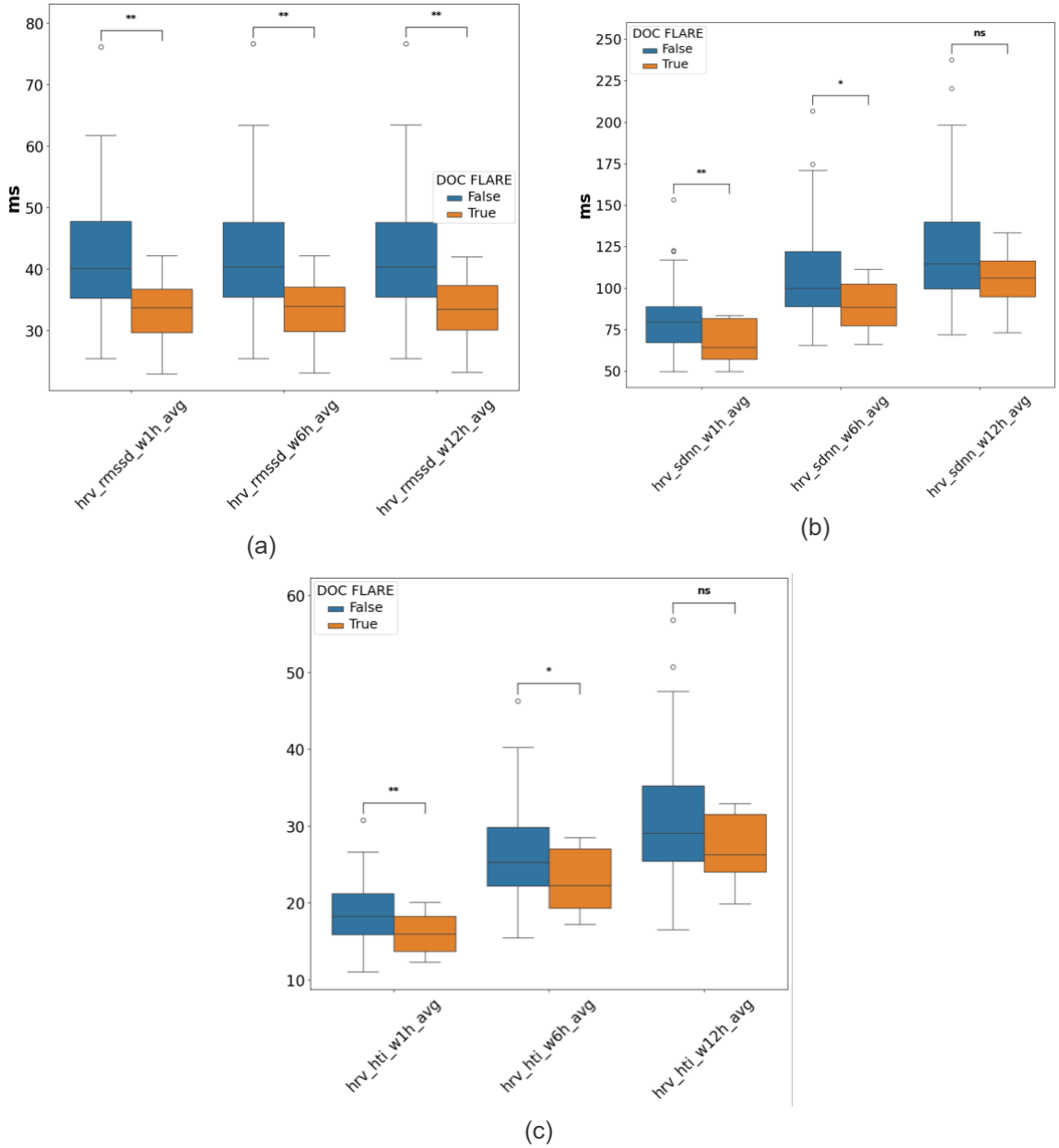


Figure 66 Boxplots of average-aggregated time-domain HRV markers: (a) RMSSD, (b) SDNN, and (c) HTI across 1-, 6-, and 12-hour windows (w=1, 6, 12 h) in the study cohort grouped by doctor-reported flare status. Two stars (**) indicate $p < 0.01$, one star (*) indicates $p < 0.05$, and ns means not significant. These markers suggest reduced HRV irregularity is associated with flare status as defined by doctors.

When using patient-reported flare status (PAT FLARE) as the target variable, the positive group (flare) included 8 patients, while the negative group (no flare) included 48. The distribution of HRV markers as median [IQR], along with Mann–Whitney U test results, are presented in **Tables 46**, **47**, and **48** for markers aggregated over 1-hour, 6-hour, and 12-hour windows, respectively. Significant differences are observed for ***time-domain HRV metrics (RMSSD, SDNN, and triangular index)***. Averaging aggregation performed better across all time-domain metrics and window lengths, whereas standard deviation aggregation showed significant effects primarily for the triangular index. Window length does affect the outcomes. The boxplots of average- and standard deviation-based aggregated time-domain HRV markers are shown in **Figure 67** and **Figure 68**, respectively.

Table 46 HRV markers aggregated from 1-hour windows, with their median [IQR] distributions in the study cohort grouped by patient-reported flare status (Yes/No), along with associated p-values indicating statistical differences. Markers with p-values less than 0.05 are indicated by purple background shading.

HRV marker	PATIENT FLARE Yes (N=8) Median [IQR]	PATIENT FLARE No (N=48) Median [IQR]	p value
Time domain markers			
hrv_rmssd_w1h_avg	32.941 [29.331, 36.438]	40.635 [35.413, 47.475]	**0.006**
hrv_rmssd_w1h_std	8.147 [7.242, 8.885]	9.852 [8.545, 11.831]	**0.049**
hrv_sdnn_w1h_avg	61.313 [56.018, 67.859]	79.963 [68.601, 91.325]	**0.002**
hrv_sdnn_w1h_std	19.999 [18.394, 23.585]	23.458 [20.578, 30.751]	0.065
hrv_hti_w1h_avg	14.870 [13.450, 16.219]	18.559 [16.440, 21.303]	**0.005**
hrv_hti_w1h_std	4.838 [4.266, 5.412]	5.447 [4.951, 6.425]	**0.049**
Frequency domain markers			
hrv_ulf_w1h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.413
hrv_ulf_w1h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.400
hrv_vlf_w1h_avg	0.002 [0.002, 0.002]	0.002 [0.001, 0.002]	0.774
hrv_vlf_w1h_std	0.002 [0.002, 0.002]	0.002 [0.002, 0.002]	0.863
hrv_lf_w1h_avg	0.001 [0.001, 0.001]	0.001 [0.001, 0.001]	0.973
hrv_lf_w1h_std	0.001 [0.001, 0.002]	0.001 [0.001, 0.002]	0.845
hrv_hf_w1h_avg	0.001 [0.001, 0.001]	0.001 [0.001, 0.001]	0.899
hrv_hf_w1h_std	0.001 [0.001, 0.001]	0.001 [0.001, 0.002]	0.936
hrv_lfhf_w1h_avg	1.449 [1.253, 1.611]	1.624 [1.320, 1.905]	0.121
hrv_lfhf_w1h_std	0.552 [0.441, 0.744]	0.591 [0.418, 1.000]	0.512
hrv_tp_w1h_avg	0.004 [0.004, 0.004]	0.004 [0.003, 0.004]	0.881
hrv_tp_w1h_std	0.005 [0.004, 0.005]	0.005 [0.004, 0.005]	0.774
Non-linear markers			
hrv_dfa1_w1h_avg	1.041 [1.012, 1.089]	1.083 [1.048, 1.127]	0.115
hrv_dfa1_w1h_std	0.115 [0.104, 0.140]	0.109 [0.090, 0.141]	0.440
hrv_dfa2_w1h_avg	1.026 [1.011, 1.049]	1.031 [1.006, 1.052]	0.991
hrv_dfa2_w1h_std	0.159 [0.152, 0.165]	0.150 [0.142, 0.167]	0.440

Table 47 HRV markers aggregated from 6-hour windows, with their median [IQR] distributions in the study cohort grouped by patient-reported flare status (Yes/No), along with associated p-values indicating statistical differences. Markers with p-values less than 0.05 are indicated by purple background shading.

HRV marker	PATIENT FLARE Yes (N=8) Median [IQR]	PATIENT FLARE No (N=48) Median [IQR]	p value
Time domain markers			
hrv_rmssd_w6h_avg	33.259 [29.687, 36.736]	40.680 [35.655, 47.143]	**0.007**
hrv_rmssd_w6h_std	5.849 [4.967, 7.000]	6.918 [5.688, 8.832]	0.152
hrv_sdnn_w6h_avg	86.697 [75.859, 90.872]	100.754 [90.113, 126.761]	**0.007**
hrv_sdnn_w6h_std	21.151 [19.132, 23.451]	25.052 [20.655, 31.256]	0.081
hrv_hti_w6h_avg	20.947 [18.995, 22.802]	26.307 [23.027, 29.731]	**0.003**
hrv_hti_w6h_std	5.459 [4.842, 6.257]	6.795 [5.713, 8.234]	**0.012**
Frequency domain markers			
hrv_ulf_w6h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.224
hrv_ulf_w6h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.512
hrv_vlf_w6h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.605
hrv_vlf_w6h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.413
hrv_lf_w6h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.705
hrv_lf_w6h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.468
hrv_hf_w6h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.973
hrv_hf_w6h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.589
hrv_lfhf_w6h_avg	1.416 [1.232, 1.598]	1.590 [1.312, 1.889]	0.198
hrv_lfhf_w6h_std	0.306 [0.250, 0.455]	0.386 [0.264, 0.665]	0.315
hrv_tp_w6h_avg	0.001 [0.000, 0.001]	0.000 [0.000, 0.001]	0.527
hrv_tp_w6h_std	0.001 [0.000, 0.001]	0.000 [0.000, 0.001]	0.637
Non-linear markers			
hrv_dfa1_w6h_avg	1.045 [1.025, 1.092]	1.078 [1.054, 1.131]	0.182
hrv_dfa1_w6h_std	0.062 [0.059, 0.095]	0.070 [0.052, 0.093]	0.688
hrv_dfa2_w6h_avg	1.103 [1.094, 1.110]	1.105 [1.078, 1.134]	0.954
hrv_dfa2_w6h_std	0.088 [0.084, 0.095]	0.077 [0.071, 0.091]	0.146

Table 48 HRV markers aggregated from 12-hour windows, with their median [IQR] distributions in the study cohort grouped by patient-reported flare status (Yes/No), along with associated p-values indicating statistical differences. Markers with p-values less than 0.05 are indicated by purple background shading.

HRV marker	PATIENT FLARE Yes (N=8) Median [IQR]	PATIENT FLARE No (N=48) Median [IQR]	p value
Time domain markers			
hrv_rmssd_w12h_avg	33.013 [29.960, 36.779]	40.546 [35.813, 47.415]	**0.005**
hrv_rmssd_w12h_std	4.379 [4.051, 5.051]	5.227 [4.442, 6.990]	0.077
hrv_sdnn_w12h_avg	101.915 [89.978, 112.487]	118.159 [102.09, 140.43]	**0.036**
hrv_sdnn_w12h_std	20.407 [16.067, 22.897]	22.696 [19.297, 30.044]	0.182
hrv_hti_w12h_avg	25.330 [23.170, 26.630]	29.680 [26.182, 35.403]	**0.011**
hrv_hti_w12h_std	5.261 [4.541, 6.639]	6.653 [5.499, 8.491]	**0.039**
Frequency domain markers			
hrv_ulf_w12h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.282
hrv_ulf_w12h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.159
hrv_vlf_w12h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.497

HRV marker	PATIENT FLARE Yes (N=8) Median [IQR]	PATIENT FLARE No (N=48) Median [IQR]	p value
hrv_vlf_w12h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.827
hrv_lf_w12h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.542
hrv_lf_w12h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.637
hrv_hf_w12h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.792
hrv_hf_w12h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.845
hrv_lfhf_w12h_avg	1.403 [1.235, 1.629]	1.535 [1.278, 1.850]	0.293
hrv_lfhf_w12h_std	0.254 [0.207, 0.398]	0.298 [0.225, 0.438]	0.557
hrv_tp_w12h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.304
hrv_tp_w12h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.527
Non-linear markers			
hrv_dfa1_w12h_avg	1.051 [1.037, 1.099]	1.073 [1.052, 1.128]	0.304
hrv_dfa1_w12h_std	0.053 [0.037, 0.071]	0.046 [0.037, 0.073]	0.809
hrv_dfa2_w12h_avg	1.116 [1.112, 1.126]	1.112 [1.090, 1.148]	0.573
hrv_dfa2_w12h_std	0.060 [0.058, 0.062]	0.055 [0.050, 0.063]	0.326

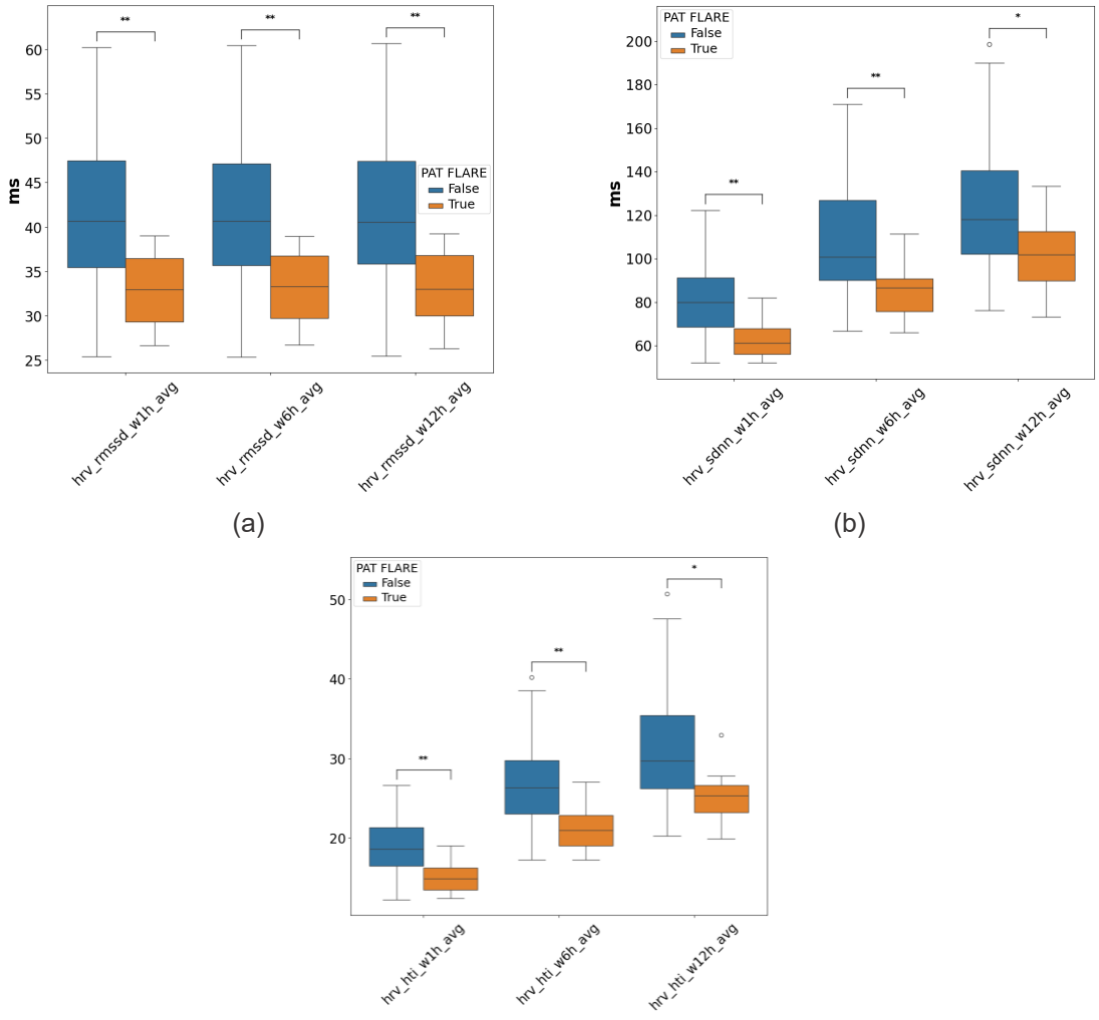


Figure 67 Boxplots of average-aggregated time-domain HRV markers: (a) RMSSD, (b) SDNN, and (c) HTI across 1-, 6-, and 12-hour windows (w=1, 6, 12 h) in in the study cohort grouped by patient-reported flare status. Two stars (**) indicate $p < 0.01$, one star (*) indicates $p < 0.05$, and ns means not significant. These markers suggest reduced HRV irregularity is associated with flare status as defined by patients.

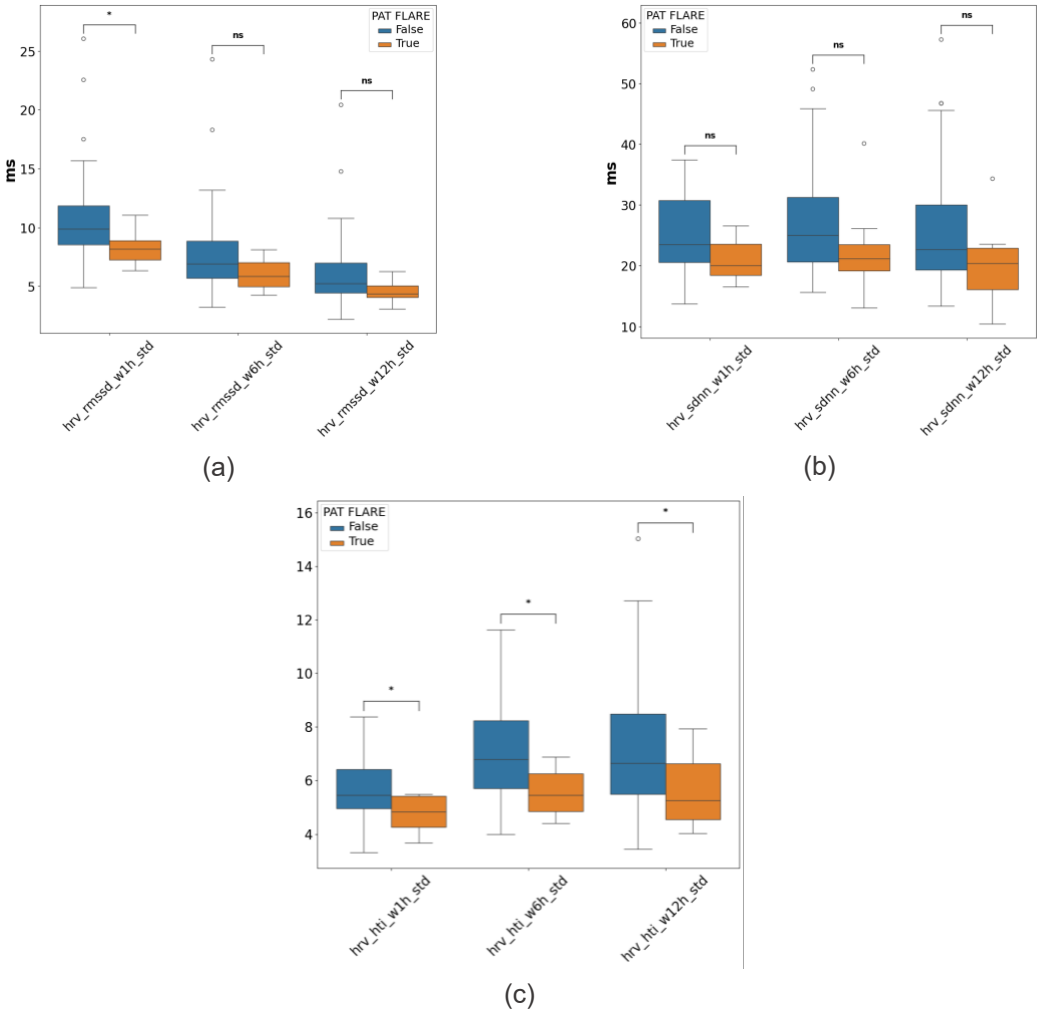


Figure 68 Boxplots of standard deviation-aggregated time-domain HRV markers: (a) RMSSD, (b) SDNN, and (c) HTI across 1-, 6-, and 12-hour windows (w=1, 6, 12 h) in in the study cohort grouped by patient-reported flare status. Two stars (**) indicate $p < 0.01$, one star (*) indicates $p < 0.05$, and ns means not significant. These markers suggest reduced HRV irregularity is associated with flare status as defined by patients.

When using inflammatory flare status, which is defined as positive when CRP > 3 mg/dL, as the target variable, the positive group (flare) included 3 patients, while the negative group (no flare) included 31. The distribution of HRV markers as median [IQR], along with Mann–Whitney U test results, is presented in **Table 49** for markers aggregated over 1-hour. When aggregation over 6- and 12-hour windows is applied no marker shows statistical significance. Significant differences are observed in ***SDNN average and DFAα2 standard deviation markers*** for the 1-hour window. The boxplots of these HRV markers are shown in **Figure 69**.

Table 49 HRV markers aggregated from 1-hour windows, with their median [IQR] distributions in the study cohort grouped by inflammatory flare status (Yes/No), along with associated p-values indicating statistical differences. Markers with p-values less than 0.05 are indicated by purple background shading.

HRV marker	INFLAMMATORY FLARE Yes (N=3) | Median [IQR]	INFLAMMATORY FLARE No (N=31) | Median [IQR]	p value
Time domain markers			
hrv_rmssd_w1h_avg	33.138 [31.320, 34.900]	41.407 [34.345, 48.019]	0.136
hrv_rmssd_w1h_std	9.024 [7.693, 10.111]	10.371 [8.673, 12.664]	0.380
hrv_sdnn_w1h_avg	63.259 [60.162, 64.032]	79.313 [66.302, 90.443]	0.058
hrv_sdnn_w1h_std	18.639 [17.849, 18.697]	23.587 [19.560, 31.669]	0.034

HRV marker	INFLAMMATORY FLARE Yes (N=3) | Median [IQR]	INFLAMMATORY FLARE No (N=31) | Median [IQR]	p value
hrv_hti_w1h_avg	15.311 [14.693, 15.329]	18.860 [15.778, 21.692]	0.092
hrv_hti_w1h_std	4.580 [4.406, 4.635]	5.693 [4.947, 6.886]	0.058
Frequency domain markers			
hrv_ulf_w1h_avg	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.645
hrv_ulf_w1h_std	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.645
hrv_vlf_w1h_avg	0.002 [0.002, 0.002]	0.002 [0.001, 0.002]	0.380
hrv_vlf_w1h_std	0.002 [0.002, 0.002]	0.002 [0.002, 0.002]	0.214
hrv_lf_w1h_avg	0.001 [0.001, 0.001]	0.001 [0.001, 0.001]	0.348
hrv_lf_w1h_std	0.001 [0.001, 0.002]	0.001 [0.001, 0.002]	0.819
hrv_hf_w1h_avg	0.001 [0.001, 0.001]	0.001 [0.001, 0.001]	0.775
hrv_hf_w1h_std	0.001 [0.001, 0.001]	0.001 [0.001, 0.001]	0.819
hrv_lfhf_w1h_avg	1.776 [1.668, 2.073]	1.653 [1.359, 1.865]	0.449
hrv_lfhf_w1h_std	0.937 [0.779, 1.133]	0.649 [0.439, 0.893]	0.262
hrv_tp_w1h_avg	0.004 [0.004, 0.005]	0.004 [0.004, 0.005]	0.318
hrv_tp_w1h_std	0.005 [0.004, 0.005]	0.005 [0.004, 0.005]	0.731
Non-linear markers			
hrv_dfa1_w1h_avg	1.100 [1.078, 1.139]	1.100 [1.045, 1.127]	0.688
hrv_dfa1_w1h_std	0.137 [0.118, 0.153]	0.106 [0.092, 0.133]	0.262
hrv_dfa2_w1h_avg	0.968 [0.967, 0.985]	1.032 [1.000, 1.053]	0.049
hrv_dfa2_w1h_std	0.160 [0.152, 0.164]	0.151 [0.141, 0.164]	0.564

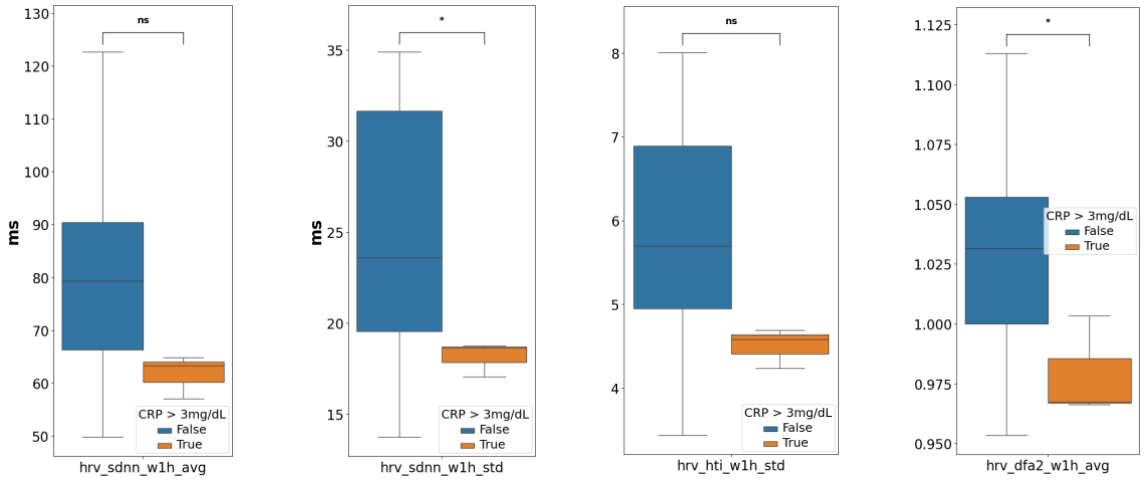


Figure 69 Boxplots of time-domain HRV markers: (a) average-based aggregated SDNN, (b) standard deviation-based aggregated SDNN, (c) standard deviation-based aggregated HTI and (d) average-based aggregated DFA α2, across 1-hour windows in in the study cohort grouped by inflammatory flare status (CRP > 3mg/dL). One star (*) indicates $p < 0.05$, and ns means not significant. The time-domain (SDNN and HTI) features show that lower HRV is associated with inflammatory flare status. DFA α2 feature (non-linear) equals to one indicates normal HRV. In this figure, DFA α2 less than one may be associated with increased inflammation.

11.3.2 Statistical correlation tests

In the statistical correlation assessment between the HRV markers and DAPSA score the sample size was 37. Results of Spearman test, i.e., Spearman ρ and the respective p-value, are presented in **Table 50**. Significant medium correlations are observed in ***HF power***, especially for window of 6 hours. The respective scatter plots (**Figure 70**) display the joint

distributions of statistically significant HRV markers with DAPSA score. It is evident that the correlation is skewed by a few supposed outliers.

Table 50 Results of statistical correlations between HRV markers and DAPSA score. Each entry shows Spearman's ρ with the corresponding p-value (in parentheses). Statistically significant correlations ($p < 0.05$) are highlighted in magenta.

	Window =1h		Window =6h		Window =12h	
	AVG	STD	AVG	STD	AVG	STD
RMSDD	0.05 (0.751)	0.09 (0.580)	0.06 (0.701)	0.09 (0.593)	0.07 (0.672)	0.07 (0.651)
SDNN	-0.03 (0.854)	-0.14 (0.399)	-0.01 (0.917)	-0.06 (0.702)	-0.02 (0.866)	-0.05 (0.754)
Triangular index	-0.01 (0.911)	-0.10 (0.550)	-0.02 (0.883)	-0.07 (0.675)	-0.01 (0.944)	0.01 (0.981)
ULF power	0.20 (0.229)	0.13 (0.412)	0.07 (0.666)	0.10 (0.549)	0.11 (0.483)	0.13 (0.408)
VLF power	0.27 (0.101)	0.26 (0.116)	0.27 (0.100)	0.21 (0.201)	0.25 (0.125)	0.27 (0.100)
LF power	0.21 (0.209)	-0.01 (0.956)	0.24 (0.136)	0.26 (0.112)	0.18 (0.275)	0.25 (0.127)
HF power	0.16 (0.344)	0.03 (0.849)	**0.36 (0.026)**	**0.40 (0.011)**	0.30 (0.062)	**0.36 (0.026)**
LF/HF	-0.05 (0.747)	-0.11 (0.509)	-0.02 (0.860)	-0.11 (0.502)	-0.03 (0.834)	-0.02 (0.881)
Total power	0.28 (0.090)	0.05 (0.742)	0.25 (0.124)	0.24 (0.136)	0.18 (0.283)	0.22 (0.190)
DFA α1	0.05 (0.767)	-0.13 (0.421)	0.05 (0.766)	-0.10 (0.536)	0.03 (0.841)	-0.15 (0.369)
DFA α2	-0.14 (0.376)	-0.08 (0.621)	-0.26 (0.111)	0.20 (0.220)	-0.22 (0.187)	0.17 (0.299)

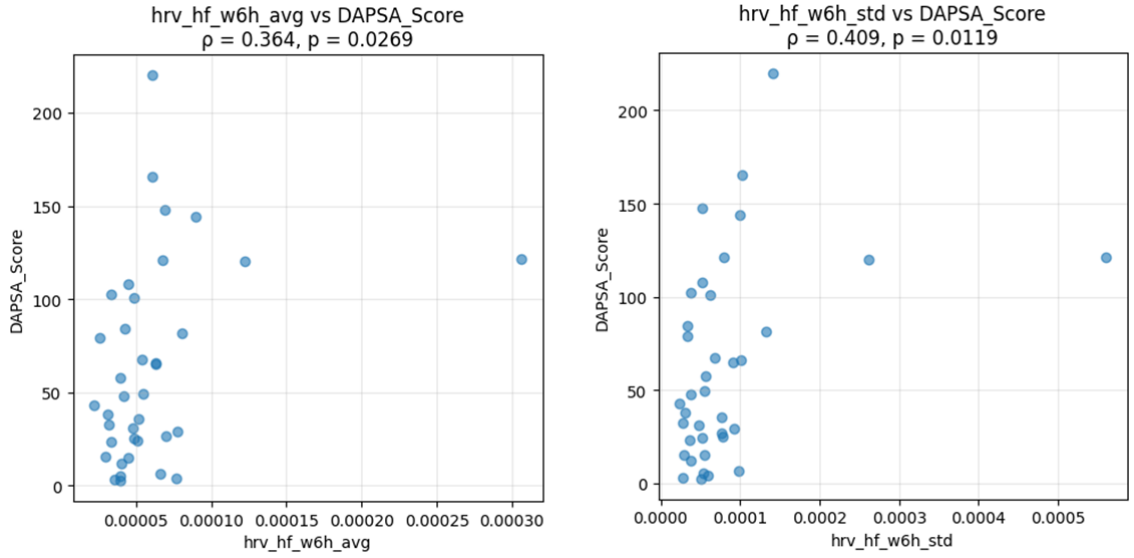


Figure 70 Scatter plots of HF power HRV markers and DAPSA score.

In the statistical correlations between the HRV markers and PASDAS and PSAID scores the sample sizes was 42 and 57, respectively. The results of Spearman tests show no statistically significant correlation between these scores and HRV metrics. The Spearman ρ and the respective p values are presented in Appendix VI Supplementary material of HRV validation.

In the statistical correlation between the HRV markers and CRP score the sample size was 34. Results of Spearman test, i.e., Spearman ρ and the respective p-value, are presented in **Table 51**. Significant medium correlations are observed in ***time-domain HRV metrics (RMSSD, SDNN, and triangular index)***, when average aggregation is applied. Among the window lengths, the 1-hour window yielded the slightly strongest effects. The respective scatter plots (**Figure 71**) display the joint distributions of statistically significant HRV markers with CRP score. It is worth noting that the correlations between the time-domain HRV metrics and the CRP have a negative sign, something that has been observed in previous relevant studies (Williams et al., 2019; Adam et al., 2023).

Table 51 Results of statistical correlations between HRV markers and CRP score. Each entry shows Spearman's ρ with the corresponding p-value (in parentheses). Statistically significant correlations ($p < 0.05$) are highlighted in magenta.

	Window =1h		Window =6h		Window =12h	
	AVG	STD	AVG	STD	AVG	STD
RMSSD	**-0.39 (0.019)**	-0.10 (0.563)	**-0.38 (0.023)**	-0.05 (0.738)	**-0.38 (0.025)**	-0.06 (0.707)
SDNN	**-0.38 (0.025)**	-0.22 (0.191)	**-0.36 (0.031)**	-0.25 (0.151)	-0.33 (0.052)	-0.19 (0.261)
Triangular index	**-0.38 (0.026)**	-0.28 (0.103)	**-0.37 (0.027)**	-0.21 (0.230)	**-0.35 (0.041)**	-0.16 (0.361)
ULF power	-0.09 (0.601)	-0.00 (0.990)	-0.10 (0.547)	-0.00 (0.957)	-0.07 (0.692)	-0.06 (0.722)
VLF power	0.21 (0.221)	0.12 (0.491)	0.16 (0.359)	0.12 (0.469)	0.11 (0.527)	0.11 (0.508)
LF power	0.04 (0.785)	-0.14 (0.408)	0.12 (0.474)	0.15 (0.394)	0.05 (0.758)	0.05 (0.755)
HF power	-0.03 (0.851)	-0.16 (0.338)	0.16 (0.365)	0.05 (0.736)	0.05 (0.774)	0.12 (0.479)
LF/HF	0.13 (0.458)	0.18 (0.295)	0.11 (0.508)	0.16 (0.345)	0.09 (0.598)	0.19 (0.265)
Total power	0.15 (0.379)	-0.10 (0.540)	0.09 (0.588)	0.10 (0.564)	-0.00 (0.997)	-0.01 (0.950)
DFA α1	0.03 (0.838)	0.23 (0.180)	0.02 (0.875)	0.17 (0.312)	0.03 (0.846)	0.12 (0.490)
DFA α2	-0.28 (0.108)	-0.05 (0.746)	-0.17 (0.328)	0.14 (0.409)	-0.10 (0.568)	0.13 (0.438)

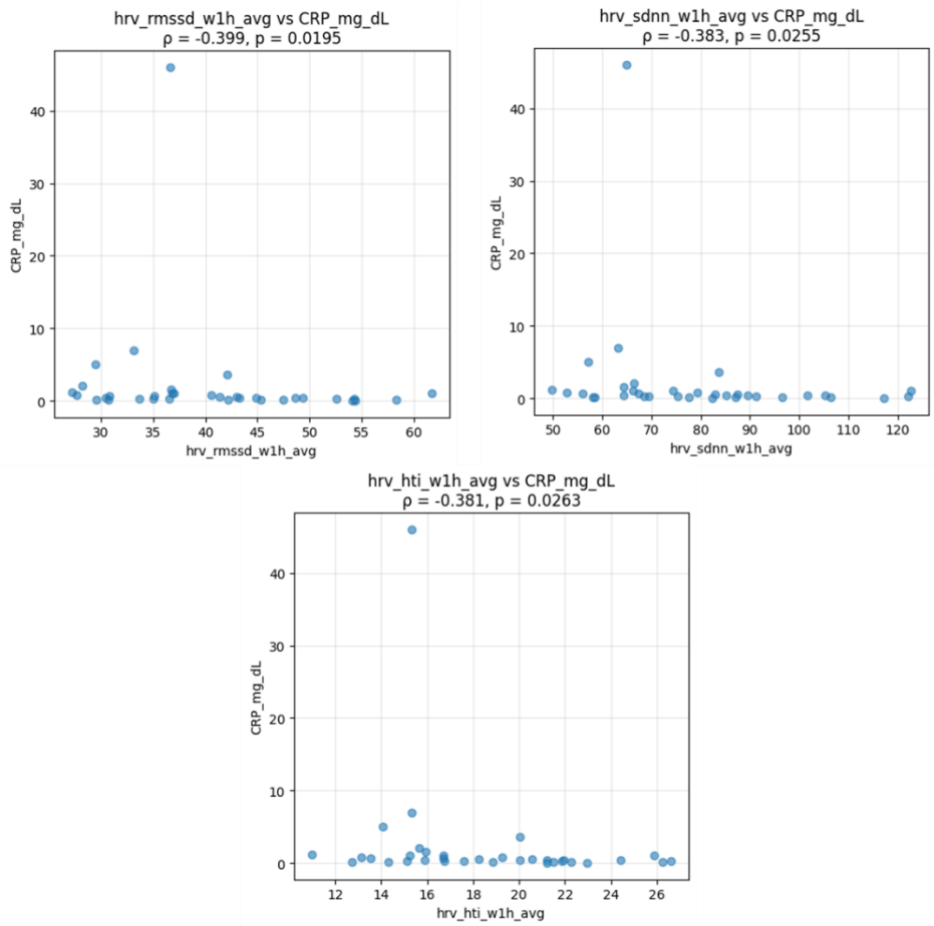


Figure 71 Scatter plots of time-domain HRV markers (RMSSD, SDNN, and triangular index) and CRP score. The red line indicates the linear regression of respective HRV markers and CRP score.

The sample sizes mentioned in this section are slightly different from the ones reported in **Table 40** due to the exclusion of some patients because they have not sufficient BBI recordings as mentioned in the Verification section (§ 11.2.3).

11.4 Discussion

11.4.1 Interpretation

The results indicate an association between HRV markers and disease status. Although stronger outcomes were observed from group-based comparisons using flare status, correlations between HRV markers and clinical scores also provided promising findings. In particular, time-domain markers such as RMSSD, SDNN, and the triangular index appear more closely associated with inflammation. These markers showed a medium negative correlation with CRP, consistent with literature, indicating that reduced HRV reflects suppressed parasympathetic function and diminished activation of anti-inflammatory mechanisms.

The type of aggregation also influenced results. Average aggregation was more frequently associated with significant comparisons and correlations than standard deviation. Regarding time windows, the 1-hour window yielded more significant results with smaller p-values compared to longer windows. A likely explanation is that shorter windows generate more HRV data points, and subsequent aggregation produces more representative markers.

11.4.2 Limitations

The main limitation is the small number of positive cases (patients with flare). Sample size was constrained by three factors: patient recruitment, adherence to smartwatch use, and availability of clinical variables. Increasing the sample size, and thus the number of positive cases, would likely strengthen the statistical evidence for the association between HRV markers and disease-related variables.

We also hypothesize that HRV markers may be less effective in representing disease status at a single time point, but more useful in capturing changes over time. This theory is supported by the role of HRV in reflecting anti-inflammatory activation, which may represent cumulative effect in regulating inflammation. Testing this hypothesis would require longitudinal data with at least two disease status time points.

Another limitation is that our analysis did not account for additional variables such as demographics, medical history, and treatment. Since these factors influence inflammation, their inclusion could complement HRV markers in disease modelling.

Finally, the computational cost of our pipeline is non-negligible. On average, processing each patient required 20 minutes on a workstation (AMD Ryzen 9 3900X 12-core CPU, 32 GiB RAM). This limited experimentation with processing parameters. However, the findings of this study will allow us to make more informed parameter choices in future analyses.

11.5 Work until D3.6

In the coming months, the work will focus on extending the analysis with an expanded dataset that incorporates existing and newly collected data. The processing pipeline will be extended to include additional elements, such as new aggregation types (e.g., skewness and kurtosis), shorter window lengths, grouping BBI segments based on sleep/awake status, and additional non-linear HRV metrics. Demographic, medical history, in-app daily questionnaires and treatment data will also be incorporated. Finally, as the dataset will include follow-up clinical visits, we will investigate associations with changes in disease status over time, rather than status at a single time point

12 Garmin All-day metrics

12.1 Background and objectives

PsA is a chronic inflammatory disease characterized by episodic flares of joint and skin symptoms, fatigue, and impaired quality of life. Recent studies have demonstrated the feasibility and clinical utility of extracting daily metrics from commercial smartwatches (e.g., Fitbit) to model and predict health outcomes in inflammatory conditions. Rao et al. (2023) used daily aggregated features such as step count, heart rate, and sleep stages to predict patient-reported outcomes in rheumatoid arthritis, showing that passively monitored physical activity and heart rate minima are informative for modelling weekly disease fluctuations. Similarly, Wu et al. (2021) developed a deep learning model to predict chronic obstructive pulmonary disease (COPD) exacerbations using daily physiological signals, including heart rate and activity levels, and their first- and second-order derivatives. This work underscored the predictive potential of daily trends and dynamic variability in wearable data.

In the context of arthritis (RA, gout), Gandrup et al. (2024) focused on daily patient-reported symptom scores and derived summary statistics (mean, SD, slope) to forecast flares in RA, indicating the value of temporal aggregation across days. Elmagboul et al. (2020) also supported this by demonstrating that daily step counts, when aggregated and compared across flare versus non-flare periods, were significantly associated with gout flare occurrence.

Together, these studies build a rationale for systematically deriving and evaluating **all-day metrics**—such as daily means, totals, and temporal changes—from smartwatch data streams to serve as **digital correlates of inflammatory activity**.

This work aims to build upon prior studies that have demonstrated the feasibility of using wearable data to predict or classify disease flares in rheumatoid arthritis (RA), gout, and COPD. Specifically, we aim to extend these approaches to psoriatic arthritis (PsA) by extracting all-day physiological and behavioural features from smartwatch data streams, including step count, heart rate, sleep stages, and activity levels. From these raw data, we derive temporal summary statistics such as daily means, standard deviations, intra-week slopes, and first- and second-order differences to capture short-term trends. We then investigate associations between these features and PsA-related inflammation, focusing on physician-reported flare episodes and Minimal Disease activity (MDA) clinical status.

12.2 Methods

12.2.1 Dataset(s)

A new dataset was compiled from PsA patients (n=110) enrolled in the iPROLEPSIS PDPID study. This dataset is notable for two key reasons:

(1) it uses the same hardware device—a GARMIN smartwatch—to collect raw data across all participants, ensuring consistency in data acquisition, and

(2) it includes only individuals diagnosed with psoriatic arthritis, allowing for focused analysis within this patient population. The variables collected are summarized in **Table 52**.

Table 52 All-Day Metrics variables. These variables are collected through Garmin smartwatches in the PDPID study. For each variable, a short description, the range of values, and the sampling period are provided. 'bodyBattery' and 'stressScore' are proprietary Garmin metrics.

Variable	Description	Range of Values	Sampling Period
bodyBattery	Estimated body battery level reflecting energy reserves	0-100	Per minute
beatsPerMinute	Heart rate reading in beats per minute	Positive integer	Per Minute
stressScore	Stress level estimated from physiological signals	0-100	Per minute
steps	Number of steps taken during the activity session	Positive integer	Per activity session
activityType	Categorical label for type of physical activity	SEDENTARY, WALKING, GENERIC, UNMONITORED	Per activity session
walkingTime	Time spent walking	Time in minutes	Per day
SedentaryProportion	Proportion of time spent in sedentary activity	0-1	Per Day

12.2.2 Development

Feature Extraction

We collected raw time-series data from Garmin smartwatch devices, specifically step counts, heart rate, activity type (sedentary or active), stress score, and body battery (the two latter being Garmin proprietary metrics). These data were processed to compute daily aggregates: total step counts, total walking time, proportion of the day spent sedentary, and daily averages for heart rate, stress score, and body battery.

From these daily time series, we extracted features using the tsfresh Python library (**Table 53**). We applied a curated set of features selected based on prior studies that demonstrated predictive value for clinical outcomes using wearable data. These literature-derived features included first- and second-order statistical moments (e.g., mean, variance, standard deviation), extrema (minimum, maximum), trend coefficients (intercept, slope, r-value), and change dynamics (mean change, mean absolute change). We also included “change quantiles” tsfresh feature to approximate the second-order derivative over time. These features were computed for each modality: daily steps, walking time, sedentary proportion, body battery, stress score, and heart rate. We filtered this set of features based on their relevance to the target variable using three approaches: statistical tests, Recursive Feature Elimination with Cross Validation (RFECV) and Elastic Net regression.

Table 53 Descriptions of Selected tsfresh Time Series Features -This table provides clear explanations for several tsfresh features, including some that are more complex or less intuitive, such as second-order differences and linear trend attributes.

Feature Name	Explanation
mean_abs_change	The average of the absolute differences between consecutive time series values. Measures volatility or smoothness of the series.
linear_trend_attr="rvalue"	The correlation coefficient (r-value) from a linear regression of the time series. Indicates how strongly and in what direction the data trends.

Feature Name	Explanation
linear_trend_attr="slope"	The slope of the best-fit line to the time series. Shows the rate and direction of change (positive = increasing trend, negative = decreasing).
linear_trend_attr="intercept"	The intercept of the best-fit line. It's the estimated value at time index 0, i.e., the starting height of the trend line.
change_quantiles__f_agg= "mean"_isabs=True_qh=0.5_ql=0.0 ***(2nd-order diff)***	Computes the mean of the absolute second-order differences between two quantiles (here, 0.5 and 0.0). Measures the curvature or acceleration in the series.

Targets Used

We considered thirteen (13) clinical targets derived from patient-reported outcomes and clinician assessments, specifically: MDA, DAPSA, PASDAS, HAQ total score and walking sub-score, GRCQ, PSS, PSAID pain/fatigue/sleep/functional scores and patient response on perceived stiffness, doctor's response on flare status (Yes/No). All targets were extracted from the corresponding clinical forms.

Imputation

Missing values in the extracted feature matrix were imputed using the mean value across all patients for a given feature. This imputation was applied uniformly to all features before statistical testing and/or model training. Imputation was necessary due to variability in wearable data completeness across patients.

Although this method does not account for inter-individual variability, it preserves the overall feature distribution and avoids introducing complex imputation artifacts. More sophisticated methods such as multivariate imputation or model-based techniques (e.g., KNN or MICE) could be considered in future work, especially if missingness patterns are found to correlate with clinical outcomes.

Concerning the clinical targets, patients with missing values in the targets were dropped from the corresponding analysis.

12.2.3 Verification

Our data collection is based on the Garmin Vivoactive 5 wearable, which offers continuous monitoring of physiological and behavioural variables such as heart rate, physical activity, and stress. In particular, we rely on two proprietary metrics provided by Garmin: the stress score, that is based on Firstbeat Analytic algorithms and compares current HRV to resting HRV[7], and the body battery, which combines stress, HRV, and activity levels into an overall energy index.

Independent validation studies support the accuracy of these proprietary metrics. Notably, a study comparing Garmin's stress score (GSS) with ECG-derived HRV measured using the Polar H10 chest strap found that GSS reliably distinguished rest from mental arithmetic stress and correlated significantly with RMSSD and SD2/SD1 HRV indices (Saha et al., 2022.). Another investigation confirmed that GSS could classify stress states with an AUROC of 0.961, comparable to clinical-grade ECG systems (Abou Faycal et al., 2024).

In terms of activity and heart rate tracking, a systematic review covering 32 studies found that Garmin wearables demonstrate good-to-excellent validity and reliability for step counts, with most studies reporting correlation coefficients above 0.75 and mean absolute percentage error (MAPE) below 10% (Evenson et al., 2015)

[7] van der Mee, D. J., Koyuncu, Z., & Lemmers-Jansen, I. L. J. (2025). Are you stressed or just excited? What the Garmin Stress Score can say about your mood: Journal: Journal of Affective Disorders. *Journal of Affective Disorders Reports*, 100974.

Heart rate measurements were found to be less consistent across activity levels but generally acceptable for free-living data collection.

These findings provide strong justification for the inclusion of stress, energy, and physical activity measures in our study. Additionally, Garmin provides detailed documentation on how body battery and stress score are calculated[8] .

To supplement Garmin's proprietary sensor calibration and documentation, we implement the following internal procedures:

1. **Wear-time filtering**: We treat a bodyBattery value of 0 as indicative of the smartwatch operating in battery-saving mode, which disables continuous tracking. During the PDPID study, we empirically confirmed that enabling battery saver leads to data corruption. These episodes are treated as missing data and excluded from analysis.
2. **Step-rate anomaly detection:** We observed during development-phase before PDPID study launch that some devices recorded implausibly high daily step counts. Hourly step rates occasionally exceeded physiologically realistic values for walking humans. To mitigate this, we excluded data segments in which the hourly step rate exceeded 150 steps per minute, and computed daily tallies based only on periods of regular activity.

12.2.4 Validation

Feature Selection with statistical hypothesis testing

We performed univariate feature selection independently for each clinical target using statistical hypothesis testing. Depending on the task type (regression or classification), either the Kendall tau correlation or the ANOVA F-test was used to assess the significance of each feature. To correct for multiple comparisons, we applied Benjamini–Hochberg false discovery rate (FDR) correction.

Feature selection with Recursive Feature Elimination with Cross Validation (RFECV)

We developed feature selection and modelling pipeline to evaluate the predictive utility of wearable-derived features for both classification and regression tasks. The pipeline was implemented in Python using scikit-learn. All targets were aligned with the wearable feature matrix by matching patient identifiers, ensuring only patients with non-missing target values and valid feature vectors were retained.

Feature Selection with RFECV

We used a Random Forest (RF)-based (classifier or regressor) RFECV strategy to identify the most informative subset of wearable features for each target. RF models were embedded in an RFECV wrapper with 5-fold cross-validation (cv) and used default impurity-based feature importance. RFECV was applied after imputation (using mean value). Feature selection was performed on the full training set for each target and not re-optimized in each cross-validation fold.

Model Training and Evaluation

The features selected through RFECV were then used to train a Random Forest model configured with hyperparameters optimized for small datasets. Specifically, the model employed 200 decision trees to enhance stability and generalization, while limiting the number of features considered at each split to the square root of the total number of features, a common heuristic that helps reduce overfitting in high-dimensional settings. Classification performance was evaluated using 5-fold stratified cross-validation, reporting metrics including

[8] Garmin International. *Vivosmart 4 Owner's Manual*. https://www8.garmin.com/manuals/webhelp/GUID-5D183A14-BB43-4A9B-B441-5F824214CE40/EN-US/GUID-0BB5B9AA-5F7D-4367-95EC-DEC82395A90F.html

balanced accuracy, macro-averaged precision, recall, and F1-score. For regression tasks, we used 5-fold cross-validation, reporting root mean squared error (RMSE) and R^2 score.

Feature Importance

Feature importances were extracted from the final trained model in each fold and averaged across cross-validation folds. For classification tasks, confusion matrices were aggregated and visualized to summarize prediction performance.

Feature selection with Elastic Net Regression

We developed predictive models for clinical outcomes using Elastic Net regularized logistic regression, focusing on two binary endpoints: MDA and Physician-assessed Flare. The binary classification problem is easier to handle with the low amount of data available to us at this point, in contrast elastic net regression models for continuous disease activity scores did not fit well (these results were omitted in this report).

Model Training and Evaluation

We performed model training using 5-fold stratified cross-validation. During each fold, the model was fit using the balanced accuracy as the objective for hyperparameter tuning. The pipeline included mean imputation, standard scaling, and an Elastic Net-penalized logistic regression model (python scikit-learn library). After fitting, we evaluated performance using balanced accuracy, macro-averaged precision, recall, and F1 score on each held-out fold.

Feature Selection

To interpret model outputs, we extracted the non-zero coefficients from the fitted Elastic Net model. These represent features retained after L1-based shrinkage and are considered the most informative for classification. The selected features were ranked by the absolute value of their coefficients.

12.3 Results

Dataset Volume and Missingness Overview

An overview of data collection and missingness, in the two-weeks observation period of this study, is given in **Figure 72** and **73** . To assess wear-time, we verify whether bodyBattery values are recorded every minute and whether those values are positive. When the battery saver mode is enabled on a device, bodyBattery is fixed at ‘0’ and data collection is disrupted. **Figure 72**-top shows the percentage of wear-time per patient, sorted in descending order. **Figure 72**-bottom shows the distribution of total wear-time percentages across all patients. The mean wear-time is 73.6%, and the central 90% of patients have wear-time between 16.2% and 99.8%. **Figure 73** displays a binary heatmap of wear-time per patient at 30-minute resolution: each row represents a patient, and each column represents a 30-minute time bin over the two-week window. White regions indicate that at least 20 minutes of valid data were collected in the bin; black regions indicate missing or unusable data.

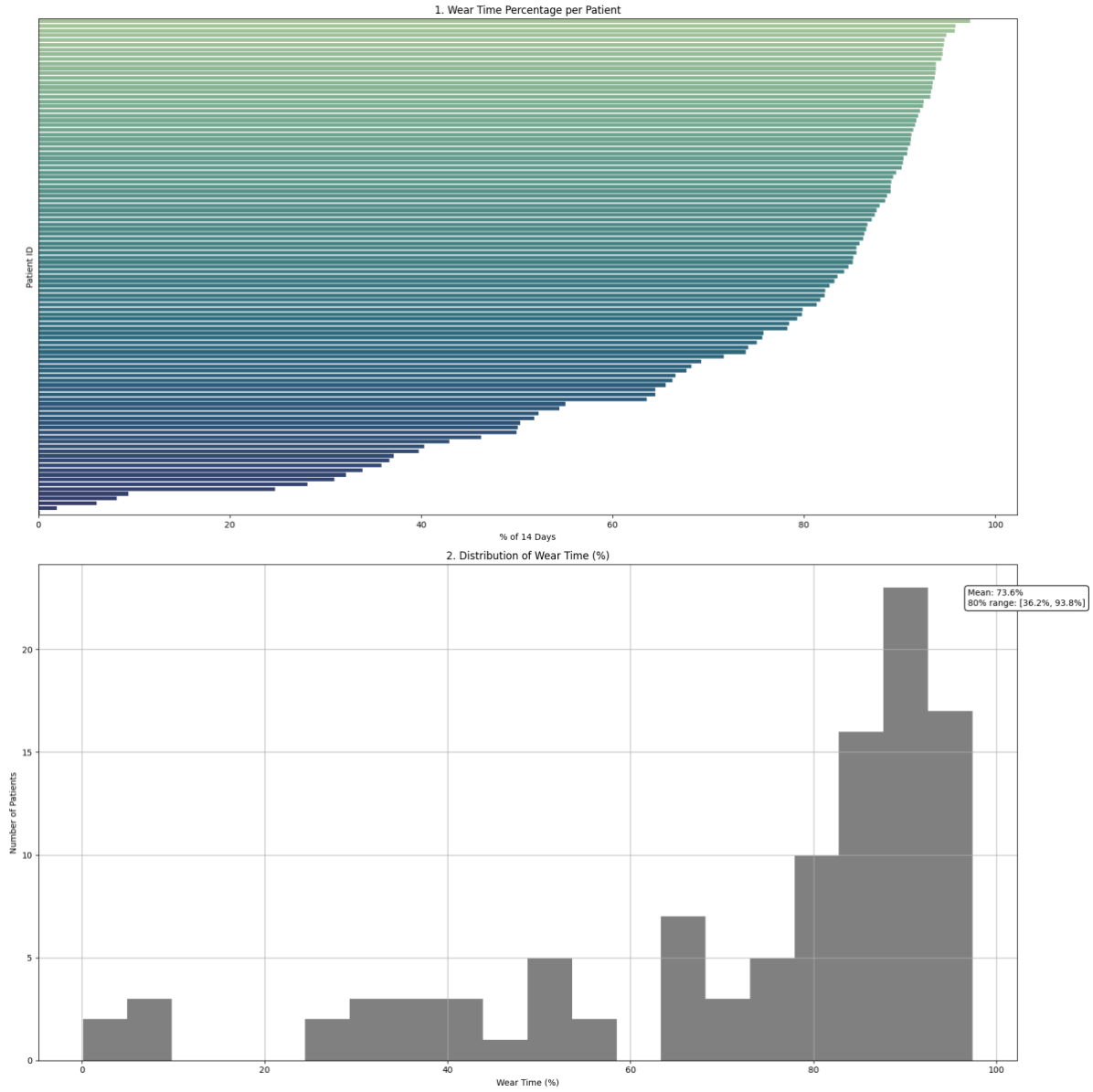


Figure 72 (Top) Wear-Time Percentage per Patient. Horizontal bar plot showing total wear-time percentage for each patient, sorted in descending order. Data loss shown in the diagram combines both participant non-compliance and technical issues". (Bottom) Distribution of Wear Time Across Patients. Histogram of the percentage of observed data per patient. Mean wear-time is 73.6%, with 77.1% of patients above 60%.

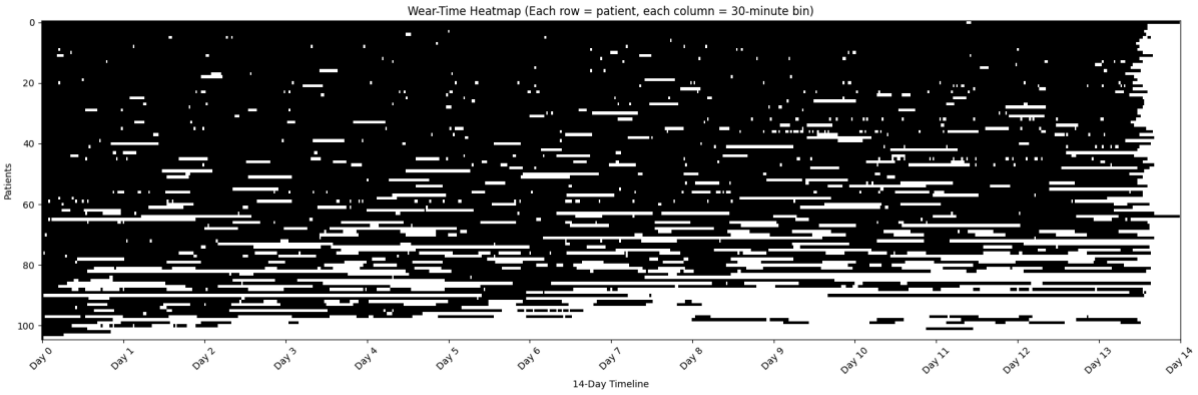


Figure 73 Wear-Time Heatmap. Each row corresponds to a patient, and each column to a 30-minute time bin across the two-week observation period. White = observed (≥20 valid minutes); black = missing.

Importantly, **77.1% of patients have wear-time above 60%**, corresponding to at least 8.4 days of usable data out of the 14-day window. This confirms that the majority of patients contribute substantial wearable data for further analysis.

Beyond overall wear-time coverage, several patterns emerge from the data visualizations. First, a clear subgroup of patients shows near-complete data coverage (>90% wear-time), providing high-quality continuous recordings that can serve as a benchmark for downstream analyses. In contrast, a smaller subset of patients presents with very low adherence (<20% wear-time), forming a long-left tail in the distribution. These low-coverage cases may require special handling, such as exclusion or imputation strategies. The wear-time heatmap also reveals non-random missingness patterns: many patients exhibit extended blocks of missing data, suggesting device removal or syncing issues, while others show sporadic gaps likely reflecting intermittent wear. Additionally, some patients show minimal or no data collection at the beginning of the two-week period, pointing to possible delays in device activation or initial syncing problems. These patterns are informative for refining inclusion criteria, improving onboarding protocols, and selecting appropriate data imputation techniques.

Overview of Physical Activity in the Study Population

To provide an overview of population-level activity characteristics, we examine distributions of key wearable-derived features related to physical activity, including daily steps, sedentary proportion, and walking time (**Figure 74**). These features are visualized with histograms to illustrate variability across participants. Physical activity is a clinically relevant domain in psoriatic arthritis, as reduced mobility, fatigue, and altered activity patterns are common disease manifestations. Describing these distributions therefore provides important context for interpreting subsequent predictive modelling results and frames the digital phenotyping approach in clinically meaningful terms.

Fluctuations in raw activity metrics over 14-day observation periods, illustrated here with daily step count and daily walking time at the patient level (**Figure 75**), provide context for the summary statistics used in later analyses. Two representative patients are shown, one during a flare and another outside a flare. Aggregating features over 14-day windows is intended to smooth out short-term variability unrelated to disease (e.g. weekends versus weekdays) and to highlight more sustained patterns potentially attributable to inflammation (e.g. reduced activity during flare episodes).

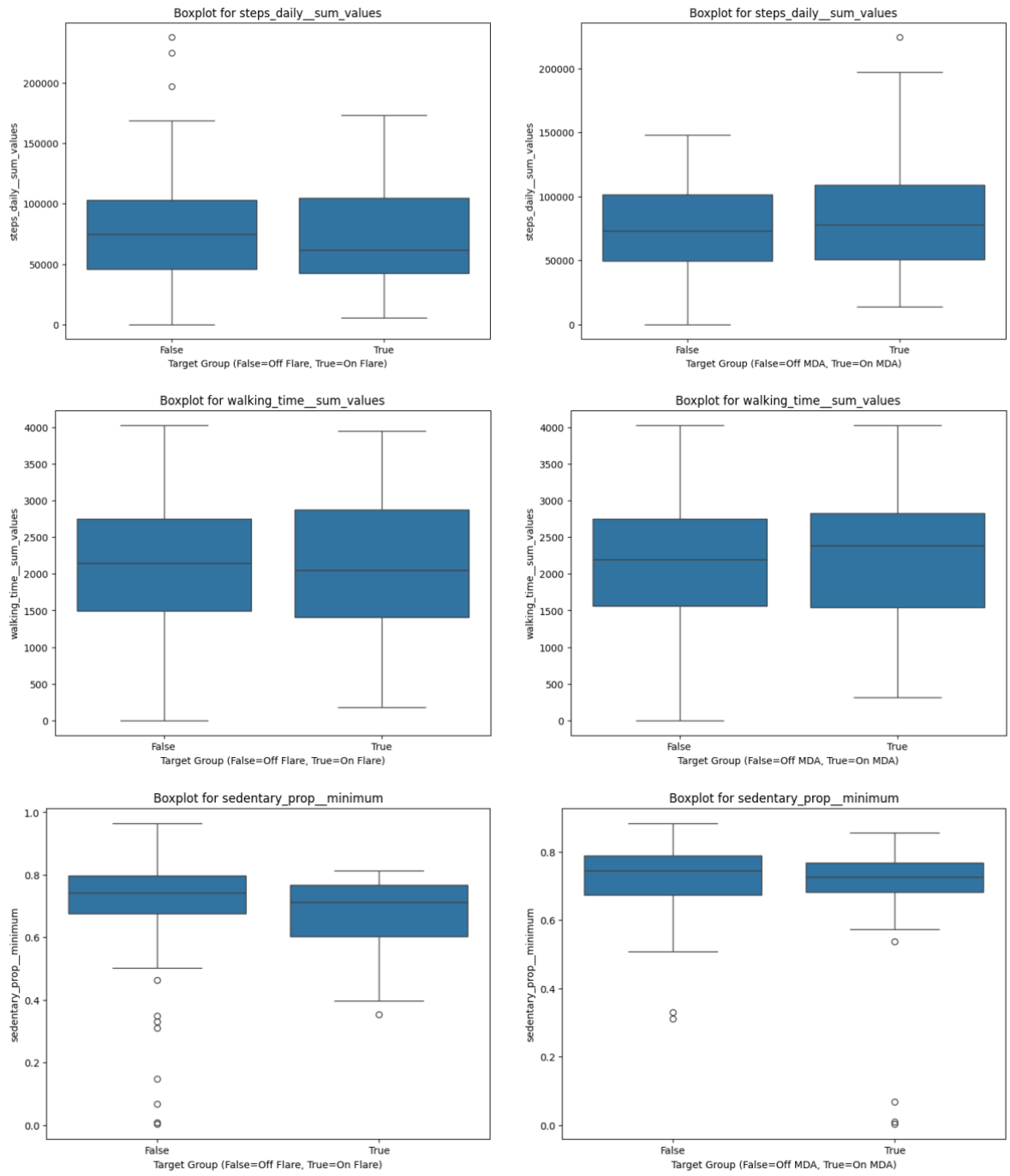


Figure 74 Distribution of physical activity features for both analysis targets. The left column shows flare status (DOC_FLARE) and the right column shows MDA status. Boxplots display group-wise distributions for selected features: total daily steps over the two-week observation period, total walking time, and minimum proportion of time spent sedentary.

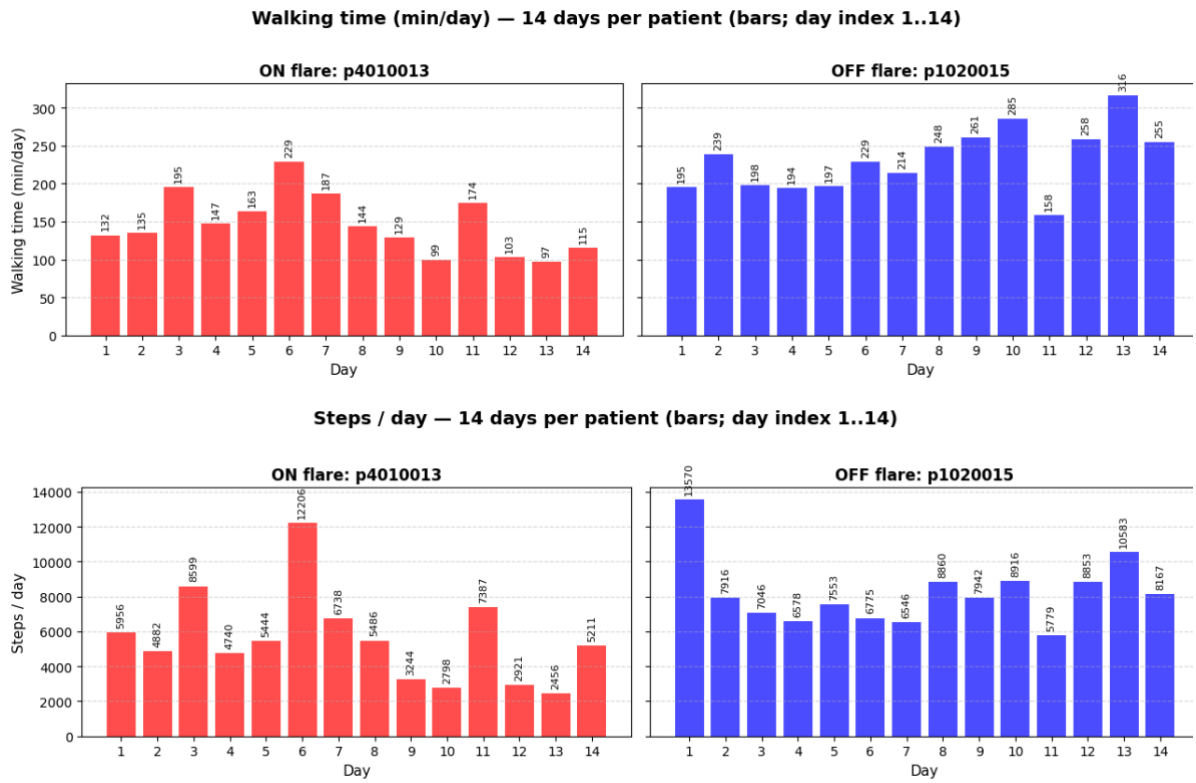


Figure 75 Example of raw activity fluctuations over 14 consecutive days for two patients, one experiencing a flare (left panels, red) and one not in flare (right panels, blue). Daily walking time (top) and daily step count (bottom) are shown. The plots illustrate short-term variability in activity levels and the potential for flare states to manifest as reduced activity.

Feature selection with statistical tests

The full set of extracted features and raw p-values (before applying FDR correction to check for significance) for association with physician flare are given in **Figure 76** and are tabulated in the Appendix (**Table 74**).

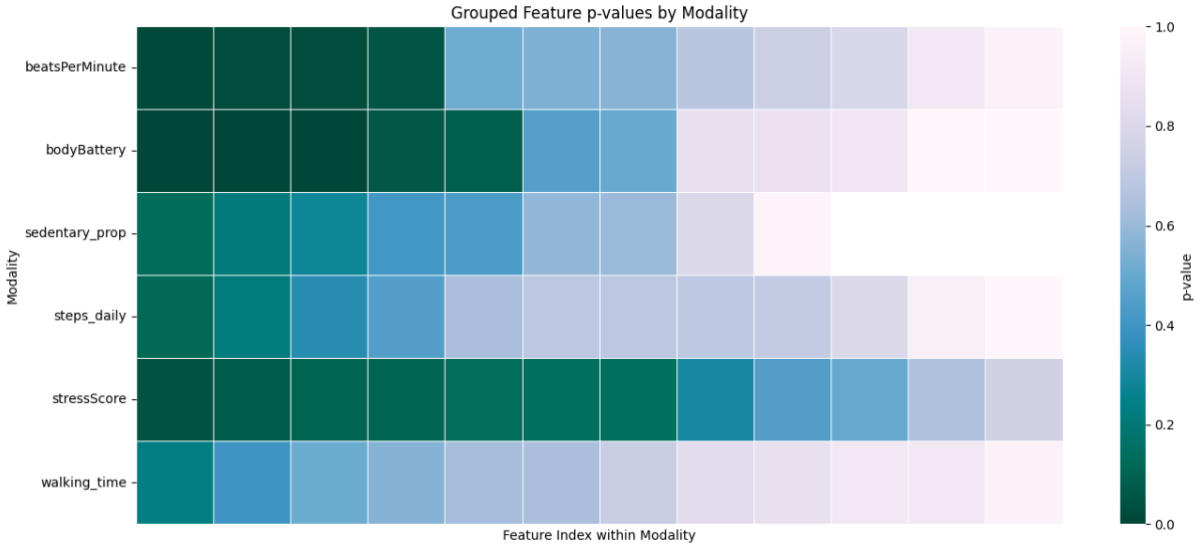


Figure 76 p-values of extracted features grouped by modality. Within each row, features of the same modality are displayed from left to right in ascending order of p-value. While these p-values did not reach statistical significance after FDR comparison, relative significance of each modality can be inferred from this chart.

We observe that stress, heart rate, and physical activity are all relevant to PsA flare, as features from all these modalities reach low (dark green in **Figure 76**) values. However, Garmin's proprietary 'body battery' index appears to capture activity levels more effectively than features based on steps, sedentary time, or walking duration. Applying FDR correction to raw p-values reduces them below p=.05 significance threshold, however with additional participants expected to be enrolled in the PDPID study in the coming months, we anticipate reaching statistical significance in future analyses.

Feature selection with RFECV

With a limited sample size of 110 patients, regression performance was close to random, with R-squared values near zero, indicating poor generalizability. Consequently, we focused our evaluation primarily on classification performance using RFECV (**Table 54**).

Table 54 RFECV classification performance for predicting Minimal Disease Activity (MDA) and physician-assessed flare (DOC_FLARE). The table reports 5-fold cross-validated balanced accuracy, precision, recall, and F1 score (mean ± standard deviation), along with the number of selected features for each target.

Metric	MDA	DOC_FLARE
Number of Features Selected	35	5
Balanced Accuracy	0.595 ± 0.112	0.500 ± 0.065
F1 Score	0.586 ± 0.107	0.475 ± 0.093
Precision	0.629 ± 0.145	0.502 ± 0.212
Recall	0.595 ± 0.112	0.500 ± 0.065

Despite the data constraints, classification performance exceeded random baseline levels for both prediction targets: MDA and physician-assessed flare (DOC_FLARE), without any prior feature pre-selection. The selected features varied between the two targets, underscoring the distinct predictive patterns associated with MDA and physician-assessed flare.

A table with the full list of selected features and respective feature importances can be found in the Appendix (**Table 75**)**.** The top five features selected are given in **Figure 77**, which show than mainly physical activity modalities are selected as important (bodybattery predominantly, and steps).

Figure 77 Top-five selected features, grouped by modality, with RFECV. Features selected for outcome MDA (left) and physician's flare (right).

Feature Selection with Elastic Net Regression

We evaluated the performance of Elastic Net–regularized logistic regression models for predicting Minimal Disease Activity (MDA) and physician-assessed flare (DOC_FLARE).

Models were assessed via 5-fold cross-validation, with results reported as mean ± standard deviation (**Table 55**). Top ten important features are plotted in **Figure 78**. We note that when a patient is experiencing a flare, they are not in MDA (Minimal Disease Activity), and vice versa. Therefore, features that are positively associated with flare should be negatively associated with MDA. This pattern is reflected in the results: heart rate (bpm) features are consistently linked with high disease activity in both the MDA and DOC_FLARE analyses, while features related to steps and stress are more positively associated with remission—that is, achieving MDA.

Table 55 Elastic Net classification performance for predicting Minimal Disease Activity (MDA) and physician-assessed flare (DOC_FLARE)

Metric	MDA	DOC_FLARE
Number of Features Selected	39	24
Balanced Accuracy	0.524 ± 0.053	0.497 ± 0.118
F1 Score	0.456 ± 0.112	0.488 ± 0.117
Precision	0.424 ± 0.151	0.500 ± 0.126
Recall	0.524 ± 0.053	0.497 ± 0.118

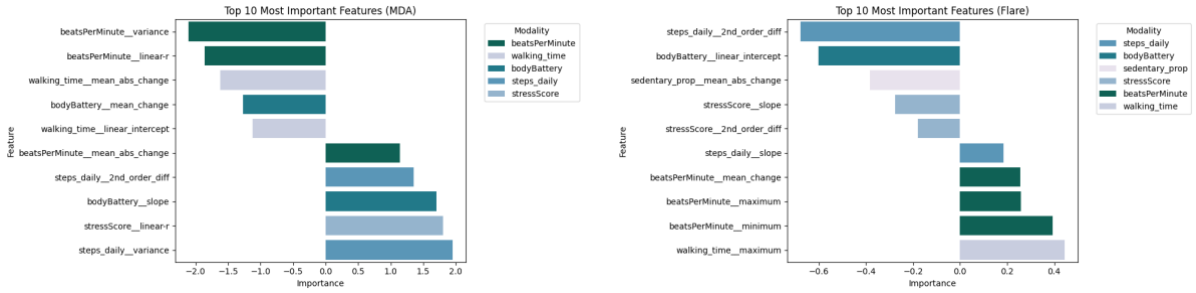


Figure 78 Top-ten selected features, grouped by modality, with Elastic Net. Features selected for outcome MDA (left) and physician's flare (right).

Table 56 Selected regression coefficients highlighting cross-domain feature contributions

Domain	Feature (MDA)	Coefficient	Feature (DOC_FLARE)	Coefficient
Activity	steps_daily__variance	+1.95	steps_daily__change_quantiles	–0.68
Cardiovascular	beatsPerMinute__variance	–2.11	beatsPerMinute__minimum	+0.39
Stress	stressScore__maximum	+0.80	stressScore__linear_trend__attr_"slope"	–0.28
Recovery	bodyBattery__linear_trend__attr_"slope"	+1.71	bodyBattery__linear_trend__attr_"intercept"	–0.60

For the MDA classification task, the logistic regression model achieved a balanced accuracy of 0.52 ± 0.05, with corresponding precision and recall of 0.42 ± 0.15 and 0.52 ± 0.05, respectively. The overall F1 score was 0.45 ± 0.11. Despite its simplicity, the model retained 39 features after Elastic Net regularization, spanning activity (e.g., steps_daily__variance), cardiovascular signals (e.g., beatsPerMinute__variance), and recovery metrics (e.g., bodyBattery__linear_trend__attr_"slope" and stressScore__maximum). Both positive and negative coefficients were observed, highlighting mixed directional associations between features and the MDA outcome. A detailed list of selected features and their regression coefficients is provided in the Appendix (**Table 76**), while a subset highlighting cross-domain contributions is shown in **Table 56**.

In contrast, the DOC_FLARE model yielded slightly lower performance, with a balanced accuracy of 0.49 ± 0.11, precision of 0.50 ± 0.12, recall of 0.49 ± 0.11, and F1 score of 0.48 ± 0.11. The selected feature set was more compact, with 24 non-zero coefficients, and prominently included features such as steps_daily__change_quantiles, bodyBattery__intercept, and beatsPerMinute__minimum. The relatively high variance across folds, along with the modest average performance, reflects both class imbalance and the complexity of flare prediction. Full regression coefficients are reported in the Appendix (**Table 76**), while a subset highlighting cross-domain contributions is shown in **Table 56**.

Overall, these results indicate that logistic regression with Elastic Net regularization can identify informative wearable-derived features for MDA prediction, whereas associations with the DOC_FLARE outcome were more difficult to establish. Performance metrics remained near-chance for DOC_FLARE, suggesting potential limitations in label quality or model expressiveness. Increasing the volume of data will help us improve regression performance and make stronger associations with selected features. A table with the full list of selected features and respective regression coefficients can be found in the Appendix (**Table 76**).

12.4 Discussion

12.4.1 Interpretation

Smartwatch all-day metrics summarizing activity, stress, and overall energy levels were aggregated into 14-day windows using statistical descriptors (means, standard deviations, intra-week slopes, and first- and second-order differences). These were linked to markers of PsA inflammation (physician-reported flare, MDA). While some associations emerged, causal links and stronger inferences will require larger cohorts and longer follow-up.

Comparison of Feature Selection: RFECV vs. Logistic Regression

A comparison of the feature sets selected by RFECV and Elastic Net logistic regression reveals both convergence and divergence in the signals considered predictive for MDA and DOC_FLARE.

Both approaches identified features from the domains of physical activity, cardiovascular function and recovery (e.g. stress score, body battery) as relevant for predicting MDA. Daily step variance and maximum values, variability in heart rate, and markers of recovery consistently appeared across methods. For DOC_FLARE, both models highlighted changes in daily steps, reduced body battery, and alterations in walking activity as predictive. These features reflect fluctuations in physical capacity and recovery, which are consistent with clinical reports of fatigue and reduced activity during flare episodes. Logistic regression captured a broader set of weaker signals, whereas RFECV emphasized a smaller group of high-impact

features, suggesting a trade-off between clinical interpretability and model sensitivity to subtle patterns.

Together, these observations suggest that while both models agree on key physiological domains—particularly activity and recovery—they differ in the granularity and breadth of the features retained. RFECV offers a more interpretable and performance-driven selection, whereas Elastic Net may better accommodate dispersed weak signals in complex outcomes like flare.

12.4.2 Limitations

Limitations of the current analysis include the small sample size (n = 110), a single clinical visit (baseline), and a 14-day data collection window. The ongoing PDPID study observation of patients and the ongoing enrolment of patients in PDPID will effectively lift these limitations in upcoming updated versions of this deliverable.

12.5 Work until D3.6

At the current stage of the study, no features passed the significance threshold after correcting for multiple hypothesis testing. This limitation is likely attributable to the relatively small sample size and the restriction to baseline clinical data alone. To address this, we plan to repeat the feature selection and modelling procedures as additional data become available. Specifically, we anticipate improved statistical power through (1) the enrolment of additional patients, (2) the inclusion of data from follow-up clinical visits, and (3) the use of patient reported flares as target outcomes, which are expected to occur more frequently and thus provide a denser target distribution.

Ultimately, our goal is to evaluate the predictive value of passively collected smartwatch data for flare detection and forecasting in PsA, with the broader aim of advancing disease monitoring through digital health tools.

13 Conclusions

The first report on digital phenotyping for PsA demonstrates the feasibility of developing multi-sensor digital biomarkers to complement clinical and patient-reported outcomes. A summary of the main outcomes of this report is shown in **Table 57**.

Table 57 Deliverable D3.2 summary of the main outcomes.

Joint Range of Motion
• RoM scores correlated with swollen/tender joints and hand disability status based on HAQ, showing that digital exercise tests reflect joint issues. • They can serve as indicators of pain/swelling during daily tasks.
Joint Swelling (Photographs)
• Logistic regression using joint "effective width" with age, BMI, sex, and finger joint type (PIP or DIP) improves swelling detection. • Effective width was the strongest predictor. Promising for monitoring swelling in PsA**.**
Nail Psoriasis Detection
• Hand nail detection worked well; foot nail detection was poor. Classification performance modest (58% accuracy). Needs foot-specific data.
Typing Dynamics
• Typing speed, errors, and accelerometer features correlated with tender hand joints, pain, and hand function. • Reliability moderate to good. • Typing may serve as a passive biomarker.
Physical Activity & Gait from Smartphone IMU
• Walking bouts, cadence, stride regularity, and symmetry can be extracted reliably. • Shows feasibility of smartphone-based gait tracking.
Physical Activity & Gait from Smartwatch Accelerometers
• Flare periods are linked to shorter/fewer walking bouts, lower cadence, and larger activity fragmentation. • Better sleep and physical activity are negatively correlated with disease measures.
Intestinal Motility
• Design of the processing pipelines that will be applied in the smart belt study.
Digital Inflammation Assessment based on Heart Rate Variability
• Preliminary differences in HRV features were observed between patient groups divided according to flare status (as defined by both doctors and patients). • Potential to track inflammation and flare status.
Garmin All-day Metrics
• Both RFECV and Elastic Net methods consistently identified physical activity and recovery features, such as step counts/variance, and body battery, as key predictors for MDA and flare.

Key takeaways:

- A comprehensive framework for digital biomarker development has been implemented, combining clinical requirements, trustworthy AI principles, and usability considerations.

- Initial analyses across several modalities (movement, hand function, nail condition, cardiovascular response, gastrointestinal activity, sleep, and stress) highlight the potential of dBMs to capture meaningful aspects of PsA.
- Preliminary correlations between digital measures and established outcomes are promising but limited by the dataset's size and inclusion of clinical data from a single clinical visit.
- Several technical challenges were identified, including variability in wearable data capture, the need for robust feature selection, and ensuring patient privacy and acceptability.

Future steps:

- Expand the PDPID cohort and include longitudinal follow-up data to increase statistical power and enable flare detection.
- Refine and revalidate dBM pipelines as additional data become available, with emphasis on clinically interpretable outcomes.
- Strengthen predictive modelling of disease activity and flare occurrence.
- Integrate validated dBMs into the miPROLEPSIS patient and clinician app, contributing to proactive, personalised PsA management.

This deliverable provides a strong foundation for advancing digital phenotyping in PsA and sets the stage for subsequent project milestones.

References

Abou Faycal, H., Gallino, C., Schobel, J., & Fornasini, F. (2024). Classifying stress using Garmin wearables: Evaluation against ECG HRV features. arXiv preprint, arXiv:2505.05694. https://arxiv.org/abs/2505.05694

Acar, B., Savelieva, I., Hemingway, H., & Malik, M. (2000). Automatic ectopic beat elimination in short-term heart rate variability measurement. Computer methods and programs in biomedicine, 63(2), 123-131.

Adam, J., Rupprecht, S., Künstler, E. C., & Hoyer, D. (2023). Heart rate variability as a marker and predictor of inflammation, nosocomial infection, and sepsis–A systematic review. Autonomic Neuroscience, 249, 103116.

Bennett, C. C., Ross, M. K., Baek, E., Kim, D., & Leow, A. D. (2022). Smartphone accelerometer data as a proxy for clinical data in modeling of bipolar disorder symptom trajectory. Npj Digital Medicine, 5(1), 181. https://doi.org/10.1038/s41746-022-00741-3

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ digital medicine, 3(1), 18.

Chan, S., Hang, Y., Tong, C., Acquah, A., Schonfeldt, A., Gershuny, J., & Doherty, A. (2024). CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. Scientific Data, 11(1), 1135.

Clinical Trials Transformation Initiative. (2022). Developing novel endpoints generated by digital health technology for use in clinical trials. *Clinical Trials Transformation Initiative*.

Coates, L. C., Fransen, J., & Helliwell, P. S. (2010). Defining minimal disease activity in psoriatic arthritis: a proposed objective target for treatment. Annals of the rheumatic diseases, 69(01), 48-53.

Czech, M. D., & Patel, S. (2019). GaitPy: an open-source python package for gait analysis using an accelerometer on the lower back. Journal of Open Source Software, 4(43), 1778.

DeVore, G. R. (2017). Computing the Z score and centiles for cross-sectional analysis: a practical approach. *Journal of Ultrasound in Medicine*, *36*(3), 459-473.

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., ... & Wareham, N. J. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. PloS one, 12(2), e0169649.

Dures, E., Bowen, C., Brooke, M., Lord, J., Tillett, W., McHugh, N., & Hewlett, S. (2019). Diagnosis and initial management in psoriatic arthritis: a qualitative study with patients. *Rheumatology advances in practice*, *3*(2), rkz022.

Elmagboul, N., Coburn, B. W., Foster, J., Mudano, A., Melnick, J., Bergman, D., Yang, S., Chen, L., Filby, C., Mikuls, T. R., Curtis, J. R., & Saag, K. (2020). *Physical activity measured using wearable activity tracking devices associated with gout flares*. Arthritis Research & Therapy, 22, Article 181

Evenson, K. R., Goto, M. M., & Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. Journal of Medical Internet Research, 17(8), e214. https://doi.org/10.2196/jmir.4998

Föll, S., Maritsch, M., Spinola, F., Mishra, V., Barata, F., Kowatsch, T., ... & Wortmann, F. (2021). FLIRT: A feature generation toolkit for wearable data. Computer Methods and Programs in Biomedicine, 212, 106461.

Gadaleta, M., & Rossi, M. (2018). Idnet: Smartphone-based gait recognition with convolutional neural networks. Pattern Recognition, 74, 25-37.

Gandrup, J., Selby, D. A., & Dixon, W. G. (2024). *Classifying self-reported rheumatoid arthritis flares using daily patient-generated data from a smartphone app: Exploratory analysis applying machine learning approaches*. JMIR Formative Research, 8, e50679. https://doi.org/10.2196/50679

Gill, T. K., Shanahan, E. M., Tucker, G. R., Buchbinder, R., & Hill, C. L. (2020). Shoulder range of movement in the general population: age and gender stratified normative data using a community-based cohort. BMC musculoskeletal disorders, 21, 1-9.

Goldsack, J. C., Coravos, A., Bakker, J. P., Bent, B., Dowling, A. V., Fitzer-Attas, C., ... & Dunn, J. (2020). Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *npj digital Medicine*, *3*(1), 55. https://doi.org/10.1038/s41746-020-0260-4

Goldsack, J. C., Dowling, A. V., Samuelson, D., Patrick-Lake, B., & Clay, I. (2021). Evaluation, acceptance, and qualification of digital measures: from proof of concept to endpoint. *Digital biomarkers*, *5*(1), 53-64. https://doi.org/10.1159/000514730

Google AI for Developers, 2024a, Hand landmarks detection guide https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker

Google AI for Developers, 2024b, Pose landmarks detection guide https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker

Gossec, L., Orbai, A. M., Coates, L. C., Gladman, D. D., Ogdie, A., Pelligra, C. G., ... & de Wit, M. (2024). Validity and score interpretation of the 12-item Psoriatic Arthritis Impact of Disease: an analysis of pooled data from two phase 3 trials of bimekizumab in patients with psoriatic arthritis. RMD open, 10(1), e003548.

Hamy, V., Garcia-Gancedo, L., Pollard, A., Myatt, A., Liu, J., Howland, A., ... & Crouthamel, M. (2020). Developing smartphone-based objective assessments of physical function in rheumatoid arthritis patients: the PARADE study. *Digital biomarkers, 4(1), 26-44.* Hamy, V., Garcia-Gancedo, L., Pollard, A., Myatt, A., Liu, J., Howland, A., ... & Crouthamel, M. (2020). Developing smartphone-based objective assessments of physical function in rheumatoid arthritis patients: the PARADE study. Digital biomarkers, 4(1), 26-44.

Hayano, J., & Yuda, E. (2021). Assessment of autonomic function by long-term heart rate variability: beyond the classical framework of LF and HF measurements. *Journal of physiological anthropology*, *40*(1), 21.

Helliwell, P. S., FitzGerald, O., Fransen, J., Gladman, D. D., Kreuger, G. G., Callis-Duffin, K., ... & Walsh, J. A. (2013). The development of candidate composite disease activity and responder indices for psoriatic arthritis (GRACE project). Annals of the rheumatic diseases, 72(6), 986-991.

Hernández-Hernández, M. V., Sánchez-Pérez, H., Luna-Gómez, C., Ferraz-Amaro, I., & Díaz-González, F. (2021). Impact of disease activity on physical activity in patients with psoriatic arthritis. Arthritis Care & Research, 73(12), 1834-1844.

Houts, C. R., Patrick-Lake, B., Clay, I., & Wirth, R. J. (2020). The path forward for digital measures: suppressing the desire to compare apples and pineapples. Digital Biomarkers, 4(Suppl. 1), 3-12. https://doi.org/10.1159/000511586

Hügle, T., Caratsch, L., Caorsi, M., Maglione, J., Dan, D., Dumusc, A., ... & Kalweit, M. (2022). Dorsal finger fold recognition by convolutional neural networks for the detection and monitoring of joint swelling in patients with rheumatoid arthritis. *Digital biomarkers*, *6*(2), 31-35.

Iakovakis, D., Hadjidimitriou, S., Charisis, V., Bostantzopoulou, S., Katsarou, Z., & Hadjileontiadis, L. J. (2018). Touchscreen typing-pattern analysis for detecting fine motor skills decline in early-stage Parkinson's disease. Scientific Reports, 8(1), 7663. https://doi.org/10.1038/s41598-018-25999-0

Immanuel, S., Teferra, M. N., Baumert, M., & Bidargaddi, N. (2023). Heart rate variability for evaluating psychological stress changes in healthy adults: a scoping review. *Neuropsychobiology*, *82*(4), 187-202.

Ji, C., Wang, H., Bao, C., Zhang, L., Ruan, S., Zhang, J., Gong, T., and Cheng, B. (2021). Challenge of nail psoriasis: an update review. *Clinical Reviews in Allergy & Immunology*, *61*(3), 377-402.

Kaggle (2021). Nail Disease Detection dataset: https://www.kaggle.com/code/reubenindustrustech/nail-disease-detection, Visited: 2025-02-26

Kaggle (2024). https://www.kaggle.com/datasets/nikhilgurav21/nail-disease-detection-dataset/data, Visited: 2025-08-01

Karas, M., Olsen, J., Straczkiewicz, M., Johnson, S. A., Burke, K. M., Iwasaki, S., ... & Onnela, J. P. (2024). Tracking amyotrophic lateral sclerosis disease progression using passively collected smartphone sensor data. *Annals of Clinical and Translational Neurology*, *11*(6), 1380-1392.

Kirk, C., Küderle, A., Micó-Amigo, M. E., Bonci, T., Paraschiv-Ionescu, A., Ullrich, M., ... & University Of Northumbria Hume Emily 61 Megaritis Dimitrios 61 Vogiatzis Ioannis 61. (2024). Mobilise-D insights to estimate real-world walking speed in multiple conditions with a wearable device. Scientific reports, 14(1), 1754.

Knitza, J., Gupta, L., & Hügle, T. (2024). Rheumatology in the digital health era: status quo and quo vadis?. *Nature Reviews Rheumatology*, *20*(12), 747-759.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Konstantinidis, D., Papastratis, I., Dimitropoulos, K., & Daras, P. (2023). Multi-manifold attention for vision transformers. *IEEE Access*, *11*, 123433-123444.

Lam, K., Meijer, K., Loonstra, F., Coerver, E., Twose, J., Redeman, E., Moraal, B., Barkhof, F., de Groot, V., Uitdehaag, B., & Killestein, J. (2021). Real-world keystroke dynamics are a potentially valid biomarker for clinical disability in multiple sclerosis. Multiple Sclerosis Journal, 27(9), 1421–1431. https://doi.org/10.1177/1352458520968797

Liphardt, A. M., Manger, E., Liehr, S., Bieniek, L., Kleyer, A., Simon, D., Tascilar, K., Sticherling, M., Rech, J., Schett, G., & Hueber, A. J. (2020). Similar Impact of Psoriatic

Arthritis and Rheumatoid Arthritis on Objective and Subjective Parameters of Hand Function. ACR open rheumatology, 2(12), 734–740.

Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., ... & Qiao, Y. (2023). Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(10), 12581-12600.

Lorenzetti, S., Ostermann, M., Zeidler, F., Zimmer, P., Jentsch, L., List, R., ... & Schellenberg, F. (2018). How to squat? Effects of various stance widths, foot placement angles and level of experience on knee, hip and trunk motion and loading. BMC Sports Science, Medicine and Rehabilitation, 10, 1-11.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. arXiv 2019. *arXiv preprint arXiv:1906.08172*, *5*.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., ... & Chen, S. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior research methods, 53(4), 1689-1696.

Malik, M. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the European Society of Cardiology and the North American Society for Pacing and Electrophysiology. Annals of Noninvasive Electrocardiology, 1(2), 151-181.

Markus (2024). NailDetection Dataset. Roboflow Universe, https://universe.roboflow.com/markus-gbzwm/naildetection-7b1cw, Visited: 2025-02-26.

Mastoras, R.-E., Iakovakis, D., Hadjidimitriou, S., Charisis, V., Kassie, S., Alsaadi, T., Khandoker, A., & Hadjileontiadis, L. J. (2019). Touchscreen typing pattern analysis for remote detection of the depressive tendency. Scientific Reports, 9(1), 13414. https://doi.org/10.1038/s41598-019-50002-9

Matias, P., Araújo, R., Graça, R., Henriques, A. R., Belo, D., Valada, M., ... & Nunes, F. (2024). COTIDIANA Dataset–Smartphone-Collected Data on the Mobility, Finger Dexterity, and Mental Health of People With Rheumatic and Musculoskeletal Diseases. IEEE Journal of Biomedical and Health Informatics.

McGagh, D., McGowan, N., Hinds, C., Saunders, K. E., & Coates, L. C. (2023). Actigraphy-derived physical activity levels and circadian rhythm parameters in patients with psoriatic arthritis: relationship with disease activity, mood, age and BMI. Therapeutic Advances in Musculoskeletal Disease, 15, 1759720X231174989.

McGagh, D., Song, K., Yuan, H., Creagh, A. P., Fenton, S., Ng, W. F., ... & Coates, L. C. (2024). Digital health technologies to strengthen patient-centred outcome assessment in clinical trials in inflammatory arthritis. The Lancet Rheumatology. http://doi.org/10.1016/s2665-9913(24)00186-3

Micó-Amigo, M. E., Bonci, T., Paraschiv-Ionescu, A., Ullrich, M., Kirk, C., Soltani, A., ... & Mobilise-D consortium. (2023). Assessing real-world gait with digital technology? Validation, insights and recommendations from the Mobilise-D consortium. Journal of neuroengineering and rehabilitation, 20(1), 78.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on

fairness, accountability, and transparency (pp. 220-229). https://doi.org/10.1145/3287560.3287596

Ngeo, J. G., Tamei, T., & Shibata, T. (2014). Continuous and simultaneous estimation of finger kinematics using inputs from an EMG-to-muscle activation model. Journal of neuroengineering and rehabilitation, 11, 1-14.

Ning, E., Cladek, A. T., Ross, M. K., Kabir, S., Barve, A., Kennelly, E., ... & Leow, A. (2023, April). Smartphone-derived Virtual Keyboard Dynamics Coupled with Accelerometer Data as a Window into Understanding Brain Health: Smartphone Keyboard and Accelerometer as Window into Brain Health. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-15). https://doi.org/10.1145/3544548.3580906

Ogdie, A., Coates, L. C., & Mease, P. (2020). Measuring outcomes in psoriatic arthritis. *Arthritis care & research*, *72*, 82-109.

Ozoani, Ezi and Gerchick, Marissa and Mitchell, Margaret. Model Card Guidebook. Hugging Face, 2022. https://huggingface.co/docs/hub/en/model-card-guidebook

Phatak, S., Chakraborty, S., & Goel, P. (2023). Computer vision detects inflammatory arthritis in standardized smartphone photographs in an Indian patient cohort. *Frontiers in Medicine*, *10*, 1280462.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. Pattern recognition, 106, 107404.

Rao, K., Speier, W., Meng, Y., Wang, J., Ramesh, N., Xie, F., Su, Y., Nowell, W. B., Curtis, J. R., & Arnold, C. (2023). Machine learning approaches to classify self-reported rheumatoid arthritis health scores using activity tracker data: Longitudinal observational study. JMIR Formative Research, 7, e43107. https://doi.org/10.2196/43107

Saha, A., Avram, R., Li, L., Mason, T., O'Keefe, S., Lu, Y., ... & Marcus, G. M. (2022). Accuracy of the Garmin stress score in detecting mental stress during rest and arithmetic tasks compared to HRV-derived metrics. JMIR Formative Research, 6(11), e41209. https://doi.org/10.2196/41209

Sassi, R., Cerutti, S., Lombardi, F., Malik, M., Huikuri, H. V., Peng, C. K., ... & Macfadyen, R. (2015). Advances in heart rate variability signal analysis: joint position statement by the e-Cardiology ESC Working Group and the European Heart Rhythm Association co-endorsed by the Asia Pacific Heart Rhythm Society. Ep Europace, 17(9), 1341-1353.

Schoels, M., Aletaha, D., Funovits, J., Kavanaugh, A., Baker, D., & Smolen, J. S. (2010). Application of the DAREA/DAPSA score for assessment of disease activity in psoriatic arthritis. Annals of the rheumatic diseases, 69(8), 1441-1447.

Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. Frontiers in public health, 258.

Small, S. R., von Fritsch, L., Doherty, A., Khalid, S., & Price, A. (2022). OxWalk: Wrist and hip-based activity tracker dataset for free-living step detection and gait recognition.

Stamate, C., Saez Pons, J., Weston, D., & Roussos, G. (2021). PDKit: A data science toolkit for the digital assessment of Parkinson's Disease. PLoS computational biology, 17(3), e1008833.

Straczkiewicz, M., James, P., & Onnela, J. P. (2021). A systematic review of smartphone-based human activity recognition methods for health research. NPJ Digital Medicine, 4(1), 148.

Straczkiewicz, M., Huang, E. J., & Onnela, J. P. (2023a). A “one-size-fits-most” walking recognition method for smartphones, smartwatches, and wearable accelerometers. NPJ Digital Medicine, 6(1), 29.

Straczkiewicz, M., Keating, N. L., Thompson, E., Matulonis, U. A., Campos, S. M., Wright, A. A., & Onnela, J. P. (2023b). Open-Source, Step-Counting Algorithm for Smartphone Data Collected in Clinical and Nonclinical Settings: Algorithm Development and Validation Study. JMIR cancer, 9(1), e47646.

Sztyler, T., & Stuckenschmidt, H. (2016, March). On-body localization of wearable devices: An investigation of position-aware activity recognition. In 2016 IEEE international conference on pervasive computing and communications (PerCom) (pp. 1-9). IEEE.

Than, T., San, A., & Myint, T. T. (2012). Biokinetic study of the wrist joint. Int. J. Collab. Res. Intern. Med. Public Health, 4(5), 450-458.

Tracey, K. J. (2002). The inflammatory reflex. Nature, 420(6917), 853-859.

Tudor-Locke, C., Ducharme, S. W., Aguiar, E. J., Schuna Jr, J. M., Barreira, T. V., Moore, C. C., ... & Staudenmayer, J. (2020). Walking cadence (steps/min) and intensity in 41 to 60-year-old adults: the CADENCE-adults study. International Journal of Behavioral Nutrition and Physical Activity, 17(1), 137.

Vesel, C., Rashidisabet, H., Zulueta, J., Stange, J. P., Duffecy, J., Hussain, F., ... & Leow, A. (2020). Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A BiAffect iOS study. Journal of the American Medical Informatics Association, 27(7), 1007-1018. https://doi.org/10.1093/jamia/ocaa057

Walton, M. K., Powers III, J. H., Hobart, J., Patrick, D., Marquis, P., Vamvakas, S., ... & Burke, L. B. (2015). Clinical outcome assessments: conceptual foundation—report of the ISPOR clinical outcomes assessment–emerging good practices for outcomes research task force. *Value in Health*, *18*(6), 741-752. https://doi.org/10.1016/j.jval.2015.08.006

Webster, D. E., Haberman, R. H., Perez-Chada, L. M., Tummalacherla, M., Tediarjo, A., Yadav, V., ... & Scher, J. U. (2024). Clinical validation of digitally acquired clinical data and machine learning models for remote measurement of psoriasis and psoriatic arthritis: a proof-of-concept study. *The Journal of Rheumatology*, *51*(8), 781-789.

Williams, D. P., Koenig, J., Carnevali, L., Sgoifo, A., Jarczok, M. N., Sternberg, E. M., & Thayer, J. F. (2019). Heart rate variability and inflammation: a meta-analysis of human studies. Brain, behavior, and immunity, 80, 219-226.

Wu, C.-T., Li, G.-H., Huang, C.-T., Cheng, Y.-C., Chen, C.-H., Chien, J.-Y., Kuo, P.-H., Kuo, L.-C., & Lai, F. (2021). *Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: Development and cohort study*. JMIR mHealth and uHealth, 9(5), e22591. https://doi.org/10.2196/22591

Yuan, H., Chan, S., Creagh, A. P., Tong, C., Acquah, A., Clifton, D. A., & Doherty, A. (2024). Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. NPJ digital medicine, 7(1), 91.

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C. L., & Grundmann, M. (2020). Mediapipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214.

Appendix I Model cards

I.1 Model card template

Model name		
Provide a 1-2 sentence summary of what the model is.		
Model Details		
Model description	*Provide basic details about the model. This includes the architecture, version, if it was introduced in a paper, if an original implementation is available, and the creators. Any copyright should be attributed here. General information about training procedures, parameters, and important disclaimers can also be mentioned in this section.*	
Developed by	*List (and ideally link to) the people who built the model.*	
Finetuned from	*If this model has another model as its base, link to that model here.*	
Intended Use		
Direct use	*Explain how the model can be used without fine-tuning, post-processing, or plugging into a pipeline.* *An example code snippet is recommended.*	
Downstream use	*Explain how this model can be used when fine-tuned for a task or when plugged into a larger ecosystem or app.* *An example code snippet is recommended.*	
Out-of-scope use	*List how the model may foreseeably be misused (used in a way it will not work for) and address what users ought not do with the model.*	
Bias, Risks, and Limitations		
Bias	*A bias is a stereotype or disproportionate performance (skew) for some subpopulations.*	
Risk	*A risk is a socially-relevant issue that the model might cause.*	
Limitation	*A limitation is a likely failure mode that can be addressed following the listed Recommendations.*	
Recommendations	*What are recommendations with respect to the foreseeable issues?*	
Training Details		
Training data	*Write 1-2 sentences on what the training data is.*	
Preprocessing	*Filtering, resizing/rewriting (depending on the modality), etc.*	
Evaluation		
Testing data	*Describe testing data*	
Factors	*What are the foreseeable characteristics that will influence how the model behaves? Evaluation should ideally be disaggregated across these factors in order to uncover disparities in performance.*	
Metrics	*What metrics will be used for evaluation?*	
Results	*Results should be based on the Factors and Metrics defined above.*	
Versions		
Version No.	**Release date**	**Description**
v1	10-Nov-2024	+++

I.2 Models in swollen joint analysis

Swollen Joint Classifier	
A logistic regression model that characterises interphalangeal (IP) joints as swollen or not. The logistic regression model uses the effective width of the joint, a metric reflecting joint thickness, as its primary predictor, with age, sex, BMI, and joint type (i.e. proximal IP or distal IP) included as covariates.	
Model Details	
Model description	The logistic regression model that classifies an IP joint as swollen or not is formulated as: $logit\big(P(Swollen = 1)\big) = \beta_0 + \beta_1 \cdot EFW + \beta_2 \cdot Age + \beta_3 \cdot BMI + \beta_4 \cdot MALE + \beta_5 \cdot DIP,$ where *EFW* (effective width), *Age*, and *BMI* (body mass index) are numerical variables, while *MALE* and *DIP* (distal IP) are binary ones. Effective width is a metric to represent the thickness of an IP joint proposed by (Webster et. al, 2024). Effective width calculation involves developing a pipeline of computer vision algorithms that segment individual fingers, detect joint landmarks, and compute distances between points of interest. In this pipeline, existing AI models are used, i.e. MediaPipe Hand Landmarker (Lugaresi et. al, 2019) for hand landmarks detection and U2-Net pretrained for human segmentation (Qin et. al, 2020) to segment hands. In this version, the swollen joint classifier was validated employing k-fold cross-validation, with each validation fold containing precisely one patient exhibiting at least one swollen joint. The model has not been published yet.
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki)
Finetuned from	No finetuning
Intended Use	
Direct use	The model takes as input 1. the effective width of an IP joint, extracted via a computer vision pipeline applied to a hand photograph, and 2. personal information (age, body mass index, and sex). It outputs whether the joint is swollen. The model can be integrated into a broader procedure that assesses all IP joints of a hand and may be used in systems for remote monitoring of joint conditions in individuals with joint-related health issues.
Downstream use	This model does not have a downstream use.
Out-of-scope use	The swollen joint classifier has no clear out-of-scope uses. The effective width pipeline requires hand photos that include the wrist, with a background that contrasts with the skin tone.
Bias, Risks, and Limitations	
Bias	The model performance may vary across skin tones, with potential bias toward lighter skin
Risk	No socially-relevant issue is associated with this model
Limitation	Classify only IP joints, i.e. joints of hand fingers (excluding thumb)
Recommendations	End-users should not rely solely on the model's output and should assess their joints independently. Developers should include a disclaimer stating that the model is experimental and its results must not be used for clinical decision-making.

Training Details		
Training data	The dataset consists of 85 patients with psoriatic arthritis, and it was evenly distributed by sex (42 male, 43 female), with a mean age of 52.6 years (SD 11.8) and a mean BMI of 28.94 kg/m2 (SD 6.43). Of the 85 participants, ten had at least one swollen interphalangeal joint, with a total of 40 swollen interphalangeal joints reported in the cohort.	
Preprocessing	No preprocessing needed	
Evaluation		
Testing data	In this version, the model was validated via cross-validation. No evaluation took place.	
Factors	-	
Metrics	-	
Results	-	
Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept development. Experimental stage.

I.3 Models in typing dynamics

Hand tenderness (TJC hand >0) from smartphone typing dynamics and accelerometer	
Binary logistic regression (RIDGE) model predicting any tender hand joint (TJC hand >0) vs none from typing dynamics (custom Android keyboard) paired with in-typing-session smartphone accelerometer.	
Model Details	
Model description	• Architecture: Logistic regression (RIDGE); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters C 0.1; penalty l2; solver liblinear; max_iter 5000; class_weight balanced. • Training setup: Nested CV—outer LOSO; inner 5-fold (subject-wise) for tuning. • Feature set (Typing & Acc. v1): Daily aggregates covering hold/flight times stats, speed stats, error rate, chars/word per session, phone orientation, motion intensity/variability, motion dynamics. Covariates: age, sex, education, handedness, typing style (one- vs two-handed). • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score participant-level summaries to get probability of TJC hand >0.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age, education, digital literacy, language, typing style, and phone handling.
Risk	False negatives may delay attention; false positives may prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, stylus/pen users excluded (may matter in the wild); one- vs two-handed typing not explicitly modeled.
Recommendations	Expand/longitudinal data, add medication/context covariates, and calibrate.
Training Details	
Training data	• Data & labels: n=46 PsA participants (PDPID study) after valid-week QC; TJC Hand > 0 n=19, TJC Hand = 0 n=27; window = first 14 days post-baseline. • Features referenced: Hold time, flight time and speed stats, error rate, chars/word per session; phone orientation, motion intensity/variability, motion dynamics; Full list provided in **Table 18** and **Table 19**. • Covariates: age, sex, education, handedness, typing style. • Model selection: Nested CV—outer LOSO (subject-wise) for generalization; inner 5-fold subject-level CV for hyperparameter tuning; per subject predictions aggregated by majority voting. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	• Typing sessions: keyboard shown→hidden; keep sessions with ≥50 characters. Derive hold time, flight time, typing speed (each:

	mean/SD/skew/kurtosis/slope), error rate (backspaces/total chars), characters/word. QC: hold ∈ [0, 500] ms; flight within 1st–99th pct of real-world distribution; negative flights removed; flight times mean-centered per session. • Accelerometer pairing: align each session to the concurrent segment; resample to 20 Hz. Orientation/tilt (low-pass ≤3 Hz: tilt mean/SD, x/z medians); motion intensity/variability (low-pass on linear: mean/SD/skew, entropy, energy, AUC); dynamic movement (band-pass 3–9 Hz: jerk mean/SD, peak-to-peak, peak time). • Aggregation: per-participant daily medians of all session-level features.
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=46; majority voting on per-subject predictions).
Factors	Demographics (age, sex, education, handedness, typing style); not modeled in v1: stylus/pen users, medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.73 [0.57, 0.88]; κ 0.52 [0.26, 0.78]; F1 0.76 [0.63, 0.89].

Versions		
Version No.	**Release date**	**Description**
v1	10-Nov-2024	Proof-of-concept; nested LOSO; Ridge Logistic Regression; Typing & Acc v1 features; experimental stage.

Hand tenderness (TJC hand >0) from smartphone typing dynamics	
Binary logistic regression (LASSO) model predicting any tender hand joint (TJC hand >0) vs none from typing dynamics (custom Android keyboard).	
Model Details	
Model description	• Architecture: Logistic regression (LASSO); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters C 0.1; penalty l1; solver liblinear; max_iter 5000; class_weight balanced. • Training setup: Nested CV—outer LOSO; inner 5-fold (subject-wise) for tuning. • Feature set (Typing v1): Daily aggregates covering hold/flight times stats, speed stats, error rate, and chars/word per session. Covariates: age, sex, education, handedness, typing style (one- vs two-handed). • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score participant-level summaries to get probability of TJC hand >0.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	

Bias	Possible performance differences across age, education, digital literacy, language, typing style, and phone handling.
Risk	False negatives may delay attention; false positives may prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, stylus/pen users excluded (may matter in the wild); one- vs two-handed typing not explicitly modeled.
Recommendations	Expand/longitudinal data, add medication/context covariates, and calibrate.
Training Details	
Training data	• Data & labels: n=46 PsA participants (PDPID study) after valid-week QC; TJC Hand > 0 n=19, TJC Hand = 0 n=27; window = first 14 days post-baseline. • Features referenced: Hold/flight times and speed stats, error rate, and chars/word per session. Full list provided in **Table 18**. • Covariates: age, sex, education, handedness, typing style. • Model selection: Nested CV—outer LOSO (subject-wise) for generalization; inner 5-fold subject-level CV for hyperparameter tuning; per subject predictions aggregated by majority voting. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	• Typing sessions: keyboard shown→hidden; keep sessions with ≥50 characters. Derive hold time, flight time, typing speed (each: mean/SD/skew/kurtosis/slope), error rate (backspaces/total chars), characters/word. QC: hold ∈ [0, 500] ms; flight within 1st–99th pct of real-world distribution; negative flights removed; flight times mean-centered per session. • Aggregation: per-participant daily medians of all session-level features.
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=46; majority voting on per-subject predictions).
Factors	Demographics (age, sex, education, handedness, typing style); not modeled in v1: stylus/pen users, medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.70 [0.53, 0.85]; κ 0.39 [0.16, 0.63]; F1 0.69 [0.56, 0.81].

Versions		
Version No.	**Release date**	**Description**
v1	10-Nov-2024	Proof-of-concept; nested LOSO; LASSO Logistic Regression; Typing v1 features; experimental stage.

Hand tenderness (TJC hand >0) from smartphone accelerometer during typing.	
Binary logistic regression (RIDGE) model predicting any tender hand joint (TJC hand >0) vs none from smartphone accelerometer during typing.	
Model Details	
Model description	• Architecture: Logistic regression (RIDGE); Python 3.11.13, scikit-learn 1.7.2.

	• Hyperparameters C 0.1; penalty l2; solver liblinear; max_iter 5000; class_weight balanced. • Training setup: Nested CV—outer LOSO; inner 5-fold (subject-wise) for tuning. • Feature set (Acc-only. v1): Daily aggregates covering phone orientation, motion intensity/variability, and motion dynamics. Covariates: age, sex, education, handedness, typing style (one- vs two-handed). • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score participant-level summaries to get probability of TJC hand >0.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age, education, digital literacy, language, typing style, and phone handling.
Risk	False negatives may delay attention; false positives may prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, stylus/pen users excluded (may matter in the wild); one- vs two-handed typing not explicitly modeled.
Recommendations	Expand/longitudinal data, add medication/context covariates, and calibrate.
Training Details	
Training data	• Data & labels: n=46 PsA participants (PDPID study) after valid-week QC; TJC Hand > 0 n=19, TJC Hand = 0 n=27; window = first 14 days post-baseline. • Features referenced: Phone orientation, motion intensity/variability, and motion dynamics; Full list provided in **Table 19**. • Covariates: age, sex, education, handedness, typing style. • Model selection: Nested CV—outer LOSO (subject-wise) for generalization; inner 5-fold subject-level CV for hyperparameter tuning; per subject predictions aggregated by majority voting. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	• Typing sessions: keyboard shown→hidden; keep sessions with ≥50 characters. • Accelerometer pairing: align each session to the concurrent segment; resample to 20 Hz. Orientation/tilt (low-pass ≤3 Hz: tilt mean/SD, x/z medians); motion intensity/variability (low-pass on linear: mean/SD/skew, entropy, energy, AUC); dynamic movement (band-pass 3–9 Hz: jerk mean/SD, peak-to-peak, peak time). • Aggregation: per-participant daily medians of all session-level features.
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=46; majority voting on per-subject predictions).

Factors	Demographics (age, sex, education, handedness, typing style); not modeled in v1: stylus/pen users, medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.70 [0.52, 0.86]; κ 0.48 [0.22, 0.73]; F1 0.73 [0.60, 0.87].

Versions		
Version No.	**Release date**	**Description**
v1	10-Nov-2024	Proof-of-concept; nested LOSO; RIDGE Logistic Regression; Acc-only v1 features; experimental stage.

I.4 Models in smartphone physical activity and gait

Flare (physician) from smartphone accelerometer	
Binary ElasticNet logistic regression model predicting physician-reported PsA flare vs. no flare from weekly physical-activity features derived from smartphone tri-axial accelerometer data.	
Model Details	
Model description	• Architecture: Logistic regression (ElasticNet); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters C 1.0; l1_ratio 0.9; penalty elasticnet; solver saga; max_iter 5000; class_weight balanced. • Training setup: Nested CV—outer LOSO; inner 5-fold for tuning. • Feature set (Smartphone PA v1): Weekly aggregates covering activity volume, fragmentation, and mobility; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly feature bundles to obtain flare probability.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device/firmware, phone-carriage behavior.
Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, device/placement heterogeneity, class imbalance.
Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=82 PsA participants (PDPID study) after valid-week QC; flare n=20, no-flare n=72; window = first 14 days post-baseline. • Daily volume (steps; minutes by class; ENMO), fragmentation (bout lengths & transitions), mobility (≥10-min walking bouts, cadence stats); summarized weekly. Full list provided in **Table 26**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Raw tri-axial acceleration resampled to 10 Hz; reduced to vector magnitude (ENMO (mg)). Detect walking via published algorithm (Straczkiewicz et al., 2023a); CWT to estimate dominant step frequency; derive per-second steps, then aggregate to 1-min. Classify minutes as Inactive (0 steps), Light (0 < steps/min < 100), MVPA (≥ 100 steps/min).

Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=92).
Factors	Demographics (sex, age, BMI), disease duration; data availability compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.60 [0.46, 0.74]; κ 0.22 [0.04, 0.42]; F1 0.45 [0.29, 0.59].

Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; Elastic-Net Logistic Regression; Smartphone PA v1 features; experimental stage.

Flare (patient) from smartphone accelerometer	
Binary XGBoost classifier predicting patient-reported PsA flare vs. no flare from weekly physical activity features derived from smartphone tri-axial accelerometer data.	
Model Details	
Model description	• Architecture: Gradient-boosted decision trees (XGBoost); Python 3.11.13, xgboost 3.0.5. • Hyperparameters: n_estimators 200, learning_rate 0.1. max_depth 7; min_child_weight 1; subsample 0.6; colsample_bytree 0.8; gamma 0.5; scale_pos_weight 1; objective binary:logistic; eval_metric auc; tree_method hist; random_state/seed 42. • Training setup: Nested CV—outer LOSO; inner 5-fold for tuning. • Feature set (Smartphone PA v1): Weekly aggregates covering activity volume, fragmentation, and mobility; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly feature bundles to obtain flare probability.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device/firmware, phone-carriage behavior.
Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, device/placement heterogeneity, class imbalance.

Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=68 PsA participants (PDPID study) after valid-week QC; flare n=12, no-flare n=56; window = first 14 days post-baseline. • Daily volume (steps; minutes by class; ENMO), fragmentation (bout lengths & transitions), mobility (≥10-min walking bouts, cadence stats); summarized weekly. Full list provided in **Table 26**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Raw tri-axial acceleration resampled to 10 Hz; reduced to vector magnitude (ENMO (mg)). Detect walking via published algorithm (Straczkiewicz et al., 2023a); CWT to estimate dominant step frequency; derive per-second steps, then aggregate to 1-min. Classify minutes as Inactive (0 steps), Light (0 < steps/min < 100), MVPA (≥ 100 steps/min).
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=68).
Factors	Demographics (sex, age, BMI), disease duration; data availability compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.77 [0.64, 0.88]; κ 0.37 [0.19, 0.59]; F1 0.53 [0.35, 0.70].
Versions	

Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; XGBoost; Smartphone PA v1 features; experimental stage.

MDA (achieved) from smartphone accelerometer	
Binary XGBoost classifier predicting Minimal Disease Activity (MDA) achieved vs. not achieved from weekly physical-activity features derived from smartphone tri-axial accelerometer data.	
Model Details	
Model description	• Architecture: Gradient-boosted decision trees (XGBoost); Python 3.11.13, xgboost 3.0.5. • Hyperparameters: n_estimators 200, learning_rate 0.05. max_depth 5; min_child_weight 10; subsample 1.0; colsample_bytree 1.0; gamma 0; scale_pos_weight 10; objective binary:logistic; eval_metric auc; tree_method hist; random_state/seed 42. • Training setup: Nested CV—outer LOSO; inner 5-fold for tuning. • Feature set (Smartphone PA v1): Weekly aggregates covering activity volume, fragmentation, and mobility; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).

Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly Score weekly feature bundles to obtain probability of MDA achieved.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device/firmware, phone-carriage behavior.
Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, device/placement heterogeneity.
Recommendations	Expand/longitudinal data, add medication/context covariates, and calibrate.
Training Details	
Training data	• Data & labels: n=58 PsA participants (PDPID study) after valid-week QC; MDA (achieved) n=29, MDA (not achieved) n=29; window = first 14 days post-baseline. • Daily volume (steps; minutes by class; ENMO), fragmentation (bout lengths & transitions), mobility (≥10-min walking bouts, cadence stats); summarized weekly. Full list provided in **Table 26**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Raw tri-axial acceleration resampled to 10 Hz; reduced to vector magnitude (ENMO (mg)). Detect walking via published algorithm (Straczkiewicz et al., 2023a); CWT to estimate dominant step frequency; derive per-second steps, then aggregate to 1-min. Classify minutes as Inactive (0 steps), Light (0 < steps/min < 100), MVPA (≥ 100 steps/min).
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=58).
Factors	Demographics (sex, age, BMI), disease duration; data availability compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.70 [0.56, 0.83]; κ 0.52 [0.31, 0.73]; F1 0.76 [0.63, 0.87].
Versions	

Version No.	**Release date**	**Description**

v1	1-Sep-2025	Proof-of-concept; LOSO training; XGBoost; Smartphone PA v1 features; experimental stage.

Lower-extremity tenderness from smartphone accelerometer	
Binary logistic regression (LASSO) model predicting presence vs. absence of tender lower-extremity joints (clinical TJC68 restricted to lower extremity) from weekly smartphone-derived physical-activity features.	
Model Details	
Model description	• Architecture: Logistic regression (LASSO); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters C 0.1; penalty l1; solver liblinear; max_iter 5000; class_weight balanced. • Training setup: Nested CV—outer LOSO; inner 5-fold for tuning. • Feature set (Smartphone PA v1): Weekly aggregates covering activity volume, fragmentation, and mobility; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly feature bundles to obtain probability of lower-extremity tenderness.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device/firmware, phone-carriage behavior.
Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, device/placement heterogeneity.
Recommendations	Expand/longitudinal data, add medication/context covariates, and calibrate.
Training Details	
Training data	• Data & labels: n=95 PsA participants (PDPID study) after valid-week QC; TJC Lower>0 n=45, TJC Lower=0 n=72; window = first 14 days post-baseline. • Daily volume (steps; minutes by class; ENMO), fragmentation (bout lengths & transitions), mobility (≥10-min walking bouts, cadence stats); summarized weekly. Full list provided in **Table 26**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning; subject-wise splits

	prevent leakage. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Raw tri-axial acceleration resampled to 10 Hz; reduced to vector magnitude (ENMO (mg)). Detect walking via published algorithm (Straczkiewicz et al., 2023a); CWT to estimate dominant step frequency; derive per-second steps, then aggregate to 1-min. Classify minutes as Inactive (0 steps), Light (0 < steps/min < 100), MVPA (≥ 100 steps/min).
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=95).
Factors	Demographics (sex, age, BMI), disease duration; data availability compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.65 [0.53, 0.76]; κ 0.31 [0.13, 0.49]; F1 0.62 [0.50, 0.74].
Versions	

Version No.	Release date	Description
v1	1-Sep-2025	Proof-of-concept; LOSO training; L1 Logistic Regression; Smartphone PA v1 features; experimental stage.

Pain severity from smartphone accelerometer	
Binary logistic regression (LASSO) model predicting no/mild pain (PSAID Pain ≤ 4) vs. moderate/severe pain (PSAID Pain > 4) from weekly smartphone-derived physical-activity features.	
Model Details	
Model description	• Architecture: Logistic regression (LASSO); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters C 10; penalty l1; solver liblinear; max_iter 5000; class_weight balanced. • Training setup: Nested CV—outer LOSO; inner 5-fold for tuning. • Feature set (Smartphone PA v1): Weekly aggregates covering activity volume, fragmentation, and mobility; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly feature bundles to obtain pain severity probability.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device/firmware, phone-carriage behavior.

Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, device/placement heterogeneity, class imbalance.
Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=71 PsA participants (PDPID study) after valid-week QC; no/mild pain n=48, moderate/severe pain n=23; window = first 14 days post-baseline. • Daily volume (steps; minutes by class; ENMO), fragmentation (bout lengths & transitions), mobility (≥10-min walking bouts, cadence stats); summarized weekly. Full list provided in **Table 26**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning; subject-wise splits prevent leakage. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Raw tri-axial acceleration resampled to 10 Hz; reduced to vector magnitude (ENMO (mg)). Detect walking via published algorithm (Straczkiewicz et al., 2023a); CWT to estimate dominant step frequency; derive per-second steps, then aggregate to 1-min. Classify minutes as Inactive (0 steps), Light (0 < steps/min < 100), MVPA (≥ 100 steps/min).
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=71).
Factors	Demographics (sex, age, BMI), disease duration; data availability compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.58 [0.46, 0.71]; κ 0.23 [0.05, 0.42]; F1 0.54 [0.38, 0.68].

Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; L1 Logistic Regression; Smartphone PA v1 features; experimental stage.

Fatigue severity from smartphone accelerometer	
Binary logistic regression model (LASSO) predicting no/mild fatigue (PSAID fatigue ≤ 4) vs moderate/severe fatigue (PSAID fatigue > 4) from weekly smartphone-derived physical-activity features.	
Model Details	
Model description	• Architecture: Logistic regression (LASSO); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters C 1.0; penalty l1; solver liblinear; max_iter 5000; class_weight balanced.

	• Training setup: Nested CV—outer LOSO; inner 5-fold for tuning. • Feature set (Smartphone PA v1): Weekly aggregates covering activity volume, fragmentation, and mobility; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly feature bundles to obtain fatigue severity probability.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device/firmware, phone-carriage behavior.
Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, device/placement heterogeneity, class imbalance.
Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=71 PsA participants (PDPID study) after valid-week QC; no/mild fatigue n=47, moderate/severe fatigue n=24; window = first 14 days post-baseline. • Daily volume (steps; minutes by class; ENMO), fragmentation (bout lengths & transitions), mobility (≥10-min walking bouts, cadence stats); summarized weekly. Full list provided in **Table 26**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Raw tri-axial acceleration resampled to 10 Hz; reduced to vector magnitude (ENMO (mg)). Detect walking via published algorithm (Straczkiewicz et al., 2023a); CWT to estimate dominant step frequency; derive per-second steps, then aggregate to 1-min. Classify minutes as Inactive (0 steps), Light (0 < steps/min < 100), MVPA (≥ 100 steps/min).
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=71).
Factors	Demographics (sex, age, BMI), disease duration; data availability compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.59 [0.46, 0.73]; κ 0.28 [0.08, 0.51]; F1 0.54 [0.39, 0.68].

Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; L1 Logistic Regression; Smartphone PA v1 features; experimental stage.

Stiffness severity from smartphone accelerometer	
Binary logistic regression (LASSO) model predicting no/mild stiffness (≤ 4) vs. moderate/severe stiffness (> 4) from weekly smartphone-derived physical-activity features.	
Model Details	
Model description	• Architecture: Logistic regression (LASSO); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters C 0.1; penalty l1; solver liblinear; max_iter 5000; class_weight balanced. • Training setup: Nested CV—outer LOSO; inner 5-fold for tuning. • Feature set (Smartphone PA v1): Weekly aggregates covering activity volume, fragmentation, and mobility; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly feature bundles to obtain stiffness severity probability.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device/firmware, phone-carriage behavior.
Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, device/placement heterogeneity.
Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=95 PsA participants (PDPID study) after valid-week QC; no/mild stiffness n=52, moderate/severe stiffness n=43; window = first 14 days post-baseline. • Daily volume (steps; minutes by class; ENMO), fragmentation (bout lengths & transitions), mobility (≥10-min walking bouts, cadence stats); summarized weekly. Full list provided in **Table 26**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models.

	• Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Raw tri-axial acceleration resampled to 10 Hz; reduced to vector magnitude (ENMO (mg)). Detect walking via published algorithm (Straczkiewicz et al., 2023a); CWT to estimate dominant step frequency; derive per-second steps, then aggregate to 1-min. Classify minutes as Inactive (0 steps), Light (0 < steps/min < 100), MVPA (≥ 100 steps/min).
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=95).
Factors	Demographics (sex, age, BMI), disease duration; data availability compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.70 [0.59, 0.81]; κ 0.41 [0.24, 0.58]; F1 0.67 [0.55, 0.78].

Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; L1 Logistic Regression; Smartphone PA v1 features; experimental stage.

I.5 Models in smartwatch physical activity and gait

Flare (physician) from smartwatch accelerometer	
A binary classifier that predicts whether a PsA patient is in flare or not (physician-reported ground truth), using weekly aggregated physical-activity features extracted from smartwatch accelerometer data (Garmin Vivoactive 5).	
Model Details	
Model description	• Architecture: Gradient-boosted decision trees (XGBoost); Python 3.11.13, xgboost 3.0.5. • Hyperparameters: n_estimators=200, learning_rate=0.1. max_depth: 7; min_child_weight: 1; subsample: 0.6; colsample_bytree: 0.8; gamma: 1.0; scale_pos_weight: 1; objective: binary:logistic; eval_metric: auc; tree_method: hist; random_state/seed: 42 • Training setup: Nested CV—outer LOSO; inner 5-fold for tuning. • Feature set (PA v1): Weekly aggregates covering volume, fragmentation, morning activity, sleep, walking; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Model can be used to score weekly feature bundles and output a probability of flare.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device types, or wear-time compliance.
Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, reliance on consistent wear-time, class imbalance.
Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=86 PsA participants (PDPID study) after valid-week QC; flare n=16, no-flare n=70; window = first 14 days post-baseline. • Features (PA v1): Daily volume, intensity (ENMO/cadence), fragmentation (bout stats & transitions—incl. MVPA→Sed), sleep, morning activity; summarized weekly. Full list provided in **Table 34**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.

Preprocessing	Wrist tri-axial signals (24/7) segmented into 30-s windows; activity recognition via DCNN HAR—SSL pre-trained on UK Biobank (Yuan et al., 2024) fine-tuned on Capture-24 (Chan et al., 2024) with HMM; parallel walking/steps model—OxWalk (Small et al., 2023)); per-window outputs: class, ENMO, steps, walking flag.
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=86).
Factors	Demographics (sex, age, BMI), disease duration/severity; wear-time compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.66 [0.51, 0.81]; κ 0.30 [0.07, 0.56]; F1 0.46 [0.28, 0.65].

Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; XGBoost; PA v1 features; experimental stage.

Flare (patient) from smartwatch accelerometer	
A binary classifier that predicts whether a PsA patient is in flare or not (patient-reported ground truth), using weekly aggregated physical-activity features extracted from smartwatch accelerometer data (Garmin Vivoactive 5).	
Model Details	
Model description	• Architecture: Logistic regression (RIDGE); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters: C: 100; penalty: l2; solver: liblinear; max_iter: 5000; class_weight: balanced. • Training setup: Nested CV—outer LOSO; inner 5-fold for tuning. • Feature set (PA v1): Weekly aggregates covering volume, fragmentation, morning activity, sleep, walking; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Model can be used to score weekly feature bundles and output a probability of flare.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device types, or wear-time compliance.

Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, reliance on consistent wear-time, class imbalance.
Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=64 PsA participants (PDPID study) after valid-week QC; flare n=10, no-flare n=54; window = first 14 days post-baseline. • Features (PA v1): Daily volume, intensity (ENMO/cadence), fragmentation (bout stats & transitions—incl. MVPA→Sed), sleep, morning activity; summarized weekly. Full list provided in **Table 34**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Wrist tri-axial signals (24/7) segmented into 30-s windows; activity recognition via DCNN HAR—SSL pre-trained on UK Biobank (Yuan et al., 2024) fine-tuned on Capture-24 (Chan et al., 2024) with HMM; parallel walking/steps model—OxWalk (Small et al., 2023)); per-window outputs: class, ENMO, steps, walking flag.
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=64).
Factors	Demographics (sex, age, BMI), disease duration/severity; wear-time compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.82 [0.65, 0.94]; κ 0.44 [0.18, 0.73]; F1 0.55 [0.31, 0.78].

Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; L2 Logistic Regression; PA v1 features; experimental stage.

MDA (achieved) from smartwatch accelerometer	
Binary logistic regression model predicting Minimal Disease Activity (MDA) achieved vs. not achieved from weekly physical-activity features derived from smartwatch accelerometer data (Garmin Vivoactive 5).	
Model Details	
Model description	• Architecture: Logistic regression (LASSO); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters: C: 1.0; penalty: l1; solver: liblinear; max_iter: 5000; class_weight: balanced. • Training setup: Nested CV—outer LOSO; inner 5-fold for tuning.

	• Feature set (PA v1): Weekly aggregates covering volume, fragmentation, morning activity, sleep, walking; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly feature bundles to obtain probability of MDA achieved.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device types, or wear-time compliance.
Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, reliance on consistent wear-time.
Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=53 PsA participants after valid-week QC; MDA (achieved) n=25, MDA (not achieved) n=28; window = first 14 days post-baseline. • Features (PA v1): Daily volume, intensity (ENMO/cadence), fragmentation (bout stats & transitions—incl. MVPA→Sed), sleep, morning activity; summarized weekly. Full list provided in **Table 34**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Wrist tri-axial signals (24/7) segmented into 30-s windows; activity recognition via DCNN HAR—SSL pre-trained on UK Biobank (Yuan et al., 2024) fine-tuned on Capture-24 (Chan et al., 2024) with HMM; parallel walking/steps model—OxWalk (Small et al., 2023)); per-window outputs: class, ENMO, steps, walking flag.
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=53).
Factors	Demographics (sex, age, BMI), disease duration/severity; wear-time compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.66 [0.50, 0.80]; κ 0.38 [0.15, 0.59]; F1 0.69 [0.53, 0.83].

Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; L1 Logistic Regression; PA v1 features; experimental stage.

Pain severity from smartwatch accelerometer	
Binary ElasticNet logistic regression predicting no/mild pain (PSAID pain ≤4) vs. moderate/severe pain (>4) using weekly physical-activity features from smartwatch accelerometer data (Garmin Vivoactive 5).	
Model Details	
Model description	• Architecture: Logistic regression (ElasticNet); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters: C: 0.1; l1_ratio: 0.5; penalty: "elasticnet"; solver: "saga"; max_iter: 5000; class_weight: "balanced". • Training setup: Nested CV — outer LOSO; inner 5-fold for tuning. • Feature set (PA v1): Weekly aggregates of volume, fragmentation, morning activity, sleep, walking; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly feature bundles to obtain pain probability.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device types, or wear-time compliance.
Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, reliance on consistent wear-time, class imbalance.
Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=65 PsA participants after valid-week QC with valid PSAID pain labels; no/mild pain n=43, moderate/severe pain n=22; window = first 14 days post-baseline. • Features (PA v1): Daily volume, intensity (ENMO/cadence), fragmentation (bout stats & transitions—incl. MVPA→Sed), sleep, morning activity; summarized weekly. Full list provided in **Table 34**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models.

	• Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Wrist tri-axial signals (24/7) segmented into 30-s windows; activity recognition via DCNN HAR—SSL pre-trained on UK Biobank (Yuan et al., 2024) fine-tuned on Capture-24 (Chan et al., 2024) with HMM; parallel walking/steps model—OxWalk (Small et al., 2023)); per-window outputs: class, ENMO, steps, walking flag.
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=65).
Factors	Demographics (sex, age, BMI), disease duration/severity; wear-time compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.64 [0.48, 0.78]; κ 0.31 [0.08, 0.53]; F1 0.56 [0.39, 0.71].

Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; ElasticNet Logistic Regression; PA v1 features; experimental stage.

Fatigue severity from smartwatch accelerometer	
Binary logistic regression (LASSO) model predicting no/mild fatigue (PSAID fatigue ≤4) vs. moderate/severe fatigue (>4) using weekly physical-activity features from smartwatch accelerometer data (Garmin Vivoactive 5).	
Model Details	
Model description	• Architecture: Logistic regression (LASSO); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters: C: 100; penalty: l1; solver: liblinear; max_iter: 5000; class_weight: balanced. • Training setup: Nested CV — outer LOSO; inner 5-fold for tuning. • Feature set (PA v1): Weekly aggregates of volume, fragmentation, morning activity, sleep, walking; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly feature bundles to obtain fatigue probability.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device types, or wear-time compliance.

Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, reliance on consistent wear-time, class imbalance.
Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=65 PsA participants after valid-week QC with valid PSAID fatigue labels; no/mild fatigue n=43, moderate/severe fatigue n=22; window = first 14 days post-baseline. • Features (PA v1): Daily volume, intensity (ENMO/cadence), fragmentation (bout stats & transitions—incl. MVPA→Sed), sleep, morning activity; summarized weekly. Full list provided in **Table 34**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Wrist tri-axial signals (24/7) segmented into 30-s windows; activity recognition via DCNN HAR—SSL pre-trained on UK Biobank (Yuan et al., 2024) fine-tuned on Capture-24 (Chan et al., 2024) with HMM; parallel walking/steps model—OxWalk (Small et al., 2023)); per-window outputs: class, ENMO, steps, walking flag.
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=65).
Factors	Demographics (sex, age, BMI), disease duration/severity; wear-time compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.64 [0.51, 0.77]; κ 0.32 [0.14, 0.50]; F1 0.61 [0.45, 0.75].

Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; L1 Logistic Regression; Smartphone PA v1 features; experimental stage.

Stiffness severity from smartwatch accelerometer	
Binary logistic regression (LASSO) model predicting no/mild stiffness (self-reported stiffness ≤4) vs. moderate/severe self-reported stiffness (>4) using weekly physical-activity features from smartwatch accelerometer data (Garmin Vivoactive 5).	
Model Details	
Model description	• Architecture: Logistic regression (RIDGE); Python 3.11.13, scikit-learn 1.7.2. • Hyperparameters: C: 10; penalty: l2; solver: liblinear; max_iter: 5000; class_weight: balanced. • Training setup: Nested CV — outer LOSO; inner 5-fold for tuning.

	• Feature set (PA v1): Weekly aggregates of volume, fragmentation, morning activity, sleep, walking; covariates: age, sex, disease duration, BMI. • Licensing/Attribution: Not yet published (PoC).
Developed by	Signal Processing and Biomedical Technology Unit team (Aristotle University of Thessaloniki).
Finetuned from	N/A (trained from scratch).
Intended Use	
Direct use	Score weekly feature bundles to obtain stiffness probability.
Downstream use	Integration into the iPROLEPSIS ecosystem.
Out-of-scope use	Not a diagnostic tool without oversight; not validated beyond training devices/populations.
Bias, Risks, and Limitations	
Bias	Possible performance differences across age/sex/BMI, device types, or wear-time compliance.
Risk	False negatives delay attention; false positives prompt unnecessary actions; distribution shift can degrade performance.
Limitation	Small, single-visit cohort (14-day window), no medication adjustment, reliance on consistent wear-time.
Recommendations	Expand/longitudinal data, add medication/context covariates, address imbalance, and calibrate.
Training Details	
Training data	• Data & labels: n=89 PsA participants after valid-week QC with valid self-reported stiffness labels; no/mild stiffness n=48, moderate/severe stiffness n=41; window = first 14 days post-baseline. • Features (PA v1): Daily volume, intensity (ENMO/cadence), fragmentation (bout stats & transitions—incl. MVPA→Sed), sleep, morning activity; summarized weekly. Full list provided in **Table 34**. • Validity & imputation: Week valid if ≥72 h covering all 24 hours; missing minutes imputed by same time-of-day mean (weekday/weekend fallbacks). • Covariates: Sex, age, BMI, disease duration included in all models. • Model selection: Nested CV—outer LOSO for generalization; inner 5-fold subject-level CV for hyperparameter tuning. Final model is retrained on all data with majority-voted hyperparameters.
Preprocessing	Wrist tri-axial signals (24/7) segmented into 30-s windows; activity recognition via DCNN HAR—SSL pre-trained on UK Biobank (Yuan et al., 2024) fine-tuned on Capture-24 (Chan et al., 2024) with HMM; parallel walking/steps model—OxWalk (Small et al., 2023)); per-window outputs: class, ENMO, steps, walking flag.
Evaluation	
Testing data	Outer-loop LOSO (each eligible subject held out once; n=89).
Factors	Demographics (sex, age, BMI), disease duration; wear-time compliance, device/firmware, seasonality; (not modeled v1) medication exposure/changes.
Metrics	Metrics: AUROC, F1, Cohen's κ at a threshold chosen via inner CV; report point estimate with 95% CI.
Results	AUROC 0.54 [0.41, 0.66]; κ 0.19 [0.01, 0.39]; F1 0.61 [0.46, 0.73].

Versions		
Version No.	**Release date**	**Description**
v1	1-Sep-2025	Proof-of-concept; LOSO training; L2 Logistic Regression; Smartphone PA v1 features; experimental stage.

Appendix II Information on the SmartBelt device

The EGG measuring band, originally developed in i-PROGNOSIS project (included in the European Union's Horizon 2020 research and innovation programme), is composed by an 8-Channel biosignalsplux Hub, which is a specialized and highly reliable physiological signal acquisition device designed by PLUX Wireless Biosignals.

The measuring band, which is the structural element of the device, incorporates several different types of sensors to ensure a deeper understanding of gastrointestinal events in a non-invasive way. The core technology and functioning of these sensors remains consistent across different applications of the device, with their positioning influencing the location and tissues or organs that are being studied but not the type of signal being studied.

Tthe configuration includes a bipolar electrogastrographic (EGG) sensor and 4 acoustic (ACU) transducers. These sensors acquire the relevant EGG signals, detect patient movements that might introduce noise, and record bowel sounds, respectively.

In this section, a set of technical specifications regarding the EGG measuring band components will be presented.

II.1. biosignalsplux | 8-Channel Hub

biosignalsplux is wireless physiological signal acquisition device that supports multi-channel data acquisitions in a totally synchronized way, ensuring an effective reliability through the high-resolution channels that it contains.

Figure 79 biosignalsplux hub highlighting each type of ports through the colour code: **a)** standard analogical ports; **b)** optional analogical ports; **c)** ground/reference electrode port; **d)** synchronization/digital port and **e)** power button.

As demonstrated on **Figure 79**, **biosignalsplux** is formed by analogical and digital ports. The available 8 analog channels offer freedom to the researcher, but the standard capabilities of a **biosignalsplux** hub can be expanded, using the available synchronization port which enables a multi-device acquisition with up to 3 hubs, reaching a total of 24 real-time data channels.

Table 58 List of biosignalsplux ports and respective meaning.

Available Ports	4	Generic Analogical Channels
	4	Optional Analogical Channels
	1	Digital I/O Port [Synchronization Port]
	1	Ground Port

The previous list establishes the observable electronical architecture from the user point of view, however, **biosignalsplux** also supports some powerful functionalities dependent not only on the hardware specifications but also on the firmware that controls it and the software responsible for managing the data acquisition process.

Table 59 Highlighting some of the main functionalities of the biosignalsplux hub.

Functionalities and Notable Specifications		
Scheduling Functionality	16 Gb	Internal Memory [ensuring the storage of up to 111h of data]
Battery Testing Thresholds	10h	Streaming
	24h	Logging
Analog to Digital Conversion Resolution	8-16 bits	Capable of distinguish analogical events that differ from each other 0.02 mV
Supported Sampling Rates	4000 Hz	Multi-channel data acquisition up to 4000 Hz sampling rates

II.2 Electrogastrography (EGG) Sensor

The EGG sensors use a standard bipolar configuration, which improves the quality of the acquired physiological data, increasing the signal-to-noise ratio (SNR), due to a differential estimation of changes in electrical voltage with origin in the gastric walls and that spreads along the skin surface. It has a sensor gain of 6114, common mode rejection ratio (CMRR) of 100 dB and an energy consumption of 3 mA.

Sensor and Technical Specifications

Each EGG sensor is composed of two exploratory electrodes (red and black cables) and a third reference electrode (cable in the centre), as demonstrated in **Figure 80**. Each exploratory electrode will measure the electric potential of the placement site in relation to the conditions on the reference electrode (which should be positioned carefully to avoid having an unstable electric tension as reference, due to unbalanced collection of signal components). The difference of values between the two exploratory electrodes will then be amplified giving origin to our bipolar signal.

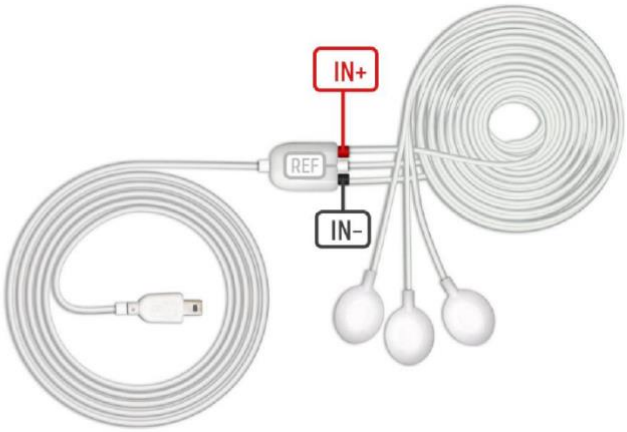


Figure 80 EGG sensor highlighting the two exploratory electrodes (red and black cables) and the respective reference electrode (cable).

Characteristics of the EGG signal

The EGG signal can be classified as a "slow" signal, taking into consideration that almost the entire informational content of EGG is contained inside a strict range of small frequency components: [0.05 - 0.20] Hz [15]–[17] or [0.0083 - 0.15] Hz [3].

Regarding the signal amplitude, it may present non-amplified values between 50-500 uV [18], however, due to the high gain of PLUX EGG sensor and the baseline removal, the normal range of acquired values will be located between -0.27 mV and 0.27 mV.

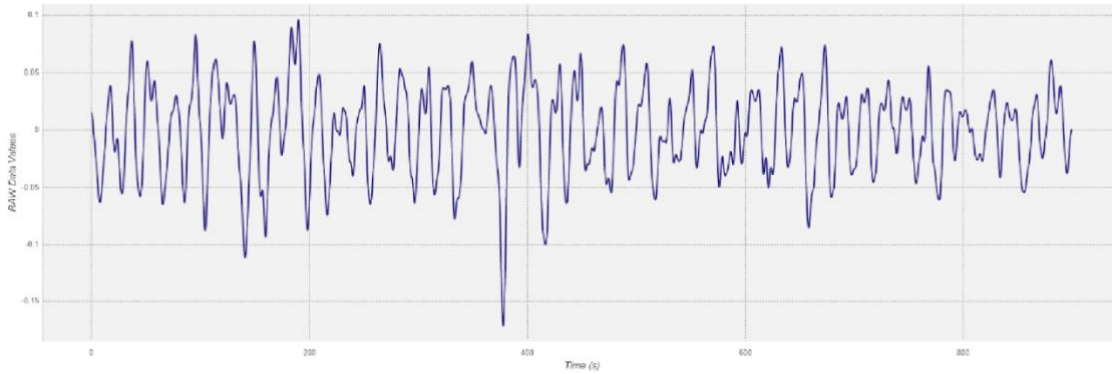


Figure 81 EGG signal (collected with the EGG measuring band) where the slow wave morphology is identifiable with long-term variations.

II.3. Acoustic sensor

With the Acoustic (ACU) sensor, the EGG measuring band is empowered with the possibility of sound collection in a wide frequency range. In terms of the EGG measuring band focus, the device will be able to acquire the sounds generated due to gastric motility related events, i.e., bowel sounds.

By using this sensor and through signal processing methodologies some future correlations may be established between the data collected from this sensor and the EGG sensors.

Sensor and Technical Specifications

ACU sensor is formed by a microphone encapsulated in a white box with a small orifice in the middle to provide a channel for sound propagation, reaching the internal sensing area (**Figure 82**). It has a sensor gain of 410, a working voltage range of 0-3V and a sound detection bandwidth of 29-1591 Hz.

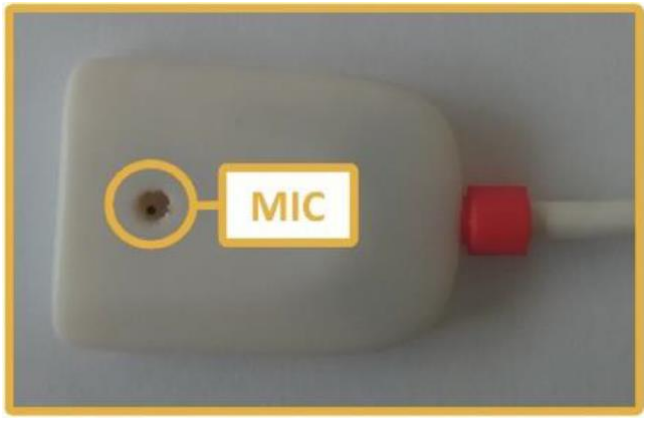


Figure 82 ACU sensor presenting the small orifice on the middle of the encapsulation where the microphone (MIC) is located.

II.4. Instructions for Use

Before each use, it should be ensured that the device is fully charged.

1. Electrode placement on band:

- Place the electrodes on the elastic EGG measuring band.
- Ensure electrodes are securely attached to the band.

2. Band placement:

- Carefully place the elastic band with the attached electrodes around the patient's abdomen
- Adjust and fasten the band to ensure a snug but comfortable fit.
- Verify that all electrodes make good contact with the patient's skin.
- Check that all connections to the devices are secure.

II.5. Training required for investigational device utilization

A brief initial demonstration by a trained professional should be sufficient for proper application and use of the device. Training should cover proper placement of the band, operation of the recording equipment, and guidelines for patient instruction if necessary.

II.6. Purpose

The primary purpose of this device is to monitor gastrointestinal motility through EGG and assess the normality of gastrointestinal myoelectrical activity. Through these direct measurements, the device will contribute valuable data that can assist healthcare providers in evaluating the efficacy of therapeutic interventions for Crohn's disease, monitoring disease progression over time, and potentially improving personalized patient care.

Appendix III Supplementary material on typing dynamics

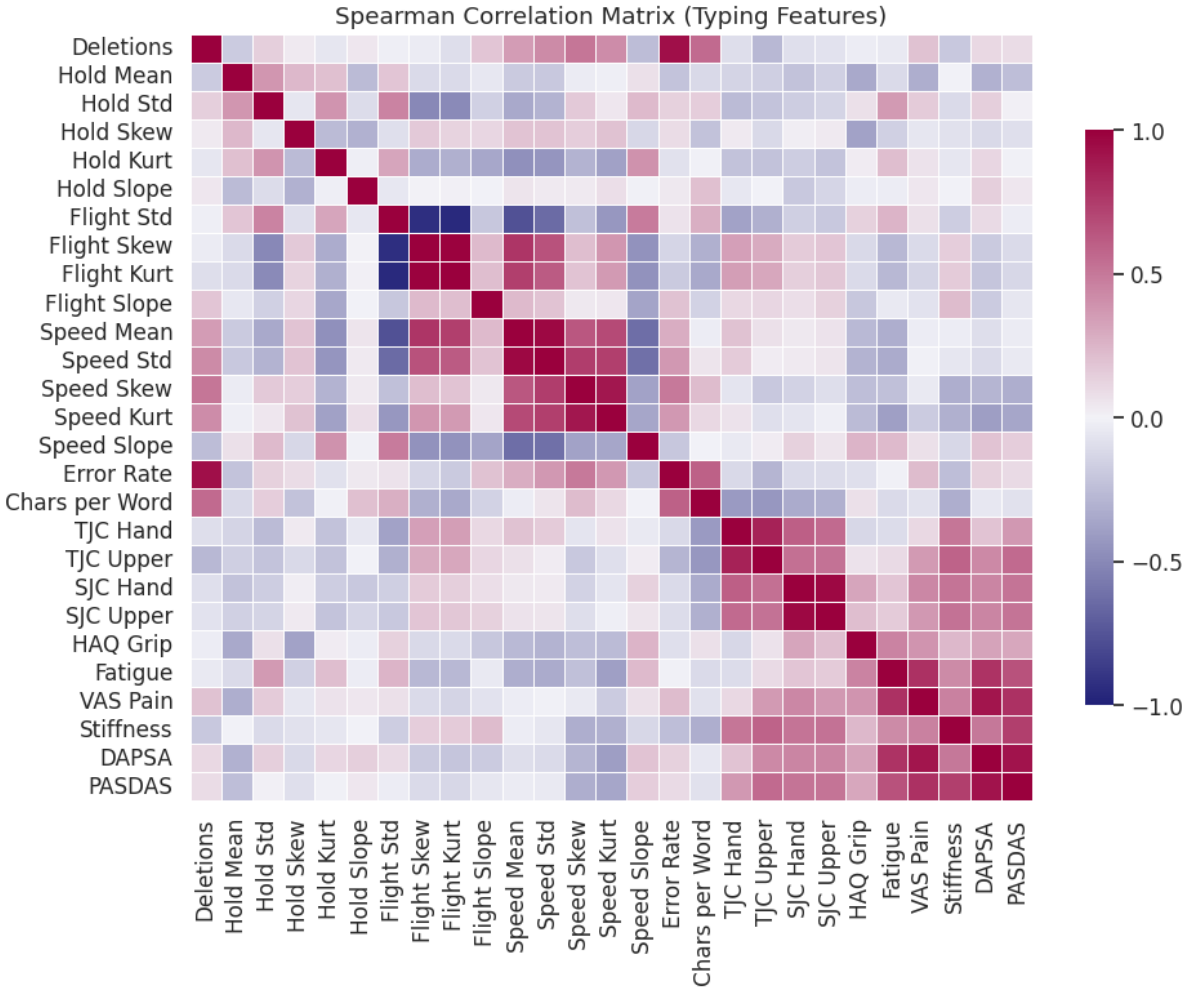


Figure 83 Spearman correlation matrix between typing-derived features and clinical or PRO outcomes. Asterisks (*) indicate statistically significant associations ($p < 0.05$).

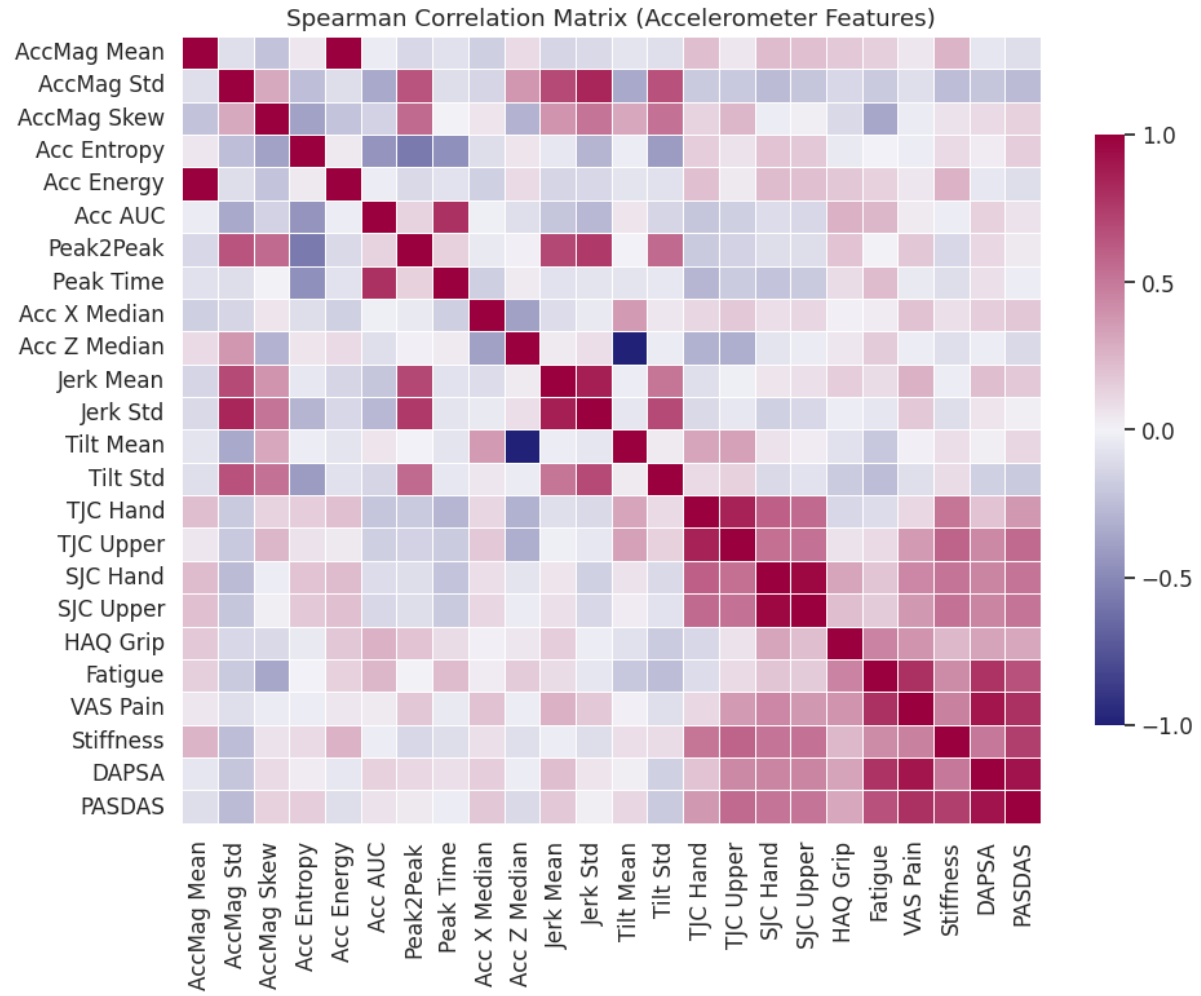


Figure 84 Spearman correlation matrix between accelerometer-derived features and clinical or PRO outcomes.

Table 60 Univariate ROC analysis results for typing and accelerometer features in relation to TJC Hand status. The table reports AUC values along with 95% confidence intervals for each feature.

Typing features	AUC [95%CI]	Accel/meter features	AUC [95%CI]
Flight Kurt	0.719 [0.559, 0.856]	Tilt Mean	0.713 [0.540, 0.862]
Flight Skew	0.713 [0.558, 0.865]	AccMag Mean	0.630 [0.462, 0.791]
Speed Mean	0.639 [0.472, 0.788]	Acc Energy	0.626 [0.442, 0.791]
Speed Std	0.614 [0.458, 0.772]	Acc Entropy	0.619 [0.438, 0.779]
Speed Kurt	0.567 [0.394, 0.725]	AccMag Skew	0.583 [0.417, 0.755]
Flight Slope	0.546 [0.371, 0.730]	Tilt Std	0.534 [0.349, 0.716]
Hold Skew	0.522 [0.348, 0.692]	Acc X Median	0.520 [0.344, 0.690]

Appendix IV Supplementary material on physical activity and gait from smartphone sensors

Table 61 Group comparisons of smartphone-derived physical-activity features by HAQ Walk difficulty (0 = no difficulty vs. >0 = any difficulty).

Feature	HAQ Walk = 0 (n=48) (Median [IQR])	HAQ Walk >0 (n=24) (Median [IQR])	p-value
Activity volume			
Steps	3565 [1936, 4888]	2989 [1679, 4441]	
Min Steps	352 [152, 1168]	500 [200, 821]	
Max Steps	8766 [5031, 11864]	7082 [4455, 8370]	
Walking (min)	105 [72, 153]	115 [64, 169]	
Min Walking (min)	28 [16, 62]	36 [15, 48]	
Max Walking (min)	200 [150, 272]	238 [137, 292]	
Mean ENMO (mg)	11 [8, 14]	11 [9, 15]	
Inactive (min)	1335 [1287, 1368]	1325 [1271, 1376]	
Light (min)	91 [61, 143]	109 [62, 165]	
MVPA (min)	8 [3, 15]	4 [2, 5]	*
Activity fragmentation			
Avg Inactive (min)	40 [30, 56]	34 [27, 53]	
Avg Light (min)	2.28 [1.99, 2.75]	2.08 [1.9, 2.32]	
Avg MVPA (min)	1.54 [1.01, 2.23]	1.07 [0.77, 1.37]	*
Max Inactive (min)	391 [350, 440]	395 [354, 426]	
Max Light (min)	10 [8, 16]	10 [7, 13]	
Max MVPA (min)	3 [1, 6]	2 [1, 2]	*
Inactive → Light	0.03 [0.02, 0.04]	0.03 [0.02, 0.05]	
Light → Inactive	0.47 [0.37, 0.54]	0.49 [0.44, 0.54]	
Inactive bouts/hr	55 [53, 57]	55 [53, 57]	
Walk bouts (≥10min)	2 [1, 3]	1 [1, 2]	
Mobility			
Top 1 Cadence	111 [97, 119]	108 [98, 11]	
Top 30 Cadence	66 [43, 82]	52 [37, 61]	
95th Cadence	93 [7, 107]	75 [62, 85]	*
*Mann-Whitney U test: *p-value<0.05 **p-value<0.001*			

Table 62 Group comparisons of smartphone-derived physical-activity features by lower-extremity Tender Joint Count (TJC Lower) (0 = none vs. >0 = any tender joints).

Feature	TJC Lower = 0 (n=50) (Median [IQR])	TJC Lower > 0 (n=45) (Median [IQR])	p-value
Activity volume			
Steps	3213 [2382, 4856]	2194 [1387, 4045]	*
Min Steps	680 [270, 1281]	339 [124, 877]	*
Max Steps	8883 [5220, 11954]	6410 [3574, 8079]	*
Walking (min)	108 [81, 162]	85 [58, 124]	
Min Walking (min)	36 [23, 66]	24 [12, 45]	*
Max Walking (min)	223 [155, 280]	178 [116, 267]	
Mean ENMO (mg)	11 [9.113, 14]	9.239 [7.211, 13]	
Inactive (min)	1332 [1278, 1359]	1355 [1316, 1382]	
Light (min)	98 [70, 148]	78 [56, 121]	
MVPA (min)	5.929 [3.214, 13]	3.786 [1.357, 8.429]	*
Activity fragmentation			
Avg Inactive (min)	38 [29, 47]	40 [28, 62]	
Avg Light (min)	2.301 [2.03, 2.921]	1.996 [1.838, 2.407]	*
Avg MVPA (min)	1.485 [0.979, 2.122]	1.177 [0.625, 1.738]	
Max Inactive (min)	400 [342, 441]	404 [376, 468]	
Max Light (min)	11 [8.125, 17]	9 [6.857, 12]	*
Max MVPA (min)	2.357 [1.32, 4.857]	1.786 [0.786, 3.214]	
Inactive → Light	0.03 [0.024, 0.038]	0.029 [0.017, 0.04]	
Light → Inactive	0.466 [0.366, 0.508]	0.525 [0.459, 0.602]	**
Inactive bouts/hr	56 [53, 57]	56 [55, 58]	
Walk bouts (≥10min)	1.607 [1.089, 3.268]	1.071 [0.5, 1.714]	*
Mobility			
Top 1 Cadence	101 [93, 114]	97 [83, 107]	*
Top 30 Cadence	63 [53, 77]	52 [38, 70]	*
95th Cadence	85 [71, 97]	73 [59, 89]	*
*Mann-Whitney U test: *p-value<0.05 **p-value<0.001*			

Table 63 Group comparisons of smartphone-derived physical-activity features by physician-reported flare status.

	Flare (physician)		
Feature	**No (n=72) (Median [IQR])**	**Yes (n=20) (Median [IQR])**	**p-value**
Activity volume			
Steps	3117 [1867, 4548]	2560 [1792, 3882]	
Min Steps	498 [204, 1188]	341 [71, 834]	
Max Steps	7630 [4699, 10020]	6204 [3928, 9293]	
Walking (min)	102 [70, 158]	103 [55, 155]	
Min Walking (min)	29 [16, 58]	32 [9, 55]	
Max Walking (min)	190 [137, 282]	214 [134, 267]	
Mean ENMO (mg)	11 [8.646, 14]	11 [8.283, 13]	
Inactive (min)	1338 [1282, 1370]	1337 [1285, 1385]	
Light (min)	86 [62, 147]	96 [51, 151]	
MVPA (min)	5.25 [3.075, 13]	3.973 [1.179, 9.268]	
Activity fragmentation			
Avg Inactive (min)	39 [29, 56]	33 [26, 73]	
Avg Light (min)	2.175 [1.916, 2.685]	2.092 [1.979, 2.475]	
Avg MVPA (min)	1.424 [0.919, 2.132]	0.947 [0.501, 1.512]	*
Max Inactive (min)	395 [346, 450]	416 [387, 459]	
Max Light (min)	9.451 [7.143, 15]	9.679 [7.768, 14]	
Max MVPA (min)	2.143 [1.345, 4.212]	1.411 [0.675, 2.929]	
Inactive → Light	0.029 [0.021, 0.038]	0.033 [0.017, 0.045]	
Light → Inactive	0.481 [0.394, 0.537]	0.511 [0.458, 0.573]	
Inactive bouts/hr	56 [53, 57]	56 [54, 58]	
Walk bouts (≥10min)	1.414 [0.786, 2.357]	1.286 [0.589, 2.071]	
Mobility			
Top 1 Cadence	101 [90, 112]	100 [81, 105]	
Top 30 Cadence	60 [47, 75]	56 [38, 73]	
95th Cadence	81 [65, 96]	78 [60, 88]	
*Mann-Whitney U test: *p-value<0.05 **p-value<0.001*			

Table 64 Group comparisons of smartphone-derived physical-activity features by patient-reported flare status.

	Flare (patient)		
Feature	**No (n=56) (Median [IQR])**	**Yes (n=12) (Median [IQR])**	**p-value**
Activity volume			
Steps	2999 [1847, 4376]	1859 [834, 2866]	*
Min Steps	509 [212, 1168]	341 [32, 792]	
Max Steps	7126 [4699, 9580]	4868 [2055, 6100]	*
Walking (min)	100 [69, 151]	86 [39, 129]	
Min Walking (min)	30 [20, 58]	34 [5.25, 52]	
Max Walking (min)	190 [141, 270]	156 [100, 263]	
Mean ENMO (mg)	11 [7.71, 14]	9.442 [7.159, 11]	
Inactive (min)	1340 [1289, 1371]	1354 [1311, 1401]	
Light (min)	85 [61, 134]	83 [32, 127]	
MVPA (min)	5.179 [3.186, 10]	1.75 [0.125, 3.674]	*
Activity fragmentation			
Avg Inactive (min)	39 [29, 57]	36 [29, 104]	
Avg Light (min)	2.12 [1.935, 2.431]	2.092 [1.808, 2.486]	
Avg MVPA (min)	1.424 [0.963, 2.021]	0.622 [0.071, 0.953]	*
Max Inactive (min)	400 [341, 456]	421 [380, 481]	
Max Light (min)	9.286 [7.268, 14]	9.571 [4.429, 14]	
Max MVPA (min)	2.143 [1.413, 3.786]	0.786 [0.071, 1.384]	*
Inactive → Light	0.029 [0.02, 0.038]	0.031 [0.015, 0.04]	
Light → Inactive	0.482 [0.406, 0.535]	0.529 [0.498, 0.678]	*
Inactive bouts/hr	56 [54, 57]	56 [54, 58]	
Walk bouts (≥10min)	1.286 [0.893, 2.304]	0.786 [0.179, 1.679]	
Mobility			
Top 1 Cadence	102 [90, 112]	89 [80, 100]	*
Top 30 Cadence	61 [47, 74]	44 [34, 57]	*
95th Cadence	82 [67, 96]	66 [53, 77]	*
*Mann-Whitney U test: *p-value<0.05 **p-value<0.001*			

Table 65 Group comparisons of smartphone-derived physical-activity features by MDA status (achieved vs. not achieved).

	MDA (achieved)		
Feature	**Yes (n=29) (Median [IQR])**	**No (n=29) (Median [IQR])**	**p-value**
Activity volume			
Steps	3163 [1757, 4546]	3071 [2208, 4851]	
Min Steps	525 [281, 914]	403 [141, 1194]	
Max Steps	7714 [4529, 9393]	7837 [5050, 12371]	
Walking (min)	106 [66, 143]	105 [76, 152]	
Min Walking (min)	32 [23, 53]	31 [16, 61]	
Max Walking (min)	184 [127, 267]	216 [155, 270]	
Mean ENMO (mg)	11 [7.211, 14]	11 [8.461, 14]	
Inactive (min)	1334 [1297, 1374]	1335 [1288, 1364]	
Light (min)	89 [63, 129]	93 [61, 141]	
MVPA (min)	3.929 [2.643, 13]	6.5 [3.429, 17]	
Activity fragmentation			
Avg Inactive (min)	36 [28, 53]	41 [31, 56]	
Avg Light (min)	2.045 [1.86, 2.281]	2.372 [2.028, 2.85]	*
Avg MVPA (min)	1.328 [0.843, 2.046]	1.518 [1.051, 2.134]	
Max Inactive (min)	388 [339, 423]	396 [347, 439]	
Max Light (min)	9.286 [7.143, 12]	11 [7.643, 17]	
Max MVPA (min)	2 [1.214, 3.714]	2.5 [1.429, 4.846]	
Inactive → Light	0.03 [0.022, 0.04]	0.028 [0.02, 0.036]	
Light → Inactive	0.508 [0.442, 0.539]	0.476 [0.389, 0.525]	
Inactive bouts/hr	56 [54, 57]	56 [54, 56]	
Walk bouts (≥10min)	794 [609, 1149]	1038 [824, 1522]	
Mobility			
Top 1 Cadence	102 [90, 112]	100 [88, 109]	
Top 30 Cadence	57 [40, 75]	64 [52, 76]	
95th Cadence	83 [63, 95]	81 [73, 93]	
*Mann-Whitney U test: *p-value<0.05 **p-value<0.001*			

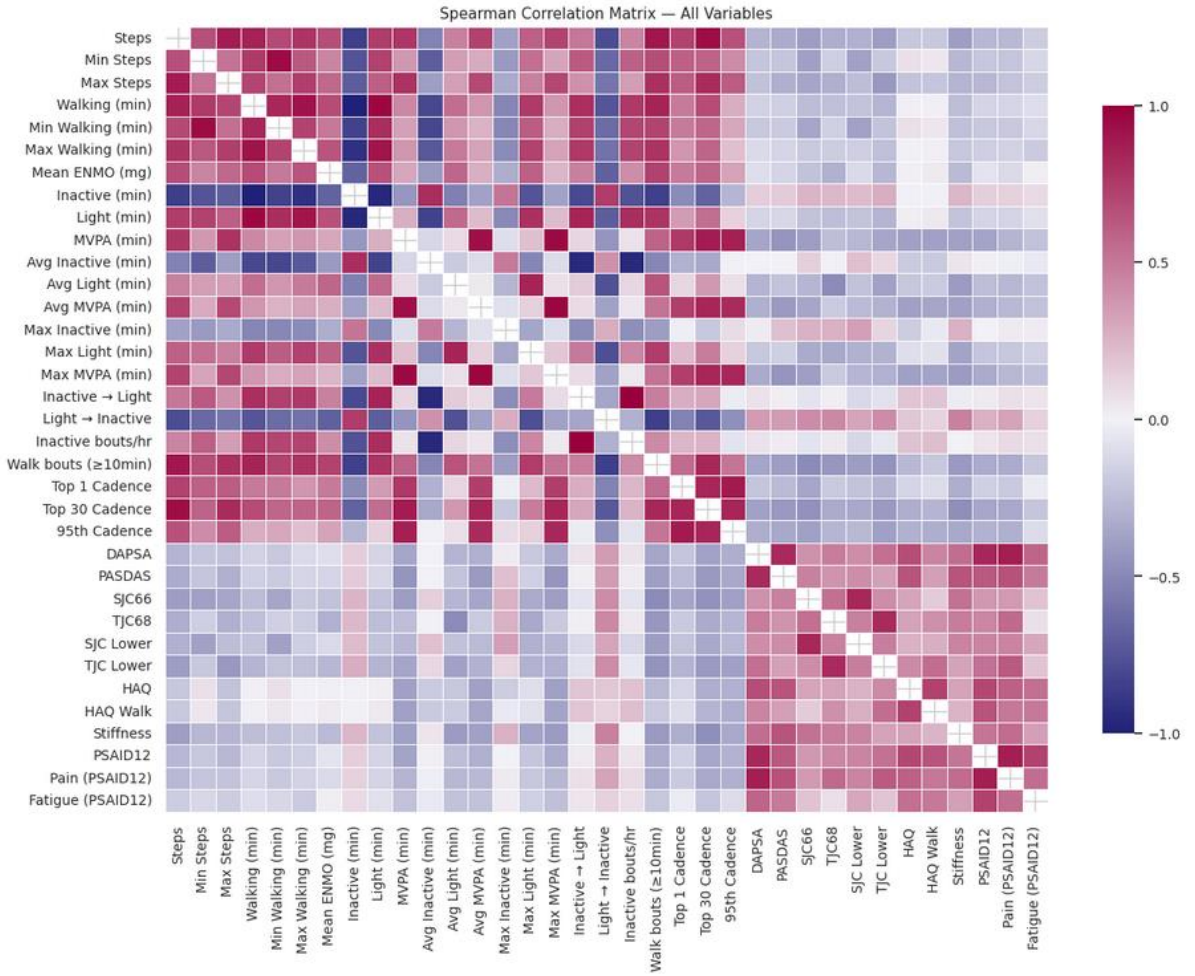


Figure 85 Full Spearman correlation matrix: smartphone-derived physical-activity features vs. clinical/PRO outcomes.

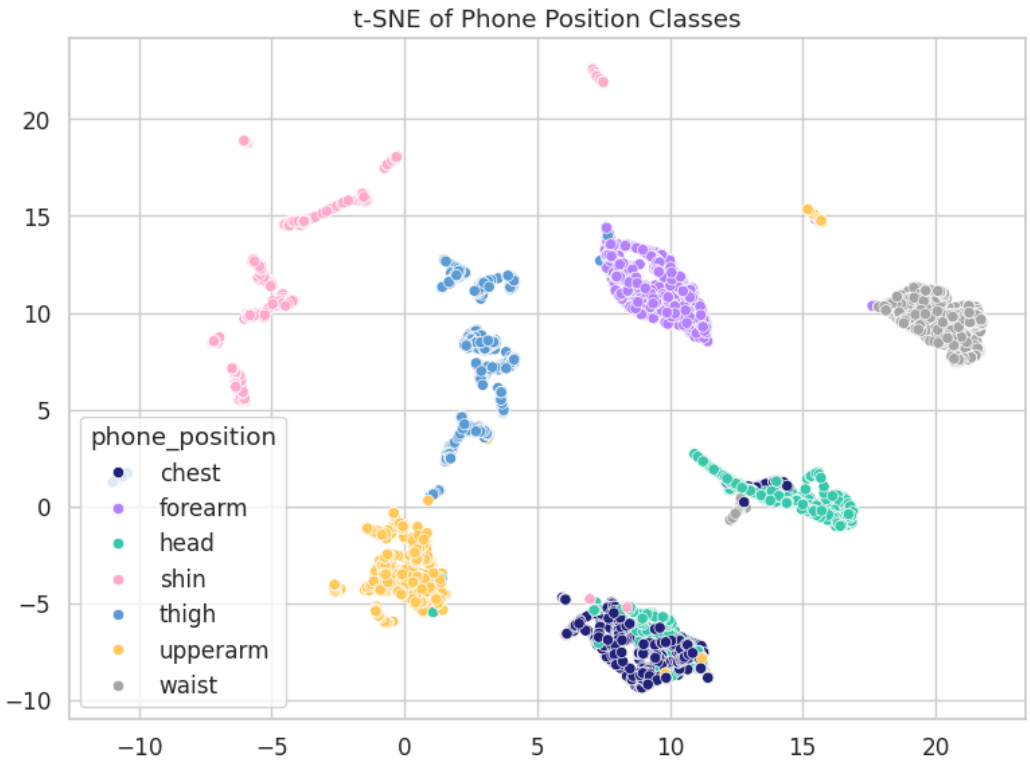


Figure 86 Two-dimensional t-SNE embedding of 10-s accelerometer/gyroscope windows used for on-body device localization.

Table 66 5-fold (subject-level) cross-validation results of a Random Forest on-body device-localization classifier trained on the RealWorld dataset.

Label	Precision	Recall	F1-Score	Support
Chest	0.85	0.79	0.82	921
Forearm	0.92	0.99	0.96	921
Head	0.76	0.86	0.8	921
Shin	1	0.93	0.96	921
Thigh	0.93	0.91	0.92	921
Upperarm	0.92	0.92	0.92	921
Waist	0.98	0.93	0.95	921
Accuracy			0.9	6447
Macro avg	0.91	0.9	0.9	6447
Weighted avg	0.91	0.9	0.9	6447

Table 67 Group comparisons of smartphone-derived gait features by HAQ Walk difficulty (0 = no difficulty vs. >0 = any difficulty).

Feature	HAQ Walk = 0 (n=13) (Median [IQR])	HAQ Walk >0 (n=10) (Median [IQR])	p-value
Step Reg	0.51 [0.49, 0.65]	0.57 [0.51, 0.63]	
Stride Reg	0.68 [0.63, 0.71]	0.62 [0.55, 0.65]	
Gait Symm	0.05 [0.03, 0.12]	0.05 [0.05, 0.10]	
Step Med	0.54 [0.52, 0.57]	0.55 [0.53, 0.58]	
Step MAD	0.03 [0.02, 0.03]	0.03 [0.02, 0.03]	
Stride Med	1.10 [1.04, 1.13]	1.11 [1.07, 1.16]	
Stride MAD	0.02 [0.02, 0.02]	0.02 [0.02, 0.03]	
Cadence	11 [106, 116]	108 [103, 113]	
*Mann-Whitney U test: *p-value<0.05 **p-value<0.001*			

Table 68 Group comparisons of smartphone-derived gait features by lower-extremity Tender Joint Count (TJC Lower) (0 = none vs. >0 = any tender joints).

Feature	TJC Lower = 0 (n=16) (Median [IQR])	TJC Lower > 0 (n=13) (Median [IQR])	p-value
Step Reg	0.52 [0.47, 0.62]	0.63 [0.55, 0.67]	
Stride Reg	0.63 [0.54, 0.68]	0.65 [0.59, 0.71]	
Gait Symm	0.07 [0.04, 0.11]	0.05 [0.04, 0.07]	
Step Med	0.54 [0.52, 0.56]	0.56 [0.53, 0.57]	
Step MAD	0.03 [0.02, 0.04]	0.02 [0.02, 0.03]	*
Stride Med	1.09 [1.04, 1.12]	1.12 [1.07, 1.14]	
Stride MAD	0.02 [0.02, 0.03]	0.02 [0.02, 0.03]	
Cadence	111 [107, 117]	107 [106, 112]	
*Mann-Whitney U test: *p-value<0.05 **p-value<0.001*			

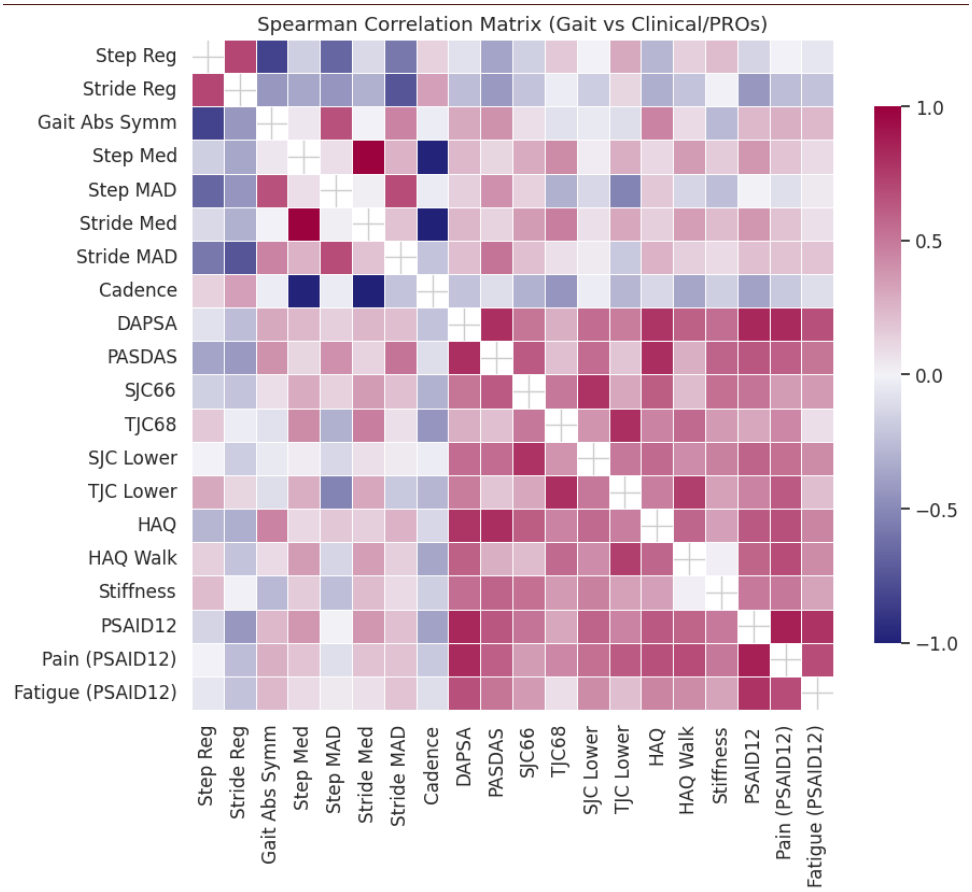


Figure 87 Full Spearman correlation matrix: smartphone-derived gait features vs. clinical/PRO outcomes.

Appendix V Supplementary material on physical activity from smartwatch

Table 69 Group comparisons of smartwatch-derived features by physician-reported flare status.

	Flare (physician)		
Feature	**No (n=70) (Median [IQR])**	**Yes (n=16) (Median [IQR])**	**p-value**
Activity volume			
Mean ENMO (mg)	22 [18, 25]	22 [19, 23]	
Sleep (min)	823 [712, 982]	897 [766, 1042]	
Sedentary (min)	518 [432, 601]	538 [496, 595]	
Light (min)	447 [363, 592]	404 [371, 539]	
MVPA (min)	48 [26, 73]	33 [28, 57]	
Steps	8226 [6145, 11277]	7709 [5349, 9032]	
Walking (min)	355 [279, 414]	320 [259, 371]	
Activity fragmentation			
Avg Sleep (min)	5.904 [4.837, 7.348]	7.185 [4.677, 8.666]	
Max Sleep (min)	212 [149, 271]	273 [189, 326]	
Avg Sedentary (min)	2.603 [2.245, 3.121]	2.976 [2.257, 3.246]	
Max Sedentary (min)	24 [18, 32]	25 [21, 34]	
Avg Light (min)	2.773 [2.341, 3.314]	2.758 [2.142, 3.433]	
Max Light (min)	26 [20, 34]	23 [20, 27]	
Avg MVPA (min)	1.047 [0.939, 1.276]	0.937 [0.843, 1.212]	
Max MVPA (min)	4.607 [3.116, 6.562]	2.973 [1.911, 6.652]	
Pr Sedentary→MVPA	0.008 [0.005, 0.012]	0.008 [0.007, 0.01]	
Pr MVPA→Sedentary	0.128 [0.093, 0.164]	0.177 [0.09, 0.224]	
Pr Sedentary→Light	0.09 [0.066, 0.107]	0.085 [0.072, 0.103]	
Pr Light→Sedentary	0.102 [0.085, 0.125]	0.123 [0.087, 0.159]	
Sedentary bouts/hr	22 [18, 25]	22 [21, 25]	
Morning activity			
Morn. ENMO 30min	29 [24, 38]	30 [26, 36]	
Morn. ENMO 60min	28 [22, 39]	29 [19, 37]	
Morn. ENMO 120min	29 [23, 36]	29 [19, 34]	
Sleep			
Sleep duration (h)	8.012 [6.99, 9.907]	8.438 [7.913, 11]	
Rest efficiency	1.246 [1.053, 1.384]	1.247 [1.056, 1.502]	
#Sleep episodes	19 [9.942, 37]	17 [11, 38]	
Rest fragmentation	2.542 [1.572, 5.776]	2.238 [1.362, 4.592]	

Feature	Flare (physician)		p-value
	No (n=70) (Median [IQR])	Yes (n=16) (Median [IQR])	
Sleep movements (#)	18 [9.154, 36]	16 [10, 37]	
Sleep movements/hr	2.542 [1.572, 5.776]	2.238 [1.362, 4.592]	
Walking			
Walk bouts (≥10min)	9.571 [7.286, 12]	9 [6.446, 10]	
Top 1 cadence	106 [98, 112]	101 [96, 110]	
Top 30 cadence	80 [69, 90]	76 [66, 90]	
95th cadence	76 [68, 90]	71 [62, 86]	
*Mann-Whitney U test: *p-value<0.05 **p-value<0.001*			

Table 70 Group comparisons of smartwatch-derived features by patient-reported flare status.

Feature	Flare (patient)		p-value
	No (n=54) (Median [IQR])	Yes (n=10) (Median [IQR])	
Activity volume			
Mean ENMO (mg)	22 [19, 25]	19 [15, 23]	
Sleep (min)	801 [692, 968]	859 [694, 965]	
Sedentary (min)	508 [411, 576]	507 [418, 576]	
Light (min)	459 [393, 600]	386 [355, 542]	
MVPA (min)	48 [28, 63]	28 [26, 31]	
Steps	8257 [6437, 11470]	6174 [4802, 7728]	*
Walking (min)	361 [287, 431]	303 [248, 377]	
Activity fragmentation			
Avg Sleep (min)	5.844 [4.465, 7.411]	5.064 [3.804, 6.698]	
Max Sleep (min)	189 [125, 261]	202 [165, 263]	
Avg Sedentary (min)	2.601 [2.148, 3.09]	2.383 [1.903, 2.81]	
Max Sedentary (min)	24 [18, 32]	23 [17, 25]	
Avg Light (min)	2.873 [2.579, 3.417]	2.345 [1.857, 3.258]	
Max Light (min)	28 [21, 34]	24 [16, 27]	
Avg MVPA (min)	1.042 [0.925, 1.265]	0.875 [0.727, 0.919]	*
Max MVPA (min)	4.391 [3.116, 6.071]	2.161 [1.696, 2.911]	*
Pr Sedentary→MVPA	0.007 [0.004, 0.011]	0.009 [0.007, 0.011]	
Pr MVPA→Sedentary	0.127 [0.082, 0.164]	0.206 [0.172, 0.25]	*
Pr Sedentary→Light	0.092 [0.073, 0.106]	0.092 [0.082, 0.116]	
Pr Light→Sedentary	0.099 [0.082, 0.119]	0.137 [0.1, 0.165]	
Sedentary bouts/hr	21 [17, 24]	21 [17, 24]	

	Flare (patient)		
Feature	**No (n=54) (Median [IQR])**	**Yes (n=10) (Median [IQR])**	**p-value**
Morning activity			
Morn. ENMO 30min	30 [24, 38]	28 [19, 33]	
Morn. ENMO 60min	31 [23, 39]	27 [17, 32]	
Morn. ENMO 120min	30 [24, 37]	27 [19, 31]	
Sleep			
Sleep duration (h)	7.646 [6.897, 9.789]	7.976 [7.208, 8.271]	
Rest efficiency	1.227 [1.029, 1.354]	1.14 [0.999, 1.391]	
#Sleep episodes	21 [13, 37]	25 [15, 48]	
Rest fragmentation	2.805 [1.692, 5.897]	3.582 [2.204, 6.634]	
Sleep movements (#)	20 [12, 36]	24 [14, 47]	
Sleep movements/hr	2.805 [1.692, 5.897]	3.582 [2.204, 6.634]	
Walking			
Walk bouts (≥10min)	10 [7.411, 13]	8.071 [5.375, 9.929]	
Top 1 cadence	105 [99, 112]	97 [84, 102]	*
Top 30 cadence	81 [70, 90]	68 [54, 77]	*
95th cadence	76 [68, 88]	64 [49, 72]	*
*Mann-Whitney U test: *p-value<0.05 **p-value<0.001*			

Table 71 Group comparisons of smartwatch-derived features by MDA status (achieved vs. not achieved).

	MDA (achieved)		
Feature	**Yes (n=25) (Median [IQR])**	**No (n=28) (Median [IQR])**	**p-value**
Activity volume			
Mean ENMO (mg)	22 [19, 24]	22 [18, 25]	
Sleep (min)	728 [655, 924]	866 [737, 975]	
Sedentary (min)	545 [429, 598]	452 [369, 529]	
Light (min)	416 [374, 548]	490 [435, 602]	
MVPA (min)	33 [27, 61]	43 [27, 62]	
Steps	7718 [6485, 11250]	9407 [7780, 10435]	
Walking (min)	348 [290, 372]	366 [342, 414]	
Activity fragmentation			
Avg Sleep (min)	5.512 [4.106, 6.371]	5.973 [5.087, 6.792]	
Max Sleep (min)	186 [82, 257]	232 [164, 316]	*
Avg Sedentary (min)	2.532 [2.081, 2.93]	2.521 [1.928, 2.893]	
Max Sedentary (min)	22 [18, 31]	24 [16, 28]	

	MDA (achieved)		
Feature	**Yes (n=25) (Median [IQR])**	**No (n=28) (Median [IQR])**	**p-value**
Avg Light (min)	2.731 [2.242, 3.185]	2.932 [2.615, 3.547]	
Max Light (min)	25 [19, 33]	30 [24, 37]	
Avg MVPA (min)	0.947 [0.819, 1.206]	1.064 [0.955, 1.256]	
Max MVPA (min)	3.189 [2.42, 5.795]	4.75 [3.071, 6.393]	
Pr Sedentary→MVPA	0.009 [0.006, 0.013]	0.007 [0.005, 0.01]	
Pr MVPA→Sedentary	0.15 [0.117, 0.22]	0.105 [0.078, 0.13]	*
Pr Sedentary→Light	0.089 [0.063, 0.11]	0.099 [0.075, 0.11]	
Pr Light→Sedentary	0.109 [0.086, 0.126]	0.1 [0.075, 0.108]	
Sedentary bouts/hr	23 [18, 25]	19 [15, 22]	*
Morning activity			
Morn. ENMO 30min	32 [26, 38]	24 [24, 35]	
Morn. ENMO 60min	30 [25, 40]	26 [21, 35]	
Morn. ENMO 120min	31 [27, 43]	27 [22, 36]	
Sleep			
Sleep duration (h)	7.848 [6.577, 8.742]	8.537 [7.406, 10]	
Rest efficiency	1.205 [0.981, 1.308]	1.224 [1.071, 1.454]	
#Sleep episodes	30 [13, 60]	19 [10, 30]	
Rest fragmentation	4.264 [2.088, 8.272]	2.379 [1.633, 4.237]	
Sleep movements (#)	29 [12, 59]	18 [9.231, 30]	
Sleep movements/hr	4.264 [2.088, 8.272]	2.379 [1.633, 4.237]	
Walking			
Walk bouts (≥10min)	9.429 [7.351, 12]	10 [7.929, 12]	
Top 1 cadence	104 [98, 112]	106 [100, 113]	
Top 30 cadence	78 [69, 90]	80 [73, 94]	
95th cadence	75 [66, 88]	79 [71, 92]	
*Mann-Whitney U test: *p-value<0.05 **p-value<0.001*			

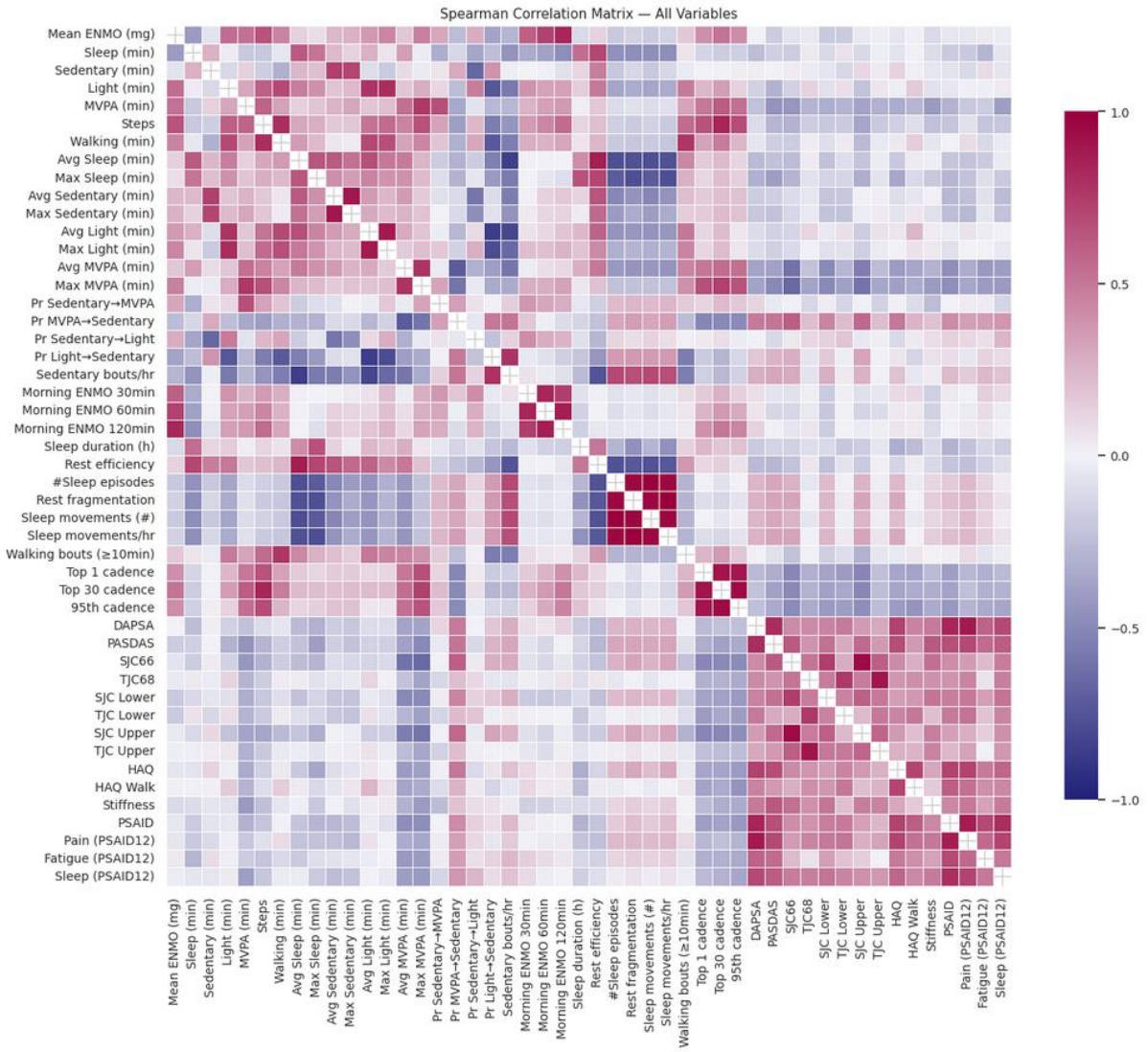


Figure 88 Full Spearman correlation matrix: smartwatch-derived features vs. clinical/PRO outcomes.

Appendix VI Supplementary material of HRV validation

Table 72 Results of statistical correlations between HRV markers and PASDAS score. Each entry shows Spearman's ρ with the corresponding p-value (in parentheses).

	Window =1h		Window =6h		Window =12h	
	AVG	STD	AVG	STD	AVG	STD
RMSSD	-0.25 (0.100)	-0.13 (0.405)	-0.256 (0.101)	-0.06 (0.703)	-0.25 (0.102)	-0.05 (0.726)
SDNN	-0.23 (0.129)	-0.260 (0.096)	-0.17 (0.270)	-0.139 (0.379)	-0.14 (0.355)	-0.09 (0.555)
Triangular index	-0.23 (0.131)	-0.246 (0.116)	-0.16 (0.293)	-0.16 (0.284)	-0.16 (0.309)	-0.10 (0.509)
ULF power	0.07 (0.642)	0.09 (0.564)	0.01 (0.937)	-0.01 (0.971)	-0.04 (0.759)	0.01 (0.919)
VLF power	0.23 (0.128)	0.25 (0.098)	0.18 (0.228)	0.11 (0.458)	0.14 (0.367)	0.20 (0.203)
LF power	0.19 (0.206)	0.01 (0.903)	0.20 (0.198)	0.17 (0.273)	0.12 (0.441)	0.18 (0.232)
HF power	0.07 (0.614)	-0.025 (0.874)	0.26 (0.091)	0.26 (0.085)	0.17 (0.261)	0.26 (0.085)
LF/HF	0.10 (0.493)	0.16 (0.291)	0.11 (0.484)	0.19 (0.222)	0.09 (0.561)	0.21 (0.172)
Total power	0.24 (0.121)	0.03 (0.826)	0.18 (0.238)	0.13 (0.386)	0.04 (0.759)	0.11 (0.460)
DFA α1	0.10 (0.513)	0.05 (0.716)	0.09 (0.565)	0.06 (0.659)	0.08 (0.579)	0.04 (0.794)
DFA α2	-0.27 (0.079)	0.036 (0.820)	-0.15 (0.332)	0.19 (0.213)	-0.11 (0.449)	0.14 (0.349)

Table 73 Results of statistical correlations between HRV markers and PSAID score. Each entry shows Spearman's ρ with the corresponding p-value (in parentheses).

	Window =1h		Window =6h		Window =12h	
	AVG	STD	AVG	STD	AVG	STD
RMSSD	-0.08 (0.524)	-0.07 (0.585)	-0.08 (0.527)	-0.05 (0.699)	-0.08 (0.515)	-0.03 (0.774)
SDNN	-0.07 (0.564)	-0.13 (0.327)	-0.04 (0.746)	-0.04 (0.746)	-0.04 (0.721)	-0.06 (0.652)
Triangular index	-0.10 (0.441)	-0.11 (0.391)	-0.09 (0.504)	-0.10 (0.420)	-0.08 (0.519)	-0.05 (0.696)
ULF power	0.06 (0.606)	0.13 (0.314)	-0.02 (0.850)	-0.03 (0.793)	-0.07 (0.573)	-0.10 (0.455)
VLF power	0.09 (0.498)	0.08 (0.542)	0.09 (0.494)	0.12 (0.347)	0.02 (0.851)	0.07 (0.575)
LF power	0.02 (0.881)	-0.10 (0.458)	0.06 (0.653)	0.14 (0.296)	0.02 (0.881)	0.05 (0.689)
HF power	0.02 (0.855)	-0.07 (0.569)	0.14 (0.283)	0.19 (0.146)	0.11 (0.408)	0.11 (0.406)
LF/HF	-0.03 (0.804)	-0.01 (0.946)	-0.05 (0.696)	-0.02 (0.840)	-0.04 (0.744)	-0.01 (0.889)
Total power	0.09 (0.493)	-0.06 (0.616)	0.09 (0.500)	0.13 (0.321)	0.00 (0.996)	-0.02 (0.881)
DFA α1	-0.02 (0.832)	-0.04 (0.734)	-0.03 (0.783)	-0.01 (0.948)	-0.04 (0.759)	-0.03 (0.815)
DFA α2	0.01 (0.924)	-0.07 (0.585)	-0.01 (0.910)	0.12 (0.358)	-0.01 (0.888)	-0.02 (0.853)

Appendix VII Supplementary material on All-Day Metrics

Table 74 Full set of n=65 time-series features extracted from wearable data using the tsfresh Python library. Each feature was statistically tested for association with the target variable *Physician-assessed flare* using univariate hypothesis testing. Reported p-values reflect raw significance levels prior to correction for multiple comparisons.

Feature	p value
bodyBattery__linear_trend__attr_"intercept"	0.002008
bodyBattery__minimum	0.0035
bodyBattery__mean	0.007264
beatsPerMinute__minimum	0.015385
beatsPerMinute__mean	0.02023
beatsPerMinute__linear_trend__attr_"intercept"	0.020685
stressScore__linear_trend__attr_"intercept"	0.047148
beatsPerMinute__maximum	0.052044
bodyBattery__sum_values	0.061914
stressScore__minimum	0.07603
bodyBattery__maximum	0.084775
stressScore__linear_trend__attr_"slope"	0.09939
stressScore__mean_change	0.101132
steps_daily__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.113993
sedentary_prop__percentage_of_reoccurring_values_to_all_values	0.126308
stressScore__mean	0.134639
stressScore__linear_trend__attr_"rvalue"	0.143638
stressScore__maximum	0.143647
sedentary_prop__minimum	0.210704
steps_daily__mean_change	0.216909
walking_time__mean_change	0.229705
sedentary_prop__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.284417
stressScore__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.301794
steps_daily__linear_trend__attr_"rvalue"	0.343019
walking_time__linear_trend__attr_"rvalue"	0.396954
sedentary_prop__mean	0.406422
sedentary_prop__median	0.430659
stressScore__variance	0.44559
steps_daily__linear_trend__attr_"slope"	0.450634
bodyBattery__linear_trend__attr_"slope"	0.455713
stressScore__standard_deviation	0.502811
bodyBattery__linear_trend__attr_"rvalue"	0.502815
walking_time__linear_trend__attr_"slope"	0.508199
beatsPerMinute__linear_trend__attr_"slope"	0.51361
beatsPerMinute__linear_trend__attr_"rvalue"	0.543944
walking_time__minimum	0.560808

Feature	p value
beatsPerMinute__mean_change	0.563691
sedentary_prop__maximum	0.586516
sedentary_prop__variance	0.604208
walking_time__maximum	0.621952
steps_daily__linear_trend__attr_"intercept"	0.639937
walking_time__mean	0.63994
stressScore__mean_abs_change	0.652035
beatsPerMinute__standard_deviation	0.682696
steps_daily__variance	0.688893
steps_daily__mean_abs_change	0.688893
steps_daily__standard_deviation	0.688893
steps_daily__sum_values	0.707609
walking_time__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.723307
beatsPerMinute__variance	0.739188
stressScore__sum_values	0.75195
beatsPerMinute__sum_values	0.784141
sedentary_prop__mean_change	0.797123
steps_daily__maximum	0.797125
walking_time__mean_abs_change	0.829821
bodyBattery__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.856193
walking_time__linear_trend__attr_"intercept"	0.856199
bodyBattery__mean_abs_change	0.869451
bodyBattery__mean_change	0.90273
beatsPerMinute__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.9094
walking_time__standard_deviation	0.909407
walking_time__variance	0.909407
steps_daily__minimum	0.956309
beatsPerMinute__mean_abs_change	0.969744
walking_time__sum_values	0.969745
sedentary_prop__mean_abs_change	0.983189
steps_daily__mean	0.989913
bodyBattery__variance	1
bodyBattery__standard_deviation	1

Table 75 RFECV selected features and feature importances for targets MDA and physician-assessed flare (DOC_FLARE).

Selected Feature (MDA)	Importance	Selected Feature (DOC_FLARE)	Importance
bodyBattery__linear_trend__attr_"slope"	0.054838	bodyBattery__standard_deviation	0.233888
stressScore__maximum	0.049562	bodyBattery__minimum	0.220279
steps_daily__variance	0.047917	steps_daily__change_quantiles__f_agg_"mean"__is	0.193188

Selected Feature (MDA)	Importance	Selected Feature (DOC_FLARE)	Importance
		abs_True__qh_0.5__ql_0.0	
steps_daily__maximum	0.047609	walking_time__minimum	0.180342
sedentary_prop__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.046465	bodyBattery__mean	0.172302
sedentary_prop__mean_change	0.043399		
sedentary_prop__variance	0.038998		
sedentary_prop__maximum	0.038424		
steps_daily__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.035479		
steps_daily__linear_trend__attr_"rvalue"	0.033686		
bodyBattery__mean_change	0.032087		
beatsPerMinute__standard_deviation	0.030487		
steps_daily__mean_change	0.02918		
bodyBattery__linear_trend__attr_"rvalue"	0.02732		
bodyBattery__mean_abs_change	0.027242		
steps_daily__linear_trend__attr_"intercept"	0.027105		
beatsPerMinute__linear_trend__attr_"intercept"	0.026905		
walking_time__linear_trend__attr_"rvalue"	0.026687		
stressScore__variance	0.026301		
walking_time__maximum	0.025563		
walking_time__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.025352		
stressScore__linear_trend__attr_"rvalue"	0.024619		
bodyBattery__linear_trend__attr_"intercept"	0.024308		
sedentary_prop__mean	0.023238		
stressScore__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.022901		
beatsPerMinute__mean_change	0.021498		
steps_daily__linear_trend__attr_"slope"	0.021224		
stressScore__minimum	0.021066		
bodyBattery__sum_values	0.020146		
bodyBattery__maximum	0.019901		
beatsPerMinute__variance	0.016897		

Selected Feature (MDA)	Importance	Selected Feature (DOC_FLARE)	Importance
walking_time__linear_trend__attr_"intercept"	0.014677		
walking_time__mean_change	0.014553		
beatsPerMinute__minimum	0.014366		

Table 76 Features selected by Elastic Net logistic regression for predicting MDA and physician-assessed flare (DOC_FLARE).

Selected Features (MDA)	Coefficient	Selected Features (DOC_FLARE)	Coefficient
beatsPerMinute__variance	-2.11063	steps_daily__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	-0.67809
steps_daily__variance	1.952609	bodyBattery__linear_trend__attr_"intercept"	-0.60121
beatsPerMinute__linear_trend__attr_"rvalue"	-1.86025	walking_time__maximum	0.443538
stressScore__linear_trend__attr_"rvalue"	1.80804	beatsPerMinute__minimum	0.393369
bodyBattery__linear_trend__attr_"slope"	1.706367	sedentary_prop__mean_abs_change	-0.38175
walking_time__mean_abs_change	-1.61836	stressScore__linear_trend__attr_"slope"	-0.27629
steps_daily__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	1.351002	beatsPerMinute__maximum	0.258942
bodyBattery__mean_change	-1.27366	beatsPerMinute__mean_change	0.254922
beatsPerMinute__mean_abs_change	1.140823	steps_daily__linear_trend__attr_"slope"	0.185195
walking_time__linear_trend__attr_"intercept"	-1.12332	stressScore__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	-0.17922
walking_time__sum_values	1.021628	bodyBattery__variance	0.136815
sedentary_prop__maximum	1.01288	bodyBattery__mean_abs_change	0.131209
stressScore__maximum	0.799269	walking_time__linear_trend__attr_"rvalue"	0.097129
sedentary_prop__percentage_of_reoccurring_values_to_all_values_1.0	0.745862	bodyBattery__minimum	-0.09129
walking_time__mean_change	-0.73901	steps_daily__mean_change	0.088021
stressScore__linear_trend__attr_"slope"	0.712437	bodyBattery__mean	-0.08668
bodyBattery__variance	0.702694	walking_time__sum_values	0.084759
bodyBattery__mean_abs_change	-0.68988	sedentary_prop__percentage_of_reoccurring_values_to_all_values_1.0	0.077366
beatsPerMinute__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.599042	stressScore__sum_values	-0.07431

bodyBattery__linear_trend__attr_"rvalue"	0.549604	beatsPerMinute__mean	0.061671
stressScore__mean_change	0.520844	steps_daily__linear_trend__attr_"rvalue"	0.030176
beatsPerMinute__mean_change	-0.5202	walking_time__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	-0.02956
sedentary_prop__mean_abs_change	0.496308	walking_time__mean_change	0.019645
beatsPerMinute__linear_trend__attr_"slope"	-0.43606	walking_time__minimum	0.005476
walking_time__standard_deviation	-0.35904		
bodyBattery__maximum	-0.35346		
steps_daily__minimum	0.32507		
walking_time__variance	-0.32433		
steps_daily__mean_abs_change	0.31527		
walking_time__linear_trend__attr_"rvalue"	0.289716		
sedentary_prop__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.280915		
sedentary_prop__mean_change	0.211306		
sedentary_prop__median	-0.19503		
beatsPerMinute__minimum	-0.14347		
steps_daily__maximum	0.128799		
stressScore__change_quantiles__f_agg_"mean"__isabs_True__qh_0.5__ql_0.0	0.071469		
steps_daily__linear_trend__attr_"rvalue"	-0.06777		
bodyBattery__sum_values	-0.05718		
steps_daily__linear_trend__attr_"intercept"	-0.0376		
walking_time__minimum	0.033568		