D4.5 - Proof-of-Concept Testing and Validation in chronic pain management in everyday life - Phase 2

Project No	GA824160
Project Acronym	EnTimeMent
Project full title	ENtrainment & synchronization at multiple TIME scales in the MENTal foundations of expressive gesture
Instrument	FET Proactive
Type of action	RIA
Start Date of project	1 January 2019
Duration	48 months

EnTimeMent D4.5

Distribution level	[PU][1]
Due date of deliverable	Month 36
Actual submission date	26 January 2022
Deliverable number	4.5
Deliverable title	Proof-of-Concept Testing and Validation in chronic pain management in everyday life
Type	ORDP (Open Research Data Pilot)
Status & version	
Number of pages	39
WP contributing to the deliverable	4
WP / Task responsible	UCL
Other contributors	ALL
Author(s)	UCL
EC Project Officer	Christiane Wilzeck
Keywords	**Analysis, Chronic pain, Data collection, Modelling, Movement, Physical activity, Physiotherapist, Time**

[1] **PU** = Public, **PP** = Restricted to other programme participants (including the Commission Services), **RE** = Restricted to a group specified by the consortium (including the Commission Services), **CO** = Confidential, only for members of the consortium (including the Commission Services).

Contents

Abbreviations

EU	European Union
EC	European Commission
WP	Work Package

EnTimeMent D4.5

1   Introduction

This deliverable summarises the findings of studies that aim to develop the scenarios for the chronic pain use case. It comprises three main parts:

- **Movement data collection in participants' homes and understanding of home activity challenges:** This section focuses on the collection of a novel dataset that captures body movement and corresponding self-report of people with chronic pain as they carry out in their own home functional activities of value to them. The purpose of the dataset is for advancing investigation on the automatic detection of pain and related experiences in spontaneous activities of people with chronic pain. The study reported also covers interviews with the participants with chronic pain to provide further insight into the roles that technology could play in supporting self-management.

- **Understanding of relevant ground truth for building movement dataset that is rudimentary to the development of technology for self-management in the context of physical activity:** This section reports two studies. The first study takes advantage of interviews with participants in their home from the above section and probes the use of chatbots as a way to gather self-report from participants while at the same time creating valuable moments for reflection and companionship. The second study explores with physiotherapists a more refined language of pain-related protective behaviour than the one presented in pain literature (Keefe and Block 1982).

- **Modelling of physical activity related pain behaviour**: This section also covers two studies. The first study reports improvement on automatic recognition of pain-related behaviour based on the fusion of body movement and muscle activity in a graphical structure model that reflects natural body configuration. The software is reported in D3.4. The second study presents initial experiments on automatic classification of pain levels based on the new dataset.

2   A Home Study and The EmoPain@Home EnTimeMent Dataset

The main aim of the home study was to create an open dataset (called *EmoPain@Home EnTimeMent* dataset) capturing pseudonymized movement data of people engaged in everyday activities in their homes. There are still very few movement datasets that reflect in-the-wild situations (one of the findings from our datasets survey reported in D3.8), and there are almost no datasets that capture body movements of people with chronic pain in their daily life. The primary purpose of this new dataset is for investigation of automatic recognition of pain and related experiences from body movement.

Additional aims of the study included: 1) to further understand (beyond the survey findings reported in previous deliverable D4.4) the types of movement that would be useful to record in homes of people with chronic pain to inform technology-based intervention; 2) to

understand how we can develop self-annotation approaches to acquire richer labels; and 3) since the study started during the COVID-19 lockdown, it further afforded us the opportunity to explore how studies could be remotely run in participant homes with minimal intrusiveness. Despite its challenges, this has been an interesting journey.

An overview of the EmoPain@Home dataset is given in D3.8. This section describes the protocol used in the study and analyses of the dataset. Preliminary modelling of the dataset is reported in Section 4.3. Further modelling will be pursued in the next year of work.

2.1 Protocol

The study was approved by the UCL Research Ethics Committee (Project Ref: 5095/001).

2.1.1   Recruitment

Participants were recruited through open-call advertisements on social media (i.e., Twitter), as well as by directly getting in touch with relevant support groups and gatekeepers within different pain organisations. All potential participants had been extensively briefed before signing up to the project, either via public presentations by the members of the research group, or by discussing with the researcher an information pack prepared for the study. Participants signed a standard consent form with additional levels of optional consents. The optional consents covered the following:

1. use of anonymous data (sensor data, text responses from diaries, self-reports, interviews) by other researchers;
2. use of photos withxin dissemination (written) projects;
3. use of photos, videos and audio in presentations;
4. use of personal information for secondary analysis by the research group;
5. option to be contacted to participate in follow-up studies within the project, or future studies of a similar nature.

All responses were captured via a REDCap form, which was subsequentially stored in a UCL Data Safe Haven (hereafter, DSH) folder for higher data protection level (https://www.ucl.ac.uk/isd/services/file-storage-sharing/data-safe-haven-dsh). DSH is used for very sensitive data, e.g., clinical data.

10 participants were recruited, but one withdrew, resulting in 9 participants taking part in the study. The participants comprised 5 female and 5 male between the ages of 27 and 59 (1 participant in their 20s, 2 participants in their 30s, 3 participants in their 40s, and 4 participants in their 50s; mean=45.11, standard deviation=11.50). All participants recruited self-identified as living with musculoskeletal chronic pain (CP) involving the lower back area, although the severity of their condition greatly varied. 4 of the participants reported sciatica as their CP condition; 2 participants reported CP resulting from an old spinal injury; and the 3 others reported other CP. See Table 1 for details on the activities the participants were engaged in during the data recordings.

7

Table 1. Types of activities recorded. The * on the ID means that the person took part in both supervised and unsupervised sessions. Activities in bold were only carried out in unsupervised sessions, whereas activities underlined were carried out in both supervised and unsupervised contexts

ID	Challenging activities	Activities not challenging
P002*	<u>Changing bed sheets</u> <u>Washing up</u> Loading washing machine Unloading washing machine Clean windows	Power walking Sweeping floors <u>Dusting</u> **Hoovering**
P003*	Hoovering <u>Washing up</u> Clean bathroom Unloading shopping Cleaning windows **Tidying up**	<u>Washing up</u> <u>Unloading washing machine</u> <u>Loading washing machine</u>
P004*	<u>Hoovering</u> <u>Changing bed sheets</u> Vacuuming inside of car Watering garden	Clean bathroom Dusting inside of car **Preparing food** **Cleaning parrot cage**
P005	Painting shelves Painting wall Power walking Cleaning bathroom	-
P006	Changing bed sheets	Unload dishwasher Organising boxes
P007	Unloading washing machine Unloading dishwasher Changing bed sheets	Loading washing machine Loading dishwasher Tidying up
P008	Washing up Hanging clothes to dry Hoovering Changing bed sheets Cleaning windows	Yoga Unloading washing machine
P009	Ironing Preparing food	Filing documents

P010	Tidying up	-

2.1.2   Sensing kit and other material

Body movement sensors: The Notch sensors (www.wearnotch.com) (see Figure 1) and the accompanying smartphone app were used to capture the movement of participants. The Notch sensors were used as they are low-cost, wearable and run on most smartphones. This allowed us to simply send the sensors and a smartphone by post to the participants, and let the participant wear the sensors on their own. This set up also allowed participants to be free to move between rooms as they would normally do. The cost and the setup reflect more closely a possible real-life scenarios, both in terms of affordability and ease of use.

Figure 1. Notch sensors used

6 Notch sensor modules were used. These were placed on one side of the person's body. Single-sided positioning was motivated by previous literature [Olugbade et al. 2018, 2019], which showed that getting data from just one side of the body could be sufficient for automatic detection of pain and related states while reducing potential interference between sensors typical of larger network of sensors. We only used movement sensors due to the already high effort required of participants in terms of managing the sensors while in potentially painful scenarios. In discussion with the clinical psychologist on the team, we concluded that as a first data collection in the wild with only remote interaction with the researcher, the setup should not be any more demanding for the participants.

Each participant wore the set of 6 Notch modules at the following positions: 1 x right wrist; 1 x right upper arm; 1 x chest; 1 x waist; 1 x right thigh; 1 x right calf (see **Error! Reference ource not found.**). Each Notch module is equipped with 3 inertia sensors (accelerometer, gyroscope and compass), which allowed us to measure joint positions and 3D Euler angles at a sampling frequency of 40 Hz.
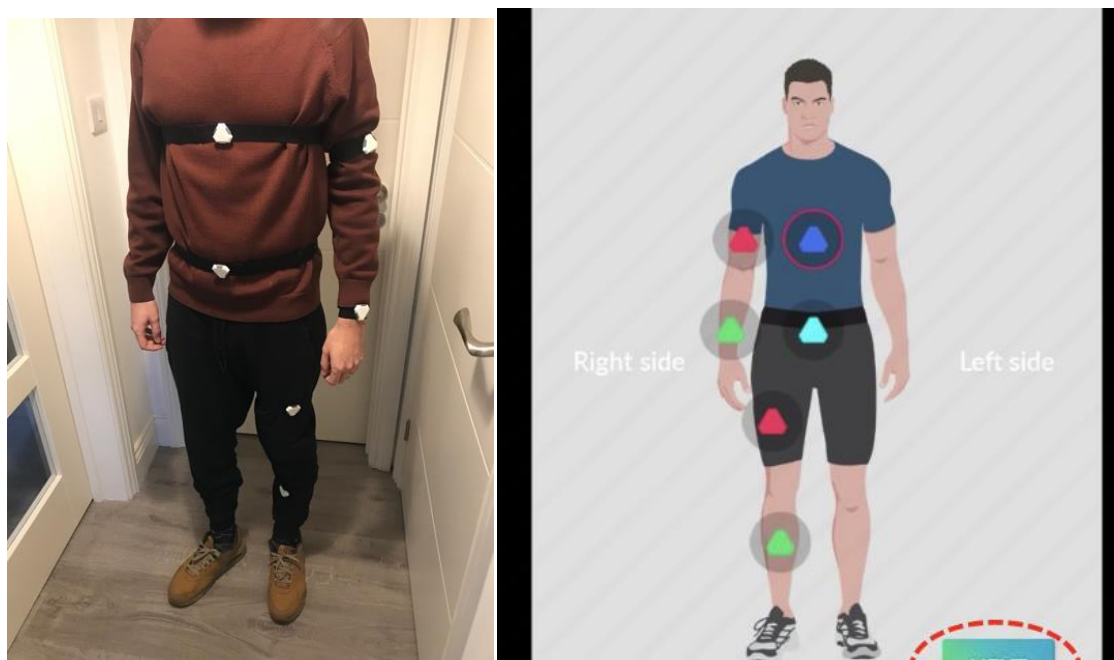
Figure 2. Sensors positioning (on body and in diagrammatic form)

Participants were fully trained on how to wear the sensors and to how to calibrate and turn on and off the sensors, as well as on how to record, download and share recordings, but they carried out each of these tasks under the remote supervision of the researcher (i.e. via a videoconferencing system). In some instances, one of the motion capture modules failed to come on when turned on; in that case, the participant was asked to record using the 5 modules that worked (on the right wrist, right upper arm, chest, waist, and right thigh).

Video camera and microphone: The webcam and microphone of the participant's computer used for the videoconference with the researcher was used to record the sessions for qualitative analyses as well as to capture the self-reports. The video and audio data were livestreamed by the researcher on their own computer, and the data was recorded at the researcher end. A private videoconference connection was used to maximise confidentiality of the data, and the recorded data was immediately moved to a more secure location, e.g. a hardware-encrypted offline disk, or the UCL DSH. It was positioned in such a way to maximise the view of the activity. However, as the activity was ubiquitous, the activity would not always be in the range of the cameras. The participant's voice self-report would be continuously captured. The visibility from the video camera was also dependent on the size of the room, and the furniture configurations.

Diary: A diary (**Error! Reference source not found.**) was prepared to help participants reflect n their days and their plans for a given day. The diary was used during the full period of the

study to understand how their home activities are affected by their pain level and general affective states. The aim was to

Figure 3. Diary instance

understand in more details how a pain and mood aware wearable device could better support people in planning and engaging with their home activities. The diary served a double purpose: on the one hand, it was a reflection tool for participants to preliminary think about some of the topics discussed during the interviews; on the other hand, it functioned as a comparative tool for the researcher to get a sense of how participants' perceptions of their pain, confidence, worries, plans, etc. varied throughout the day, and between supervised moments (i.e. interviews), and unsupervised ones (i.e. diary).

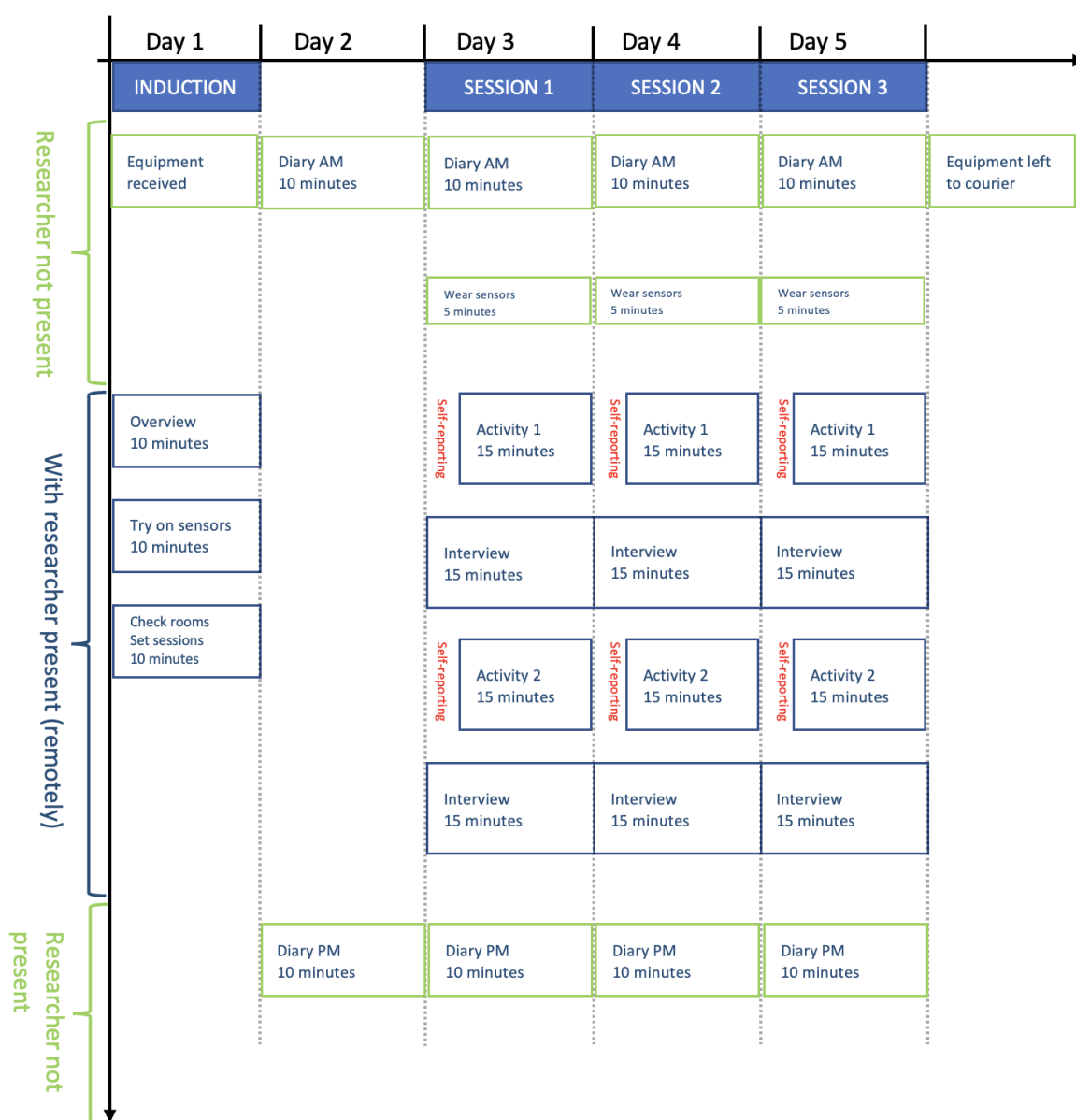|	Day 1	Day 2	Day 3	Day 4	Day 5	
	INDUCTION		SESSION 1	SESSION 2	SESSION 3	
Researcher not present	Equipment received	Diary AM 10 minutes	Diary AM 10 minutes	Diary AM 10 minutes	Diary AM 10 minutes	Equipment left to courier
			Wear sensors 5 minutes	Wear sensors 5 minutes	Wear sensors 5 minutes	
With researcher present (remotely)	Overview 10 minutes		Activity 1 15 minutes	Activity 1 15 minutes	Activity 1 15 minutes	
	Try on sensors 10 minutes		Interview 15 minutes	Interview 15 minutes	Interview 15 minutes	
	Check rooms Set sessions 10 minutes		Activity 2 15 minutes	Activity 2 15 minutes	Activity 2 15 minutes	
			Interview 15 minutes	Interview 15 minutes	Interview 15 minutes	
Researcher not present		Diary PM 10 minutes	Diary PM 10 minutes	Diary PM 10 minutes	Diary PM 10 minutes	

Figure 4. Schematic overview of data collection protocol. In actuality, the days may not be consecutive, but spread across a slightly longer period according to participants' availability

2.1.3   Data collection protocol

Participants were given the option to sign up for just one or two recording sessions. All participants except one had 1 half-hour training session and 3 hour-long recording sessions. All study sessions within this leg of the study were carried out with the researcher present. The entire study was conducted remotely and in compliance with Covid-19 health and safety protocols. Participants were sent all the equipment needed via courier, and interviews and observations were conducted via Microsoft Teams. All participants were reimbursed for their time at the standard rate of £10 per hour. Each session was split into two main parts: activities and interview.

Activities: Participants were asked to perform indoor household activities that they would normally carry out as part of their day-to-day routines. Participants performed between 1 and 3 activities each session, with each activity lasting between 5 and 30 minutes, and with an average of 15 minutes per activity. Activities included chores as varied as cleaning surfaces, hoovering floors, painting, washing dishes, loading/unloading washing machines, etc (see Table 1). Before starting each activity, participants were asked to label it as either 'challenging' or 'not challenging'; it should be noted that, they had been instructed by the researcher that 'challenging activities' should not be understood as activities that go out of their routine, or that could cause any harm to them.

While carrying out each activity, participants' movements were recorded via video and motion capture sensors (see Section 2.1.2). Participants self-reported their levels of pain, worry, and confidence every minute while performing the activities. In particular, they responded to the following prompt from the researcher: "What is your level of pain, worry, and confidence at this moment in time?". Pain and worry were assessed on an ascending numerical scale from 0 to 10, where 0 equalled to 'no pain (or worry)' and 10 to 'very severe pain (or worry)'; confidence was assessed using an ordinal scale of {no confidence; less than average confidence; average confidence; more than average confidence; max confidence}. In order to avoid disruption to the overall flow of the physical activity, the long-form question above was replaced by a shorter prompt (i.e. *'time'*).

Interview: Participants took part in 30-minute semi-structured interviews during each session. Interview scripts were prepared in advance, but such script served as merely a set of cues, with most of the interviews being adapted on the go according to the flow of the conversation and responses from the interviewee. Nonetheless, throughout the interviews, four main topics were explored:
   1. Participants' understandings and descriptions of what makes an activity challenging or not challenging;
   2. Explorations of the self-reported experiences (worry, pain, confidence), questioning what these constructs mean to participants;

3. Discussion regarding the process and manner of self-reporting itself, questioning its usefulness for participants and how it could be improved;

4. Discussion on expectations and hopes regarding the future use of technology for chronic pain management.

While the researcher strived to dedicate roughly an equal amount of time to each topic, the actual time allotted greatly varied and was adapted to how much each participant was interested and engaged in each topic.

Diary: Additionally, participants were asked to fill in a diary twice a day while taking part in the study (see Figure 3). This was completely optional, and, while reminders were sent by the researcher twice daily (AM and PM), participants were never coerced into filling it in, nor they were penalised for not doing it. In total, for those who completed the diary in every instance, 8 records were collected.

Unsupervised activities: Lastly, half of the participants took part in an extra leg of the study a few months after their first experience. As part of this, they used the sensors in 4 additional 30-minute recording sessions, carrying out 2 activities in each session. Out of these, 2 sessions were completely unsupervised, i.e. they carried out activities on their own and with only the body movement sensors (and so video and audio were not recorded), self-reporting only at the beginning and end of each activity. An unsupervised approach was followed for multiple reasons: 1) to speed up the data collection process since participants could carry out activities at their own leisure, without needing to make suitable arrangements for both them and the researcher; 2) to better understand how comfortable potential users would be in using wearable technology on their own; 3) to get a sense of the experience of the participants when not continuously self-reporting. (2) and (3) were further discussed with participants themselves through a brief interview at the end of the last extra session. We further envisaged that the unsupervised approach could be important to give participants the possibility to carry out activities in parts of the house that were considered more private, or during more private moments of the day (e.g., in the morning), but which may also be critical to self-management (e.g., starting the day may be harder as the body has to warm up). For unsupervised sessions, activities were discussed in advance to ensure that people would not overdo in order to fulfil the requirements for the project; the researcher also ensured that someone was present in the home while such activities were carried out, so as to minimise potential risks of injury.

2.1.4   Data storage

As a result of this study, the following data have been produced:
1. Pseudonymised motion-capture movement data (.bvh and .csv files);
2. Pseudonymised numerical self-reports;
3. Pseudonymised interview transcripts (identifiable information such as names, locations, etc., were removed);

4. Interview audio and video recordings (only open to members of the research team).

2.2 Qualitative analysis

2.2.1   Methodology

All audio data from the interviews were transcribed and analysed, i.e. summarised and grouped into themes and codes according to the principles of latent and reflexive thematic analysis (Joffe & Yardley, 2004). Reflexive analysis was chosen in order to be flexible and enable codes and themes to emerge from specific research questions used to probe the data (Braun & Clarke, 2012, 2014). After an initial overarching analysis, the data were divided into the following initial themes and codes:

1. Sociality of chronic pain
   i. emotional and physical support;
   ii. accountability;
   iii. encouragement;
   iv. check-in, monitoring and feedback;
   v. conversations and distractions
2. Role of technology
   i. personalised care;
   ii. reminders and planners;
   iii. cognitive load and trust;
   iv. awareness and knowledge
3. Pacing and planning
   i. cautions and worries;
   ii. reassurance;
   iii. grounding;
   iv. confidence;
   v. know-how;
   vi. planning strategies;
   vii. difficulties;
   viii. breathing
4. Self-reflection and monitoring
   i. challenging vs not challenging;
   ii. dangers;
   iii. awareness;
   iv. emotions and monitoring

This initial analysis was then subsequently refined under two different concepts developed in two separate analyses. The first analysis focuses on the self-reporting process and experience, and this analysis has now been published in (Bi et al. 2021), and an overview is given in Section 3.1. The second analysis targets the multiple timescales of movement decision making, and is

15

currently being prepared for publication, and a succinct overview is sketched out in Section 2.2.2.

2.2.2   Movement as decision-making (Buono et al., in prep)

This analysis was aimed at providing deeper insights (beyond previous work such as Olugbade et al. 2019; Singh et al. 2017) to better understand the strategies deployed and the challenges faced by participants in carrying out everyday activities. We framed the analysis around the concept of decision-making: while for people not living in CP, movement in such scenarios is almost instantaneous and unconscious (for experimental evidence, please see: Libet et al., 1983; D'Ostilio & Garraux, 2012), the interviews highlighted that people with CP tend make several *explicit* assessments before and during execution of movement, e.g. should I move at all or not? how should I move? what (body part) should I move and at what speed? etc.

Our analysis showed that such decision-making moments happen at two distinct, but intertwined, timescales – the micro-timescale of the 'now', and the macro-timescale of the entire day and/or week. As far as the former is concerned, good decision-making seems to be enhanced by pacing, whereas in the latter, planning seems to be of paramount importance. In terms of strategies used in decision making, participants reported a heightened capacity based on grounding and mindful techniques (Mason & Hargreaves, 2001; Nielson et al., 2013) to strategise for real-time decisions, with particular importance given to body scans, self-reporting and correct breathing tempo. These tools heightened the actor's capacity to connect to their body, be aware of its conditions, and accommodate its needs accordingly. For longer-term decisions, participants stressed the importance of strategic planning, reporting how it follows similar principles of detached awareness (describing the ability to make adaptable projects and set achievable goals) and of foregrounding needs and challenges in advance.

However, while participants seemed all fairly aware of how to improve their decision-making abilities, and thus verbalised about the above strategies extensively, they all highlighted challenges to their actual incorporation of these strategies in their routines. Such challenges comprised for instance: 1) a general lack of trust of their own judgements and intuitive assessments gained through grounding; 2) difficulties in re-adjusting plans on the go; 3) lack of awareness of when pacing and planning is particularly needed; 4) cognitive toll these strategies take on the person in pain.

The analysis further clustered together responses and comments which covered participant discussion of how technology could aid in resolving some of the challenges posed by pacing and planning, thus proving itself a support for decision-making. In particular, the data highlights that participants envisioned a system that could prompt them with reminders to pay attention to their bodies, their breathing, etc. when this is most needed (e.g. when expecting pain, worried, or distracted, etc.). Such reminders would aim to enhance the user's ability to make strategic decisions, but not impose a way of doing on the user. In fact, all

participants strongly advocated for solutions which would maintain and enhance their sense of agency. Furthermore, participants described the usefulness of a system which could also aid them at the macro-timescale, for instance a smart planner which would be able to make recommendations (e.g. postponing an activity, splitting it into chunks, carrying something out for a shorter time, etc.) based on the user's levels of pain, confidence, emotional state, etc.

2.3 Descriptive Statistics

In this section, we give an overview of the distribution of activities and labels in the EmoPain@Home dataset.

2.3.1   Data Instance Definition

To enable quantitative analysis, we segmented the dataset by defining the one-minute time window preceding each self-report as a data instance. Thus, for each participant activity a^k we extracted n_k instances $\{m_1^k, m_2^k, \dots m_i^k, \dots, m_{n_k}^k\}$ such that

$$m_i^k = \left\{ a_{t_{i_k}-60}^k, a_{t_{i_k}-59}^k, \dots, a_{t_{i_k}-1}^k \right\}$$

where t_{i_k} is the time (in seconds) of the ith ESM self-report l_i^k of a^k, $k \in \{1, 2, \dots, K\}$; and instance m_i^k is specified by:

- the kinematic data captured for the given participant in activity a^k particularly the data recorded between second $t_{i_k} - 60$ and $t_{i_k} - 1$,
- the (pain, worry, and confidence) experience labels for time t_{i_k} based on the participant's self-report at that time,
- the activity label for a^k,
- the level of challenge self-reported by the participant for activity a^k, and
- the id for the participant.

We excluded participant activities where the time of the self-report was not noted, e.g. when the participant did not say it out loud (or it could not otherwise be heard) at the time of the experience itself, or during the unsupervised activities.

2.3.2   Distribution of Data Instances by Activity type

As can be seen in Figure 5, there is a wide variety of activity types (26 in total) in the dataset. While most of the activities are exclusive to individual participants (e.g. only P8 had a yoga activity), a number of activities are common to 3 or more participants. For instance, 5 participants had *changing bed sheets* activities, and 3 participants had *bathroom cleaning* activities. However, in general, most of the activities are represented by fewer than 20 data instances.

These have implication for computational modelling of the data instances. First, it will be challenging to build a model that generalizes to unseen participants who carry out activity types not represented in the training set. Second, the disproportionate number of activity

types (disproportionate to the total number of data instances) already introduces a peculiar level of complexity to the learning process for the model given that the expression of target experiences (e.g. low level confidence) can vary not only across people but further across activity types for each person.
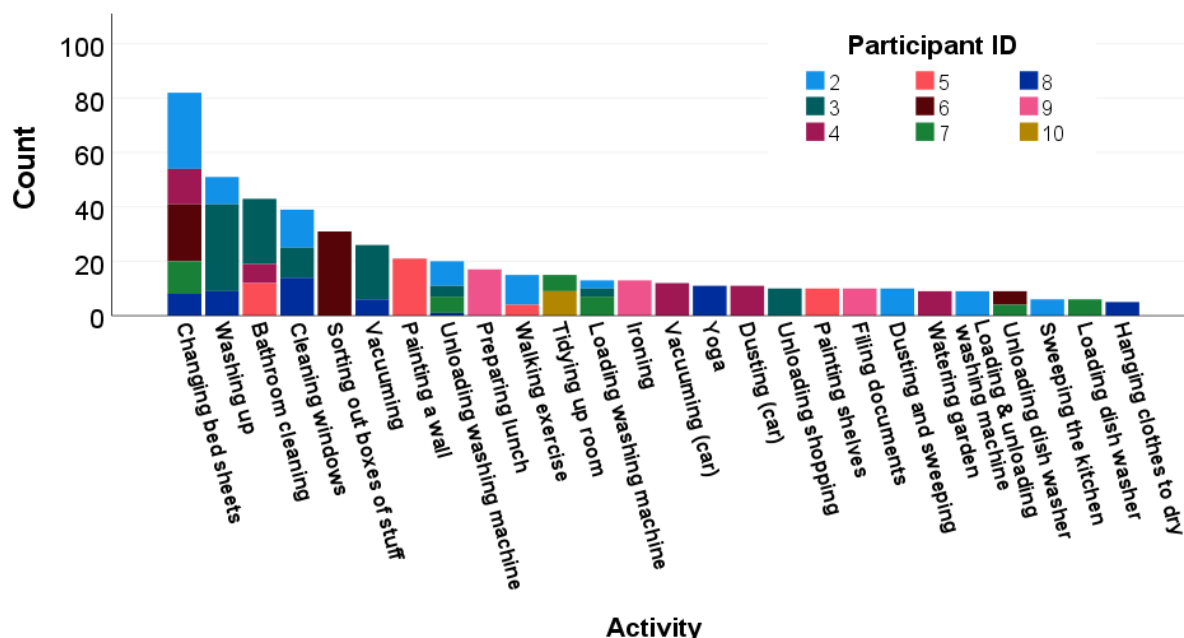

Figure 5. Distribution of data instances in the EmoPain@Home dataset by activity type.

2.3.3   Distribution of Data Instances by Experience Labels

For pain experience (Figure 6 top-left), most of the data instances have mid level values with a peak in number of instances at pain level 4. Most of the data instances are samples from activities that the participants reported that they typically find challenging, and although the highest level of pain reported for instances from the non-challenging activities was only 6, instances from the challenging activities had pain level as low as 0.

For worry experience (Figure 6 top-right), the distribution of data instances is right skewed with a peak at worry level 2. As with pain, instances from the challenging activities had all (integer) worry levels from 0 to 10 whereas instances from the non-challenging activities did not have worry levels higher than 5.

Confidence experience (Figure 6 bottom), similar to the finding for worry, has a skew on the positive side with its peak at 4. Instances from the challenging activities have confidence labels between the two ends of the confidence scale (i.e. between 1 and 5), while instances from the non-challenging activities have confidence labels between 3 and 5.
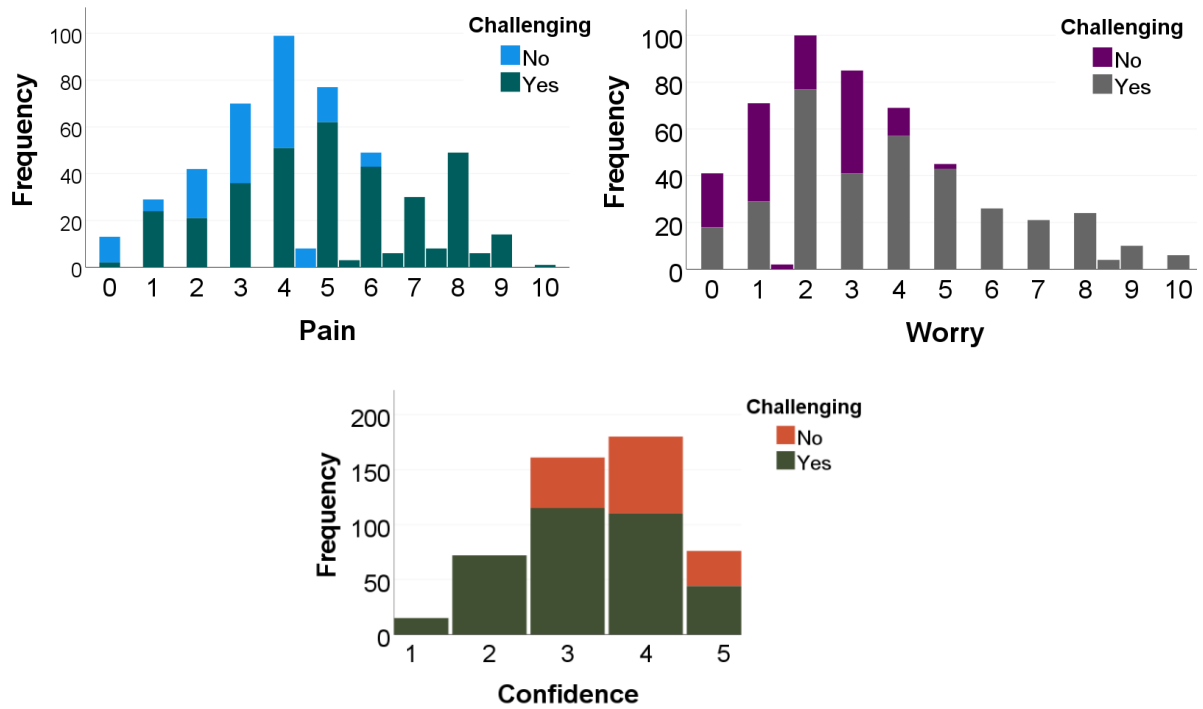
Figure 6. Distribution of the data instances in the EmoPain@Home dataset by experience labels.

To further understand the variability of each experience label within session, we plot the given experience label separately for each different participant session activity. Figure 7, Figure 8, and Figure 9 show the variabilities for pain, worry, and confidence respectively. The plots suggest that there is higher variability in the challenging activities for all three labels. The plots additionally show differences in the range of values across participants

Figure 7. Distribution of pain levels of data instances in the EmoPain@Home dataset by participant activity per session.

Figure 8. Distribution of worry levels of data instances in the EmoPain@Home dataset by participant activity per session.

Figure 9. Distribution of confidence levels of data instances in the EmoPain@Home dataset by participant activity per session.

3   What is the Ground Truth?

3.1 Chatbot opportunities for rich continuous labelling of pain experience in ubiquitous settings (Bi et al. 2021)

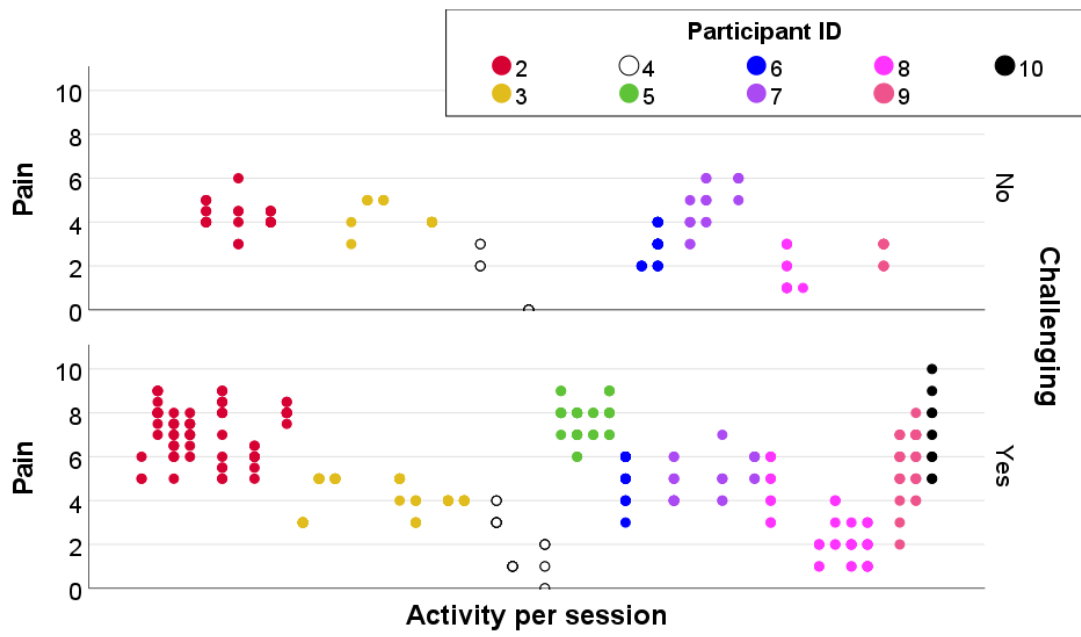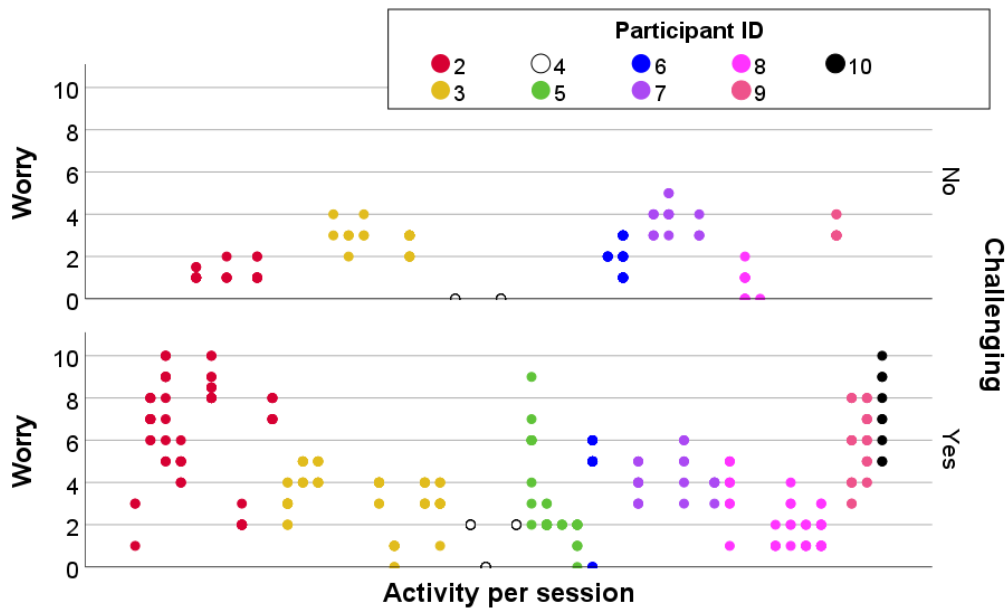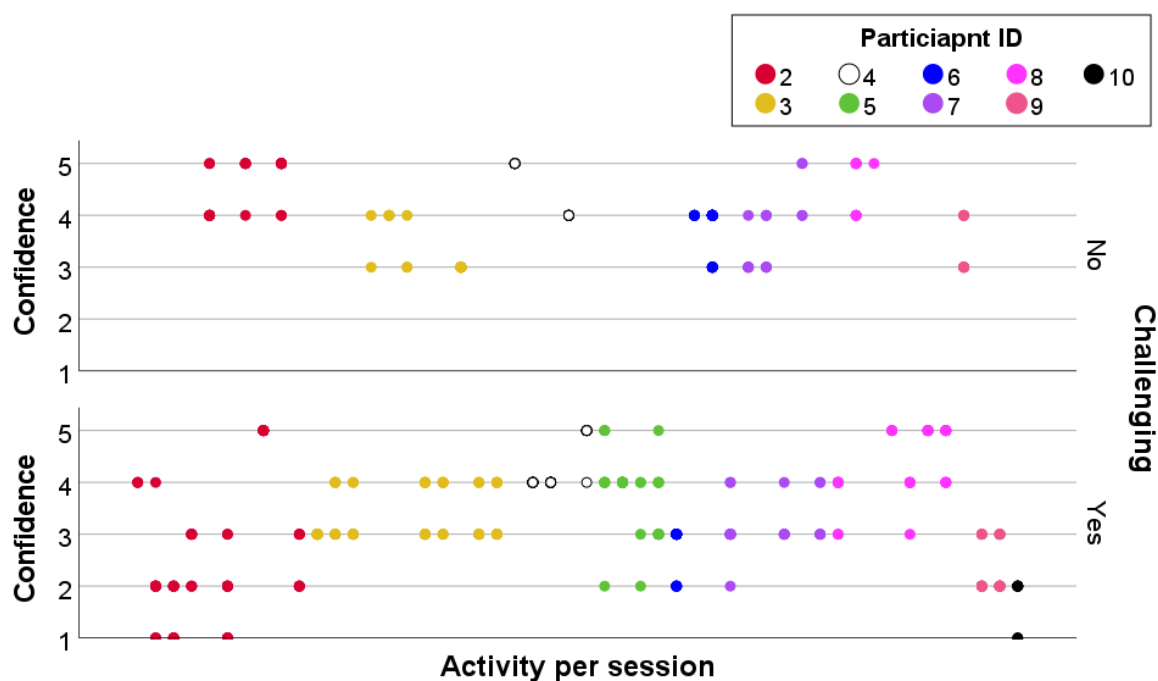Building systems that recognize how a person feels requires, or at least benefits from, having a dataset that is fully or partially labelled (i.e. ground truth to be modelled). Gathering the ground truth in ubiquitous settings is not an easy task, as it is generally not possible to simply use observers to label the data post recording as activities in everyday settings cannot be easily and/or fully captured using video cameras. Hence, self-reports are the typically used approach. However, asking people to self-report is generally considered quite demanding and disruptive. In the context of CP, the continuous focus on pain and negative states may further lead to exacerbate disengagement with physical activity [Olugbade et al. 2019]. In this study, the interview data from the EmoPain@Home dataset have been analysed alongside a dataset comprising long-distance runners (which was reported in deliverable D4.10), with a focus on the opportunities that self-reporting provided in both instances. In particular, our data highlighted how self-reporting could be obtained through a conversational agent, such as a chatbot, which could further enhance the element of companionship which participants in both studies seemed to highlight; in turn, such sense of companionship could stimulate participants to self-report beyond numerical (pain) ratings. In doing so, a conversational approach to labelling tasks could lead to more fine-grained and richer ground truth combining required ratings with further qualifiers of one's experience. Such richer self-report may lead to richer datasets annotation for machine learning while the chat-based self-reporting

becomes a useful activity to participants in longitudinal data collection activities. The following considerations summarise the content of a published paper presented at ICMI 2021 – Workshop of Automatic Assessment of Pain (Bi et al. 2021).

3.1.1   Leveraging companionship towards richer labelling and maximising subjects' benefits

First, participants in the EmoPain@Home dataset highlighted how self-reporting allowed them to have a better grasp of their present condition. In particular, they felt that the voice-based interactions underpinning self-reporting facilitated a sense of immediacy and accountability, leading to a feeling of being paid attention and cared for. Second, they often did not limit themselves to just reporting numerical/ordinal values as instructed, but rather presented small exegeses regarding the reasons behind the values provided (e.g. "my pain is at this level right now because I have just moved in this specific way"). In this sense then, the social element provided by the physical presence of a human asking participants to self-report seemed to provide important benefits in two directions. On the one hand, participants felt compelled towards providing additional and contextual cues regarding their reports, which could be leveraged by machine learning (ML) specialists towards richer ground truth and more precise labelling. For instance, such secondary, contextual and experiential data could be analysed to further validate the reliability of numeric self-report – i.e. checking for consistency and correlation between self-reported scalar values and semantic meanings of the verbalisations. They could also be used to create ML-based models that can differentiate between for example different types of pain sensations, location of pain or capture the relationship between pain and anxiety levels (Olugbade et al. 2019b; Rivas et al. 2021). On the other hand, and circularly, such more fine-grained analyses (only emerging insofar as a sense of social companionship was felt) allowed participants to more clearly benefit from self-reporting: self-reporting allowed them to engage in physical activities equipped with heightened knowledge and attentional focus. Moreover, such a stance allowed them to see self-reporting in a more positive light – not as something boring and cumbersome, but rather as helpful and stimulating.

3.1.2   From human presence to machinic presence: chatbot opportunities and challenges

From an initial analysis of the data presented above, a chatbot with self-reporting capabilities emerged as a valuable possibility to foster and heighten the sense of social companionship. First, a chatbot would retain the elements of immediacy and accountability participants said drove their perceived benefits – they felt that the fact that they were asked in the moment allowed them to be more mindful, and the fact that they had to report to someone/something listening made them more thoughtful regarding their judgements. Second, a chatbot would be able to prompt speech-based cues with a natural-sounding voice and following natural and interactive speech patterns.

Our analyses are however mindful of the fact that self-reporting cues in the EmoPain@Home study were provided by a human actor (i.e. the researcher), and thus the question of whether

or not participants would grasp the same sense of social companionship from a chatbot/prompting system is an open one. A growing body of work (e.g. Araujo et al. 2018) suggests that chatbots, if well designed, do lead to be perceived as having some form of agency, thus allowing people to engage in useful conversations regarding their issues and problems (e.g. applications for supporting the management of mental health). Our analysis in this sense pinpoints some important preliminary remarks made by participants, which pose the question of what design practices should be of paramount importance when designing a chatbot system which mimics agency. As a matter of fact, most participants seemed uncertain about replacement of the human prompts with a technical solution – citing in particular how they would not form social relationship/bond with a chatbot. More interestingly however, such dismissal of a technological replacement was attributed not to the artificial nature of a chatbot *per se*, but rather because they felt a chatbot could not efficiently engage in small talk with them, beyond merely asking them to self-report every minute. Participants in this sense recognised that the sense of companionship at the basis of the heightened benefits of self-reporting did not stem merely from the (virtual) presence of another human, but rather from relaxed non-functional chat talk between participant and researcher during the activity sessions. To clarify, these conversations were never focused on the activities at hand (e.g. "why are you hoovering this way?"), nor was it a request for further elaboration upon the self-reporting (e.g. "why is your pain higher now?"), or suggestions regarding movements (e.g. "maybe you should wash the dishes in this way"). Rather, they often engaged in small talk around trivial topics, e.g. the weather, their days, common interests, news, etc. What participants seemed unsure of then is whether or not a machine would be able to replicate and take part in such trivial conversation.

Further work will be needed to devise and design a chatbot solution that does not simply prompt participants to self-report, but which is also able to naturally engage in conversation that might go beyond self-reporting – because the perceived benefits of said system seem to strongly rely on such trivial conversational dimensions. Before that, and in order to achieve that, future studies will necessarily seek to more properly understand what would count as 'small talk' (i.e. conversations which are not cognitively taxing, as they would take away from task completion, but that still afford the formation of some kind of sociable relationship), as well as what elements of the conversations participants had with the researcher helped in the process. A critical evaluation of the data collected will be carried out in early 2022: all conversations happening during the activity phases of the sessions will be analysed and summarily clustered, in order to better understand the kinds of topics these conversations veered towards.

3.2 Observational Study with Physiotherapists (Williams et al., in prep)

People with chronic pain develop ways of moving in order to avoid or mitigate pain or anticipated strain or injury. We had previously explored such pain behaviours with machine learning (e.g. Wang et al. 2019), but here sought to enrich our understanding and achieve

better integration across movement by collecting descriptions by pain-experienced physiotherapists of videoed standard movements by people with chronic pain. We asked specifically about protective ways of moving called 'guarding', and about 'flow' of the movement, here loosely conceptualised as an activity state in which skills matched the challenge, with timeless awareness of the movement but without self-conscious control (Nakamura & Csikszentmihalyi, 2002). Both these fit the possibilities envisaged in a review by Sterling and Keefe (2021) of the opportunities offered by digital technology in investigating mechanisms associated with pain and disability. We also asked physiotherapists whether and how they could use the motion capture output from the same patients, represented in a stick figure. An overview of the aims and protocol design of the study was reported in interim stage as in deliverable D4.4. Here, we provide a description of the methods and report the findings of the study.

3.2.1   Methods

The transfer of this study from in-person plans to remote interviewing, necessitated by the onset of the coronavirus pandemic, required a change to ethics permission and imposed a delay in the start of recruiting. Seventeen of the 18 pain-specialist physiotherapists invited agreed to be interviewed remotely on Microsoft Teams, but the recording of one failed so the final sample was 16 UK physiotherapists. They had worked in chronic pain for a mean of 10 full-time years (adjusting for part-time work and career breaks); all spent at least half their time working in chronic pain currently, 12 full-time. Interviews were transcribed and independently analysed by two researchers.

Each physiotherapist observed eight video clips of people with chronic pain performing movement instances from a larger sample; the physiotherapist watched two people each for reaching forward with arms horizontal in standing position, bending down towards the toes in standing position, standing from sitting, and sitting from standing. All these movements can be challenging for people with chronic low back pain. The videos were from the EmoPain dataset (Aung et al. 2016), chosen for the quality of the video, and to provide as diverse as possible of patients and pain-related behaviours. Each video (most between 10 and 20 seconds) was played multiple times to enable in-depth description of the movement. Videos were randomly assigned across physiotherapists, so each video was described by a minimum of four and a maximum of eight physiotherapists.

Each physiotherapist first watched each video, and then was asked questions about the video, which was repeated while the physiotherapist answered questions.
1. What do you notice seeing the whole movement?
2. How would you describe it? Stiff? Slow? Guarded? Interrupted? Cautious?
3. At what point do you start to think 'this person has difficulty? has pain? [The participant watched the video and said "now" at the point where they first identified a problem.]
4. Do particular behaviours trigger that thought?

5. Does the way you see the movement change from that point, or is it just an event in the whole movement?
6. Would you describe the movement as having 'flow'?
7. What (single piece of) advice do you think would change the movement?

After completing the commentaries on the videos, physiotherapists were shown on computer screen the output of the motion capture of two of four of the patients doing the movements shown in the video (see Figure 10). Where possible, one of the displays showed the same patient doing the same movement as that physiotherapist had seen on video, otherwise selection was random. The displays could be rotated through 360 degrees. Physiotherapists were asked whether the display could be clinically useful, and their answers were recorded and transcribed.
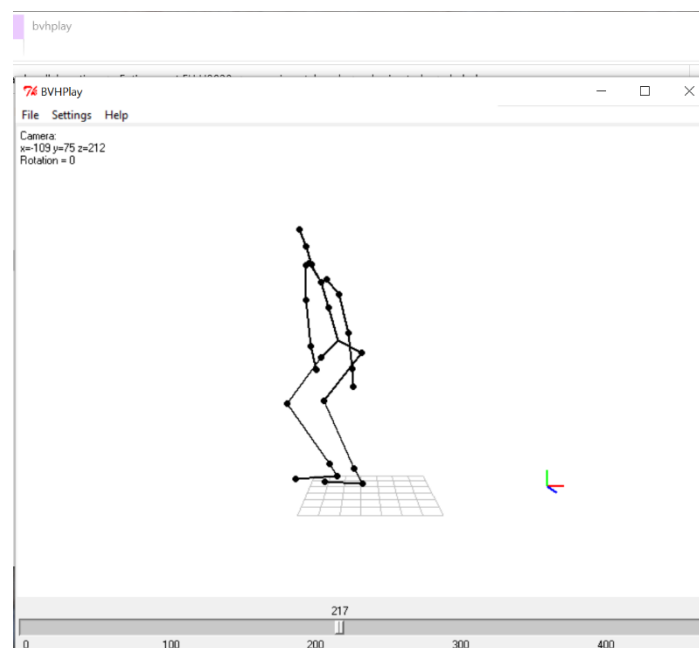
Figure 10. Still from mocap output of movement

The data were subjected to two forms of analysis. First, a grid was constructed of each patient for each movement, and the responses of the different physiotherapists to the same patient/same movement were compared. Second, the entire transcripts were analysed thematically by two independent researchers, one the interviewer, and the other not involved in the process of the study.

3.2.2   Results

The two qualitative analyses largely agreed so the two versions have been combined here.

Guarding: Guarding was answered fairly categorically, sometimes qualified as only for part of the movement, or part of the body, or emphasised, as in "very guarded". These were coded as 0 for not guarded, 1 for partial guarding (part of body or part of movement), 2 for definite guarding, and 3 for very marked guarding.

25

Agreement on guarding was not high, with four instances of complete agreement, three of them on no guarding, and one on guarding. Five further instances were rated in only two adjacent categories of rating, and seven instances in three adjacent categories. The remaining six (with four to eight physiotherapists in each) were clear disagreements.

Although agreement was moderate at best, physiotherapists were very consistent in how they described guarding so no constituent themes were extracted, but their understanding of guarding is described here. Guarding was usually distinguished from stiffness, slowness, and bracing. Although guarded movements often appeared stiff and slow, physiotherapists distinguished between stiffness due to physical condition from guarding: *"rather than him himself physically being stiff, it's the guarding that makes it look so stiff."* [PH15]

Physiotherapists often used terms such as 'cautious', 'protective', 'avoidant' in association with their opinion on guarding: *"Guarding is … people look reluctant to do it: they're cautious and apprehensive about doing it"* [PH10]. Physiotherapists often specified the part of body that was not moving as expected, most often the spine or part of the spine, and on other parts of the body that appeared to compensate for the noted restriction: *"it looks like he's quite cautious with the spine there - and that's why maybe he's compensating with his legs."* [PH5]. There was an implication of habitual or even deliberate caution, and breath-holding was reported several times as an indication that the patient viewed was focused in an anxious way on the movement.

Flow: The question for each video "does this have flow?" was answered often with a clear "yes" or "no", but also with a qualification that there was some flow, or flow for part but not all of the movement, or flow in some parts of the body but not others. These were coded as 0 for no flow, 1 for some flow, and 2 for "yes", and agreement calculated. The overall level of agreement for reach-forward was 67%, for sit-to-stand 65%, for stand-to-sit 59%, and for forward-bend 74%. There were three cases, all of *no flow*, with total agreement. Four themes were extracted from physiotherapists' descriptions.

The first theme was ***overall understanding of movement***: whether or not a movement was judged to have flow, physiotherapists contextualised the movements they observed within an understanding of the particular patient and the experimental setting. More certainty in their judgements would have required information about or from the patient, or the opportunity to observe more closely: there was almost no reference to normative movement.

The second theme, ***restriction of movement,*** emerged where there was *no flow* or *some flow*. Where there was judged to be no flow, restriction was described in terms of rigidity, stiffness, or balance; "he keeps his back quite arched and quite rigid" [PH3], or of there being very little movement in the spine or joints. Where there was some flow, comments often included contextualisation of the restriction, as in the first theme. "He definitely looks like he's trying not to… aggravate his back… he obviously can stand up without using his arms." [PH10] However, as identified above in the distinction between guarding and stiffness, rigidity was

more often implied as a protective, even deliberate, lack of movement, whereas stiffness was attributed to age or limited flexibility.

Related to the theme of restricted movement, and mainly occurring for *no flow* or *some flow*, was a third theme, ***flow as tempo***. Thus a movement performed slowly could still have flow, depending on its other qualities: "She's actually got quite, you know, quite good movement. She's just… she's slow in doing it." [PH12]. Time and timing of movements was an important dimension of judgement, but without reference to normal movement speed, rather to hesitation ("moving with hiccups" [PH13]), or uneven tempo (which was often associated with judgements of guarding), both often interpreted as unwillingness to do the movement. This theme also encompassed a few judgements concerning effort (where the movement demanded more of the patient than the patient was confident of performing), or movement performed too fast and appearing "jerky".

The last but crucial theme is **flow as natural**, occurring mainly when physiotherapists judged a movement to have flow. Flow was sometimes described in aesthetic terms ("lovely"), or as not deliberate or consciously controlled: "intuitive", "happy", "confident", "comfortable". Other descriptions used terms that implied an even speed with no hesitation or sudden change in velocity ("smooth", "fluid"), or normality of speed or less often of range ("moving freely and able to reach up and move to a point that would be expected" [PH16]). By contrast, no flow or very little flow was often described in terms of awareness, deliberate protection, guarding, consciousness of movement, conscious control of movement, or avoidance, albeit contextualised in the patient's age, build and condition. These terms did not appear where some or good flow was described. However, the first theme is pertinent here: the judgement that the movement had flow did not require faultless movement, but a movement could show some difficulty while still having a naturalness and ease that made for flow. "OK, there's also a tiny bit of an area that he is trying to avoid as he does it, but on the whole it's a much more fluid movement." [PH12]. A further element, beyond the movement itself, was that movements judged to have flow expressed confidence in ability to perform the movement, even in an adapted form. This confidence was not a self-conscious performance, but an in-the-moment intuitively guided, rather than controlled, act. Flow seemed to be about having control while allowing the body to explore. Some of the advice associated with judging a movement not to have (much) flow was just "to let go".

Mocap representation of patients' movement: Physiotherapists' answers to questions about the usefulness of the animated motion capture output, shown in an interactive presentation that they could rotate and expand, were influenced by their overall feelings about computerised visual information compared to direct clinical contact, particularly in the context of working remotely during the COVID pandemic. Some were enthusiastic – the figures were "fun" and "cool"; at the other extreme, "I hate this virtual world that we're living in right now" [PH9]. Only two of the 15 physiotherapists asked saw no advantage of the motion capture output over seeing a patient in person or on video.

The other 13 physiotherapists, while clear that the motion capture output could not substitute for the patient, could all see benefits for physiotherapist and/or for patients, from finding that it provided independent information that was clearer than on the video to feeling that it added little. Of those who said that they found extra information of value in the stick figure movement, one felt that a side view (the motion capture figure could be rotated) was a particular help, and several mentioned being able to see 'quality of movement' without distractions of the expressive person as an asset: "So absolutely you can see in this picture that there's limited flexion at the hip joints, for example. So yeah, it does give you some information… you can certainly see speed of movement, you can see order of movement within the body." [PH16]. Those physiotherapists who were less enthusiastic held that the information of most value to them came from asking the patient about his or her fears, beliefs, and goals, and from facial expression and breathing, rather than from observing the patient's movement. Benefits for patients were identified by eight physiotherapists, mostly about enabling better awareness of movement and movement patterns and habits by sharing the motion capture output with the patient, "I could see patients could be quite interested in that, because it's not threatening" [PH8], but also feedback of progress, with patients "enjoying seeing the change" [PH4]. Two physiotherapists suggested additional uses for the motion capture information in sports injury rehabilitation.

3.2.3   Discussion

Although physiotherapists were in many instances willing only to make judgements conditional on assumptions about the state of the patient in the video, being unable to interact with the patient as they would do clinically, they confidently identified both guarding and flow in the movements they observed on video.

For guarding, although estimated between-physiotherapist agreement was only moderate, their elaborations were highly consistent: guarding referred to reluctance or caution in movement, distinct from stiffness, an inference about patients' physical condition, and from slowness. Exactly how guarding intention was expressed behaviourally requires further exploration but justifies our focus on it separately from other 'pain behaviours'.

Flow also only showed modest agreement between physiotherapists, but interestingly was not incompatible with some stiffness, slowness, nor even with guarding, although the presence of guarding was usually associated with judgements of limited flow at best. Attribution of flow appeared to be a more holistic judgement of the movement at a higher level than that of guarding. The four themes extracted referred to individual context of the judgement, to restricted movement, to a smooth tempo across phases of movement, and to a naturalness and unselfconsciousness that is close to definitions and descriptions of flow in the literature. Although there was some reference to how a movement might be performed by someone without pain, judgements were not normative. It may be that in the context of chronic pain, flow for some respondents represented a relative lack of any protective behaviour that might interrupt or disrupt it, while still being slower than normal.

The relationship of flow to tempo of movement was particularly intriguing and constituted an important part of physiotherapists' judgements. Flow appeared to be characterised by the absence of sudden changes in velocity of movement, or of complete stops, although the whole envelope of some moves involved acceleration and deceleration but with seamless transitions. There had to be a harmony between different parts of the body and of the movement. The movement could be performed slowly, slower than 'normal', yet still have flow. Incompatible with flow were stops, pauses, and hesitation, often to take more control of a movement in a cautious or less confident way, suggesting anxiety either about capacity to complete the movement or to do so without exacerbating pain unduly. Further, restricted breathing appeared to be related to judgements about flow and uneven tempo, although this would bear more detailed investigation using specific sensors. Both monitoring breathing and using it consciously were recommended as ways to increase fluidity of movement, and to restore rhythmic harmony.

This study had several limitations. One was the absence of background information on the patient shown in the video, and of the possibility of closer observation that would reveal facial expression and breathing patterns sought by physiotherapists to contribute to their judgements. Wearable sensors, of course, could render filming unnecessary and operate in people's own environments and chosen activities. Quantitative judgements of guarding and flow (using numerical rating scales) would have allowed better calculation of dis/agreement, but might have disrupted the descriptions elicited from physiotherapists. A further point is that most of the participating physiotherapists were known to the interviewing author, and some were past or current colleagues; all were aware that the author was a psychologist, and this may have inhibited some types of comment and stimulated others.

How could these findings inform the design of technical systems to support people with chronic pain in moving more easily, more confidently, and with less pain and anxiety? Guarding is a possible target for further training of algorithms, since if it is related more to the individual's beliefs and expectations, consistent with many other studies in chronic pain, it needs cognitive not physical challenge. This could be supplemented by attention to breath-holding that a majority of physiotherapists looked for as a sign of anxiety about the movement. Fears and worries about movement causing pain, strain or injury are a major contributor to guarding, and a system focused only on 'correct' or normative movement will be seriously limited by failing to address these.

Work is progressing to identify whether the information provided here will provide sufficient ground truth for machine learning of guarding and of flow. It confirmed for us the importance of integrating emotional state and beliefs into understanding movement, as the physiotherapists did, although it may require more granular information than collected in this study, or confirmatory testing of provisional hypotheses on different patient videos from those used here. Emotional information is inherent in movement quality, providing an overall narrative for which flow emerges as a useful concept. It clarified for us that flow does not

represent either normal speed of movement or constant velocity, and (although physiotherapists differed on this) can even be compatible with some protective behaviour.

Experienced physiotherapists are a scarce resource compared to the prevalence of chronic disabling pain (Fayaz et al. 2016), but those trained in cognitive and behavioural methods applied to pain can provide flexible and relevant prompts and help in the moment, personalised to the patient's age and state, whether exploring patients' fears or directly facilitating physical movement. Additionally, in the predictable and safe setting of the physiotherapy clinic, patients may find it much easier to experiment and to take (what they perceive to be) risks with movement than they do when alone in their own homes. Technology can be designed to extend the support and even intervention provided by the pain-specialist physiotherapist into the patient's own environments, available at any time.

Using findings from this study, such a system might be trained to pay more attention to interruptions, hesitations, and abrupt changes in velocity of movement. It is helpful to identify here the clear association between flow and a movement having its own rhythm and harmony, rather than between flow and speed. A chatbot might prompt users to consider what fears or expectations they had about a movement, before or during it, and rather than trying to correct elements of a movement, instead suggesting modifications in order to build up to the full movement from an easier version and challenging or questioning fearful expectations. It might also pay attention to breath-holding or changes in breathing likely to be indicative of anxiety or apprehension. It could encourage exploration of movement in a fluid and even playful way.

4   Modelling Approaches for Continuous Pain Movement Data

4.1 Exploring Recurrent Neural Networks for Modelling of Protective Behaviour (Wang et al. 2021)

We show for the first time that neural networks based on long short-term memory (LSTMNN) (Hochreiter and Schmidhuber 1997; Gers et al. 1999) support activity-independent recognition of pain-related behaviour. We also explore and identify through experiments, statistical analysis and reflections on pain behaviour characteristics based on the EmoPain dataset, critical parameters/processes for data segmentation, label definition, and data augmentation. This work is published in (Wang et al. 2021b) and has informed studies reported in previous deliverables as well as further modelling reported in Section 4.2.

4.2 Further Modelling of Protective Behaviour: Integrating movement and muscle activity data (Cen 2021)

Here, we present three new architectures explored for protective behaviour. This set of architectures build on the *Hierarchical Human Activity Recognition and Protective Behaviour Detection* (HAR-PBD) architecture (Wang et al. 2021) described in deliverables D1.7

(theoretical background), D3.7 (architecture description), and D2.2 (modelling results), by integrating muscle activity data with body movement data in the PBD module of the Hierarchical HAR-PBD architecture. This work has been done within the MSc student project of Guanting Cen (Cen 2021).

4.2.1 Hierarchical Human Activity Recognition and Protective Behaviour Detection (HAR-PBD) v2 Architectures

The PBD module of the original Hierarchical HAR-PBD architecture (Hierarchical HAR-PBD v1) takes in as input data 3D positions of anatomical joints over a period t_1 to t_2 and the type of activity being performed during the period. The PBD module of the Hierarchical HAR-PBD v2 architectures include electromyography (EMG) sensor data as a third input.

Similar to the Hierarchical HAR-PBD v1, the v2 architectures are based on graphical convolution networks (GCN) (Kipf and Welling 2017) and LSTMNNs. In the PBD module of the v1, for each time t_i, each joint is represented as a node specified by the 3D position of the joint together with the activity type being performed at t_i, and all joints are nodes in a graph with an adjacency matrix defining direct connections between anatomical joints.

v2-early architecture: The first of the v2 architectures is an early fusion model where a new node representing each location of recorded muscle activity is added to the same graph, and each of these nodes is specified by a scalar value representing the activity of the muscle at time t_i. The anatomical location of the muscle determines the update to the adjacency matrix to include the corresponding muscle node. For example, where the node for a right trapezius muscle (upper back) is included in the graph, the adjacency matrix reflects connections from the right shoulder and the top spine joints to the muscle. In order to have similar dimensionality as other nodes of the graph, the muscle nodes are padded to a 3D value; the activity type is also included as the fourth dimension of the node.

v2-late architecture: The second v2 architecture is a late fusion model in which there are two different PBD modules (and so two different graphs) for the joints position data and muscle activity data respectively. The predictions for both modules are then concatenated and input to a fully connected neural network for the final classification.

v2-mid-level architecture: The third v2 architecture is a mid-level fusion model. In this architecture, the joints positions and muscle activity data have separate GCNs as well as a shared fully connect neural network. At each layer, the GCN for the joint positions data and the GCN for the muscle activity data both feed into a common network. Figure 11 illustrates the GCN structure of the PBD module of this architecture. The output for each layer i of the shared network is defined at time t as:

$$\hat{h}_t^i = \sigma\left(W^i\left(\hat{h}_t^{i-1} + \sum_{m=1}^{M} {}^m h_t^i \right) \right)$$

31

where $\sigma(.)$ is the activation function, $\hat{h}_t^i$ is the output of shared network layer i at time t, $^m h_t^i$ is the output of the ith layer of the GCN for the mth modality at time t, and W^i is a matrix of trainable weights.

In a more developed version of the architecture, the output of each shared network is fed into an attention submodule before being passed to the next shared network layer. The output of the attention submodule is specified as O_i:

$$O_t^i = A_t^i V_t^i$$

where $A_t^i = softmax\left(\frac{Q_t^i K_t^i}{\sqrt{c}}\right)$,

$$Q_t^i = h_t^i W^Q,$$

$$K_t^i = h_t^i W^K,$$

$$V_t^i = h_t^i W^V,$$

and W^Q, W^K, W^V are trainable weights.
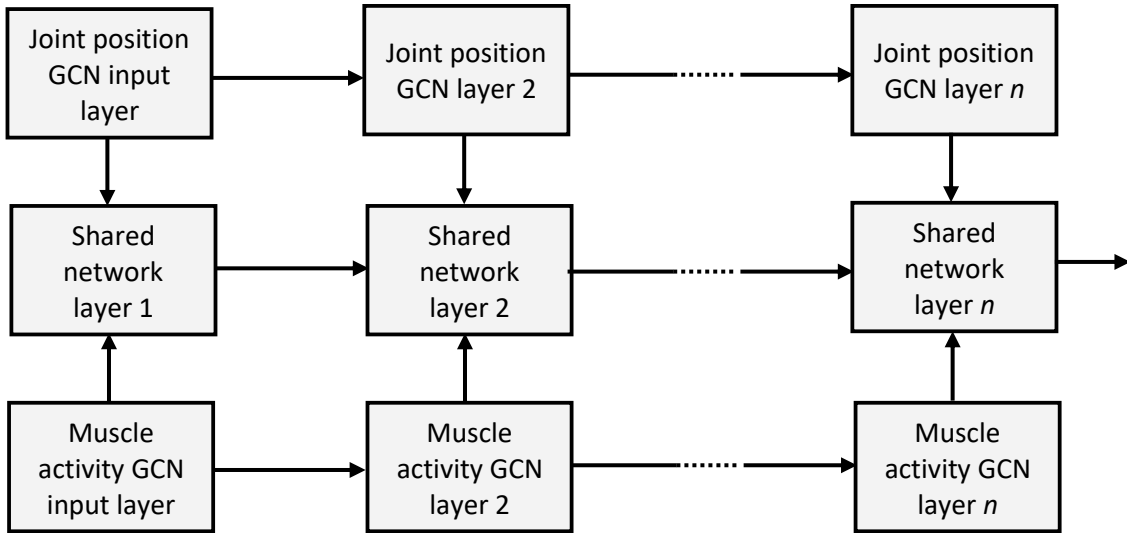
Figure 11. The GCN structure of the PBD module of the Hierarchical HAR-PBD v2-mid-level architecture

4.2.2   Dataset and data preparation

The EmoPain dataset (Aung et al. 2016) consists of 3D joints positions data and muscle activity data from four locations. The data was captured from both participants with chronic pain and healthy participants while they performed exercise movements such as sit-to-stand, but only data from people with chronic pain were used in the experiments reported here. The dataset

contains time-continuous protective behaviour annotations from clinicians. See (Wang et al. 2021) for further details about how a single protective behaviour label is derived from the annotations for six different behaviour categories: guarding, hesitation, bracing/support, abrupt action, limping, rubbing or stimulating the affected region

Each captured session in the EmoPain dataset is a sequency of activities (e.g. sit-to-stand, forward reach) for a single participant and there are multiple sessions for some participants. Similar to the approach of (Wang et al. 2021), a sliding window of length of 3s (180 frames) and 50% overlapping ratio was used to segment the sessions. A segment was labelled as *protective* if 50% or more of the frames in the segment have protective behaviour labels, and *not protective* otherwise. The same data augmentation methods used in (Wang et al. 2021) and additionally reported in deliverable D2.2, i.e. random dropout and Gaussian noise, were used. This resulted in roughly 35,000 segments in total.

Since the difference between the v1 model and the v2 architectures is the PBD modules, the HAR module for predicting the activity type was simulated using the ground truth activity label in the experiments reported here.

4.2.3   Parameters

A leave-one-subject-out cross-validation (LOSOCV) method was used to understand how well each model generalised to unseen subjects.

For the *v2-late* and *v2-mid-level* models where separate GCNs are used for the joint positions and muscle activity data, the GCN for the joint positions data is made up of 3 layers each with 16 units while the GCN for the muscle activity data has 3 layers with 6 units each.

For the *v2-early* and *v2-mid-level* models that have a single LSTMNN for both types of data, the LSTMNN is made up of 3 LSTM layers each with 24 units, whereas for the *v2-late* model in which the two types of data have separate LSTMNNs, both LSTMNNs have 3 layers but while the joint positions LSTM layers have 24 units each, the muscle activity LSTM layers have 8 units each.

Each of the three v2 models was trained for 100 epochs. The same loss function described in (Wang et al. 2021) was used to address class imbalance.

4.2.4   Results

Table 2 shows the performance of the three v2 models for protective behaviour detection. We further include the performance of a compatible v1 model where the HAR is also simulated using the ground truth activity labels in the dataset. The results show that integrating the muscle activity data with the joint positions data improves the performance of the Hierarchical HAR-PBD architecture; however, this is only true when mid-level fusion is done. Further statistical analysis suggests that the difference in performance between the v1 and v2-mid-level models is statistically significant, $p<0.05$ ($F=7.477$ and $p<0.001$ for difference

between means across the four models, and $p<0.0001$ for Bonferroni-corrected pairwise comparisons between the v2-mid-level and v1 models). Difference between the v2-mid-level and v2-early models was also statistically significant ($p<0.001$). A late-fusion is only marginally better than use of the joint positions data alone, while early fusion does not have any advantage over the v1 model based on only joints positions data. For the mid-level fusion, performance drops when attention mechanism is included in fusing the joint positions and muscle activity embeddings, although it still outperforms the v1, v2-early, and v2-late models.

Table 2. HAR-PBD v2 Results

Hierarchical HAR-PBD Model	Accuracy	F1 Score
v1 [Wang et al. 2021]	0.89	0.82
v2-early	0.89	0.82
v2-late	0.90	0.83
v2-mid-level	0.93	0.88
v2-mid-level with attention	0.92	0.86

4.3 Preliminary Modelling of Pain Levels in the EmoPain@Home Dataset

4.3.1   Feature Extraction

We extracted 6 sets of features (see Table 3) based on previous work on automatic detection of pain levels, mood, and confidence in (Olugbade et al. 2018, 2019): speed, jerk, energy, amount of movement, minimum distance between joints, and range of joint angle:

— For speed, jerk, energy, and amount of movement features, we further extracted features per anatomical joint (6 joints X 4 features = 24)

— The minimum distance feature was on the other computed between each joint and the forearm as we wanted to capture self-adaptor behaviour (i.e. self-touching). We excluded the minimum distance for the (ipsilateral) upper arm as this is expected to be constant since they shared the same bone and bones are rigid objects (4 joints X 1 feature = 4)

— For the range of joint angle, we computed the range for the knee and hip angles (2 joints X 1 feature = 2), with

   o   the knee angle defined as the angle between the vector through the lower trunk and hip joints, and the vector through the hip and ankle joints, and

   o   the hip angle defined as the angle between the vector through the hip and knee joints, and the vector through the knee and ankle joints

For each data instance m_i^k (see Section 2.3), we extracted features from two windows:

- window c_i^k between time $t = t_{i_k} - 1$ and $t = t_{i_k}$, which represents the time window of the data instance, and
- window cc_i^k between time $t = 0$ and $t = t_{i_k}$, which represents the cumulative time from the beginning of the activity

To extract features, we divided each window into 3 non-overlapping segments, and then computed the features on each segment. Thus, each data instance is represented as a 180-dimension vector (3 segments X 2 windows X 24+4+2 features).

Table 3. Features extracted from data instances in the EmoPain@Home dataset

Feature	Description
Speed	3D speed of a given joint
Jerk	3D jerk (i.e. acceleration over time) of a given joint, to capture the smoothness of movement of the joint
Energy	3D kinetic energy of a given joint (assuming unit mass)
Amount of movement	The sum of the 3D displacement of a given joint over time
Minimum distance between joints	The minimum 3D displacement between a given joint and the forearm joint, to capture self-adaptor behaviour
Range of joint angle	The range of angles for a given joint

4.3.2   Methods

We explored automatic classification of pain into two levels: low level pain (i.e. pain less than or equal to 5 on the pain scale defined in Section 2.1.3) and high level pain (i.e. pain reported to be higher than 5). More fine-grained levels of pain could not be explored due to the limited size of the data, which is typical of annotated data captured in the wild from specific population groups, that is, beyond the general population.

Evaluation approach: Although the LOSOCV evaluation approach used in the study reported in Section 4.2 is standard for assessing the ability of an automatic detection model to generalise to unseen subjects, as highlighted in Section 2.3.1 the differences in activity types across participants is expected to be significant challenge in generalising to unseen subjects for the EmoPain@Home dataset. Indeed, initial experiments based on LOSOCV confirmed

this. Thus, we followed the leave-one-out cross-validation (LOOCV) which is the gold standard when the aim is not to evaluate generalisation to new subjects.

Learning algorithm: Given the high dimensionality of the feature set (n=180, see Section 4.3.1 for details) relative to the data size, we explored the use of the Random Subspaces ensemble method (Ho 1998) where a random subset of the features are used to train weak learners and the prediction is based on an aggregate of the predictions of these learners. We used decision trees as the learners, and we set the features to be selected with replacement. We explored c=5, 10, 50, and 100 trees using a nested LOOCV.

4.3.3   Results

Table 4 shows the results for the pain level classification comparing four different maximum number of features to build each tree in the bagging classifier: n, n/2, n/5, and n/10. The classification performance was very high in all four cases, but as expected, the use of n (i.e. 180) features performed the worst due to the high dimensionality. Although n/2 (i.e. 90) and n/10 (i.e. 18) showed slightly better performance, n/5 (i.e. 36) led to the highest F1 scores suggesting that it is indeed valuable to minimise the dimensionality of the input data although there is a bound beyond which this leads to diminishing returns. Table 5, Table 6, Table 7, and Table 8 present the confusion matrices for the 4 different maximum feature sizes.

Table 4. Pain level classification results

Maximum number of features to build each tree	F1 score (low level pain)	F1 score (high level pain)
n = 180	0.88	0.86
n/2 = 90	0.88	0.88
n/5 = 36	0.90	0.90
n/10 = 18	0.88	0.88

Table 5. Pain level classification results for maximum number of features per tree = n

		PREDICTED	
		Low level pain	High level pain
TRUE LABEL	Low level pain	**99**	10
	High level pain	16	**101**

Table 6. Pain level classification results for maximum number of features per tree = n/2

		PREDICTED	
		Low level pain	High level pain
TRUE LABEL	Low level pain	**96**	13
	High level pain	14	**103**

Table 7. Pain level classification results for maximum number of features per tree = n/5

		PREDICTED	
		Low level pain	High level pain
TRUE LABEL	Low level pain	**99**	10
	High level pain	13	**104**

Table 8. Pain level classification results for maximum number of features per tree = n/10

		PREDICTED	
		Low level pain	High level pain
TRUE LABEL	Low level pain	**96**	13
	High level pain	14	**103**

Future work will explore the relevance of each of the n features for pain classification. Further, we will consider other learning algorithms beyond the decision-tree based bagging classifier. We will additionally investigate the automatic classification of the other two labels in the EmoPain@Home dataset, i.e. worry and confidence.

5 References

Araujo, T., 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Computers in Human Behavior, 85, pp.183-189.

Aung MS, Kaltwang S, Romera-Paredes B, Martinez B, Singh A, Cella M, Valstar M, Meng H, Kemp A, Shafizadeh M, Elkins AC. The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal EmoPain dataset. IEEE Transactions on Affective Computing 2016;7(4):435-451.

Bi T, Buono RA, Olugbade T, Singh A, Holloway C, Costanza E, Williams ACdeC, Gold NE, Berthouze N. Towards Chatbot-Supported Self-Reporting for Increased Reliability and Richness of Ground Truth for Automatic Pain Recognition: Reflections on Long-Distance Runners and People with Chronic Pain. In the Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21). October 18-22, 2021. Montréal, QC, Canada. ACM, New York, NY, USA.

Braun, V., Clarke, V., 2012. Thematic analysis. In Cooper, H., Camic, P.M., Long, D.L., Panter, A.T., Rindskopf, D., Sher, K.J. (eds.) APA Handbook of Research Methods in Psychology, Research Designs, vol. 2, pp. 57–71. American Psychological Association, Washington

Braun, V., Clarke, V., 2014: Thematic *analysis*. In Teo, T. (ed.) Encyclopedia of Critical Psychology, pp. 1947–1952. Springer, New York

D'Ostilio, K. and Garraux, G., 2012. Brain mechanisms underlying automatic and unconscious control of motor action. Frontiers in Human Neuroscience, 6.

Cen, Guanting. 2021. Exploring Multimodal Fusion for Protective Behavior Detection in Continuous Data. MSc Disserttion, University College London.

Fayaz A, Croft P, Langford RM, Donaldson LJ, Jones GT. Prevalence of chronic pain in the UK: a systematic review and meta-analysis of population studies. BMJ 2016;6:e010364.

Gers, F., Schmidhuber, J. and Cummins, F. (1999). Learning to forget: continual prediction with LSTM. In 1999 Ninth International Conference on Artificial NeuralNetworks ICANN 99. (Conf. Publ. No. 470), Vol. 2. 850–855 vol.2. https://doi.org/10.1049/cp:19991218.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), pp.832-844.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory.Neural Comput.9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Joffe, H., & Yardley, L., 2004. Content and thematic analysis. In D. Marks & L. Yardley (eds.). Research methods for clinical and health psychology. London: Sage.

Keefe, F.J. and Block, A.R., 1982. Development of an observation method for assessing pain behavior in chronic low back pain patients. Behavior therapy.

Kipf, T., and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In 5th International Conference on Learning Representations (ICLR).

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (Readiness-Potential): The unconscious initiation of a freely voluntary act. Brain, 106, 623–642.

Mason O, Hargreaves I. A qualitative study of mindfulness-based cognitive therapy for depression. Br J Med Psychol. 2001 Jun;74(Pt 2):197-212.

Nakamura J, Csikszentmihalyi M. The concept of flow. In: Snyder CR, Lopez SJ (eds.), Handbook of Positive Psychology. Oxford, OUP, 2002, p89-105.

Nielson, Warren R, Jensen, Mark P, Karsdorp, Petra A, Vlaeyen, Johannes W.S. (2013). Activity Pacing in Chronic Pain, The Clinical Journal of Pain: Volume 29 - Issue 5 - p 461-468

Olugbade, T. A., Bianchi-Berthouze, N., Marquardt, N., & Williams, A. C. (2018). Human Observer and Automatic Assessment of Movement Related Self-Efficacy in Chronic Pain: from Exercise to Functional Activity. IEEE Transaction on Affective Computing, 11(2), 214–229.

Olugbade, T. A., Singh, A., Bianchi-Berthouze, N., Marquardt, N., Aung, M. S. H., & Williams, A. (2019). How Can Affect Be Detected and Represented in Technological Support for Physical Rehabilitation? Transactions on Computer-Human Interaction, 26(1), 1–29.

Olugbade, T., Berthouze, N., & Williams, A. (2019b). The relationship between guarding, pain and emotion. Pain Reports. 494)

Rivas, J.J., del Carmen Lara, M., Castrejon, L., Hernandez-Franco, J., Orihuela-Espina, F., Palafox, L., Williams, A., Berthouze, N. and Sucar, E.. (2021). Multi-label and multimodal classifier for affectve states recognition in virtual rehabilitation. IEEE Transactions on Affective Computing

Singh, A., Bianchi-Berthouze, N. and Williams, A.C., 2017, May. Supporting everyday function in chronic pain using wearable technology. In Proceedings of the 2017 CHI Conference on human factors in computing systems (pp. 3903-3915).

Sterling M, Keefe FJ. Physical rehabilitation research and pain science. PAIN 2021;162(11):2621-4.

Wang C, Olugbade TA, Mathur A, Williams ACdeC, Lane ND, Bianchi-Berthouze N. Recurrent network based automatic detection of chronic pain protective behavior using mocap and sEMG data. In Proceedings of the 23rd International Symposium on Wearable Computers 2019;Sept:225-230.

Wang, C., Williams, A., Lane, N., and Bianchi-Berthouze, N. (2021). "Leveraging Activity Recognition to Enable Protective Behavior Detection in Continuous Data." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5 (2): 1–24.

Wang, C., Olugbade, T. A., Mathur, A., Williams, A. C. D. C., Lane, N. D., & Bianchi-Berthouze, N. (2021b). Chronic-Pain Protective Behavior Detection with Deep Learning. ACM Transactions on Computing for Healthcare. doi:10.1145/344906