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

Tao Bi
University College London, t.bi@ucl.ac.uk

Raffaele Andrea Buono
University College London, raffaele.buono.18@ucl.ac.uk

Temitayo Olugbade
University College London, temitayo.olugbade.13@ucl.ac.uk

Aneesha Singh
University College London, aneesha.singh@ucl.ac.uk

Catherine Holloway
University College London, c.holloway@ucl.ac.uk

Enrico Costanza
University College London, e.costanza@ucl.ac.uk

Amanda C de C Williams
University College London, amanda.williams@ucl.ac.uk

Nicolas E. Gold
University College London, n.gold@ucl.ac.uk

Nadia Berthouze
University College London, nadia.berthouze@ucl.ac.uk

Pain is a ubiquitous and multifaceted experience, making the gathering of ground truth for training machine learning system particularly difficult. In this paper, we reflect on the use of voice-based Experience Sampling Method (ESM) approaches for collecting pain self-reports in two different real-life case studies: long-distance runners, and people living with chronic pain performing housework activities. We report on the reflections emerging from these two qualitative studies in which semi-structured interviews were used to exploratively gather initial insights on how voice-based ESM could affect the collection of self-reports as ground truth. While frequent ESM questions may be considered intrusive, most of our participants found them useful, and even welcomed those question prompts. Particularly, they found that such voice-based questions facilitated in-the-moment self-reflection, and stimulated a sense of companionship leading to richer self-reporting, and possibly more reliable ground truth. We will discuss the ways in which participants benefitted from subjective self-reporting leading to an increased awareness and self-understanding. In addition, we make the case for the possibility of building a chatbot with ESM capabilities in order to gather more enhanced, refined but structured ground truth that combines pain ratings and their qualification. Such rich ground truth can provide could be seen as more reliable, as well as contributing to more refined machine learning models able to better capture the complexity of pain experience.

CCS CONCEPTS • **Human-centered computing • Human computer interaction (HCI) • Empirical studies in HCI**

Additional Keywords and Phrases: Pain, Exertion, Self-reporting, Chatbot, Experience Sampling Method, Voice-based ESM, Ground Truth, Machine Learning, Automatic Recognition System for Pain, Affective Computing, Distance-Running, Chronic Pain

ACM Reference Format:

Tao Bi, Raffaele A. Buono, Temitayo Olugbade, Aneesha Singh, Catherine Holloway, Enrico Costanza, Amanda C de C Williams, Nicolas E. Gold, Nadia Berthouze. 2021. 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 Pai. In *the Proceedings of the 2021*

International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Candada. ACM, New York, NY, USA, 10 pages.

1 Introduction

Pain can be ubiquitous, emerging at any point or context in daily life. It is heavily intertwined with physical activities and exertion, but also emotions, mood and behaviours (as well as social elements of daily routine) can influence pain appraisal. For instance, long-distance runners can perceive heightened levels of pain as either a boost to their stamina (pain as a form of pride), or as an obstacle towards completion of the track (pain as developing worries) [1]. Similarly, people with chronic low back pain may develop fear of housework chores, due to anticipating pain increase [2]. Such appraisals of pain are thus highly dependent on other factors beyond simply assessed pain levels [2-5]. Aside from the inherent complexity of analysing pain experience given this multifaceted nature, standard observational techniques and external data collection practices [6] might be, under specific circumstances, simply not practical, or at least sub-optimal. Aside from the issue of observation and subsequent labelling/annotation being time and resource intensive, our two case studies exemplify how these methodological approaches might be limited or not applicable in certain instances: on the one hand, runners cannot be physically observed in-situ, unless operating in specific scenarios (e.g. treadmill); on the other hand, the ubiquity of pain in people with chronic low back pain means that standard observational settings might not shed light to more mundane scenarios where presence of researchers for the purpose of observation and/or recording might not be a feasible option: participants might carry out activities at 'non-social' hours, or in spatial contexts that might be radically different from the ones observed in clinical settings.

In such contexts where observation might be impractical, *in-situ* (i.e. in-the-moment) self-reporting [7] is a standard practice, and a more *efficient* data collection approach. This experience capture method allows participants to elaborate upon their experiences across a temporal and spatial axis. Moreover, in-situ approaches attempt to avoid potential risks of *recalling bias* which might be prominent in other forms of self-reporting (e.g. post-activity self-reporting), as well as attempting to provide more fine-grained data given the shorter gap between experience and prompt. In-situ self-reporting can also be conceived in a way as to understand pain experiences in a nuanced and multi-faceted manner – allowing participants to elaborate and assess upon different contextual elements which might play a part in their understanding of pain (e.g. confidence, emotional state, exertion, etc.).

One drawback to in-situ self-reporting has been the nuisance and intrusion it represents, which has led to the use of sparse self-reporting events. We argue that for certain situations, instead such approach may lead to a very fine grained (e.g., even every minute) and possibly more insightful ground truth, if the approach is well designed. However, precisely how to design such approach for continuous use in everyday life scenarios is still an open question. This is the question we have started to address in this paper. Here we show that ESM-based self-reporting not only provided high-quality data regarding pain experiences in an efficient and feasible manner for machine learning researchers. In addition, in-situ self-reporting has also proved a particularly *useful* instrument of self-discovery for participants themselves. Such engagement with self-reporting and such increased awareness of one's pain and body may in return further contribute to reliability and richness of the ground truth for building the recognition systems.

This paper is developed as a series of reflections on our prompting of research participants for in-the-moment, subjective self-reports of pain and several other subjective experiences. In particular, it is based upon two main research contexts: 1) people performing long-distance running; 2) people living with chronic lower back pain performing daily functional activities in an indoor home context. We consider these case studies to be particularly relevant since previous research showed that runners were particularly receptive to being assisted by an automatic recognition systems which could aid them in understanding their overall physical condition, as well as allowing them to share live such experiential data with spectators and coaches alike, in order to receive better and more personalised support during a race [1]. The case of runners can also be considered representative for other sports and physical performances. Similarly, it has been argued that people who live with chronic pain could benefit from automatic recognition systems which could equip them with better pain management skills and insights to be deployed in their daily life [2, 4]. In addition, pain self-management is required in many long-term conditions (e.g., post-stroke, multiple sclerosis). In these two studies, we took a fully qualitative approach by conducting in-depth semi-structured interviews. Through these, we sought to explore how participants perceive and understand the dialogical mechanism underpinning voice-based approaches to ESM, as well as how such approaches could affect behaviours and the self-reporting of ground truth labels. Additionally, such insight-heavy interviews allowed us to preliminarily enquire

regarding the factors to be considered when designing voice-based prompting systems. In this sense then, the current paper is solely focused around better understanding and positioning the insights on potentially enhanced ground truth labels and reported benefits provided by the participants, rather than quantitatively evaluating the degree to which such benefits led to improvements of self-reported ground truth data: while we consider such analysis necessary and of paramount importance, it is outside of the scope of this paper.

Building such pain automatic recognition system requires ground truth labels of people's pain state to train the machine learning model: collection of such ground truth labels remains a challenge in the field of affective computing and machine learning [8, 9]. As mentioned above, such ground truth collection is *particularly* arduous in ubiquitous contexts such these ones, making the issue of exploring and thinking through different data collection techniques even more pressing.

Despite the differences in activity type, population group, and physical setting, preliminary analyses of captured data, as well as participant feedback in both studies, suggest that participants perceived important and often unexpected benefits from self-reporting their levels of pain at set intervals: these ranged from improved self-awareness, enhanced understanding of the experience, better understanding of the compositionality of movements, as well as a heightened understanding of how other facets of pain experience (e.g. worries, confidence, exertion, etc.) influence variation in pain and its understanding. In short, in-situ self-reporting stimulated richer and more profound *self-reflection* on which participants could act to better understand their pain and the activities performed. Furthermore, in the long-distance running case study, several participants stated that the computer-generated prompt for self-reporting provided them with a sense of *social companionship* often lacking while running. Such insight informed our design of the self-reporting system for the chronic pain case study, and also prompted preliminary discussions regarding the possibility of constructing a chatbot that could leverage such feelings of companionship while collecting self-reporting data.

In this paper we will first present participants' discussions of the benefits they found in in-the-moment self-reporting using a voice-based ESM system. We then further provide a preliminary discussion considering the following: 1) how a tailored and context-dependent approach to designing self-reporting systems could maximise retention of self-reporting and perceived benefits of it for participants, which would in turn translate into more representative and fine-grained data for researchers to work on; 2) how the sense of companionship highlighted by participants could be further integrated and enhanced within a technical system, such as a chatbot, which could maintain the benefits stemming from ESM-like capabilities, while also extending the benefits participants felt from interacting with a human researcher; 3) how subjective reflections stemming from self-reporting could contribute towards more precise ground truth labels enriched by qualifiers for building systems of pain recognition and assessment.

2 Study design

Within both studies, we based self-report design upon the Experience Sampling Method (ESM), understood as a longitudinal research method to gather participants' feelings in the moment at set intervals [10]. We adapted our implementation of the ESM to be attentive to the specific contextual needs of our studies. While many existing ESM systems (e.g., [11-14], see a review in [10]) require people to physically interact with a device to self-report via their hands, such a solution would not be optimal within our scenarios, given that maintaining mobility and free hands are a priority, so as not to disrupt the flow of ongoing physical activities. For this reason, in these studies we sought to explore the use of a voice-based ESM. This novel approach thus affords minimally invasive interaction in terms of bodily movements and gestures needed to successfully interact with the system. In addition, we have extended the ESM design by combining typical self-reporting rating questions with think-aloud descriptions of experience. While participants in the first study were openly invited to talk back to the system to provide additional details, participants in the second study were not explicitly asked to, but were not stopped by the researcher if and when they started to do so. As mentioned previously, we engaged in in-depth interviews in order to first understand the benefits participants perceived from self-reporting, as well as to explore how a voice-based self-reporting system could impact the quality of ground truth data gathered. At the current stage, the research remains insights-oriented, and we thus focused on a smaller sample size: we understand that such limited number of participants cannot provide any robust and objective direction to make claims. Nonetheless, we believe such insights are important starting points to both rethink design imperatives, as well as to further advance our research.

2.1 Study 1

The study with long-distance runners was conducted first. The study involved 11 runners (five males, six females) who often run either more than 5 km, or for at least 30 minutes. We selected this distance/time lower bound as it allowed us to prompt participants to self-report multiple times. Their running location varied – outdoor track (3), indoor treadmill (6), outdoor park (3). The voice-based ESM system was developed as an Android application. During the run, participants wore smart earbuds, and the phone was attached to a waistband. The self-reports were captured by the microphone of the earbuds, and then stored on the application.

Every minute runners were prompted to self-report by a computer-generated voice (built using TtS software). The prompt we used was the following question: "What is your level of pain, exertion, desire to stop, and your emotional valence?". Beforehand, we trained the participants to verbally rate these states: pain, exertion, desire to stop with ascending numerical values between 1 and 5 (1 means not at all; 5 means maximum level); and emotional valence on an ordinal scale comprising three points: {negative, neutral, positive}. In addition, between prompted self-report ratings, we invited participants, if they wished, to think aloud as if talking to the ESM system. We additionally collected behavioural and physiological data using sensors: head motion data captured with accelerometers embedded in the smart earbuds, blood volume pulse and electrodermal activity signals based on a smart bracelet, and foot pressure data using smart insoles.

We then interviewed them, using a semi-structured approach, about their experience of using and interacting with the voice-based ESM system during their run. An initial thematic analysis [15, 16] of the interviews revealed that participants enjoyed the conversational element provided by the ESM system, and in fact desired that the system more closely mimics human-like conversations. To start to explore the possibility of this, in the second study, we replaced the computer-generated prompting system with the human researcher delivering the ESM prompts.

2.2 Study 2

As noted above, the ESM design in the second study was based on preliminary findings in the first. The second study involved 10 participants (5 male, 5 female) who self-identified as living with musculoskeletal chronic pain involving the low back. These participants were asked to perform indoor household activities that they would normally do as part of their daily routines, while observed remotely by the researcher. Activities lasted between 10 and 30 minutes, and included chores as varied as cleaning surfaces, hoovering floors, painting walls, washing dishes, loading/unloading washing machines.

Participants were prompted every minute by the researcher. To avoid disruptions, the question **"**What is your level of pain, worry, and confidence in tasks at this moment in time**?"** was replaced by a shorter word, *'time'*. *Pain* and *worry* were assessed on an ascending numerical scale from 0 to 10 where 0 = no pain and 10 very severe; *confidence* was assessed using an ordinal scale of {no confidence, less than average confidence, average confidence, more than average confidence, max confidence}. We also collected movement data while the participants performed activities. For this purpose, they wore a set of 6 Notch motion capture sensors, positioned on the right wrist, right upper arm, chest, waist, right thigh, right calf. We also video-recorded all activities for further movement analysis and annotation.

Following the activity, participants took part in a semi-structured interview. As with the first study, this was to gain insight into their understanding of the constructs that they were reporting on, as well as their opinions on the ESM. Building upon the first study, we additionally enquired about the perceived benefits of a dialogical ESM approach, as well as about how they thought technology could help them manage their chronic pain condition. The interview data were analysed following several steps of thematic analysis [15, 16].

As mentioned above, the most salient difference between the two studies was regarding what was prompting participants to self-report – with the first study relying on an automated system, and the second one on human-generated prompts. Furthermore, in the second study, the researcher often engaged in short conversations in between self-reports, either by responding to participants' comments, or by making situational remarks. We will further comment on the importance of this later on, but it is relevant to remark here that this was just small-talk about participants' hobbies, how the day was going, observable memorabilia around the environment, etc.

In this paper, we do not focus on the sensor data, but on understanding how a valid and rich ground truth can be gathered by participants themselves at the same time as the data tracked by sensors. Hence, thematic analysis was used to analyse the interview data and the think-aloud data gathered during the data collection. The dataset is not the focus of this paper.

3 Findings

In this section, we will be presenting some of the most salient findings regarding the benefits participants felt that they gained from reporting their state while carrying out physical activities, *in spite of* the voice-based ESM initially being developed just for the purpose of collecting data useful to the researchers. We then reflect the effect that the perceived benefits have on the level of granularity and reliability of the ground truth for building pain recognition systems. We also provide new insights on how a voice-based ESM should be designed to enhance such outcomes. We report these findings under four main themes below. In reporting the direct quotes from participants, RP# will be used to indicate runner participants and CP# for the participants living with chronic pain.

3.1 No negative reaction to fine-grain self-reporting.

It is worth first pointing out that *no negative reaction* was observed in the participants in attending to and verbally elaborating upon their pain experience so frequently [17]. As far as runners are concerned, they saw the self-reporting as a *positive distraction* (from the default negative reaction to pain), rather than a negative obsession (over possibly heightened pain):

> "It's funny, because I thought that they [i.e. self-reporting prompts] might make me engage more with the pain. […] But actually, [...] if I hadn't been listening to anything at all, I would have been more like 'oh, when can I stop [running]? When can I get out of the pain?'. While actually listening to the questions was the distraction from such negative thoughts." (RP005, Female, 22)

Similarly, none of the participants living with chronic pain felt that thinking about pain *while* in pain hindered their ability to carry out activities. Quite the contrary, participants felt that reporting about pain *in the moment* allowed them to implement previously learned thought patterns more successfully. CP003 (Male, 56) for example reported that the days in which pain is more manageable are those when he is able to acknowledge the pain, talking to himself and say "hello pain, I know you're here. You're welcome in". Self-reporting allowed him to extend this attitude to the period *while* he was doing house chores. Other participants were initially more sceptical about the ESM approach, only to then see benefits in their use of it. CP009 (Female, 59) for instance said at the end of the study:

> "To be honest, at first I wanted to ask you [the researcher] not to do the assessment thing. I was worried about constantly referring back to my pain […]. But actually, it's never been a paralysing thought: even when the self-reporting made me aware that things were bad… even that allowed me to think 'okay, where is the bad coming from?' […] I believe I could actually make some positive changes through these assessments and awareness."

It is worth noting that chronic pain study participants all seemed aware of the underlying dangers of negative reactivity – i.e. self-reporting making them hyper-aware of pain. They all highlighted how self-reporting could be a useful instrument for self-reflection only insofar as one already inhabits a specific detached mindset where pain is acknowledged and made amenable to observation, without allowing it to take over. It is thus important to stress how self-reporting should only be considered as one of the many strategies within a larger and more complex pain management toolkit [17].

3.2 Self-reporting triggering self-reflection.

It is precisely this idea of self-report triggering ***self-reflection*** outlined by CP009 that participants in general found as the most beneficial aspect of the self-reporting exercise. Particularly, self-reflection was articulated according to 4 main parameters: a. awareness of one's body and movements; b. temporal and comparative understanding of pain and its fluctuations; c. deeper understanding of the relationship between pain and the other experiences self-reported; d. ability to better verbalise such acquired knowledge.

3.2.1 Awareness of one's body and movements

As far as the topic of knowledge of pain and movement is concerned, participants involved in the running study brought up the fact that self-reporting translated into a heightened attentional focus on the present moment, with particular regards to body movements and bodily sensations. In such a state, they felt it was easier for strategies of self-regulation and self-intervention to trigger. For instance, RP005 explained that without a self-reporting prompting system her

thoughts would wander off, losing focus and therefore making it harder to intervene on one's performance. In this sense, RP003 (Female, 28) emphasised how the self-reporting prompts helped her in constantly referring back to the body, to how it moves and what it needs:

> "Maybe you're thinking something completely different… Then, at some point, I had that input [from the ESM], and I start to think 'okay, let's check my posture. I feel like my shoulders are a bit tight. So, I need to release and relax the shoulders and the arms'. I was saying that out loud. […] So [through self-reporting] I will make sure that I'm running in a very economical way." [RP003]

Similarly, many of the participants in the chronic pain study shared that self-reporting became a tool that equipped them with the awareness and knowledge to subjectively decompose a given activity, and reflect upon how each movement flowed throughout their body. CP003 cogently described referred this as the capacity to see movements in their *granularity*, understanding how at each given time-point specific gestures have a direct and immediate impact on one's understanding of their body and its condition. CP007 (27, Male) articulates the positive benefits of such granular insights as follows:

> "One thing I noticed when you were making me do the self-reporting, it made me realise how I normally… when I do tasks, I do them in a sort of very blurred state. I'm not aware of what pain is being caused by, I'm not aware of the movements I'm doing, how I'm doing them, how I feel about them and what they do to me. But [by self-reporting] I started noticing how the specific movements would cause pain in my body. And more than noticing it, I could feel it. This is something I am not able to pay attention to most of the time. If someone said to me now 'check your body right now, where is the pain and where does it come from?' – I could tell you […]. But a lot of the time I'm completely disconnected and I wouldn't be able to." [CP007]

Such granular understanding seems to have allowed participants to disentangle themselves from the idea that an activity necessarily generates pain, and might therefore become something to be indefinitely put off. Rather, self-reflection through self-reporting allowed them to start to unpack movements as sequential and granular. Such an understanding was particularly prominent in the reflections CP004 (48, Female) shared regarding cleaning the bathroom. She felt that she tends to find the activity challenging, and often puts it off as much as she can. While doing the activity as part of the study, she found the task much less daunting. She explained her belief that the self-reporting allowed her to constantly check that her pain, worries and confidence were at levels she deemed 'okay' for the vast majority of the activity. In this sense then, CP004 thought that self-reporting functioned as a way for her to step out of the resolute thought pattern in which cleaning the bathroom (as an unspecified, unscrutinised series of movements) is perceived as painful, daunting and to be avoided. It equipped her with the knowledge that specific movements and habits within the activity cause her elevated levels of worries and pain. This was tangible and personalised knowledge that she felt she could act upon: for instance, she mentioned she could discuss with a therapist how to tackle (physically and mentally) the particular movements associated with cleaning the tiles, as well as more radical solutions such as retiling the room.

3.2.2 Temporal and comparative understanding of pain and its fluctuations

Participants found self-reporting useful not just in terms of gaining in-the-moment awareness. They also stressed that it facilitated a comparative process to understand the temporal unfolding of pain. This was particularly pronounced among runners, with many of them stating that self-reporting helped them to be more aware of the transitory nature of pain. Some of them (RP008, Male, 26; RP011, Female, 28) stated that they were able to mentally develop a sort of visualised graph that allowed them to record, store and recall the fluctuations of their pain status. Such a mental exercise could help them have a more insightful and personalised perspective on their running experience, as well as allowing them to better evaluate their performance. In this sense, self-reporting functioned as a way for them to think about their conditions in more relative, rather than absolute, terms, evaluating progress systematically. For example, RP011 mentioned:

> "It made me compare throughout the session. Before I was at exertion level 1 [...] then I started getting a bit tired. Is it now 2? I was rationally thinking: is this how tired I can get? can I get more tired? Can I get less tired? So, it's not only a number. […] It has more nuance, more richness. [...] I thought I ran too much, as I was so tired at the end. But [when I think back to my self-reports across the run] I realize that I started at exertion

level 1, and I finished at level 3. I actually didn't push myself as much as I thought. [...] So, this question help[s] me say, 'you're tired, but it's not the worst that you can be'". [RP011]

This capacity to follow the temporal changes in one's condition bestowed a renewed sense of security and achievement in some runners. For instance, RP007 (Female, 23) said that, by being able to visualise and foresee the full experience:

"I felt like I could understand how I'm changing over time. Even though you'll see it wasn't like a massive change, there was a general pattern of being okay at the beginning and then feeling tired and then stopping a bit and walking, then being okay again and then being tired. [...] It was also good to be able to see that." [RP007]

3.2.3 Deeper understanding of the relationship between pain and the other experiences

Particularly in the context of people living with chronic pain, many participants elaborated on the usefulness of self-reporting in terms of allowing them to be more *aware* of the complex, and often confusing, connections between their levels of pain and their emotions in the moment. Participants thought that being asked to assess their worries and confidence gave them the chance to reflect about how these elements impact one another and their felt (and reported) pain, again in a much more granular way. In this sense, CP004 mentioned that for her managing pain successfully equates with understanding when she starts getting too worried about pain, and that worry becomes a paralysing fear. The issue stopping many people from successful management, she recognises, is in being able to pick up those paralysing thoughts as they emerge, in being able to recognise the intimate relationship between worries becoming fears, and pain. She stated:

"You can't do anything about a thought, you can't stop that thought from stopping you, unless you catch yourself doing it first. Having to think every minute to tell you about my worries helped me catch those patterns in the first place. It allowed me to catch myself saying 'am I worrying about this? Oh yes, I am. What am I worrying about? Can I do anything about it?'. That's the moment where I can then use all the other strategies, we talked about to counteract irrational fears". [CP004]

From CP004's remarks, what self-reporting seems to be adding to the equation is the possibility to operate upon emotions *as they emerge*, rather than pre-emptively or retroactively. It allowed her to truly 'catch the thought' *in the moment*.

Reflecting more explicitly on how self-reporting made her more aware in the situations she has practically explored with the researchers, CP006 (43, Female) came to similar conclusions. She felt self-reporting allowed her to start to unpack the moments and instances in which her confidence is impacted by physical exercises, which in turn allowed her to question whether those thoughts were related to pain itself, or to previous and/or different experiences, and what could be done about such instances.

It is also worth noting here that the design of our self-reporting protocol was ultimately *closed*: participants were restricted in the labels they were self-reporting on (the labels are four types of state decided from our previous study [1, 18]), and they were required to answer the prompts with just absolute values they had memorised in advance. While the choices of labels were motivated by previous studies which highlighted how worry and confidence are important factors in the experience of pain in people living with chronic pain [4, 19], and exertion is central for people practicing sports [1, 18], some participants highlighted how there were indeed other factors they thought would be relevant to pay attention to. Almost all participants in the chronic pain study mentioned they would have liked to be asked to self-report upon their emotional state. In this sense, CP009 mentioned:

"Yes, I definitely think you picked the right words, nothing to object there. I also think that I would have liked more freedom to tell you how I am feeling. Because that varies a lot while I'm doing chores, and it has an impact on my pain, and I often struggle to figure out why and how my mood swings. So for instance, when I'm ironing I often get depressed, and that makes my pain worse. I never seem to be able to figure out when that happens, and I think that being asked about how I'm feeling could help me figure that out. And then who knows? Maybe that could help tackling the pain side of things." [CP009]

In this sense then, a more open-ended approach to self-reporting might be taken, for instance by allowing participants to more freely report upon what they believe is relevant in a given circumstance. Using a language that

makes sense to people in specific situations and emotional circumstances may also be critical to contribute to a more reliable and fine-grained ground truth, where 'fine grained' is here understood in terms of the *type* and *quality* of the experience, rather than simply temporal frequency (of the report) or intensity (of the reported levels).

3.2.4 Ability to verbalise in detail acquired knowledge

Lastly, both studies showed how a voice-based, more dialogical setting translated into more expansive self-reporting, with participants often elaborating beyond scalar values they were required to provide. In the study with runners, where participants were encouraged to think aloud the thoughts in their minds (in order to gather additional and contextual information regarding one's experience), all participants mentioned self-reporting naturally extends itself to more refined explanations around the mental processes underlying one's evaluation. For instance, RP003 stated: "You probably wouldn't want to just limit yourself to a number – I also want to say 'I reply with [this number] *because* [this arose; this movement happened; etc.]".

More interestingly, in the study with people living with chronic pain, participants have never been encouraged nor required to provide additional information beyond answering to the self-reporting prompts with the scales they had memorised. Despite this, in many instances (with all participants doing it at least once) they provided, alongside their self-report, a brief exegesis to explain particularly noteworthy changes in the self-reported values between consecutive sampling points As we will argue in our discussion, such insights are not just important verbalisations that could aid users towards better awareness of bodily changes, but could also be successfully integrated as qualitative data for better refinement of ground truth labels.

3.3 Self-reporting fostering a sense of companionship.

Alongside the heightened capacity for self-reflection, runners surprisingly highlighted how voice interaction via the ESM system provided them with a sense of a socially shared running experience, as well as a sense of ***companionship***. RP010 (Male, 29) for instance mentioned that the voice interactions felt like "talking to my friend", commenting that "it's good, it's talking to me like a friend. It keeps me entertained, occupied". Along similar lines, RP003 said that "it was interesting – having these inputs… like this question every now and then. It felt like […] a companion". Interestingly, this social function seemed to enhance positive emotional attachment to the system, making runners feel better in return. In this sense, RP005 (Female, 40) mentioned:

> "It prompted me to talk. […] And I found that talking to the device made me feel happier. Whereas if I had just been quiet, I wouldn't have been so happy. I don't know why that was the case. It feels like… I have someone to talk to even though you're just talking out loud" [RP005]

Some runners also advocated for a more human-like, personalised ESM system. They expressed their preference for a human voice, rather than a computer-generated one. Particularly, some mentioned that a familiar voice (e.g. a friend, family, or fellow runner) would be particularly helpful in augmenting the self-reflective behaviours outlined above. For instance, RP003 said:

> "I think if it's like a person that you know… I think it would be an interesting experience. It would feel more familiar, instead of being a synthetic voice. […] So it's like they are asking how *you* are doing during the run. [It would be helpful] especially if it was someone that you normally share your running experience with" [RP003]

Participants in the chronic pain study (who were prompted by a human actor) commented even more emphatically regarding the potential for companionship afforded by self-reporting. CP003, for instance, thought that having someone, or something, to relay feedback to allowed him to feel closer to an ideal of personalized care. Particularly, he felt that self-reporting could be a good avenue and opportunity for both self-monitoring and the feel that someone is listening and continually 'checking in' on their status. He further elaborated:

> "There is huge potential here for input […] [for] that extra anything. It feels like you are talking to someone, that you might be getting a little bit of feedback. Or maybe what you're doing wrong, what you could be doing better. Just feels like something checking in on you remotely. That could be very positive." [CP003]

Similarly, CP002 (37, Female) felt that self-reporting served self-reflective purposes similar to written diaries she had been tasked to complete in the past. Unlike diaries however, self-reporting to a physically present human gave her a feeling of social presence, of someone being there to listen, and therefore a heightened sense of *accountability*. Because someone was on the other end listening, she felt that she could not just pretend, or give answers without much thought.

Lastly, CP005 (44, Female) further elaborated on the idea of the self-reporting as a positive (social) distraction mentioned previously. The house activity that she carried out in the session was a small painting job in her flat. At the end of the activity, the researcher remarked that she had been working for almost 30 minutes. She was positively surprised by this observation, and mentioned to the researcher that it had never happened that she could paint for such a long amount of time without major breaks, and that she never managed to paint an entire wall in one sitting. Further reflecting on the experience, she said:

> "I feel like it's because you were there asking me things, and then we got to talk about other things… it just made me feel like pain was there, but it was also not everything. I felt distracted, but still focused on my work. It's similar to when I'm doing stuff in the house with [wife]: her voice calms me and distracts me, and we get to talk about stuff, and all of a sudden I realise I've been at it for almost an hour" [CP005].

4 Discussion

Through our two studies, we have shown that a voice-based ESM, complemented by participants' additional think-aloud reflections, has led to more engaging self-reporting experience, one which stimulated self-awareness and triggered self-reflection, thus providing immediate and tangible benefits and value to the participants themselves. These two factors have shown the potential for obtaining more considered, and hence possibly more reliable ratings. In addition, the think-aloud exegeses complemented such ratings with rich contextual and highly subjective qualifications of the rated experience, thus providing further information for a fine-grained ground truth. Unfortunately, the analysis of such free text poses challenges to the Natural Language Processing communities, particularly in terms of how to automatically transform such free text into a ground truth suitable to train a pain recognition machine learning model. To address this problem, our studies also suggested that the voice-based ESM approach should embrace the dialogical structure of a chatbot to facilitate think aloud moments by leveraging a sense of social companionship. On the one hand, companionship could contribute to the engagement level by fulfilling a user need, as well as making the think aloud self-reflection process more natural. On the other hand, the dialogical structure provided by a chatbot could further help people to elaborate their self-reports in a more structured dialogical format. Such structure would enable to more easily extract the information (e.g., comparative ratings, location of the pain, type of pain) to train pain recognition systems, or even simply validate and refine the ground truth. In the next section, we briefly discuss the above points.

4.1 Maximising benefits of self-reporting.

As emerged from participants' experiences, self-reporting has the potential to be more than just an efficient data collection tool for researchers, and rather provided tangible and integrative benefits for participants too. Participants in both studies valued how synchronous self-reporting facilitated self-awareness and self-understanding, thus assigning value to the ESM exercise beyond merely completing an assigned task. This could be a reason for participants to (pro-)actively engage in self-reporting, which is also highlighted by existing ESM literature: participants' retention is maximized when the ESM activity is perceived as relevant and of value to them [20, 21].

As shown in the findings (see theme 3.2), participants found value in the enhanced opportunities for self-reflection and introspection emerging through self-reporting. This is in line with previous studies which highlighted that ESM helped people to be aware of the underlying meanings of, or reasons for, their feelings, judgements, and behaviours, by encouraging them to reflect on otherwise unnoticed events [22], and the studies that showed the self-reporting in enabling reflection can the cycle of people's unhealthy habits [23]. Despite such consensus, further work is necessary to better outline detailed guidelines for designing an ESM system that leverages such findings, particularly in our specific study contexts (i.e. running and chronic pain).

Our findings suggest that facilitating self-awareness and self-understanding necessarily involves careful consideration around different design steps. For example, how should the ESM system kickstart a person's introspection process? Speech, either from computer-generated or human-generated voice messages, worked in our

studies – not only because it is minimally intrusive to the functional activities considered here but also because it seemed to evoke a sense of social association, mimicking the dynamics of a conversation. It is particularly important in this sense that the ESM system prompts be clear and easily understandable by participants. We believe that further studies are necessary to explore and assess how the dialogue between participants and the ESM system (potentially a chatbot) should be designed, particularly in terms of starting the conversation, as well as accounting for other context-dependent factors: questions such as when and how often to send question prompts, the kinds of phrases to be used, the speaking style and tone to be used will have to be attuned to the specific scenarios the agent will be deployed in. All of these considerations could be further explored, particularly through empirical studies and comparative usability tests, in order to maximise the perceived benefits of the system for participants in real-life scenarios. Such understanding will further help us design a chatbot-supported ESM system towards stimulating users to kickstart an introspection journey towards mindful self-awareness.

As shown in findings 3.2.1, we observed that our participants often conducted a mental checking of their body (e.g., shoulder, arm, etc.) and current activity performing immediately before reporting their experience. If such examinations were skipped, the self-reported experience might be biased by several factors, including the memory of the pain from the previous moment, or the anticipation of pain in a future moment. The ESM system could play a role here to counsel participants by promoting body scanning. Such scanning guidance should be tailored to contextual and individual differences. For instance, people with chronic pain might need more focus on certain parts of their body, or during certain types of movements. Runners often need to investigate their full body postures to make sure that there is as little as possible negative impact on any injured body part. Such mindful scanning activities could be integrated into a chatbot with ESM capabilities.

More importantly, findings (see theme 3.2.2) in our studies seem to suggest that, once participants acquired knowledge of their in-the-moment experience, their interests grew further, towards rethinking and comparing both their past experiences and future possibilities. Runners for instance relied on the comparison with previously self-reported pain (e.g., a mile before, or at the beginning or the run) to rate their current pain level; they also often referred back to the self-reporting in its entirety to draw conclusions regarding their performance. Such comparative behaviours for assessing subjective experience have been studied as ordinal/relative ranking (e.g., higher pain, lower pain) [24-26], which has been proved to be more reliable than absolute rating (e.g., level 3 or 4) as ground truth for machine learning training. In this sense, it might be relevant in the future to explore an ESM system that guides participants towards relative ranking and self-reporting. A chatbot with ESM features could ask questions to guide participants to conduct comparative subjective evaluation that are often considered more reliable [27], but in a context that makes sense and that is useful to the user. Particularly, further research should be conducted to study and evaluate how the choice between absolute rating, relative rating, or a hybrid approach could affect the user experience (e.g., user's cognitive load, memory effect) of self-reporting, and in turn affect the reliability of pain self-reports as ground truth for machine learning training. Leaving aside these important questions around the modality of rating, we believe ESM-stimulated prompts could be at the basis of a system which could provide a quick and easy overview of previous reports – giving for instance a pain summary. The appropriateness of any representation formats needs to be investigated particularly considering usage contexts: people with chronic pain might visualise such data through a larger screen at home, whereas runners might find it more useful to have such schematics read out loud. It will also be important to assess whether providing such summary might affect users' perception of their current state, possibly making a potential benefit into a dangerous hinderer.

Finally, the use of the chatbot could enable to transform the free think aloud used in our study in a more structured conversation. A chatbot can extract end-users' expressions as parameters that can be easily used to perform computational and logical processing[28], which can enable to clarify and verify fine grained aspects of the ground truth.

In short, we believe that design of ESM approaches to data collection could and should be attentive to maximising perceived benefits for participants, rather than focusing solely on improving efficiency and data quality for researchers. This will inevitably entail a more tailored and attuned design, one which fosters and leverages the potential for self-reflection afforded by self-reporting in the moment and is attentive to the contextual challenges of different audiences and scenarios. People's introspective experience discussed above could create space for machine learning algorithms to continuously learn [29] from the adynamic affective experience. A chatbot could leverage its AI capabilities to guide more structured conversational interactions with a person in order to facilitate such reflective experience when

conducting ESM tasks. This will hopefully in turn improve retention and the quality of perceived benefits, as well as providing researchers more expansive self-reporting to construct ground truth labels.

4.2 Leveraging social companionship within ESM designs

As shown in the theme 3.3, participants strongly alluded to feeling a sense of companionship from using the ESM system in both studies. What is clear from the participant interviews is that the *immediacy* of self-reporting, as well as the use of the *voice,* facilitated this dialogical form of interaction. In stark contrast to asynchronous forms of self-reporting (e.g., post-activity written diaries), voice-based ESM prompts provided participants with a stronger sense of accountability, leading to a feeling of being followed and cared for. Runners saw the voice ESM prompt as from a friend and hinted at the possibility of constructing a system which shows richer human traits and communicative capabilities. For instance, RP008 expressed the desire for a more natural and improvised conversational element by saying "This could also be a good conversational assistant – […] maybe you could find certain phrases or words that are more common between people".

It is critical to additionally note the importance of humanness in the ESM system, a factor that was strongly highlighted whether we used a computer-generated prompt or an actual physically-present human prompt. It is further valuable to highlight that, while findings from the participants who were exposed to the computer-generated prompt (i.e. the runners) clearly indicate that an artificially-intelligent ESM system with human-like capabilities could be satisfactory, those exposed to the prompt from a human (i.e. participants with chronic pain) seemed somewhat unsure about whether the system would be a benefit to them to the fullest extent without a real human. For example, when explicitly asked about the possibility of replacing the role of the human researcher with a machine, CP002 was particularly sceptical:

> "To be completely honest, I think self-reporting would still be beneficial – and as I said I already tried to mentally implement it when we were not together. But I don't think it would be as good as it is when I'm doing it with *you*. […] Because I feel we built a relationship. […] We got to talk about things while I was doing chores: I know we like the same videogames, we both watch Grey's Anatomy. That really helped. […] Could a machine give me that same sense of talking to someone? I don't know."

Other participants commented along similar lines. CP003 mentioned for instance that in an ideal scenario the researcher would be there all the time, but knowing that that is physically impossible, a technological companion would suffice, albeit perhaps not as successfully.

Nevertheless, our findings do suggest value in further exploring the idea of embedding this prompting system within a smart chatbot which could integrate human conversational capabilities to enhance the outlined sense of companionship. Our findings on the participants in the chronic pain study show that human-initiated ESM prompts are beneficial, with many being surprised by the ease with which they managed to complete tasks, and some explicitly attributing this accomplishment to their heightened self-awareness triggered by the self-reporting task.

While they partly attributed the efficacy of the prompts to the physical presence of the researcher, it is important to stress that the efficacy of the system did not rely *solely* on physical presence. As highlighted by the excerpt above, participants in fact recognised that part of the reason why they found self-reporting exercises particularly useful was because they had managed to engage in 'idle talks' with the researcher in between self-reporting prompts, building some form of a social relationship with him. To clarify, while activities were ongoing, the content of these conversations was never focused on the activities at hand (e.g. "why are you hoovering that 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, the participant and researcher often engaged in small talks regarding trivial topics. In one instance, the researcher noticed a gaming console, and the two started talking about videogames that they had recently played. In another instance, the participant asked the researcher about his opinions regarding the coffee he had purchased while unloading a food shopping deliver.

These trivial conversations seemed to have further strengthened the sense of companionship provided by the ESM-enhanced tasks. This may in turn motivate participants to engage more in the self-reporting tasks, thus contributing to richer datasets for researchers. This is in line with previous research [30] which argued that a sense of a socially shared experience contributes to a higher compliance rate of ESM tasks. As pointed in existing literature, traditional computer-generated questionnaire is dull and lacking elements to motivate participants, whereas a chatbot-delivered

questionnaire could provide more engaging conversational interactions, which can lead to higher quantity and quality of participants' responses to the ESM questions [31].

We thus argue that, in order to truly leverage the sense of social companionship highlighted by participants, it will be necessary to more attentively enquire and attend to this trivial conversational dimension emerged in the study on chronic pain, since our data show how perceived benefits of the system seem to rely upon it. In this sense then, further studies are required to explore the following questions: 1) what is understood as 'small talk'? i.e. a conversation which is not too cognitively taxing, but still affords the formation of some kind of sociable relationship; 2) what elements of the conversations with the researcher helped participants establish social ties with him? 3) what role does the personality of the researcher play? 4) how could participant's personality type affect the perceived companionship?

Once the nature of such sense of connectedness between participant and researcher is more properly understood, such insights could inform the design of an ESM-powered chatbot which could closely mimic and attempt to replicate the kinds of conversations participants came to expect from their time spent with the researcher. This will in turn lead to further work solely focused on exploring naturalness of the chatbot, which will necessarily have to be intelligent enough to understand when to cut conversations short as not to detract the user from self-reporting, but also in other circumstances perhaps allow more conversation in lieu of the self-reporting. In this sense, a critical focal point will be that of understanding natural conflicts between assigned task (i.e. self-reporting) and idle time (i.e. the conversational element), thus devising systems that are able to resolve and adapt to such conflicts in a human-like manner.

In summary, we believe building a chatbot which is able to engage in interactions with users in a much more natural, responsive and reactive manner might be beneficial in two ways. On the one hand, it would facilitate self-reflective behaviours leading to more precise and granular self-reporting data to be used in the context of machine learning. On the other hand, it would leverage the sense of companionship and distraction outlined by participants, allowing them to be physically active for longer and in a more mindful manner.

4.3 Using self-reporting data as ground truth labels: reliability and challenges.

As discussed above, the voice-based ESM approach deployed facilitated deeper self-reflection, further enhancing the reliability of self-reported pain ratings. First, such self-reflection and introspection worked as an initial consolidation step for participants themselves to realise how they actually felt *before* giving the rating to their pain level, because, as highlighted above, self-reporting allowed them to start to unpack how pain changed, or how it was associated with specific movements, etc. This self-reflective attitude stimulated by this ESM approach could potentially reduce the randomness of ground truth values due to lack of thorough assessment, and thus lead to more accurate and reliable ground truth for machine learning training.

What is more, participants also often thought aloud (see theme 3.2.4) secondary and more nuanced explanations beyond the given scales, providing brief additional explanations around the self-reported values. We believe that such secondary data recordings could be analysed to further validate the reliability of numeric self-reports. For instance, consistency and correlation between the self-reported scalar values and the semantic meanings of the secondary verbalisations could be scrutinised, opening the space for researchers to understand if a given set of data is valid, or if it requires further exploration. Additionally, qualitative analysis (e.g., thematic analysis, content analysis) could provide additional linguistic cues (e.g., pain location, cause, emotional reaction to pain, etc) as multiple labels [32] together with the scalar values towards building a more contextualised pain recognition system. Machine learning researchers could also take advantage of such expansive data provided by participants' self-reflection process to further triangulate the reliability of the self-reported pain ratings and descriptions across a variety of formats. Such analysis could be conducted by human researchers, or by advanced NLP algorithms [33] in the future.

Beyond questions around reliability of self-reported scalar values, there is the main challenge of how to match discrete self-reports (as ground truth) with continuous, concomitant sensor data. In our study, participants were prompted to assess and self-report the pain that they felt *in the moment,* rather than over a previous period of time. This means that pain was assessed discretely. However, other data collected by the sensors are continuous. Pain report at one single discrete point cannot inform whether the *current* pain had already existed, when it started, or when it ended. Although increasing the self-report sampling rate may give us a more fine-grained view, it would come with the demerit of increasing the participant's workload. In fact, it is technically impossible for the self-report or ground truth sampling rate to match the sensor sampling rate. This itself prompts several questions on how to utilize self-report data as ground truth for labelling sensor data: what should be the time scale accounted for by each self-report? what

should be the starting and ending point of this time window? This is a problem that machine learning researchers are already dealing with (e.g. a multiple instance learning method for addressing the time ambiguity of emotional responses [34]). What our results show is that highly frequent reports are acceptable possibly facilitating the collection of larger datasets on ubiquitous settings as it may happen in the context of video-based large datasets enabling the use of semi-supervise learning to further improve performances. Such problem also opens space for a chatbot based ESM system to guide participants to provide more accurate and temporally granular information at each self-report. For instance, the ESM system could further ask participants: "when did you feel this pain?" or "How many seconds ago did you feel it?" We speculatively suggest that a chatbot could ask such questions when needed [35]. More importantly, participants' responses gathered via chatbot are more structured compared to a free-style or open-ended elaboration, which makes it easier for machine learning researcher to further process and analyse.

While our findings suggest the possibility of rethinking the content and format of the prompts, further analysis is necessary to establish if more punctual questions would diminish participants' compliance. Moreover, it is necessary to assess and devise strategies to incorporate such open-ended responses into precise ground truth labels.

5 Limitations

While in situ self-reporting revealed itself a particularly convenient methodological avenue within the contexts analysed, we nonetheless observed specific scenarios in which participants struggled, were not able, or preferred not to self-report. For instance, some runners expressed discomfort in self-reporting via their voice in a public space (i.e. the gym), fearing they might be judged. Moreover, there is the possibility of limited verbal capabilities when experiencing higher levels of exertions or when in pain. With the chronic pain study participants, there were a few occasions of delay in self-reporting due to the urgency or time-sensitivity of specific house tasks. There were other situations in which environmental or equipment-related challenges posed a barrier to self-reporting – for example, participants found it harder to hear the researcher's prompts when vacuuming. Lastly, we are aware that voice-based self-reporting might not be viable with certain population groups, e.g. preverbal children, people with speech impairments or cognitive disabilities, etc. We believe that these limitations could encourage further studies to explore the degree to which the basic principles of our approach could be remodelled within different circumstances and within populations with different needs. In addition, this current study only discussed the potential improved reliability and richness of ground truth labels, validation methods (either qualitative or quantitative measurements) should be further explored in future studies.

6 Conclusion

This paper reflected on the use of voice-based, in situ self-reporting approaches for assisting people's self-reporting of pain. We showed that self-reporting could be beneficial for participants themselves, which would in turn provide more fine-grained, detailed, and possibly more reliable, ground truth. We presented qualitative findings from two case studies: people self-reported pain to computer-generated ESM prompt during distance running, and people with chronic pain self-reported to human-generated ESM prompt during house activities. At this stage, our research was solely focused on gathering experiential insights from participants to better understand what elements are positively valued when engaging in in-situ self-reporting: while our findings overwhelmingly pointed us towards the value of building such a system, it is necessary to further stress that these initial considerations (as well as the technical system stemming out of these) will inevitably have to be evaluated through quantitative controlled studies. Nonetheless, the findings showed that participants felt better able to engage in self-reflective behaviours towards better understanding and awareness of their pain and emotional state as a result of interacting and responding to voice-based ESM prompts. Furthermore, they felt the sense of social companionship provided by the prompting system, which helped them engage in activities for longer and with a less negative disposition. Such engagement with self-reporting, and the increased self-knowledge stemming from it, appear to create spaces and opportunities for improving the reliability and quality of ground truth ratings for building automatic pain recognition systems. We also preliminarily highlighted some of the challenges that lie ahead in order to build self-reporting systems which are attentive to the needs of specific audiences, to maximise compliance and perceived benefits, and in turn the quality of the data provided. Lastly, we enquired into the possibilities and challenges of designing a chatbot which could stimulate users towards an introspective journey of mindful self-awareness. Particularly, we focused on the importance of better understanding seemingly trivial elements of human-to-human conversation, which seemed to have played a key role in participants' sense of social and emotional

attachment to the prompting system. We hypothesised that the structured, but not fully repetitive, dialogue of chatbots could help extract ground truth information from an apparently fluid conversation. Such chatbot could also be used to adapt the questioning and the conversation through terminology that is meaningful to the user. Naturally, in depth longitudinal studies (with more controlled and quantitative evaluations) with the use of a chatbot are needed to confirm our initial results. That is indeed our next step in this journey.

ACKNOWLEDGMENTS

Tao Bi was supported by a studentship deriving from a grant awarded by the London Legacy Development Corporation (C02955) to the Global Disability Innovation Hub, University College London. The work was also supported by the EU Future and Emerging Technologies Proactive Programme H2020 (Grant No. 824160: EnTimeMent - https://entimement.dibris.unige.it/).

REFERENCES

1. Bi, T., et al., Understanding the Shared Experience of Runners and Spectators in Long-Distance Running Events, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019, ACM: Glasgow, Scotland Uk. p. 1-13.
2. Singh, A., N. Bianchi-Berthouze, and A.C. Williams. *Supporting everyday function in chronic pain using wearable technology*. in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017.
3. Felipe, S., et al. *Roles for personal informatics in chronic pain*. in *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. 2015. IEEE.
4. Singh, A., *Staying active despite pain: Investigating feedback mechanisms to support physical activity in people with chronic musculoskeletal pain*. 2016, UCL (University College London).
5. Werner, P., et al., *Automatic recognition methods supporting pain assessment: A survey.* IEEE Transactions on Affective Computing, 2019.
6. Labus, J.S., F.J. Keefe, and M.P. Jensen, *Self-reports of pain intensity and direct observations of pain behavior: when are they correlated?* Pain, 2003. **102**(1-2): p. 109-124.
7. Yan, X., et al., *Toward Lightweight In-situ Self-reporting: An Exploratory Study of Alternative Smartwatch Interface Designs in Context.* Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020. **4**(4): p. 1-22.
8. Gao, N., et al., *Investigating the Reliability of Self-report Survey in the Wild: The Quest for Ground Truth.* arXiv preprint arXiv:2107.00389, 2021.
9. Healey, J. *Recording Affect in the Field: Towards Methods and Metrics for Improving Ground Truth Labels*. in *Affective Computing and Intelligent Interaction*. 2011. Berlin, Heidelberg: Springer Berlin Heidelberg.
10. Van Berkel, N., D. Ferreira, and V. Kostakos, *The experience sampling method on mobile devices.* ACM Computing Surveys (CSUR), 2017. **50**(6): p. 1-40.
11. Intille, S., et al., *µEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch*, in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2016, Association for Computing Machinery: Heidelberg, Germany. p. 1124–1128.
12. Vaizman, Y., et al., *ExtraSensory App: Data Collection In-the-Wild with Rich User Interface to Self-Report Behavior*, in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, Association for Computing Machinery. p. Paper 554.
13. Timmermann, J., W. Heuten, and S. Boll. *Input methods for the Borg-RPE-scale on smartwatches*. in *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. 2015.
14. Adams, A.T., et al. *Keppi: A tangible user interface for self-reporting pain*. in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018.
15. Braun, V., V. Clarke, and P. Weate, *Using thematic analysis in sport and exercise research.* Routledge handbook of qualitative research in sport and exercise, 2016: p. 191-205.
16. Braun, V. and V. Clarke, *Using thematic analysis in psychology.* Qualitative research in psychology, 2006. **3**(2).
17. Adams, P., et al. *Supporting the self-management of chronic pain conditions with tailored momentary self-assessments*. in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017.
18. Bi, T. *Wearable Sensing Technology for Capturing and Sharing Emotional Experience of Running*. in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 2019. IEEE.
19. Olugbade, T., N. Bianchi-Berthouze, and A.C.d.C. Williams, *The relationship between guarding, pain, and emotion.* Pain reports, 2019. **4**(4).
20. Ghosh, S., B. Mitra, and P. De. *Towards Improving Emotion Self-report Collection using Self-reflection*. in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020.
21. Silvia, P.J., et al., *Missed beeps and missing data: Dispositional and situational predictors of nonresponse in experience sampling research.* Social Science Computer Review, 2013. **31**(4): p. 471-481.
22. Sloboda, J.A., S.A. O'Neill, and A. Ivaldi, *Functions of music in everyday life: An exploratory study using the Experience Sampling Method.* Musicae scientiae, 2001. **5**(1): p. 9-32.
23. Pinder, C., et al., *Digital behaviour change interventions to break and form habits.* ACM Transactions on Computer-Human Interaction (TOCHI), 2018. **25**(3): p. 1-66.
24. Yannakakis, G.N., R. Cowie, and C. Busso. *The ordinal nature of emotions*. in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2017. IEEE.
25. Martinez, H.P., G.N. Yannakakis, and J. Hallam, *Don't classify ratings of affect; rank them!* IEEE transactions on affective computing, 2014. **5**(3): p. 314-326.
26. Yannakakis, G.N. and H.P. Martinez. *Grounding truth via ordinal annotation*. in *2015 international conference on affective computing and intelligent interaction (ACII)*. 2015. IEEE.
27. Yannakakis, G.N. and H.P. Martínez, *Ratings are overrated!* Frontiers in ICT, 2015. **2**: p. 13.
28. *Dialogflow Documenation*. Available from: https://cloud.google.com/dialogflow/es/docs/intents-training-phrases#annotation.

29. Churamani, N., S. Kalkan, and H. Gunes. *Continual Learning for Affective Robotics: Why, What and How?* in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2020. IEEE.
30. Chang, Y.-L., Y.-J. Chang, and C.-Y. Shen. *She is in a Bad Mood Now: Leveraging Peers to Increase Data Quantity via a Chatbot-Based ESM*. in *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 2019.
31. Te Pas, M.E., et al., *User experience of a chatbot questionnaire versus a regular computer questionnaire: prospective comparative study.* JMIR Medical Informatics, 2020. **8**(12): p. e21982.
32. Zhang, M.-L. and Z.-H. Zhou, *A review on multi-label learning algorithms.* IEEE transactions on knowledge and data engineering, 2013. **26**(8): p. 1819-1837.
33. Yu, C.H., A. Jannasch-Pennell, and S. DiGangi, *Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability.* Qualitative Report, 2011. **16**(3): p. 730-744.
34. Romeo, L., et al., *Multiple instance learning for emotion recognition using physiological signals.* IEEE Transactions on Affective Computing, 2019.
35. Oh, K.-J., et al. *A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation*. in *2017 18th IEEE international conference on mobile data management (MDM)*. 2017. IEEE.