Predicting Health States of Patients with Chronic Pain from Cellphone Usage Data

Maya Stemmer MAYA.STEMMER@POST.RUNI.AC.IL
The Data Science Institute, Reichman University, Israel

Lior Ungar LIOR.UNGAR@SHEBA.HEALTH.GOV.IL
Pain Unit, Sheba Medical Center, Tel Hashomer, Israel

Talia Friedman TALIA.FRIEDMANN@GMAIL.COM
The Data Science Institute, Reichman University, Israel

Lihi Bik LIHI.BIK@POST.RUNI.AC.IL
The Data Science Institute, Reichman University, Israel

Yotam Hadari YOTAM.HADARI@SHEBA.HEALTH.GOV.IL
Pain Unit, Sheba Medical Center, Tel Hashomer, Israel

Itamar Efrati ITAMAR.EFRATI@POST.RUNI.AC.IL
The Data Science Institute, Reichman University, Israel

Yarden Rachamim YARDEN.RACHAMIN@POST.RUNI.AC.IL
The Data Science Institute, Reichman University, Israel

Lior Carmi LIOR.CARMI@POST.RUNI.AC.IL
The Data Science Institute, Reichman University, Israel

Shai Fine SHAI.FINE@RUNI.AC.IL
The Data Science Institute, Reichman University, Israel

Abstract

This study followed patients suffering from chronic pain and aimed to predict their health states. To this end, we conducted a clinical study in which patients were digitally monitored via clinically validated questionnaires (SF-36 and EQ-5D) and continuously collected cellphone usage data. We present a novel two-step approach for utilizing the immense amounts of unlabeled cellular logs in a supervised, binary classification problem and predicting patient-reported outcomes from objective cellphone usage data. Reaching an accuracy of 0.827 for women and 0.898 for men, our classification results show the feasibility of using cellphone monitoring data for patient state prediction. Such a capability may enrich periodic clinical assessments with frequent digital follow-ups, assist in disease management for chronic patients, and raise awareness whenever necessary.

Institutional Review Board (IRB) This study was approved by the IRB of Sheba Medical Center in Israel.

Data and Code Availability In this paper, we analyze data from a clinical study at the Pain Unit of Sheba Medical Center in Israel. Due to consent restrictions, this data cannot be shared. The code is customized to process this data and includes details that may reveal aspects of the data. Moreover, the code operates in a highly regulated environment that restricts any export. Hence, we cannot share the code as well. However, the pipeline is heavily based on the scikit-learn Python package (Pedregosa et al., 2011), and we provide sufficient details to reconstruct the pipeline.

1. Introduction

Chronic pain exerts an enormous burden on society, healthcare systems, and individuals, affecting more than 30% of people worldwide (Cohen et al., 2021).

It significantly disrupts patient routines and reduces their quality of life. Lately, machine learning solutions and data-driven approaches have been used in chronic pain research in attempts to ease the burden (Santana et al., 2020).

This work explores the feasibility of using cellphone usage data to predict the states of patients suffering from chronic pain. For this purpose, we carried out a six-month observational clinical study at the Pain Unit of Sheba Medical Center involving participants exclusively diagnosed with chronic pain. The experimental setup involved continuously gathering and analyzing cellular usage data (logs of app usage and cellphone events) from a cohort of 91 participants.

In addition, patients were asked to fill out two types of health-related quality of life (HRQL) questionnaires to monitor their health states. With the guidance of the research assistant, they manually filled out the SF-36 questionnaire (Ware et al., 1993) three times during the study - in the beginning, middle, and end. On their own time and ideally every week, they were asked to digitally fill out the EQ-5D-5L questionnaire (Rabin and Charro, 2001), which is shorter and clinically validated for mobile devices (Kamstra et al., 2022). We provide a detailed statistical analysis of the patient-reported outcomes measured by these two questionnaires.

The information provided by the patients in their EQ-5D reports served as the basis for our supervised prediction task. Based on the relevant value set (Pickard et al., 2019), we converted the five dimensions of the EQ-5D questionnaire (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) into a continuous health index. Then, we performed binary classification on the joint HRQL measure. Using the accepted cutoff of 0, indicating a health state as bad as death (Devlin et al., 2020a), we divided the states into worse-than-death health states (0) and otherwise (1). In this binary setting, we aimed to determine whether a patient's health state was so severe that they considered it to be worse than death.

We present a novel two-step method for utilizing the immense amounts of unlabeled cellular logs in a supervised classification problem. First, we construct a new representation of the cellphone usage data using classic and deep unsupervised methods, such as clustering, manifold learning, auto-encoding, and self-learning. The new representation is designed to capture typical patterns of behavior, as reflected by cellphone usage, and highlight deviations from these patterns on a weekly basis. Then, we apply supervised classification algorithms to predict patients' health states depending on the resemblance between their current usage and their normal behavior.

Interestingly, our data analysis points to engagement differences between men and women, as the female participants tended to fill out the EQ-5D and SF-36 questionnaires more often than the male participants. Nonetheless, in terms of health state *severity*, no statistically significant differences were found between the reported health states of the two genders in either questionnaire. In our tested settings, gender-specific classification models consistently outperformed generic (gender-agnostic) models.

Our classification results, reported on a held-out test set, demonstrate promising, statistically validated outcomes, which support the viability of utilizing cellphone monitoring data to predict chronic pain states. Such capability can potentially enhance regular clinical evaluations with frequent digital follow-ups, assist in disease management for chronic patients, and prompt awareness whenever necessary.

This study offers two main contributions. The first is methodological in the form of the two-step method we suggest for transforming mobile sensing data into predictors in a supervised classification problem. An earlier version of this method[1] was also applied in a previous study aiming to predict the imminent suicide risk of adolescents previously diagnosed with depression (Stemmer et al., 2024). The method's ability to provide good predictions for different populations and indications shows its adaptability and generalization power.

The second key aspect is the clinical contribution. Given the inherent challenges in pain assessment (Hadjiat and Arendt-Nielsen, 2023) and the growing need for digital biomarkers to facilitate quantitative, data-driven, and evidence-based clinical measurement (Rejula et al., 2021; Hadjiat and Perrot, 2022), digital monitoring offers a more objective, continuous, and scalable solution for evaluating and managing pain, including severe pain episodes (EFIC, 2022). Notably, due to the high accuracy of our model's results, it may assist clinicians in remotely detecting severe pain episodes and implementing timely interventions for patients in greatest need. This might allow the patients acute intervention without the effort of coming to the clinic (Bartels et al., 2024).

1. In this study, we enhanced the method and added two additional transformations, PCT and DCT, that will be explained in the sequel.

The rest of the paper is organized as follows: In section 2, we provide a literature review of the two questionnaires used in the study, SF-36 (2.1) and EQ-5D (2.2), and how mobile sensing data was previously used for predicting patient-reported outcomes (2.3). In section 3, we explain the clinical study we conducted (3.1), describe in detail the collected data and how they were used (3.2, 3.3), and provide descriptive statistics of the population (3.4). In section 4, we detail the two-step mechanism, the unsupervised learning representation step (4.1) and the supervised classification step (4.2). In section 5, we present the classification results (5.2) and discuss the findings' implications (5.3). We conclude in section 6.

2. Background and Related work

2.1. SF-36 for Measuring Chronic Pain

The SF-36 questionnaire is a useful monitoring tool for patients with chronic pain (Elliott et al., 2003; Lin et al., 2024). It measures HRQL using 36 questions within eight concepts: physical functioning, bodily pain, role limitations due to physical health problems, role limitations due to personal or emotional problems, emotional well-being, social functioning, energy/fatigue, and general health perceptions. Each domain is measured on a scale from 0 (meaning worst possible health) to 100 (meaning perfect health) (Ware Jr, 2000).

The SF-36 questionnaire has been widely used for measuring HRQL in patients with chronic pain (Járomi et al., 2021; Vereščiagina et al., 2007). As can be expected, patients suffering from chronic pain reported significantly lower SF-36 scores compared to the general population (Vereščiagina et al., 2007; Torrance et al., 2009). Previous studies showed that mean bodily pain scores of patients with chronic pain range from approximately 35 in severe cases when the patients suffer from multiple pain sites (Dominick et al., 2011) to 64 in other cases (Járomi et al., 2021; Iguti et al., 2021). The equivalent means of the control groups who did not suffer from chronic pain were over 80 (Dominick et al., 2011; Iguti et al., 2021).

2.2. EQ-5D as a Patient-Reported Outcome

The EQ-5D questionnaire is a standardized, generic measure of HRQL (Group, 1990). It is widely used in clinical and economic appraisal and population health surveys. EQ-5D assesses health status in five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. In the EQ-5D-3L version, each dimension has three levels: no problems, some problems, and extreme problems. In the EQ-5D-5L version, designed to improve the instrument's sensitivity and reduce ceiling effects, each dimension has five levels: no problems, slight problems, moderate problems, severe problems, and extreme problems (Oemar, 2013; Hernández Alava et al., 2023).

EQ-5D is designed for self-completion by respondents and can hence be referred to as a patient-reported outcome measure (Devlin et al., 2020b). Patients can complete the questionnaire themselves to provide information about their current health status and how it changes over time. The patient indicates their health state by choosing the most appropriate statement in each dimension, and their decision is associated with a 1-digit number that expresses the level selected for that dimension. The digits for the five dimensions can be combined into a 5-digit number describing the patient's health state. In the EQ-5D-5L version, the best health state is represented by 11111, and the worst by 55555 (Pan et al., 2022).

The health state captured by the EQ-5D is often converted into a single index, summarizing the patient's health using a country-specific value set (Oemar, 2013). A value set consists of weights attached to each level in each dimension, and a designated formula converts the EQ-5D health state into a single value on a scale anchored at 1 (meaning full health) and 0 (meaning a state as bad as death). The scale allows negative values to be assigned to health states considered worse than dead. In the absence of a designated value set for a specific country, the analysis should be based on a value set for a similar country (Devlin et al., 2020a).

2.3. Mobile App Data for Chronic Pain Prediction

Advances in mobile sensing technologies enable the bridging of current gaps in traditional health monitoring (Kelly et al., 2017). While clinical assessments are collected periodically, mobile sensing data are collected in real-time, 24/7, and may provide instantaneous insights (Mullick et al., 2022). Furthermore, sensing data are objective, in contrast to the subjective responses to questionnaires and clinical assessments, which also rely on patient memory and compliance (Morshed et al., 2019).

Recently, mobile apps and sensing data have been used in chronic pain research. While substantial work has shown the benefits of using mobile apps for pain monitoring and management (Aung et al., 2016; Zhao et al., 2019; Thomson et al., 2024), studies focusing on chronic pain *prediction* are still limited. Dorris et al. (2024) demonstrated the power of passive wearable measures by predicting chronic pain from daily actigraphy movement data using a deep neural network. Santana et al. (2020) compared multiple machine learning algorithms to differentiate between chronic pain patients and healthy individuals based on self-report and pain sensitivity data.

Rahman et al. (2018) used a designated app, *Manage My Pain*, to predict the pain volatility of patients with chronic pain. Similar to our pipeline, they started by clustering the data using the k-means algorithm. However, their prediction models were built based on patient-generated data describing their health rather than regular cellphone usage data collected objectively in the background.

Rabbi et al. (2018) investigated the physical activities of patients suffering from chronic pain using a mobile app, *MyBehaviorCBP*. They clustered location trace logs to recognize patient routine behavior and recommend physical activities similar to the patient's usual behavior. While they utilize clustering to identify typical behavior patterns, their work fundamentally differs from ours in several aspects: Their objective was to enhance chronic pain self-management rather than predicting pain states; they collected and clustered different types of data - location traces and GPS, as opposed to cellphone usage data; and their modeling technique differs from the two-step method we suggest in this study.

3. Clinical Study, Data Collection, and Experimental Design

In this section, we explain the experiment (3.1), describe the data we collected from passive cell phone logs (3.2) as well as the labels we generated from EQ-5D self-reports (3.3), and provide descriptive statistics on the participants and the questionnaires (3.4).

3.1. Recruitment and Participant Breakdown

This study included adults who regularly attended physician appointments at a local Pain Clinic. They were recruited by their physician during appointments or by research staff during treatment visits. The study and its goals were explained to obtain the patient's consent.

Ninety-one patients consented to the study and were guided to download the iFeel app. Once downloaded, the app initiated collecting relevant passively sensed data from the mobile phone and EQ-5D questionnaires. Research assistants trained the participants on filling out the questionnaires. Each patient filled out their first questionnaire in the presence of the research assistants. Afterward, patients could fill out the questionnaires as much as they liked on their own time. Per the research protocol, they were asked to do so weekly and received weekly reminders from the research assistants. The research assistants also contacted participants during the follow-up period in case of missing active or passive data collection and assisted with technical or adherence issues.

In addition, participants manually filled out the SF-36 questionnaire in the presence of the research assistant three times during the study - in the beginning, middle, and end.

3.2. The iFeel App: Real-Time Mobile Data

iFeel is an innovative digital health research platform that enables passive and active digital monitoring and provides continuous objective measurements for any disorder. iFeel harnesses mobile phone data to find associations between digital markers and HRQL obtained via periodic EQ-5D self-reports. The passively sensed data is collected continuously, 24/7.

For this study, the data consisted of communication patterns (number of phone calls, duration of incoming and outgoing calls, communication app usage duration), device usage (power on/off, doze mode in/out, number of device screen opens and locks, Wi-Fi connections, Bluetooth connections, battery usage, network mobile on/off, airplane mode on/off), and app usage (name of app and seconds of active app usage). All the collected information had non-identifiable information and fully complied with the General Data Protection Regulation (GDPR). No app-specific data, content, texts, or other sensitive data was collected. Per the study's protocol, data were collected over six months from the intake date.

The study assessment battery, including demographic information and self-report outcome measures, was completed online through a secure online survey. The mobile app ensures user privacy and anonymity with proper security data server protection software (e.g., no data could be paired to a specific user). Privacy and information security comply with the European Union privacy requirements. The data did not include any other identifying information. The research coordinator recorded the key to identifying details on a secure log, which was kept separately and password protected.

3.3. Analysis of EQ-5D

The EQ-5D-5L questionnaire was available for patients to fill out on the iFeel app on a weekly basis. This questionnaire is clinically validated for self-completion on mobile devices in our country and mother tongue (Horowitz et al., 2010).

Since no value set has been published for our country, recent studies have used the United States (US) value set (Edelstein et al., 2023; Elsinga et al., 2023; Hafi et al., 2021; Ahmad et al., 2020). It considers pain/discomfort to be the most critical dimension and is, therefore, suitable for our investigation of patients with chronic pain. The US value set is more stringent, with values ranging from -0.573 to 1, compared to the United Kingdom value set (Pan et al., 2022) that has been used before (Abu Farha et al., 2017; Khatib et al., 2018). Considering our patients with chronic pain were more likely to have worse health compared to the general population, we decided to use the US value set.

We converted all health states to indices using the US value set. Then, we classified the values into two classes, indicating whether a patient's health state is worse than death: class 0, for negative indices, representing health states worse than death; and class 1, for indices that are at least 0, representing regular (not necessarily good) health.

3.4. Statistical Analysis and Train-Test Split

The patient sample included participants aged 18-80, with an average age of 60. Female participants accounted for 52.7% of the participants (48/91), and males accounted for 47.3% (43/91). As illustrated in Figure 1, men and women reported similar USA index levels in their EQ-5D questionnaires during the study. The same was true for the pain domain. Two Generalized Mixed Linear Models (GLMMs) (Chen, 2024) confirmed that there were indeed no statistically significant gender differences in the reported USA index levels (P=0.376) and pain levels (P=0.487).

On some occasions, patients filled out more than one EQ-5D questionnaire per day. In such cases, we

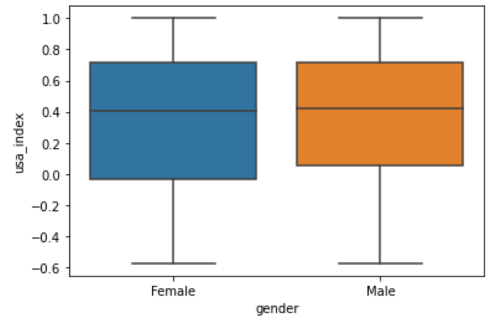


Figure 1: Box Plots of EQ-5D Health per Gender

kept the latest one. Still, each participant had multiple observations - health states reported on different dates and their corresponding cellular usage data. To avoid data leakage, all labels of the same patient were assigned to either the train or the test set.

To avoid bias, we aligned male-female and class "0" out of total proportions between the sets. Multiple Z-tests for independent samples showed no difference in gender or class "0" between the train and test sets (P>0.1). Table 1 presents the number of labels for males and females in each class and each set.

A total of 188 SF-36 questionnaires were filled out during the study, 86 at the beginning, 43 in the middle, and 59 at the end. Regarding the pain domain, patients reported extremely low outcomes (indicating severe pain) with an average of 29.6. A third GLMM confirmed that there were no statistically significant differences in the reported pain levels between men and women (P=0.788) throughout the study. However, patients reported better health (less pain) in the middle and the end of the study compared to the beginning (P=0.002 and 0=0.001, respectively). Figure 2 shows the upward trend of the mean SF-36 pain levels within the three study phases.

To assess the correlation between the two questionnaires, specifically in their ability to measure pain, we investigated the reported outcomes of patients who filled out both questionnaires within the same week. Overall, 113 patients filled out the EQ-5D and the SF-36 questionnaires within the same week: 84 on week 0 when the study started, 19 on week 9/10 in the middle of the study, and 10 on week 18, the final week of the study. The two measures use different scales - SF-36 is scaled between 0 (worst health) and 100 (best health), whereas the EQ-5D-5L is scaled between 1 (best health) and 5 (worst health). Hence, we calculated the Spearman correlation between the

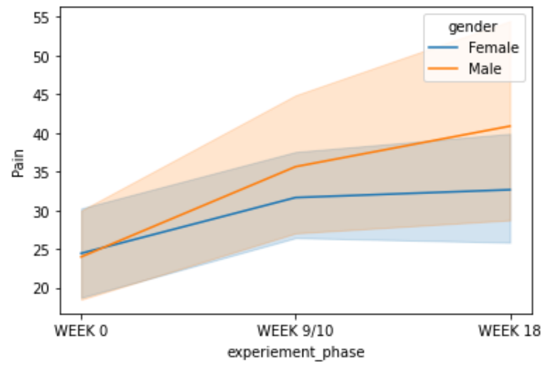


Figure 2: Mean SF-36 Pain by Experiment Phase

SF-36 bodily pain concept and the *reversed* ED-5D-5L pain/discomfort domain, which was 0.57.

To provide insights not just in terms of the EQ-5D index but also in SF-36 pain scales, we assessed the influence of our binary EQ-5D labels on the reported SF-36 pain scores. Comparing the mean SF-36 pain scores across our two label classes, revealed the binary index's ability to distinguish between SF-36 pain levels. The mean SF-36 pain score was significantly lower for the negative EQ-5D indexes (13.7) compared to the non-negative ones (30.7).

Table 1: Train-Test Split

		Class 0	Class 1	All
Train	Male	21 (26.9%)	57 (73.1%)	246
	Female	45 (26.8%)	123 (73.2%)	
Test	Male	13 (22.0%)	46 (78.0%)	134
	Female	27 (36.0%)	48 (64.0%)	

4. EQ-5D Prediction from Cellular Data

We aimed to predict EQ-5D health indices from measurements passively collected by the iFeel app from patients' cell phones. These included app usage (app type and duration) and cellphone events (e.g., the screen on/off) and were collected continuously during the clinical study.

Our aim and setting pose a few inherent challenges:

- Data's nature – many apps and events can be monitored using the iFeel app. However, at each point, only a handful are active. Thus, the data is highly sparse. Also, we did not collect any specific app data or content to preserve privacy. Only indications for app usage were used.
- Dataset size – data was collected continuously, 24/7, for six months, but for a relatively small number of subjects. Thus, we had a reasonable amount of data per patient (intra-patient examples) but a relatively small amount of cross-patient data (inter-patient sequences). Also, there was a fair amount of missing data due to connectivity issues, version updates, etc.
- Labels – our labels were based on patient subjective responses to the EQ-5D questionnaire. We had various numbers of labels per patient, as each patient decided when and how many times to fill out their questionnaire. Moreover, the self-reports reflected patients' own views of their health with their unique scaling.
- Data and label alignment – syncing between labels derived from the (higher level) EQ-5D responses and the (low level) cellphone features is also challenging: An EQ-5D response reflects periodical patient condition. In contrast, the high-frequency cell phone data demonstrates the patient's instantaneous activities and state. Hence, the self-assessment labels and the collected cellphone measurements act at different time scales and reflect different levels of information. Thus, duplicating the same label over multiple adjacent feature vectors or labeling just one instance and utilizing semi-supervised methods were unsuitable solutions.

We addressed these challenges via a carefully designed preprocessing step, followed by a two-step classification method to maximize the utilization of labeled and unlabeled data. At the preprocessing step, we first mapped the various apps into predefined app types. Then, we aggregated the data on an hourly basis and constructed feature vectors where each feature recorded hourly usage (in seconds) of a specific app type. Applying these actions reduced the feature space dimension and sparsity.

To sync between the labels and the feature vectors, we grouped the feature vectors from the preceding week for each self-report (label). Thus, a labeled example was a set of 181 feature vectors with a single EQ-5D index score, i.e., the set's label. It should be noted that for all patients, most weeks had no label since the patient did not fill out a questionnaire.

Hence, we ended with many unlabeled feature vectors and a few labeled sets.

To overcome these unique characteristics of our problem, we designed a dedicated two-step method for learning behavioral patterns and deviations. First, we use unsupervised learning techniques to utilize the immense amounts of unlabeled cellphone usage data. As described in subsection 4.1, we cluster the vectors and change their representation to capture hourly behavioral patterns. Then, in 4.2, we train supervised classification models to predict the class of the labeled vectors.

4.1. Unsupervised Representation Learning

Our pipeline starts by clustering the (unlabeled) feature vectors using the K-means clustering algorithm. The outcome is a set of clusters, each reflecting a typical hourly behavior for this patient population. The pipeline includes two optional transformations preceding the clustering step. Both are designed to refine the clusters by reducing dimensionality and sparsity and making the clusters more distinct.

The first transformation, termed Pre-Clustering Transformation (PCT), is a deep autoencoding, followed by Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) learning of the encoder's latent representation. The PCT component transforms the sparse, high-dimensional raw feature vectors into a dense, lower-dimensional feature vectors embedded in a smooth manifold. This transformation contributes to the stability of the proceeding clustering procedure.

The second transformation, termed Discriminative Clustering Transformation (DCT), starts by clustering and mapping the feature vectors to the nearest clusters. It provides a pseudo label (the mapped cluster) for each feature vector. Then, we train a two-header autoencoder-classifier model (a.k.a multi-task learning (Zhang and Yang, 2021)) and extract the latent representation. The DCT component essentially implements a self-learning procedure designed to increase the separability of the latent representation before the (original) unsupervised clustering procedure.

In the next pipeline step, after clustering, each feature vector is mapped to the closest cluster, as measured by Mahalanobis distance (De Maesschalck et al., 2000). The distances from each cluster were binned into ten quantiles, and each vector ascribed to the cluster belonged to one of them. Thus, each vector was mapped to a specific distance quantile in its closest cluster. Unique vectors (representing deviations from common behavioral patterns) were mapped to higher quantiles. As a result, each labeled week (with a self-assessment) was projected onto the obtained clusters, such that each hourly vector within the week was assigned to its corresponding quantile in the nearest cluster.

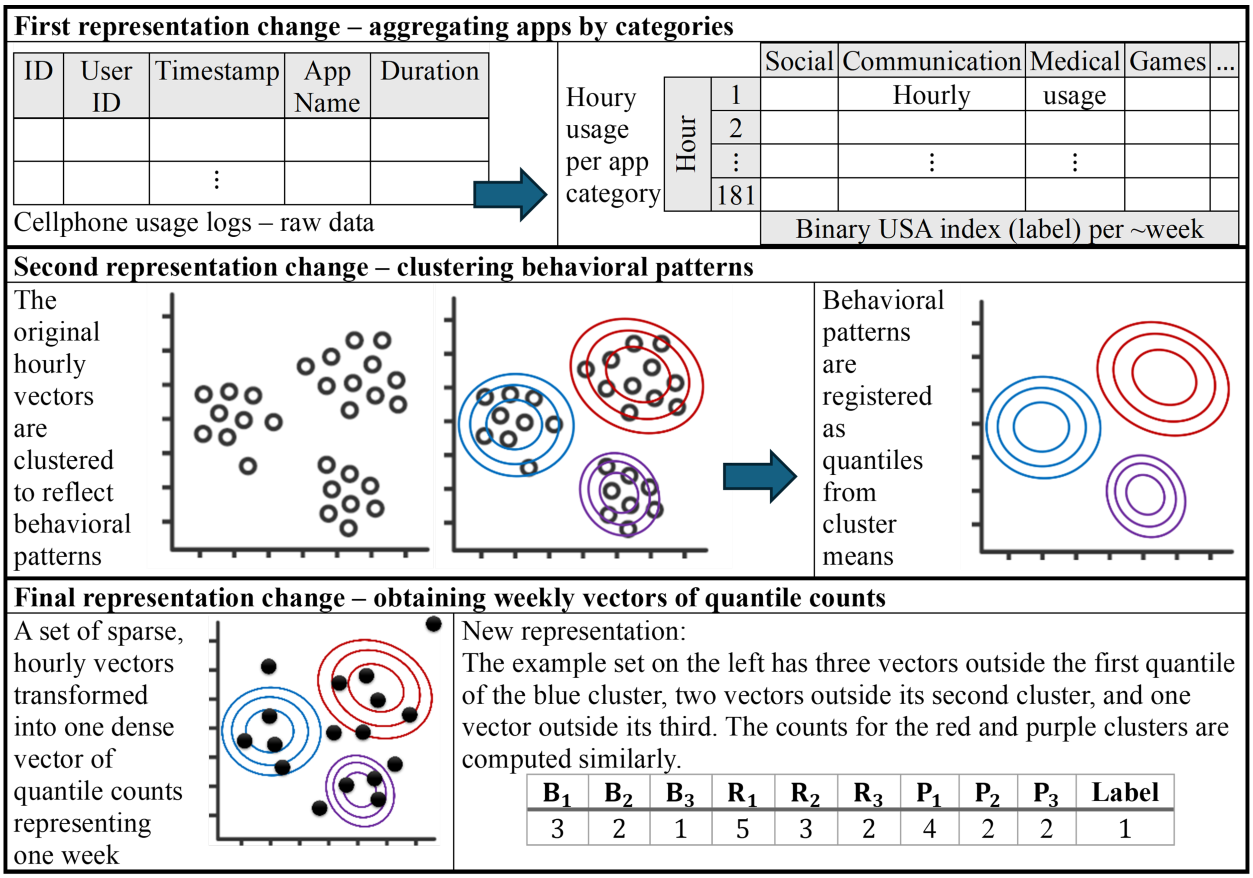
Finally, every labeled set of 181 hourly feature vectors was represented by a single vector of counts. The counts recorded the number of vectors assigned to each quantile of each cluster. The outcome was a labeled dense vector representing an entire week, where each entry corresponded to a specific quantile holding the aggregated counts of hourly feature vectors that were mapped to higher quantiles. In essence, this serves as an empirical approximation of the feature vectors' inverse CDF, capturing typical behavioral patterns and deviations throughout the week. Figure 3 illustrates the construction of representation from the raw data, through the behavioral clustering, to the count vector.

4.2. Supervised Classification

Standard supervised learning methods and best practices were applied at this step. After further preprocessing the data using feature selection, traditional classification methods were applied: logistic regression (LR), random forest (RF), and k-nearest neighbors (kNN). The outcome was a model designed to predict a patient's state given behavioral patterns from the preceding week, as captured by their cellphone usage.

We trained separate models for men and women, as well as gender-agnostic models for comparison. For each gender - Male, Female, All - all possible combinations of pre-clustering transformation, K, and classification model were evaluated.

For model evaluation, we used a test set formed by cellphone data and self-assessment labels of patients excluded from both the representation learning and the classifier training. Hyperparameter optimization was performed using grid search cross-validation on the training set, while the held-out test set was used to evaluate model performance on unseen data. In section 5, we report performance on the test set and provide statistical guarantees using Fisher and Barnard's tests.



B_1	B_2	B_3	R_1	R_2	R_3	P_1	P_2	P_3	Label
3	2	1	5	3	2	4	2	2	1

Figure 3: Two-step Representation. Mapping hourly vectors to behavioral clusters. B_i, R_i, P_i: Number of vectors outside the i-th quantile of the blue, red, and purple clusters, respectively.

5. Results and Discussion

5.1. Representation Results

The raw cellphone data were mapped to 56-dimension feature vectors representing hourly app type usage (in seconds). After cleansing, we had an unlabeled dataset of 95165×56 vectors.

Using the k-means clustering algorithm with 50 initial starts, we transformed the labeled example sets into $K \times 9$-dimension feature vectors in the representation change phase, and ended up with 246 labeled vectors in the training set and 134 labeled vectors in the test set.

5.2. Classification Results

We present the results of gender-specific modeling, which outperformed models trained on both men's and women's data. Table 2 presents the best classification results for each gender (Male, Female, All) and each pipeline (Standard, PCT, DCT, or both) with the chosen clustering K (16, 24, 32, 64, or 128) and classification model (RF, LR, or kNN). Note that the "Standard" pipeline refers to the clustering phase without any optional preceding variants, PCT or DCT, as described in subsection 4.2. To help identify powerful settings, the highest values for each metric appear in bold. The table reflects the robustness of our classification results, as they are similar across all settings.

For the men, the **PCT with $K = 32$ and kNN** setting achieved the best results. The $K = 32$ was chosen for the K-means clustering algorithm, and the kNN model used the k=5 nearest neighbors for the classification. Even though it did not achieve the best recall on class 0 or precision on class 1, it performed best according to all the other metrics and appeared to be the most stable setting. Overall, the kNN algorithm performed well in our task, as it frequently appeared in the classification results table.

As for the women, the most stable setting was **PCT with $K = 32$ and RF**. It significantly outperformed all other models in every F1-score measure while maintaining high precision and recall values.

Both Male and Female models consistently outperformed the gender-agnostic models (All). While gender-specific modeling should be considered when applicable, gender-agnostic modeling can help with non-binary cases or when app users prefer not to mention their gender.

Fisher and Barnard's tests were used to determine the classification significance of our best settings: PCT with kNN for the men and PCT with RF for the women. As both tests showed highly statistically significant results (Fisher test P$\leq 5\times10^{-6}$, Barnard's test P$\leq 2 \times 10^{-6}$), it is unlikely to receive our predictions purely by chance.

5.3. Discussion

Our analysis and results point to differences between men and women in terms of willingness to share health-related information but not in their reported health. Even though the participants' male-female ratio was close to 1 (43 vs. 48, respectively), there were significantly more self-reports by women, 243 EQ-5D and 111 SF-36, compared to 137 and 77 by men. Nonetheless, in terms of pain *severity*, there were no statistically significant gender differences in either questionnaire.

The binary USA index of the EQ-5D questionnaire turned out to be a useful predictor for severe pain in the SF-36 questionnaire. The mean SF-36 pain score was significantly lower for the negative EQ-5D indexes (13.7) compared to the non-negative ones (30.7), and according to the GLMM, the binary index highly affected the SF-36 pain levels (P<0.001).

The study population - patients suffering from chronic pain, reported extremely low SF-36 pain scores. This finding is consistent with previous work using the SF-36 questionnaire to monitor patients with chronic back pain, showing that the physical and mental component scales of the chronically disabled back pain population were significantly lower compared to the United States norms for back pain/sciatica patients (Gatchel et al., 1998).

In the study's setting, the labels were subjective as they were generated from patient-reported outcomes; there were multiple labels per patient, and each patient had a different number of labels, filling out a different number of questionnaires throughout the study. Moreover, gender-specific classification models consistently outperformed generic (gender-agnostic) models in all tested settings. Hence, personalized, patient-specific models are worth testing to improve the predictions further.

Several studies have demonstrated individual differences in smartphone-derived behaviors, highlighting the need to account for age and gender in digital phenotyping models (Zhang et al., 2022; Laiou et al., 2022). For example, females tend to use smartphones

Table 2: Best Classification Results per Gender and Pre-clustering Transformation

	Setting			F1-Score			Class 0		Class 1		AUC
Gender	Pipeline	K	Model	Acc.	M. avg.	W. avg.	Prec.	Recall	Prec.	Recall	ROC
Male	Standard	24	kNN	0.797	0.607	0.760	0.600	0.231	0.815	0.957	0.594
Male	PCT	32	kNN	**0.898**	**0.832**	**0.891**	**0.889**	0.615	0.900	**0.978**	0.797
Male	DCT	16	kNN	0.831	0.699	0.810	0.714	0.385	0.846	0.957	0.671
Male	PCT+DCT	24	LR	0.831	0.777	0.838	0.588	**0.769**	**0.929**	0.848	**0.809**
Female	Standard	16	kNN	0.787	0.755	0.780	0.762	0.593	0.796	**0.896**	0.744
Female	PCT	32	RF	**0.827**	**0.807**	**0.824**	**0.792**	0.704	0.843	**0.896**	**0.800**
Female	DCT	32	RF	0.773	0.767	0.778	0.639	**0.852**	**0.897**	0.729	0.791
Female	PCT+DCT	24	LR	0.787	0.778	0.790	0.667	0.815	0.881	0.771	0.793
All	Standard	24	LR	0.754	0.704	0.753	0.590	**0.575**	**0.821**	0.830	**0.702**
All	PCT	16	kNN	0.739	0.637	0.714	0.609	0.350	0.766	**0.904**	0.627
All	DCT	32	LR	**0.761**	**0.706**	**0.757**	**0.611**	0.550	0.816	0.851	0.701
All	PCT+DCT	128	kNN	0.724	0.624	0.702	0.560	0.350	0.761	0.883	0.616

Acc. = Accuracy; M. avg. = Macro average; W. avg. = Weighted average; Prec. = Precision.

for longer periods than males, and younger individuals are more likely to engage with entertainment and social media apps, while older adults primarily use smartphones for information retrieval or as a traditional communication tool (Andone et al., 2016). While demographic factors have not always influenced engagement or compliance with mobile health interventions (Ross et al., 2020), they may still shape passively collected behavioral patterns, which could impact model predictions.

In the context of chronic pain, smartphone behavioral patterns may also be influenced by pain severity and disability levels. For instance, individuals with higher pain intensity or greater pain-related disability have been found to engage more with mobile health tools (Ross et al., 2020). Additionally, lifestyle factors such as working remotely may shape digital behaviors (e.g., exhibiting low mobility and frequent phone use). Yet, these patterns may not directly correlate with pain state (Leaning et al., 2024). Despite these potential biases, our study accounts for gender-specific variations, ensuring that predictions remain sensitive to demographic differences while leveraging the advantages of smartphone-based monitoring.

In this study, a static modeling approach that uses a fixed layout of hourly clusters was used to represent behavior patterns. Labels of the same user were considered independent, even though self-assessments likely correlate. Feature vectors constructed from cellphone usage data were also dependent due to the inherent sequential nature of the data. Both dependencies suggest room for stochastic models that learn patient behavior over time. Liu et al. (2025) derived digital phenotyping from wearable devices to identify individuals with mental health disorders, demonstrating the advantage of using time series to capture the data's time-varying nature and preserve temporal patterns. In future work, we intend to explore the contribution of a dynamic modeling approach that accounts for transitions between behavior clusters.

It should be noted that the iFeel app is only supported by Android mobile phones. Hence, this study focused on Android mobile users, which limits the representativeness and generalizability of the findings. However, the two-step method described in this paper is adaptable and applicable to different domains. An earlier version of the method, without the PCT and DCT transformers for clustering enhancement, was successfully used in a previous study aiming to predict the imminent suicide risk of adolescents previously diagnosed with depression (Stemmer et al., 2024). A new massive-scale study, currently recruiting teenagers from different schools across multiple countries, will integrate the method to predict the risk of problematic internet use. This ongoing project uses a similar mobile app to track user usage patterns, which is also available on iOS. Hence, the findings might extend to other smartphone platforms, allowing the model deployment in diverse, real-world healthcare settings.

The clinical insights of our results are twofold. The first highlights the prediction of the SF-36 severe score (36 questions) out of the EQ-5D/USA Index (5 questions), allowing clinicians to identify severe cases via rapid and digitally validated questionnaires. The second is the ability to detect patients with severe pain conditions via passive and remote tools.

The findings serve as a strong foundation for real-world model deployment. Clinicians will greatly benefit from the predictors derived from the research as an additional step toward more objective pain and quality-of-life assessments for pain patients. By using objective and continuous parameters, these predictors will allow for a more precise real-time evaluation of the effectiveness of both pharmacological and interventional treatments, improving the cost-effectiveness of various treatments.

The strength of the study's predictors lies in their ability to enhance both pain and quality of life assessments for patients and clinicians alike. We envision the model as a resource that can be integrated into clinicians' workflows, providing objective tools for more accurate evaluations of treatment effectiveness. One of the main challenges in chronic pain management is the infrequency of patient visits, which typically occur once every few months, making it difficult to consistently track patient progress (de C Williams et al., 2020). The breakthrough of real-time monitoring offered by these predictors provides clinicians with continuous and up-to-date data, allowing them to adjust pharmacological and interventional treatments more effectively, thus improving outcomes (Rejula et al., 2021). Simultaneously, patients benefit from a system that gives them feedback on their condition, enabling them to track their progress in real time between visits.

6. Conclusion

This study uses passively collected cellphone usage data to predict the state of patients suffering from chronic pain, as captured by subjective self-reported EQ-5D questionnaires. It suggests a two-step approach for utilizing unlabeled cellular logs in a classification problem via unsupervised learning. Previous studies have used machine learning to predict patient EQ-5D scores based on patient and disease characteristics (Lee et al., 2014; Zrubka et al., 2022), but to the best of our knowledge, this is the first study to do so based on cellphone logs.

Unlike self-reports, which are subjective measures to assess health states in general and pain specifically, cellphone monitoring provides objective indicators crucial for more accurate predictions. Moreover, prediction by *passive* cellphone usage data, as used in this study, is less burdensome than self-reports, which rely on compliance over time. The objective, passive measures overcome the variance in patient perception of their own health and pain.

The high prediction levels reached in this study provide additional validation for incorporating digital monitoring in the clinical field. It may equip clinicians with tools to assess patient health status constantly and objectively, identify risks, and provide timely interventions.

The setting of our study posed three interesting challenges: the relatively small labeled dataset compared to the massive amounts of unlabeled cellphone data, the different time scales and information levels of self-assessments and cellphone measurements, and the subjectivity of the labels. The first two are addressed in this paper with representation learning. The latter is left for future research.

References

Nader H Abu Farha, Mohammed T Khatib, Husam Salameh, and Sa'ed H Zyoud. Cancer-related post-treatment pain and its impact on health-related quality of life in breast cancer patients: a cross sectional study in palestine. *Asia Pacific Family Medicine*, 16:1–9, 2017.

Qais T Ahmad, Jaffar H Saffarini, Ahmad M Samara, Dima S Jabri, Zaina H Safarini, Yousra M Banijaber, Ahmad Jaradat, Faris Abushamma, and Sa'ed H Zyoud. The impact of lower urinary tract symptoms on the quality of life during pregnancy: a cross-sectional study from palestine. *BMC urology*, 20:1–10, 2020.

Ionut Andone, Konrad Błaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. How age and gender affect smartphone usage. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: adjunct*, pages 9–12, 2016.

Min S Hane Aung, Faisal Alquaddoomi, Cheng-Kang Hsieh, Mashfiqui Rabbi, Longqi Yang, John P Pollak, Deborah Estrin, and Tanzeem Choudhury. Leveraging multi-modal sensing for mobile health: a case review in chronic pain. *IEEE journal of selected topics in signal processing*, 10(5):962–974, 2016.

Sara Laureen Bartels, Angeliki Pelika, Afra S Taygar, and Rikard K Wicksell. Digital approaches to chronic pain: A brief meta-review of ehealth interventions-current evidence and future directions. *Current Opinion in Psychology*, page 101976, 2024.

Siyuan Chen. *Modeling Aging Trajectories and Urban-rural Disparities in Chronic Pain among Middle-aged and Older Chinese: Evidence from the Generalized Estimating Equations (GEE) and Generalized Linear Mixed Models (GLMMs)*. PhD thesis, Ghent University, 2024.

Steven P Cohen, Lene Vase, and William M Hooten. Chronic pain: an update on burden, best practices, and new advances. *The Lancet*, 397(10289):2082–2097, 2021.

Amanda C de C Williams, Emma Fisher, Leslie Hearn, and Christopher Eccleston. Psychological therapies for the management of chronic pain (excluding headache) in adults. *Cochrane database of systematic reviews*, (8), 2020.

Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.

Nancy Devlin, David Parkin, and Bas Janssen. *Methods for analysing and reporting EQ-5D data*. Springer Nature, 2020a.

Nancy Devlin, David Parkin, Bas Janssen, Nancy Devlin, David Parkin, and Bas Janssen. An introduction to eq-5d instruments and their applications. *Methods for analysing and reporting EQ-5D data*, pages 1–22, 2020b.

Clare Dominick, Fiona Blyth, and Michael Nicholas. Patterns of chronic pain in the new zealand population. *The New Zealand Medical Journal (Online)*, 124(1337), 2011.

Hannah Dorris, Jenny Oh, and Nicholas Jacobson. Wearable movement data as a potential digital biomarker for chronic pain: an investigation using deep learning. *Physical Activity and Health*, 8(1), 2024.

Michael Edelstein, Paul Kuodi, Yanay Gorelik, Hiba Zayyad, Ofir Wertheim, Karine Beiruti Wiegler, Kamal Jabal, Amiel Dror, Jelte Elsinga, Saleh Nazzal, et al. Association between bnt162b2 vaccination and health-related quality of life up to 18 months post-sars-cov-2 infection in israel: A cross sectional survey. 2023.

European Pain Federation EFIC. The societal impact of pain, 2022. URL https://europeanpainfederation.eu/sip/.

Thomas E Elliott, Colleen M Renier, and Jeanette A Palcher. Chronic pain, depression, and quality of life: correlations and predictive value of the sf-36. *Pain medicine*, 4(4):331–339, 2003.

Jelte Elsinga, Paul Kuodi, Haneen Shibli, Yanay Gorelik, Hiba Zayyad, Ofir Wertheim, Kamal Abu Jabal, Amiel Dror, Saleh Nazzal, Daniel Glikman, et al. Changes in quality of life following sars-cov-2 infection among jewish and arab populations in israel: A cross-sectional study. *International Journal of Public Health*, 68:1605970, 2023.

Robert J Gatchel, Peter B Polatin, Tom G Mayer, Richard Robinson, and Jeffrey Dersh. Use of the sf-36 health status survey with a chronically disabled back pain population: strengths and limitations. *Journal of Occupational Rehabilitation*, 8:237–246, 1998.

The EuroQol Group. Euroqol-a new facility for the measurement of health-related quality of life. *Health policy*, 16(3):199–208, 1990.

Yacine Hadjiat and Lars Arendt-Nielsen. Digital health in pain assessment, diagnosis, and management: Overview and perspectives. *Frontiers in Pain Research*, 4:1097379, 2023.

Yacine Hadjiat and Serge Perrot. Cancer pain management in french-speaking african countries: assessment of the current situation and research into factors limiting treatment and access to analgesic drugs. *Frontiers in Public Health*, 10:846042, 2022.

Eba'a Hafi, Ro'ya Soradi, Sarah Diab, Ahmad M Samara, Marah Shakhshir, Malik Alqub, and Sa'ed H Zyoud. Nutritional status and quality of life in diabetic patients on hemodialysis: a cross-sectional study from palestine. *Journal of Health, Population and Nutrition*, 40:1–11, 2021.

Mónica Hernández Alava, Steve Pudney, and Allan Wailoo. Estimating the relationship between eq-5d-5l and eq-5d-3l: results from a uk population study. *Pharmacoeconomics*, 41(2):199–207, 2023.

Einav Horowitz, Ifat Abadi-Korek, Mordechai Shani, and Joshua Shemer. Eq-5d as a generic measure of health-related quality of life in israel: reliability, validity and responsiveness. *IMAJ-Israel Medical Association Journal*, 12(12):715, 2010.

Aparecida Mari Iguti, Margareth Guimarães, and Marilisa Berti Azevedo Barros. Health-related quality of life (sf-36) in back pain: a population-based study, campinas, são paulo state, brazil. *Cadernos de saude publica*, 37:e00206019, 2021.

Melinda Járomi, Brigitta Szilágyi, Anita Velényi, Eleonóra Leidecker, Bence László Raposa, Márta Hock, Petra Baumann, Pongrác Ács, and Alexandra Makai. Assessment of health-related quality of life and patient's knowledge in chronic non-specific low back pain. *BMC Public Health*, 21:1–8, 2021.

Regina JM Kamstra, André Boorsma, Tanja Krone, Robin M van Stokkum, Hannah M Eggink, Ton Peters, and Wilrike J Pasman. Validation of the mobile app version of the eq-5d-5l quality of life questionnaire against the gold standard paper-based version: Randomized crossover study. *JMIR Formative Research*, 6(8):e37303, 2022.

Daniel Kelly, Kevin Curran, and Brian Caulfield. Automatic prediction of health status using smartphone-derived behavior profiles. *IEEE journal of biomedical and health informatics*, 21(6): 1750–1760, 2017.

Sohaib T Khatib, Mohammad K Hemadneh, Samer A Hasan, Emad Khazneh, and Sa'ed H Zyoud. Quality of life in hemodialysis diabetic patients: a multicenter cross-sectional study from palestine. *BMC nephrology*, 19:1–9, 2018.

Petroula Laiou, Dzmitry A Kaliukhovich, Amos A Folarin, Yatharth Ranjan, Zulqarnain Rashid, Pauline Conde, Callum Stewart, Shaoxiong Sun, Yuezhou Zhang, Faith Matcham, et al. The association between home stay and symptom severity in major depressive disorder: preliminary findings from a multicenter observational study using geolocation data from smartphones. *JMIR mHealth and uHealth*, 10(1):e28095, 2022.

Imogen E Leaning, Nessa Ikani, Hannah S Savage, Alex Leow, Christian Beckmann, Henricus G Ruhé, and Andre F Marquand. From smartphone data to clinically relevant predictions: A systematic review of digital phenotyping methods in depression. *Neuroscience & Biobehavioral Reviews*, 158: 105541, 2024.

Soo-Kyoung Lee, Youn-Jung Son, Jeongeun Kim, Hong-Gee Kim, Jae-Il Lee, Bo-Yeong Kang, Hyeon-Sung Cho, and Sungin Lee. Prediction model for health-related quality of life of elderly with chronic diseases using machine learning techniques. *Healthcare informatics research*, 20(2):125, 2014.

Huang-Li Lin, Wei-Yang Lee, Wei-Yu Chiang, Tsai-Sheng Fu, Wen-Chien Chen, and Ching-I Hung. Bodily pain and vitality are the key factors in the disability of chronic low back pain patients under short form 36 base study: a five-year cohort study. *Health and Quality of Life Outcomes*, 22(1): 88, 2024.

Jason J Liu, Beatrice Borsari, Yunyang Li, Susanna X Liu, Yuan Gao, Xin Xin, Shaoke Lou, Matthew Jensen, Diego Garrido-Martín, Terril L Verplaetse, et al. Digital phenotyping from wearables using ai characterizes psychiatric disorders and identifies genetic associations. *Cell*, 188(2):515–529, 2025.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K D'Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–21, 2019.

Tahsin Mullick, Ana Radovic, Sam Shaaban, Afsaneh Doryab, et al. Predicting depression in adolescents using mobile and wearable sensors: Multimodal machine learning–based exploratory study. *JMIR Formative Research*, 6(6):e35807, 2022.

MJ Oemar. Basic information on how to use the eq-5d-5l instrument. Technical report, EQ-5D-5L User Guide Version 20, 2013.

Tianxin Pan, Brendan Mulhern, Rosalie Viney, Richard Norman, Janel Hanmer, and Nancy Devlin. A comparison of propr and eq-5d-5l value sets. *Pharmacoeconomics*, pages 1–11, 2022.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

A Simon Pickard, Ernest H Law, Ruixuan Jiang, Eleanor Pullenayegum, James W Shaw, Feng Xie, Mark Oppe, Kristina S Boye, Richard H Chapman, Cynthia L Gong, et al. United states valuation of eq-5d-5l health states using an international protocol. *Value in health*, 22(8):931–941, 2019.

Mashfiqui Rabbi, Min SH Aung, Geri Gay, M Cary Reid, and Tanzeem Choudhury. Feasibility and acceptability of mobile phone–based auto-personalized physical activity recommendations for chronic pain self-management: Pilot study on adults. *Journal of medical Internet research*, 20 (10):e10147, 2018.

Rosalind Rabin and Frank de Charro. Eq-sd: a measure of health status from the euroqol group. *Annals of medicine*, 33(5):337–343, 2001.

Quazi Abidur Rahman, Tahir Janmohamed, Meysam Pirbaglou, Hance Clarke, Paul Ritvo, Jane M Heffernan, and Joel Katz. Defining and predicting pain volatility in users of the manage my pain app: analysis using data mining and machine learning methods. *Journal of medical Internet research*, 20(11): e12001, 2018.

V Rejula, J Anitha, RV Belfin, and J Dinesh Peter. Chronic pain treatment and digital health era-an opinion. *Frontiers in Public Health*, 9:779328, 2021.

Edgar L Ross, Robert N Jamison, Lance Nicholls, Barbara M Perry, and Kim D Nolen. Clinical integration of a smartphone app for patients with chronic pain: retrospective analysis of predictors of benefits and patient engagement between clinic visits. *Journal of medical Internet research*, 22(4): e16939, 2020.

Alex Novaes Santana, Charles Novaes de Santana, and Pedro Montoya. Chronic pain diagnosis using machine learning, questionnaires, and qst: a sensitivity experiment. *Diagnostics*, 10(11):958, 2020.

Maya Stemmer, Shira Barzilay, Itamar Efrati, Talia Friedman, Lior Carmi, Mishael Zohar, Anat Brunstein Klomek, Alan Apter, and Shai Fine. Predicting adolescent suicide risk from cellphone usage data and self-report assessments. *Proceedings of the 57th Hawaii International Conference on System Sciences*, 2024.

Cynthia J Thomson, Hanna Pahl, and Luisa V Giles. Randomized controlled trial investigating the effectiveness of a multimodal mobile application for the treatment of chronic pain. *Canadian Journal of Pain*, 8(1):2352399, 2024.

Nicola Torrance, Blair H Smith, Amanda J Lee, Lorna Aucott, Amanda Cardy, and Michael I Bennett. Analysing the sf-36 in population-based research. a comparison of methods of statistical approaches using chronic pain as an example. *Journal of Evaluation in Clinical Practice*, 15(2):328–334, 2009.

Kotryna Vereščiagina, Kazys Vytautas Ambrozaitis, and Bronius Špakauskas. Health-related quality-of-life assessment in patients with low back pain using sf-36 questionnaire. *Medicina*, 43(8):607, 2007.

John E Ware, Kristin K Snow, Mark Kosinski, and Barbara Gandek. Sf-36 health survey. *Manual and interpretation guide*, 2, 1993.

John E Ware Jr. Sf-36 health survey update. *Spine*, 25(24):3130–3139, 2000.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

Yuezhou Zhang, Amos A Folarin, Shaoxiong Sun, Nicholas Cummins, Srinivasan Vairavan, Rebecca Bendayan, Yatharth Ranjan, Zulqarnain Rashid, Pauline Conde, Callum Stewart, et al. Longitudinal relationships between depressive symptom severity and phone-measured mobility: dynamic structural equation modeling study. *JMIR mental health*, 9(3):e34898, 2022.

Peng Zhao, Illhoi Yoo, Robert Lancey, and Ebby Varghese. Mobile applications for pain management: an app analysis for clinical usage. *BMC Medical Informatics and Decision Making*, 19:1–10, 2019.

Zsombor Zrubka, István Csabai, Zoltán Hermann, Dominik Golicki, Valentina Prevolnik-Rupel, Marko Ogorevc, László Gulácsi, and Márta Péntek. Predicting patient-level 3-level version of eq-5d index scores from a large international database using machine learning and regression methods. *Value in Health*, 25(9):1590–1601, 2022.