Emerging Applications

Predicting Subjective Recovery from Lower Limb Surgery Using Consumer Wearables

Marta Karas^{a, b} Nikki Marinsek^a Jörg Goldhahn^c Luca Foschini^a
Ernesto Ramirez^a Ieuan Clay^a

^aEvidation Health Inc., San Mateo, CA, USA; ^bDepartment of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA; ^cInstitute of Translational Medicine, Department of Health Sciences and Technology, Eidgenössische Technische Hochschule (ETH), Zurich, Switzerland

Keywords

Person-generated health data · Activity monitors · Conventional wearable devices · Influence and/or predict health-related outcomes · Quantifiable physiological and behavioral data

Abstract

Introduction: A major challenge in the monitoring of rehabilitation is the lack of long-term individual baseline data which would enable accurate and objective assessment of functional recovery. Consumer-grade wearable devices enable the tracking of individual everyday functioning prior to illness or other medical events which necessitate the monitoring of recovery trajectories. **Methods:** For 1,324 individuals who underwent surgery on a lower limb, we collected their Fitbit device data of steps, heart rate, and sleep from 26 weeks before to 26 weeks after the self-reported surgery date. We identified subgroups of individuals who self-reported surgeries for bone fracture repair ($n = 355$), tendon or ligament repair/reconstruction ($n = 773$), and knee or hip joint replacement ($n = 196$). We used linear mixed models to estimate the average effect of time relative to surgery on daily activity measurements while adjusting for gender, age, and the participant-specific activity baseline. We used a sub-cohort of 127 individuals with dense wearable data who underwent tendon/ligament surgery and employed XGBoost to predict the self-reported recovery time. **Results:** The 1,324 study individuals were all US residents, predominantly female (84%), white or Caucasian (85%), and young to middle-aged (mean age 36.2 years). We showed that 12 weeks pre- and 26 weeks post-surgery trajectories of daily behavioral measurements (steps sum, heart rate, sleep efficiency score) can

Ieuan Clay
Research, Analysis and Learning
Evidation Health Inc.
San Mateo, CA 94401 (USA)
iclay@evidation.com

capture activity changes relative to an individual's baseline. We demonstrated that the trajectories differ across surgery types, recapitulate the documented effect of age on functional recovery, and highlight differences in relative activity change across self-reported recovery time groups. Finally, using a sub-cohort of 127 individuals, we showed that long-term recovery can be accurately predicted, on an individual level, only 1 month after surgery (AUROC 0.734, AUPRC 0.8). Furthermore, we showed that predictions are most accurate when long-term, individual baseline data are available. **Discussion:** Leveraging long-term, passively collected wearable data promises to enable relative assessment of individual recovery and is a first step towards data-driven intervention for individuals.

© 2020 The Author(s)
Published by S. Karger AG, Basel

Introduction

Digital health is delivering value across healthcare applications, including in safety monitoring, diagnostic and therapeutic applications, and particularly in monitoring responses to therapeutic intervention [1]. Such digital measures are becoming increasingly sophisticated in terms of their development for clinical usage [2], and we are seeing the first accepted and qualified examples already emerging [3, 4].

Simultaneously, consumer-facing digital technologies such as mobile devices and wearable sensors have become ubiquitous, enabling collection of person-generated health data (PGHD) on virtually all aspects of individual lifestyles and behaviors, with unprecedented potential to add rich insight on everyday human life to traditional health research [5]. Crucially, the term PGHD highlights that data are collected through all of an individual's healthcare journey, including before they become a patient [6].

Such low cost, pervasive measurement tools offer significant advantages, for example circumventing traditional access/institutional barriers to health services [7], which are globally recognized as continuing to be a major impediment to accessing healthcare [8]. However, engagement with healthcare systems, and therefore monitoring, typically only begins when an individual is diagnosed or their symptoms otherwise become so severe that they seek care [9, 10]. Prospective natural history studies are relatively rare [11], thus a further advantage of PGHD captured via consumer-grade wearables is that prediction or forecasting of outcomes can leverage data collected prior to the diagnosis or event, enabling early detection and treatment by “funneling” high-risk individuals towards proactive screening. Previous work has demonstrated the power of this idea by demonstrating that long-term, relative changes in vital sign and activity measures can be used to assess the effectiveness of weight-loss surgery [12].

Typical trajectories in functional recovery from lower limb surgeries typically take around 6 months, for example knee and hip replacement [13], or hip fracture surgery [14]. Recovery trajectories longer than 6 months are typically seen as abnormal and a trigger for further intervention [15]. Assessment of recovery is highly challenging, primarily because canonical practice provides no personalized baseline to which functional recovery can be compared. Equally, subjective (i.e., patient-reported) assessment of recovery is challenging due to individual reference perceptions and expectations of what “normal” (i.e., fully recovered function) is [16]. While evidence exists that increasing activity during rehabilitation improves recovery outcomes [15, 17], triggering these interventions is thus practically difficult.

Here, we apply the concept of assessing recovery from a range of lower limb surgeries relative to a personal baseline derived from long-term passive monitoring with consumer wearables. We establish that PGHD from consumer-grade technologies can capture, and be used to predict, long-term recovery trajectories. This work may help to identify patients at

risk for delayed rehabilitation early enough to trigger additional or more targeted rehabilitation interventions. Personalized recommendations based on individualized baseline data can be a major contribution of PGHD towards virtual healthcare.

Materials and Methods

Data Collection

Achievement, a product of Evidation Health, is an online platform where people can connect their digital health tools, including wearable activity trackers and fitness apps. Achievement members agree to be contacted with study opportunities when they create an Achievement account [18]. Participants are rewarded for participating in studies by answering survey questions and contributing data from connected devices. This platform enables rapid recruitment of participants to specific studies [19], where consent for all research is granted on a per use basis.

Data were collected from a previously cited study surveying participant experience relating to surgery and medical devices [12]. Briefly, participants were asked about which surgeries they had experienced, and for the most recent surgery, the type of surgery, the date of surgery, and the time required for recovery. The full survey is included in online supplementary Note 1 (for all online suppl. material, see www.karger.com/doi/10.1159/000511531). Between May 5 and September 21, 2018, 200,325 individuals consented to take part in the study. A total of 50,938 participants reported they underwent a medical procedure, out of which 4,312 reported at least 1 of the 3 lower limb procedures itemized in the survey (surgery to repair a bone fracture, tendon or ligament repair/reconstruction surgery, or knee or hip joint replacement surgery). The initial dataset consisted of 3,740 participants reporting lower limb surgery as their most recent surgery.

Data Processing

The participant filtering process is illustrated in Figure 1. From the initial dataset, participants who had multiple unique answers to questions about the most recent procedure type, or recovery time, or who provided an implausible recovery time label were filtered out (for example, reported recovery time of “3–5 months” where the procedure date was <3 months from the survey date). The resulting data set consisted of 3,485 participants.

Next, with the participants’ permission, their activity data were linked for the time window from 182 days (26 weeks) before to 182 days after the self-reported surgery date. In order to ensure consistency in data quality across the participants, only participants who had any Fitbit device data available in the observation window ($n = 1,336$) were kept. Fitbit devices have been validated and reported as reliable for capturing steps, heart rate, and sleep data [18–21]; these 3 data modalities were used to get daily aggregates of various activity and behavioral statistics (see details in online suppl. Note 2). All 3 modalities are known to be of relevance to post-surgical recovery [20–23].

Furthermore, only participants for which age and gender data were available ($n = 1,324$) were kept. Most of the participants had steps ($n = 1,276$) and sleep ($n = 1,211$) data available, fewer participants had heart rate data ($n = 901$). At this point, no participant exclusion criterion due to missing data was applied; missing data in the statistical analysis part of this work was addressed by the choice of modeling approach, as we describe below.

Prediction of recovery time required further data filtering to ensure a higher data density on a participant level so that predictions could be made for each individual. This was achieved by restricting the data sets to participants who: (a) did not have continuous periods of missing steps data longer than 28 days, and (b) had at least 50% of observation window days with

Fig. 1. Illustration of the study participant filtering process. The flow chart demonstrates the number of participants across 3 lower limb surgery types: surgery to repair a bone fracture (“Bone frac.”), tendon or ligament repair/reconstruction surgery (“Tendon”), or knee or hip joint replacement surgery (“Knee/hip”).

steps data available. Data coverage in the full statistical analysis data sample ($n = 1,324$) and filtered sample ($n = 295$) is illustrated in online supplementary Note 3.

In order to ensure maximal data quality in the reporting of surgery dates, cases with a high likelihood of misreporting were systematically identified using a change point detection methodology. This approach was adapted from Zeileis et al. [24]; a function was fit based on the cohort-level model, and excluded instances where the function strongly fit but the self-reported and function-reported surgery date were more than 28 days apart. The process is described in more detail in online supplementary Note 4. After applying the rule, 217 out of 295 participants remained. Finally, only participants who reported completion of the recovery were kept. The final predictive modeling sample had 197 participants.

Statistical Modeling of Wearable PGHD

To estimate the impact of the medical procedures on steps count, heart rate, and sleep, the statistical analysis focused on 3 activity features: total number of steps, 95th percentile heart rate, and sleep efficiency (the proportion of minutes asleep of the total time in bed) during the main sleep. The baseline time period was defined as weeks from 26 (the earliest week in the observation window) to 13 weeks before the surgery; the upper limit of 13 weeks before the surgery was chosen in order to account for potential cases of a relatively long time from injury to surgery (an average of 13 weeks from injury to surgery was reported in patients with chronic Achilles tendon rupture, where more than half of the cases had tendon rupture after failure of conservative treatment [25]). In all visualizations in the manuscript, the “week 0” label denotes a 7-day period starting on the self-reported surgery day. Daily activity measurements were modeled with a linear mixed effect model (LMM), fitting a separate

model for each activity feature and surgery type sub-cohort. The outcome was defined as the participant- and day-specific activity measurement. The baseline period and each week in the range from 12 before to 26 after the surgery were represented by an indicator variable. The model was adjusted for fixed effects of age, age and relative week interaction, gender, month of the year, weekend day versus weekday, and participant-specific random effects (baseline activity and weekend day vs. weekday).

To further estimate trajectories of activity across time of recovery groups, the above model was extended by adding indicator variables for self-reported recovery time groups and for recovery group and relative week interaction. The choice of using day-level activity measurements and employing LMM with participant-specific intercept allowed us to avoid enforcing minimal data coverage or performing missing data imputation. Importantly, by including participants with missing data, we sought to increase the statistical power and avoid biasing our population-level estimates of activity.

Prediction of Self-Reported Recovery Times

To demonstrate the utility of wearable PGHD in predicting long-term trajectories of mobility recovery, the experiment was designed to evaluate performance of classifying self-reported recovery time labels. The machine learning task setup is described in more detail in online supplementary Note 5. In short, the model's performance was compared in 6 scenarios which differed in assumed availability of PGHD from wearable sensors: (1) no post-operative, no pre-operative, (2) no post-operative, 6 months (full) pre-operative, (3) 4 weeks post-operative, no pre-operative, (4) 4 weeks post-operative, 6 months pre-operative, (5) 6 months (full) post-operative, no pre-operative, (6) 6 months post-operative, 6 months pre-operative; in each case, demographics (age, gender) information was used.

Due to relatively small sample sizes for bone fracture and knee/hip replacement surgery predictive data sets ($n = 46$ and 26 , respectively; Table 1), the experiment was narrowed down to analyzing the tendon/ligament surgery group only ($n = 125$), and the task was cast as a binary classification of a participant into a faster ("0–2 months"; coded as negative case) and slower ("≥3 months"; coded as positive case) track of mobility recovery. The classification models were trained with the Extreme Gradient Boosting (XGBoost) [26] algorithm and evaluated in the 100-repeat holdout procedure. Alternative algorithms, including random forest with data imputation and feature preselection, and LASSO logistic regression were explored in preliminary stages of analysis and they did not yield performance results better than XGBoost (data not shown). The AUROC (area under the ROC curve) and AUPRC (area under the precision-recall curve) values, obtained on holdout test set across the 100 repetitions, are reported.

Results

Study Participants

Table 1 shows a summary of the participant demographics and self-reported recovery time for the statistical modeling sample ($n = 1,324$) and predictive modeling sample ($n = 197$). Data are summarized for whole sample cohorts ("all") and for respective strata by surgery type. Participants included in the statistical analysis sample were predominantly female (84%), white or Caucasian (85%), college-educated (62%), and young to middle-aged (mean 36.2 years, SD 12.9), closely in line with the distribution skewness we observed for the whole user base of the Achievement platform (77% female, 88% white or Caucasian, mean age 33 years [27]). The mean age varied across the surgery type sub-cohorts, from 32.9 years in bone fracture surgery to 47.7 years in knee/hip joint replacement surgery; for comparison, the

Table 1. Participant demographics and self-reported recovery time for statistical modeling sample and predictive modeling sample

	Statistical modeling set				Predictive modeling set			
	all (n = 1,324)	bone frac. (n = 355)	tendon (n = 773)	knee/hip (n = 196)	all (n = 197)	bone frac. (n = 46)	tendon (n = 125)	knee/hip (n = 26)
Gender								
Female	1,117 (84)	307 (86)	648 (84)	162 (83)	169 (86)	41 (89)	106 (85)	22 (85)
Male	203 (15)	47 (13)	122 (16)	34 (17)	28 (14)	5 (11)	19 (15)	4 (15)
Other	4 (<1)	1 (<1)	3 (<1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Race								
White or Caucasian	1,119 (85)	302 (85)	649 (84)	168 (86)	168 (85)	37 (80)	110 (88)	21 (81)
Black or African American	52 (4)	15 (4)	30 (4)	7 (4)	5 (3)	1 (2)	4 (3)	0 (0)
Hispanic or Latino	61 (5)	14 (4)	39 (5)	8 (4)	8 (4)	2 (4)	4 (3)	2 (8)
Other	45 (3)	10 (3)	26 (3)	9 (5)	6 (3)	3 (7)	2 (2)	1 (4)
Unavailable	47 (4)	14 (4)	29 (4)	4 (2)	10 (5)	3 (7)	5 (4)	2 (8)
Age, years								
mean (SD)	36.2 (12.9)	32.9 (11.1)	34.9 (11.7)	47.7 (14.4)	37.3 (11.6)	34.2 (8.9)	36.1 (11.1)	48.4 (12.4)
median (min.–max.)	34 (18–77)	31 (18–70)	33 (18–70)	51 (18–77)	35 (18–71)	32 (19–53)	34 (18–64)	49 (24–71)
Education								
Doctorate, MD	29 (2)	5 (1)	19 (2)	5 (3)	4 (2)	0 (0)	4 (3)	0 (0)
Graduate degree	236 (18)	58 (16)	146 (19)	32 (16)	33 (17)	8 (17)	22 (18)	3 (12)
College degree (AS or BS)	515 (39)	133 (37)	317 (41)	65 (33)	89 (45)	24 (52)	58 (46)	7 (27)
Some college	311 (23)	87 (25)	183 (24)	41 (21)	42 (21)	7 (15)	26 (21)	9 (35)
Trade or vocational training	77 (6)	23 (6)	33 (4)	21 (11)	8 (4)	4 (9)	3 (2)	1 (4)
High school diploma/GED	107 (8)	31 (9)	48 (6)	28 (14)	13 (7)	1 (2)	8 (6)	4 (15)
No high school diploma	9 (1)	5 (1)	4 (1)	0 (0)	1 (1)	0 (0)	1 (1)	0 (0)
Unavailable	40 (3)	13 (4)	23 (3)	4 (2)	7 (4)	2 (4)	3 (2)	2 (8)
Recovery time								
<1 month	260 (20)	61 (17)	164 (21)	35 (18)	37 (19)	6 (13)	26 (21)	5 (19)
1–2 months	257 (19)	80 (23)	137 (18)	40 (20)	51 (26)	16 (35)	29 (23)	6 (23)
3–5 months	292 (22)	96 (27)	151 (20)	45 (23)	52 (26)	15 (33)	28 (22)	9 (35)
6–12 months	209 (16)	38 (11)	142 (18)	29 (15)	53 (27)	7 (15)	41 (33)	5 (19)
1 year or longer	28 (2)	8 (2)	17 (2)	3 (2)	4 (2)	2 (4)	1 (1)	1 (4)
I never fully recovered	33 (2)	8 (2)	20 (3)	5 (3)	0 (0)	0 (0)	0 (0)	0 (0)
I'm still recovering	245 (19)	64 (18)	142 (18)	39 (20)	0 (0)	0 (0)	0 (0)	0 (0)

Values are n (%) or as stated. Data are summarized for the whole sample cohort (“all”) and by strata by lower limb surgery types: surgery to repair a bone fracture (“bone frac.”), tendon or ligament repair/reconstruction surgery (“tendon”), or knee or hip joint replacement surgery (“knee/hip”). Age at the time of procedure was estimated based on information from a patient ID-linked survey at a different time point than the medical event survey.

Fig. 2. Change in activity features in subsequent weeks from week 12 before to week 26 after surgery compared to the average value in the baseline period (from week 26 to week 13 before the surgery). Horizontal plot panels correspond to 3 daily features: total number of steps, 95th percentile heart rate, and sleep efficiency during the main sleep. Vertical plot panels correspond to 3 lower limb surgery types: bone fracture, tendon or ligament repair, and knee or hip replacement. The colors and error bars correspond to p values and 95% CIs of the model coefficient estimate for the effect of a relative week compared to baseline, respectively. The “week 0” label (x -axis) denotes a 7-day period starting on a self-reported surgery day.

average age for total hip arthroplasty and total knee arthroplasty patients was reported as 65 years [28] and 67 years [29], respectively. The most common self-reported time of recovery fell between 1 and 5 months for bone fracture and knee/hip replacement surgery, and from 1 to 12 months for tendon/ligament repair surgery.

Demographic data summaries for participants included in the predictive modeling sample closely follow the distribution of the analysis data sample. The percentages of self-reported time groups changed mostly due to the fact that this sample excluded individuals who had not reported completion of the recovery.

Different Types of Lower Limb Surgery Have Distinct Recovery Profiles

Figure 2 summarizes the resulting cohort-level model fit, showing for each surgery type changes relative to baseline for representative features from step, heart rate, and sleep data (daily step count, 95th percentile heart rate, and sleep efficiency, respectively) from 12 weeks

before to 26 weeks after surgery. The trajectories are shown for a “typical” cohort individual (female aged 40 years, with an average baseline activity level among otherwise similar ones). The model-estimated values of activity are also summarized in Table 3 of online supplementary Note 8.

At baseline, the estimated average daily measurement values varied very slightly across 3 surgery type sub-cohorts: 8,900, 8,905, and 8,815 for daily sum of steps, 103.9, 102.9, and 103.8 for the 95th percentile of heart rate (bpm), and 60.4, 57.6, and 57.7 for sleep efficiency – for 3 surgery type sub-cohorts (bone fracture repair, tendon or ligament repair/reconstruction, knee or hip joint replacement), respectively. As expected, all surgeries resulted in significant changes in activity, typically reducing daily step counts by 3,000–4,000 steps in the week following surgery, returning to near baseline levels over 8–12 weeks. All surgeries also resulted in reductions in sub-maximal heart rate, which generally returned to baseline levels within 4–8 weeks, and reductions in sleep efficiency which remained throughout the 12 weeks post-surgery. Activity and heart rate data were generally observed to be less variable than sleep data, possibly due to poorer nighttime data coverage and a relatively low accuracy of current models for estimating sleep metrics from consumer wearables [30].

In addition to these general similarities, patterns were also observed that distinguished the 3 surgery groups and which correspond to distinct best practices [31]. For example, a significant pre-surgical reduction in steps sum and heart rate levels was seen in the 2–3 weeks prior to bone fracture surgeries, whereas for tendon and ligament surgeries this reduction was already apparent 8–10 weeks prior to surgery, and for knee or hip replacement the reduction was stronger (more than 1,000 steps) and observable 3–4 weeks prior to surgery. Distinct post-surgical recovery trajectories were also observed; for example, the effect of bed rest in bone fracture and joint replacement surgeries was visible immediately post-surgery, while tendon/ligament repair surgery patients recovered to baseline activity more slowly than the 2 other groups, which agrees with a slightly higher proportion of self-reported “6–12 months” time of recovery for this group (Table 1). Confirming the validity of the model, we observed that the known effect of age on recovery trajectories [32] was captured (see online suppl. Note 6).

Recovery Trajectories Associated with Self-Reported Recovery Times

To verify that PGHD from wearable sensors can capture differences in activity across recovery groups, an extended statistical model (see Materials and Methods) was used. Figure 3 shows the estimated average trajectories of daily number of steps across 3 self-reported recovery time groups across the 3 lower limb surgeries. Values are shown for a “typical” cohort individual (female aged 40 years, with an average baseline activity level among similar ones). The upper panel shows absolute activity (steps) values, the lower plots show change with respect to the model-estimated baseline. In the 1- to 4-week post-operative period, absolute values of activity distinguish the recovery groups, especially for bone fracture and tendon/ligament repair groups. Furthermore, we observed a complementary signal in the trajectory of relative change compared to the baseline, particularly for the tendon/ligament repair sub-cohort, where differences between the recovery time groups were visible both before and after surgery. For the knee/hip replacement surgery sub-cohort (the smallest sub-cohort), a relatively higher variability of fitted values was observed; the resulting patterns may possibly represent the effects of a mixture of different knee and hip replacement procedures which cannot be disentangled based on the survey conducted.

Model-estimated values of activity are also summarized in Table 4 of online supplementary Note 8. For completion, activity trajectories estimated across 2 and 4 self-reported recovery time groups are included in online supplementary Note 7.

Fig. 3. Estimated trajectories of daily number of steps across 3 self-reported recovery time groups in subsequent weeks from 12 weeks before to 26 weeks after surgery. The upper plots show absolute values of activity, the bottom plots show change with respect to the model-estimated baseline. Vertical plot panels correspond to 3 lower limb surgery types: bone fracture, tendon or ligament repair, and knee or hip replacement. The color of a point/line corresponds to the self-reported recovery time group. The “week 0” label (x-axis) denotes a 7-day period starting on a self-reported surgery day.

Wearable PGHD Can Be Used to Predict Recovery Trajectories

Table 2 summarizes the results of the experiment to discriminate participants who self-reported faster (“0–2 months”) versus slower (“≥3 months”) functional recovery trajectories across 6 scenarios in which different data availability was assumed: demographic data only; individual baseline data only; 1 month post-surgery with and without an individual baseline, and 6 months post-surgery with and without individual baseline. The analysis focused on the tendon/ligament surgery group ($n = 125$) as the bone fracture ($n = 46$) and knee/hip replacement ($n = 26$) groups were too small to robustly train and test a predictive model.

Demographic variables (age, gender) themselves were not discriminative between faster and slower recovery track patients, attaining a median AUROC of 0.489 (mean 0.473, SD 0.108; Table 2). This aligns with high demographic similarity between the recovery groups, for example in the tendon/ligament surgery group the sample mean of age was very similar in the faster and slower recovery tracks: 36.6 (SD 10.9) and 35.9 (SD 11.3), respectively.

In the 4 weeks post-operative scenarios, the scenario with pre-operative activity data available attained a higher AUROC (median 0.734, mean 0.724, SD 0.095) than in the scenario without pre-operative data (AUROC median 0.701, mean 0.705, SD 0.089). Compared to the 4 weeks post-operative scenarios, the 6 months post-operative scenarios yielded results slightly worse when pre-operative activity data were available (median 0.721, mean 0.71, SD

Table 2. Performance of predictive models in the task of discriminating participants between a faster (“0–2 months”) and slower (“≥3 months”) track of mobility recovery

Experiment scenario (post-/pre-operative wearable PGHD availability)	Tendon or ligament repair/reconstruction surgery			
	AUROC		AUPRC	
	mean (SD)	median (min.–max.)	mean (SD)	median (min.–max.)
(1) No post-op, no pre-op	0.473 (0.108)	0.489 (0.114–0.669)	0.569 (0.059)	0.563 (0.428–0.727)
(2) No post-op, 6 months pre-op	0.497 (0.096)	0.498 (0.266–0.734)	0.596 (0.068)	0.596 (0.453–0.741)
(3) 4 weeks post-op, no pre-op	0.705 (0.089)	0.701 (0.510–0.929)	0.784 (0.076)	0.799 (0.598–0.947)
(4) 4 weeks post-op, 6 months pre-op	0.724 (0.095)	0.734 (0.442–0.942)	0.795 (0.077)	0.800 (0.576–0.960)
(5) 6 months post-op, no pre-op	0.712 (0.084)	0.716 (0.542–0.929)	0.798 (0.067)	0.806 (0.628–0.937)
(6) 6 months post-op, 6 months pre-op	0.710 (0.096)	0.721 (0.435–0.942)	0.786 (0.080)	0.791 (0.542–0.960)

Results are shown across 6 experiment scenarios in which different data availability was assumed (in each case, age and gender demographic information was used), and across surgery types considered.

0.096) and slightly better without pre-operative activity data (median 0.716, mean 0.712, SD 0.084).

The features relative to baseline and those calculated from weeks immediately around the surgery were observed to be particularly important in driving the predictive power (Fig. 4). Taken together, these results suggest that 4 weeks post-operative activity data already carry substantial information predictive of a patient’s long-term recovery, and that the discriminative power of a model using 4 weeks post-operative activity data may be improved when pre-operative data are available.

Discussion

Using PGHD from 1,324 individuals, we have demonstrated that self-reported functional recovery trajectories can be accurately modeled based on data from consumer wearable devices describing everyday function from up to 6 months prior to surgery to 6 months post-surgery. We have shown that typical recovery trajectories from different types of surgery can be distinguished, for example the 2–4 weeks of immobilization following bone fracture surgery versus immediate remobilization of patients following tendon surgery. Support for the validity of our model was demonstrated by capturing the known impact of age on functional recovery [33]. We also saw that retrospective, self-reported recovery trajectories were clearly differentiated in terms of recovery trajectories, for example by the “depth” of functional limitation immediately post-surgery. We also observed that pre-surgery, long-term baseline function, and functional decreases immediately prior to surgery also differentiated the groups.

Prediction of long-term outcomes is highly important because early intervention, for example increasing exercise, is hypothesized to improve recovery outcomes [15, 17]. Indeed, our observations provide evidence that higher levels of activity prior to surgery correspond with better functional recovery post-surgery. The accurate prediction of outcomes is often not possible, as pre-surgery risk factors and demographics, without any functional baseline data, do not provide sufficient predictive power, for example 2-year risk of knee replacement revision [34]. We demonstrated that passively collected, consumer-grade wearable data can provide baseline data to accurately predict long-term recovery trajectories. Furthermore, we show that such predictions can be made only 1 month after surgery, early enough to inform

Fig. 4. Shapley Additive Explanations (SHAP) obtained from the hand-tuned XGBoost model fitted to data of all participants in the tendon/ligament surgery group, assuming 4 weeks post-operative and 6 months pre-operative availability of PGHD from wearable sensors. The SHAP values are shown for the top 20 most impactful predictors. “BS” denotes predictors defined as a ratio of value derived from a particular week(s) period to value derived from the baseline period.

alterations to physiotherapy regimes, for example specific targeting of “pre-habilitation” [35]. Recent work has also shown that this approach may have value in other therapeutic interventions, for example in oncology [36].

Limitations

While we believe the results presented here to be robust, we are aware of limitations of the analysis conducted. First, the data used are primarily based on self-reported dates and recovery times. We have made efforts to be conservative in our selection of data to ensure maximal quality, in part enabled by the large scale of data collection. Future work will focus on integrating data across a wider range of consumer devices. Second, the lack of more specific information about causes for surgical intervention prevents further clustering or data analysis. This might be addressed in a prospective, well-designed data collection.

Furthermore, we are also blind to many hidden or otherwise unknown causal factors; for example, the speed of recovery may be linked to comorbidities, or to geography and socio-economic status, which may underlie access to proactive physiotherapy and a need and timeline of return to work. Previous work has shown that returning to work can have a signif-

ificant impact on observed real-world data without having a strong link to functional recovery [10]. Again, a prospective study would address these limitations.

Finally, the generalizability of the approach is still to be demonstrated. Our predictive modeling is only based on a convenience sample of tendon and ligament surgeries, and individuals who have relatively dense wearable data. Future work will attempt to recruit bigger and more diverse cohorts from other surgery types to test the generalizability of the approach and to address other demographic biases in the enrolled population, which is tendentially young and female. This prospective study will also allow calibration of the models presented here [37].

Outlook

Prospective evaluation of the results and methods presented here, in a broader group of participants and integrating a broader range of devices, will further demonstrate how consumer devices can become an even more powerful tool in digital health, enabling objective assessment of recovery and personalized recommendations based on an individualized baseline. Ultimately, we hope that individual baseline data can become an informative tool in triggering early engagement with healthcare.

Acknowledgements

The authors would like to thank the participants, members of the Evidation Health Achievement Platform, for their participation in this research. The authors would also like to thank Celeste Scotti for his advice and Raghu Kainkaryam for technical input.

Statement of Ethics

This study received expedited review and IRB approval from Solutions IRB (protocol ID 2019/04/14). Waiver of informed consent was granted by the IRB due to the retrospective design of the study.

Conflict of Interest Statement

M.K. has previously received payment for internships carried out at Novartis, and received payment from Evidation Health for the internship during which this work was carried out. N.M. is currently an employee of and holds stock options in Evidation Health. J.G. was previously an employee of Novartis. L.F. is a cofounder and employee of Evidation Health. E.R. is currently an employee of and holds stock options in Evidation Health. I.C. was previously an employee of Novartis and is currently an employee of and holds stock options in Evidation Health. He has received payment for lecturing on Digital Health at ETH Zurich and FHNW Muttensz. He is an Editorial Board Member at Karger Digital Biomarkers and a founding member of the Digital Medicine Society.

Funding Sources

This work received no direct funding and was otherwise entirely self-funded by Evidation Health Inc.

Author Contributions

M.K. and I.C. contributed to the conception and design of the work, data analysis and interpretation. N.M. contributed to the design of the work, data analysis and interpretation. J.G. contributed to the design of the work and data interpretation. L.F. and E.R. contributed to the conception and design of the work and data interpretation. All authors contributed to drafting and revision of the manuscript and approved the final version.

References

- 1 Cohen AB, Dorsey ER, Mathews SC, Bates DW, Safavi K. A digital health industry cohort across the health continuum. *NPJ Digit Med*. 2020 May;3(1):68.
- 2 Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *NPJ Digit Med*. 2020 Apr;3(1):55.
- 3 Haberkamp M, Moseley J, Athanasiou D, de Andres-Trelles F, Elferink A, Rosa MM, et al. European regulators' views on a wearable-derived performance measurement of ambulation for Duchenne muscular dystrophy regulatory trials. *Neuromuscul Disord*. 2019 Jul;29(7):514–6.
- 4 Izmailova ES, Wagner JA, Ammour N, Amondikar N, Bell-Vlasov A, Berman S, et al. Remote digital monitoring for medical product development. *Clin Transl Sci*. 2020. doi: 10.1111/cts.12851.
- 5 Coravos A, Goldsack JC, Karlin DR, Nebeker C, Perakslis E, Zimmerman N, et al. Digital Medicine: A Primer on Measurement. *Digit Biomark*. 2019 May;3(2):31–71.
- 6 Website [Internet]. [cited 2020 Sep 4]. Available from: https://healthpolicy.duke.edu/sites/default/files/2019-11/rwd_reliability.pdf.
- 7 Labrique AB, Wadhwani C, Williams KA, Lamptey P, Hesp C, Luk R, et al. Best practices in scaling digital health in low and middle income countries. *Global Health*. 2018 Nov;14(1):103.
- 8 OECD iLibrary. Health at a Glance 2019: OECD Indicators. OECD; 2019. <https://doi.org/10.1787/4dd50c09-en>.
- 9 Appelboom G, Taylor BE, Bruce E, Bassile CC, Malakidis C, Yang A, et al. Mobile Phone-Connected Wearable Motion Sensors to Assess Postoperative Mobilization. *JMIR Mhealth Uhealth*. 2015 Jul;3(3):e78.
- 10 Mueller A, Hoefling H, Nuritdinow T, Holway N, Schieker M, Daumer M, et al. Continuous Monitoring of Patient Mobility for 18 Months Using Inertial Sensors following Traumatic Knee Injury: A Case Study. *Digit Biomark*. 2018 Aug;2(2):79–89.
- 11 Piau A, Wild K, Mattek N, Kaye J. Current State of Digital Biomarker Technologies for Real-Life, Home-Based Monitoring of Cognitive Function for Mild Cognitive Impairment to Mild Alzheimer Disease and Implications for Clinical Care: systematic Review. *J Med Internet Res*. 2019 Aug;21(8):e12785.
- 12 Ramirez E, Marinsek N, Bradshaw B, Kanard R, Foschini L. Continuous Digital Assessment for Weight Loss Surgery Patients. *Digit Biomark*. 2020 Mar;4(1):13–20.
- 13 Bindawas SM, Graham JE, Karmarkar AM, Chen NW, Granger CV, Niewczyk P, et al. Trajectories in functional recovery for patients receiving inpatient rehabilitation for unilateral hip or knee replacement. *Arch Gerontol Geriatr*. 2014 May-Jun;58(3):344–9.
- 14 Dyer SM, Crotty M, Fairhall N, Magaziner J, Beaupre LA, Cameron ID, et al.; Fragility Fracture Network (FFN) Rehabilitation Research Special Interest Group. A critical review of the long-term disability outcomes following hip fracture. *BMC Geriatr*. 2016 Sep;16(1):158.
- 15 Arnold JB, Walters JL, Ferrar KE. Does Physical Activity Increase After Total Hip or Knee Arthroplasty for Osteoarthritis? A Systematic Review. *J Orthop Sports Phys Ther*. 2016 Jun;46(6):431–42.
- 16 Stevens-Lapsley JE, Schenkman ML, Dayton MR. Comparison of self-reported knee injury and osteoarthritis outcome score to performance measures in patients after total knee arthroplasty. *PM R*. 2011 Jun;3(6):541–9.
- 17 Magaziner J, Chiles N, Orwig D. Recovery after Hip Fracture: Interventions and Their Timing to Address Deficits and Desired Outcomes—Evidence from the Baltimore Hip Studies. *Nestle Nutr Inst Workshop Ser*. 2015;83:71–81.
- 18 Achievement [Internet]. [cited 2020 Jul 6]. Available from: www.myachievement.com.
- 19 Deering S, Grade MM, Uppal JK, Foschini L, Juusola JL, Amdur AM, et al. Accelerating Research With Technology: Rapid Recruitment for a Large-Scale Web-Based Sleep Study. *JMIR Res Protoc*. 2019 Jan;8(1):e10974.
- 20 Su X, Wang DX. Improve postoperative sleep: what can we do? *Curr Opin Anaesthesiol*. 2018 Feb;31(1):83–8.
- 21 Abeler K, Friberg O, Engström M, Sand T, Bergvik S. Sleep Characteristics in Adults With and Without Chronic Musculoskeletal Pain: The Role of Mental Distress and Pain Catastrophizing. *Clin J Pain*. 2020 Sep;36(9):707–15.
- 22 Altman MT, Knauert MP, Pisani MA. Sleep Disturbance after Hospitalization and Critical Illness: A Systematic Review. *Ann Am Thorac Soc*. 2017 Sep;14(9):1457–68.

- 23 O'Driscoll R, Turicchi J, Beaulieu K, Scott S, Matu J, Deighton K, et al. How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. *Br J Sports Med*. 2020 Mar;54(6):332–40.
- 24 Zeileis A, Leisch F, Hornik K, Kleiber C. Strucchange: an R package for testing for structural change in linear regression models. *J Stat Softw*. 2002. doi: 10.18637/jss.v007.i02.
- 25 Lin Y, Yang L, Yin L, Duan X. Surgical Strategy for the Chronic Achilles Tendon Rupture. *BioMed Res Int*. 2016; 2016:1416971.
- 26 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM; 2016; pp. 785–94.
- 27 Kumar S, Tran JL, Ramirez E, Lee WN, Foschini L, Juusola JL. Design, recruitment, and baseline characteristics of a virtual 1-year mental health study on behavioral data and health outcomes: observational Study. *JMIR Ment Health*. 2020 Jul;7(7):e17075.
- 28 Passias PG, Bono OJ, Bono JV. Total Knee Arthroplasty in Patients of Advanced Age: A Look at Outcomes and Complications. *J Knee Surg*. 2020 Jan;33(1):1–7.
- 29 Bang H, Chiu YL, Memtsoudis SG, Mandl LA, Della Valle AG, Mushlin AI, et al. Total hip and total knee arthroplasties: trends and disparities revisited. *Am J Orthop (Belle Mead NJ)*. 2010 Sep;39(9):E95–102.
- 30 Liang Z, Chapa-Martell MA. Accuracy of Fitbit Wristbands in Measuring Sleep Stage Transitions and the Effect of User-Specific Factors. *JMIR Mhealth Uhealth*. 2019 Jun;7(6):e13384.
- 31 Cioppa-Mosca J, Cahill JB, Tucker CY. Postsurgical rehabilitation guidelines for the orthopedic clinician. Amsterdam: Mosby Elsevier; 2006.
- 32 Dolata J, Pietrzak K, Manikowski W, Kaczmarczyk J, Gajewska E, Kaczmarek W. Influence of age on the outcome of rehabilitation after total hip replacement. *Pol Orthop Traumatol*. 2013 May;78:109–13.
- 33 Kuroda Y, Saito M, Çınar EN, Norrish A, Khanduja V. Patient-related risk factors associated with less favourable outcomes following hip arthroscopy. *Bone Joint J*. 2020 Jul;102-B(7):822–31.
- 34 El-Galaly A, Grazal C, Kappel A, Nielsen PT, Jensen SL, Forsberg JA. Can Machine-learning Algorithms Predict Early Revision TKA in the Danish Knee Arthroplasty Registry? *Clin Orthop Relat Res*. 2020 Jun;478(9):2088–101.
- 35 Santa Mina D, Clarke H, Ritvo P, Leung YW, Matthew AG, Katz J, et al. Effect of total-body prehabilitation on postoperative outcomes: a systematic review and meta-analysis. *Physiotherapy*. 2014 Sep;100(3):196–207.
- 36 Jonker LT, Hendriks S, Lahr MM, van Munster BC, de Bock GH, van Leeuwen BL. Postoperative recovery of accelerometer-based physical activity in older cancer patients. *Eur J Surg Oncol*. 2020, Online ahead of print.
- 37 Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019 Dec;17(1):230.