How accurate is accurate enough? - An Evaluation of Commercial Fitness Trackers for Individual Health Management

Completed Research

Anne-Katrin Witte
Technische Universität Berlin
a.witte@tu-berlin.de

Kim Janine Blankenhagel
Technische Universität Berlin
k.blankenhagel@tu-berlin.de

Jakob J. Korbel
Technische Universität Berlin
jakob.j.korbel@tu-berlin.de

Rüdiger Zarnekow
Technische Universität Berlin
ruediger.zarnekow@tu-berlin.de

Abstract

Despite their extensive functionality in the area of physiological data collection and high user adoption, commercial fitness trackers are rarely used in a medical context to improve costly patient care. Therefore, this paper investigates under which circumstances fitness trackers are applicable to give reasonable behavior change advice for patient's individual health management. To do so we defined daily routine application scenarios and evaluated the accuracy of 15 commercial fitness trackers in a field study for step count, continuous heart rate and sleep duration measurement. In this context we conclude that the sensor technology of the devices is partly accurate enough, depending on the test scenario and device. To effectively enhance individual health management, we suggest a definition of context-aware application scenarios and standardized procedures for accuracy tests. However, the challenges of the current devices regarding susceptibility to outliers and low software and hardware transparency need to be considered.

Keywords

Fitness Tracker, Individual Health Management, Field Study, Mobile Health.

Introduction

Currently, the healthcare system is facing ever-increasing challenges including patients with chronic conditions, risk of medical errors, cost of hospitalization and the ageing population. For this reason, there is a need to design a sustainable system for healthcare delivery which provides not only good acute care in emergency units but also supports outpatients at home, especially the ones with long term conditions (Baig et al. 2017). The global health expenditure rises constantly and is expected to reach a new maximum of over 12 trillion US $ in 2022 (World Economic Forum 2014), thereby most of the spending goes to non-communicable diseases which mainly describe chronic disorders (World Health Organization 2018). Even though these conditions are largely preventable and have high potential for interventions, patients with a chronic or long term condition spend 99% of their time managing their disease individually by themselves and only a few hours with a physician (Eaton et al. 2015). The total number of people with a long term condition is likely to rise continuously due to the demographic transition, which says that the population aged over 65 years will reach a number near 2 billion people by 2050 (United Nations 2015). To facilitate a better use of resources, patients need to be "activated". This means that they have skills, knowledge and confidence to effectively manage their individual health condition. Therefore "activated" patients have a higher chance to adopt healthy behaviors and for that reason experience better care and health outcomes (Hibbard and Greene 2013).

Wearable systems for patient monitoring are an effective tool to detect, manage or even prevent chronic conditions. Their applicability in a huge variety of contexts and their expansion for health data collection

holds the potential to engage the user and offer self-management for chronic diseases (Baig et al. 2017). Currently there are two kinds of healthcare wearable devices on the market, one for fitness purposes and one for medical application. On the one hand there are fitness trackers, which are generally defined as wearable technology that is worn on the wrist (Swan 2012) and help young and healthy users to track their daily lifestyle data (Gao et al. 2015). Fitness trackers are constantly evolving and can nowadays comprise several sensors that measure different values e.g. accelerometer and gyroscopes for physical activity or photoplethysmography for optical heart rate measurement (Henriksen et al. 2018). On the other hand medical devices are aimed at elderly and unhealthy patients and designed to manage a certain disease e.g. diabetes (Gao et al. 2015).

Wearable healthcare devices can be a great solution to reduce the rising costs in the medical area (Behkami and Daim 2012), with possible savings up to 305 billion US $ in the United States through Remote Patient Monitoring and Telehealth (Roman 2015). But to achieve these potential cost savings, it is necessary for a majority of individuals to adopt wearable healthcare devices (Roman 2015). The main challenges for the adoption of wearable technology in healthcare are accuracy, security, accessibility, cost, compatibility, acceptability, interpretation and technological/form factor issues (Dunn, Runge, and Snyder 2018). While medical (research-grade) devices promise high accuracy, their outer appearance is often bulky, it is hard for the patient to set up the device unaccompanied, they are expensive and are not compatible with established standards. Especially in the area of self-management it is important to reach a high wear time compliance, which can be enhanced using unobtrusive, versatile and stylish devices (Ferguson et al. 2015). Commercially available fitness devices fulfil these requirements. They have a high potential for quick user adoption because they are „less obtrusive than standard research-grade devices" and offer additional feedback for the user (Lee et al. 2014). This comprehensive feedback functionality of fitness trackers is one of their strengths, which may be sufficient for the user to enable behavior modification (Ferguson et al. 2015). That fitness trackers are able to support behavior change, education and increase patient's self-efficacy through immediate feedback and visible reminders was shown in a study where adults with chronic medical conditions were given fitness tracker to increase their physical activity (Gualtieri et al. 2016). In general, a majority of people have a positive association with fitness trackers in a medical context. Surveys indicated that 65 % of respondents expressed excitement about their doctors providing fitness trackers (PwC 2016) and 48,2% non-user US adults are willing to use a free fitness tracker provided by their physician (Technology Advice 2014).

While some studies argue that commercial devices are not appropriate for medical purposes (e.g. commercial pedometers for telemonitoring subjects with motion disability (Giansanti et al. 2008)), there are several related works that evaluated fitness trackers positively in predefined, mostly laboratory scenarios. These studies concluded e.g. a valid measurement of sleep and heart rate in healthy adolescents with typical sleep patterns using Fitbit Charge HR (de Zambotti et al. 2016) or comparable performance characteristics to a standard actigraph in the estimation of sleep duration during a single night by Jawbone UP3 (Cook et al. 2018). Hence in specific contexts these devices are assessed as "reasonably valid for measuring the respective variables" (Ferguson et al. 2015) and their characteristics of wide availability, easy usage, low-cost and techno-attractiveness are emphasized which may make these fitness devices an attractive alternative e.g. to standard actigraphy in monitoring daily sleep-wake rhythms (de Zambotti et al. 2016; de Zambotti, Claudatos, et al. 2015). We aim to contribute to this field and evaluate the strengths and limitations of commercial fitness trackers to enhance patient self-management. For this purpose, we assess *under which conditions commercial fitness trackers are sufficiently accurate to give reasonable behavior change advice for individual health management*. To do so we conducted a field study, which represents the usual activities our participant group performs during a regular office day and compared the results with those of previous studies. Our approach is structured as follows: chapter 2 introduces related literature in the field of commercial fitness tracker evaluation. Based on this, chapter 3 describes the research methodology and how we structured the field study. Following, the results are presented in chapter 4 and discussed in chapter 5. We finish with our conclusion and future outlook in chapter 6.

Related Work

In the area of commercial fitness technology there are already several publications evaluating fitness trackers (see Table 1). This is done with different foci and using diverse testing approaches. Many contributions started to test commercial fitness trackers regarding various criteria e.g. accuracy of heart

Author	Test Models	Year	#	Sleep (SL) - Heart rate (HR) - Steps (ST) - Energy expenditure (EE)	Lab	Scenario
(de Zambotti, Baker, et al. 2015)	Jawbone UP	2015	65	**SL**: mean differences ± SD: total sleep time (TST) −10.0 ± 20.5 min	**yes**	medical (overnight sleep laboratory recording)
(Jo et al. 2016)	Basis Peak (BP), Fitbit Charge HR (FBH)	2015	24	**HR**: mean bias -2.5 bpm (BP) and -8.8 bpm (FBH)	**yes**	**fitness** (e.g. 60W/120W cycling, walking, running, resisted arm raises, etc.)
(Ferguson et al. 2015)	7 fitness trackers e.g. Fitbit One/ Zip, Jawbone UP, Withings Pulse	2015	21	**ST**: strong correlation (r > 0.8); **SP**: strong correlation (r > 0.8); EE: moderate-strong correlation (r = 0.74-0.81)	no	**general** (wear all devices simultaneously for 48 hours)
(El-Amrawy and Nounou 2015)	17 fitness tracker e.g. Apple Watch, Samsung Gear Fit/Gear 1/Gear 2/Gear S, FitBit Flex, etc.	2015	4	**HR**: accuracy: 99,9% - 92,8%; **ST**: accuracy: 99,1% - 79,8%	**(yes)**	**general** (walking 200, 500, and 1,000 steps, 40 times)
(Wallen et al. 2016)	Apple Watch, Fitbit Charge HR, Samsung Gear S, Mio Alpha	2016	22	**HR**: 0.67–0.95 correlation with reference value (RV) EE: 0.16–0.86 correlation with RV	**yes**	**fitness** (1-hr protocols involving rest, walking, running on treadmill, etc.)
(de Zambotti et al. 2016)	Fitbit Charge HR	2016	32	**SL**: accuracy (91%); **HR**: average FBH (59.3 ± 7.5 bpm), average ECG (60.2 ± 7.6 bpm, p < 0.001)	**yes**	medical (1 night in sleep laboratory)
(Kaewkannate and Kim 2016)	Withings Pulse (WP), Misfit Shine, Jawbone Up24, Fitbit Flex	2016	7	**ST**: WP accuracy of 99.90 % and repeatability of 0.86	**(yes)**	**fitness** (48m indoor walk, 10 times; 1min 8km/h treadmill running, 5 times; walking up and down 4 flights of stairs)
(de Zambotti, Claudatos, et al. 2015)	Jawbone UP	2016	28	**SL**: detecting sleep (0.97), detecting wake (0.37), TST (26.6±35.3min)	**yes**	medical (1-2 nights in the sleep laboratory)
(Kroll et al. 2016)	Fitbit Charge HR	2016	50	**HR**: average bias −1.14 bpm	**(yes)**	medical (intensive care unit patients for 24 hours)
(Montoye et al. 2016)	Fitbit One/ Zip/ Flex (FBF), Jawbone UP24	2016	30	**ST**: (household): except FBF underestimation by 35%–64%, (exercise): within 4% of RV; EE: severely underestimation	**yes**	**general** (structured protocol: 3 sedentary, 4 household, 4 exercise activities)
(Shah et al., 2017)	Jawbone UP4, Fitbit Charge HR	2017	3	**HR**: significantly different, even at daily level; **ST**: 0.82–0.93 correlation coefficient; EE: 0.71–0.85 correlation coefficient	**(yes)**	**general** (crewmembers wear devices in tandem for part of 12-month HI-SEAS mission IV)
(Parak et al. 2017)	PulseOn	2017	24	**HR**: running: 1.9% mean absolute percentage error (MAPE); EE: MAPE above aerobic threshold = 6.7%, lighter intensity = 16.5%	**(yes)**	fitness (submaximal self-paced outdoor running test and maximal voluntary exercise test)
(Sirard et al. 2017)	Movband (MB), Sqord (SQ) and Zamzee (ZZ)	2017	14/ 16	**ST**: significantly associated (r=.79); EE: Spearman correlation coefficients (AG, MB, SQ, and ZZ) were .87, .61, .87, and .60	yes/ no	**general** (phase 2: 9 structured activities (e.g. quiet sitting), Phase 3: wearing all 4 devices 4 consecutive days)
(Shcherbina et al. 2017)	6 fitness trackers e.g. Apple Watch, Fitbit Surge, Microsoft Band, etc.	2017	60	**HR**: lowest error cycle ergometer 1.8%, highest error walking 5.5%; EE: median error rates across tasks: 27.4% to 92.6%	**yes**	**fitness** (standardized exercise protocol 44min: sitting, walking, running, cycling)
(Reddy et al. 2018)	Fitbit Charge 2 (FB), Garmin vívosmart HR+ (G)	2018	20	**HR**: G relative error (RE): −3.3% (SD 16.7), FB RE: −4.7% (SD 19.6); EE: FB: −19.3% [SD 28.9], G: −1.6% [SD 30.6], P<.001	**yes**	**general/fitness** (1: maximal oxygen uptake test, free-weight resistance circuit; 2. laboratory visit: interval training session, daily activities)
(Ummels et al. 2018)	9 fitness trackers e.g. Fitbit Flex/One (FBO), Jawbone UP24	2018	130	**ST**: low correlation (range:−.02 to .33), mean difference: FBO −29.7 (SD 155.10)	**yes**	medical (28-33 minutes activity protocol based on free living tasks e.g. sitting)
(Cook et al. 2018)	Jawbone UP3	2018	43	**SL**: overestimation TST (39.6 min, P < .0001), sleep detection (sensitivity = 0.97)	**yes**	medical (1 night sleep in laboratory)
(Pelizzo et al. 2018)	Fitbit Charge HR	2018	30	**HR**: continuous ECG (r = 0.99)	**(yes)**	medical (preoperatively and undergoing laparoscopy/open surgery)
(Jones et al. 2018)	Fitbit Flex, ActiGraph GT3X+	2018	30	**ST**: MAPE: ≤1% (8-14 km/h), Standard Error of Measurement FBF ≤ 7 steps (8-14 km/h) and 9-19 steps (16 km/h)	**yes**	**fitness** (treadmill protocol at jogging and running speeds, 8 km/h to 16 km/h)
(Munck et al. 2018)	Fitbit One/Charge HR, Garmin Vivofit 2, Jawbone UP3/UP24	2018	22	**ST**: (except Jawbone UP24) percentage error was below 20% at all gait speeds	**yes**	medical (treadmill walking test: 3 sessions at 1, 2, and 3 minutes at different gait speeds)
(Burton et al. 2018)	Fitbit Flex/Charge HR	2018	31/ 30	**ST**: Intra Class Correlation's (ICC:0.86,95%CI:0.76,0.93), underestimation of steps	yes/ no	**general**/medical (Phase 1: two-minute-walk-test, 2 times; Phase 2: 14 days in a free-living environment)

Table 1. Related literature of commercial fitness tracker evaluation

rate (Jo et al. 2016), step count (Kaewkannate and Kim 2016), sleep duration (de Zambotti, Baker, et al. 2015) or energy expenditure measurement (Parak et al. 2017). The test setups are designed differently, including general scenarios with household activities, fitness and lifestyle oriented training protocols or medically motivated test arrangements. The general scenarios often assess predefined activity sequences from daily life according to their accuracy or correlation with reference values (Montoye et al. 2016; Sirard et al. 2017). In contrast to this, fitness setups include various training scenarios e.g. running or resistance exercises (Jo et al. 2016; Reddy et al. 2018) and evaluate the fitness tracker's performance during (high) physical activity. Medical test scenarios often emerge from acute care settings e.g. heart rate accuracy for patients at the intensive care unit (Kroll et al. 2016). Overall, each study reaches (slightly) different results for measurement accuracy, depending on the fitness tracker model and test setup. Still, many of the previous works conclude that there is high accuracy or strong correlation for at least one device and/or test setup within their study. Across all fitness tracker evaluations, only general tendencies can be identified e.g.

heart rate measurement gets inaccurate at high motion intensities, energy expenditure measurements are often imprecise, fitness trackers frequently overestimate sleep duration and underestimate step count. The majority of the studies is performed in a laboratory environment. Although this makes it possible to eliminate disturbance variables, there is a call to evaluate fitness trackers in the normal home environment, especially for sleep measurement (de Zambotti, Claudatos, et al. 2015). Additionally, previous studies stress that the accuracy of consumer-based devices is dependent on the activity that is performed (Montoye et al. 2016) and that there should be awareness of the strengths and limitations of these devices (Shcherbina et al. 2017). For this reason, we want to extend the body of knowledge and perform a field study that identifies under which conditions fitness trackers are sufficiently accurate to give reasonable behavior change advice in an individual health management setting. In contrast to many previous studies that evaluated fitness devices with a general or fitness purpose (see Table 1), our field study is performed with a background for individual health management. It differs from existing studies in a medical context, by being conducted in the everyday environment of participants.

Methodology

The evaluation of commercial fitness trackers is carried out as a field study that measures the patient's values in a real life context. So far there are predominantly laboratory studies in the area of fitness trackers which show results with a high internal validity. In contrast to that there should also be field investigations producing test results with a high external validity (Bortz and Döring 2006). To do so field studies take place in the natural environment of the subject group, but their naturalness goes at the expense of internal validity because the control of disturbing factors is only partly possible (Bortz and Döring 2006). This disadvantage can be accepted if the influencing factors of the investigation are controlled carefully. It should be noted that evaluation studies are typically field studies because the effectiveness of the devices should be tested under real and not under artificial, lab-like conditions (Bortz and Döring 2006).

Figure 1. Selection of commercial fitness trackers

Continuous monitoring can be especially beneficial for patients with chronic conditions (Roman 2015), but while fitness trackers nowadays offer a wide range of monitoring functionality, not all measured parameters are important for medical treatment. To choose the vital parameter measurements for our field study we compared the measurement capabilities of fitness trackers to important parameters that help patients to manage a chronic condition in general. Exercise is an effective treatment for many chronic diseases (Hoffmann et al. 2016), so we chose the measurement of the volume of physical activity (step count) and its intensity (heart rate). Insufficient sleep is also linked to the development and management of chronic conditions (Centers for Disease Control and Prevention (CDC) 2018), so we decided to include sleep duration as well. Based on these three parameters patients can be given reasonable behavior change advice to generally improve their chronic disease. Therefore, they provide the basis for the choice of commercial devices. Other influencing factors were their interoperability with mobile phone operating systems and their availability on the market. The final sample includes 15 fitness trackers that were bought between May 2017 and July 2018 (if available) in an electronics store or otherwise online. We assume that every device manufacturer uses proprietary hardware (sensors) and/or software (analysis algorithms). For this reason, we considered as many different manufacturers as possible. All fitness trackers are compatible with established mobile operating systems e.g. iOS and Android. Only the Apple Watch does not fulfil this requirement, still we included the device in the test scenarios due to its high popularity in the medical research field (El-Amrawy and Nounou 2015; Veerabhadrappa et al. 2018; Wallen et al. 2016). Figure 1 shows the exact device models that we evaluated, arranged according to their purchase order. In the following the devices are referenced by their manufacturer name (see Figure 1 bold words) for reasons of clarity.

For every relevant measurement parameter, a different scenario was developed. This took place at the office/work place and at home of the study participants. The experiments were conduct with 6 healthy persons (male: 3; female: 3; average age: 27,5 years) who all participated voluntarily and gave verbal informed consent. While testing, some elementary rules were followed, to control influencing factors due to incorrect device handling. One participant only wears one fitness tracker per wrist. If a person wears multiple devices on one wrist, the light of the PPG sensors measuring the pulse may influence one another (Ray Maker (DC Rainmaker) 2017). Additionally, the side of the wrist was customized according to the current participant (dominant/non-dominant wrist) to consider possible changes in the analysis algorithm. The manual was studied to make sure the tracker is in the right position on the wrist for the best measurement results. The battery was charged beforehand to ensure there were no interruptions while testing. To use the companion app, iPhones as well as Android Phones were used to synchronize the data to a central user account. The detailed test setup of the scenarios is described in the following. During each test scenario a maximum of two devices was worn, one on the left wrist and the other one on the right wrist.

Step count. We chose step count as a scenario that represents the amount of daily activity during a usual work day. For this reason, our test environment was an indoor hallway at the workplace's top floor (ca. 50 m length) that is usually walked several times during a day. The distance was walked for 20 times by each participant, which represents the average physical activity throughout 8-hours of work. Each participant received a list to fill in the fitness tracker models as well as the manually obtained step count. During walking, each participant counted the steps taken. After each iteration the value displayed on the fitness tracker monitor was noted as well as the manually obtained reference step count.

Heart rate. This value represents the activity/stress intensity of the participants during work. The experiment was divided into three phases with different intensities: resting (e.g. during desk work), standing (e.g. during a presentation) and slow running (e.g. hurrying from one appointment to another). Each participant received a heart rate chest strap (Medisana Heart Rate Belt Art. 99645) that was used as a reference value. The heart rate chest strap was connected to a mobile phone via Bluetooth that displayed the beats per minute (bpm) measurement in real time on the screen. The bmp value of the fitness tracker was directly read from the display of the device itself. Each phase lasted for one minute, which we defined as a reasonable period for the fitness tracker to calculate the bpm value. The heart rate was noted every ten seconds (seven measurement values per phase). Between each of the three phases was a 30-second-break so the analysis algorithm of the fitness tracker could adjust its bpm calculation. Between each of the five iterations was a three minutes break so the participant's heart rate dropped to a normal level after slow running.

Sleep duration. The duration of night's sleep shows the length of the rest period between two (working) days. The participants spent three consecutive nights sleeping with the same combination of two fitness tracker devices at home in their familiar environment. Each morning they filled out a questionnaire that is designed to repeatedly evaluate the experience of last night's sleep in a hospital setting and has already proven to be reliable (Ellis et al. 1981). The sleep duration measured by the fitness tracker was obtained after synchronizing with the companion app, because the data is not analyzed/displayed on the device itself. The results from the questionnaire are used as a reference value to be compared with the fitness tracker measurement.

Results

For each test scenario the mean absolute deviation as well as the mean absolute deviation in % was calculated. Additionally, it was measured if the values are mostly below/above the reference value or if certain patterns could be identified. To assess the measurement's accuracy, a maximum acceptable deviation from previous works is given as a comparison.

Step Count. For the steps the TomTom measured most accurate (1,85% deviation), whereas the Healbe is positioned last with 57,67% deviation within the iterations (see table 2). For the step count in our opinion the overall daily activity within a certain timeframe is more important for patient advice than deviations of single steps within a short distance. In this context we also calculated the overall step count for all 20 iterations and compared it to the reference value. The results are depicted in figure 2 and show that many fitness trackers compensate measurement errors within different iterations with the same amount of positive and negative false values that neutralize each other in the absolute step count. Overall, 8 out of 15

n = 20	Apple Watch	Fitbit	Fitbit Ionic	Garmin vivoactive	Garmin vivosmart	Healbe	Huawei	Jawbone	Microsoft	Oregon	Philips	Polar	**Tom Tom**	Withings	Xiaomi
mean absolute deviation in %	19,25	4,04	14,96	2,88	10,67	57,67	10,63	14,48	8,04	4,70	17,61	51,66	**1,85**	4,50	37,26
mean absolute deviation in steps	13,40	2,55	11,20	2,15	28,00	41,60	7,35	10,50	5,80	3,40	15,20	44,05	**1,45**	2,85	27,00

Table 2. Test results for step count

devices underestimated the total step count and the underestimation is on average higher (176,13 steps) than the overestimation (42,43 steps). Missing data (e.g. when the fitness tracker did not recognize the steps) were included in the calculation. In a previous quality evaluation of commercial pedometers, Tudor-Locke et al. (2006) conclude that research grade pedometers should not exceed 1% error most of the time when walking on a treadmill (3mph). Even though the TomTom nearly fulfils this requirement, many other devices exceed this threshold by far.

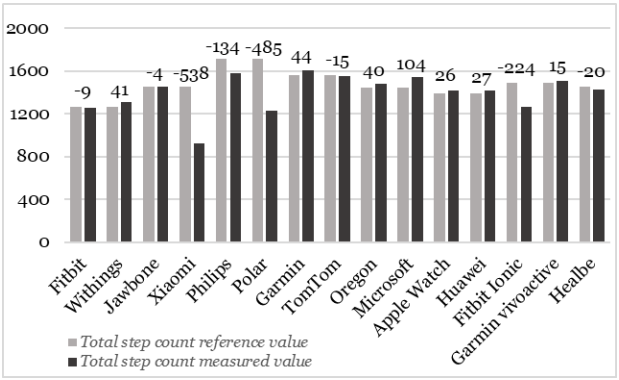

Figure 2. Absolute deviation of step count

Heart Rate. For the heart rate we consider the mean absolute deviation in bmp to be the most important measurement parameter. While the mean absolute deviation in % is influenced by the heart rate value itself (5 bmp deviation has a higher influence on lower than on higher heart rates when calculating the deviation in %) this is not the case when looking at the deviation in bmp. Individual health management advice can either be given by the average heart rate throughout the day or based on the current value in the immediate situation. The results of the heart rate field study are displayed in table 3. It shows that the Huawei performed best in our tests with a mean absolute deviation of 5,22 bpm. The worst performance is shown by the Healbe which differs 18,60 bpm on average from the reference value. Figure 3 shows the detailed results for the commercial fitness trackers that performed best (Huawei and Philips). When looking at the graphical representation in general it can be said that commercial fitness trackers often have problems to adapt to changes in the heart rate (transition period from resting to standing etc.) and that the

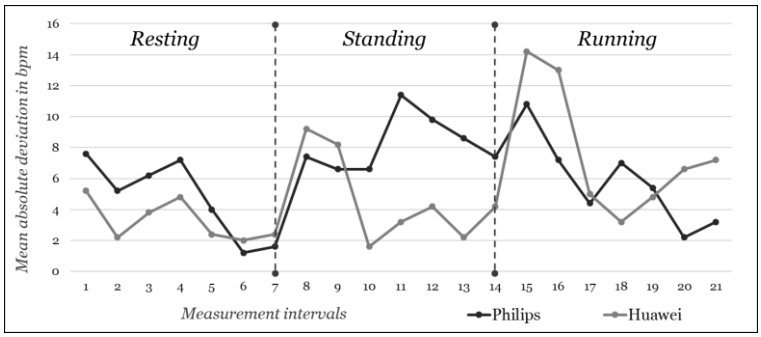

Figure 3. Mean absolute deviation in bpm for Huawei and Philips (over 5 iterations)

n = 105	Fitbit	Polar	Xiaomi	TomTom	Garmin vivosmart	Microsoft	Garmin vivoactive	Fitbit Ionic	Apple Watch	**Huawei**	Healbe	Philips
mean absolute deviation in %	11,77	10,18	14,03	13,61	9,74	13,79	8,40	5,21	9,90	**5,57**	17,48	7,47
mean absolute deviation in bpm	10,58	8,06	14,31	14,65	8,42	11,92	10,43	6,25	9,78	**5,22**	18,60	6,24
Total error	2	3	23	10	0	1	5	2	14	**3**	0	0

Table 3. Test results for continuous heart rate

measurement accuracy decreases with an increasing heart rate. Additionally, the high error rate (maximum error rate: Xiaomi = 21,9%) represents the failure of the fitness trackers to calculate a bpm value at all. These missing values were excluded from the calculation. In previous studies concerning heart rate measurement accuracy, an acceptable error range of 5% compared to 12-lead ECG-monitoring (Shcherbina et al. 2017) or a discrepancy of 5 bpm of continuous monitoring devices from nurse measurement (Weenk et al. 2017) are assumed as acceptable. The Huawei nearly reaches this requirement, but the other fitness trackers differ in up to 12,48%/13,60 bpm from this threshold.

Sleep Duration. The sleep duration has been measured over three consecutive nights. The deviation of the sleep duration for every single night in minutes is depicted in figure 4. For every device the mean absolute deviation is displayed in the bar chart as well. Based on this assessment the Huawei performed best with an average deviation of 10 minutes per night. The TomTom scored worst with a mean absolute deviation of 371,67 minutes. Overall, 8 out of 14 fitness trackers underestimated the sleep duration. Comparing the values of single outliers, the overestimations are much higher (e.g. 525 minutes for TomTom) than the underestimations (180 minutes for Philips). In the existing literature a difference of ≤ 30 min of total sleep time between two devices is defined as a satisfactory outcome, when comparing two commercially available actigraphs to polysomnography measurements (Meltzer et al. 2012). Looking at figure 4, only three devices fulfil this criterion.

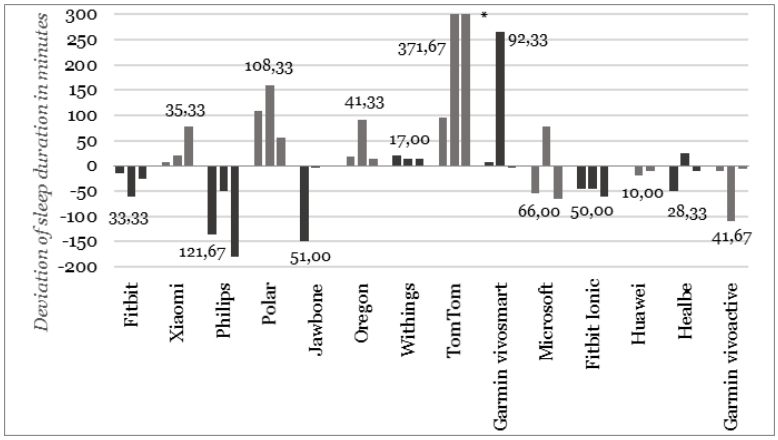

Figure 4. Deviation of sleep duration in minutes
(*very high values (495 minutes, 525 minutes) excluded for reasons of clarity)

Discussion

Our field study evaluated 15 commercial fitness trackers in a real life scenario regarding their applicability for reasonable behavior change advice. The results contribute to the literature, by giving further insides in test results and test scenarios of sensors for individual health management. In agreement with previous studies in this area, we could not determine the one fitness tracker that measures best for every test scenario. Still our tendencies mostly correspond with the results from existing studies e.g. heart rate gets inaccurate at higher bpm values and step count is often underestimated. In contrast to existing laboratory studies stating an overestimation of sleep duration, we observed that there are more fitness trackers underestimating this value, even though the outliers of overestimation can be extremely high. Overall, we could identify two main findings within our study that help practitioners to evaluate and develop feasible sensor applications for chronic disease management. We will elaborate on these in the following.

As a first result, *context-aware application scenarios* need to be designed to apply commercial fitness trackers for medical behavior change advice in practice. Our study showed that even in an (uncontrolled) real-life context there was at least one device per test scenario that complies with predefined (health related) standards. Still, this highly depends on the test setup and the device choice, so overall the accuracy is influenced by the specific application scenario. Kroll et al. (2016) recommend to identify patients that are most suitable for heart rate monitoring by commercial fitness devices. Following this notion, we propose to focus on the particular characteristics of different patient groups, device models and diseases. By assuring the best *patient-device-disease-fit*, optimal application scenarios can be developed. Especially the evaluation of step count measurement shows that there are differences in the accuracy per single measurement (per 50 m walked) and throughout the aggregated measurement (total step count of 20 iterations). Concerning reasonable interventions, it might not always be necessary to know if the patient walked exactly 45 or 55 steps but how much the patient moved throughout the whole day. For this reason, we should reconsider if commercial sensors need to be as accurate as medical sensors to be able to give

useful and appropriate health behavior change advice for the patient. Further research should consider *maximal deviation values for reasonable interventions*. To illustrate this, we apply our fitness tracker measurements for the scenario of depressive patients, which often rely on self-management due to under-provision of therapy places. Reduction of drive with increased fatigue and sleep disorders are symptoms for depression (cf. ICD-10-GM-2018 F32.-). A patient can get useful advice for interventions e.g. "to take a walk" when the total step count is below 1000 in the afternoon or "to keep in mind sleep hygiene" if last night's sleep duration was below 4 hours, without relying on absolutely accurate measurements. To make sure there is no harm for the patient when measurement errors or outliers occur, the maximal deviations and interventions need to be defined carefully. Still, helping patients on the basis of approximated vital parameters measurements might be better than to leave them alone in managing their chronic condition until accurate devices with a high usability are approved for medical treatment.

Secondly, *standardized test procedures* and protocols are necessary to evaluate sensor accuracy. While we showed that some commercial fitness trackers have the potential to accurately measure parameters for behavior change advice, there are still some issues that we identified in the process of fitness tracker evaluation. There is low transparency from device manufacturing companies (Shcherbina et al. 2017) regarding their different analysis algorithms and hardware conditions. These are not publicly accessible but highly influence the measurement results, which is shown by the different results of our field study. There is missing transparency whether the quality of the hardware sensors or the analysis algorithms causes the high deviation of some measurement results. Especially questionable are aggregated assessments of the tracker e.g. sleep quality/efficiency that are not displayed as "raw" data on the fitness tracker itself (de Zambotti, Claudatos, et al. 2015) and not defined in a standardized way. That's why we recommend to further research the *influence of analysis algorithms and hardware sensors on measurement accuracy*. Most (commercial and medical) sensors and algorithms are designed to measure standard values of healthy patients. Especially in the medical area there are many outliers when physiological data of ill patients are required. This is e.g. also the case in the context of depression where the measurement of heart rate might be complicated due to comorbidities. Many depressive patients do have heart diseases which may influence the heart rate measurement of fitness trackers. To be applicable in a medical context in practice the analysis algorithms of the devices need to be adjusted to interpret the sensor measurements correctly. Next to this there is in general a high susceptibility to outliers. Most of the sensors are still very error-prone and their measurement accuracy can be influenced by a patient's physical appearance e.g. skin-color and movement e.g. tension of arm muscles (Jo et al. 2016). In both cases the *sensor algorithms need to be individually adjusted to the user's characteristics*, which should be further investigated.

Conclusion and Future Research

Summing up, we could identify some test setups and fitness tracker models where the measurements were accurate enough to be used for reasonable behavior change advice. If the application scenario is designed clearly and standardized test procedures are developed, also commercial devices can enhance individual health management e.g. for patients with chronic conditions. The limitations of this study need to be acknowledged. The selected fitness trackers show a current market overview of the sensor performance at the moment but will likely change with the emergence of new technologies in the future. The test scenarios we designed have a low internal validity and are likely influenced through disturbance variables because they were conducted in an uncontrolled setting. This is especially true for the sleep environment in our field study, which can be influenced by various aspects at the participants' home. Still we see this as a strength of our study because it displays the real use case of the technology. The study participant group was very small, young and healthy which makes it hard to generalize the findings. We suggest future research to develop standardized test procedures and define application scenarios that are aware of the patient-device-disease context. Additionally, the current challenges of fitness tracker's transparency and outliers should be investigated, to be able to enhance individual health management effectively.

REFERENCES

Baig, M. M., Gholamhosseini, H., Moqeem, A. A., Mirza, F., & Lindén, M. 2017. "A Systematic Review of Wearable Patient Monitoring Systems – Current Challenges and Opportunities for Clinical Adoption," Journal of Medical Systems, (41), pp. 1-9.

Behkami, N. A., & Daim, T. U. 2012. "Research Forecasting for Health Information Technology (HIT), using technology intelligence," Technological Forecasting & Social Change, (79), pp. 498-508.

Bortz, J., & Döring, N. 2006. "Forschungsmethoden und Evaluation für Human- und Sozialwissenschaften", Springer Berlin Heidelberg.

Burton, E., Hill, K. D., Lautenschlager, N. T., Thøgersen-Ntoumani, C., Lewin, G., Boyle, E., & Howie, E. 2018. "Reliability and validity of two fitness tracker devices in the laboratory and home environment for older community-dwelling people," BMC Geriatrics, pp. 1–12.

Centers for Disease Control and Prevention (CDC). 2018. Sleep and Chronic Disease. Retrieved February 27, 2019, from https://www.cdc.gov/sleep/about_sleep/chronic_disease.html.

Cook, J. D., Prairie, M. L., & Plante, D. T. 2018. "Ability of the Multisensory Jawbone UP3 to Quantify and Classify Sleep in Patients With Suspected Central Disorders of Hypersomnolence: A Comparison Against Polysomnography and Actigraphy," Journal of Clinical Sleep Medicine (14:5), pp. 841-848.

de Zambotti, M., Baker, F. C., & Colrain, I. M. 2015. "Validation of Sleep-Tracking Technology Compared with Polysomnography in Adolescents," Sleep 38(9), pp. 1461–1468.

de Zambotti, M., Baker, F. C., Willoughby, A. R., Godino, J. G., Wing, D., Patrick, K., & Colrain, I. M. 2016. "Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents," Physiology and Behavior (158), pp. 143–149.

de Zambotti, M., Claudatos, S., Inkelis, S., Colrain, I. M., & Baker, F. C. 2015. "Evaluation of a Consumer Fitness-Tracking Device to Assess Sleep in Adults: Evaluation of Wearable Technology to Assess Sleep," Chronobiol Int. (32:7), pp. 1024–1028.

Dunn, J., Runge, R., & Snyder, M. 2018. "Wearables and the medical revolution," Personalized Medicine. (15:5), pp. 429–448.

Eaton, S., Roberts, S., & Turner, B. 2015. "Delivering person centred care in long term conditions," The BMJ - Spotlight: Patient Centred Care (181), pp. 1–4.

El-Amrawy, F., & Nounou, M. I. 2015. "Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial?," Healthcare Informatics Research (21:4), pp. 315–320.

Ellis, B. W., Johns, M. W., Lancaster, R., Raptopoulos, P., Angelopoulos, N., & Priest, R. G. 1981. "The St. Mary's Hospital Sleep Questionnaire: A Study of Reliability," Sleep (4:1), pp. 93–97.

Ferguson, T., Rowlands, A. V, Olds, T., & Maher, C. 2015. "The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study," International Journal of Behavioral Nutrition and Physical Activity (12:42), pp. 1–9.

Gao, Y., Li, H., & Luo, Y. 2015. "An empirical study of wearable technology acceptance in healthcare," Industrial Management & Data Systems (115:9), pp. 1704–1723.

Giansanti, D., Maccioni, G., & Morelli, S. 2008. "An experience of health technology assessment in new models of care for subjects with Parkinson's disease by means of a new wearable device," Telemedicine and E-Health (14:5), pp. 467–472.

Gualtieri, L., Rosenbluth, S., & Phillips, J. 2016. "Can a Free Wearable Activity Tracker Change Behavior? The Impact of Trackers on Adults in a Physician-Led Wellness Group," JMIR Research Protocols (5:4).

Henriksen, A., Mikalsen, M. H., Woldaregay, A. Z., Muzny, M., Hartvigsen, G., Hopstock, Grimsgaard, S. 2018. "Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables," Journal of Medical Internet Research (20:3), pp. 11-19.

Hibbard, B. J. H., & Greene, J. 2013. "What The Evidence Shows About Patient Activation: Better Health Outcomes and Care Experiences; Fewer Data On Costs," Health Affairs (32:2), pp. 207–214.

Hoffmann, T. C., Maher, C. G., Briffa, T., Sherrington, C., Benell, K., Alison, J., Singh, M.F. 2016. "Prescribing exercise interventions for patients with chronic conditions," CMAJ (188:7), pp. 510–518.

Jo, E., Lewis, K., Directo, D., Kim, M. J. Y., Dolezal, B. A. 2016. "Validation of biofeedback wearables for photoplethysmographic heart rate tracking," Journal of Sports Science and Medicine (15:3), pp. 540–547.

Jones, D., Crossley, K., Dascombe, B., Hart, H. F., Kemp, J. 2018. "Validity and Reliability of the Fitbit Flex TM and Actigraph GT3X + at Jogging and Running Speeds," (13:5), pp. 860–870.

Kaewkannate, K., & Kim, S. 2016. "A comparison of wearable fitness devices," BMC Public Health (16:1).

Kroll, R. R., Boyd, J. G., & Maslove, D. M. 2016. "Accuracy of a Wrist-Worn Wearable Device for Monitoring Heart Rates in Hospital Inpatients: A Prospective Observational Study," Journal of Medical Internet Reserach (18), pp. 1-11.

Lee, J. M., Kim, Y., & Welk, G. J. 2014. "Validity of consumer-based physical activity monitors," Medicine and Science in Sports and Exercise (46:9), pp. 1840–1848.

Meltzer, L. J., Walsh, C. M., Traylor, J., & Westin, A. M. L. 2012. "Direct Comparison of Two New Actigraphs and Polysomnography in Children and Adolescents," Sleep, pp. 12–15.

Montoye, A. H. K., Nelson, M. B., Kaminsky, L. A., Dickin, D. C., & Montoye, A. H. K. 2016. "Validity of consumer-based physical activity monitors for specific activity types," Medicine & Science in Sports & Exercise (48:8), pp. 1619–1628.

Munck, K., Christensen, M. H., Tahhan, A., Dinesen, B. I., Spindler, H., Hansen, J., Nielsen, O.W., Leth, S. 2018. "Evaluation of self-trackers for use in telerehabilitation," Journal of Usability Studies (13:3), pp. 125–137.

Parak, J., Uuskoski, M., Machek, J., & Korhonen, I. 2017. "Estimating Heart Rate, Energy Expenditure, and Physical Performance With a Wrist Photoplethysmographic Device During Running," JMIR Mhealth Uhealth (5:7), pp.1-12.

Pelizzo, G., Guddo, A., Puglisi, A., De Silvestri, A., Comparato, C., Valenza, M., Bordonaro, M., Calcaterra, V. 2018. "Accuracy of a Wrist-Worn Heart Rate Sensing Device during Elective Pediatric Surgical Procedures," Children (5), pp. 1–7.

PwC. 2016. The Wearable Life 2.0 - Connected living in a wearable world. Retrieved from https://www.pwc.com/us/en/industry/entertainment-media/assets/pwc-cis-wearables.pdf.

Ray Maker (DC Rainmaker). 2017. Thoughts on the wearables studies (including The Stanford Wearables study). Retrieved from https://www.dcrainmaker.com/2017/06/thoughts-on-the-wearables-studies-including-the-stanford-wearables-study.html.

Reddy, R. K., Pooni, R., Zaharieva, D. P., Senf, B., El, J., Dassau, E., … Rickels, M. R. 2018. "Accuracy of Wrist-Worn Activity Monitors During Common Daily Physical Activities and Types of Structured Exercise: Evaluation Study," JMIR Mhealth Uhealth (6:12), pp. 1-18.

Roman, D. H. (2015). The Digital Revolution comes to US Healthcare. Goldman Sachs. Retrieved from https://www.wur.nl/upload_mm/0/f/3/8fe8684c-2a84-4965-9dce-550584aae48c_Internet of Things 5 - Digital Revolution Comes to US Healtcare.pdf.

Shah, Y., Dunn, J., Huebner, E., & Landry, S. 2017. "Wearables data integration: Data-driven modeling to adjust for differences in Jawbone and Fitbit estimations of steps, calories, and resting heart-rate," Computers in Industry (86), pp. 72–81.

Shcherbina, A., Mattsson, C. M., Waggott, D., Salisbury, H., Christle, J. W., Hastie, T., … Ashley, E. A. 2017. "Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort," Journal of Personalized Medicine (7:3), pp.1-12.

Sirard, J. R., Masteller, B., Freedson, P. S., Mendoza, A., & Hickey, A. 2017. "Youth Oriented Activity Trackers: Comprehensive Laboratory- and Field-Based Validation" Journal of Medical Internet Research (19:7), pp. 1-13.

Swan, M. 2012. "Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0," Journal of Sensor and Actuator Networks (1:3), pp. 217–253.

Technology Advice. 2014. "Wearable Technology & Preventative Healthcare - Trends in Fitness Tracking among U.S. Adults".

Tudor-Locke, C., Sisson, S. B., Lee, S. M., Craig, C. L., Plotnikoff, R. C., & Bauman, A. 2006. "Evaluation of quality of commercial pedometers," Canadian Journal of Public Health, pp.10-15.

Ummels, D., Beekman, E., Theunissen, K., Braun, S., & Anna, J. 2018. "Counting Steps in Activities of Daily Living in People With a Chronic Disease Using Nine Commercially Available Fitness Trackers: Cross-Sectional Validity Study," JMIR Mhealth Uhealth (6:4), pp. 1–14.

United Nations. 2015. World Population Ageing. Retrieved from http://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf.

Veerabhadrappa, P., Moran, M. D., Renninger, M. D., Rhudy, M. B., Dreisbach, S. B., & Gift, K. M. 2018. "Tracking Steps on Apple Watch at Different Walking Speeds," Journal of General Internal Medicine (33:6), pp. 795–796.

Wallen, M. P., Gomersall, S. R., Keating, S. E., Wisløff, U., & Coombes, J. S. 2016. "Accuracy of heart rate watches: Implications for weight management," PLoS ONE (11:5), pp. 1–9.

Weenk, M., van Goor, H., Frietman, B., Engelen, L. J., van Laarhoven, C. J. H. M., Smit, J., Bredie, S. J., van de Belt, T. H. 2017. "Continuous Monitoring of Vital Signs Using Wearable Devices on the General Ward: Pilot Study," JMIR MHealth and UHealth (5:7). p. e91.

World Economic Forum. 2014. Health Systems Leapfrogging in Emerging Economies - Project Paper.

World Health Organization. 2018. What are noncommunicable diseases? Retrieved from http://www.euro.who.int/en/health-topics/noncommunicable-diseases/ncd-background-information/what-are-noncommunicable-diseases.