Comparison of accuracy of activity measurements with wearable activity trackers in wheelchair users: a preliminary evaluation

Vergleich der Genauigkeit von Aktivitätsmessungen mit tragbaren Aktivitätstrackern bei Rollstuhlfahrern: Eine vorläufige Evaluation

Abstract

Background: Central nervous system diseases or injuries such as spinal cord injury (SCI) are often associated with a severe impairment of ambulatory function and result in wheelchair dependency. Typical long-term complications of wheelchair users include pressure injuries, spasticity, musculoskeletal pain and psychological issues. It is hypothesized that the occurrence of these complications is related to the level of physical activity (PA). While a low level of PA often results in skin or cardiovascular problems, a very high PA level results in shoulder pain and muscular fatigue. However, current evidence for this hypothesis is based on qualitative data from patient interviews. To investigate the relation between PA and health complications in a more objective manner, we propose the use of wearable activity trackers. As a first step, accuracy of common trackers – Apple Watch Series 4 and Fitbit Flex 2 – for quantification of wheelchair pushes was determined. For the Apple Watch, two different conditions are compared: out of the box usage and usage with GPS calibration.

Methods: We used a 200 m outdoor test course. Healthy subjects were asked to propel with a wheelchair along the course while ground truth was captured by manually counting the wheelchair pushes. This procedure was conducted twice: once with the GPS-calibrated Apple Watch and Flex 2 in parallel and once with the non-calibrated Apple Watch. We analyzed the reproducibility of the ground truth measurement method by calculating the interrater reliability for two rater roles. To compare for accuracy of the activity trackers, we calculated differences between trackers and ground truth and analyzed them in Bland-Altman plots. To conclude on the need for GPS calibration of the Apple Watch in future studies, we did an equivalence test.

Results: Twenty subjects without motor impairments participated in driving the test course. We found a reproducibility of the ground truth measurement of ICC(2,1)=0.981 (CI=0.96<ICC<0.99). The percentage of error for the calibrated Apple Watch is 13.9%, for the uncalibrated Apple Watch 22.8% and for the Flex 2 59.7%. Bland-Altman plots indicate a tendency for a higher error for test series with a higher number of pushes for the Flex 2. The equivalence test was significant for the defined equivalence boundaries of 15%.

Conclusions: The ICC of the ground truth measurements indicates a high reproducibility of manual counting. Percentages of error show the highest accuracy for the calibrated Apple Watch. However, the results of the significant equivalence test suggest that the Apple Watch can be used without complex and time-consuming calibration. Due to extremely high error, PA tracking by Fitbit Flex 2 cannot be recommended for wheelchair users. This work is intended to serve as a basis for a future evaluation study of an activity tracker, for which we recommend using the uncalibrated Apple Watch in chronic wheelchair users. The evaluated

Nils-Hendrik Benning[1]
Petra Knaup[1]
Rüdiger Rupp[2]

1 Institute of Medical Biometry and Informatics, Heidelberg University, Heidelberg, Germany

2 Spinal Cord Injury Center, Heidelberg University Hospital, Heidelberg, Germany

tracker is intended to be used in a larger healthcare registry for follow-up of SCI patients.

Keywords: medical informatics, fitness trackers, physical activity, wheelchairs, spinal cord injuries

Zusammenfassung

Hintergrund: Erkrankungen des zentralen Nervensystems oder Verletzungen wie z.B. eine Querschnittlähmung (QSL) sind oft mit einer schweren Beeinträchtigung der Gehfunktion verbunden und führen zur Abhängigkeit vom Rollstuhl. Typische Langzeitkomplikationen von Rollstuhlfahrern sind Dekubitus, Spastizität, Muskel-Skelett-Schmerzen und psychische Probleme. Es wird vermutet, dass das Auftreten dieser Komplikationen mit dem Maß der körperlichen Aktivität (PA) zusammenhängt. Während ein niedriges PA-Niveau häufig zu Haut- oder Herz-Kreislauf-Problemen führt, führt ein sehr hohes PA-Niveau zu Schulterschmerzen und muskulärer Ermüdung. Die bisherige Evidenz für diese Hypothese basiert jedoch auf qualitativen Daten aus Patientenbefragungen. Um den Zusammenhang zwischen PA und gesundheitlichen Komplikationen objektiver zu untersuchen, schlagen wir die Verwendung von tragbaren Aktivitätstrackern vor. In einem ersten Schritt wurde die Genauigkeit der verbreiteten Tracker – Apple Watch Series 4 und Fitbit Flex 2 – zur Quantifizierung von Rollstuhlschüben bestimmt. Für die Apple Watch werden zwei verschiedene Zustände verglichen: Die Nutzung im Auslieferungszustand und die Nutzung mit GPS-Kalibrierung.

Methoden: Wir haben einen 200 m langen, realitätsnahen Testparcours genutzt. Gesunde Probanden wurden gebeten, mit einem Rollstuhl die Strecke zu fahren, während manuelles Zählen der Rollstuhlschübe als Goldstandard erfasst wurde. Dieses Verfahren wurde zweimal durchgeführt: Einmal mit der GPS-kalibrierten Apple Watch und dem Flex 2 parallel und einmal mit der nicht kalibrierten Apple Watch. Wir analysierten die Reproduzierbarkeit der Goldstandard-Methode, indem wir die Interrater-Reliabilität für zwei Rater-Rollen berechnet haben. Um die Genauigkeit der Aktivitätstracker zu vergleichen, berechneten wir die Differenzen zwischen Trackern und dem Goldstandard und analysierten sie in Bland-Altman-Plots. Um auf die Notwendigkeit einer GPS-Kalibrierung der Apple Watch in zukünftigen Studien zu schließen, führten wir einen Äquivalenztest durch.

Ergebnisse: Zwanzig Personen ohne motorische Beeinträchtigung nahmen an der Fahrt auf dem Testparcours teil. Wir stellten eine Reliabilität der Goldstandardmessung von ICC(2,1)=0,981 fest (CI=0,96<ICC<0,99). Der Fehlerprozentsatz für die kalibrierte Apple Watch beträgt 13,9%, für die unkalibrierte Apple Watch 22,8% und für das Flex 2 59,7%. Die Bland-Altman-Plots zeigen eine Tendenz zu einem höheren Fehler für Testreihen mit einer höheren Anzahl von Schüben beim Flex 2. Der Äquivalenztest war für die definierten Äquivalenzgrenzen von 15% signifikant.

Schlussfolgerungen: Der ICC der Goldstandardmessung weist auf eine hohe Reliabilität der manuellen Zählung hin. Die Fehleranteile zeigen die höchste Genauigkeit für die kalibrierte Apple Watch. Die Ergebnisse des signifikanten Äquivalenztests legen jedoch nahe, dass die Apple Watch ohne aufwendige und zeitintensive Kalibrierung verwendet werden kann. Aufgrund des extrem hohen Fehlers kann das PA-Tracking von Fitbit Flex 2 für Rollstuhlfahrer nicht empfohlen werden. Diese Arbeit soll als Grundlage für eine zukünftige Evaluationsstudie eines Aktivitätstrackers dienen, für die wir die Verwendung der unkalibrierten Apple Watch bei chronischen Rollstuhlfahrern empfehlen. Der evaluierte

Tracker soll in einem größeren Register für die Nachsorge von SCI-Patienten eingesetzt werden.

Schlüsselwörter: medizinische Informatik, Fitnesstracker, körperliche Aktivität, Rollstühle, Rückenmarkverletzungen

Introduction

Neurological diseases or trauma such as stroke, spinal cord injury (SCI) or multiple sclerosis often result in severe impairments of the walking ability. To compensate for the loss of ambulation, wheelchairs are used to regain at least some mobility. The current estimate of the number of wheelchair users in Germany is 1.35 million [1]. Generally, the use of wheelchairs entails several side effects, such as pressure injuries [2], pain [3] and psychosocial impacts [4]. All these side effects lead to a deterioration in quality of life for patients with neuromuscular disorders [5].

More specifically, Tweedy et al. [6] reported first determinants for side effects like shoulder pain or depressive symptoms in people with SCI: They are less likely to occur, if a given minimum amount of physical activity (PA) is regularly executed. Jörgensen et al. [7] found a generally improved cardiovascular health for people with SCI, who have a lifestyle with a higher amount of PA. Tawashy et al. [3] found a negative correlation between PA and pain, fatigue as well as depressive symptoms. On the other hand, Akbar et al. [8] reported a higher probability for rotator cuff tears in persons with SCI, who regularly do overhead sports. Crespo-Ruiz et al. [9] found a deteriorated response of tissues for persons with SCI and high amounts of PA. This could result in pressure injuries.

However, in all above-mentioned studies PA data has been captured by asking subjects about their lifestyle. For example, in the study by Tawashy et al. [3] PA data has been captured using PARA-SCI, an interview-based 3-day recall format, which is based on a questionnaire about the PA on the day of the interview. For representativeness, PARA-SCI needs 3 interviews in each patient, taking 20–30 minutes [10] and collects only highly subjective data biased by the self-perception of the patients. This way of data collection is not just resource-consuming, but often also inaccurate [11]. Furthermore, side effects typically develop over a long period after the initial inpatient rehabilitation phase [12]. Therefore, only long-term collection of PA allows any conclusions on the dependency between PA and the health status. The inaccuracy and the spotlight character of interview-based PA data collections calls their practical value into question.

To overcome these limitations, wearable activity measuring devices are considered as a promising tool for quantitative research [13]. Such devices, like fitness trackers, are considered as a more accurate data source [11], also enabling the objective examination of physical parameters over a long period of time.

To ensure a high quality for wearable-generated data, we compared the accuracy of different devices. We chose the number of wheelchair pushes as a pendant to the step count in pedestrians as primary variable. The rationale for this decision is that this parameter represents a clinically relevant, easily quantifiable and comparable integrative unit of measurement. It also enables patients to interpret their PA for themselves, which offers the potential of implementing patient participation mechanisms, like providing patients with feedback about their level of PA potentially leading to a more active lifestyle. For recording of the number of wheelchair pushes, different hardware is available: wheelchair-specific activity trackers (programmed to track pushes) and activity trackers being unspecific to wheelchair users (tracking steps). The latter represent mass products with a broad range of models, including many with relatively low prices [14]. Availability of wearable activity trackers specifically designed for PA measurements in wheelchair users seems to be limited, resulting in very low scientific evidence. An exception is the Apple Watch which offers the ability to track wheelchair pushes since the release of WatchOS 3 in 2016. Apple Watch supports wheelchair push tracking out of the box, but additionally offers the possibility for calibration of the watch to specific wheelchair usage patterns. We evaluated Apple Watch as a wheelchair-specific option for assessing PA together with a representative device of wheelchair unspecific trackers in a competitive setting against the ground truth.

This work represents a preparation for a future evaluation study of an activity tracker, intended to be used for long-term PA measurement in a healthcare registry (ParaReg) for long-term follow-up of SCI patients. The aim of this study was

1. to analyze which activity tracker offers the higher accuracy for counting pushes of wheelchair drivers and
2. to come up with recommendations about the need for calibration of the Apple Watch in a subsequent evaluation study.

Methods

Materials

For the comparison, we used an Apple Watch Series 4 with WatchOS 5.3.2. To our knowledge, there are no other commercially available wearable activity trackers offering specialized signal processing for wheelchair users. To activate and configure the Apple Watch, an Apple iPhone is needed, therefore an iPhone 7 was used. As an activity tracker unspecific to wheelchair users we chose a Fitbit Flex 2. This device is capable of tracking steps in pedestrians. To reliably measure the ground truth
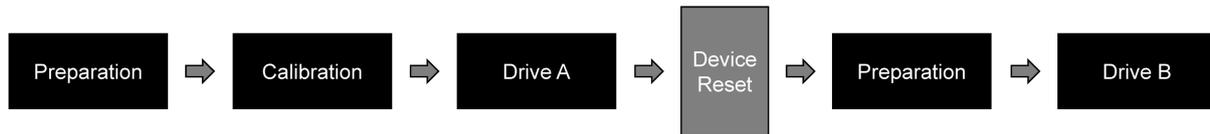
of push count, a digital tally counter was used. The utilized wheelchair model was a Sopur Easy 300.

Participants

We did not include wheelchair-dependent patients as study participants, but rather involved healthy subjects who temporarily used a wheelchair for our study. Only subjects without known diseases that could influence physical behavior were included. They were recruited in the environment of a university research institute. All participants had to be at least 18 years old and the participation was voluntary.

Setting

To simulate everyday life use, an outdoor test course on a university campus was defined (see Figure 1). It consisted of a mostly smooth surface (paving stones) with positive and negative pitches. The number of right- and left-turns (each 90°) was matched not to introduce a systematic error due to the fact that trackers are only worn on one arm. The total distance of the course was approximately 200 m.

Figure 1: Schematic course layout

Procedure

The procedure was split into 6 steps (Figure 2).

1. The materials and subjects were prepared.
2. The Apple Watch was calibrated. This calibration necessitates a 20-minute drive, which is considered to also have a familiarization effect on the participants wheelchair usage.
3. Test drive A with the Flex 2 and calibrated Apple Watch was done.
4. The calibration of the Apple Watch was deleted.
5. The preparation step was repeated, because the devices were removed from the arm for the reset.
6. A second test series (drive B) with uncalibrated Apple Watch was done, to compare the trackers accuracy in both the calibrated and uncalibrated states.

Preparation

Participants were asked to wear both trackers on their non-dominant arm. The Apple Watch was placed on the right side from the Flex 2 (Apple Watch distal on left arm, proximal on right arm), to avoid unwanted interaction between the side-mounted buttons of Apple Watch and Flex 2. Next, participants were asked to sit down in the wheelchair. The footrest was set to an appropriate height (if needed). The participants were given time to get used to the wheelchair by driving on the course.

Calibration

Afterwards the recommendations of Apple Inc. to increase accuracy of the Apple Watch were applied [15]. First, demographic data of the participant was configured in the Apple Health App: birthdate, gender, body height and weight. Apple Inc. further recommends a calibration of the Apple Watch with the help of GPS data, to increase accuracy [16]. For this calibration the workout mode was activated as recommended [16] and a free 20 minute drive across a university campus was done. This drive contained no indoor or longer roofed sections, which could limit access to the GPS service. Fitbit does not provide information on any kind of calibration; thus, it was worn in idle mode while the calibration.

Drive A

The wheelchair was placed in a starting position where it could not move by itself, so that the participant was able to put his arms in his lap. Before reading baseline values of the daily counts (daily step count for Flex 2 and daily push count for Apple Watch), the participant kept this pose for 60 seconds to allow the trackers to set a stable baseline value. Daily push count of Apple Watch was then read via the "Activity" menu on the watch itself. The daily step count of Flex 2 was read via the Fitbit app, because the tracker has no display. After the initial values were recorded, the participants were asked to start the drive on the test course with a speed comfortable for them. They were also asked to count the pushes with the non-dominant arm. A push was defined by a movement with the hand to the front while holding the handlebar of the wheel. Potential backwards movements or braking postures did not count as a push. The examiner walked next to the participant during the whole drive and also counted the pushes, using a tally counter. After completing the course, the participants resumed the resting po-

Figure 2: Procedure overview

Table 1: Recorded variables of the test drives

Test series	Variable	Description
Drive A	Ground truth (subject)	Number of pushes counted by subject
	Ground truth (examiner)	Number of pushes counted by examiner
	Apple Watch (calibrated)	Number of pushes measured by calibrated Apple Watch 4
	Flex 2	Number of pushes measured by Fitbit Flex 2
Drive B	Ground truth (subject)	Number of pushes counted by subject
	Ground truth (examiner)	Number of pushes counted by examiner
	Apple Watch (uncalibrated)	Number of pushes measured by uncalibrated Apple Watch 4
	Flex 2	Number of pushes measured by Fitbit Flex 2

sition and waited for 60 seconds. In parallel, the examiner recorded the manually counted push counts both from the examiner and the participant. After the 60 seconds were completed, he also recorded the daily push counts from the trackers, as described above.

Device reset

While conducting the experiments neither iOS nor WatchOS offered a function to delete the calibration, making it necessary to completely reset both the Apple Watch and the paired iPhone. This reset takes approximately 20 minutes. Therefore, drive B was planned for a later time on the same day or no later than the day after.

Drive B

This drive was done with the reset (uncalibrated) Apple Watch. Therefore, the step preparation was needed as well, before the course drive could be repeated. For comparability, again both trackers, Apple Watch and Flex 2, were worn.

Variables

Table 1 shows the recorded variables. The counts of pushes in the course were calculated from the daily counts of the trackers. This was done on a paper form used during the experiments by subtracting the daily step/push count at the end of the course from the daily step/push count before starting the course.

Analysis

General

To analyze the reproducibility of the ground truth measurement method, we calculated the interrater reliability (IRR) between the counted pushes of the study parti-

cipants and the examiner. Therefore, we calculated an intraclass correlation coefficient (two-way random effects, absolute agreement, single rater/measurement) estimate and its 95% confident interval [17] with GNU R package "irr" version 0.84.1 [18] based on a single rating, absolute agreement, and two-way random effects model. For the IRR, we pooled the data from the calibrated and uncalibrated recordings because the calibration is not expected to influence manual counting by subject and examiner. For further analysis, we used the mean value of the counted pushes from the participant and the examiner as ground truth.

In order to analyze the reproducibility between drive A and B, we calculated the test-retest reliability between both drives for the percentages of error of Flex 2. Test-retest reliability was calculated as an intraclass correlation coefficient (two-way mixed effects, absolute agreement, multiple raters/measurements) estimate and its 95% confident interval with the above-mentioned software package. We only performed this test for the Flex 2 because it is applied in the same state during both drives (no difference due to calibration).

Accuracy of activity trackers for counting pushes of wheelchair drivers

To compare the overall accuracy of the examined trackers, we calculated the sum of absolute differences (SOAD) of pushes between each tracker and the corresponding ground truth. The SOAD were made assessable by dividing the SOAD of each test series by the corresponding cumulated count of pushes (ground truth), which results in the percentage of error. To examine data for systematic effects like tendencies for deviation in case of higher push counts, Bland-Altman plots were created for each test series.

Table 2: Differences between the pushes counted by the trackers and the ground truth

Activity tracker	Range of differences	SOAD	Percentage of error
Apple Watch (calibrated)	+3 – +40	271	13.9%
Apple Watch (uncalibrated)	−20 – +46	401	22.8%
Flex 2 (drive A)	+105 – +184	2,890	148.4%

Need for calibration of the Apple Watch in a subsequent evaluation study

The Apple Watch can be used in an uncalibrated and calibrated state. It is stated that the calibration increases accuracy [15]. On the other hand, it increases the time and physical efforts for the study participant – in particular those with motor impairments – in an evaluation study significantly. The need for a continuous 20-minute drive excludes participants with more severe motor impairments who are not able to propel a manual wheelchair for such a prolonged time. Therefore, we consider a calibration effect of 15% (or smaller) as neglectable, because we want to analyze long-term trends in PA, which are expected to be greater than this effect size. To proof that a calibration is not needed with respect to this boundary, we performed an equivalence test. We used a test of one-sided significance (TOST). The null hypothesis is the presence of a calibration effect exceeding the equivalence bounds. The alternative hypothesis is a calibration effect within the equivalence bounds, or the absence of a measurable effect. The equivalence bounds were defined by 85% and 115% of the mean number of pushes (from all data present).

Results

Sample

We included 20 subjects according to the inclusion and exclusion criteria. Eighteen subjects (90%) were male and 2 (10%) were female. The average age was M=30.8 years (SD=6.7 years), ranging from 25 to 47 years. In 17 participants (85%) the trackers were mounted on the left arm and in 3 of them (15%) on the right arm.

Interrater reliability between subjects and examiner

Through pooling the ground truth data from drive A and drive B, the sample size for calculation of ICC is 40 with n=1 for the examiner role and n=20 for the subject role. The ICC results in a value of 0.981. The 95% confidence interval is 0.96<ICC<0.99.

Test-retest reliability for Flex 2

The test-retest reliability between drive A and drive B for the percentages of error of Flex 2 was calculated based on a sample size of 20. The result of the ICC calculation is 0.785 with a confidence interval of 0.468<ICC<0.914.

Absolute and relative deviation from ground truth

We analyzed the range of differences, the SOAD as well as the percentage of error for all test series. The results are presented in Table 2.

Differences range from −20 to +184 pushes for all test series. The ranges for the two series with Apple Watch are smaller than the one of Flex 2. The same comparison applies for SOAD, where both Apple Watch values are less than a quarter of the Flex 2 SOAD. Percentage of error ranges from 13.9% up to 148.4%, where Flex 2 has the highest error share.

Bland-Altman plots for visualization of deviations from ground truth

The differences for the Bland-Altman plots were calculated by subtracting ground truth from tracker measurement. Thus, a positive difference represents an over-estimation of the number of wheelchair pushes by the tracker and a negative difference means an under-estimation. The plots can be found in Figure 3, Figure 4 and Figure 5.
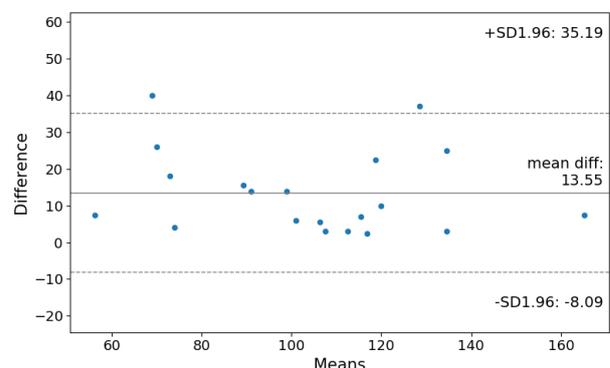

Figure 3: Bland-Altman plot of calibrated Apple Watch (difference of pushes: Apple Watch vs. averaged ground truth while drive A)
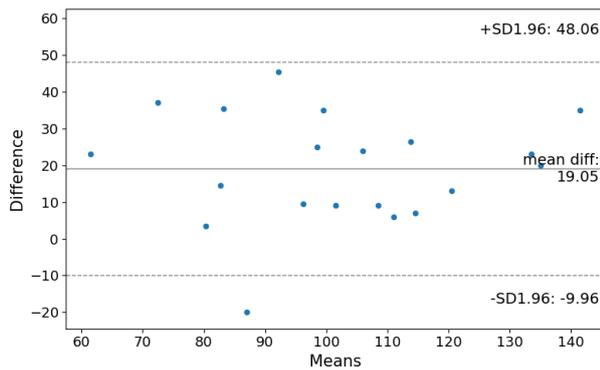
Figure 4: Bland-Altman plot of uncalibrated Apple Watch (difference of pushes: Apple Watch vs. averaged ground truth while drive B)
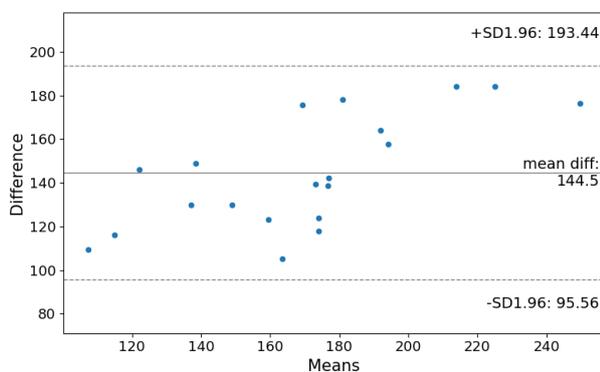

Figure 5:-Bland-Altman plot of Flex 2 (difference of pushes: Flex 2 vs. averaged ground truth while drive A)

Equivalence test for calibrated and uncalibrated use of the Apple Watch

The mean count of pushes from all available test series is 103.125. 15% margins for lower and upper equivalence bound are –15.4688 and 15.4688. The lower bound t-value is 2.35 and the p-value is 0.012. For the upper bound, we found a t-value of –4.95 and a p-value of 0.000008. Degrees of freedom were 38 and the 95% confidence interval was –14.077 for lower bound and 3.077 for the upper bound. The equivalence test is significant within the equivalence bounds, given a significance level of 0.05.

Discussion

Percentage of error for Flex 2 is rather high compared to Apple Watch. For Flex 2, we measured a percentage of error of 59.7%, which is considered as not acceptable for research use. Apple Watch shows a much higher accuracy with a percentage of error of 22.8% in uncalibrated test series. The differences are in the range from –20 to 46, so there is a tendency to over-estimate. There is an even lower percentage of error (13.9%) when calibrating the Apple Watch first. For the calibrated watch, we found a range of differences from 3 to 40 pushes, so it generally

over-estimates the count. The differences for both Apple Watch series compare well to the absolute results from Case et al. [19], who evaluated the accuracy of wearables for step counting while walking. They reported percentages of error ranging from –22.7% to –1.5%. Thus, the scale of difference is comparable. However, we found an over-estimation for wheelchair users, whereas Case et al. [19] found an under-estimation for pedestrians. In an article of An et al. [20] wrist-worn step counters were investigated at normal walking speeds. They found percentages of error of 0.7% to 17.4%. Only the accuracy of calibrated Apple Watch lies within this boundaries, the other two series indicate a higher difference. To our knowledge, there are no studies examining the accuracy of wheelchair-specific activity trackers.

Subsidiary, Apple Watch (calibrated) provides the highest accuracy in our test series. Apple Watch (uncalibrated) measures wheelchair pushes less accurate, but still on an acceptable level. The accuracy of the Flex 2 designed for step counting seems to be unusable to reliably measure data in wheelchair users for research purposes. Bland-Altman plots confirm the high error rate of the Flex 2. All test series contain an over-estimation of at least 100 pushes. There is a tendency for a higher percentage of error for Flex 2 test series with a higher push count: Series with a higher mean push count than 180 generally have a difference above the mean difference of 144.5, whereas series with a lower mean push count tend to have differences lower than the mean difference. We suspect a higher amount of vibrations due to the increased push count, causing the detection of even more steps. This non-linear effect limits the correction of data by assuming the presence of a systematic error. Bland-Altman plots of both calibrated and uncalibrated Apple Watch do not indicate such a systematic error.

The equivalence test shows that there is no significant effect of calibrating the Apple Watch within the equivalence bounds of 15%. As we strive for a physical activity estimate, enabling us to keep track of long-term changes in behavioral patterns, we consider these equivalence bounds to be rational. This indicates that the calibration can be neglected for future identically structured evaluations. However, the calibration could be advantageous for users who plan a long-term personal use of Apple Watch.

The utilized materials contain popular commercially available activity trackers [21]. Thus, they have the potential to be broadly available in many study participants, reducing the potential budget of research trials. Especially the Apple Watch offers numerous lifestyle features, which have the potential to motivate wheelchair users to continuously use the device. This long-term usage is highly relevant for the generation of research data. On the other hand, this motivational aspect might introduce a bias in respect to assessing the typical activity patterns without the use of an activity tracking device. Furthermore, Apple Watch is a rather cost-intensive activity tracker, which could limit accessibility for some groups.

There are limitations in our study: The study sample is not representative neither for a national normal population nor for the wheelchair user population with respect to gender and age. The sample size is reasonable for a preliminary study. It is also sufficient to provide an estimate for the variability of such measurements, which enables a reliable sample-size estimation. However, we did not include real wheelchair users. Thus, our subjects were not used to the wheelchair, which might alter the movement patterns compared to experienced wheelchair users. All test series were performed on the same test course, which ensures comparability between the series. Our course does not fully represent a real-world scenario, but it was not set up in a laboratory or other controlled environments. Thus, our results might translate to real-world scenarios; however, they do still not account for all influences from everyday life, such as varying surfaces or different turning degrees. We organized the procedure with the Apple Watch in a calibrated state first and a subsequent drive in uncalibrated state. This causes more effort because a time-consuming device reset is necessary in between. However, the time spent was worth it, because a changed procedure would deteriorate the comparability of the test series: The calibration drive incorporates a 20-minute drive with the wheelchair, during which we observed a significant progress in driving experience. Thus, a measurement before and after the calibration drive would be less comparable. By measuring all test series after the calibration drive, we achieve a higher comparability.

The interrater reliability between the study participants and the examiner indicates an excellent consistency between the two raters [17]. This suggests manual counting of wheelchair pushes to be a reproducible method to measure the ground truth. However, when discrepancies of counted pushes appeared, the examiner was confident that he could not underestimate/overestimate the correct count of pushes that much. The subjects on the other hand often expressed uncertainty about their counted pushes, because they had to concentrate on propelling the wheelchair as well. Even though the consistency is very satisfying, we propose counting exclusively by the examiner for future work. Another approach could be video recording the wheelchair's handlebar and count the pushes offline. The higher effort of this approach could be compensated by a more accurate measurement of the ground truth, because ambiguous movements could be viewed several times or also in slow motion. Additionally, more than one examiner may be involved in the counting of the wheelchair pushes.

The results of the test-retest reliability indicate a good reliability between the measurement of error between drive A and B [17]. This could be considered to reflect the reliability of the whole procedure, because Flex 2 was used in both drives without any modifications. However, a dedicated test-retest reliability study with the Apple Watch is recommended for future work.

Conclusions

Generally, the examined tracking devices for measurement of the number of wheelchair pushes seem to provide a lower accuracy than the devices counting steps in pedestrians. The ranking from best to poorest accuracy is: Apple Watch (calibrated), Apple Watch (uncalibrated), Fitbit Flex 2. Considering equivalence bounds of 15% as neglectable, we found that the calibration of the Apple Watch has no significant effect on the accuracy of push counting. Thus, we propose an evaluation study for the Apple Watch with experienced wheelchair users without a preceding calibration.

Notes

Competing interests

The authors declare that they have no competing interests.

Funding

References

1. Bedarf an barrierefreien Wohnungen in Deutschland. In: nullbarriere.de [Internet]. 2020 [cited 2020 Feb 27]. Available from: https://nullbarriere.de/bedarf-barrierefreie-wohnung.htm

2. Sprigle S, McNair D, Sonenblum S. Pressure Ulcer Risk Factors in Persons with Mobility-Related Disabilities. Adv Skin Wound Care. 2020 Mar;33(3):146-54. DOI: 10.1097/01.ASW.0000653152.36482.7d

3. Tawashy AE, Eng JJ, Lin KH, Tang PF, Hung C. Physical activity is related to lower levels of pain, fatigue and depression in individuals with spinal-cord injury: a correlational study. Spinal Cord. 2009 Apr;47(4):301-6. DOI: 10.1038/sc.2008.120

4. Singh H, Scovil CY, Yoshida K, Oosman S, Kaiser A, Jaglal SB, Musselman KE. Capturing the psychosocial impacts of falls from the perspectives of wheelchair users with spinal cord injury through photo-elicitation. Disabil Rehabil. 2020 Jan 6:1-10. DOI: 10.1080/09638288.2019.1709911

5. Pousada García T, Groba González B, Nieto Rivero L, Pereira Loureiro J, Díez Villoria E, Pazos Sierra A. Exploring the Psychosocial Impact of Wheelchair and Contextual Factors on Quality of Life of People with Neuromuscular Disorders. Assist Technol. 2015;27(4):246-56. DOI: 10.1080/10400435.2015.1045996

6. Tweedy SM, Beckman EM, Geraghty TJ, Theisen D, Perret C, Harvey LA, Vanlandewijck YC. Exercise and sports science Australia (ESSA) position statement on exercise and spinal cord injury. J Sci Med Sport. 2017 Feb;20(2):108-15. DOI: 10.1016/j.jsams.2016.02.001

7. Jörgensen S, Svedevall S, Magnusson L, Martin Ginis KA, Lexell J. Associations between leisure time physical activity and cardiovascular risk factors among older adults with long-term spinal cord injury. Spinal Cord. 2019 May;57(5):427-33. DOI: 10.1038/s41393-018-0233-5

8. Akbar M, Brunner M, Ewerbeck V, Wiedenhöfer B, Grieser T, Bruckner T, Loew M, Raiss P. Do overhead sports increase risk for rotator cuff tears in wheelchair users? Arch Phys Med Rehabil. 2015 Mar;96(3):484-8. DOI: 10.1016/j.apmr.2014.09.032

9. Crespo-Ruiz B, del-Ama AJ, Jiménez-Díaz FJ, Morgan J, de la Peña-González A, Gil-Agudo ÁM. Physical activity and transcutaneous oxygen pressure in men with spinal cord injury. J Rehabil Res Dev. 2012;49(6):913-24. DOI: 10.1682/jrrd.2011.05.0087

10. Ginis KA, Latimer AE, Hicks AL, Craven BC. Development and evaluation of an activity measure for people with spinal cord injury. Med Sci Sports Exerc. 2005 Jul;37(7):1099-111. DOI: 10.1249/01.mss.0000170127.54394.eb

11. Ma JK, McCracken LA, Voss C, Chan FHN, West CR, Martin Ginis KA. Physical activity measurement in people with spinal cord injury: comparison of accelerometry and self-report (the Physical Activity Recall Assessment for People with Spinal Cord Injury). Disabil Rehabil. 2020 Jan;42(2):240-6. DOI: 10.1080/09638288.2018.1494213

12. Hossain MS, Islam MS, Rahman MA, Glinsky JV, Herbert RD, Ducharme S, Harvey LA. Health status, quality of life and socioeconomic situation of people with spinal cord injuries six years after discharge from a hospital in Bangladesh. Spinal Cord. 2019 Aug;57(8):652-61. DOI: 10.1038/s41393-019-0261-9

13. Riffenburg KM, Spartano NL. Physical activity and weight maintenance: the utility of wearable devices and mobile health technology in research and clinical settings. Curr Opin Endocrinol Diabetes Obes. 2018 Oct;25(5):310-4. DOI: 10.1097/MED.0000000000000433

14. Henriksen A, Haugen Mikalsen M, Woldaregay AZ, Muzny M, Hartvigsen G, Hopstock LA, Grimsgaard S. Using Fitness Trackers and Smartwatches to Measure Physical Activity in Research: Analysis of Consumer Wrist-Worn Wearables. J Med Internet Res. 2018 Mar;20(3):e110. DOI: 10.2196/jmir.9157

15. Apple Inc. Get the most accurate measurements using your Apple Watch [Internet]. 2019 [cited 2020 Mar 4]. Available from: https://support.apple.com/de-de/HT207941

16. Apple Inc. Calibrating your Apple Watch for improved Workout and Activity accuracy [Internet]. 2019 [cited 2020 Mar 4]. Available from: https://support.apple.com/en-us/HT204516

17. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155-63. DOI: 10.1016/j.jcm.2016.02.012

18. Gamer M, Lemon J, Singh P. irr: Various Coefficients of Interrater Reliability and Agreement [Internet]. Version 0.84.1. 2019 Jan 26 [cited 2020 Mar 5]. Available from: https://cran.r-project.org/package=irr

19. Case MA, Burwick HA, Volpp KG, Patel MS. Accuracy of smartphone applications and wearable devices for tracking physical activity data. JAMA. 2015 Feb;313(6):625-6. DOI: 10.1001/jama.2014.17841

20. An HS, Jones GC, Kang SK, Welk GJ, Lee JM. How valid are wearable physical activity trackers for measuring steps? Eur J Sport Sci. 2017 Apr;17(3):360-8. DOI: 10.1080/17461391.2016.1255261

21. Canalys. North American wearables market hits US$2.0 billion in Q2 2019 [Internet]. 2019 [cited 2020 Mar 6]. Available from: https://www.canalys.com/newsroom/north-america-wearables-q2-2019

Corresponding author:

Nils-Hendrik Benning
Universitätsklinikum Heidelberg, Institut für Medizinische Biometrie und Informatik, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany
Nils.Benning@med.uni-heidelberg.de