Measuring steps with the Fitbit activity tracker: an inter-device reliability study

Manon L. Dontje, Martijn de Groot, Remko R. Lengton, Cees P. van der Schans & Wim P. Krijnen

To cite this article: Manon L. Dontje, Martijn de Groot, Remko R. Lengton, Cees P. van der Schans & Wim P. Krijnen (2015) Measuring steps with the Fitbit activity tracker: an inter-device reliability study, Journal of Medical Engineering & Technology, 39:5, 286-290, DOI: 10.3109/03091902.2015.1050125

To link to this article: https://doi.org/10.3109/03091902.2015.1050125
INNOVATION

Measuring steps with the Fitbit activity tracker: an inter-device reliability study

Manon L. Dontje*,1, Martijn de Groot1,2, Remko R. Lengton2, Cees P. van der Schans2,3 and Wim P. Krijnen2,4

1Hanse University of Applied Sciences, Quantified Self Institute, Groningen, The Netherlands, 2Hanze University of Applied Sciences, Research and Innovation Group in Health Care and Nursing, Groningen, The Netherlands, 3Department of Rehabilitation, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands, and 4Department of Statistics, University of Groningen, Groningen, The Netherlands

Abstract
Activity trackers like Fitbit are used for self-tracking of physical activity by an increasing number of individuals. Comparing physical activity scores with peers can contribute to the desired behavioural change. However, for meaningful social comparison a high inter-device reliability is paramount. This study aimed to determine the inter-device reliability of Fitbit activity trackers in measuring steps. Ten activity trackers (Fitbit Ultra) were worn by a single person (male, 46 years) during eight consecutive days. Inter-device reliability was assessed on three different levels of aggregation (minutes, hours, days) with various methods, including intra-class correlation coefficient (ICC), Bland-Altman plots, limits of agreement (LOA) and Mixed Model Analysis. Results showed that the inter-device reliability of the Fitbit in measuring steps is good at all levels of aggregation (minutes, hours, days), but especially when steps were measured per day. This implies that individuals can reliably compare their daily physical activity scores with peers.

Keywords
Accelerometry, Fitbit, objective measurement, physical activity, reliability

History
Received 13 February 2015
Revised 3 May 2015
Accepted 6 May 2015

1. Introduction

Regular physical activity has multiple beneficial health effects. For example, it reduces the risk of many chronic diseases such as cardiovascular disease and diabetes and it improves psychosocial well-being and quality-of-life [1,2]. Healthy adults are recommended to accumulate on average 10 000 steps per day to maintain or improve health [3]. Despite the well-known health benefits of regular physical activity, many people fail to meet the recommended level of physical activity [4].

To support the inactive population in increasing their physical activity, numerous activity trackers have been developed. Not only expensive accelerometers which can be used in research, but also the number of activity trackers accessible for individual consumers is increasing rapidly. Such activity trackers are often small, low-cost and user friendly. Activity trackers measure physical activity of the user, often expressed as the number of steps. Self-monitoring of daily physical activity may increase awareness about one’s own activity level, which is a pre-requisite for behavioural change [5,6].

By using activity trackers, people can compare their own physical activity with that of their peers, which can further enhance behavioural change. However, comparing scores requires sufficient inter-device reliability. Also, when activity trackers are being used in research, a high inter-device reliability is one of the pre-requisites [7]. Reliability of accelerometers is often examined by focusing on raw activity counts, but raw activity counts are often difficult to interpret, especially with regard to physical activity guidelines. Raw activity counts can be summarized and converted into more meaningful variables such as number of steps per minute, hour or day. For each of these derived variables it is important to determine the inter-device reliability, because each variable can be used for social comparison and research-related purposes.

One of the most popular recent activity trackers on the consumer market is Fitbit (Fitbit Inc., San Francisco, CA). To date, only little is known about the validity and reliability of this device [8,9]. Research suggests that the inter-device reliability for measuring steps on a treadmill is high [8,9], but there is no information on inter-device reliability for measuring steps in free-living conditions yet. Therefore, the aim of this study was to determine the inter-device reliability of Fitbit activity trackers in measuring steps during free-living conditions, over different units of time.
2. Methods

2.1. Design

Ten activity trackers (Fitbit Ultra, Fitbit Inc., San Francisco, CA) were worn by a single person (one of the researchers, male, 46 years) during 8 consecutive days, except during sleep and water-based activities (e.g. showering, bathing). The trackers were attached to the trouser, equally distributed over the left and right side pocket.

2.2. Measurements

The Fitbit (Fitbit Ultra) is a small device which uses a three dimensional accelerometer to measure movement of the user. The main outcome measure is steps, which can be combined with user data to calculate among others distance walked and energy expenditure. Scores are presented on the display and can be uploaded to the Fitbit website, where an overview of physical activity is presented together with a comparison with friends. The Fitbit API was used to extract all minute-to-minute data per day for each device. Since each device detects number of steps per minute, data per device consisted of the number of steps over $8 \times 24 \times 60 = 11 \, 520$ min in total.

2.3. Statistical analyses

Statistical analysis was performed by the statistical programming language R [10]; $p$ values $\leq 0.05$ were considered significant. Three levels of aggregation were distinguished: minutes, hours and days. For each aggregation level the reliability was examined in six steps, to be described subsequently below.

The first step was to examine whether the devices were equal in detecting movement vs no movement. Therefore, data were dichotomized (movement vs no movement) at the minute and hour level and the proportion of movement of each device was calculated. The proportions of movement of the 10 devices were compared with each other and equality was tested with the proportions test.

The second step was to obtain an impression on the degree of normality of the 10 activity trackers by analysing the measurements per minute by the Shapiro-Wilk normality test. The third step was to calculate the two-way intra-class correlation coefficients (ICC) for absolute agreement in order to estimate the proportion of variance attributable to the devices [11]. The ICC’s were calculated by comparing 45 pairs of distinct devices (10 devices make 45 possible twin combinations).

The fourth step was to determine the concordance correlation coefficient (CCC). This measure is comparable to the ICC, except that the CCC combines coefficients of precision and accuracy into a coefficient that measures the agreement to the identity line [12]. Values can range between −1 and 1, indicating perfect disagreement to perfect agreement.

The fifth step was to visualise the agreement between each pair of Fitbits with Bland Altman plots. The limits of agreement (LOA) were also computed, to determine boundaries that contain 95% of the measurements [13,14]. The total deviation index (TDI) was computed to determine boundaries such that 95% of the measurements are within the boundary from their target values [12]. Bland Altman plots were also used to visualise a possible increase of variability of the measurements with size, outliers, as well as existence of systematic bias.

The sixth step was to test any systematic effects in mean number of steps due to any of the devices. Therefore, Mixed Models Analyses were conducted, taking the devices as fixed effects and the aggregation level as random effects.

The afore-mentioned agreement coefficients were calculated for all 45 distinct pairs of Fitbits. There was no ‘‘gold standard’’ to compare each individual tracker with. Even so, to examine the reliability of the Fitbit we determined the agreement coefficients also by comparing the measurements of each Fitbit with the ‘‘true value’’ of measurement. The true value of measurement was approximated by the average of the remaining nine measurements. To obtain an impression of closeness between the average and each single Fitbit the mean difference (mean compared with the averaged remaining nine devices), percentage relative error ((mean difference/mean) *100), ICC, CCC and its precision and accuracy measures, LOA and TDI were calculated. The LOA relative to the mean were calculated as well, to see whether there is an increase in precision with the aggregation level.

3. Results

The first step showed that the 10 devices were (almost) completely equal in detecting movement vs no movement when analysed in minutes as well as in hours ($p = 0.98$ and $p = 1.00$, respectively). This analysis was not applicable for aggregation level day, because there was no day without registered movement.

The Shapiro-Wilk normality tests indicated that normality of the measurements of the 10 activity trackers increased with level of aggregation. In 54% of the measurements of the 10 activity trackers over all minutes normality was not rejected, over all hours this was 84% and over all days this was 100%.

Third, the median (range) ICC was 0.90 (0.83–0.96), 1 (0.99–1) and 1 (0.99–1) for level of aggregation minutes, hours and days, respectively.

The CCC had a median (range) of 0.90 (0.83–0.96), 1 (0.99–1) and 0.99 (0.98–1) for level of aggregation minutes, hours and days, respectively.

Figure 1 presents examples of Bland Altman plots between two devices at three different levels of aggregation, including the corresponding LOA. For aggregation level minute, the median (range) LOA over the 45 distinct pairs of devices was 27.90 (17.47–36.86) steps, which indicates that 95% of the measurements were within the boundaries of 27.90 steps above and below the mean difference. For aggregation level hour, the median (range) LOA was 124.80 (62.62–229.80) steps. For aggregation level day, the median (range) LOA was 693.00 (228.40–1173.00) steps, which indicates that 95% of the measurements were within the boundaries of 693 steps above and below the mean difference. The median (range) TDI was 27.91 (17.54–36.89) steps, 129.00 (62.85–235.50) and 962.40 (244.6–1524.0) steps for level of aggregation minutes, hours and days, respectively.

The sixth step was to test any systematic effects in the number of steps depending on the different devices with
Mixed Model Analyses. There was a significant effect for devices when measured in minutes ($F = 8.816; p < 0.0001$), implying a significant systematic measurement difference. Five devices contributed significantly to the model, of which the size of the effect ranged from $-1.33$ to $-0.52$ steps (Table 1). There was also a significant effect for devices when measured in hours ($F = 8.133; p < 0.0001$). As shown in Table 1, four devices contributed significantly to the model, of which the size of the effect ranges from $-16.26$ to $-32.78$ steps. The Mixed Model Analysis also showed a significant effect for devices when measured in days ($F = 4.328; p < 0.001$). As shown in Table 1, only two devices contributed significantly to the model, of which the size of effect ranges from $-499.88$ to $-295.63$ steps.

The final step to examine reliability of the Fitbit was to calculate the agreement coefficients in comparison with the ‘‘true value’’ of measurement. Per level of aggregation, the approximated true mean steps per minute, mean difference (mean compared with the averaged remaining nine devices), percentage relative error, LOA, TDI, CCC, ICC and the LOA relative to the mean are presented in Table 2. For aggregation level minutes, the LOA relative to the mean ranged between 0.44–0.79, indicating that the LOA were ~times smaller in size than the mean number of steps per minute. For aggregation level hours, the LOA relative to the mean ranged between 0.07–0.24, indicating that the LOA were more than 5-times smaller in size than the mean number of steps per hour. Also, for aggregation level days, the LOA relative to the mean ranged between 0.02–0.06, indicating that the LOA were more than 25-times smaller in size than the mean number of steps per day.
The aim of this study was to determine the inter-device reliability of Fitbit Ultra activity trackers in measuring steps during free-living conditions, over different units of time. The results show that the inter-device reliability of the Fitbit Ultra is replaceable by another within reliable standards.

This is based on rigorous analysis using various methods for assessing reliability of 10 Fitbit activity trackers. The reliability of the devices was examined at three different levels of aggregation over time, i.e., minutes, hours, and days, with various methods. First, the results indicate that the 10 devices were equal in detecting movement vs no movement.

### 4. Discussion

The aim of this study was to determine the inter-device reliability of Fitbit Ultra activity trackers in measuring steps during free-living conditions, over different units of time (minutes, hours, days). The results show that the inter-device reliability was (reasonably) good at all levels of aggregation, but especially when steps were counted per day. This indicates that social comparison is reliable for daily measurements and that the Fitbit Ultra is replaceable by another within reliable standards.
This means roughly that, when the user was moving, it was reliably detected by each device, whether activity was measured in steps per minute or per hour. The results showed that most reliability coefficients improved with increasing level of aggregation (from minutes to days). For example, the degree of normality increased from 54% (aggregation level minutes) to 100% (aggregation level days). The ICCs and CCCs increased from 0.90 to unity with increasing levels of aggregation. This indicates that the agreement between the different devices is excellent, especially for measuring steps per day. This is in line with other studies [8,9]. Generally, ICC > 0.80 and CCC > 0.95 are considered acceptable levels of agreement [15–17]. The increasing reliability with increasing level of aggregation could also be concluded from the Bland–Altman plots, as the number of outliers decreased when the level of aggregation increased from minute to day. Moreover, the LOA relative to the mean decreased enormously as the level of aggregation increased from minute to day. When steps were counted per minute, the LOA were 2-times smaller than the mean number of steps. When steps were counted per hour, the LOA were 5-times smaller than the mean number of steps. Also, when steps were counted per day, the LOA were ~25-times smaller than the mean number of steps.

The aforementioned methods showed reasonably high reliability, although the mixed model analyses indicated significant Fitbit effects for a minority of cases. All of these seem small in size from a clinical point of view. All of these effects are responsible for the content and writing of the paper.

The analysis of method comparison studies indicated significant Fitbit effects for a minority of cases. All of these seem small in size from a clinical point of view. The maximum effects are ~1, 26 and 402 steps for aggregation level minute, hour and day, respectively. The maximum difference for all devices is 3.3%, which seems clinically acceptable. It comes close to the 3% error of miscounting which is maximally accepted by the quality standards of the Japanese industry of pedometers [18]. In general, an error percentage of less than 10% is considered acceptable for step counters [18]. Another study also found a low measurement error for Fitbits [9]. They compared steps of Fitbits to manual step count during walking on a treadmill and found that the measurement error was less than 1.3%. Taking the results of both studies together, the measurement error of Fitbit activity trackers seems acceptable.

When interpreting the results of the present study, it should be taken into account that only 10 activity trackers were examined. However, due to the design of this study 45 pairs of measurements could be examined with various methods. Nevertheless, for future research it is recommended to include a larger number of activity trackers, on a range of different subjects and activities for inter- and intra-device comparisons. A strength of the current study is that reliability was examined with several methods which all have their own distinct value, but when used together are more worthwhile than the sum of parts. Another strength is that the 8 days of measurement included different types of physical activity in real life, from light intensity physical activity to vigorous physical activity. In conclusion, the inter-device reliability of the Fitbit Ultra in measuring steps in free-living conditions is good, especially when measured per day. This implies that social comparison is reliable for daily measurements and that the Fitbit Ultra can be replaced by another Fitbit Ultra within reliable standards.

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References