Influence of Accelerometer Calibration on the Estimation of Objectively Measured Physical Activity: The Tromsø Study

Marc Weitz,1 Bente Morseth,2 Laila A. Hopstock,3 and Alexander Horsch1

1Department of Computer Science, UiT The Arctic University of Norway, Tromso, Norway; 2School of Sport Sciences, UiT The Arctic University of Norway, Tromso, Norway; 3Department of Health and Care Sciences, UiT The Arctic University of Norway, Tromso, Norway

Accelerometers are increasingly used to observe human behavior such as physical activity under free-living conditions. An important prerequisite to obtain reliable results is the correct calibration of the sensors. However, accurate calibration is often neglected, leading to potentially biased results. Here, we demonstrate and quantify the effect of accelerometer miscalibration on the estimation of objectively measured physical activity under free-living conditions. The total volume of moderate to vigorous physical activity (MVPA) was significantly reduced after post hoc auto-calibration for uniaxial and triaxial count data, as well as for Euclidean Norm Minus One and mean amplitude deviation raw data. Weekly estimates of MVPA were reduced on average by 5.5, 9.2, 45.8, and 4.8 min, respectively, when compared to the original uncalibrated estimates. Our results indicate a general trend of overestimating physical activity when using factory-calibrated sensors. In particular, the accuracy of estimates derived from the Euclidean Norm Minus One feature suffered from uncalibrated sensors. For all modalities, the more uncalibrated the sensor was, the more MVPA was overestimated. This might especially affect studies with lower sample sizes.

Keywords: ActiGraph, accelerometry, counts, auto-calibration, physical activity intensity, hip

Accelerometers are an indispensable tool in various disciplines to measure physical activity. Application ranges from small-scale studies (Deraas et al., 2021; Hoaas et al., 2016) to large-scale population studies like US NHANES (Troiano et al., 2008), the UK Biobank (Doherty et al., 2017), the German National Cohort (Leitzmann et al., 2020), and the Norwegian Tromsø Study (Sagelv et al., 2019). This development has been facilitated by the increasing availability of sensors that can store multiday to multiweek recordings. Additionally, the growing awareness of social desirability biases in self-reported physical activity (Sallis & Saelens, 2000) has contributed to the increased use of accelerometers.

Accelerometers record physical activity signals mechanically and convert them into electrical signals (i.e., voltage, resistance, or capacitance; Van Hees et al., 2009). These analog signals are then digitized and stored on the sensor in a compressed format of gravitational units (1 g = 9.80665 m·s⁻²). Some manufacturers, like ActiGraph, aggregate the raw data on either the sensor (older models) or post hoc in their accompanying software (newer models) into custom units called counts. How ActiGraph counts are calculated has not been known until recently (Neishabouri et al., 2022) and has been subject to speculation for many years (Brønd et al., 2017). Furthermore, the non-disclosure has been criticized for preventing study comparability (Van Hees et al., 2016), and researchers aimed to provide sensor-independent aggregation algorithms (Brønd et al., 2017; John et al., 2019).

An important prerequisite to accurately measure movements is the correct calibration of the sensors. Usually, accelerometers are initially calibrated as part of the manufacturing process such that under nonmovement at exact orientation of any of the three axes toward the Earth’s center of gravity, the sensor will accurately output ±1 g for this axis and 0 g for the other two axes (Lötters et al., 1998). The manufacturer’s calibration can later be controlled for by repeating the procedure and evaluating whether the sensor correctly outputs ±1 g or 0 g for each direction. However, this procedure is cumbersome and often neglected in practice (Van Hees et al., 2014). Van Hees et al. (2014) investigated the effects of auto-calibration on two features derived from the raw acceleration signal of wrist-worn accelerometers: The Euclidean Norm Minus One (ENMO) and the band-pass filtered Euclidean norm (BFEN). The authors showed that auto-calibration had a significant effect on the average and distribution of both features, even though the effect was stronger for ENMO than for BFEN. Despite being originally developed for wrist-worn accelerometers, the calibration procedure can be assumed to be independent of the sensor placement and has been applied to, for example, hip placement (Hildebrand et al., 2014).

When using accelerometers, the most common goal is to measure the physical activity of the observed subject. Physical activity, in this context, is often synonymously used with moderate to vigorous physical activity (MVPA), an intensity class defined from energy expenditure (Freedson et al., 1998). MVPA has become an important measurement due to its potentially health-enhancing effects (Garber et al., 2011) and is used in many physical activity regulations, for example, by the World Health Organization (WHO). For that reason, we focus on the estimation of MVPA in this study and assume that our results will also hold for energy expenditure. A significant amount of research has focused on finding correlates between the signals measured by accelerometers and MVPA (Freedson et al., 1998; Hildebrand et al., 2014; Sasaki et al., 2011; Vähä-Ypyä et al., 2015). The goal of these studies was to find appropriate thresholds that can classify accelerometer-based measures into distinct categories of energy expenditure either on...
count data (Freedson et al., 1998; Sasaki et al., 2011), the ENMO of the raw data (Hildebrand et al., 2014), or the mean amplitude deviation (MAD) of the raw data (Vähä-Yypä et al., 2015). However, as MVPA is defined by energy expenditure, thresholds depend significantly on the demographics; for an overview of different thresholds, see Migueles et al. (2017).

To our knowledge, no study has yet examined the effect of sensor calibration on the estimation of MVPA. In particular, the effect on count-based estimates is still unknown due to the long-lasting nondisclosure of the count algorithm. We assume that auto-calibration will reduce the time spent in MVPA significantly as auto-calibration seems to generally lower the average acceleration (Van Hees et al., 2014). We further assume a reduction in MVPA minutes for the count-based data due to a certain similarity to the BFEN feature studied by van Hees et al. (2014). However, given the structural differences in both features, the latter hypothesis appears less certain than the former.

**Methods**

**Data**

We used cross-sectional raw acceleration data from the seventh survey of the population-based Tromsø Study (n = 6,115). The average age of the participants was 63 years, and 53.6% were women. A detailed description of the sample and data set has been published previously (Hopstock et al., 2022; Sagelv et al., 2019). The raw acceleration data were collected by ActiGraph wGT3X-BT monitors worn for 7 consecutive days on the right hip at a sampling frequency of 100 Hz. Data collection was conducted in 2015–2016 and was approved by the Regional Committee for Medical Research Ethics (REC North ref. 2014/940) and the Norwegian Data Protection Authority, and all participants gave written informed consent. Furthermore, all methods were carried out in accordance with relevant guidelines and regulations (i.e., Declaration of Helsinki).

**Procedure**

We used the auto-calibration procedure proposed by van Hees et al. (2014) with the implementation from GGIR (Migueles et al., 2019) to calibrate the raw acceleration data. Subsequently, we calculated the ENMO values as:

\[
ENMO = \max \left( \sqrt{y^2 + x^2 + z^2} - 1.0 \right),
\]

and the MAD values as:

\[
MAD = \frac{1}{n} \left| r_i - \bar{r} \right|,
\]

with

\[
r_i = \sqrt{x_i^2 + y_i^2 + z_i^2},
\]

and the ActiLife counts using the Agcounts Python package (ActiGraph, LLC, 2022). ENMO was calculated on 1-s epochs (Hildebrand et al., 2014), while MAD values were calculated on 6-s epochs (Vähä-Yypä et al., 2015). We applied demographic-appropriate cut points to data of four types: (a) the uniaxial count data (Freedson et al., 1998), (b) the vector magnitude of the triaxial count data (Sasaki et al., 2011), (c) the ENMO of the triaxial raw data (Hildebrand et al., 2014), and (d) the MAD of the triaxial raw data (Vähä-Yypä et al., 2015) on the uncalibrated and the auto-calibrated raw data, leading to a total of eight different estimates of MVPA per participant. We only included participants for whom the calibration algorithm was successful (99.43% of the days). We excluded 1,085 (2.49%) not fully recorded days (usually at the beginning or at the end of the recording, depending on the initialization). Furthermore, we used the nonwear time detection algorithm proposed by van Hees et al. (2011) and excluded all days with >180 min of nonwear time (2931 days/6.89% of the remaining days). For the remaining data, we calculated the average time spent in MVPA per day for each participant with more than four valid days according to the aforementioned criteria (n = 5,827). Pairwise t tests were conducted to test for differences between the uncalibrated and auto-calibrated MVPA estimates. Furthermore, regression models were used to quantify the effect of device calibration error on the observed differences.

**Results**

Minutes of MVPA per day were significantly reduced after post hoc auto-calibration for all investigated data types, uniaxial count data, \((n, 5,826) = 61.13, p < .001\); triaxial count data, \((n, 5,826) = 75.30, p < .001\); ENMO data, \((n, 5,826) = 30.36, p < .001\); and MAD data, \((n, 5,826) = 72.35, p < .001\), respectively. In total, the average number of minutes in MVPA was reduced by 47 s for the uniaxial count data, by 78 s for the triaxial count data, by 393 s (about 6.5 min) for the ENMO data, and by 41 s for the MAD data (see Figure 1a and Supplementary Table S1 [available online]). For weekly estimation, as often used in WHO adherence studies, a reduction on average by 5.5, 9.2, 45.8, and 4.8 min was observed when compared to the original estimates. The mean absolute percentage deviation (MAPD) was 4.53%, 4.63%, 19.76%, and 1.56%, respectively.

A hierarchical regression analysis (Figure 1b) further demonstrated the effect of the accelerometer’s calibration error on the observed difference in MVPA minutes. Incremental F tests showed significant influence for all four data types, uniaxial count data, \(F(1) = 49.64, p < .001\); triaxial count data, \(F(1) = 103.43, p < .001\); ENMO data, \(F(1) = 39.73, p < .001\); and MAD data, \(F(1) = 21.21, p < .001\), respectively, indicating that regardless of data type, the difference between uncalibrated and auto-calibrated data increases with the estimated calibration error of the sensor. Bland-Altman plots for the four MVPA estimates can be found in the supplementary material (Supplementary Figure S3 [available online]).

For a comparison beyond MVPA, averaged counts per minute for the uni- and triaxial count data, averaged ENMO and MAD values in mg (1mg = .001g), and averaged monitor-independent movement summary (MIMS) units (John et al., 2019) are shown in Table 1. All metrics’ averages were only marginally reduced by 1%–2% with the exception of the ENMO metric which was almost halved from 19.93 mg to only 10.65 mg.

**Discussion**

This study’s aim was to examine and quantify the effect of sensor calibration on the estimation of physical activity. After calibrating the raw acceleration data from the seventh survey of the Tromsø Study, we found significant reductions of time spent in MVPA across all data types. However, despite being significant, the average daily time spent in MVPA was only reduced by 47 s and 78 s per day for the uni- and triaxial count data, respectively. In weekly statistics, which are often relevant for WHO guidelines’ adherence studies, this still resembles only 5.5 and 9.2 min in total.
Given the substantial amount of noise, for example, by variations in energy expenditure, the clinical relevance of these reductions can be challenged. Simultaneously, the ENMO data type revealed a reduction by over 6 min per day, adding up to almost 46 min per week. Note that this corresponds to 14%–28% of the total volume recommended by the WHO of 150–300 min of MVPA per week (World Health Organization, 2020).

Despite focusing on MVPA, it is worth noting that the average ENMO acceleration was almost halved after auto-calibration (from 19.93 mg to 10.65 mg) indicating that not only the higher levels of physical activity are affected. In contrast, the averages of the other metrics were only barely affected (see Table 1). This also includes the MAD metric in which the average of the acceleration is subtracted instead of 1 g in order to account for the gravity component in the signal. Subtracting the average does not only seem to make MAD more resilient for MVPA estimation but generally more robust for different calibrations. However, as it is well known that the ENMO metric requires accurate calibration, it is doubtful that this robustness provides a significant advantage of the MAD metric. Especially as MAD, at least in this study, seems to overestimate MVPA compared to triaxial count-based data, making it difficult to compare its estimates with previous research. It is worth mentioning that with the Tromsø Study, an adult and elderly (40–84 years) population was investigated. Physical activity research has shown itself to be demographic specific with methods varying significantly between age groups (Migueles et al., 2017). For that reason, the quantity of available thresholds for both count and raw data is not surprising. The thresholds used in this study have been selected based on the age distribution of the sample. The thresholds of Freedson et al. (1998) and Sasaki et al. (2011) are well established in physical activity research and are the established thresholds in the Tromsø Study. The threshold for the ENMO data has been chosen due to age-related considerations as well but has not been used in any other previous study based on data from the Tromsø Study. However, Hildebrand et al. (2014) were among the first to propose ENMO-based MVPA cut points which, since then, are commonly used. Another recent study also proposed a remarkably similar threshold for the same age group (Sanders et al., 2019). However, to which extent our results generalize to other age groups (i.e., children and younger adults) and cut points remains difficult to judge.

An important observation of our study is the huge difference between the substantial reduction for the ENMO data type in comparison to the small reduction for the two count data types. We believe this difference can be explained by the band-pass filtering step in the ActiGraph count algorithm (Neishabouri et al., 2022). In this case, our results are also in line with the results reported by van Hees et al. (2014) where the band-pass filtered BFEN features were comparably less but still significantly affected by the calibration procedure. Given that the MAD metric, compared to the ENMO metric, does not depend on gravity, it is likely to assume that the band-pass filtering step is also able to account for the gravity component in the signal. However, we are not aware how band-pass filtering is solving this problem. The origin of the observed calibration error remains unclear, though. Accelerometers are calibrated during the manufacturing process and should, thus, not show such a substantial variation as observed in the Tromsø Study. The initial suspicion that the accelerometers would lose their accuracy over time could not be verified. We analyzed one data set of a continuous recording with the same sensor over 7 years and could not find any decline in accuracy (see Supplementary Figure S1 [available online]). However, the explanatory power of this study is limited.

**Figure 1** — Auto-calibration reduces the estimated minutes of MVPA on average and as a function of the device’s calibration error. The average number of minutes in MVPA was significantly reduced by 47 s for the uniaxial count data, by 78 s for the triaxial count data, by 393 s (about 6.5 min) for the ENMO data, and by 41 s for the MAD data. The difference between auto-calibrated and factory-calibrated data increased with the estimated calibration error of the device. (a) Average daily MVPA estimates. (b) Difference between auto-calibrated and factory-calibrated data as a function of the device’s calibration error. ENMO = Euclidean Norm Minus One; MAD = Mean Amplitude Deviation; MVPA = moderate to vigorous physical activity. *** indicates a p-value < .001.
given the $n = 1$ sample size. Another source of the observed variation might be temperature as we could observe a tendency toward lower estimated calibration errors during summer (see Supplementary Figure S2 [available online]). However, van Hees et al. (2014) only found a small nonsignificant improvement when utilizing temperature data in their study. A similar replication for our data was not possible due to the lack of temperature data recorded by the sensors.

As the observed tendency is also antiproportional to the measured physical activity (which has been higher during summer compared to winter), another explanation is differences in the utilized data to estimate the calibration error. The auto-calibration method uses all 10-s epochs with a standard deviation of $<13$ mg to estimate the calibration error and the correction coefficients. Thus, enough of such epochs is a requirement for accurate estimation. GGIR provides feedback about whether enough epochs were found to perform auto-calibration, but not about how many epochs were detected. In this study, we only included participants that had enough of these epochs according to GGIR (which were 99.41%), but it is still possible that the quality of the estimates varies significantly between participants and the requirements of GGIR to perform the auto-calibration are too low. On the other hand, the high percentage of participants for which auto-calibration could be performed is also likely to stem from the wear protocol of 7 consecutive days as it seems reasonable that this provides enough nonwear and nonmovement periods to estimate calibration.

### Conclusions

In practice, accelerometers are often not ideally calibrated and tend to overestimate physical activity. However, we found that physical activity is less overestimated when using counts than when analyzing raw data. Whether the overestimation for count data is clinically relevant or not requires specific assessment depending on the objectives of the study concerned. The overestimation using ENMO data derived from the raw signal is so substantial that auto-calibration should always be applied. The overestimation increases with the device’s calibration error. This might be especially influential for small-scale studies as the calibration error seems to vary substantially between sensors. The calibration error also does not appear to increase over time, but a seasonal trend of smaller calibration errors during summer could be observed. However, the summer months are also the months of highest physical activity.

### References


