OPEN

Standardizing Analytic Methods and Reporting in Activity Monitor Validation Studies

GREGORY J. WELK[1], YANG BAI[2], JUNG-MIN LEE[3], JOB GODINO[4], PEDRO F. SAINT-MAURICE[5], and LUCAS CARR[6]

[1]Department of Kinesiology, Iowa State University, Ames, IA; [2]Department of Rehabilitation and Movement Science, University of Vermont, Burlington, VT; [3]College of Physical Education, Kyung Hee University, Yong-in, KOREA; [4]Department of Family Medicine and Public Health, School of Medicine, University of California, San Diego, La Jolla, CA; [5]Metabolic Epidemiology Branch, National Cancer Institute, NIH, HHS, Rockville, MD; and [6]Department of Health and Human Physiology, University of Iowa, Iowa City, IA

ABSTRACT

WELK, G. J., Y. BAI, J.-M. LEE, J. GODINO, P. F. SAINT-MAURICE, and L. CARR. Standardizing Analytic Methods and Reporting in Activity Monitor Validation Studies. *Med. Sci. Sports Exerc.*, Vol. 51, No. 8, pp. 1767–1780, 2019. **Introduction:** A lack of standardization with accelerometry-based monitors has made it hard to advance applications for both research and practice. Resolving these challenges is essential for developing methods for consistent, agnostic reporting of physical activity outcomes from wearable monitors in clinical applications. **Methods:** This article reviewed the literature on the methods used to evaluate the validity of contemporary consumer activity monitors. A rationale for focusing on energy expenditure as a key outcome measure in validation studies was provided followed by a summary of the strengths and limitations of different analytical methods. The primary review included 23 recent validation studies that collectively reported energy expenditure estimates from 58 monitors relative to values from appropriate criterion measures. **Results:** The majority of studies reported weak indicators such as correlation coefficients (87%), but only half (52%) reported the recommended summary statistic of mean absolute percent error needed to evaluate actual individual error. Fewer used appropriate tests of agreement such as equivalence testing (22%). **Conclusions:** The use of inappropriate analytic methods and incomplete reporting of outcomes is a major limitation for systematically advancing research with both research grade and consumer-grade activity monitors. Guidelines are provided to standardize analytic methods and reporting in these types of studies to enhance the utility of the devices for clinical mHealth applications. **Key Words:** ACCELEROMETERS, CALIBRATION, VALIDATION, CRITERION VALIDITY, EQUIVALENCE TESTING

Wearable activity monitors that track user's daily behavior continue to be popular among consumers, and industry experts project continued growth of the consumer sector over time (1,2). One projection suggested that the number of connected wearable devices worldwide would exceed 900 million by the year 2021 (an approximate tripling of the values from 2016) (3). The strongest areas of growth are expected to be in the continued integration with Smartwatch technology, and in mobile health (mHealth) applications (apps) designed for clinical applications (4,5). One estimate suggests that the use of mHealth apps would save the US health care system approximately US $7 billion per year if they were integrated into standard treatments for conditions, such as diabetes, asthma, and cardiac rehabilitation (6). Wearable activity monitors represent only a fraction of a broader spectrum of "wearable health technology" but they are frequently emphasized due to the continued popularity of affordable, commercially available devices (7).

Although the potential is high, many challenges remain in trying to incorporate wearable activity monitors into standard clinical care. The promise and challenges of integrating wearable activity monitors into clinical applications have been fully described in a publication titled "The Wild West: A Framework to Integrate mHealth Software Applications and Wearables to Support Physical Activity Assessment, Counseling and Interventions for Cardiovascular Disease Risk Reduction" (8). The article provides a blueprint of the key steps involved in this integration process. Raw data from wearable monitors must first be uploaded into a robust data management system to enable it to be processed in standardized ways. The complex raw data must then be converted into appropriate summary indicators to facilitate integration of data into electronic medical records (EMR) systems. Once integrated, the data can then be

Address for correspondence: Gregory J. Welk, Ph.D., Department of Kinesiology, Iowa State University, 257 Forker Building, Ames, IA, 50011; E-mail: gwelk@iastate.edu.
Submitted for publication October 2018.
Accepted for publication February 2019.

0195-9131/19/5108-1767/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0000000000001966

used to inform clinical decisions made by the health care team. Although considerable research has been done to develop and evaluate wearable monitors, there still are no viable examples of devices being fully deployed and integrated into EMR systems for clinical applications. The health care industry has already adopted various standards for data sharing (e.g., Health Level Seven) and there are promising data processing tools available to facilitate the integration of data into EMR applications in a standardized way (e.g., Fast Healthcare Interoperability Resources). Thus, the primary stumbling block with the use of wearable monitors in clinical settings is arguably with the lack of standardization in metrics and formats available from the devices.

Addressing these fundamental measurement challenges on the input side of the mHealth framework (8) is essential to ensure accurate outcomes and effective clinical counseling at the output side. The international Exercise is Medicine (EIM) initiative (www.exerciseismedicine.org) led by the American College of Sports Medicine is specifically focused on the incorporation of physical activity (PA) as a "vital sign" in clinical medicine and to facilitate effective referral to community-based programming. Self-report measures have already been successfully incorporated into EMR systems in many medical systems, but this is viewed as a preliminary screening tool (8), due to established limitations of report-based measures (9). Wearable monitors clearly offer potential to overcome the limitations of report-based measures, but their potential has yet to be realized (10). Thus, to capitalize on the promise of mHealth applications, it is essential to develop ways to facilitate the standardized use of wearable devices to assess PA in clinical applications.

Before standardized use can be promoted in practice, it is necessary to first standardize the evaluation of the monitors. This point was specifically emphasized in a recent position statement of the American Heart Association (AHA) designed to facilitate the routine assessment and promotion of PA in health care settings (11). Currently, the market for wearable monitors is driven by market forces, but to harness the potential of monitors in mHealth applications it is essential to develop standards and guidelines to govern potential use in clinical settings. To facilitate broader integration into mHealth applications (and support clinical decisions made by health care teams), the field must eventually move toward agnostic monitoring methods that enable outcomes to be reported in a consistent way regardless of the device being used. Researchers may defend a certain choice of monitor or method for a specific application, but broader integration into health care settings requires standardized (and accurate) reporting of PA metrics by devices and across time (11).

The purpose of this article is to review current analytic methods used to evaluate wearable activity monitors and provide guidelines to standardize methods in future research. Considerable research has already been conducted on the utility of various research-grade monitors, but the variability in the ways in which they are evaluated and compared has made it difficult to compare outcomes. This is further compounded by the inherent complexities of working with (and evaluating) newer lines of wearable activity monitors. Therefore, the overall focus is on the extent to which studies have utilized sound methods and on the provision of guidelines to promote more comparable metrics in the future. The article has four major sections that build toward the specific recommendations: 1) The first section provides a rationale for focusing on PA and energy expenditure (EE), as well as an overview of the evolution of monitoring technology in these devices. 2) The second section summarizes the strengths and limitations of various statistical indicators and analytic methods used to examine validity and measurement agreement to justify the approaches used in the studies we reviewed. 3) The third section describes the results of the review and explains how the choice of indicators and methods can influence outcomes. 4) The fourth section provides specific recommendations and considerations for future evaluation studies to promote standardization and more systematic evaluation and refinement in the future. Although emphasis is placed on the assessment of PA and EE, the concepts can be related to other indicators to promote standardization and harmonization of wearable technology for a variety of applications.

CONSIDERATIONS FOR ASSESSING PA BEHAVIOR

Evaluating compliance with PA guidelines. The primary need for most mHealth applications is a way to systematically evaluate compliance with established PA guidelines. The recent AHA recommendations (11) specifically emphasized the importance of "*...validation efforts tied to guideline-recommended PA metrics for newer technologies*" (p. e12). The current Physical Activity Guidelines for Americans (PAGA) emphasize that volume of PA is most relevant for health benefits and recommend at least 150 min·wk^{-1} of moderate-intensity aerobic activity, or 75 min·wk^{-1} of vigorous-intensity activity, or an equivalent combination of both (12,13). The underlying scientific review by the 2018 Physical Activity Guidelines Review Committee (14) emphasized the strong dose–response relationship that exists between PA and health. They also specifically indicated that greater volumes of moderate and vigorous PA (MVPA) are associated with reduced risk of many chronic diseases, as well as weight gain and dementia. However, although the total volume of activity was emphasized, the report clarified that the relative gain in benefits is larger for those who are currently inactive or not achieving the minimum recommendations. Thus, the scientific report (14) and the PAGA document (12) highlight the importance of the overall volume of PA, as well as the need for individualized prescriptions and goal setting.

A unifying metric for evaluating compliance with PA guidelines is to compute accumulated METs across the week in units of "MET-minutes." The use of MET-minutes enables the volume of PA to be captured without having to categorize activities as minutes of MPA or minutes of VPA. As stated in the PAGA report, "The recommendation that adults do 150 to

300 min of moderate-intensity physical activity or 75 to 150 min of vigorous-intensity physical activity are both equivalent to doing about 500 to 1000 MET·min·wk^{-1}" (p. 109). Earlier statements and guidelines (e.g., Surgeon General's Report; American College of Sports Medicine/Centers for Disease Control and Prevention guidelines) established separate recommendations for moderate PA (MPA) and vigorous PA (VPA). This led to conceptual challenges and confusion because it was possible for people to be active but not meet either the MPA or the VPA target. The current guidelines provide examples of how MPA and VPA can be used to achieve the recommended levels, but the use of MET-minutes is the underlying metric that enables activity to be quantified/equated in a systematic way.

From a measurement perspective, a primary advantage of standardization based on EE is that there are accepted gold standard methods available to quantify the energy costs of activity under both laboratory and field conditions (i.e., direct and indirect calorimetry). Although doubly labeled water is widely endorsed as the criterion measure for EE, the use of indirect calorimetry enables data to be examined by time for more robust analyses. Another major advantage is that EE values can easily be converted to an absolute metric, METs, which can then be partitioned into intensity categories using established thresholds (sedentary, 1.0–1.5 METs, light, 1.5–3.0; moderate, 3.0–6.0; and vigorous, >6.0). Accumulated MET-minutes throughout the day provides a useful summary indicator, but the ability to create daily profiles of activity patterns using the four categories provides considerable utility for education and behavior change applications. A final advantage of MET conversions is that individual PA data can be expressed relative to fitness if estimates of maximal aerobic capacity are available. A moderate-intensity activity such as walking (~3.0 METs) can be perceived as vigorous for a low fit individual but light for a high fit individual. This issue explains some of the discrepancies between monitor-based estimates and report-based estimates of PA (15). Although rarely considered in population-based research, it is an important consideration for screening, counseling, and referral in clinical applications.

A traditional concern expressed about the use of METs is that people may have difficulty understanding their meaning. However, this is not a reason to deemphasize them or to resort to other indicators that are more interpretable. A useful (and parallel) example of a different vital sign that is routinely (and consistently) assessed in clinical settings is blood pressure. The accepted unit of blood pressure (mm Hg) is hardly an interpretable indicator yet the public has come to understand the differences and recommended thresholds for systolic and diastolic blood pressure. Patients are encouraged to have their blood pressure checked regularly and various devices enable users to check blood pressure at home using automated cuffs. For clinical applications (and tracking in EMR systems), medical staff typically rely on measurements obtained by a trained expert using a standardized method. The unifying aspect is the standardized use of millimeters of mercury as the unit to capture and report blood pressure. Although devices may vary in their accuracy, the use of a common indicator makes it possible for clinicians and patients to discuss a common number. Clinicians are then able to make decisions based on where the patient falls in relation to the blood pressure recommendations. To enable and support clinical EIM applications using accelerometry-based activity monitors, it is critical to establish similar standardization for indicators of PA. The standardized use of MET-minutes provides the most defensible option because it is used as the basis for official PA guidelines. Values can still be expressed to the public in terms of minutes of MVPA but the estimates need to be derived from MET-minutes to promote standardization.

The tracking and reporting of steps is also common in devices and apps, but there are limitations for clinical applications because steps only capture locomotor activity (i.e., ambulation). Steps are a metric that can be easily translated to the general audience but it is not clear yet how many steps per day are necessary/recommended (i.e., dose) to improve health (16). The available evidence suggests that it is the accumulated EE associated with steps and other activities that contribute health benefits, and not steps *per se* (13). Because public health guidelines are based on MET-minutes, it is critical to develop methods that can estimate the energy cost of free-living activity with reasonable accuracy. Steps provide advantages for daily monitoring and for behavior change applications but available standards are not sufficient for clinical mHealth applications.

Evolution of monitor technology for assessing PA. The use of accelerometry-based activity monitors for research on PA goes back at least 30 yr, but the use of monitors for consumer and clinical applications is a rather new phenomenon. Fitbit Inc. (San Francisco, CA), one of the most popular activity monitor companies, was launched in 2007 and shipped more than 22 million units in 2016 alone (17). Many factors have contributed to the growth of this technology sector, but one driving force was the inclusion of accelerometer technology in a variety of consumer products (e.g., smartphones, watches) which promoted innovation and competition. With a more competitive price-point, companies were able to develop monitoring devices that enabled consumers to easily track their daily PA behaviors. The ubiquity of smartphones and integrated connectivity through Bluetooth has ushered in many opportunities for interventions and mHealth applications.

Although the use of activity monitors is now common in society, there are challenges that limit effective use in research and clinical applications (18,19). The complexity is compounded by the variability in monitor types and technologies, as well as by monitor placement and sampling methods. Numerous articles have been written on the technical aspects of various monitoring technologies and even more have been published on methods to process and interpret data from these devices (20,21). However, little attention has been given to the specific considerations needed to use monitors in mHealth applications. A previous review of consumer monitors provided direct comparisons of different monitors (22) but it is difficult

to make comparisons between studies due to differences in the methods and monitors being compared. Therefore, the focus in this review is on how decisions about the indicators and analytic methods influence conclusions that are drawn from monitor validation studies. Emphasis is placed on the utility of monitors to assess PA and EE because, as mentioned above, these indicators have the most relevance for assessing compliance with PA guidelines.

Calibration methods for estimating EE. The process of converting raw data into estimates of PA and EE is generally referred to as "calibration." Early research aimed at calibrating accelerometry-based activity monitors relied on linear regression equations to convert counts from accelerometer to EE. The equations were generally developed under laboratory conditions, but the accuracy did not generalize to free-living activities (23). New equations were developed for different ages and populations (24–26), but cutpoints developed from these equations were highly disparate, leading to the widely noted "cutpoint conundrum" (27,28). It became evident that single equations cannot accurately predict EE for different types and intensities of activity—or for detecting sedentary behaviors (SB) (29). Thus, more complex statistical modeling techniques started to be used to differentiate walking and running activities from other lifestyle activities (30,31). To overcome the overestimation of sedentary and light PA EE, inactivity thresholds were also introduced to assign EE values instead of applying regression models for this determination (32). Continued efforts were made to improve the accuracy of EE estimation by using more robust, naturalistic study designs (e.g., including a more diverse range of activities), and by using different machine learning approaches to aid in pattern recognition (33–35). The complex machine learning models provide more flexibility and more fully utilize the rich data from accelerometer outputs, but a detailed summary of these methods is beyond the scope of this review.

Refining calibration methods has been complicated by the multiple locations used by contemporary monitors (e.g., hip, wrist, thigh, ankle) as well as by efforts to combine heart rate with accelerometer data. Extensive lines of research with the multisensor, SenseWear Armband demonstrated improved accuracy with the inclusion of heat-related sensors to supplement the primary accelerometer signal (36,37). The company that developed the SenseWear (BodyMedia) was bought by Jawbone (Jawbone, San Francisco, CA) to presumably be used in other devices but different technologies and analytic methods have evolved over time. Contemporary wearable activity monitors have clearly capitalized on insights from early research, but the proprietary procedures used by each company are typically not disclosed. Most devices likely use triaxial accelerometers and probably use robust pattern recognition procedures to differentiate type and intensity of activities. Many wrist-worn monitors now incorporate photoplethysmography, which enables the detection of heart rate, but it is not clear how the data is used to improve estimation of EE. Research with the Actiheart monitor documented advantages of incorporating heart rate data with accelerometer data (38,39), but the advantages of consumer monitors that integrate heart rate data are less clear. A direct comparison of two heart rate enabled consumer monitors by Bai et al. (40) demonstrated stronger accuracy for the Apple Watch (Apple, Cupertino, CA) compared to the Fitbit Charge HR (Fibit Inc. San Francisco, CA). However, the precision of estimates from the newer device was actually worse than an earlier version of the Fitbit tested under relatively similar conditions (41). A challenge in interpreting any of the newer monitors is the lack of transparency about the specific algorithms or methods. Thus, it cannot even be assumed that the heart rate data are really used within the prediction algorithms or whether error is reduced. A recent study by Montoye et al. (42), for example, indicated that Fitbit products measuring heart rate yielded different estimates of EE compared to Fitbit products without heart rate. Without standardization it is difficult to interpret the disparate findings. Although companies have different methods and approaches for estimating PA and EE, the accuracy and utility of their estimates should be directly comparable if common metrics and indicators are used in evaluation. The next section will review the strengths and limitations of different approaches to aid in output standardization.

OVERVIEW OF MEASUREMENT AGREEMENT METHODS AND INDICATORS

Measurement error principles. A factor contributing confusion in the literature is the lack of attention given to fundamental measurement principles and terminology. For example, the ubiquity of monitoring technology in research has led many to erroneously imply that monitors are "measuring" PA. In actuality, traditional accelerometry-based activity monitors are *measuring* acceleration (i.e., change in the speed of movement over time) and this raw data is used to *estimate* EE or PA (15). Pedometers can count steps and heart rate monitors can directly sense heart beat; however PA cannot be directly measured with accelerometry-based devices but is instead inferred or estimated. Thus, although it is true that these devices are used as activity "measures" (i.e., measure used as a noun), they are not technically measuring PA (i.e., measure used as a verb). The lack of attention given to this fundamental issue has hampered research because it has deemphasized the significant measurement error associated with current methods and estimates generated from these devices.

Despite the popularity of accelerometry-based monitors, these measures are indeed affected by both systematic (i.e., bias) and random error (i.e., within-person random error). The systematic error reflects systematic differences between a criterion measure (e.g., double labeled water) and the alternative measure being tested (e.g., accelerometer). Random error results from the variability associated with repeated assessments (i.e., within-person assessments). It influences the reliability or precision of a measure, but studies have demonstrated that most of the within-person variation in PA is attributable to individual factors (i.e., day-to-day/season variability), and not the technical features built into these devices (43). Systematic error relates

closely to validity or accuracy of a measure and this measurement property has been more extensively studied with activity-based monitors (18,19,43). Although random error has received less attention in the literature we will focus our discussion on systematic error/validity of activity monitor devices.

Although many studies have reported on the validity of prediction equations and algorithms for assessing EE and PA, it has proven difficult to compare the accuracy of the outcomes (44). This has been attributed to choice of the criterion method, the research study design, the sample population, the nature/diversity of the activities evaluated, and the way data are analyzed/reported (45). The features of the calibration protocol also have a direct influence on the nature of the error and how they perform in free-living studies. For example, calibration studies of accelerometers using purely ambulatory activities are known to produce substantially lower estimates of MVPA when compared with accelerometers whose protocols included a more diverse set of activities (e.g., mixture of free-living activities) (46). These challenges have direct implications for the calibration/validation of wearable devices.

Researchers have previously called for more standardization in methods (47,48) and have started to embrace the use of open-source methods; however, little information is available about the specific protocols/methods used to calibrate/validate newer lines of wearable activity monitors. The need to protect intellectual property is understandable, but the lack of transparency has made it difficult to systematically advance the science of monitor calibration while understanding the implications for validity as described above. Regardless of the methods used, it is important to emphasize that the process of calibration is quite complex and that there is considerable measurement error with all current methods.

Indicators and analytic techniques for evaluating agreement. Many articles erroneously conclude that a device or method has been "validated" based on a single statistical test or by the use in a published article. However, systematically evaluating validity and agreement requires a more comprehensive analytic approach (49–51). It is also important to emphasize that a method can be "accurate" for group level estimations, but lack precision needed for individual level estimations. Because mHealth applications are almost always aimed at individual applications, the degree of individual error is a particularly important consideration. Thus, it is also important to make clear distinctions between individual and group level error when interpreting data from accelerometry-based monitors (52). A brief summary of mathematical and statistical metrics is provided below to set the stage for the review.

The most commonly reported indicator is mean error (ME) or mean bias, which is calculated by averaging the difference between the criterion and the estimate (i.e., $EE_{Criterion} - EE_{monitors}$). The indicator of mean percentage error (MPE) standardizes the error by expressing the error as a percentage deviation from the criterion (i.e., average EE difference from two measures divided by $EE_{Criterion}$). These two validation indicators reflect group-level agreement. In contrast, mean absolute percentage error (MAPE) uses the absolute value of the EE difference before dividing by the $EE_{Criterion}$. Root mean square error (RMSE), which was originally used to quantify the differences between sample and population values predicted by a model, is also frequently used in validation studies. Both MAPE and RMSE provide indicators of individual agreement.

Mean absolute percentage error provides a particularly useful comparison of individual agreement because it accounts for each individual participant's error while avoiding cancellation of errors from underestimation and overestimation. For instance, if error is 15% for one participant (i.e., overestimation) and −13% for another participant, the MPE would be 1%, whereas the MAPE would be 14%. The standardized use of MAPE is strongly recommended because it reflects the error expected at the individual level and because it can be directly compared across studies and monitors, regardless of the magnitude of the values. Absolute error (i.e., ME) will be larger for studies evaluating higher intensity activities or those conducted for longer periods of time; however, by expressing error in relative terms (using MAPE), indicators can be directly compared, regardless of the protocol used.

Although MAPE can reflect the magnitude of error it is still critical to quantify the overall direction (i.e., overestimation or underestimation). Many PA validation studies have used the Bland–Altman method (53) for this purpose, but there are many misconceptions with the application and interpretation of this method (54). Although Bland–Altman plots have clear utility, a limitation is that they do not provide a way to empirically (or statistically) evaluate agreement. The information helps understand the nature and source of the error, but it proves challenging to draw definitive conclusions based on the distributions in these plots.

Standard statistical tests such as *t*-tests and ANOVA are also frequently used to compare two measures and assess group agreement. However, these tests are designed to test for differences rather than agreement. The failure to reject the null hypotheses of "no difference" simply cannot be used to infer agreement or equivalence. Moreover, because the significance of these tests is directly influenced by sample size, studies with larger and more robust samples are inherently more likely to detect statistically significant differences than studies with fewer participants—regardless of the size of the difference. Thus, these tests are powered in the wrong direction to test agreement. Several articles have explicitly called for the use of "equivalence testing" for studies evaluating agreement (55,56). The approach essentially flips the null hypothesis and this enables zones of equivalence to be established *a priori* and to be tested statistically. Readers interested in a more comprehensive explanation of equivalence testing and its applications are referred to the article by Dixon et al. (55).

A step beyond simply visualizing the error is to move toward the use of measurement error models that can help to adjust for known bias and error. These methods are widely used in the nutrition literature to address errors in nutrition assessments and to a less extent used in PA research (57–59). They are more commonly used to address error in report-based

measures but the same applications can be used to model and control for error in monitor-based methods. A detailed review of measurement error methods is beyond the scope of this review but it is important to acknowledge the potential to statistically adjust for measurement error as opposed to evaluating error.

The main point of this section is to document the various analytic techniques commonly used to evaluate agreement. The subsequent review of validation studies provides a summary of measurement properties associated with various devices using the statistical indicators just described.

REVIEW OF ANALYTIC METHODS IN MONITOR VALIDATION STUDIES

A review of monitor validation studies was conducted to systematically evaluate the impact of analytic methods on outcomes from validation studies. Because the focus of this article is on clinically related mHealth applications, the review was restricted to wearable activity monitors that are currently being explored for mHealth integration. The distinction is blurry but Bluetooth capability and links with smartphone technology distinguish this type of monitor from traditional research grade devices. Emphasis in the review was also placed on the evaluation of EE (as opposed to steps or other indicators) because the focus is on evaluating compliance with PA guidelines. To enable direct evaluation of the analytic methods it was necessary to further restrict the review to studies that used a viable criterion measure of EE (i.e., doubly labeled water or direct/indirect calorimetry).

The PubMed Search criteria (ran through May 31, 2018) used the following keyword combinations: 1. (Consumer monitors) AND (validity OR validation OR accuracy) and 2. (PA monitors OR trackers) AND (validity OR validation OR accuracy). Only articles evaluated EE as outcome and used indirect or direct calorimetry or doubly labeled water as criterion were included. Abstracts and conference proceedings were excluded. The literature search identified 23 articles published between 2013 and 2018 that met the criteria (40,41,60–80). Table 1 summarized key characteristics of the studies including sample size, criterion measure, monitors evaluated, settings, and protocol design. The sample sizes ranged from 13 to 60 with more than half of the studies (12 out of 23) reporting sample sizes ranging from 20 to 30. Most of the studies (21 of 23) used indirect calorimetry as the criterion measure, whereas two studies used a metabolic chamber and/or doubly labeled water. A total of 58 different monitors were evaluated including several lines of Fitbit products (Zip, One, Ultra, Flex, Charge, Charge HR, Charge HR 2, Blaze, and Surge), Apple Watch series 1 and 2, Garmin products (Vivofit, Vivosmart, Vivosmart HR, Vivoactive, Forerunner), Polar products (H7 and A360), and Jawbone products (Up, Up24, and Up3). All monitors are listed in Table 1.

With regard to study design, three studies evaluated EE under free-living settings (60,69,73). Among the 20 studies conducted in controlled laboratory settings, eight utilized a semistructured design allowing participants to self-select the mode and/or intensity of activities (40,41,61,68,71,76–78). Five of the remaining 12 studies included structured activities beyond aerobic exercise such as lifestyle activities, resistance exercise, or sedentary activities (62,63,66,73,79).

Table 2 summarizes the various analytic methods and indicators used in the identified studies. The table reveals considerable variability with regard to the reporting of different indicators of agreement with the ME being the most commonly reported indicator (65%) followed by MAPE (52%), RMSE (26%), and MPE (22%). Each indicator provides different information and the lack of comprehensive reporting makes it difficult to compare outcomes across studies. It is encouraging that reporting of MAPE is increasingly common, but reporting of MPE is also important to document the direction of error and the magnitude of group-level error in a standardized way.

The inferential indicators in the reviewed studies include correlations, Bland–Altman plots as well as indicators of statistical differences (e.g., *t*-test or ANOVA) or equivalence (e.g., equivalence testing). The most commonly used statistical metrics to assess the validity are Pearson product-moment correlation coefficient (or Spearman Rank correlation as nonparametric measure) (20 of 23 studies or 87%) and paired *t*-test or ANOVA (or other nonparametric test for mean difference) (17 of 23 studies or 74%). Bland–Altman plots along with 95% limits of agreement were used by 16 studies (70%), whereas equivalence testing was used in only five studies (22%). Correlations clearly have limitations for evaluating agreement because two indicators can be associated with each other but yield very different estimates.

The tests for both difference tests and equivalence tests are dependent on sample sizes and arbitrary thresholds of significance so the most defensible indicator for ongoing comparison of monitor precision is the MAPE. Table 3 provides detailed results of the 10 studies that used indirect calorimetry as the criterion and reported MAPE values as part of the evaluation (40,41,62,65,66,70,71,73,75–79). Three studies that met the criterion were not included in the table because they reported MAPE from more than two activities or intensities without an overall summary (70,71,75). The MAPE values ranged from 9% (62) to 64% (79), but it is important to point out that MAPE values can be influenced by the monitors as well as by the nature of the research protocol used in the study. Studies that included an array of activities (including exercise and lifestyle activities) had MAPE values ranging from 9% to 64% with a median of 18% (40,41,62,65,66,77). Studies that only examined aerobic activities (or evaluated aerobic activities separately) had MAPE values ranging from 7% to 67% with a median value of 33% (40,41,66,76,78). The exception is one study that specifically examined the accuracy during cycling which revealed much higher errors (21% to 75%) based on seven monitors (79). Two studies examined resistance activities and the MAPE values ranged from 29% to 57% (41,79). These results demonstrate a wide variability in MAPE values but also document the inherent value in being able to directly compare precision of each monitor with a single unifying metric.

TABLE 1. Summary of the characteristics of reviewed studies.

First Author Name (Year)	Sample Size	Criterion	Monitors	Setting	Activity
Dannecker (2013)	19	Room Calorimeter	Footwear-based monitor, Actical, ActiGraph, IDEEA, DirectLife, and Fitbit Classic	4-h stay	Walking, cycling, stepping
Noah (2013)	23	Indirect Calorimetry (COSMED)	Fitbit Classic and Fitbit Ultra	6-min bouts of treadmill walking, jogging and stair stepping	3.5 mph (5.63 km·h^{-1}) at 0% incline (walk), 3.5 mph (5.63 km·h^{-1}) at 5% incline (walk-incline) and 5.5 mph (8.85 km·h^{-1}) at 0% incline (jog).
Gusmer (2014)	21	Indirect Calorimetry (CPX Ultima)	FitBit Ultra and ActiGraph	Two 30-min phases of walking (slow and brisk) on a treadmill	
Lee (2014)	60	Indirect Calorimetry (Oxycon)	BodyMedia FIT, Fitbit Zip, Fitbit One, Jawbone Up, ActiGraph, DirectLife, NikeFuel Band, and Basis B1 Band	Eight different types of activity monitors simultaneously while completing a 69-min protocol	Resting EE for 15 min, 1) sedentary (reclining, writing at a computer), 2) walking (treadmill walking at 2.5 mph, treadmill brisk walking at 3.5 mph, self-paced overground walking, and self-paced overground walking with 15 kg backpack), 3) running (treadmill jogging at 5.5 mph, treadmill running at 6.5 mph), and 4) moderate-to-vigorous activities (ascending and descending stairs, stationary bike, elliptical exercise, Wii tennis play, and playing basketball with researchers).
Sasaki (2015)	20	Indirect Calorimetry (Oxycon)	Fitbit Classic	6 min activity and 4 min resting	Walking at 3.0 mph and 4.0 mph at 5% grade, and jogging at 5.5 mph at 0% grade. Office work, driving, carrying a box, carrying groceries, and ascend/descend stairs; 2) cycling at 300 kg·min^{-1}, golf, tennis, and basketball; or 3) dusting, laundry, vacuuming, raking, and gardening
Diaz (2015)	23	Indirect Calorimetry (CPX Ultima)	Three hip-based Fitbit One (two on the right, one on the left hip) and two wrist-based Fitbit Flex (one on the right and left wrists)	24 min. Each stage was 6 min in duration.	A four-stage treadmill exercise protocol consisting of walking at slow (1.9 mph), moderate (3.0 mph), and brisk (4.0 mph) paces; and jogging (5.2 mph).
Tucker (2015)	24	Indirect Calorimetry (Oxycon)	Nike Fuelband and Armband	Two, 60-min semistructured routines	12 activities selected from compendium of physical activities
Bai (2015)	52	Indirect Calorimetry (Oxycon)	Fitbit Flex, Jawbone Up24, Misfit Shine, Nike Fuelband SE, ActiGraph GT3X+ and BodyMedia Core	20 min of self-selected sedentary activity, 25 min of aerobic exercise, and 25 min of resistance exercise, with 5 min of rest between each activity.	
Nelson (2016)	30	Indirect Calorimetry (COSMED)	Fitbits One, Zip, and Flex and Jawbone UP24	Lying on a bed for 10 min; 10 other activities were performed for 5 min each	Three sedentary, four household, and four ambulatory/exercise) chosen by researchers from a list of 21 activities
Wallen (2016)	22	Indirect Calorimetry (MetaMax 3B)	Apple Watch, Fitbit Charge HR, Samsung, Gear S and Mio Alpha	58 min protocol	Activities at rest (lying, sitting, standing) and exercise (walking, cycling)
Alsubheen (2016)	13	Indirect Calorimetry (CA-10 Carbon Dioxide and PA-10 Oxygen Analysers)	Garmin Vivofit	1 h of self-selected activity each day for 2 d	Session 1: three 10 min walking conditions at a self-selected pace (ranging from 4.0 to 7.2 km·h^{-1}), with inclines of 0, 5 and 10%. Session 2: 1 h of OTP including computer work, reading articles and writing
Murakami (2016)	19	Metabolic chamber and Doubly labeled water	Jawbone UP24, Fitbit Flex, Misfit Shine, Epson Pulsence Pulsence PS-100, Garmin Vivofit, TANITA AM-160, OMRON CaloriScanHJA-403C, and Withings PulseO2, OMRON Active style Pro HJA-350IT, Panasonic Actimarker EW4800, SUZUKEN Lifecorder EX, and ActiGraph GT3X	24-h living in metabolic chamber and 15 d of free living.	24-h indirect calorimetry: under a standardized protocol simulating normal daily life, which included three meals, deskwork, watching TV, housework, treadmill walking, and sleeping. DLW: 15 free-living days.
Dooley (2017)	62	Indirect Calorimetry (Parvo Medics)	Apple Watch, Fitbit Charge HR, and Garmin Forerunner 225	Laboratory conditions	A 10-min seated baseline assessment; separate 4-min stages of light-, moderate-, and vigorous-intensity treadmill exercises; and a 10-min seated recovery period.
Imboden (2017)	30	Indirect Calorimetry (COSMED)	Fitbit One, Fitbit Zip, Fitbit Flex, Jawbone UP24	80 min protocol	Performing ≥12 activities from a list of 21 choices; 1) sedentary activities, 2) household activities, and 3) ambulatory and cycling activities

(continued next page)

TABLE 1. *(Continued)*

First Author Name (Year)	Sample Size	Criterion	Monitors	Setting	Activity
Shcherbina (2017)	60	Indirect Calorimetry (COSMED)	Apple Watch, Basis Peak, Fitbit Surge, Microsoft Band, Mio Alpha 2, PulseOn, and Samsung Gear S2	5 min of activity	Faster walking (4.0 mph at 0.5% incline), slow running (average speed 5.7 mph at 0.5% incline, range 4.5–6.5 mph), faster running (average speed 6.9 mph at 0.5% incline, range 4.8–9.0 mph), low intensity cycling (average work rate 88 W, range 50–100 W), intense cycling (average work rate 160 W, range 80–225 W)
Chowdhury (2017)	30	Indirect Calorimetry (COSMED) and Bodymedia armband	Microsoft Band, Apple Watch and Fitbit Charge HR, Jawbone UP24, BodyMedia Core and individually calibrated Actiheart	Both controlled laboratory conditions (simulated activities of daily living and structured exercise) and over a 24-h period in free-living conditions	24-min protocol comprising of four activities of 5 min duration (seated and typing, loading dishes, sweeping, stairs). four exercise of 10 min (treadmill, walking, cycling, jogging)
Price (2017)	14	Indirect Calorimetry (Parvo Medics)	Fitbit One, Garmin Vivofit, and Jawbone UP	Laboratory conditions	Walked at 0.70, 1.25, 1.80 $m \cdot s^{-1}$ and ran at 2.22, 2.78, 3.33 $m \cdot s^{-1}$ on a treadmill
Wahl (2017)	20	Indirect Calorimetry (MetaMax 3B)	Bodymedia Sensewear, Beurer AS 80, Polar Loop, Garmin Vivofit, Garmin Vivosmart, Garmin Vivoactive, Garmin Forerunner 920XT, Fitbit Charge, Fitbit Charge HR, Xaomi MiBand, Withings Pulse Ox	Four 5 min stages of different constant velocities (4.3; 7.2; 10.1; 13.0 $km \cdot h^{-1}$), a 5-min period of intermittent velocity, and a 2.4-km outdoor run (10.1 $km \cdot h^{-1}$)	
Parak (2017)	24	Indirect Calorimetry (Metalyzer 3B)	PulseOn	The subjects were instructed to run at a self-determined pace for at least 20 min, targeting moderate to vigorous subjectively assessed intensity, and to run 5 km.	
Woodman (2017)	28	Indirect Calorimetry (Oxycon)	Basis Peak and Garmin Vivofit and three Withings Pulse	Supine for 10 and 5 min of the other 10 activities with 2 min transition.	1. Supine rest 2. Computer 3. Folding clothes in a seated position 4. Sweeping a floor 5. Treadmill walking at 80.5 $m \cdot min^{-1}$ and 7% incline 6. Ascending and descending stairs 7. Walking at a self-selected pace 8. Running at a self-selected pace 9. Seated rest 10. Cycling at a self-selected pace 11. Cycling on ergometer at 100 W
Xie (2018)	44	Indirect Calorimetry (COSMED)	Apple Watch 2, Samsung Gear S3, Jawbone Up3, Fitbit Surge, Huawei Talk Band B3, and Xiaomi Mi Band 2	Walking: walk on a 400-m track for two laps; Running: run on a 400-m; Cycling: ride three predetermined trips	
Boudreaux (2018)	50	Indirect Calorimetry (TrueOne-2400)	Apple Watch Series 2, Fitbit Blaze, Fitbit Charge 2, Polar H7, Polar A360, Garmin Vivosmart HR, TomTom Touch, and Bose SoundSport Pulse (BSP) headphones	Separate trials of graded cycling and three sets of four resistance exercises at a 10 repetition-maximum load	Cycling: 5-min rest period, 2-min stages at 50 rpm, beginning at 300 $kpm \cdot min^{-1}$ and increasing by 150 $kpm \cdot min^{-1}$ until exhaustion, followed by a 5-min cool down. Resistance exercise: two upper body exercises (chest press, latissimus dorsi (lat) pulldown) and two lower body exercises (leg extension and leg curl).
Bai (2018)	39	Indirect Calorimetry (Oxycon)	Apple Watch 1 and Fitbit Charge HR	20 min of sedentary activity, 25 min of aerobic exercise, and 25 min of light intensity PA	

We also calculated "percent error" using mean difference ($EE_{criterion} - EE_{monitor}$) divided by mean $EE_{monitor}$ from six studies that reported mean difference and MAPE (40,41,62,65,77,79). This information was added in Table 3 to compare with MAPE to document the point that a smaller mean difference at the group level does not imply small individual level error as evaluated by MAPE. As shown in Table 3, MAPE is consistently greater than percent error with values as large as 10 times the magnitude of percent error. Monitors with similar MAPE can also have very different percent errors. For instance, in the Lee et al. (62) study, both Fitbit Zip and Fitbit One had around 10% MAPE but percent error for Fitbit Zip was −3.7%, whereas Fitbit One had a value of 7.3%. Thus, smaller mean difference is not necessarily related to high validity because of measurement error cancellation from combining underestimation and overestimation.

Unfortunately, only one article reported both MAPE and MPE. As we mentioned in the earlier section, it is worth reporting both MAPE and MPE because they provide validity information from different perspectives. MAPE is usually greater than MPE except for cases where monitors consistently underestimate or overestimate EE across all the participants. In the Bai et al. article, both Apple Watch and Fitbit Charge HR had larger MAPE compared with MPE that Apple Watch had 15.2% MAPE and 7.6% MPE, whereas Fitbit Charge HR had MAPE of 32.9% and MPE of −29.6% for estimating overall

TABLE 2. Summary of statistics used in EE validation study.

First Author Name (Year)	Mean ± SD/SE	ME (ME)	Mean Percent Error (MPE)	Mean Absolute Percent Error (MAPE)	Root Mean Squared Error (RMSE)	Limit of Agreement	Correlation	*t*-Test or ANOVA	Bland–Altman Plots	Equivalence Testing	Other
Dannecker (2013)	√				√			√			Regression
Noah (2013)	√						√	√			
Gusmer (2014)	√						√	√	√		
Lee (2014)	√	√		√	√	√	√	√	√	√	Bland–Altman intercept and slope
Sasaki (2015)	√	√	√			√	√	√	√		
Diaz (2015)	√	√					√				
Tucker (2015)	√	√		√			√	√	√	√	
Bai (2015)	√	√		√		√	√		√	√	95% CI, effect size
Nelson (2016)	√			√	√						MAE, 95% CI, Friedman, Dunn's test
Wallen (2016)	√	√				√	√		√		
Alsubheen (2016)	√							√			Regression
Murakami (2016)	√	√					√	√			
Dooley (2017)	√			√			√	√	√		Cohen's *d*
Imboden (2017)	√	√	√	√		√	√	√	√		
Shcherbina (2017)			√		√		√	√	√		PCA, GEE, Regression, Median error
Chowdhury (2017)	√	√			√	√	√	√	√	√	MSE, MAE
Price (2017)	√	√	√			√	√	√	√		Regression
Wahl (2017)	√	√		√		√	√				95% CI, typical error
Parak (2017)		√		√		√	√	√	√		MAE
Woodman (2017)	√	√		√			√	√			ICC, 95% CI
Xie (2018)				√			√	√			
Boudreaux (2018)	√	√		√		√	√	√			Mean difference, 95% CI
Bai (2018)	√	√	√	√	√	√	√		√	√	

GLMN, generalized linear mixed modeling; GEE, general estimating equation; PCA, principal component analysis; MAE, mean absolute error; 95% prediction interval; LOA, 95% limits of agreement; ICC, intraclass correlation coefficient.

EE (40). The monitors underestimated EE among some participants and overestimated EE among other participants. Overall, Apple Watch underestimated EE (7.6% of MPE) and Fitbit Charge HR overestimated EE (−29.6% of MPE).

RECOMMENDATIONS FOR STANDARDIZATION AND DATA HARMONIZATION

Considerable attention has been placed on the potential of various accelerometry-based activity monitors designed to estimate PA and EE in free-living conditions. This feature has led to growing interest in various mHealth applications aimed at clinical evaluations and behavior change, but the consensus in the literature is that the current generation of monitors are not yet sufficiently accurate (40,79,81,82). The devices clearly provide value to consumers, but greater accuracy is needed for screening, tracking, and clinical decision-making. Researchers have continued to explore ways to improve the accuracy of these estimates and most monitor companies are likely conducting parallel research. The key to advancing work in this area is to promote standardization in the ways that monitors are evaluated and how results are reported. This section will summarize guidelines for standardization and strategies for data harmonization.

Guidelines for standardization. The review provides documentation of the advantages and disadvantages of different metrics/indicators and directly documents the impact that evaluation methods have on interpretations that can be drawn from different studies. The use of consistent and comprehensive reporting of data would facilitate comparisons and enable new technologies to be directly compared to existing methods. If new monitors and processing algorithms have lower MAPE and better overall measurement properties than previous devices evaluated in similar ways the advantages can be directly quantified. Key design elements for validation studies adapted from previous guidelines (47,48) are summarized below:

- Diverse sample: validation studies should include a heterogeneous sample to test monitors across different individual characteristics, including: sex, Age, and body size. Individual characteristics might not be as critical when determining group-level agreement; however, variability in accuracy across individuals is very relevant when considering the applications of devices for individual assessments/feedback.
- Appropriate sampling of daily behavior: monitors should be evaluated under free-living conditions or with simulated conditions that capture activities of daily living in a natural way. Validation studies relying only on ambulatory activities cannot provide an accurate validation assessment of the monitor for other activities of daily living that are likely to occur throughout the day (e.g., upper-body activities).
- Selection of Criterion Measure: An appropriate criterion measure is needed to enable error to be directly evaluated. The use of portable indirect calorimetry devices is recommended because they can be used in simulated free-living designs and enable error to be evaluated for discrete periods of time. The use of doubly labeled water and direct observation also have value as criterion measures but each has inherent limitations.
- Standardized protocols and wear locations: wearable activity monitors are designed to be worn in specific locations and in certain ways. Therefore, it is important to standardize locations/orientations as well as fit. It may seem convenient to place multiple monitors on the same wrist for

TABLE 3. Percent error and mean absolute percentage errors.

	Lee (2014)		Tucker (2015)		Bai (2015)		Nelson (2016)		Chowdhury (2017)		Parak (2017)		Woodman (2017)		Boudreaux (2018)		Bai (2018)		Xie (2018)	
Brand Name/ Author (Year)	**MAPE**	**Percent Error**	**MAPE**	**Percent Error**	**MAPE**	**Percent Error**	**MAPE**	**Percent Error**	**MAPE**	**Percent Error**	**MAPE**	**Percent Error**	**MAPE**	**Percent Error**	**MAPE**	**Percent Error**	**MAPE**	**Percent Error**	**MAPE**	**Percent Error**
BodyMedia Mini	9%	5%	18%	2%						NA		NA								NA
BodyMedia Core					15%	−11%			33%											
Fitbit Zip	10%	−4%					36%													
Fitbit One	10%	7%					24%													
Fitbit Flex					17%	−6%	29%													
Fitbit Charge HR									36%								33%	−22%		
Fitbit Charge HR 2															75% (C); 48% (R)	42% (C); 8% (R)				
Fitbit Surge																			67%	
Fitbit Blaze															72% (C); 49% (R)	39% (C); 5% (R)				
Apple Watch 1																	15%	10%		
Apple Watch 2															21% (C); 43% (R)	−100% (C); −87% (R)			49%	
Garmin VivoFit													45%	45%						
Garmin Vivosmart HR															63% (C); 57% (R)	−8% (C); −31% (R)				
Polar H7															30% (C); 43% (R)	−51% (C); −76% (R)				
Polar A360															38% (C); 53% (R)	−29% (C); −50% (R)				
Jawbone UP	12%	7%																		
Jawbone UP24					18%	8%	33%		36%											
Jawbone UP3																			28%	
NikeFuel Band	13%	2%	16%	−1%																
NikeFuel Band SE					17%	15%														
Basis Band	24%	24%																		
Basis Peak													27%	−7%						
Withings Pulse Ox Hip													40%	42%						
Withings Pulse Ox Wrist													64%	64%						
Withings Pulse shirt collar													41%	42%						
Pulse One											17% (L) 7% (H)									
Actiheart									20%											
Microsoft Band									40%											
TomTom Touch															41% (C); 52% (R)	−49% (C); −33% (R)				
Misfit Shine					30%	−25%														
Actigraph GT3X+	13%	9%			17%	3%														
DirectLife	13%	10%																		
Huawei Talk Band B3																			32%	
Samsung Gear S3																			38%	
Ledongli																			39%	
Xiaomi Mi Band 2																			40%	
Dongdong																			48%	

MAPE, mean absolute percentage error; C, cycling; R, resistance exercise; L, light intensity; H, high-intensity.

comparison but the altered locations may lead to systematic differences if it deviates from manufacturer's recommendations. For example, wear location for multisensor devices that detect heart rate at the wrist must be positioned in specific ways to provide an accurate signal.
- Inclusion of reference monitors/metrics: studies should consider including more established monitors and methods when evaluating new monitors/methods. Reference monitors cannot be considered true criterion measures in these studies, but using various devices simultaneously provides a more robust way to understand differences in validity and to justify the relative differences between estimates.

The combination of the recommendations listed above requires both financial and logistic efforts, but consideration of these features will provide for a more comprehensive design. The unique aspect of the present review is that it directly documented the importance of using standardized statistical tests and comprehensive reporting of indicators. Thus, there is a similar need for standardization in analytic methods. A recent article by our team was developed as a guide to facilitate more standardized analyses and reporting of monitor validation studies (81). Readers are also encouraged to review the technical aspects of equivalence testing described previously (55) because this approach is central to the recommendations. The essential features of a comprehensive evaluation are summarized below:

- Reporting relevant metrics: comprehensive reporting of all relevant metrics is important to enable comparisons across monitors and studies. This is particularly important for studies reporting EE indicators because they can be reported in both absolute and relative terms. Absolute metrics of mean difference (i.e., ME) have been widely used in the past validity studies, but emphasis should be placed on metrics that account for duration or intensity of the testing protocol. In some cases, the errors expressed at minute or day level are acceptable, which also provide possible ways to make comparisons between studies.
- Documenting error: the direction of error for group-level estimation should be reported using MPE and the magnitude of error for individual level estimation should be reported using MAPE. Bland–Altman plots can be used to complement these analyses and to determine if the devices provide similar accuracy across a range of activity levels. Error also needs to be reported for both individual activities and aggregated set of free-living activities. Reporting the estimated error for both will help understand what activities add the most error and how these are likely to impact estimation of total activity or EE.
- Focusing on equivalence: the use of equivalence testing is strongly recommended to provide a way to directly evaluate measurement agreement. Specific criteria needed to document statistical "equivalence" would require consensus but guidelines for chemical bioequivalence are typically risk-based (83). Because the risk of misclassification of PA is low, a relatively liberal criteria of 15% to 20% error could be a viable target for equivalence (55). However, an advantage of equivalence testing is that it is possible to calculate the actual bounds within which the monitors are statistically equivalent to the criterion. This would make it possible to document, for example, that measurement error is less than 12% (i.e., between 88% and 112% of reference measures (i.e., 88% to 112%). The implications of error vary depending on the application (e.g., counseling, surveillance vs test of associations) so readers are referred to several primers on equivalence testing (55,56) for broader discussion on these issues.

CONSIDERATIONS FOR AGNOSTIC REPORTING OF MONITOR DATA

A documented barrier in mHealth applications is the variability in ways that data are summarized and reported by companies and the lack of resolution in data exports. A position article on the complexities of using monitors in mHealth applications summarized the major challenges in incorporating data from monitors into EMR systems (8). A key to this process is to enable systems to import standardized outcomes and common units. The standardized reporting of MET-minutes and evaluation of compliance with PA guidelines provides one form of standardization but other features are needed to achieve the agnostic monitoring of PA data for mHealth applications. Companies should clearly have freedom to innovate, but some standardization is needed to ensure that data from different devices can populate the same indicators in EMR

- Reporting wear time: monitor companies should enable reporting of wear time to ensure that collected data are representative of a day and that sufficient data are collected to generalize to a typical week.
- High-resolution outputs: monitor companies should provide high resolution export capabilities to enable exporting of data on a minute-by-minute basis. This is critical for coding minutes into the four established intensities (sedentary, light, moderate, and vigorous) and for more robust visualizations of PA profiles. Accumulated steps over a whole day or other summary indicators can have utility for consumers but reporting of data on a minute-by-minute basis is needed to convert outcomes into MET-minutes so that compliance with PA guidelines can be evaluated.
- Open API: monitor companies should provide "open API" to enable integration into data aggregation utilities and tools to facilitate standardized processing. Without methods to directly extract the raw data, it will not be possible to achieve the goals of more agnostic reporting of PA data from monitors.

Of course, for any of this to occur, it is necessary to identify effective incentives that encourage monitor companies to adhere to these guidelines. A potential incentive could be some type of certification or documentation that a monitor has met the basic criteria needed for inclusion in mHealth applications.

There are examples of organizations that document standards for electronic equipment and for food so there could be opportunities to establish minimal standards needed for inclusion in clinical applications such as those envisioned by the EIM initiative (8). Guidelines should be established to ensure that indicators correspond with established Logical Observation Identifiers Names and Codes indicators used in electronic medical systems. Standardized data formats, such as Health Level Seven International, are already used by most health care systems, and tools such as Regenstrief Logical Observation Identifiers Names and Codes Mapping Assistant and Fast Healthcare Interoperability Resources enable efficient integration. The lucrative and strategic opportunities for integrating devices into mHealth applications will continue to drive innovation, but it is important for professional organizations and researchers to help ensure that the monitors used in these applications have sufficient validity. The Consumer Technology Association has taken steps to facilitate standardized evaluation of some indicators from monitoring devices but collaboration with research-based organizations would promote transparency and quality control. Preliminary guidelines for companies seeking a basic level of endorsement could include documentation of the technology and sensors used in the devices as well as published reports on the reliability of signals and outputs. Descriptions of calibration methods used to establish the values should be provided as well as evidence of external validation based on the protocols and methods described here. The AHA position statement on physician counseling (11) referenced roles that organization, such as the Healthcare Information and Management Systems Society and Open mHealth.org can play in this process. Leadership is clearly needed to drive this agenda, but the guidelines here provide a starting point to facilitate standardized reporting procedures in research.

SUMMARY AND CONCLUSIONS

The focus of this review was on the impact of analytic methods on the outcomes and interpretations of validation studies on accelerometry-based activity monitors. Emphasis was placed on considerations for evaluating PA behavior, but the analytic methods and reporting guidelines would also apply for research on SB. The current PAGA report emphasizes the mantra "move more—sit less" (12) so it is also important to consider whether devices can effectively capture SB. This is especially important considering the evolving paradigm shift toward 24-h epidemiology and 24-h movement guidelines that emphasize "healthy movement profiles" as opposed to specific behaviors (84). The recommended use of METs as a unifying metric is consistent with these new concepts because tracking of MET-minutes enables the creation of profiles that capture the full intensity spectrum (sedentary, light, moderate, and vigorous). However, there is continued need for innovation in methods to obtain more accurate estimates of these indicators. The main point of the article is that standardized analytic methods and reporting of outcomes are critical to systematically advance research on PA and SB with these devices.

The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation, and statement that results of the present study do not constitute endorsement by ACSM.

Conflicts of Interest and Source of Funding: The authors declare no conflicts of interest. The authors acknowledge partial funding support for the publication of this work from the Mobilize Center, a National Institutes of Health (NIH) Big Data to Knowledge Center of Excellence supported by NIH grant U54 EB020405.

REFERENCES

1. "Worldwide wearables market forecast to nearly double by 2021, according to IDC." International Data Corporation, 21 June. 2017, https://www.idc.com/getdoc.jsp?containerId=prUS42818517.
2. Witkowski Wallace. "Smartwatch growth surges ahead of apple, Fitbit offerings." MarketWatch.com,1 September. 2017, https://www.marketwatch.com/story/smartwatch-growth-surges-ahead-of-apple-fitbit-offerings-2017-08-31.
3. "Connected wearable devices worldwide 2016–2021." Statista, 2018, https://www.statista.com/statistics/487291/global-connected-wearable-devices/.
4. Burke LE, Ma J, Azar KM, et al. Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American Heart Association. *Circulation*. 2015; 132(12):1157–213.
5. Knight E, Stuckey MI, Prapavessis H, Petrella RJ. Public health guidelines for physical activity: is there an app for that? A review of android and apple app stores. *JMIR Mhealth Uhealth*. 2015; 3(2):e43.
6. "The growing value of digital health, evidence and impact on human health and the healthcare system." IQVIA, 7 November. 2017, https://www.iqvia.com/institute/reports/the-growing-value-of-digital-health.
7. Haghi M, Thurow K, Stoll R. Wearable devices in medical internet of things: scientific research and commercially available devices. *Healthc Inform Res*. 2017;23(1):4–15.
8. Lobelo F, Kelli HM, Tejedor SC, et al. The wild wild west: a framework to integrate mHealth software applications and wearables to support physical activity assessment, counseling and interventions for cardiovascular disease risk reduction. *Prog Cardiovasc Dis*. 2016;58(6):584–94.
9. Shephard RJ. Limits to the measurement of habitual physical activity by questionnaires. *Br J Sports Med*. 2003;37(3):197–206; discussion 06.
10. Loiselle CG, Ahmed S. Is connected health contributing to a healthier population? *J Med Internet Res*. 2017;19(11):e386.
11. Lobelo F, Rohm Young D, Sallis R, et al. Routine assessment and promotion of physical activity in healthcare settings: a scientific statement from the American Heart Association. *Circulation*. 2018;137(18): e495–522.
12. US Department of Health and Human Services. *Physical Activity Guidelines for Americans*. 2nd ed. Washington, DC; 2018.
13. US Department of Health and Human Services. *Physical Activity Guidelines for Americans*. Washington, DC; 2008. pp. 15–34.
14. Physical Activity Guidelines Advisory Committee. *Physical Activity Guidelines Advisory Committee Scientific Report*. Washington, DC: U.S. Department of Health and Human Services; 2018.
15. Welk GJ. Harmonizing monitor- and report-based estimates of physical activity through calibration. *Kinesiol Rev*. 2019;8(1):16–24.
16. Bassett DR Jr, Toth LP, LaMunion SR, Crouter SE. Step counting: a review of measurement considerations and health-related applications. *Sports Med*. 2017;47(7):1303–15.

17. "Number of Fitbit devices sold worldwide from 2010 to 2017." Statista, 2018, https://www.statista.com/statistics/472591/fitbit-devices-sold/.
18. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. *Br J Sports Med*. 2014;48(13):1019–23.
19. Migueles JH, Cadenas-Sanchez C, Ekelund U, et al. Accelerometer data collection and processing criteria to assess physical activity and other outcomes: a systematic review and practical considerations. *Sports Med*. 2017;47(9):1821–45.
20. Freedson P, Bowles HR, Troiano R, Haskell W. Assessment of physical activity using wearable monitors: recommendations for monitor calibration and use in the field. *Med Sci Sports Exerc*. 2012;44(1 Suppl 1):S1–4.
21. Bassett DR, Troiano RP, McClain JJ, Wolff DL. Accelerometer-based physical activity: total volume per day and standardized measures. *Med Sci Sports Exerc*. 2015;47(4):833–8.
22. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int J Behav Nutr Phys Act*. 2015;12:159.
23. Welk GJ, Blair SN, Wood K, Jones S, Thompson RW. A comparative evaluation of three accelerometry-based physical activity monitors. *Med Sci Sports Exerc*. 2000;32(9 Suppl):S489–97.
24. Freedson PS, Melanson E, Sirard J. Calibration of the computer science and applications, Inc. accelerometer. *Med Sci Sports Exerc*. 1998;30(5):777–81.
25. Hendelman D, Miller K, Baggett C, Debold E, Freedson P. Validity of accelerometry for the assessment of moderate intensity physical activity in the field. *Med Sci Sports Exerc*. 2000;32(9 Suppl):S442–9.
26. Swartz AM, Strath SJ, Bassett DR Jr, O'Brien WL, King GA, Ainsworth BE. Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Med Sci Sports Exerc*. 2000; 32(9 Suppl):S450–6.
27. Brazendale K, Beets MW, Bornstein DB, et al. Equating accelerometer estimates among youth: the Rosetta stone 2. *J Sci Med Sport*. 2016;19(3):242–9.
28. Bornstein DB, Beets MW, Byun W, et al. Equating accelerometer estimates of moderate-to-vigorous physical activity: in search of the Rosetta stone. *J Sci Med Sport*. 2011;14(5):404–10.
29. Bassett DR Jr, Ainsworth BE, Swartz AM, Strath SJ, O'Brien WL, King GA. Validity of four motion sensors in measuring moderate intensity physical activity. *Med Sci Sports Exerc*. 2000;32(9 Suppl): S471–80.
30. Crouter SE, Clowers KG, Bassett DR Jr. A novel method for using accelerometer data to predict energy expenditure. *J Appl Physiol (1985)*. 2006;100(4):1324–31.
31. Crouter SE, Bassett DR Jr. A new 2-regression model for the Actical accelerometer. *Br J Sports Med*. 2008;42(3):217–24.
32. Heil DP. Predicting activity energy expenditure using the Actical activity monitor. *Res Q Exerc Sport*. 2006;77(1):64–80.
33. Pober DM, Staudenmayer J, Raphael C, Freedson PS. Development of novel techniques to classify physical activity mode using accelerometers. *Med Sci Sports Exerc*. 2006;38(9):1626–34.
34. Rothney MP, Neumann M, Beziat A, Chen KY. An artificial neural network model of energy expenditure using nonintegrated acceleration signals. *J Appl Physiol (1985)*. 2007;103(4): 1419–27.
35. Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *J Appl Physiol (1985)*. 2009;107(4):1300–7.
36. Jakicic JM, Marcus M, Gallagher KI, et al. Evaluation of the SenseWear pro armband to assess energy expenditure during exercise. *Med Sci Sports Exerc*. 2004;36(5):897–904.
37. Arvidsson D, Slinde F, Larsson S, Hulthen L. Energy cost of physical activities in children: validation of SenseWear armband. *Med Sci Sports Exerc*. 2007;39(11):2076–84.
38. Brage S, Ekelund U, Brage N, et al. Hierarchy of individual calibration levels for heart rate and accelerometry to measure physical activity. *J Appl Physiol (1985)*. 2007;103(2):682–92.
39. Brage S, Westgate K, Franks PW, et al. Estimation of free-living energy expenditure by heart rate and movement sensing: a doubly-labelled water study. *PLoS One*. 2015;10(9):e0137206.
40. Bai Y, Hibbing P, Mantis C, Welk GJ. Comparative evaluation of heart rate-based monitors: apple watch vs Fitbit charge HR. *J Sports Sci*. 2018;36(15):1734–41.
41. Bai Y, Welk GJ, Nam YH, et al. Comparison of consumer and research monitors under Semistructured settings. *Med Sci Sports Exerc*. 2016;48(1):151–8.
42. Montoye AH, Vusich J, Mitrzyk J, Wiersma M. Heart rate alters, but does not improve, calorie predictions in Fitbit activity monitors. *J Meas Phys Behav*. 2018;1(1):9–17.
43. Kelly P, Fitzsimons C, Baker G. Should we reframe how we think about physical activity and sedentary behaviour measurement? Validity and reliability reconsidered. *Int J Behav Nutr Phys Act*. 2016; 13:32.
44. Dowd KP, Szeklicki R, Minetto MA, et al. A systematic literature review of reviews on techniques for physical activity measurement in adults: a DEDIPAC study. *Int J Behav Nutr Phys Act*. 2018;15(1):15.
45. Welk GJ. Principles of design and analyses for the calibration of accelerometry-based activity monitors. *Med Sci Sports Exerc*. 2005; 37(11 Suppl):S501–11.
46. Matthews CE, Keadle SK, Berrigan D, et al. Influence of accelerometer calibration approach on moderate-vigorous physical activity estimates for adults. *Med Sci Sports Exerc*. 2018;50:2285–91.
47. Bassett DR Jr, Rowlands A, Trost SG. Calibration and validation of wearable monitors. *Med Sci Sports Exerc*. 2012;44(1 Suppl 1):S32–8.
48. Welk GJ, McClain J, Ainsworth BE. Protocols for evaluating equivalency of accelerometry-based activity monitors. *Med Sci Sports Exerc*. 2012;44(1 Suppl 1):S39–49.
49. Staudenmayer J, Zhu W, Catellier DJ. Statistical considerations in the analysis of accelerometry-based activity monitor data. *Med Sci Sports Exerc*. 2012;44(1 Suppl 1):S61–7.
50. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc*. 2009;41(1):3–13.
51. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. *PLoS One*. 2012;7(5):e37908.
52. Hanneman SK. Design, analysis, and interpretation of method—comparison studies. *AACN Adv Crit Care*. 2008;19(2):223–34.
53. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–10.
54. Bunce C. Correlation, agreement, and Bland–Altman analysis: statistical analysis of method comparison studies. *Am J Ophthalmol*. 2009; 148(1):4–6.
55. Dixon PM, Saint-Maurice PF, Kim Y, Hibbing P, Bai Y, Welk GJ. A primer on the use of equivalence testing for evaluating measurement agreement. *Med Sci Sports Exerc*. 2018;50(4):837–45.
56. Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Soc Psychol Personal Sci*. 2017;8(4): 355–62.
57. Nusser SM, Beyler NK, Welk GJ, Carriquiry AL, Fuller WA, King B MN. Modeling errors in physical activity recall data. *J Phys Act Health*. 2012;9(1 Suppl):S56–67.
58. Tooze JA, Troiano RP, Carroll RJ, Moshfegh AJ, Freedman LS. A measurement error model for physical activity level as measured by a questionnaire with application to the 1999–2006 NHANES questionnaire. *Am J Epidemiol*. 2013;177(11):1199–208.
59. Neuhouser ML, Di C, Tinker LF, et al. Physical activity assessment: biomarkers and self-report of activity-related energy expenditure in the WHI. *Am J Epidemiol*. 2013;177(6):576–85.

60. Dannecker KL, Sazonova NA, Melanson EL, Sazonov ES, Browning RC. A comparison of energy expenditure estimation of several physical activity monitors. *Med Sci Sports Exerc*. 2013;45(11):2105–12.
61. Gusmer R, Bosch T, Watkins A, Ostrem J, Dengel D. Comparison of FitBit® ultra to ActiGraph™ GT1M for assessment of physical activity in young adults during treadmill walking. *Open Sports Med J*. 2014;8(1).
62. Lee JM, Kim Y, Welk GJ. Validity of consumer-based physical activity monitors. *Med Sci Sports Exerc*. 2014;46(9):1840–8.
63. Sasaki JE, Hickey A, Mavilia M, et al. Validation of the Fitbit wireless activity tracker for prediction of energy expenditure. *J Phys Act Health*. 2015;12(2):149–54.
64. Diaz KM, Krupka DJ, Chang MJ, et al. Fitbit®: an accurate and reliable device for wireless physical activity tracking. *Int J Cardiol*. 2015;185:138–40.
65. Tucker WJ, Bhammar DM, Sawyer BJ, Buman MP, Gaesser GA. Validity and reliability of Nike + Fuelband for estimating physical activity energy expenditure. *BMC Sports Sci Med Rehabil*. 2015;7:14.
66. Nelson MB, Kaminsky LA, Dickin DC, Montoye AH. Validity of consumer-based physical activity monitors for specific activity types. *Med Sci Sports Exerc*. 2016;48(8):1619–28.
67. Wallen MP, Gomersall SR, Keating SE, Wisloff U, Coombes JS. Accuracy of heart rate watches: implications for weight management. *PLoS One*. 2016;11(5):e0154420.
68. Alsubheen SA, George AM, Baker A, Rohr LE, Basset FA. Accuracy of the vivofit activity tracker. *J Med Eng Technol*. 2016;40(6):298–306.
69. Murakami H, Kawakami R, Nakae S, et al. Accuracy of wearable devices for estimating total energy expenditure: comparison with metabolic chamber and doubly labeled water method. *JAMA Intern Med*. 2016;176(5):702–3.
70. Dooley EE, Golaszewski NM, Bartholomew JB. Estimating accuracy at exercise intensities: a comparative study of self-monitoring heart rate and physical activity wearable devices. *JMIR Mhealth Uhealth*. 2017;5(3):e34.
71. Imboden MT, Nelson MB, Kaminsky LA, Montoye AH. Comparison of four Fitbit and jawbone activity monitors with a research-grade ActiGraph accelerometer for estimating physical activity and energy expenditure. *Br J Sports Med*. 2018;52(13):844–50.
72. Shcherbina A, Mattsson CM, Waggott D, et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *J Pers Med*. 2017;7(2). doi: 10.3390/jpm7020003.
73. Chowdhury EA, Western MJ, Nightingale TE, Peacock OJ, Thompson D. Assessment of laboratory and daily energy expenditure estimates from consumer multi-sensor physical activity monitors. *PLoS One*. 2017;12(2):e0171720.
74. Price K, Bird SR, Lythgo N, Raj IS, Wong JY, Lynch C. Validation of the Fitbit one, Garmin Vivofit and jawbone UP activity tracker in estimation of energy expenditure during treadmill walking and running. *J Med Eng Technol*. 2017;41(3):208–15.
75. Wahl Y, Düking P, Droszez A, Wahl P, Mester J. Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions. *Front Physiol*. 2017;8:725.
76. Parak J, Uuskoski M, Machek J, Korhonen I. Estimating heart rate, energy expenditure, and physical performance with a wrist photoplethysmographic device during running. *JMIR Mhealth Uhealth*. 2017;5(7):e97.
77. Woodman JA, Crouter SE, Bassett DR Jr, Fitzhugh EC, Boyer WR. Accuracy of consumer monitors for estimating energy expenditure and activity type. *Med Sci Sports Exerc*. 2017;49(2):371–7.
78. Xie J, Wen D, Liang L, Jia Y, Gao L, Lei J. Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: comparative study. *JMIR Mhealth Uhealth*. 2018; 6(4):e94.
79. Boudreaux BD, Hebert EP, Hollander DB, et al. Validity of wearable activity monitors during cycling and resistance exercise. *Med Sci Sports Exerc*. 2018;50(3):624–33.
80. Adam Noah J, Spierer DK, Gu J, Bronner S. Comparison of steps and energy expenditure assessment in adults of Fitbit tracker and ultra to the actical and indirect calorimetry. *J Med Eng Technol*. 2013;37(7): 456–62.
81. DeShaw KJ, Ellingson L, Bai Y, Lansing J, Perez M, Welk G. Methods for activity monitor validation studies: an example with the Fitbit charge. *Journal for the Measurement of Physical Behaviour*. 2018;1(3):130–5.
82. Kooiman TJ, Dontje ML, Sprenger SR, Krijnen WP, van der Schans CP, de Groot M. Reliability and validity of ten consumer activity trackers. *BMC Sports Sci Med Rehabil*. 2015;7:24.
83. U.S. Department of Health and Human Services Food and Drug Administration. *Guidance for Industry, Statistical Approaches to Establishing Bioequivalence*. Rockville, MD; 2001.
84. Tremblay MS, Carson V, Chaput J-P, et al. Canadian 24-hour movement guidelines for children and youth: an integration of physical activity, sedentary behaviour, and sleep. *Appl Physiol Nutr Metab*. 2016;41(6):S311–27.