EVIDENCE ANALYSIS PROJECT:

VALIDITY AND UTILITY OF ACTIVITY MONITORS TO IMPROVE HEALTH OUTCOMES

Alana Degeneffe, RDN, CD

Submitted in partial fulfillment of the requirements for the degree

Master of Science in Dietetics

Mount Mary University

December 2018

Approved by

Megan Baumler, PhD, RD
Professor, Graduate Program in Dietetics

Tara LaRowe, PhD, RDN
Professor, Graduate Program in Dietetics

EVIDENCE ANALYSIS PROJECT:
VALIDITY AND UTILITY OF ACTIVITY MONITORS TO IMPROVE HEALTH OUTCOMES

Alana Degeneffe, RDN, CD

December 2018

ABSTRACT

BACKGROUND: Worldwide obesity rates have nearly tripled since 1975 and are projected to increase further. The prevention and treatment of obesity has focused on pharmacological, educational, and behavioral interventions with limited success overall. A recent trend and possible technological intervention, activity monitors are wearable accelerometer-based devices aimed for the consumer market. Activity monitoring technology offers promise for improving adherence and weight loss outcomes.

OBJECTIVE: To critically analyze and appraise current evidence on the validity and utility of activity monitors to improve health outcomes in adults.

DESIGN: The Academy of Nutrition and Dietetics' (AND) Evidence Analysis Library (EAL) project.

METHODS: The five steps included in the AND's Evidence Analysis process are 1) Formulate the Evidence Analysis Question, 2) Gather and Classify the Evidence, 3) Critically Appraise Each Article, 4) Summarize the Evidence, and 5) Write and Grade the Conclusion Statement.

RESULTS: Using the PubMed database, a total of 18 studies were identified relating to activity monitor validity and utility in adults. Six studies were excluded because they assessed participants with more than one health condition, the predictability of body weight changes, or activity monitors that were not consumer-grade devices. Twelve studies were included for further analysis. Of the 12 studies, seven were validity studies and five were randomized controlled trials, non-randomized crossover trials, and a cross sectional study exploring activity monitor utility. Not all studies analyzed activity monitor validity or utility in improving health outcomes within the same parameters. Four validity studies were conducted under controlled, research environments utilizing gold standard comparisons, while three validity studies were conducted in free-living environments as activity monitors are intended for, utilizing common field-based devices as comparisons. Utility studies did not evaluate activity monitors as independent interventions. Instead, most utility studies incorporated activity monitors as self-monitoring tools integrated into behavioral interventions with positive outcomes.

CONCLUSION: Current, consumer-grade activity monitors exhibit moderate validity on average, tend to estimate step counts accurately, underestimate heart rate and energy expenditure, overestimate time asleep, and are more accurate at rest than during activity. Adults who utilize current, consumer-grade activity monitors as combined interventions may experience

a clinically meaningful increase in steps, physical activity, and weight loss. This conclusion was graded *II, Fair*.

ACKNOWLEDGMENTS

I would like to thank my parents, grandparents, and fiancé for instilling a fierce work ethic in me, believing I can achieve anything I set my mind to, and encouraging me through my entire academic journey. I would not be where I am today without your love and support.

Thank you to my colleagues and clients for inspiring me to pursue this research topic.

Last but certainly not least, thank you to Megan Baumler, my professor and advisor for sticking with me through this process. I truly appreciate all the time and effort you dedicated to guiding and improving my Evidence Analysis project.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION

Overweight and obesity are defined as abnormal, excessive fat accumulation that may impair health. In 2016, nearly two billion adults (39 percent of the world's population) were overweight. Of these, over 650 million adults (13 percent of the world's population) were obese. The prevalence has also risen among children. In 2016, over 18 percent of the world's population of children were overweight or obese. Since 1975, obesity rates have nearly tripled, and are projected to increase further (World Health Organization, 2018).

Personal behaviors play a dominant role in preventing and treating obesity (Hruby & Hu, 2015), as the fundamental cause of obesity is an energy imbalance between calories consumed and calories expended. Lifestyle adaptations can result in energy imbalances that promote obesity, such as consuming calorically-dense diets without adequate physical activity. Physical activity is defined as any body movement produced by the skeletal muscles that requires energy expenditure. Therefore, physical activity is an essential component of weight management. However, an estimated 31 percent of adults worldwide do not meet the recommended levels. Physical inactivity contributes to obesity, chronic illnesses, and other healthcare concerns (World Health Organization, 2018), while obesity has drastic impacts on morbidity, mortality, and economic burden (Hruby & Hu, 2015).

A recent trend and possible technological intervention, activity monitors are wearable accelerometer-based devices aimed for the consumer market. Most activity monitors have displays for immediate health feedback, including step counts, elevation or stairs climbed, distance traveled, heart rate, calories burned, active time, and time asleep (Ferguson, Rowlands, Olds, & Maher, 2015). Additionally, many activity monitors have associated websites, smartphone applications, and smart scales that synchronize with the wearable devices. Users can

track their workouts, water consumption, caloric intake, and weight fluctuations through the synchronizing options (Ross & Wing, 2016).

Technology-based tools create awareness by assisting individuals in tracking their weight-related behaviors (Ross & Wing, 2016). Activity monitoring technology shows potential for facilitating self-motivation, self-monitoring, self-efficacy, and positive behavior change, all of which are essential qualities for successful behavioral weight management treatments (Ferguson et al., 2015). Overall, activity monitoring technology offers promise for improving adherence and weight loss outcomes (Ross & Wing, 2016). However, the usefulness of activity monitors depends highly on their accuracy.

Rationale

The prevention and treatment of obesity has focused on pharmacological, educational, and behavioral interventions, with limited success overall (Townshend & Lake, 2017). However, the implementation of physical activity programs with self-motivating, self-monitoring mechanisms may be efficacious in suppressing the global physical inactivity and obesity epidemics. Furthermore, efforts to increase physical activity and control obesity will result in tremendous economic savings and remarkable health benefits (Lewis, Lyons, Jarvis, & Baillargeon, 2015). The purpose of this Evidence Analysis project is to critically analyze and appraise current evidence on the validity and utility of activity monitors to improve health outcomes in adults. A systematic review of existing literature has been conducted, and conclusion statements have been determined based on the findings of the review.

Potential Significance

Results from this Evidence Analysis project will indicate whether activity monitors are useful additions to healthy lifestyle interventions. The findings may also contribute to future

research and product development. With progress, activity monitor functions should reach reference device standards, and may have practical clinical applications for all populations. Dietitians and other healthcare professionals could provide evidence-based recommendations regarding appropriate activity monitor selection. With automatic data transfer of validated activity monitors, healthcare professionals could accurately track clients' physical activity habits, heart rate, energy expenditure, sleep patterns, food and beverage consumption, and weight fluctuations. Direct access to such data would save an immense amount of time, provide a holistic picture of health for each client, and enable healthcare professionals to personalize care more than ever before. Future technologies may be able to provide an early warning of disease, aid in diagnosis and treatment, and contribute to a deeper understanding of human health (Savage, 2017). Hopefully, the individualized approach to care by means of utilizing activity monitor data will result in better health outcomes for all populations, ultimately leading to a decreased prevalence of overweight and obesity, chronic illnesses, mortality, and healthcare costs.

Objectives

A. Identify the validity of current, consumer-grade activity monitor functions
B. Identify the health outcome improvements when current, consumer-grade activity monitors are utilized
C. Compare 10 current, best-selling, commercially-available activity monitor product details in a supplemental comparison report

Research Questions

1. How valid are the functions of current, consumer-grade activity monitors compared to research-grade devices?

2. Does physical activity improve and weight loss occur as a result when adults utilize current, consumer-grade activity monitors?

Research Hypotheses

1. Current, consumer-grade activity monitor functions will be within 25 percent error range of research-grade devices
2. Adults who utilize current, consumer-grade activity monitors will experience an increase in physical activity and weight loss as a result

Limitations

Data has been collected exclusively from published research. Due to the nature of an Evidence Analysis project, two limitations are the reliance on, and lack of pertinent research articles. In regard to objective C of this Evidence Analysis project, financial constraint was an additional limitation, restricting data collection to manufacturer websites only.

Delimitations

Data has been collected from the most current studies due to the rapidly evolving activity monitor market. Validity data was limited further to the activity monitors and specific functions tested in each study, even though updated activity monitor models may currently be on the market. Likewise, utility data must take the results, strengths, and limitations of activity monitor validity data into consideration.

Assumptions

This Evidence Analysis project assumes the accuracy and honesty of all published research.

Definition of Terms

- **Accelerometry**: the electromechanical measurement of acceleration and deceleration in a part of, or the entire body during the performance of a task. A common, noninvasive procedure used to capture physical activity intensity using a wearable device called an accelerometer.
- **Actigraphy**: a method of monitoring body movements over time to determine periods of rest versus activity. A common, noninvasive procedure used to detect sleep disorders by using a wearable device called an actigraph.
- **Electrocardiography**: a noninvasive procedure used to record electrical changes in the heart. The record, which is called an electrocardiogram, shows the series of waves that relate to the electrical impulses that occur during each beat of the heart.
- **Indirect calorimetry**: a noninvasive procedure used to quantify energy expenditure. An individual's heat production is determined by measuring oxygen uptake and carbon dioxide output over a given period of time.
- **Polysomnography**: a noninvasive procedure used to study sleep and diagnose sleep disorders. The record, which is called a polysomnogram, shows brain waves, blood oxygen levels, heart and breathing rates, eye and leg movements, and sleep stages.
- **Utility**: the state of making practical and effective use of something
- **Validity**: the extent to which a variable or measure captures the concept it is intended to reflect

CHAPTER II: LITERATURE REVIEW

Introduction

Overweight and obesity are defined as abnormal, excessive fat accumulation that may impair health (World Health Organization, 2018). The current, most widely used criteria for classifying overweight and obesity is body mass index (BMI), which is calculated by dividing body weight in kilograms by height in meters squared (Hruby & Hu, 2015). A BMI greater than or equal to 25 is considered overweight. A BMI greater than or equal to 30 is considered obese. In 2016, nearly two billion adults (39 percent of the world's population) were overweight. Of these, over 650 million adults (13 percent of the world's population) were obese. The prevalence has also risen among children. In 2016, over 18 percent of the world's population of children were overweight or obese. Obesity rates have nearly tripled since 1975 and are projected to increase further (World Health Organization, 2018).

The fundamental cause of overweight and obesity is an energy imbalance between calories consumed and calories expended (World Health Organization, 2018). This energy imbalance is partially a result of environmental changes beyond the control of any individual. Environmental changes that promote obesity, known as obesogenic environments, have been fueled by industrialization, automation, transportation, urbanization, economic growth, sedentary lifestyles, and consuming highly processed, calorically-dense diets. Risk factors of obesity unrelated to environmental changes include lower socioeconomic status, limited education, and hereditary factors such as genetics, family history, racial, and ethnic differences (Hruby & Hu, 2015). Likewise, obesity is a major risk factor for a number of noncommunicable diseases, namely musculoskeletal disorders, diabetes, cardiovascular disease, and certain cancers. As BMI increases, the risk for theses chronic diseases also increases (World Health Organization, 2018).

Overall, obesity has drastic impacts on morbidity, mortality, healthcare costs, and economic burden. However, obesity risk factors are modifiable. In response to these conditions, personal behaviors play a dominant role in preventing and treating obesity (Hruby & Hu, 2015).

Physical activity is defined as any body movement produced by the skeletal muscles that requires energy expenditure (World Health Organization, 2018). Therefore, physical activity is an essential component of weight management. According to the 2008 Physical Activity Guidelines for Americans, adults 18 to 64 years of age should do at least 150 minutes of moderate-intensity, or 75 minutes of vigorous-intensity aerobic physical activity per week (Office of Disease Prevention and Health Promotion, 2018). However, it is estimated that 31 percent of adults worldwide do not meet these recommended levels. Physical inactivity can be attributed to an increase in the use of transportation, sedentary behavior during occupational and domestic activities, and insufficient participation in physical activity during leisure time. Strong evidence demonstrates that adults who meet the recommended levels of physical activity have lower rates of fractures, depression, diabetes, cardiovascular disease, cancer, and mortality compared to those who do not. Adults who are more active also exhibit a higher level of physical fitness, a lower BMI, and are more likely to achieve weight maintenance (World Health Organization, 2018). Thus, efforts to increase physical activity and control obesity will result in tremendous economic savings and remarkable health benefits (Lewis, Lyons, Jarvis, & Baillargeon, 2015). The implementation of physical activity programs with self-motivating, self-monitoring mechanisms may be efficacious in suppressing the global physical inactivity and obesity epidemics.

A recent trend and possible technologic intervention, activity monitors are wearable accelerometer-based devices aimed for the consumer market. Most activity monitors have

displays for immediate health feedback, including step counts, elevation or stairs climbed, distance traveled, heart rate, calories burned, active time, and time asleep (Ferguson, Rowlands, Olds, & Maher, 2015). Additionally, many activity monitors have associated websites, smartphone applications, and smart scales that synchronize with the wearable devices. Users can track their workouts, water consumption, caloric intake, and weight fluctuations through the synchronizing options (Ross & Wing, 2016).

Activity monitoring technology also shows potential for facilitating self-motivation, self-monitoring, self-efficacy, and positive behavior change, all of which are essential qualities for successful behavioral weight management treatments (Ferguson et al., 2015). Technology-based tools create awareness by assisting individuals in tracking their weight-related behaviors. Activity monitoring technology offers promise for improving adherence and weight loss outcomes (Ross & Wing, 2016). However, the usefulness of activity monitors depends highly on their accuracy. The purpose of this literature review is to critically analyze current evidence on the validity and utility of activity monitors to improve health outcomes in adults.

Background

The prevention and treatment of obesity has focused on pharmacological, educational, and behavioral interventions, with limited success overall (Townshend & Lake, 2017).

Pharmacotherapy

The history of weight loss drugs has seen the rise and fall of numerous medications that proved highly effective, but ultimately dangerous. The purpose of pharmacotherapy is not to search for a magic pill, but rather to apply a safe and effective drug regimen, in combination with improved diet and exercise to achieve a sustainable reduction in body weight. As of 2013, only three drugs were approved by the United States Food and Drug Administration (FDA) as

adjunctive therapy for chronic weight management. The FDA-approved anti-obesity medications include orlistat (brands include Alli and Xenical), lorcaserin (trade name Belviq), and phentermine/topiramate extended-release (trade name Qsymia) (Kim, Lin, Blomain, & Waldman, 2014).

Orlistat is the only FDA-approved anti-obesity medication that is available without a prescription. Orlistat acts by binding and inhibiting pancreatic and gastrointestinal lipases from breaking down dietary triglycerides into free fatty acids, which can be absorbed via fatty acid transporters expressed by the intestinal epithelial cells. Thus, orlistat decreases systemic fat absorption and reduces caloric intake. However, major adverse effects of orlistat include steatorrhea and the risk of fat-soluble vitamin deficiencies. Hepatotoxicity, nephrotoxicity, pancreatitis, and kidney stones are additional safety concerns (Kim, Lin, Blomain, & Waldman, 2014).

Serotonin is a neurotransmitter that mediates several processes in the central nervous system. One of the physiological processes that serotonin regulates is postprandial satiety through hypothalamic serotonin receptors. Lorcaserin is a selective hypothalamic serotonin receptor agonist. Therefore, lorcaserin decreases appetite and food intake, prevents weight gain, and assists in weight loss. Although, adverse effects of lorcaserin include headache, dizziness, fatigue, dry mouth, nausea, and constipation. Psychiatric disorders, cardiovascular events, and carcinogenesis are additional safety concerns (Kim, Lin, Blomain, & Waldman, 2014).

Given the complex, multifactorial etiology of obesity, it is unlikely that one weight loss drug will be sufficient to reverse the condition. Unsurprisingly, combination therapies such as phentermine/topiramate extended-release have been evaluated and show a greater potential in the treatment of obesity. Phentermine's mechanism of action is reliant on modulation of

catecholamines in the satiety centers of the hypothalamus, thus decreasing appetite. Topiramate is an alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate/kainate receptor antagonist. Modification of voltage-gated sodium and calcium channels, as well as induction of gamma-aminobutyric acid receptor-mediated inhibitory currents may also contribute to topiramate's weight loss effects by decreasing food intake and efficiency of nutrient utilization, and increasing energy expenditure. However, various adverse effects of phentermine/topiramate extended-release have been noted, such as headache, dizziness, blurry vision, paresthesia, insomnia, anxiety, depression, dry mouth, upper respiratory tract infection, nasopharyngitis, and constipation. Psychiatric disorders, cardiovascular events, and teratogenesis are additional safety concerns. The modest efficacy, undesirable adverse effects, serious health risks, questionable safety, and potential of weight regain combine to highlight the major limitations of weight loss drugs in the treatment of obesity (Kim, Lin, Blomain, & Waldman, 2014).

Medical Nutrition Therapy

The AND describes medical nutrition therapy (MNT) as an evidence-based, in-depth application of the Nutrition Care Process (NCP). The NCP entails an individualized nutrition assessment, determination of the nutrition diagnosis, determination and application of the nutrition intervention appropriate for the individual or group, routine monitoring, and evaluation to manage the disease, condition, or injury. MNT employs all domains of nutrition intervention, including food and nutrient delivery, nutrition education, nutrition counseling, and coordination of nutrition care. MNT services are provided by registered dietitians for individuals and groups utilizing meal plans, medically prescribed diets, specialized oral feedings, tube feedings, intravenous solutions, and the analysis of potential food and drug interactions. The provision of

MNT typically results in the prevention, delay, or management of diseases, conditions, or injuries (Academy of Nutrition and Dietetics, 2017).

The AND defines nutrition education as the reinforcement of basic or essential nutrition-related knowledge (Academy of Nutrition and Dietetics, 2006). In regard to weight loss and maintenance, nutrition education may cover an array of topics such as food preparation, portion control, reading nutrition labels, calorie, and carbohydrate counting. Nutrition counseling is a supportive process used to set priorities, establish goals, create individualized action plans, and promote accountability for self-care (Academy of Nutrition and Dietetics, 2006). It is common practice for registered dietitians to incorporate psychology-based techniques into nutrition counseling, such as applying the transtheoretical model, motivational interviewing, and providing positive reinforcement.

In overweight and obese adults, strong evidence supports the effectiveness of multiple visits for MNT provided by a registered dietitian. According to the AND's EAL, research demonstrates improvements in anthropometric measurements as well as biochemical data. Substantial evidence supports a reduction in weight (-0.5 kg to -9.0 kg), waist circumference (-2.0 cm to -14.0 cm), BMI (-0.2 kg/m^2 to -7.8 kg/m^2), fasting blood glucose (-5.2 mg/dL to -9.5 mg/dL), total cholesterol (-4.3 mg/dL to -59 mg/dL), LDL-cholesterol (-15 mg/dL to -47 mg/dL), HDL-cholesterol (+2.0 mg/dL to +11 mg/dL), and triglycerides (-12 mg/dL to -60 mg/dL) (Evidence Analysis Library, 2015). Despite the strong evidence to support the effectiveness of MNT, not all overweight and obese individuals have access to MNT services.

Educational Interventions

Nutrition education in school provides students with the knowledge to develop proper eating habits, along with necessary skills to make well-informed decisions regarding their health

in the future. Students learn to choose healthy foods through effective nutrition education. Nutrition education standards are not mandatory, although school districts are encouraged to use them to develop a comprehensive kindergarten through twelfth grade nutrition education curriculum. Nutrition education standards integrate health, science, and mathematics content with real-world applications through various learning activities. When students see the connection between what they are learning and real-life examples, their motivation and learning intensifies. Key concepts presented in Wisconsin's nutrition education standards promote health literacy, food safety, food preparation, healthy eating behaviors, nutrition for growth, health, and energy. Wisconsin's nutrition education standards support variety, moderation, and balance in food choices, with the fundamental goal of engaging students in their education as they make healthier choices for themselves and their families. Schools play a significant role in helping students develop healthy eating habits by providing nutritious meals and snacks through the schools' meal programs. Additionally, school district wellness policies can implement nutrition education content by establishing healthy school environments (Wisconsin Department of Public Instruction, 2009).

Physical education in school benefits both academic performance and physical activity patterns of students. Students learn to make informed decisions and understand the value of leading an active lifestyle through effective physical education. Similar to nutrition education standards, physical education standards are not mandatory, although school districts are encouraged to use them. Physical education standards provide developmental guidance for a consistent kindergarten through twelfth grade physical education curriculum. Key concepts presented in Wisconsin's physical education standards promote physical fitness, healthy physical activity behaviors, skill development, stress reduction, improved judgement, strengthened peer

relationships, goal-setting, self-monitoring, self-discipline, self-confidence, and self-esteem. Integrating listening, speaking, reading, and writing into physical skills and activities creates cross-curricular connections, making learning relevant and meaningful to students (Wisconsin Department of Public Instruction, 2010). While nutrition and physical education are necessary components of students' overall education, standards are not mandated by schools, and support is only temporary as curriculums end before students reach an independent stage of adulthood.

Behavioral Interventions

Commercial weight loss programs are popular treatment options for overweight and obese adults, although their efficacy is unclear. In 2014, Americans spent nearly $2.5 billion on commercial weight loss services, with increases projected for years to come. Currently, Weight Watchers, Nutrisystem, and Jenny Craig are the top three programs dominating the weight loss services industry. These three programs are high intensity, with a focus on goal-setting, self-monitoring, and group support. Nutrisystem and Jenny Craig also endorse low calorie meal replacements. Results of a systematic review indicated that Weight Watchers' participants lost more weight than control participants, which they sustained beyond 12 months. Researchers concluded that Weight Watchers has weight loss efficacy, yet it may not be superior to behavioral counseling. Weight Watchers was also the most cost-effective weight management strategy compared to other commercial programs. Nutrisystem demonstrated greater short-term weight loss compared to control and behavior counseling participants, however, long-term results were not identified. Jenny Craig participants sustained more weight loss than control and behavior counseling participants, although Jenny Craig was more expensive because it includes the price of meal replacements. Based on these findings, it may be practical for healthcare professionals to refer patients to Weight Watchers or Jenny Craig if they lack the time, training,

or ancillary staff to deliver behavioral counseling in their practices. However, high program costs may make commercial weight loss services unaffordable for many individuals (Gudzune et al., 2015).

An innovative, long-term approach to obesity prevention should address the obesogenic environments that promote sedentary lifestyles, and the consumption of highly processed, calorically-dense diets. Shaping obesogenic environments to better support healthful decisions has the potential to be a key aspect of a successful obesity prevention intervention. Thus, in order to develop effective environmental interventions in relation to obesity, we must understand how individuals and groups interact within their environments, in terms of physical activity and food intake (Townshend & Lake, 2017).

Technological Interventions

Strong evidence supports the role of physical activity in managing obesity and other noncommunicable diseases. In healthcare, there is general consensus that technological interventions, including activity monitors, can potentially increase physical activity in patients. A qualitative interview-based study explored how physicians prescribed activity monitors to patients with cardiometabolic diseases. Results revealed that most physicians had never prescribed activity monitors, whereas they frequently prescribed blood glucose, blood pressure, or other self-monitoring devices. Reasons for nonprescription included a lack of interest in the data collected, a lack of evidence for data accuracy, concerns about work overload possibly resulting from automatic data transfer, and the risk of patients becoming addicted to data. Current activity monitor features are popular amongst consumers, but do not meet the needs of physicians. Physicians expected future technologies to measure physical activity intensity and duration accurately while providing understandable, motivating feedback. Understanding

physicians' expectations is a preliminary step in designing future technologies that can be widely used in clinical settings and facilitate physical activity prescription. Major healthcare stakeholders, including patients, physicians, researchers, and information technology firms should be involved in developing the most effective methods for integrating activity monitors into patient care (Bellicha, Macé, & Oppert, 2017). Moving beyond fitness tracking, activity monitors, smart watches, or their successors could provide an early warning of disease, aid in diagnosis and treatment, and contribute to a deeper understanding of human health (Savage, 2017).

In recent years, several companies have emerged as leaders of the activity monitor industry. Apple, Fitbit, Garmin, Jawbone, Microsoft, Misfit, Moov, Nike, Polar, Samsung, Striiv, TomTom, and Withings are companies at the forefront of the activity monitor industry. Activity monitors differ by price, size, style, battery life, compatibility, and tracking features. Wearable devices currently on the market range from approximately $20 to $600. A supplemental report will summarize, compare, and display product details of 10 current, best-selling, commercially-available activity monitors.

Beyond immediate health feedback, activity monitors also provide immediate reinforcement. Many activity monitors allow individuals to set short-term and long-term goals and compare their self-monitoring data to their goals, which supports self-efficacy. When individuals reach their goals, they may receive virtual badges, phone notifications, or email messages as positive reinforcement, which are intended to be self-motivating (Ross & Wing, 2016). Activity monitors are typically used for personal health or behavioral weight management programs, but are rarely tested as intervention tools. Below is a discussion of the

current evidence on the validity and utility of activity monitors to improve health outcomes in adults.

Activity Monitor Validity

A validity study by Shcherbina et al. assessed the accuracy of several commercially available wrist-worn devices in estimating heart rate and energy expenditure (2017). Seven commercially-available wrist-worn devices (Apple Watch, Basis Peak, Fitbit Surge, Microsoft Band, Mio Alpha 2, PulseOn, and Samsung Gear S2) were evaluated during varying intervals of physical activity. Participants were selected based on physical characteristics, including age, height, weight, wrist circumference, BMI, skin tone, and fitness level. Sixty healthy, demographically diverse adults ranging from 21 to 64 years of age participated in this study. Participants were from the Stanford, California area, and were distributed equally with 29 males and 31 females. While performing the standardized exercise protocol, participants wore up to four devices and simultaneously underwent continuous electrocardiographic monitoring and indirect calorimetry as FDA-approved gold standard measurements of comparison. The exercise protocol involved five-minute intervals of sitting, walking, fast walking, running, fast running, cycling, and intense cycling. All 60 participants completed the study (Shcherbina et al., 2017).

Results indicated that under laboratory-controlled conditions, six of the wrist-worn devices reported heart rate within five percent error range, while the Samsung Gear S2 achieved a 5.1 percent error rate when measuring heart rate. Error in estimation of energy expenditure was considerably higher than for heart rate for all devices. None of the wrist-worn devices reported energy expenditure within 20 percent error range. Energy expenditure error rates varied from 24 percent for the Fitbit Surge to 97.7 percent for the PulseOn. Researchers concluded that in a diverse group of individuals, heart rate measurements were within acceptable error range,

while energy expenditure estimates were not. Of the seven devices tested, the Apple Watch had the most favorable error profile, while the Samsung Gear S2 had the least favorable error profile. Strengths of this study were the highly diverse sample of participants with different ages, BMIs, and skin tones, the use of numerous consumer and gold standard comparison devices, the standardized exercise protocol to stimulate low and high intensity, and examining several different activity domains collected by the devices. Although validating wrist-worn devices in a laboratory-controlled setting was a strong starting point, it was also a limitation of this study because the results cannot be generalized to free-living conditions as they are intended for. The findings of this study were consistent with previous validity studies. Wallen, Gomersall, Keating, Wisloff, and Coombes assessed the accuracy of heart rate watches, in which heart rate error was within one to nine percent of reference standards, while energy expenditure estimates differed by 43 percent from reference standards (2016). Findings of this study add to the literature on wearable devices by developing error models and proposing a clinical standard for acceptable error. The validation data is important for researchers, as well as consumers and practitioners interested in the clinical application of wrist-worn activity monitors (Shcherbina et al., 2017).

Ferguson et al. assessed the concurrent validity of a selection of consumer-grade accelerometer-based activity monitors against two research-grade multi-sensor accelerometers in free-living conditions (2015). Seven activity monitors, including the Fitbit One, Fitbit Zip, Jawbone UP, Misfit Shine, Nike Fuelband, Striiv Smart Pedometer, and Withings Pulse were compared to two research-grade devices, the BodyMedia SenseWear and ActiGraph GT3X+ over a period of 48 hours. Twenty-one healthy adults aged 20 to 59 years, comprised of 10 males and 11 females from South Australia participated in this study. Demographic data

including date of birth, gender, height, weight, and dominant hand were obtained from all participants for device initialization and calibration. All nine devices were fitted to the participant in the following locations: BodyMedia SenseWear on the left upper arm; Nike Fuelband, Jawbone UP, and Misfit Shine on the left wrist; and ActiGraph GT3X+, Fitbit One, Fitbit Zip, Withings Pulse, and Striiv Smart Pedometer on the right side of the waist on an elasticized belt. Participants were instructed to wear all nine activity monitors simultaneously. To ensure free-living conditions were represented, the wear period was not limited to weekdays or weekends, and guidelines were not provided regarding physical activity or sleep. Over the 48-hour wear period, step counting, moderate-to-vigorous intensity physical activity, sleep duration, and total daily energy expenditure data was collected and quantified. All 21 participants completed the study (Ferguson et al., 2015).

Results indicated that all consumer-grade activity monitors showed strong validity for the measurement of steps ($r = 0.94 – 0.99$) and sleep duration ($r = 0.82 – 0.92$), and moderate validity for the measurement of moderate-to-vigorous intensity physical activity ($r = 0.52 – 0.91$) and total daily energy expenditure ($r = 0.74 – 0.81$). However, the validity of the devices varied considerably within each activity construct. For example, the Misfit Shine undercounted the measurement of moderate-to-vigorous intensity physical activity (mean = 53.3 minutes) compared to the ActiGraph GT3X+ (mean = 58.5 minutes), while the Striiv Smart Pedometer overcounted the measurement of moderate-to-vigorous intensity physical activity (mean = 249 minutes) compared to the ActiGraph GT3X+. Researchers concluded that in free-living conditions, the Fitbit One, Fitbit Zip, and Withings Pulse were the strongest performers compared to the reference devices. Strengths of this study include the use of numerous consumer and reference devices, testing the devices in free-living conditions as they are designed

for, and examining several different activity variables collected by the devices. Limitations of this study were the small sample size, and varying validity if activity monitors are worn in locations other than the hip or wrist. The findings of this study were consistent with previous validity studies, which have similarly found Fitbit activity monitors to be highly valid for measuring step counts in healthy subjects. Although, the scientific evaluation of these devices is a challenge due to the rapidly evolving activity monitor market (Ferguson et al., 2015).

Activity Monitor Utility

A randomized controlled trial by Cadmus-Bertram, Marcus, Patterson, Parker, and Morey evaluated the feasibility and efficacy of integrating a Fitbit tracker and website into a physical activity intervention for postmenopausal women (2015). Fifty-one overweight or obese, inactive, postmenopausal women from the San Diego, California area participated in this study. Participants attended three appointments to receive baseline assessments, orientation, and final assessments. At the second visit, participants were randomized to a 16-week intervention group. Each participant received either a Fitbit One or a basic pedometer with printed materials and a goal-setting process. The Fitbit group received additional software installation and usage training, and a follow-up call after four weeks to evaluate progress. Both groups were asked to wear the Fitbit One or pedometer every day throughout the 16-week intervention period (112 prescribed days), walk 10,000 steps per day, and perform 150 minutes of moderate-to-vigorous intensity physical activity per week (Cadmus-Bertram et al., 2015).

Compared to baseline measurements, the Fitbit group significantly increased physical activity by 789±1,979 steps per day ($p=0.01$), 38±83 minutes of moderate-to-vigorous intensity physical activity in 10-minute bouts ($p=0.008$), and a total of 62±108 minutes of moderate-to-vigorous intensity physical activity per week ($p<0.001$). The pedometer group experienced non-

significant increases in physical activity. Feedback indicated that participants were most engaged with the Fitbit tracker, while participants were least engaged with the website. One hundred percent of women reported liking the Fitbit One, wearing it on 95 percent of intervention days. Ninety-six percent of women rated the Fitbit One as helpful, opposed to 32 percent of women who rated the pedometer as helpful. Researchers concluded that the Fitbit intervention was associated with increased steps and physical activity at 16 weeks, while no change was observed in the pedometer group. The Fitbit One was well-accepted in this sample of women, contributing to the significant increase in physical activity. The 62-minute increase of moderate-to-vigorous intensity physical activity per week observed in the Fitbit group was substantial, especially if maintained over time. Researchers believe that physical activity interventions can be strengthened by leveraging consumer technologies that align with behavior change theories. Strengths of this study include the use of baseline and final questionnaires for detailed participant feedback, use of the ActiGraph GT3X+ as a reference device during baseline and final assessments, and use of Fitbit data to corroborate adherence. Limitations include a small sample size, short intervention period, and lack of generalizability. This study's findings differ from previous activity monitor utility studies. Thompson, Kuhle, Koepp, McCrady-Spitzer, and Levine found a Fitbit with feedback did not increase physical activity among older adults (2014). The different findings may be attributed to the age range of 16-years between participants in this study (Cadmus-Bertram et al., 2015).

O'Brien et al. used a cross-sectional study design to explore the utility of a non-commercial activity monitor to characterize activity profiles in late life depression (2017). A total of 59 subjects over the age of 60 from northeast England participated in this study. Twenty-nine subjects fulfilled Diagnostic and Statistical Manual of Mental Disorders criteria for current

major depression. Thirty subjects of equivalent age without self-reported history of depression or current depression comprised the control group. Each participant received a wrist-worn activity monitor and underwent neuropsychological testing over the seven-day intervention period. Additionally, demographic information, current medications, mood evaluation, social functioning, quality of life, activities of daily living, physical, and mental wellbeing were assessed at baseline and day seven. Due to less than seven days of battery life, the initial wrist-worn activity monitor was switched with an identical, fully charged device between days two and six (O'Brien et al., 2017).

Results indicated that physical activity was significantly reduced in participants with late life depression compared to healthy controls ($p<0.001$). The difference in activity levels between the groups was greatest during the morning and early afternoon. Furthermore, participants with late life depression showed significantly slower fine motor movements ($p<0.001$), lower quality of life scores ($p<0.001$), and reduced activities of daily living ($p<0.001$) compared to healthy controls. Researchers concluded that quality of life and activities of daily living measures were strongly correlated with physical activity, while self-reported measures of loneliness and social support were not. High resolution analysis of accelerometer-derived physical activity may provide an appropriate indication of depression in older adults. Lastly, since exercise has been proposed as a treatment for individuals with depression, wearable devices may play a positive role in monitoring levels of activity when used therapeutically. Strengths of this study include the use of an unobtrusive, waterproof wrist-worn activity monitor, a cohort of currently depressed older adults, and high compliance. A limitation of this study was that causality between physical activity and other key variables could not be determined due to the cross-sectional study design. Likewise, the association between physical activity, depression,

and cognition may be interrelated. This was the first study to objectively characterize the quality of physical activity in late life depression, suggesting that when used therapeutically, wearable devices have potential to objectively monitor levels of activity (O'Brien et al., 2017).

Discussion

Both validity studies simultaneously compared seven commercial activity monitors to research-grade devices. Ferguson et al. found Fitbit activity monitors to be of highest validity, while Shcherbina et al. found the Apple Watch to have the most favorable error profile. The selection of activity monitors was different in both studies, and not all activity monitor functions were validated. When using healthy subjects, the most accurate Fitbit measurement was step counts, while the most accurate measurement for the Apple Watch was heart rate. The environment of these studies also differed from each other; the first study was conducted in a laboratory-controlled setting, while the second study was conducted in free-living conditions. When comparing activity monitor utility studies, both concentrated on older adult populations. The third study discussed used a Fitbit One, while the fourth study used a non-commercial activity monitor. Cadmus-Bertram et al. found that a Fitbit intervention was associated with significantly increased physical activity in postmenopausal, overweight or obese women compared to a pedometer intervention. O'Brien et al. used the activity monitor not as an intervention, but to assess activity levels, and found a significant reduction in general physical activity, lower quality of life, and reduced activities of daily living in depressed older adults. The results of the fourth study discussed could lead into another experimental study utilizing a commercial activity monitor as an intervention tool to increase physical activity in depressed older adults. Although both outcomes are in favor of activity monitor usage, both studies lacked evidence of long-term health benefits.

Conclusion

The purpose of this literature review was to critically analyze current evidence on the validity and utility of activity monitors to improve health outcomes in adults. Activity monitor validity studies suggest Fitbits and Apple Watches are high quality commercial devices. In healthy subjects, Fitbits count steps most accurately, and Apple Watches precisely measure heart rate. Utility studies suggest activity monitors are motivating devices that can increase physical activity, may support other positive health outcomes, and may be potential indicators used in the detection of disease. Future research must expand to keep up with the rapidly evolving activity monitor market. To improve the quality of wearable devices, all commercially-available activity monitors should be regulated by predefined boundaries of accuracy based on reference device standards (Chowdhury, Western, Nightingale, Peacock, & Thompson, 2017). Further research should evaluate which activity monitor features are most effective, examine all health outcomes associated with utilizing activity monitors, and determine which populations are most receptive to activity monitors.

Activity monitors can be useful additions to healthy lifestyle interventions. Activity monitor functions also have the potential to reach reference device standards. As research and product development progress, activity monitors may have practical clinical applications for all populations. Healthcare professionals could provide evidence-based recommendations regarding appropriate activity monitor selection. With automatic data transfer of validated activity monitors, healthcare professionals could accurately track clients' physical activity habits, heart rate, energy expenditure, sleep patterns, food and beverage consumption, and weight fluctuations. Direct access to such data would save an immense amount of time, provide a holistic picture of health for each client, and enable healthcare professionals to personalize care

more than ever before. Future technologies may be able to provide an early warning of disease, aid in diagnosis and treatment, and contribute to a deeper understanding of human health (Savage, 2017). Thus, the individualized approach to care by means of utilizing activity monitor data will result in better health outcomes for all populations, ultimately leading to a decreased prevalence of overweight and obesity, chronic illnesses, mortality, and healthcare costs.

CHAPTER III: METHODOLOGY

The AND's EAL is an online resource synthesizing the strongest, most relevant nutritional research on important dietetic practice questions. The EAL is a series of systematic reviews developed by Academy members for Academy members. Expert Academy members use a predefined approach and criteria to document each step, ensuring objectivity, transparency, and reproducibility of the Evidence Analysis process. The EAL provides bibliographies, evidence summaries, worksheets, conclusion statements, and grades, as well as recommendations, recommendation strength and narrative, algorithms, and links to evidence. The EAL enhances the credibility of the dietetics profession by assisting dietetic practitioners in utilizing evidence-based practice (Academy of Nutrition and Dietetics, 2018). This Evidence Analysis project follows the five steps of the Evidence Analysis process, described below.

Step One: Formulate the Evidence Analysis Question

The first step in the Evidence Analysis process focuses on a specific question in a defined area of practice. High quality research questions are developed using an analytical framework to identify links between factors and outcomes. The AND's NCP should serve as the framework, while the PICO format should be used to format questions. PICO is an acronym for population, intervention, comparison, and outcome (Academy of Nutrition and Dietetics, 2016). The research questions in this Evidence Analysis project are as follows.

1. How valid are the functions of current, consumer-grade activity monitors compared to research-grade devices?
2. Does physical activity improve and weight loss occur as a result when adults utilize current, consumer-grade activity monitors?

Step Two: Gather and Classify the Evidence

After the Evidence Analysis question is formulated, research must be gathered and classified. This step involves creating a search plan to conduct a thorough literature search. Ensuring that all relevant evidence is reviewed is one of the most important aspects of the Evidence Analysis process. If some evidence is missed, the conclusion statement may be misleading. The search plan should delineate the inclusion and exclusion criteria, key search terms or phrases, and outcomes necessary to conduct a thorough literature search. Research is classified by type of evidence, with classes differentiating between primary (Class A, B, C, and D) and secondary (Class M, R, and X) reports (Academy of Nutrition and Dietetics, 2016). Research articles that meet the predefined criteria will be included for further evaluation. Articles that do not meet the criteria will be excluded along with appropriate reasoning. This Evidence Analysis project adheres to the following search plan.

Research Question	1. How valid are the functions of current, consumer-grade activity monitors compared to research-grade devices? 2. Does physical activity improve and weight loss occur as a result when adults utilize current, consumer-grade activity monitors?
Date of Literature Review	September 2018
Inclusion Criteria	• Language: English • Research: primary (Class A, B, C, and D) • Year Range: 2014 to 2018 • Participant Age Range: adults 18 to 80 years • Participant Health Status: healthy or with one health condition related to diet and/or physical inactivity
Exclusion Criteria	• Language: all languages that are not English • Research: secondary (Class M, R, and X) • Year Range: prior to 2014 • Participant Age Range: children less than 18 years and adults greater than 80 years • Participant Health Status: with more than one health condition that is related or unrelated to diet and/or physical inactivity

Search Terms/Phrases	• Activity monitor • Fitness tracker
Electronic Databases	• PubMed
Included Articles	• Benedetto, S., Caldato, C., Bazzan, E., Greenwood, D., Pensabene, V., & Actis, P. (2018). Assessment of the Fitbit Charge 2 for monitoring heart rate. *PLoS ONE*, *13*(2), e0192691. doi:10.1371/journal.pone.0192691 • Cadmus-Bertram, L., Gangnon, R., Wirkus, E. J., Thraen-Borowski, K. M., & Gorzelitz-Liebhauser, J. (2017). The Accuracy of Heart Rate Monitoring by Some Wrist-Worn Activity Trackers. *Annals of Internal Medicine*, *166*(8), 610–612. doi:10.7326/L16-0353 • Cadmus-Bertram, L., Marcus, B., Patterson, R., Parker, B., & Morey, B. (2015). Randomized Trial of a Fitbit-Based Physical Activity Intervention for Women. *American Journal of Preventive Medicine*, *49*(3), 414–418. doi:10.1016/j.amepre.2015.01.020 • Chum, J., Kim, M., Zielinski, L., Bhatt, M., Chung, D., Yeung, S., … Samaan, Z. (2017). Acceptability of the Fitbit in behavioural activation therapy for depression: a qualitative study. *Evidence-Based Mental Health*, *20*(4), 128–133. doi:10.1136/eb-2017-102763 • Cook, J., Prairie, M., & Plante, D. (2017). Utility of the Fitbit Flex to Evaluate Sleep in Major Depressive Disorder: A comparison against polysomnography and wrist-worn actigraphy. *Journal of Affective Disorders*, *217*, 299–305. doi: 10.1016/j.jad.2017.04.030 • Ferguson, T., Rowlands, A., Olds, T., & Maher, C. (2015). The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. *International Journal of Behavioral Nutrition and Physical Activity, 12*, 42. doi:10.1186/s12966-015-0201-9 • Gomersall, S., Ng, N., Burton, N., Pavey, T., Gilson, N., & Brown, W. (2016). Estimating Physical Activity and Sedentary Behavior in a Free-Living Context: A Pragmatic Comparison of Consumer-Based Activity Trackers and ActiGraph Accelerometry. *Journal of Medical Internet Research*, *18*(9), e239. doi:10.2196/jmir.5531 • Gualtieri, L., Rosenbluth, S., & Phillips, J. (2016). Can a Free Wearable Activity Tracker Change Behavior? The Impact of Trackers on Adults in a Physician-Led Wellness Group. *JMIR Research Protocols*, *5*(4), e237. doi:10.2196/resprot.6534 • Naslund, J., Aschbrenner, K., Scherer, E., McHugo, G., Marsch, L., & Bartels, S. (2016). Wearable Devices and Mobile Technologies for Supporting Behavioral Weight Loss

	Among People with Serious Mental Illness. *Psychiatry Research*, *244*, 139–144. doi:10.1016/j.psychres.2016.06.056 • Maher, C., Ryan, J., Ambrosi, C., & Edney, S. (2017). Users' experiences of wearable activity trackers: a cross-sectional study. *BMC Public Health*, *17*, 880. doi:10.1186/s12889-017-4888-1 • Rosenberger, M., Buman, M., Haskell, W., McConnell, M., & Carstensen, L. (2016). 24 Hours of Sleep, Sedentary Behavior, and Physical Activity with Nine Wearable Devices. *Medicine and Science in Sports and Exercise*, *48*(3), 457–465. doi:10.1249/MSS.0000000000000778 • Shcherbina, A., Mattsson, C., Waggott, D., Salisbury, H., Christle, J., Hastie, T., … Ashley, E. (2017). Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine*, *7*(2), 3. doi:10.3390/jpm7020003
Excluded Articles with Reason	• Abrantes, A., Blevins, C., Battle, C., Read, J., Gordon, A., & Stein, M. (2017). Developing a Fitbit-Supported Lifestyle Physical Activity Intervention for Depressed Alcohol Dependent Women. *Journal of Substance Abuse Treatment*, *80*, 88–97. doi:10.1016/j.jsat.2017.07.006 o Participants used two consumer-grade activity monitors inconsistently o Assessed utility of activity monitors in participants with more than one health condition • Berendsen, B., Hendriks, M., Meijer, K., Plasqui, G., Schaper, N., & Savelberg, H. (2014). Which activity monitor to use? Validity, reproducibility and user friendliness of three activity monitors. *BMC Public Health*, *14*, 749. doi:10.1186/1471-2458-14-749 o Participants did not use a consumer-grade activity monitor • Cochrane, S., Chen, S., Fitzgerald, J., Dodson, J., Fielding, R., King, A., … Kaplan, R. (2017). Association of Accelerometry-Measured Physical Activity and Cardiovascular Events in Mobility-Limited Older Adults: The LIFE (Lifestyle Interventions and Independence for Elders) Study. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, *6*(12), e007215. doi:10.1161/JAHA.117.007215 o Participants did not use a consumer-grade activity monitor • Correa, J., Apolzan, J., Shepard, D., Heil, D., Rood, J., & Martin, C. (2016). Evaluation of the ability of three physical activity monitors to predict weight change and estimate

	energy expenditure. *Applied Physiology, Nutrition, and Metabolism*, *41*(7), 758–766. doi:10.1139/apnm-2015-0461 o Assessed predictability of body weight changes • LaMonte, M., Lewis, C., Buchner, D., Evenson, K., Rillamas-Sun, E., Di, C., … Shumaker, S. (2017). Both Light Intensity and Moderate-to-Vigorous Physical Activity Measured by Accelerometry Are Favorably Associated With Cardiometabolic Risk Factors in Older Women: The Objective Physical Activity and Cardiovascular Health (OPACH) Study. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, *6*(10), e007064. doi:10.1161/JAHA.117.007064 o Participants did not use a consumer-grade activity monitor • O'Brien, J., Gallagher, P., Stow, D., Hammerla, N., Ploetz, T., Firbank, M., . . . Olivier, P. (2017). A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression. *Psychological Medicine, 47*(1), 93-102. doi:10.1017/S0033291716002166 o Participants did not use a consumer-grade activity monitor
Summary of Articles Identified to Review	Number of Primary Research Articles Identified: 18 Number of Primary Research Articles Excluded: 6 Total Number of Primary Research Articles Included: 12

Step Three: Critically Appraise Each Article

Research articles that meet the inclusion criteria are evaluated for methodologic quality in this step. Articles are appraised individually based on the appropriateness of the study design and the quality of how the study was conducted by using the AND's Quality Criteria Checklist (QCC) and EAL worksheets. Key information is abstracted and entered into the worksheets. The QCC worksheet assesses risk of bias by asking questions related to relevance and validity, while determining if the strength of evidence overall is rated as negative, neutral, or positive quality. The EAL worksheet summarizes the methodology, results, authors' conclusion, and reviewer's comments (Academy of Nutrition and Dietetics, 2016). Studies with negative quality

evidence were not included or used in the development of either conclusion statement in this Evidence Analysis project.

Step Four: Summarize the Evidence

After the research has been critically appraised, evidence from each article is summarized and displayed. This step involves the creation of an overview table and an evidence summary. The overview table includes key information such as study designs, sample sizes, interventions, outcomes, and quality ratings, as well as citations for the included research articles. Data for the overview table is transferred from the QCC and EAL worksheets. The overview table is beneficial because it allows healthcare professionals to visually compare the studies. The evidence summary is a concise description of the overall findings of the Evidence Analysis process (Academy of Nutrition and Dietetics, 2016). Activity monitor validity and utility data from neutral and positive quality studies has been summarized and displayed in an overview table and evidence summary.

Step Five: Write and Grade the Conclusion Statement

The final step in the Evidence Analysis process involves establishing a conclusion statement to answer the research question, along with assigning a grade to the conclusion statement. The grade reflects the overall strength of the available supporting evidence in forming the conclusion statement. The grading scale used by the AND includes: Grade I (good/strong), II (fair), III (limited/weak), IV (expert opinion only), and V (not assignable) (Academy of Nutrition and Dietetics, 2016). Conclusion statements for both research questions in this Evidence Analysis project have been established based on the available supporting evidence. The overall strength of the supporting evidence has also been graded, ranging from Grade I to V.

Supplemental Comparison Report

Following the completion of the Evidence Analysis process, product details of 10 current, best-selling, commercially-available activity monitors will be summarized and displayed in a supplemental comparison report to fulfill objective C of this Evidence Analysis project. Data has been collected exclusively from product manufacturer websites to maintain the EAL's standards of objectivity, transparency, and reproducibility. If evidence is deemed adequate and conclusive, dietitians and other healthcare professionals may draft evidence-based guideline recommendations regarding appropriate activity monitor selection.

CHAPTER IV: RESULTS

Using the PubMed database, a total of 18 studies were identified relating to activity monitor validity and utility in adults. Six studies were excluded because they assessed participants with more than one health condition, the predictability of body weight changes, or activity monitors that were not consumer-grade devices. Twelve studies were included for further evaluation. Of the 12 studies, seven were validity studies and five were randomized controlled trials, non-randomized crossover trials, and a cross sectional study relating to activity monitor utility. An overview table summarizing and displaying key information from each study can be found in Appendix B. Below are brief summaries of each study, followed by the graded conclusion statements that answer both research questions.

Benedetto et al., (2018) (Quality Rating: +)

Benedetto et al., (2018) assessed in a controlled, research environment the accuracy and precision of the Fitbit Charge 2 for measuring heart rate with respect to the ProComp Infiniti T7500M, a gold standard electrocardiograph. The Fitbit Charge 2 exhibited a mean bias of -5.9 beats per minute (95% CI). The limits of agreement, which indicate the precision of individual measurements, between the Fitbit Charge 2 and ProComp Infiniti T7500M were wide. The upper limit of agreement was +16.8 beats per minute, whereas the lower limit of agreement was -28.5 beats per minute. The intraclass correlation coefficient, used as an alternative measure of agreement between the Fitbit Charge 2 and ProComp Infiniti T7500M was 0.21 (95% CI). Researchers concluded that the Fitbit Charge 2 tends to underestimate heart rate with moderate bias on average, although precision is poor for individual measurements, which could be underestimated by as much as 30 beats per minute.

Cadmus-Bertram et al., (2017) (Quality Rating: ø)

Cadmus-Bertram et al., (2017) determined the heart rate accuracy measured by four commercial, light-emitting diode-dependent, wrist-worn activity trackers. When participants rested, the limits of agreement were best for the Fitbit Surge (−5.1 to 4.5 beats per minute) and worst for the Basis Peak (−17.1 to 22.6 beats per minute). When participants exercised at 65 percent of their maximum heart rate, the limits of agreement were relatively poor for all the activity trackers, ranging from −22.5 to 26.0 beats per minute for the Mio Fuse to −41.0 to 36.0 beats per minute for the Fitbit Charge. At rest, the repeatability coefficient ranged from 4.2 beats per minute for the Fitbit Surge to 19.3 beats per minute for the Basis Peak. During exercise, the repeatability coefficient ranged from 20.2 beats per minute for the Basis Peak to 23.7 beats per minute for the Mio Fuse. Researchers concluded that all activity trackers were more accurate at rest than during moderate exercise, implying that more heart rate feature research is needed.

Cadmus-Bertram et al., (2015) (Quality Rating: +)

Cadmus-Bertram et al., (2015) evaluated within a randomized controlled trial, the feasibility and preliminary efficacy of integrating a Fitbit tracker and website into a physical activity intervention for postmenopausal, overweight or obese women. After the 16-week intervention, the Fitbit group significantly increased physical activity by 789±1,979 steps per day ($p=0.01$), moderate-to-vigorous intensity physical activity in 10-minute bouts by 38±83 minutes per week ($p=0.008$), and moderate-to-vigorous intensity physical activity by 62±108 minutes per week ($p<0.001$), compared to non-significant increases in the pedometer group (between-group p-values were 0.11, 0.28, and 0.30, respectively). The Fitbit group wore the Fitbit One on 95 percent of intervention days, 96 percent of women reported liking the website, and 100 percent

of women reported liking the Fitbit One. Although the study had confounding variables, researchers concluded that the Fitbit One was well-accepted in this sample of women, was associated with increased steps and physical activity at 16 weeks, and that physical activity interventions can be strengthened by leveraging consumer technologies that align with behavior change theories.

Chum et al., (2017) (Quality Rating: ø)

Chum et al., (2017) aimed to understand patients' perceived benefit from the Fitbit One and explore themes associated with patient experiences, as well as compare the perceived benefit, patient factors, Fitbit usage, and Beck's Depression Inventory scores. Of the 36 patients who underwent the BRAVE study and completed interviews, 23 patients found the Fitbit One to be helpful for their physical activity. Themes of positive experiences included self-awareness, peer motivation, and goal-setting opportunities. Themes of negative experiences included inconvenience, inaccuracies, discouragement, and disinterest. There was a significant relationship between total Fitbit One usage and perceived benefit. The mean number of weeks of Fitbit One use for those who found the Fitbit helpful was 18.57 and 12.27 weeks for those who did not ($p<0.001$). Interestingly, there was no significant relationship between perceived benefit of the Fitbit One and percent change in Beck's Depression Inventory scores, which contradicts previous literature supporting Fitbit use in treating depression. Researchers concluded that the Fitbit One may be useful for patients with varying characteristics, although strengths and limitations of activity trackers should be considered when implementing them to motivate patients with depression.

Cook et al., (2017) (Quality Rating: +)

Cook et al., (2017) evaluated the utility of the Fitbit Flex to estimate sleep in a well-characterized cohort of adult patients with major depressive disorder relative to gold standard polysomnography and validated actigraphy (Actiwatch-2; AW-2). Compared to polysomnography, the Fitbit Flex significantly overestimated total sleep time (mean difference of 46.0 minutes, $p<0.0001$) and sleep efficiency (mean difference of 8.1%, $p<0.0001$), significantly underestimated wake after sleep onset (mean difference of −44.0 minutes, $p<0.0001$), while sleep onset latency was quite similar (mean difference of −2.0 minutes, $p=0.72$). The Fitbit Flex demonstrated high sensitivity and accuracy with low specificity. Researchers concluded that the Fitbit Flex is not an adequate substitute for polysomnography when quantifying sleep in adults with major depressive disorder. However, the Fitbit Flex does demonstrate similar performance characteristics to a standard actigraph.

Ferguson et al., (2015) (Quality Rating: +)

Ferguson et al., (2015) assessed the concurrent validity of a selection of consumer-grade, accelerometer-based activity monitors compared to two research-grade accelerometers in free-living conditions. All activity monitors measured steps, and correlations with reference devices were very strong (r=0.94-0.99). Five activity monitors measured moderate-to-vigorous intensity physical activity, and correlations ranged from moderate to strong (r=0.52-0.91). Four activity monitors measured sleep, and all correlated strongly with the reference device (r=0.82-0.92). Five activity monitors measured total daily energy expenditure, and correlations were moderate to strong (r=0.74-0.81). Researchers concluded that the Fitbit One, Fitbit Zip, and Withings Pulse were the strongest performers.

Gomersall et al., (2016) (Quality Rating: ø)

Gomersall et al., (2016) compared Fitbit One and Jawbone UP estimates of steps, moderate-to-vigorous intensity physical activity, and sedentary behavior with data from the ActiGraph GT3X+ accelerometer in a free-living context. Correlations for steps and moderate-to-vigorous intensity physical activity were strong for both devices, although higher for the Fitbit One ($r=0.85$ for steps and $p=0.80$ for moderate-to-vigorous intensity physical activity) than for the Jawbone UP ($r=0.75$ for steps and $p=0.75$ for moderate-to-vigorous intensity physical activity). The correlation between the Jawbone UP longest idle time and ActiGraph longest sedentary bout was weak ($p=0.19$). Agreement between the Fitbit One and ActiGraph for the classification of active versus inactive time was substantial ($k=0.68$, $p<0.001$), while agreement between the Jawbone UP and ActiGraph was moderate ($k=0.52$, $p<0.001$). Due to modest accuracy and systematic bias, researchers concluded that both activity trackers are better suited as self-monitoring tools for consumers or in behavior change interventions rather than for the evaluation of research outcomes.

Gualtieri et al., (2016) (Quality Rating: ø)

Gualtieri et al., (2016) investigated the use of wearable activity trackers by adults with chronic medical conditions who have never used trackers previously. Specifically, the researchers aimed to determine (1) if participants would accept and use activity trackers to increase their physical activity; (2) if there were barriers to use besides cost and training; (3) if activity trackers would educate participants on their activity levels and support behavior change; and (4) if clinical outcomes would show improvements in participants' health. Improvements were seen in clinical outcomes, physical activity behaviors, and attitudes towards the Withings Pulse after the 12-week study. Participants lost an average of 0.5 pounds per week with a mean total weight loss of 5.97 pounds ($p=0.004$). Other clinical outcomes included a 9.2 percent

decrease in LDL levels (p=0.038), while changes in blood pressure were non-significant. All participants reported an increase in well-being, health education, physical activity, and confidence in their ability to lead more active lives. Researchers concluded that adding activity trackers to wellness groups can support education and behavior change, adding that it may be cost-effective to provide free or heavily subsidized activity trackers that lower the risk of chronic conditions compared to the healthcare costs required to treat illnesses after they develop.

Maher et al., (2017) (Quality Rating: +)

Maher et al., (2017) explored users' experiences of activity trackers, including the ease of use, patterns of use, barriers to use, and perceived usefulness for tracking and modifying lifestyle behaviors, such as physical activity, diet, and sleep. The most commonly used brand of activity tracker was Fitbit (67.5%), followed by Garmin (16.5%), Apple (3.4%), Jawbone (2.5%), Samsung (1.7%), Polar (1.3%), and other (7.1%). Participants agreed that various features on their activity trackers were useful, including step counts (95%), active minutes (76%), sleep (66%), heart rate (63%), elevation or stairs climbed (58%), and calories burned (57%), while fewer participants agreed that the food intake feature was useful (36%). Overall, 94 percent of current users and 65 percent of former users agreed that they had a positive experience with their activity tracker. Researchers concluded that in general, activity trackers are used for a substantial period of time, are viewed positively by users, and are useful tools for intervening on physical activity.

Naslund et al., (2016) (Quality Rating: +)

Naslund et al., (2016) examined whether average daily step counts measured using Fitbit Zip wearable devices was associated with weight loss and improved fitness among individuals

with serious mental illness enrolled in a six-month lifestyle program. Participants accumulated an average of 4,453.5 steps per day, with average daily step counts ranging from 1,037.6 to 11,366.3 steps. There was a significant association between participants' average daily step counts and weight loss. For every 1,000-step increase, participants experienced a decrease in weight of 1.78 pounds (p=0.0314). The relationship between average daily step counts and change in fitness was non-significant. Every 1,000-step increase corresponded to an increase of 18.79 feet on the Six-Minute Walk Test (p=0.176). Researchers concluded that wearable devices and their associated smartphone applications may serve as valuable tools for supporting weight loss efforts in individuals with serious mental illness. Additionally, providing participants with the recommendation to maintain a high average daily step count throughout participation in a lifestyle intervention may contribute to greater weight loss.

Rosenberger et al., (2016) (Quality Rating: ø)

Rosenberger et al., (2016) compared the output from commercially available wearable devices using current standards for objective measurement of sleep, sedentary behavior, light-intensity physical activity, moderate-to-vigorous intensity physical activity, and steps in a free-living environment. Mean error analyses for the devices ranged from 8.1 percent for the Actigraph GT3X+ to 16.9 percent for the GENEactiv when measuring sleep duration; 9.5 percent from the LUMOback to 65.8 percent for the GENEactiv when measuring sedentary behavior; 19.7 percent from the GENEactiv to 28.0 percent for the Fitbit One when measuring light-intensity physical activity; 51.8 percent from the Jawbone Up to 92.0 percent for the Nike Fuelband when measuring moderate-to-vigorous intensity physical activity; and 14.1 percent from the Actigraph GT3X+ to 29.9 percent for the Nike Fuelband when measuring total steps per day. Equivalence analyses indicated only one comparison device, the LUMOback was

significantly equivalent to standards when measuring sedentary behavior (90% CI). Researchers concluded that none of the commercial wearable devices provide all measures of the 24-hour activity model, which is currently only possible with research-grade devices, suggesting that the future of activity measurement should aim for accurate 24-hour measurement as a goal.

Shcherbina et al., (2017) (Quality Rating: +)

Shcherbina et al., (2017) assessed under controlled laboratory conditions the accuracy of seven commercially available wrist-worn devices in estimating heart rate and energy expenditure. The lowest error in measuring heart rate (1.8%) was observed for the cycling stage (0.9%-2.7%, 95% CI), while the highest error in measuring heart rate (5.5%) was observed for the walking stage (3.9%-7.1%, 95% CI). Error in estimation of energy expenditure was considerably higher than for heart rate for all devices. Median error rates across activities varied from 27.4 percent (24.0%-30.8%, 95% CI) for the Fitbit Surge to 92.6 percent (87.5%-97.7%, 95% CI) for the PulseOn. Researchers concluded that most wrist-worn devices reported heart rate within acceptable error range (5%), while none of the wrist-worn devices reported energy expenditure within an acceptable error range. The Apple Watch had the most favorable error profile and the Samsung Gear S2 had the least favorable error profile.

Research Question

1. Current, consumer-grade activity monitor functions will be within 25 percent error range of research-grade devices

Conclusion Statement

Current, consumer-grade activity monitors exhibit moderate validity on average, tend to estimate step counts accurately, underestimate heart rate and energy expenditure, overestimate time asleep, and are more accurate at rest than during activity.

Grade: II, Fair

This is a Grade II, Fair conclusion because the evidence consists of results from studies that answered the research question addressed, although there is uncertainty about generalizability, bias, research design flaws, and adequacy of sample sizes.

Research Question

2. Does physical activity improve and weight loss occur as a result when adults utilize current, consumer-grade activity monitors?

Conclusion Statement

Adults who utilize current, consumer-grade activity monitors as combined interventions may experience a clinically meaningful increase in steps, physical activity, and weight loss.

Grade: II, Fair

This is a Grade II, Fair conclusion because the evidence consists of results from studies that answered the research question addressed, although there is uncertainty about generalizability, bias, research design flaws, and adequacy of sample sizes.

CHAPTER V: CONCLUSION

Evidence Summary

Current, consumer-grade activity monitors exhibit moderate validity on average and provide more precise measurements at rest compared to during activity. Activity monitors tend to estimate step counts accurately, underestimate heart rate and energy expenditure, and overestimate time asleep. The Apple Watch, Fitbits, and Withings Pulse were found to be the most valid, while the Basis Peak, Jawbone UP, Nike Fuelband, PulseOn, and Samsung Gear S2 were found to be the least valid consumer-grade devices. Adults who utilize activity monitors as combined interventions may experience a clinically meaningful increase in steps, physical activity, and weight loss. Overall, activity monitors are recognized as useful additions to healthy lifestyle interventions.

Of the seven validity studies, five assessed activity monitor accuracy in healthy subjects, one evaluated participants with major depressive disorder, and one did not specify the study population beyond reporting that participants were adults. Of the five studies exploring activity monitor utility, two were randomized controlled trials, two were non-randomized crossover trials, and one was a cross sectional study. Populations included in the activity monitor utility studies were postmenopausal women, participants with overweight or obesity, one chronic medical condition, major depressive disorder, serious mental illness, and were current or former activity tracker users. Not all studies analyzed activity monitor validity or utility in improving health outcomes in adults within the same parameters. Four validity studies were conducted under controlled, research environments utilizing gold standard comparisons, while three validity studies were conducted in free-living environments as activity monitors are intended for, utilizing common field-based devices as a comparison. Utility studies did not evaluate activity

monitors as independent interventions. Instead, most utility studies incorporated activity monitors as self-motivating, self-monitoring tools integrated into behavioral interventions with positive outcomes.

Conclusion

Research hypotheses were relatively similar to the conclusion statements developed for this Evidence Analysis project. Activity monitor utility studies revealed that adults did in fact experience an increase in physical activity and a decrease in weight as a result. However, exact error profiles could not be determined due to the varied statistical analyses and methods of reporting results by activity monitor validity studies. Interestingly, Chum et al. found no significant relationship between perceived benefit of the Fitbit One and percent change in Beck's Depression Inventory scores, which contradicts previous literature supporting activity monitor use in treating depression (2017). Results from this Evidence Analysis project demonstrate that activity monitors are self-motivating, self-monitoring devices, and their usefulness is not dependent on their validity. Participants were motivated by activity monitors regardless of their accuracy. Although, for research purposes, activity monitors could be more effective if they were more accurate.

Applications to Practice

Activity monitors have current and future applications to practice. Currently, activity monitors create awareness by assisting individuals in tracking their weight-related behaviors. Personal behaviors play a dominant role in preventing and treating noncommunicable diseases, such as obesity, diabetes, cardiovascular disease, and certain cancers. As self-motivating, self-monitoring devices, activity monitors may be used as intervention tools to improve adherence and support positive behavior change. Dietitians and other healthcare professionals may

encourage the use of activity monitors as a cost-effective method to implement healthy lifestyle changes, which in turn would decrease morbidity, mortality, and healthcare costs.

Activity monitors have the potential to reach reference device standards, which in turn may result in further clinical applications for all populations. Dietitians and other healthcare professionals could provide evidence-based recommendations regarding appropriate activity monitor selection. With automatic data transfer of validated activity monitors, healthcare professionals could accurately track their clients' physical activity habits, heart rate, energy expenditure, sleep patterns, food and beverage consumption, and weight fluctuations. Direct access to such data would save an immense amount of time, provide a holistic picture of health for each client, and enable healthcare professionals to personalize care more than ever before. Future wearable technologies may be able to provide an early warning of disease, aid in diagnosis and treatment, and contribute to a deeper understanding of human health (Savage, 2017). Hopefully, the individualized approach to care by means of utilizing validated activity monitor data will result in better health outcomes for all populations, ultimately leading to a decreased prevalence of overweight and obesity, chronic illnesses, mortality, and healthcare costs.

Recommendations for Future Research

It may be speculated that current activity monitor hardware and software used to estimate activity parameters, such as step counts, elevation or stairs climbed, distance traveled, heart rate, calories burned, active time, and time asleep lack proper sophistication. Product developers should consult activity monitor users, healthcare professionals, researchers, and information technology firms to gain an in-depth understanding of current activity monitor performance,

effective methods for integrating activity monitors into clinical care, and expectations for activity monitor successors.

Future research needs to expand in this area. First, research should evaluate which activity monitor features are most effective, determine which populations are most receptive to activity monitors, and examine all health outcomes associated with utilizing activity monitors. Second, research must be more comprehensive with larger sample sizes, demographically diverse subjects, stronger study designs, and longer study durations to determine long-term effects. Third, research must be consistent, if not continuous in order to keep up with the rapidly evolving consumer-grade activity monitor market. To achieve this, researchers could collaborate in the creation of an online forum to update and share evidence objectively. Evidence should also be made public in an easily readable and understandable format to encourage transparency from activity monitor manufacturers. As a final effort to improve the quality of activity monitors and associated health outcomes, predefined boundaries of accuracy based on reference device standards should be regulated.

BIBLIOGRAPHY

Academy of Nutrition and Dietetics. (2018). About the Evidence Analysis Library. Retrieved from https://www.andeal.org/about

Academy of Nutrition and Dietetics. (2017). Definition of Terms List. Retrieved from https://www.eatrightpro.org/-/media/eatrightpro-files/practice/scope-standards-of-practice/academydefinitionoftermslist.pdf

Academy of Nutrition and Dietetics, Evidence Analysis Library. (2015). MNT: Weight Management. Retrieved from https://www.andeal.org/topic.cfm?menu=5284&cat=5230

Academy of Nutrition and Dietetics. (2016). *Evidence Analysis Manual: Steps in the Academy Evidence Analysis Process*. Retrieved from https://www.andeal.org/vault/2440/web/files/2016_April_EA_Manual.pdf

Academy of Nutrition and Dietetics. (2006). MNT Versus Nutrition Education. Retrieved from https://www.eatrightpro.org/payment/coding-and-billing/mnt-vs-nutrition-education

Bellicha, A., Macé, S., & Oppert, J. M. (2017). Prescribing of Electronic Activity Monitors in Cardiometabolic Diseases: Qualitative Interview-Based Study. *Journal of medical Internet research*, *19*(9), e328. doi:10.2196/jmir.8107

Benedetto, S., Caldato, C., Bazzan, E., Greenwood, D., Pensabene, V., & Actis, P. (2018). Assessment of the Fitbit Charge 2 for monitoring heart rate. *PLoS ONE*, *13*(2), e0192691. doi:10.1371/journal.pone.0192691

Cadmus-Bertram, L., Gangnon, R., Wirkus, E. J., Thraen-Borowski, K. M., & Gorzelitz-Liebhauser, J. (2017). The Accuracy of Heart Rate Monitoring by Some Wrist-Worn Activity Trackers. *Annals of Internal Medicine*, *166*(8), 610–612. doi:10.7326/L16-0353

Cadmus-Bertram, L., Marcus, B., Patterson, R., Parker, B., & Morey, B. (2015). Randomized

Trial of a Fitbit-Based Physical Activity Intervention for Women. *American Journal of Preventive Medicine*, *49*(3), 414–418. doi:10.1016/j.amepre.2015.01.020

Chowdhury, E., Western, M., Nightingale, T., Peacock, O., & Thompson, D. (2017). Assessment of laboratory and daily energy expenditure estimates from consumer multi-sensor physical activity monitors. *PLoS ONE*, *12*(2), e0171720. doi:10.1371/journal.pone.0171720

Chum, J., Kim, M., Zielinski, L., Bhatt, M., Chung, D., Yeung, S., … Samaan, Z. (2017). Acceptability of the Fitbit in behavioural activation therapy for depression: a qualitative study. *Evidence-Based Mental Health*, *20*(4), 128–133. doi:10.1136/eb-2017-102763

Cook, J., Prairie, M., & Plante, D. (2017). Utility of the Fitbit Flex to Evaluate Sleep in Major Depressive Disorder: A comparison against polysomnography and wrist-worn actigraphy. *Journal of Affective Disorders*, *217*, 299–305. doi: 10.1016/j.jad.2017.04.030

Ferguson, T., Rowlands, A., Olds, T., & Maher, C. (2015). The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. *International Journal of Behavioral Nutrition and Physical Activity, 12*, 42. doi:10.1186/s12966-015-0201-9

Gomersall, S., Ng, N., Burton, N., Pavey, T., Gilson, N., & Brown, W. (2016). Estimating Physical Activity and Sedentary Behavior in a Free-Living Context: A Pragmatic Comparison of Consumer-Based Activity Trackers and ActiGraph Accelerometry. *Journal of Medical Internet Research*, *18*(9), e239. doi:10.2196/jmir.5531

Gualtieri, L., Rosenbluth, S., & Phillips, J. (2016). Can a Free Wearable Activity Tracker Change Behavior? The Impact of Trackers on Adults in a Physician-Led Wellness Group. *JMIR Research Protocols*, *5*(4), e237. doi:10.2196/resprot.6534

Gudzune, K., Doshi, R., Mehta, A., Chaudhry, Z., Jacobs, D., Vakil, R., … Clark, J. (2015). Efficacy of commercial weight loss programs: an updated systematic review. *Annals of Internal Medicine*, *162*(7), 501–512. doi:10.7326/M14-2238

Hruby, A., & Hu, F. (2015). The Epidemiology of Obesity: A Big Picture. *PharmacoEconomics*, *33*(7), 673–689. doi:10.1007/s40273-014-0243-x

Kim, G., Lin, J., Blomain, E., & Waldman, S. (2014). Anti-Obesity Pharmacotherapy: New Drugs and Emerging Targets. *Clinical Pharmacology and Therapeutics*, *95*(1), 53–66. doi:10.1038/clpt.2013.204

Lewis, Z., Lyons, E., Jarvis, J., & Baillargeon, J. (2015). Using an electronic activity monitor system as an intervention modality: A systematic review. *BMC Public Health*, *15*, 585. doi:10.1186/s12889-015-1947-3

Maher, C., Ryan, J., Ambrosi, C., & Edney, S. (2017). Users' experiences of wearable activity trackers: a cross-sectional study. *BMC Public Health*, *17*, 880. doi:10.1186/s12889-017-4888-1

Naslund, J., Aschbrenner, K., Scherer, E., McHugo, G., Marsch, L., & Bartels, S. (2016). Wearable Devices and Mobile Technologies for Supporting Behavioral Weight Loss Among People with Serious Mental Illness. *Psychiatry Research*, *244*, 139–144. doi:10.1016/j.psychres.2016.06.056

O'Brien, J., Gallagher, P., Stow, D., Hammerla, N., Ploetz, T., Firbank, M., . . . Olivier, P. (2017). A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression. *Psychological Medicine, 47*(1), 93-102. doi:10.1017/S0033291716002166

Office of Disease Prevention and Health Promotion. (2018). 2008 Physical Activity Guidelines

for Americans Summary. Retrieved from https://health.gov/paguidelines/guidelines/summary.aspx

Rosenberger, M., Buman, M., Haskell, W., McConnell, M., & Carstensen, L. (2016). 24 Hours of Sleep, Sedentary Behavior, and Physical Activity with Nine Wearable Devices. *Medicine and Science in Sports and Exercise*, *48*(3), 457–465. doi:10.1249/MSS.0000000000000778

Ross, K. & Wing, R. (2016). Impact of newer self-monitoring technology and brief phone-based intervention on weight loss: a randomized pilot study. *Obesity (Silver Spring)*, *24*(8), 1653–1659. doi:10.1002/oby.21536

Savage, N. (2017). The Measure of a Man. *Cell (Published by Elsevier Inc.), 169*, 1159-1161. doi:10.1016/j.cell.2017.05.037.

Shcherbina, A., Mattsson, C., Waggott, D., Salisbury, H., Christle, J., Hastie, T., … Ashley, E. (2017). Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine*, *7*(2), 3. doi:10.3390/jpm7020003

Thompson, W., Kuhle, C., Koepp, G., McCrady-Spitzer, S., Levine, J. (2014). "Go4Life" exercise counseling, accelerometer feedback, and activity levels in older people. *Archives of Gerontology and Geriatrics*, 58(3), 314-319. doi: 10.1016/j.archger.2014.01.004

Townshend, T., & Lake, A. (2017). Obesogenic environments: current evidence of the built and food environments. *Perspectives in Public Health, 137*(1), 38-44. doi:10.1177/1757913916679860

Wallen, M., Gomersall, S., Keating, S., Wisloff, U., & Coombes, J. (2016). Accuracy of Heart

Rate Watches: Implications for Weight Management. *PLoS ONE*, *11*(5), e0154420. doi:10.1371/journal.pone.0154420

Wisconsin Department of Public Instruction. (2009). Wisconsin's Model Academic Standards for Nutrition. Retrieved from https://dpi.wi.gov/sites/default/files/imce/team-nutrition/pdf/nestandards.pdf

Wisconsin Department of Public Instruction. (2010). Wisconsin Standards for Physical Education. Retrieved from https://dpi.wi.gov/sites/default/files/imce/sspw/pdf/pewiscstandards.pdf

World Health Organization. (2018). Obesity and overweight. Retrieved from http://www.who.int/mediacentre/factsheets/fs311/en/

World Health Organization. (2018). Physical Activity. Retrieved from http://www.who.int/dietphysicalactivity/pa/en/

World Health Organization. (2018). Physical Inactivity: A Global Public Health Problem. Retrieved from http://www.who.int/dietphysicalactivity/factsheet_inactivity/en/

APPENDIX A: QUALITY CRITERIA CHECKLISTS AND EVIDENCE ANALYSIS

LIBRARY WORKSHEETS

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Benedetto, S., Caldato, C., Bazzan, E., Greenwood, D., Pensabene, V., & Actis, P. (2018). Assessment of the Fitbit Charge 2 for monitoring heart rate. *PLoS ONE*, *13*(2), e0192691. doi:10.1371/journal.pone.0192691
Study design: *Use algorithm – RCT, cohort, etc*	Validity study
Study Class (A,B,C,D)	C
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**POSITIVE (+)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To assess in a controlled, research environment the accuracy and precision of the Fitbit Charge 2 for measuring heart rate (HR) with respect to a gold standard electrocardiograph
Inclusion criteria: *requirements for study eligibility*	Healthy adult participants
Exclusion criteria (conditions that make individual ineligible)	Participants with neurological or cognitive disorders, recent musculoskeletal damage or surgery that would impair motor function, and tattoos
Recruitment	Unclear
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	NA
Description of study protocol *What happened in the study?*	Participants rode a stationary bike for 10 minutes while their HR was simultaneously recorded from each device
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	For electrocardiograph recording, the electrode placement sites were prepared by standardized procedures of cleaning, shaving, and abrading the skin to improve signal acquisition and to minimize noise artifact. Three self-adhesive electrodes were placed on the upper torso. HR data per second was converted to bpm automatically by the data acquisition software program prior to analysis. The Fitbit Charge 2 was placed on the non-dominant wrist following manufacturer instructions and was charged fully prior to testing. Participants were asked to ride a stationary bike for 10 minutes with the stated goal to raise their HR as much as possible. Participants were free to slow down and rest at any time they desired to do so. The goal of the experiment was not to evaluate the training activity, but rather to collect enough HR data spanning as wide of a range of heart beats per minute (bpm) as possible. HR was acquired simultaneously using both devices (Fitbit Charge 2, ProComp Infiniti T7500M).

Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Agreement between the Fitbit Charge 2 and the ProComp Infiniti T7500M was estimated using the Bland-Altman method. The intraclass correlation coefficient (ICC) was used as an alternative measure of agreement.
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	HR was measured continuously using both devices (Fitbit Charge 2, ProComp Infiniti T7500M) throughout the 10-minute intervention
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	HR according to Fitbit Charge 2
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Participants' level or intensity of cycling, instability or improper positioning of the devices
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	HR according to ProComp Infiniti T7500M
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	15 participants, 7 males and 8 females
Final n (attrition) *number of subjects that completed study*	15 participants, 7 males and 8 females
Age *usually mean or range*	25 to 36 years
Ethnicity (if given)	Caucasian
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	Unclear
Anthropometrics: *e.g. were groups same or different on important physical measures (BMI, fitness level)*	Weight: 56 to 82 kg Height: 155 to 185 cm BMI: 20 to 25 kg/m2
Location: *Where did the study take place? City or country*	TSW XP Lab in Treviso, Italy
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	The Fitbit Charge 2 exhibited a mean bias of -5.9 bpm (95% CI: -6.1 to -5.6 bpm). The limits of agreement (LoA), which indicate the precision of individual measurements, between the Fitbit Charge 2 and ProComp Infiniti T7500M were wide. The upper LoA was +16.8 bpm, whereas the lower LoA was -28.5 bpm. The ICC between the Fitbit Charge 2 and ProComp Infiniti T7500M was 0.21 (95% CI: 0.09 to 0.34).
Author's Conclusions	

Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	Findings are in line with those of several recent publications; the Fitbit Charge 2 presents an unchanged level of HR measurement accuracy compared to existing models of the same brand (i.e. Fitbit Charge HR). The Fitbit Charge 2 tends to underestimate HR, with moderate bias on average, although precision is poor for individual measurements, which could be underestimated by as much as 30 bpm. It may be speculated that the current algorithms for HR estimation lack proper sophistication.
Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: *zero percent attrition, stable positioning of the ProComp Infiniti T7500M* Limitations: *small sample size, participant recruitment was not discussed, blinding was not utilized, unstable positioning of the Fitbit Charge 2,* lacking a defined activity pattern for the participants in order to simulate low, medium, and intensive exercise, and lacking a variety of participants with different skin tones, BMI, and ages Funding source: this work was supported by TSW XP Lab, which only provided financial support in the form of authors' salaries and/or research materials. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Benedetto, S., Caldato, C., Bazzan, E., Greenwood, D., Pensabene, V., & Actis, P. (2018). Assessment of the Fitbit Charge 2 for monitoring heart rate. *PLoS ONE*, *13*(2), e0192691. doi:10.1371/journal.pone.0192691		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1			X	
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3		X		
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES X	NO	UNCLEAR	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1	X			
2.2 Were criteria applied equally to all study groups?	2.2				X
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3	X			
2.4 Were the subjects/patients in a representative sample of the relevant population? *The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.*	2.4			X	

3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES	NO	UNCLEAR	NA
					X
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1				X
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2				X
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3	X			
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6	X			
4. Was method of handling withdrawals described?		YES	NO	UNCLEAR	NA
					X
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2				X
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3	X			
4.4 Were reasons for withdrawals similar across groups?	4.4				X

If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.						
4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5	X				
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES	NO	UNCLEAR	NA	
				X		
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1					X
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2			X		
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3					X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4					X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5			X		
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES	NO	UNCLEAR	NA	
		X				
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1					X
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2	X				
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X				
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4	X				
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5					X
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6					X
6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups?	6.7					X

For a study to be valid and unbiased, it is important that this be yes.						
6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8	X				
7. Were <u>outcomes</u> clearly defined and the measurements valid and reliable?		YES X	NO	UNCLEAR	NA	
7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant.*	7.1	X				
7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X	
7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X	
7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X				
7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5	X				
7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6	X				
7.7 Were the measurements conducted consistently across groups?	7.7				X	
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?		YES X	NO	UNCLEAR	NA	
8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X				
8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X				
8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X				
8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4				X	
8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)? *Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.*	8.5				X	

8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.*	8.6	X			
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7				X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X			
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X			
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X			
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X			

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral* (ø) *symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Benedetto, S., Caldato, C., Bazzan, E., Greenwood, D., Pensabene, V., & Actis, P. (2018). Assessment of the Fitbit Charge 2 for monitoring heart rate. PLoS ONE, 13(2), e0192691. doi:10.1371/journal.pone.0192691
Study Design	Validity study
Class	C
Quality Rating	☒ + (Positive) ☐ - (Negative) ☐ ⊘ (Neutral)
Research Purpose	To assess in a controlled research environment the accuracy and precision of the Fitbit Charge 2 for measuring heart rate (HR) with respect to a gold standard electrocardiograph
Inclusion Criteria	Healthy adult participants
Exclusion Criteria	Participants with neurological or cognitive disorders, recent musculoskeletal damage or surgery that would impair motor function, and tattoos
Description of Study Protocol	Recruitment: Unclear Design: Participants rode a stationary bike for 10 minutes while their HR was simultaneously recorded from each device Blinding used (if applicable): NA Intervention (if applicable): For electrocardiograph recording, the electrode placement sites were prepared by standardized procedures of cleaning, shaving, and abrading the skin to improve signal acquisition and to minimize noise artifact. Three self-adhesive electrodes were placed on the upper torso. HR data per second was converted to bpm automatically by the data acquisition software program prior to analysis. The Fitbit Charge 2 was placed on the non-dominant wrist following manufacturer instructions and was charged fully prior to testing. Participants were asked to ride a stationary bike for 10 minutes with the stated goal to raise their HR as much as possible. Participants were free to slow down and rest at any time they desired to do so. The goal of the experiment was not to evaluate the training activity, but rather to collect enough HR data spanning as wide of a range of heart beats per minute (bpm) as possible.

	HR was acquired simultaneously using both devices (Fitbit Charge 2, ProComp Infiniti T7500M). Statistical Analysis: Agreement between the Fitbit Charge 2 and the ProComp Infiniti T7500M was estimated using the Bland-Altman method. The intraclass correlation coefficient (ICC) was used as an alternative measure of agreement.
Data Collection Summary	Timing of Measurements: HR was measured continuously using both devices (Fitbit Charge 2, ProComp Infiniti T7500M) throughout the 10 minute intervention Dependent Variables: HR according to Fitbit Charge 2 Independent Variables: Participants' level or intensity of cycling, instability or improper positioning of the devices Control Variables: HR according to ProComp Infiniti T7500M
Description of Actual Data Sample	Initial: 15 (7 Males 8 Females) Attrition (final N): 15 Age: 25 to 36 years Ethnicity: Caucasian Other relevant demographics: Unclear Anthropometrics: Weight: 56 to 82 kg, Height: 155 to 185 cm, and BMI: 20 to 25 kg/m2 Location: TSW XP Lab in Treviso, Italy
Summary of Results	Key Findings: The Fitbit Charge 2 exhibited a mean bias of -5.9 bpm (95% CI: -6.1 to -5.6 bpm). The limits of agreement (LoA), which indicate the precision of individual measurements, between the Fitbit Charge 2 and ProComp Infiniti T7500M were wide. The upper LoA was +16.8 bpm, whereas the lower LoA was -28.5 bpm. The ICC between the Fitbit Charge 2 and ProComp Infiniti T7500M was 0.21 (95% CI: 0.09 to 0.34). Other Findings:

Author Conclusion	Findings are in line with those of several recent publications; the Fitbit Charge 2 presents an unchanged level of HR measurement accuracy compared to existing models of the same brand (i.e. Fitbit Charge HR). The Fitbit Charge 2 tends to underestimate HR, with moderate bias on average, although precision is poor for individual measurements, which could be underestimated by as much as 30 bpm. It may be speculated that the current algorithms for HR estimation lack proper sophistication.
Reviewer Comments	*Strengths: zero percent attrition, stable positioning of the ProComp Infiniti T7500M* *Limitations: small sample size, participant recruitment was not discussed, blinding was not utilized, unstable positioning of the Fitbit Charge 2, lacking a defined activity pattern for the participants in order to simulate low, medium, and intensive exercise, and lacking a variety of participants with different skin tones, BMI, and ages*
Funding Source	This work was supported by TSW XP Lab, which only provided financial support in the form of authors' salaries and/or research materials. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive** – Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative** – Indicates that these issues have not been adequately addressed.*
⊘	***Neutral** – indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Unclear
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	No
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes

If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Yes
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Yes
2.2. Were criteria applied equally to all study groups?	2.2	N/A
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	Yes
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Unclear
3. Were <u>study groups comparable</u>?	3	N/A
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	N/A
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	N/A
3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	Yes
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	N/A
3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.)	3.5	N/A
3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	Yes
4. Was method of handling <u>withdrawals</u> described?	4	N/A
4.1. Were follow up methods described and the same for all groups?	4.1	N/A
4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)	4.2	N/A
4.3. Were all enrolled subjects/patients (in the original sample) accounted for?	4.3	Yes
4.4. Were reasons for withdrawals similar across groups	4.4	N/A
4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4.5	Yes
5. Was <u>blinding</u> used to prevent introduction of bias?	5	Unclear
5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate?	5.1	N/A
5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.)	5.2	Unclear
5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	5.3	N/A

5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	5.4	N/A
5.5. In diagnostic study, were test results blinded to patient history and other test results?	5.5	Unclear
6. Were <u>intervention</u>/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were <u>intervening factors</u> described?	6	Yes
6.1. In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	N/A
6.2. In observational study, were interventions, study settings, and clinicians/provider described?	6.2	Yes
6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	6.3	Yes
6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured?	6.4	Yes
6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described?	6.5	N/A
6.6. Were extra or unplanned treatments described?	6.6	N/A
6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups?	6.7	N/A
6.8. In diagnostic study, were details of test administration and replication sufficient?	6.8	Yes
7. Were <u>outcomes</u> clearly defined and the <u>measurements valid and reliable</u>?	7	Yes
7.1. Were primary and secondary endpoints described and relevant to the question?	7.1	Yes
7.2. Were nutrition measures appropriate to question and outcomes of concern?	7.2	N/A
7.3. Was the period of follow-up long enough for important outcome(s) to occur?	7.3	N/A
7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures?	7.4	Yes
7.5. Was the measurement of effect at an appropriate level of precision?	7.5	Yes
7.6. Were other factors accounted for (measured) that could affect outcomes?	7.6	Yes
7.7. Were the measurements conducted consistently across groups?	7.7	N/A
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?	8	Yes
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Yes
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes
8.3. Were statistics reported with levels of significance and/or confidence intervals?	8.3	Yes
8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)?	8.4	N/A
8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)?	8.5	N/A
8.6. Was clinical significance as well as statistical significance reported?	8.6	Yes
8.7. If negative findings, was a power calculation reported to address type 2 error?	8.7	N/A
9. **Are <u>conclusions supported by results</u> with biases and limitations taken into consideration?**	9	Yes
9.1. Is there a discussion of findings?	9.1	Yes
9.2. Are biases and study limitations identified and discussed?	9.2	Yes
10. Is bias due to study's <u>funding or sponsorship</u> unlikely?	10	Yes
10.1. Were sources of funding and investigators' affiliations described?	10.1	Yes
10.2. Was there no apparent conflict of interest?	10.2	Yes
MINUS/NEGATIVE (-)		

If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.
NEUTRAL (∅) *If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.*
PLUS/POSITIVE (+) *If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.*

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Cadmus-Bertram, L., Gangnon, R., Wirkus, E. J., Thraen-Borowski, K. M., & Gorzelitz-Liebhauser, J. (2017). The Accuracy of Heart Rate Monitoring by Some Wrist-Worn Activity Trackers. *Annals of Internal Medicine*, *166*(8), 610–612. doi:10.7326/L16-0353
Study design: *Use algorithm – RCT, cohort, etc*	Validity study
Study Class (A,B,C,D)	C
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**NEUTRAL (ø)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To determine the accuracy of the heart rate measured by four commercial, light-emitting diode–dependent, wrist-worn activity trackers
Inclusion criteria: *requirements for study eligibility*	Healthy adult participants
Exclusion criteria (conditions that make individual ineligible)	Participants with cardiovascular conditions
Recruitment	Unclear
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	NA
Description of study protocol *What happened in the study?*	Four activity trackers (Basis Peak, Fitbit Charge, Fitbit Surge, Mio Fuse) were selected and tested for resting and active heart rate measurement accuracy against a reference standard electrocardiograph
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	For each participant, two activity trackers were placed on each wrist. Next, participants were seated, connected to an electrocardiograph, and their resting heart rates were measured at one-minute intervals for 10 minutes. Then, participants' heart rates were measured at one-minute intervals for 10 minutes while they exercised on a treadmill at 65 percent of their maximum heart rate.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Bland-Altman plots were used to compare the heart rates measured by the electrocardiograph and by each of the activity trackers
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	Heart rate was measured at one-minute intervals throughout the 20-minute intervention using the electrocardiograph and four activity trackers (Basis Peak, Fitbit Charge, Fitbit Surge, Mio Fuse)
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	Heart rate according to Basis Peak, Fitbit Charge, Fitbit Surge, Mio Fuse

Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Participants' level of physical fitness
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	Heart rate according to electrocardiograph
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	40 participants, 20 males and 20 females
Final n (attrition) *number of subjects that completed study*	40 participants, 20 males and 20 females
Age *usually mean or range*	30 to 65 years
Ethnicity (if given)	Unclear
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	Unclear
Anthropometrics: *e.g. were groups same or different on important physical measures (BMI, fitness level)*	Mean BMI: 25.1 kg/m2
Location: *Where did the study take place? City or country*	Unclear
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	For participants at rest, the limits of agreement was best for the Fitbit Surge, which had the narrowest limits of agreement (−5.1 to 4.5 beats/min), worst for the Basis Peak (−17.1 to 22.6 beats/min), and intermediate for the Fitbit Charge (−10.5 to 9.2 beats/min) and Mio Fuse (−7.8 to 9.9 beats/min). When participants exercised at 65% of their maximum heart rate, the limits of agreement were relatively poor for all the activity trackers (Mio Fuse, −22.5 to 26.0 beats/min; Basis Peak, −27.1 to 29.2 beats/min; Fitbit Surge, −34.8 to 39.0 beats/min; and Fitbit Charge, −41.0 to 36.0 beats/min). The repeatability coefficient for the electrocardiograph was 5.3 beats/min at rest and 9.1 beats/min during exercise. In comparison, the repeatability coefficient at rest was 4.2 beats/min for the Fitbit Surge, 9.3 beats/min for the Fitbit Charge, 10.9 beats/min for the Mio Fuse, and 19.3 beats/min for the Basis Peak. During exercise, the repeatability coefficient was 20.2 beats/min for the Basis Peak, 20.6 beats/min for the Fitbit Surge, 21.6 beats/min for the Fitbit Charge, and 23.7 beats/min for the Mio Fuse.
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	Some of the activity trackers measured values for heart rate that were similar to those measured by the electrocardiograph. All of the activity trackers were more accurate at rest than during moderate exercise, performance at rest was better for some trackers than for others, and limited repeatability for each tracker caused more problems than poor agreement between each activity tracker and the electrocardiograph. Although current trackers may help persons self-monitor their daily activity, more research is needed before we can confidently conclude that the monitoring feature for heart rate is sufficient to help clinicians advise their patients

	about health issues and conduct clinical trials that require a high level of accuracy and reliability for heart rate measurement.
Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: *zero percent attrition, defined activity pattern for participants to stimulate minimal and moderate exercise* Limitations: *participant recruitment, demographics, and location of study were not discussed, and blinding was not utilized* Funding source: unclear, authors disclosed no conflicts of interest

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Cadmus-Bertram, L., Gangnon, R., Wirkus, E. J., Thraen-Borowski, K. M., & Gorzelitz-Liebhauser, J. (2017). The Accuracy of Heart Rate Monitoring by Some Wrist-Worn Activity Trackers. *Annals of Internal Medicine*, *166*(8), 610–612. doi:10.7326/L16-0353		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1			X	
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3		X		
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES	NO	UNCLEAR X	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1	X			
2.2 Were criteria applied equally to all study groups?	2.2				X
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3		X		
2.4 Were the subjects/patients in a representative sample of the relevant population?	2.4			X	

The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.					
3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES	NO	UNCLEAR	NA X
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1				X
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2				X
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3	X			
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6	X			
4. Was method of handling withdrawals described?		YES	NO	UNCLEAR	NA X
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2				X
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3	X			

4.4 Were reasons for withdrawals similar across groups? *If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.*	4.4				X
4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5	X			
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES	NO	UNCLEAR	NA
				X	
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1				X
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2			X	
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3				X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4				X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5			X	
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES	NO	UNCLEAR	NA
		X			
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1				X
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2	X			
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X			
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4	X			
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5				X
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6				X

	6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7				X
	6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8	X			
7.	**Were <u>outcomes</u> clearly defined and the measurements valid and reliable?**		YES X	NO	UNCLEAR	NA
	7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant*.	7.1	X			
	7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X
	7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X
	7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X			
	7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5	X			
	7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6	X			
	7.7 Were the measurements conducted consistently across groups?	7.7				X
8.	**Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?**		YES X	NO	UNCLEAR	NA
	8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X			
	8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X			
	8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3			X	
	8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4				X
	8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)? *Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.*	8.5				X

8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.*	8.6			X	
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7				X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X			
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2			X	
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1			X	
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X			

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral* (ø) *symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Cadmus-Bertram, L., Gangnon, R., Wirkus, E. J., Thraen-Borowski, K. M., & Gorzelitz-Liebhauser, J. (2017). The Accuracy of Heart Rate Monitoring by Some Wrist-Worn Activity Trackers. Annals of Internal Medicine, 166(8), 610–612. doi:10.7326/L16-0353
Study Design	Validity study
Class	C
Quality Rating	☐ + (Positive) ☐ - (Negative) ☒ ⊘ (Neutral)
Research Purpose	To determine the accuracy of the heart rate measured by four commercial, light-emitting diode–dependent, wrist-worn activity trackers
Inclusion Criteria	Healthy adult participants
Exclusion Criteria	Participants with cardiovascular conditions
Description of Study Protocol	Recruitment: Unclear Design: Four activity trackers (Basis Peak, Fitbit Charge, Fitbit Surge, Mio Fuse) were selected and tested for resting and active heart rate measurement accuracy against a reference standard electrocardiograph Blinding used (if applicable): NA Intervention (if applicable): For each participant, two activity trackers were placed on each wrist. Next, participants were seated, connected to an electrocardiograph, and their resting heart rates were measured at one-minute intervals for 10 minutes. Then, participants' heart rates were measured at one-minute intervals for 10 minutes while they exercised on a treadmill at 65 percent of their maximum heart rate. Statistical Analysis: Bland-Altman plots were used to compare the heart rates measured by the electrocardiograph and by each of the activity trackers
Data Collection Summary	Timing of Measurements: Heart rate was measured at one-minute intervals throughout the 20-minute intervention using the electrocardiograph and four activity trackers (Basis Peak, Fitbit Charge, Fitbit Surge, Mio Fuse)

	Dependent Variables: Heart rate according to Basis Peak, Fitbit Charge, Fitbit Surge, Mio Fuse Independent Variables: Participants' level of physical fitness Control Variables: Heart rate according to electrocardiograph
Description of Actual Data Sample	Initial: 40 (20 Males 20 Females) Attrition (final N): 40 Age: 30 to 65 years Ethnicity: Unclear Other relevant demographics: Unclear Anthropometrics: Mean BMI: 25.1 kg/m2 Location: Unclear
Summary of Results	Key Findings: For participants at rest, the limits of agreement was best for the Fitbit Surge, which had the narrowest limits of agreement (−5.1 to 4.5 beats/min), worst for the Basis Peak (−17.1 to 22.6 beats/min), and intermediate for the Fitbit Charge (−10.5 to 9.2 beats/min) and Mio Fuse (−7.8 to 9.9 beats/min). When participants exercised at 65% of their maximum heart rate, the limits of agreement were relatively poor for all the activity trackers (Mio Fuse, −22.5 to 26.0 beats/min; Basis Peak, −27.1 to 29.2 beats/min; Fitbit Surge, −34.8 to 39.0 beats/min; and Fitbit Charge, −41.0 to 36.0 beats/min). The repeatability coefficient for the electrocardiograph was 5.3 beats/min at rest and 9.1 beats/min during exercise. In comparison, the repeatability coefficient at rest was 4.2 beats/min for the Fitbit Surge, 9.3 beats/min for the Fitbit Charge, 10.9 beats/min for the Mio Fuse, and 19.3 beats/min for the Basis Peak. During exercise, the repeatability coefficient was 20.2 beats/min for the Basis Peak, 20.6 beats/min for the Fitbit Surge, 21.6 beats/min for the Fitbit Charge, and 23.7 beats/min for the Mio Fuse. Other Findings:
Author Conclusion	Some of the activity trackers measured values for heart rate that were similar to those measured by the electrocardiograph. All of the activity

	trackers were more accurate at rest than during moderate exercise, performance at rest was better for some trackers than for others, and limited repeatability for each tracker caused more problems than poor agreement between each activity tracker and the electrocardiograph. Although current trackers may help persons self-monitor their daily activity, more research is needed before we can confidently conclude that the monitoring feature for heart rate is sufficient to help clinicians advise their patients about health issues and conduct clinical trials that require a high level of accuracy and reliability for heart rate measurement.
Reviewer Comments	*Strengths: zero percent attrition, defined activity pattern for participants to stimulate minimal and moderate exercise* *Limitations: participant recruitment, demographics, and location of study were not discussed, and blinding was not utilized*
Funding Source	Unclear, authors disclosed no conflicts of interest

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive** – Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative** – Indicates that these issues have not been adequately addressed.*
⊘	***Neutral** – indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Unclear
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	No
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the research question clearly stated?	1	Yes

1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Unclear
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Yes
2.2. Were criteria applied equally to all study groups?	2.2	N/A
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	No
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Unclear
3. Were <u>study groups comparable</u>?	3	N/A
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	N/A
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	N/A
3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	Yes
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	N/A
3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.)	3.5	N/A
3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	Yes
4. Was method of handling <u>withdrawals</u> described?	4	N/A
4.1. Were follow up methods described and the same for all groups?	4.1	N/A
4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)	4.2	N/A
4.3. Were all enrolled subjects/patients (in the original sample) accounted for?	4.3	Yes
4.4. Were reasons for withdrawals similar across groups	4.4	N/A
4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4.5	Yes
5. Was <u>blinding</u> used to prevent introduction of bias?	5	Unclear
5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate?	5.1	N/A
5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.)	5.2	Unclear
5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	5.3	N/A
5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	5.4	N/A
5.5. In diagnostic study, were test results blinded to patient history and other test results?	5.5	Unclear
	6	Yes

6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?	6.1	N/A
6.1. In RCT or other intervention trial, were protocols described for all regimens studied?	6.2	Yes
6.2. In observational study, were interventions, study settings, and clinicians/provider described?	6.3	Yes
6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	6.4	Yes
6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured?	6.5	N/A
6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described?	6.6	N/A
6.6. Were extra or unplanned treatments described?	6.7	N/A
6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups?	6.8	Yes
6.8. In diagnostic study, were details of test administration and replication sufficient?		
7. Were outcomes clearly defined and the measurements valid and reliable?	7	Yes
7.1. Were primary and secondary endpoints described and relevant to the question?	7.1	Yes
7.2. Were nutrition measures appropriate to question and outcomes of concern?	7.2	N/A
7.3. Was the period of follow-up long enough for important outcome(s) to occur?	7.3	N/A
7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures?	7.4	Yes
7.5. Was the measurement of effect at an appropriate level of precision?	7.5	Yes
7.6. Were other factors accounted for (measured) that could affect outcomes?	7.6	Yes
7.7. Were the measurements conducted consistently across groups?	7.7	N/A
8. Was the statistical analysis appropriate for the study design and type of outcome indicators?	8	Yes
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Yes
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes
8.3. Were statistics reported with levels of significance and/or confidence intervals?	8.3	Unclear
8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)?	8.4	N/A
8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)?	8.5	N/A
8.6. Was clinical significance as well as statistical significance reported?	8.6	Unclear
8.7. If negative findings, was a power calculation reported to address type 2 error?	8.7	N/A
9. **Are conclusions supported by results with biases and limitations taken into consideration?**	9	Yes
9.1. Is there a discussion of findings?	9.1	Yes
9.2. Are biases and study limitations identified and discussed?	9.2	Unclear
10. Is bias due to study's funding or sponsorship unlikely?	10	Yes
10.1. Were sources of funding and investigators' affiliations described?	10.1	Unclear
10.2. Was there no apparent conflict of interest?	10.2	Yes
MINUS/NEGATIVE (-) *If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.*		
NEUTRAL (∅)		

If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.
PLUS/POSITIVE (+) *If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.*

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Cadmus-Bertram, L., Marcus, B., Patterson, R., Parker, B., & Morey, B. (2015). Randomized Trial of a Fitbit-Based Physical Activity Intervention for Women. *American Journal of Preventive Medicine*, *49*(3), 414–418. doi:10.1016/j.amepre.2015.01.020
Study design: *Use algorithm – RCT, cohort, etc*	Randomized controlled trial
Study Class (A,B,C,D)	A
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**POSITIVE (+)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To evaluate, within a randomized controlled trial, the feasibility and preliminary efficacy of integrating a Fitbit tracker and website into a physical activity intervention for postmenopausal, overweight/obese women
Inclusion criteria: *requirements for study eligibility*	Postmenopausal women, overweight/obese, perform less than 60 minutes per week of moderate-to-vigorous intensity physical activity (MVPA), could exercise safely, were regular internet users, and owned a tablet/computer
Exclusion criteria (conditions that make individual ineligible)	Women who responded, "somewhat uncomfortable," "very uncomfortable," "somewhat do not enjoy," or "very much do not enjoy" on baseline questionnaire
Recruitment	Unclear, supported by the Athena Breast Health Network
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	Unclear
Description of study protocol *What happened in the study?*	Participants were randomized to either a Fitbit or pedometer-based intervention group to determine whether the Fitbit One increased physical activity more than the pedometer in postmenopausal, overweight/obese women
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	Participants received a baseline questionnaire, anthropometric measurements, and physical activity was measured for seven days using the ActiGraph GT3X+ accelerometer. Then, participants were randomized to a 16-week web-based self-monitoring intervention (N=25) or a comparison group (N=26). Participants in the Web-Based Tracking Group received a Fitbit One, an instructional session, and a follow-up call at four weeks. The comparison group received a standard pedometer and printed materials with tips for increasing steps. All participants were asked to perform 150 minutes per week of MVPA and walk 10,000 steps per day. A final questionnaire evaluated the assigned intervention, and physical activity was measured for an additional seven days using the ActiGraph GT3X+ accelerometer.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Data were collected and analyzed according to the intent-to-treat principle. Baseline characteristics were compared using chi-square and t-tests. ActiGraph data were adjusted for number of valid days (95% had 7 valid days; 5% had 5–6 valid days). Baseline-to-16-week physical activity changes were assessed using repeated-measures ANCOVA, adjusted for age and ActiGraph daily wear time to address potential residual confounding.

Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	Participants' physical activity was measured for seven days at baseline and at 16 weeks using the ActiGraph GT3X+. Participants were asked to wear the Fitbit One or pedometer every day throughout the intervention period (112 prescribed days).
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	Physical activity changes in Pedometer Group and Web-Based Tracking Group
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Goal setting process, four-week follow-up call, Fitbit website, printed materials with tips for increasing steps
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	Physical activity according to ActiGraph GT3X+, physical activity changes were adjusted for age and ActiGraph daily wear time
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	51 participants, 0 males and 51 females
Final n (attrition) *number of subjects that completed study*	51 participants, 0 males and 51 females
Age *usually mean or range*	53 to 67 years
Ethnicity (if given)	46 participants were non-Hispanic White
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	32 participants earned a college degree or higher, all participants were comfortable using computers and the internet
Anthropometrics: *e.g. were groups same or different on important physical measures (BMI, fitness level)*	BMI: 25.7 to 32.7 kg/m2
Location: *Where did the study take place? City or country*	University of California, San Diego
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	At baseline, participants were performing 33±56 min/week of MVPA in bouts ≥10 minutes in length and accumulating 5,866±2,195 steps/day. After the 16-week intervention, the Web-Based Tracking Group increased MVPA by 62±108 min/week (p<.001), MVPA in 10-min bouts by 38±83 min/week (p=0.008), and steps by 789±1,979 (p=0.01), compared to non-significant increases in the Pedometer Group (between-group p-values: 0.11, 0.28 and 0.30, respectively). The Web-Based Tracking Group wore the tracker on 95% of intervention days; 96% reported liking the website, and 100% liked the Fitbit One.
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	The Fitbit One was well-accepted in this sample of women and was associated with increased physical activity at 16 weeks. By leveraging direct-to-consumer technologies that align with behavior change theories, researchers can strengthen physical activity interventions.

Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: *zero percent attrition, use of baseline and final questionnaires for* detailed participant feedback, use of the ActiGraph GT3X+, and use of Fitbit data to corroborate adherence Limitations: small sample size, short duration, and lack of generalizability since all participants were postmenopausal, overweight/obese women and there were several confounders such as the goal setting process, four-week follow-up call, and Fitbit website Funding source: this study was funded by NIH (1R03CA168450) and recruitment supported by the Athena Breast Health Network. This research was supported by the National Cancer Institute (1R03CA168450). The authors have no conflicts of interest to report.

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Cadmus-Bertram, L., Marcus, B., Patterson, R., Parker, B., & Morey, B. (2015). Randomized Trial of a Fitbit-Based Physical Activity Intervention for Women. *American Journal of Preventive Medicine*, *49*(3), 414–418. doi:10.1016/j.amepre.2015.01.020		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1	X			
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	X			
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES X	NO	UNCLEAR	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1	X			
2.2 Were criteria applied equally to all study groups?	2.2	X			
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3	X			
2.4 Were the subjects/patients in a representative sample of the relevant population? *The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.*	2.4			X	

3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES	NO	UNCLEAR	NA
		X			
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1	X			
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2	X			
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3	X			
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6				X
4. Was method of handling withdrawals described?		YES	NO	UNCLEAR	NA
					X
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2				X
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3	X			
4.4 Were reasons for withdrawals similar across groups?	4.4				X

If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.						
4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5					X
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES	NO	UNCLEAR		NA
				X		
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1			X		
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2			X		
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3					X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4					X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5					X
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES	NO	UNCLEAR		NA
		X				
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	X				
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2					X
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X				
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4	X				
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5	X				
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6					X
6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7	X				

	6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8				X
7.	**Were <u>outcomes</u> clearly defined and the measurements valid and reliable?**		YES X	NO	UNCLEAR	NA
	7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant.*	7.1	X			
	7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X
	7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X
	7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X			
	7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5	X			
	7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6	X			
	7.7 Were the measurements conducted consistently across groups?	7.7	X			
8.	**Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?**		YES X	NO	UNCLEAR	NA
	8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X			
	8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X			
	8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X			
	8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4	X			
	8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)? *Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.*	8.5	X			
	8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical*	8.6	X			

significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.						
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7					X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA	
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X				
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X				
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA	
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X				
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X				

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral (ø) symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Cadmus-Bertram, L., Marcus, B., Patterson, R., Parker, B., & Morey, B. (2015). Randomized Trial of a Fitbit-Based Physical Activity Intervention for Women. American Journal of Preventive Medicine, 49(3), 414–418. doi:10.1016/j.amepre.2015.01.020
Study Design	Randomized controlled trial
Class	A
Quality Rating	☒ + (Positive) ☐ - (Negative) ☐ ⊘ (Neutral)
Research Purpose	To evaluate, within a randomized controlled trial, the feasibility and preliminary efficacy of integrating a Fitbit tracker and website into a physical activity intervention for postmenopausal, overweight/obese women
Inclusion Criteria	Postmenopausal women, overweight/obese, perform less than 60 minutes per week of moderate-to-vigorous intensity physical activity (MVPA), could exercise safely, were regular internet users, and owned a tablet/computer
Exclusion Criteria	Women who responded, "somewhat uncomfortable," "very uncomfortable," "somewhat do not enjoy," or "very much do not enjoy" on baseline questionnaire
Description of Study Protocol	Recruitment: Unclear, supported by the Athena Breast Health Network Design: Participants were randomized to either a Fitbit or pedometer-based intervention group to determine whether the Fitbit One increased physical activity more than the pedometer in postmenopausal, overweight/obese women Blinding used (if applicable): Unclear Intervention (if applicable): Participants received a baseline questionnaire, anthropometric measurements, and physical activity was measured for seven days using the ActiGraph GT3X+ accelerometer. Then, participants were randomized to a 16-week web-based self-monitoring intervention (N=25) or a comparison group (N=26). Participants in the Web-Based Tracking Group received a Fitbit One, an instructional session, and a follow-up call at four weeks. The comparison group received a standard pedometer and printed materials with tips for increasing steps. All participants were asked to perform 150 minutes per

	week of MVPA and walk 10,000 steps per day. A final questionnaire evaluated the assigned intervention, and physical activity was measured for an additional seven days using the ActiGraph GT3X+ accelerometer. Statistical Analysis: Data were collected and analyzed according to the intent-to-treat principle. Baseline characteristics were compared using chi-square and t-tests. ActiGraph data were adjusted for number of valid days (95% had 7 valid days; 5% had 5–6 valid days). Baseline-to-16-week physical activity changes were assessed using repeated-measures ANCOVA, adjusted for age and ActiGraph daily wear time to address potential residual confounding.
Data Collection Summary	Timing of Measurements: Participants' physical activity was measured for seven days at baseline and at 16 weeks using the ActiGraph GT3X+. Participants were asked to wear the Fitbit One or pedometer every day throughout the intervention period (112 prescribed days). Dependent Variables: Physical activity changes in Pedometer Group and Web-Based Tracking Group Independent Variables: Goal setting process, four-week follow-up call, Fitbit website, printed materials with tips for increasing steps Control Variables: Physical activity according to ActiGraph GT3X+, physical activity changes were adjusted for age and ActiGraph daily wear time
Description of Actual Data Sample	Initial: 51 (0 Males 51 Females) Attrition (final N): 51 Age: 53 to 67 years Ethnicity: 46 participants were non-Hispanic White Other relevant demographics: 32 participants earned a college degree or higher, all participants were comfortable using computers and the internet Anthropometrics: BMI: 25.7 to 32.7 kg/m2 Location: University of California, San Diego

Summary of Results	Key Findings: At baseline, participants were performing 33±56 min/week of MVPA in bouts ≥10 minutes in length and accumulating 5,866±2,195 steps/day. After the 16-week intervention, the Web-Based Tracking Group increased MVPA by 62±108 min/week (p<.001), MVPA in 10-min bouts by 38±83 min/week (p=0.008), and steps by 789±1,979 (p=0.01), compared to non-significant increases in the Pedometer Group (between-group p-values: 0.11, 0.28 and 0.30, respectively). The Web-Based Tracking Group wore the tracker on 95% of intervention days; 96% reported liking the website, and 100% liked the Fitbit One. Other Findings:
Author Conclusion	The Fitbit One was well-accepted in this sample of women and was associated with increased physical activity at 16 weeks. By leveraging direct-to-consumer technologies that align with behavior change theories, researchers can strengthen physical activity interventions.
Reviewer Comments	*Strengths: zero percent attrition, use of baseline and final questionnaires for detailed participant feedback, use of the ActiGraph GT3X+, and use of Fitbit data to corroborate adherence* *Limitations: small sample size, short duration, and lack of generalizability since all participants were postmenopausal, overweight/obese women and there were several confounders such as the goal setting process, four-week follow-up call, and Fitbit website*
Funding Source	This study was funded by NIH (1R03CA168450) and recruitment supported by the Athena Breast Health Network. This research was supported by the National Cancer Institute (1R03CA168450). The authors have no conflicts of interest to report.

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive*** *– Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative*** *– Indicates that these issues have not been adequately addressed.*

⊘	***Neutral*** *– indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Yes
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	Yes
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Yes
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Yes
2.2. Were criteria applied equally to all study groups?	2.2	Yes
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	Yes
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Unclear
3. Were <u>study groups comparable</u>?	3	Yes
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	Yes
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	Yes
3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	Yes
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	N/A
3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.)	3.5	N/A
3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	N/A
4. Was method of handling <u>withdrawals</u> described?	4	N/A

4.1. Were follow up methods described and the same for all groups?	4.1	N/A
4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)	4.2	N/A
4.3. Were all enrolled subjects/patients (in the original sample) accounted for?	4.3	Yes
4.4. Were reasons for withdrawals similar across groups	4.4	N/A
4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4.5	N/A
5. Was <u>blinding</u> used to prevent introduction of bias?	5	Unclear
5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate?	5.1	Unclear
5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.)	5.2	Unclear
5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	5.3	N/A
5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	5.4	N/A
5.5. In diagnostic study, were test results blinded to patient history and other test results?	5.5	N/A
6. Were <u>intervention</u>/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were <u>intervening factors</u> described?	6	Yes
6.1. In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	Yes
6.2. In observational study, were interventions, study settings, and clinicians/provider described?	6.2	N/A
6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	6.3	Yes
6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured?	6.4	Yes
6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described?	6.5	Yes
6.6. Were extra or unplanned treatments described?	6.6	N/A
6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups?	6.7	Yes
6.8. In diagnostic study, were details of test administration and replication sufficient?	6.8	N/A
7. Were <u>outcomes</u> clearly defined and the <u>measurements valid and reliable</u>?	7	Yes
7.1. Were primary and secondary endpoints described and relevant to the question?	7.1	Yes
7.2. Were nutrition measures appropriate to question and outcomes of concern?	7.2	N/A
7.3. Was the period of follow-up long enough for important outcome(s) to occur?	7.3	N/A
7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures?	7.4	Yes
7.5. Was the measurement of effect at an appropriate level of precision?	7.5	Yes
7.6. Were other factors accounted for (measured) that could affect outcomes?	7.6	Yes
7.7. Were the measurements conducted consistently across groups?	7.7	Yes
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?	8	Yes
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Yes
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes

8.3. Were statistics reported with levels of significance and/or confidence intervals? 8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? 8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)? 8.6. Was clinical significance as well as statistical significance reported? 8.7. If negative findings, was a power calculation reported to address type 2 error?	8.3	Yes
	8.4	Yes
	8.5	Yes
	8.6	Yes
	8.7	N/A
9. **Are <u>conclusions supported by results</u> with biases and limitations taken into consideration?** 9.1. Is there a discussion of findings? 9.2. Are biases and study limitations identified and discussed?	9	Yes
	9.1	Yes
	9.2	Yes
10. Is bias due to study's <u>funding or sponsorship</u> unlikely? 10.1. Were sources of funding and investigators' affiliations described? 10.2. Was there no apparent conflict of interest?	10	Yes
	10.1	Yes
	10.2	Yes
MINUS/NEGATIVE (-) *If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.*		
NEUTRAL (∅) *If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.*		
PLUS/POSITIVE (+) *If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.*		

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Chum, J., Kim, M., Zielinski, L., Bhatt, M., Chung, D., Yeung, S., … Samaan, Z. (2017). Acceptability of the Fitbit in behavioural activation therapy for depression: a qualitative study. *Evidence-Based Mental Health*, *20*(4), 128–133. doi:10.1136/eb-2017-102763
Study design: *Use algorithm – RCT, cohort, etc*	Randomized controlled trial
Study Class (A,B,C,D)	A
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**NEUTRAL (ø)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To understand patients' perceived benefit from the Fitbit One and explore themes associated with patient experiences. To compare perceived benefit, patient factors, Fitbit usage, and Beck's Depression Inventory (BDI) scores.
Inclusion criteria: *requirements for study eligibility*	Patients 18 years or older with major depressive disorder were approached to participate in the Behavioural Activation Group Program in Patients with Depression (BRAVE) study. All patients with depressive disorder receiving treatment for depression, including pharmacotherapy and psychotherapy were eligible to participate. Among the 87 participants who completed the BRAVE study, 36 participants completed interviews.
Exclusion criteria (conditions that make individual ineligible)	Inability to provide written informed consent, inability to understand written and spoken English, and having a primary diagnosis other than depression
Recruitment	Patients who were attending or referred for an assessment at the mood disorders outpatient clinic
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	Research assistants collected, recharged, and synced data from each participants' Fitbit One to their online account. Personalized printouts were shared with participants to minimize data manipulation. Research assistants also conducted individual, anonymized interviews with patients who completed the BRAVE study.
Description of study protocol *What happened in the study?*	During the BRAVE study, patients were asked to carry a Fitbit One for 18-weeks. Interviews were conducted with 36 patients who completed the BRAVE study. Thematic analyses were conducted on the interviews and exploratory quantitative analyses were conducted on patient characteristics, Fitbit usage, steps recorded, perceived benefit, and BDI scores.
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	Fitbit One activity trackers were provided to all patients allocated to the behavioral activation (BA) group at the beginning of the BRAVE study, and email accounts were created for each participant. The patients were encouraged to carry the Fitbit One at all times throughout the 18 weeks to track their physical activity. At the beginning of each BA session, Fitbits were collected, recharged, and synced to the online accounts of their respective holders. The participants' weekly activity data were then exported from the Fitbit website, and a personalized printout was generated including summaries of step count, stairs climbed, and distance travelled. The printouts were shared with participants as they did not have access to the online accounts to minimize data manipulation. However, the Fitbit One can show the number of steps taken daily, which can be seen by the patient. Qualitative data

	were collected through individual, semi-structured interviews, which aimed to explore patients' experiences with their BA treatment and their Fitbit use. Additional data such as baseline BDI scores, pretherapy goals, familiarity with technology, and other patient characteristics were collected through questionnaires facilitated by a clinician during the study intake. BDI scores were collected weekly throughout the study including at completion. Exploratory analyses were conducted to examine how perceived benefit related to participants' baseline characteristics, including age, BDI scores, pretherapy physical activity goals, and smartphone use, as well as percent change in BDI scores from baseline to completion and Fitbit usage.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	The control and intervention groups were examined through chi-square tests. To examine the relationships between baseline characteristics and Fitbit use, independent t-tests were used. Between group differences were also calculated through independent t-tests. Pearson's correlation was used to compare Fitbit usage with age and baseline BDI scores.
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	Trained clinicians delivered BA to participants randomized to receive the intervention up to twice per week over 18 weeks, for a total of 28 sessions. Patients were asked to carry a Fitbit One at all times for the 18-week intervention period. Individual interviews lasted 30 to 60 minutes and were finished within two months of participants' completion of the BA program.
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	Patients' perceived benefits, Fitbit usage, BDI scores
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	BA topics, such as value assessment, goal setting, leisure education
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	Receiving treatment for depression as usual, including pharmacotherapy and psychotherapy
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	36 participants, 18 males and 18 females
Final n (attrition) *number of subjects that completed study*	36 participants, 18 males and 18 females
Age *usually mean or range*	Mean age: 53 years
Ethnicity (if given)	Unclear
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	Mean BDI score: 36.27 (scale of 0 to 63) 23 participants had prior smartphone use 26 participants attended a college, university, or higher education
Anthropometrics: *e.g. were groups same or different on important physical measures (BMI, fitness level)*	Mean BMI: 30.16 kg/m2

Location: *Where did the study take place? City or country*	Mood disorders outpatient clinic in Hamilton, Ontario, Canada
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	23 patients found the Fitbit One to be helpful for their physical activity. Themes of positive experiences included 27 patients who reported self-awareness, three patients who reported peer motivation, five patients who reported goal-setting opportunities, and 13 patients who reported enjoying using the Fitbit One. Themes of negative experiences included 12 patients who reported inconvenience, 11 patients who reported inaccuracies, two patients who reported discouragement, and 10 patients who reported disinterest. No significant differences were found in age ($p=0.72$), baseline BDI scores ($p=0.44$), percent change in BDI scores ($p=0.283$), pretherapy physical activity goals ($p=0.549$), and smartphone use ($p=0.825$) between those who did and did not find the Fitbit One helpful. However, there was a significant relationship between total Fitbit One usage and perceived benefit. The mean number of weeks of Fitbit One use for those who found the Fitbit helpful was 18.57 (SD 1.21) and 12.27 (SD 5.76) weeks for those who did not ($p<0.001$).
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	Of the 36 patients who underwent the BRAVE study and completed interviews, the majority (64%) found the Fitbit One to be helpful, identifying their Fitbit use as a factor in improving their physical activity level. Many positive themes were concordant with current literature; however, patients also reported negative aspects that may affect use. Interestingly, there was no significant relationship between perceived benefit of the Fitbit and percent change in BDI scores, which contradicts previous literature supporting Fitbit use in treating depression. These findings suggest that the Fitbit One may be useful for patients with varying characteristics. Clinicians and researchers should consider both strengths and limitations of activity trackers when implementing them to motivate patients with depression.
Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: *zero percent attrition, use of baseline data and final interviews for detailed participant feedback, and use of Fitbit data to corroborate adherence* Limitations: *small sample size compared to the number of participants who completed the BRAVE study,* difficulty interpreting exploratory findings given that the study was underpowered to test effectiveness, and there were several confounders such as combined BA intervention, study setting, and participants' restricted use of Fitbit Funding source: supported by Canadian Institutes of Health Research (CIHR) (Randomised Controlled Trials: Mentoring, code number 201303MTP-303860-182743) to conduct the BRAVE study. CIHR mentorship award (2013–2015) is to support research time. The authors declared no competing interests.

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Chum, J., Kim, M., Zielinski, L., Bhatt, M., Chung, D., Yeung, S., … Samaan, Z. (2017). Acceptability of the Fitbit in behavioural activation therapy for depression: a qualitative study. *Evidence-Based Mental Health*, *20*(4), 128–133. doi:10.1136/eb-2017-102763		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1	X			
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3		X		
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES X	NO	UNCLEAR	NA
2.2 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1	X			
2.3 Were criteria applied equally to all study groups?	2.2	X			
2.4 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3	X			
2.4 Were the subjects/patients in a representative sample of the relevant population? *The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.*	2.4			X	

3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES	NO	UNCLEAR	NA
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1			X	
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2			X	
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3		X		
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6				X
4. Was method of handling withdrawals described?		YES	NO	UNCLEAR	NA X
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2				X
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3		X		
4.4 Were reasons for withdrawals similar across groups?	4.4				X

If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.						
4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5					X
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES X	NO	UNCLEAR		NA
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1	X				
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2	X				
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3					X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4					X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5					X
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES X	NO	UNCLEAR		NA
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	X				
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2					X
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X				
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4	X				
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5	X				
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6					X
6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7					X

	6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8				X
7.	**Were <u>outcomes</u> clearly defined and the measurements valid and reliable?**		YES	NO	UNCLEAR	NA
					X	
	7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study. Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant.*	7.1	X			
	7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X
	7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X
	7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4			X	
	7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5		X		
	7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6		X		
	7.7 Were the measurements conducted consistently across groups?	7.7				X
8.	**Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?**		YES	NO	UNCLEAR	NA
					X	
	8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X			
	8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X			
	8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X			
	8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4		X		
	8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)? *Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.*	8.5		X		

8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.*	8.6	X			
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7				X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X			
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X			
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X			
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X			

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral* (ø) *symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Chum, J., Kim, M., Zielinski, L., Bhatt, M., Chung, D., Yeung, S., … Samaan, Z. (2017). Acceptability of the Fitbit in behavioural activation therapy for depression: a qualitative study. Evidence-Based Mental Health, 20(4), 128–133. doi:10.1136/eb-2017-102763
Study Design	Randomized controlled trial
Class	A
Quality Rating	☐ + (Positive) ☐ - (Negative) ☒ ⊘ (Neutral)
Research Purpose	To understand patients' perceived benefit from the Fitbit One and explore themes associated with patient experiences. To compare perceived benefit, patient factors, Fitbit usage, and Beck's Depression Inventory (BDI) scores.
Inclusion Criteria	Patients 18 years or older with major depressive disorder were approached to participate in the Behavioural Activation Group Program in Patients with Depression (BRAVE) study. All patients with depressive disorder receiving treatment for depression, including pharmacotherapy and psychotherapy were eligible to participate. Among the 87 participants who completed the BRAVE study, 36 participants completed interviews.
Exclusion Criteria	Inability to provide written informed consent, inability to understand written and spoken English, and having a primary diagnosis other than depression
Description of Study Protocol	Recruitment: Patients who were attending or referred for an assessment at the mood disorders outpatient clinic Design: During the BRAVE study, patients were asked to carry a Fitbit One for 18-weeks. Interviews were conducted with 36 patients who completed the BRAVE study. Thematic analyses were conducted on the interviews and exploratory quantitative analyses were conducted on patient characteristics, Fitbit usage, steps recorded, perceived benefit, and BDI scores. Blinding used (if applicable): Research assistants collected, recharged, and synced data from each participants' Fitbit One to their online account. Personalized printouts were shared with participants to minimize data manipulation. Research assistants also conducted individual, anonymized interviews with patients who completed the BRAVE study.

	Intervention (if applicable): Fitbit One activity trackers were provided to all patients allocated to the behavioural activation (BA) group at the beginning of the BRAVE study, and email accounts were created for each participant. The patients were encouraged to carry the Fitbit One at all times throughout the 18 weeks to track their physical activity. At the beginning of each BA session, Fitbits were collected, recharged, and synced to the online accounts of their respective holders. The participants' weekly activity data were then exported from the Fitbit website, and a personalized printout was generated including summaries of step count, stairs climbed, and distance travelled. The printouts were shared with participants as they did not have access to the online accounts to minimize data manipulation. However, the Fitbit One can show the number of steps taken daily, which can be seen by the patient. Qualitative data were collected through individual, semi-structured interviews, which aimed to explore patients' experiences with their BA treatment and their Fitbit use. Additional data such as baseline BDI scores, pretherapy goals, familiarity with technology, and other patient characteristics were collected through questionnaires facilitated by a clinician during the study intake. BDI scores were collected weekly throughout the study including at completion. Exploratory analyses were conducted to examine how perceived benefit related to participants' baseline characteristics, including age, BDI scores, pretherapy physical activity goals, and smartphone use, as well as percent change in BDI scores from baseline to completion and Fitbit usage. Statistical Analysis: The control and intervention groups were examined through chi-square tests. To examine the relationships between baseline characteristics and Fitbit use, independent t-tests were used. Between group differences were also calculated through independent t-tests. Pearson's correlation was used to compare Fitbit usage with age and baseline BDI scores.

Data Collection Summary	Timing of Measurements: Trained clinicians delivered BA to participants randomized to receive the intervention up to twice per week over 18 weeks, for a total of 28 sessions. Patients were asked to carry a Fitbit One at all times for the 18-week intervention period. Individual interviews lasted 30 to 60 minutes and were finished within two months of participants' completion of the BA program. Dependent Variables: Patients' perceived benefits, Fitbit usage, BDI scores Independent Variables: BA topics, such as value assessment, goal setting, leisure education Control Variables: Receiving treatment for depression as usual, including pharmacotherapy and psychotherapy
Description of Actual Data Sample	Initial: 36 (18 Males 18 Females) Attrition (final N): 36 Age: Mean age: 53 years Ethnicity: Unclear Other relevant demographics: Mean BDI score: 36.27 (scale of 0 to 63), 23 participants had prior smartphone use, 26 participants attended a college, university, or higher education Anthropometrics: Mean BMI: 30.16 kg/m2 Location: Mood disorders outpatient clinic in Hamilton, Ontario, Canada
Summary of Results	Key Findings: 23 patients found the Fitbit One to be helpful for their physical activity. Themes of positive experiences included 27 patients who reported self-awareness, three patients who reported peer motivation, five patients who reported goal-setting opportunities, and 13 patients who reported enjoying using the Fitbit One. Themes of negative experiences included 12 patients who reported inconvenience, 11 patients who reported inaccuracies, two patients who reported discouragement, and 10 patients who reported disinterest. No significant differences were found in age ($p=0.72$), baseline BDI scores ($p=0.44$), percent change in BDI

	scores (p=0.283), pretherapy physical activity goals (p=0.549), and smartphone use (p=0.825) between those who did and did not find the Fitbit One helpful. However, there was a significant relationship between total Fitbit One usage and perceived benefit. The mean number of weeks of Fitbit One use for those who found the Fitbit helpful was 18.57 (SD 1.21) and 12.27 (SD 5.76) weeks for those who did not (p<0.001). Other Findings:
Author Conclusion	Of the 36 patients who underwent the BRAVE study and completed interviews, the majority (64%) found the Fitbit One to be helpful, identifying their Fitbit use as a factor in improving their physical activity level. Many positive themes were concordant with current literature; however, patients also reported negative aspects that may affect use. Interestingly, there was no significant relationship between perceived benefit of the Fitbit and percent change in BDI scores, which contradicts previous literature supporting Fitbit use in treating depression. These findings suggest that the Fitbit One may be useful for patients with varying characteristics. Clinicians and researchers should consider both strengths and limitations of activity trackers when implementing them to motivate patients with depression.
Reviewer Comments	*Strengths: zero percent attrition, use of baseline data and final interviews for detailed participant feedback, and use of Fitbit data to corroborate adherence* *Limitations: small sample size compared to the number of participants who completed the BRAVE study, difficulty interpreting exploratory findings given that the study was underpowered to test effectiveness and there were several confounders such as combined BA intervention, study setting, and participants' restricted use of Fitbit*
Funding Source	Supported by Canadian Institutes of Health Research (CIHR) (Randomised Controlled Trials: Mentoring, code number 201303MTP-303860-182743) to conduct the BRAVE study. CIHR mentorship award

	(2013–2015) is to support research time. The authors declared no competing interests.

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive** – Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative** – Indicates that these issues have not been adequately addressed.*
⊘	***Neutral** – indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Yes
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	No
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Yes
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Yes
2.2. Were criteria applied equally to all study groups?	2.2	Yes
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	Yes
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Unclear
3. Were <u>study groups comparable</u>?	3	Unclear
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	Unclear
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	Unclear

3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	No
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	N/A
3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.)	3.5	N/A
3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	N/A
4. Was method of handling <u>withdrawals</u> described?	4	N/A
4.1. Were follow up methods described and the same for all groups?	4.1	N/A
4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)	4.2	N/A
4.3. Were all enrolled subjects/patients (in the original sample) accounted for?	4.3	No
4.4. Were reasons for withdrawals similar across groups	4.4	N/A
4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4.5	N/A
5. Was <u>blinding</u> used to prevent introduction of bias?	5	Yes
5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate?	5.1	Yes
5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.)	5.2	Yes
5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	5.3	N/A
5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	5.4	N/A
5.5. In diagnostic study, were test results blinded to patient history and other test results?	5.5	N/A
6. Were <u>intervention</u>/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were <u>intervening factors</u> described?	6	Yes
6.1. In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	Yes
6.2. In observational study, were interventions, study settings, and clinicians/provider described?	6.2	N/A
6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	6.3	Yes
6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured?	6.4	Yes
6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described?	6.5	Yes
6.6. Were extra or unplanned treatments described?	6.6	N/A
6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups?	6.7	N/A
6.8. In diagnostic study, were details of test administration and replication sufficient?	6.8	N/A
7. Were <u>outcomes</u> clearly defined and the <u>measurements valid and reliable</u>?	7	Unclear
7.1. Were primary and secondary endpoints described and relevant to the question?	7.1	Yes
7.2. Were nutrition measures appropriate to question and outcomes of concern?	7.2	N/A

7.3. Was the period of follow-up long enough for important outcome(s) to occur?	7.3	N/A
7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures?	7.4	Unclear
7.5. Was the measurement of effect at an appropriate level of precision?	7.5	No
7.6. Were other factors accounted for (measured) that could affect outcomes?	7.6	No
7.7. Were the measurements conducted consistently across groups?	7.7	N/A

8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?	8	Unclear
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Yes
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes
8.3. Were statistics reported with levels of significance and/or confidence intervals?	8.3	Yes
8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)?	8.4	No
8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)?	8.5	No
8.6. Was clinical significance as well as statistical significance reported?	8.6	Yes
8.7. If negative findings, was a power calculation reported to address type 2 error?	8.7	N/A
9. **Are <u>conclusions supported by results</u> with biases and limitations taken into consideration?**	9	Yes
9.1. Is there a discussion of findings?	9.1	Yes
9.2. Are biases and study limitations identified and discussed?	9.2	Yes
10. Is bias due to study's <u>funding or sponsorship</u> unlikely?	10	Yes
10.1. Were sources of funding and investigators' affiliations described?	10.1	Yes
10.2. Was there no apparent conflict of interest?	10.2	Yes

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.

NEUTRAL (∅)
If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.

PLUS/POSITIVE (+)
If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Cook, J., Prairie, M., & Plante, D. (2017). Utility of the Fitbit Flex to Evaluate Sleep in Major Depressive Disorder: A comparison against polysomnography and wrist-worn actigraphy. *Journal of Affective Disorders*, *217*, 299–305. doi: 10.1016/j.jad.2017.04.030
Study design: *Use algorithm – RCT, cohort, etc*	Validity study
Study Class (A,B,C,D)	C
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**POSITIVE (+)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To evaluate the utility of the Fitbit Flex (FBF) to estimate sleep in a well-characterized cohort of adult patients with major depressive disorder (MDD) relative to gold standard polysomnography (PSG) and validated actigraphy (Actiwatch-2; AW-2).
Inclusion criteria: *requirements for study eligibility*	Right-handed, unmedicated patients with unipolar MDD who completed an initial phone screening, in-person medical, sleep, and psychiatric evaluation. If patients met criteria for other psychiatric disorders, MDD had to be considered the primary disorder for study inclusion.
Exclusion criteria (conditions that make individual ineligible)	Smoking of >15 cigarettes per day; >3 caffeinated beverages per day; significant sleep, neurologic, or medical disorder; history of significant head trauma or loss of consciousness >30 minutes; and imminent risk of self-harm or suicide. Women who were pregnant, breastfeeding, <6 months post-partum, or planning to become pregnant during the study were excluded. Participants were also excluded if they met DSM-IV criteria for alcohol or substance abuse/dependence within the preceding 6 months.
Recruitment	A convenience sample of patients with MDD was recruited as part of a larger study investigating electroencephalographic biomarkers of sleep disturbance in neuropsychiatric disorders
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	A registered sleep technologist, blind to the FBF and AW-2 staging output, staged all sleep recordings according to standard criteria according to American Academy of Sleep Medicine criteria
Description of study protocol *What happened in the study?*	Twenty-one patients with unipolar MDD wore the FBF and AW-2 during in-laboratory PSG
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	After an initial phone screening, participants completed an in-person medical, sleep, and psychiatric evaluation that included the Structured Clinical Interview for DSM-IV (SCID), semi-structured sleep disorders evaluation, and physical exam. Participants completed additional self-report instruments including the Beck Depression Inventory (BDI-II), Pittsburgh Sleep Quality Index (PSQI), and Insomnia Severity Index (ISI). Eligible participants were then scheduled for an in-laboratory PSG at least one week but no more than one month after their in-person screening visit. All participants were instructed to maintain their usual sleep-wake schedules for the duration of their time in the study. Participants arrived at

	approximately 18:00 on the night of their PSG for set-up. A wrist-worn AW-2 and FBF were both placed adjacently on the participant's non-dominant (left) wrist. Polysomnographic data were collected using an integrated recording system that utilized a 256-channel EEG net along with other standard recording sensors including electrooculogram (EOG), sub-mental electromyogram (EMG), electrocardiogram (ECG), bilateral tibial EMG, respiratory inductance plethysmography, pulse oximetry, and a position sensor. A registered sleep technologist, blind to the FBF and AW-2 staging output, staged all sleep recordings using 30-second epochs according to standard criteria according to American Academy of Sleep Medicine criteria. Bedtimes were tailored to each participant's habitual sleep pattern, with lights-off (participant actively trying to fall asleep) occurring between approximately 22:00 and 23:00. Participants were allowed to sleep ad libitum, remaining undisturbed throughout the night and not awoken at a prescribed time the following morning. Lights-on was determined based on the participant's stated desire to terminate the nocturnal sleep period upon awakening. Polysomnography and accelerometer data were collected within a local network of computers time synchronized to an external clock. The following sleep variables were calculated for PSG, FBF, and AW-2: total sleep time (TST), sleep onset latency (SOL), wake after sleep onset (WASO), and sleep efficiency (SE). AW-2 data were analyzed utilizing the medium threshold relative to PSG, and FBF data were analyzed using both the normal and sensitive settings relative to PSG since these settings have been shown to produce the most accurate outputs.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Bland-Altman analysis was utilized to calculate the mean difference between devices (AW-2 vs. PSG; FBF-N vs. PSG; FBF-N vs. AW-2; FBF-S vs. PSG; and FBF-S vs. AW-2) for each sleep variable (TST, SOL, WASO, SE). Epoch-by-epoch analysis further evaluated sensitivity, specificity, and accuracy for the FBF and AW-2 relative to PSG. Alpha equaled 0.05 for statistical significance for all comparisons.
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	Lights-off occurred between approximately 22:00 and 23:00. All sleep recordings were collected and staged in 30-second epochs according to standard criteria. Participants were not awoken at a prescribed time the following morning.
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	Sleep variables according to FBF and AW-2
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Participants' anthropometrics or stress levels
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	Sleep variables according to PSG
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	21 participants, 4 males and 17 females
Final n (attrition)	21 participants, 4 males and 17 females

number of subjects that completed study	
Age *usually mean or range*	Mean age = 26.5 ± 4.6 years
Ethnicity (if given)	Unclear
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	Mean BDI-II score = 22.9 ± 6.8 Mean PSQI score = 8.4 ± 2.5 Mean ISI score = 14.3 ± 5.6
Anthropometrics: *e.g. were groups same or different on important physical measures (BMI, fitness level)*	Unclear
Location: *Where did the study take place? City or country*	Wisconsin Sleep Center
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	When the AW-2 was compared to PSG, AW-2 significantly overestimated TST (mean difference of 40.6 min, p=0.0004) and SE (mean difference of 7.0%, p=0.0003), while significantly underestimating SOL (mean difference of −13.5 min, p=0.012) and WASO (mean difference of −27.1 min, p=0.005). When compared epoch-by-epoch against PSG, the AW-2 displayed relatively good sensitivity (0.97 ± 0.02) and accuracy (0.87 ± 0.06), with poor specificity (0.31 ± 0.15). When the FBF-N was compared to PSG, like the AW-2, FBF-N significantly overestimated TST (mean difference of 46.0 min, p<0.0001) and SE (mean difference of 8.1%, p<0.0001), while significantly underestimating WASO (mean difference of −44.0 min, p<0.0001). However, SOL assessed by FBF-N and PSG were quite similar (mean difference of −2.0 min, p=0.72). When compared epoch-by-epoch against PSG, again like the AW-2, the FBF-N showed a high sensitivity (0.98 ± 0.02) and accuracy (0.88 ± 0.05), with low specificity (0.35 ± 0.13). Direct comparison of the FBF-N to AW-2 demonstrated significantly higher estimates of SE (mean difference of 1.1%, p=0.042) and SOL (mean difference of 11.5 min, p=0.0003) for the FBF-N, as well as significantly lower estimates of WASO (mean difference of −16.9 min, p<0.0001). FBF-N and AW-2 had comparable estimates of TST (mean difference of 5.4 min, p=0.08). When the FBF-S was compared to PSG, findings were quite different from those derived using the normal mode for the device. Relative to PSG, FBF-S significantly underestimated TST (mean difference of −86.3 min, p<0.0001) and SE (mean difference of −16.0%, p<0.0001), while significantly overestimating SOL (mean difference of 11.5 min, p=0.012) and WASO (mean difference of 74.8 min, p<0.0001). When compared epoch-by-epoch against PSG, the FBF-S displayed a modest sensitivity (0.78 ± 0.09), specificity (0.80 ± 0.17), and accuracy (0.78 ± 0.08). Similarly, when the FBF-S was compared to the AW-2, FBF-S had significantly lower estimates of TST (mean difference of −126.8 min, p<0.0001) and SE (mean difference of −22.9%, p<0.0001) with significantly higher estimates of SOL (mean difference of 24.9 min, p=0.0006) and WASO (mean difference of 101.9 min, p<0.0001).
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	The FBF demonstrated significant limitations in quantifying sleep and wake, relative to PSG. In the normal setting, the FBF significantly overestimated sleep time and efficiency, and displayed low specificity. In the sensitive setting, the FBF significantly underestimated sleep time and efficiency relative to PSG. The FBF is not an adequate substitute for PSG when quantifying sleep in MDD, however, the FBF does demonstrate similar performance characteristics to a standard actigraph, particularly in the estimation of total sleep duration, when used in the normal mode. The capabilities, limitations, and settings of the FBF should be carefully considered prior to clinical and research implementation.
Reviewer comments: *Note strengths and limitations of study;*	Strengths: *zero percent attrition, use of a variety of initial screening methods, including physical examinations and validated questionnaires, and use of AW-2 as*

identify concerns that affect study validity and generalizability—your comments should be italicized)	*an alternate measure for sleep and wake to circumvent some of the shortcomings of PSG.* Limitations: participants were young to middle aged and predominantly female, which may limit generalizability of findings. Study specifically examined outpatients with MDD, thus findings cannot be directly extended to other mood and/or sleep disorders. Also, results cannot be extended to other fitness trackers, or more current generations of the same model as these devices may have different performance characteristics. The study design leaves in question the true capabilities of the FBF as a long-term sleep measurement device. Funding source: this work was supported by grants from the National Institute of Mental Health (K23MH099234), the Brain and Behavior Research Foundation, and the American Sleep Medicine Foundation. The sources of funding for this investigation had no further role in the study design, data collection, analysis and interpretation of the data, and the decision to submit the paper for publication. The study authors have no relationship with Fitbit, and Fitbit did not supply any funding, supplies, or guidance towards this investigation, or have any bearing on the decision to submit this manuscript for publication.

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Cook, J., Prairie, M., & Plante, D. (2017). Utility of the Fitbit Flex to Evaluate Sleep in Major Depressive Disorder: A comparison against polysomnography and wrist-worn actigraphy. *Journal of Affective Disorders*, *217*, 299–305. doi: 10.1016/j.jad.2017.04.030		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1			X	
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3		X		
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES X	NO	UNCLEAR	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1	X			
2.2 Were criteria applied equally to all study groups?	2.2				X
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3	X			
2.4 Were the subjects/patients in a representative sample of the relevant population? *The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.*	2.4			X	

3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES	NO	UNCLEAR	NA X
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1				X
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2				X
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3	X			
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6	X			
4. Was method of handling withdrawals described?		YES	NO	UNCLEAR	NA X
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2				X
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3	X			
4.4 Were reasons for withdrawals similar across groups? *If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.*	4.4				X

4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5	X			
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES X	NO	UNCLEAR	NA
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1				X
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2	X			
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3				X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4				X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5			X	
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES X	NO	UNCLEAR	NA
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1				X
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2	X			
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X			
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4	X			
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5				X
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6				X
6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7				X
6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8	X			

		YES	NO	UNCLEAR	NA
7. Were <u>outcomes</u> clearly defined and the measurements valid and reliable?		X			
7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant*.	7.1	X			
7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X
7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X
7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X			
7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5	X			
7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6	X			
7.7 Were the measurements conducted consistently across groups?	7.7				X
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?		YES X	NO	UNCLEAR	NA
8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X			
8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X			
8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X			
8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4				X
8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)? *Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.*	8.5				X
8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical*	8.6	X			

significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.						
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7					X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA	
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X				
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X				
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA	
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X				
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X				

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral* (ø) *symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Cook, J., Prairie, M., & Plante, D. (2017). Utility of the Fitbit Flex to Evaluate Sleep in Major Depressive Disorder: A comparison against polysomnography and wrist-worn actigraphy. Journal of Affective Disorders, 217, 299–305. doi: 10.1016/j.jad.2017.04.030
Study Design	Validity study
Class	C
Quality Rating	☒ + (Positive) ☐ - (Negative) ☐ ⊘ (Neutral)
Research Purpose	To evaluate the utility of the Fitbit Flex (FBF) to estimate sleep in a well-characterized cohort of adult patients with major depressive disorder (MDD) relative to gold standard polysomnography (PSG) and validated actigraphy (Actiwatch-2; AW-2).
Inclusion Criteria	Right-handed, unmedicated patients with unipolar MDD who completed an initial phone screening, in-person medical, sleep, and psychiatric evaluation. If patients met criteria for other psychiatric disorders, MDD had to be considered the primary disorder for study inclusion.
Exclusion Criteria	Smoking of >15 cigarettes per day; >3 caffeinated beverages per day; significant sleep, neurologic, or medical disorder; history of significant head trauma or loss of consciousness >30 minutes; and imminent risk of self-harm or suicide. Women who were pregnant, breastfeeding, <6 months post-partum, or planning to become pregnant during the study were excluded. Participants were also excluded if they met DSM-IV criteria for alcohol or substance abuse/dependence within the preceding 6 months.
Description of Study Protocol	Recruitment: A convenience sample of patients with MDD was recruited as part of a larger study investigating electroencephalographic biomarkers of sleep disturbance in neuropsychiatric disorders Design: Twenty-one patients with unipolar MDD wore the FBF and AW-2 during in-laboratory PSG Blinding used (if applicable): A registered sleep technologist, blind to the FBF and AW-2 staging output, staged all sleep recordings according to standard criteria according to American Academy of Sleep Medicine criteria Intervention (if applicable): After an initial phone screening, participants completed an in-person medical, sleep, and psychiatric evaluation that included the Structured Clinical Interview for DSM-IV (SCID), semi-

	structured sleep disorders evaluation, and physical exam. Participants completed additional self-report instruments including the Beck Depression Inventory (BDI-II), Pittsburgh Sleep Quality Index (PSQI), and Insomnia Severity Index (ISI). Eligible participants were then scheduled for an in-laboratory PSG at least one week but no more than one month after their in-person screening visit. All participants were instructed to maintain their usual sleep-wake schedules for the duration of their time in the study. Participants arrived at approximately 18:00 on the night of their PSG for set-up. A wrist-worn AW-2 and FBF were both placed adjacently on the participant's non-dominant (left) wrist. Polysomnographic data were collected using an integrated recording system that utilized a 256-channel EEG net along with other standard recording sensors including electrooculogram (EOG), sub-mental electromyogram (EMG), electrocardiogram (ECG), bilateral tibial EMG, respiratory inductance plethysmography, pulse oximetry, and a position sensor. A registered sleep technologist, blind to the FBF and AW-2 staging output, staged all sleep recordings using 30-second epochs according to standard criteria according to American Academy of Sleep Medicine criteria. Bedtimes were tailored to each participant's habitual sleep pattern, with lights-off (participant actively trying to fall asleep) occurring between approximately 22:00 and 23:00. Participants were allowed to sleep ad libitum, remaining undisturbed throughout the night and not awoken at a prescribed time the following morning. Lights-on was determined based on the participant's stated desire to terminate the nocturnal sleep period upon awakening. Polysomnography and accelerometer data were collected within a local network of computers time synchronized to an external clock. The following sleep variables were calculated for PSG, FBF, and AW-2: total sleep time (TST), sleep onset latency (SOL), wake after sleep onset (WASO), and sleep efficiency (SE). AW-2 data were analyzed utilizing the medium threshold relative to PSG, and FBF data were analyzed using both the normal and sensitive

	settings relative to PSG since these settings have been shown to produce the most accurate outputs. Statistical Analysis: Bland-Altman analysis was utilized to calculate the mean difference between devices (AW-2 vs. PSG; FBF-N vs. PSG; FBF-N vs. AW-2; FBF-S vs. PSG; and FBF-S vs. AW-2) for each sleep variable (TST, SOL, WASO, SE). Epoch-by-epoch analysis further evaluated sensitivity, specificity, and accuracy for the FBF and AW-2 relative to PSG. Alpha equaled 0.05 for statistical significance for all comparisons.
Data Collection Summary	Timing of Measurements: Lights-off occurred between approximately 22:00 and 23:00. All sleep recordings were collected and staged in 30-second epochs according to standard criteria. Participants were not awoken at a prescribed time the following morning. Dependent Variables: Sleep variables according to FBF and AW-2 Independent Variables: Participants' anthropometrics or stress levels Control Variables: Sleep variables according to PSG
Description of Actual Data Sample	Initial: 21 (4 Males 17 Females) Attrition (final N): 21 Age: Mean age = 26.5 ± 4.6 years Ethnicity: Unclear Other relevant demographics: Mean BDI-II score = 22.9 ± 6.8, mean PSQI score = 8.4 ± 2.5, mean ISI score = 14.3 ± 5.6 Anthropometrics: Unclear Location: Wisconsin Sleep Center
Summary of Results	Key Findings: When the AW-2 was compared to PSG, AW-2 significantly overestimated TST (mean difference of 40.6 min, p=0.0004) and SE (mean difference of 7.0%, p=0.0003), while significantly underestimating SOL (mean difference of −13.5 min, p=0.012) and WASO (mean difference of −27.1 min, p=0.005). When compared epoch-by-epoch against PSG, the AW-2 displayed relatively good

	sensitivity (0.97 ± 0.02) and accuracy (0.87 ± 0.06), with poor specificity (0.31 ± 0.15). When the FBF-N was compared to PSG, like the AW-2, FBF-N significantly overestimated TST (mean difference of 46.0 min, $p<0.0001$) and SE (mean difference of 8.1%, $p<0.0001$), while significantly underestimating WASO (mean difference of −44.0 min, $p<0.0001$). However, SOL assessed by FBF-N and PSG were quite similar (mean difference of −2.0 min, $p=0.72$). When compared epoch-by-epoch against PSG, again like the AW-2, the FBF-N showed a high sensitivity (0.98 ± 0.02) and accuracy (0.88 ± 0.05), with low specificity (0.35 ± 0.13). Direct comparison of the FBF-N to AW-2 demonstrated significantly higher estimates of SE (mean difference of 1.1%, $p=0.042$) and SOL (mean difference of 11.5 min, $p=0.0003$) for the FBF-N, as well as significantly lower estimates of WASO (mean difference of −16.9 min, $p<0.0001$). FBF-N and AW-2 had comparable estimates of TST (mean difference of 5.4 min, $p=0.08$). When the FBF-S was compared to PSG, findings were quite different from those derived using the normal mode for the device. Relative to PSG, FBF-S significantly underestimated TST (mean difference of −86.3 min, $p<0.0001$) and SE (mean difference of −16.0%, $p<0.0001$), while significantly overestimating SOL (mean difference of 11.5 min, $p=0.012$) and WASO (mean difference of 74.8 min, $p<0.0001$). When compared epoch-by-epoch against PSG, the FBF-S displayed a modest sensitivity (0.78 ± 0.09), specificity (0.80 ± 0.17), and accuracy (0.78 ± 0.08). Similarly, when the FBF-S was compared to the AW-2, FBF-S had significantly lower estimates of TST (mean difference of −126.8 min, $p<0.0001$) and SE (mean difference of −22.9%, $p<0.0001$) with significantly higher estimates of SOL (mean difference of 24.9 min, $p=0.0006$) and WASO (mean difference of 101.9 min, $p<0.0001$). Other Findings:

Author Conclusion	The FBF demonstrated significant limitations in quantifying sleep and wake, relative to PSG. In the normal setting, the FBF significantly overestimated sleep time and efficiency, and displayed low specificity. In the sensitive setting, the FBF significantly underestimated sleep time and efficiency relative to PSG. The FBF is not an adequate substitute for PSG when quantifying sleep in MDD, however, the FBF does demonstrate similar performance characteristics to a standard actigraph, particularly in the estimation of total sleep duration, when used in the normal mode. The capabilities, limitations, and settings of the FBF should be carefully considered prior to clinical and research implementation.
Reviewer Comments	*Strengths: zero percent attrition, use of a variety of initial screening methods, including physical examinations and validated questionnaires, and use of AW-2 as an alternate measure for sleep and wake to circumvent some of the shortcomings of PSG.* *Limitations: participants were young to middle aged and predominantly female, which may limit generalizability of findings. Study specifically examined outpatients with MDD, thus findings cannot be directly extended to other mood and/or sleep disorders. Also, results cannot be extended to other fitness trackers, or more current generations of the same model as these devices may have different performance characteristics. The study design leaves in question the true capabilities of the FBF as a long-term sleep measurement device.*
Funding Source	This work was supported by grants from the National Institute of Mental Health (K23MH099234), the Brain and Behavior Research Foundation, and the American Sleep Medicine Foundation. The sources of funding for this investigation had no further role in the study design, data collection, analysis and interpretation of the data, and the decision to submit the paper for publication. The study authors have no relationship with Fitbit, and Fitbit did not supply any funding, supplies, or guidance towards this investigation, or have any bearing on the decision to submit this manuscript for publication.

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive** – Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative** – Indicates that these issues have not been adequately addressed.*
⊘	***Neutral** – indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Unclear
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	No
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Yes
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Yes
2.2. Were criteria applied equally to all study groups?	2.2	N/A
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	Yes
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Unclear
3. Were <u>study groups comparable</u>?	3	N/A
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	N/A
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	N/A
3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	Yes
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	N/A

3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.)	3.5	N/A
3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	Yes
4. Was method of handling <u>withdrawals</u> described?	4	N/A
4.1. Were follow up methods described and the same for all groups?	4.1	N/A
4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)	4.2	N/A
4.3. Were all enrolled subjects/patients (in the original sample) accounted for?	4.3	Yes
4.4. Were reasons for withdrawals similar across groups	4.4	N/A
4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4.5	Yes
5. Was <u>blinding</u> used to prevent introduction of bias?	5	Yes
5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate?	5.1	N/A
5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.)	5.2	Yes
5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	5.3	N/A
5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	5.4	N/A
5.5. In diagnostic study, were test results blinded to patient history and other test results?	5.5	Unclear
6. Were <u>intervention</u>/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were <u>intervening factors</u> described?	6	Yes
6.1. In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	N/A
6.2. In observational study, were interventions, study settings, and clinicians/provider described?	6.2	Yes
6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	6.3	Yes
6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured?	6.4	Yes
6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described?	6.5	N/A
6.6. Were extra or unplanned treatments described?	6.6	N/A
6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups?	6.7	N/A
6.8. In diagnostic study, were details of test administration and replication sufficient?	6.8	Yes
7. Were <u>outcomes</u> clearly defined and the <u>measurements valid and reliable</u>?	7	Yes
7.1. Were primary and secondary endpoints described and relevant to the question?	7.1	Yes
7.2. Were nutrition measures appropriate to question and outcomes of concern?	7.2	N/A
7.3. Was the period of follow-up long enough for important outcome(s) to occur?	7.3	N/A
7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures?	7.4	Yes
7.5. Was the measurement of effect at an appropriate level of precision?	7.5	Yes
7.6. Were other factors accounted for (measured) that could affect outcomes?	7.6	Yes

7.7. Were the measurements conducted consistently across groups?	7.7	N/A
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?	8	Yes
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Yes
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes
8.3. Were statistics reported with levels of significance and/or confidence intervals?	8.3	Yes
8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)?	8.4	N/A
8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)?	8.5	N/A
8.6. Was clinical significance as well as statistical significance reported?	8.6	Yes
8.7. If negative findings, was a power calculation reported to address type 2 error?	8.7	N/A
9. **Are <u>conclusions supported by results</u> with biases and limitations taken into consideration?**	9	Yes
9.1. Is there a discussion of findings?	9.1	Yes
9.2. Are biases and study limitations identified and discussed?	9.2	Yes
10. Is bias due to study's <u>funding or sponsorship</u> unlikely?	10	Yes
10.1. Were sources of funding and investigators' affiliations described?	10.1	Yes
10.2. Was there no apparent conflict of interest?	10.2	Yes

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.

NEUTRAL (∅)
If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.

PLUS/POSITIVE (+)
If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Ferguson, T., Rowlands, A., Olds, T., & Maher, C. (2015). The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. *International Journal of Behavioral Nutrition and Physical Activity, 12*, 42. doi:10.1186/s12966-015-0201-9
Study design: *Use algorithm – RCT, cohort, etc*	Validity study
Study Class (A,B,C,D)	C
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**POSITIVE (+)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To assess the concurrent validity of a selection of consumer-level accelerometer-based activity monitors compared to two research-grade accelerometers in free living conditions
Inclusion criteria: *requirements for study eligibility*	Healthy adult participants over 18 years of age, living in metropolitan Adelaide, South Australia, and could ambulate without walking aids
Exclusion criteria (conditions that make individual ineligible)	Self-reported injury or illness affecting mobility
Recruitment	Unclear, convenience sample
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	NA
Description of study protocol *What happened in the study?*	21 healthy adult participants wore seven consumer-level activity monitors (Fitbit One, Fitbit Zip, Jawbone UP, Misfit Shine, Nike Fuelband, Striiv Smart Pedometer, and Withings Pulse) and two research-grade accelerometers/multi-sensor devices (BodyMedia SenseWear, and ActiGraph GT3X+) simultaneously for 48-hours. Participants went about their daily life in free-living conditions during data collection. Four physical activity parameters were measured, including step count, moderate to vigorous physical activity (MVPA), sleep, and total daily energy expenditure (TDEE).
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	Seven consumer-level activity monitors (Fitbit One, Fitbit Zip, Jawbone UP, Misfit Shine, Nike Fuelband, Striiv Smart Pedometer, and Withings Pulse) were examined and chosen based on availability to the authors for purchase between February and August 2013. The consumer-level devices were compared with two research grade tri-axial accelerometers/multi-sensor devices (BodyMedia SenseWear Model MF and ActiGraph GT3X+), which have accepted reliability and validity as free-living measures of physical activity and sleep time. Participants attended an appointment at which demographic data (date of birth, sex, and dominant side) were obtained, with height and mass measured following standardized procedures. All nine devices were fitted to the participant in the following locations: SenseWear on the left upper arm; Fuelband, UP and Shine on the left wrist; GT3X+, One, Zip, Pulse and Striiv on the right side of the waist on an elasticized belt. Where consumer-level devices were designed for multiple wear locations, devices suitable for wrist

	wear were worn on the wrist; otherwise the device was worn on the waist. Placement order of the devices at the wrist and waist was randomized. Participants were instructed to leave all devices on simultaneously for approximately 48 hours (including sleep, but excluding showering) in order to capture a full overnight sleep episode as well as two 24-hours of activity data from midnight to midnight. The wear period was not limited to a particular period of the week (i.e. not restricted to weekdays only or weekends only), and no guidelines or restrictions on activity levels or sleep were provided, in order to ensure the study broadly represented free-living conditions. Participants were instructed how to turn sleep mode on and off for the relevant devices (Shine, Pulse, One, UP). Participants were not given access to any of the device software or account information and were also instructed not to turn off, modify, or change any device wear locations once fitted. Devices were collected after the 48-hour wear period for data collection. Data were extracted using the proprietary software for all consumer devices, in the same fashion that a consumer would utilize the software.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Participants' demographic data were analyzed descriptively. The validity of the consumer-level activity monitors relative to the research-grade devices for step count, MVPA, sleep, and TDEE was quantified using Bland-Altman analysis, median absolute difference, and Pearson's correlation. A priori power analyses were undertaken based on existing data on correlations among various research devices, which suggested that the correlation between consumer-level and research devices would be about 0.85. If the actual population correlation between consumer-level and research-grade devices was 0.85, then a target sample size of 21 would yield, in 95% of cases, a sample correlation between 0.65 and 0.94.
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	Data were collected continuously throughout the 48-hour wear period from all nine devices. However, data relating to physical activity were limited to the full calendar day (24-hour period midnight to midnight) following initialization, and data relating to sleep were limited to the first night of sleep (24-hour period midday to midday, excluding naps) following initialization.
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	Physical activity parameters (step count, MVPA, sleep, and TDEE) according to Fitbit One, Fitbit Zip, Jawbone UP, Misfit Shine, Nike Fuelband, Striiv Smart Pedometer, and Withings Pulse
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Free living conditions (e.g. participants' daily obligations, lifestyles, level of physical fitness, stress levels)
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	Physical activity parameters (step count, MVPA, sleep, and TDEE) according to BodyMedia SenseWear and ActiGraph GT3X+
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	21 participants, 10 males and 11 females
Final n (attrition) *number of subjects that completed study*	21 participants, 10 males and 11 females
Age *usually mean or range*	20 to 59 years

Ethnicity (if given)	Unclear
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	All participants were right hand dominant
Anthropometrics: *e.g. were groups same or different on important physical measures (BMI, fitness level)*	Male BMI: 27.3 ± 3.2 kg/m2 Female BMI: 25.5 ± 5.2 kg/m2
Location: *Where did the study take place? City or country*	Metropolitan Adelaide, South Australia
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	All consumer-level activity monitors measured steps, and correlations with reference devices were very strong (r = 0.94-0.99). Bland-Altman analyses suggested that three of the activity monitors slightly over-counted (Striiv, Zip, One) while four under-counted (Fuelband, Shine, Up, Pulse). Five of the activity monitors (Striiv, Shine, Up, Zip, One) were considered to measure a parameter similar or equivalent to MVPA time. Correlations between readings from the activity monitors and reference devices ranged from weak to strong (r = 0.52-0.91). Bland-Altman analyses showed large differences between the mean values reported. For example, the Shine under-counted (mean = 53.3 min of MVPA compared to reference device (GT3X+) mean = 58.5 min), while the Striiv over-counted (mean = 249 min of MVPA compared to reference device (GT3X+)). Of the five activity monitors (Shine, Up, Pulse, Zip, One) that measured TDEE, correlations with the reference devices were moderate to strong (r = 0.74- 0.81). Bland-Altman analyses suggest all activity monitors considerably underestimated TDEE compared to the reference device (SenseWear, mean = 3005 kcal), ranging from 475 kcal (One) to 898 kcal (UP). Of the four activity monitors (Shine, Up, Pulse, One) that measured minutes of sleep, all correlated strongly with the reference device (r = 0.82-0.92). Bland-Altman analyses showed all activity monitors overestimated minutes of sleep, most notably, the Shine (mean = 44 min) compared to reference device (SenseWear) mean = 424 min).
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	In free-living conditions, the consumer-level activity monitors showed strong validity for the measurement of steps and sleep duration, and moderate-to-strong valid for measurement of TDEE and MVPA. Median absolute differences were generally modest for sleep and steps, moderate for TDEE, and large for MVPA. Validity for each construct ranged widely between activity monitors, with the Fitbit One, Fitbit Zip, and Withings Pulse being the strongest performers.
Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: *zero percent attrition, the use of numerous consumer and reference devices, testing the devices in free-living conditions as they are designed for, and examining several different physical activity variables collected by the devices* Limitations: *participant recruitment was not discussed, blinding was not utilized, and validity may vary if activity monitors are worn in locations other than the hip or wrist* Funding source: unclear, authors declared no competing interests

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Ferguson, T., Rowlands, A., Olds, T., & Maher, C. (2015). The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. *International Journal of Behavioral Nutrition and Physical Activity, 12*, 42. doi:10.1186/s12966-015-0201-9		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1			X	
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3		X		
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES X	NO	UNCLEAR	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1	X			
2.2 Were criteria applied equally to all study groups?	2.2				X
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3	X			
2.4 Were the subjects/patients in a representative sample of the relevant population? *The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.*	2.4			X	

3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES	NO	UNCLEAR	NA X
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1				X
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2				X
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3	X			
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6	X			
4. Was method of handling withdrawals described?		YES	NO	UNCLEAR	NA X
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2				X
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3	X			
4.4 Were reasons for withdrawals similar across groups? *If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.*	4.4				X

4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5	X			
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES	NO	UNCLEAR	NA
				X	
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1				X
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2			X	
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3				X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4				X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5			X	
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES	NO	UNCLEAR	NA
		X			
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1				X
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2	X			
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X			
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4	X			
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5				X
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6				X
6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7				X
6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8	X			

		YES	NO	UNCLEAR	NA
7. Were <u>outcomes</u> clearly defined and the measurements valid and reliable?		X			
7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant*.	7.1	X			
7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X
7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X
7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X			
7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5	X			
7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6		X		
7.7 Were the measurements conducted consistently across groups?	7.7				X
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?		YES X	NO	UNCLEAR	NA
8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X			
8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X			
8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X			
8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4				X
8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)? *Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.*	8.5				X
8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical*	8.6	X			

significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.						
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7					X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA	
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X				
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X				
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA	
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X				
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X				

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral* (ø) *symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Ferguson, T., Rowlands, A., Olds, T., & Maher, C. (2015). The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. International Journal of Behavioral Nutrition and Physical Activity, 12, 42. doi:10.1186/s12966-015-0201-9
Study Design	Validity study
Class	C
Quality Rating	☒ + (Positive) ☐ - (Negative) ☐ ⊘ (Neutral)
Research Purpose	To assess the concurrent validity of a selection of consumer-level accelerometer-based activity monitors compared to two research-grade accelerometers in free living conditions
Inclusion Criteria	Healthy adult participants over 18 years of age, living in metropolitan Adelaide, South Australia, and could ambulate without walking aids
Exclusion Criteria	Self-reported injury or illness affecting mobility
Description of Study Protocol	Recruitment: Unclear, convenience sample Design: 21 healthy adult participants wore seven consumer-level activity monitors (Fitbit One, Fitbit Zip, Jawbone UP, Misfit Shine, Nike Fuelband, Striiv Smart Pedometer, and Withings Pulse) and two research-grade accelerometers/multi-sensor devices (BodyMedia SenseWear, and ActiGraph GT3X+) simultaneously for 48-hours. Participants went about their daily life in free-living conditions during data collection. Four physical activity parameters were measured, including step count, moderate to vigorous physical activity (MVPA), sleep, and total daily energy expenditure (TDEE). Blinding used (if applicable): NA Intervention (if applicable): Seven consumer-level activity monitors (Fitbit One, Fitbit Zip, Jawbone UP, Misfit Shine, Nike Fuelband, Striiv Smart Pedometer, and Withings Pulse) were examined and chosen based on availability to the authors for purchase between February and August 2013. The consumer-level devices were compared with two research grade tri-axial accelerometers/multi-sensor devices (BodyMedia SenseWear Model MF and ActiGraph GT3X+), which have accepted

	reliability and validity as free-living measures of physical activity and sleep time. Participants attended an appointment at which demographic data (date of birth, sex, and dominant side) were obtained, with height and mass measured following standardized procedures. All nine devices were fitted to the participant in the following locations: SenseWear on the left upper arm; Fuelband, UP and Shine on the left wrist; GT3X+, One, Zip, Pulse and Striiv on the right side of the waist on an elasticized belt. Where consumer-level devices were designed for multiple wear locations, devices suitable for wrist wear were worn on the wrist; otherwise the device was worn on the waist. Placement order of the devices at the wrist and waist was randomized. Participants were instructed to leave all devices on simultaneously for approximately 48 hours (including sleep, but excluding showering) in order to capture a full overnight sleep episode as well as two 24-hours of activity data from midnight to midnight. The wear period was not limited to a particular period of the week (i.e. not restricted to weekdays only or weekends only), and no guidelines or restrictions on activity levels or sleep were provided, in order to ensure the study broadly represented free-living conditions. Participants were instructed how to turn sleep mode on and off for the relevant devices (Shine, Pulse, One, UP). Participants were not given access to any of the device software or account information and were also instructed not to turn off, modify, or change any device wear locations once fitted. Devices were collected after the 48-hour wear period for data collection. Data were extracted using the proprietary software for all consumer devices, in the same fashion that a consumer would utilize the software. Statistical Analysis: Participants' demographic data were analyzed descriptively. The validity of the consumer-level activity monitors relative to the research-grade devices for step count, MVPA, sleep, and TDEE was quantified using Bland-Altman analysis, median absolute difference, and Pearson's correlation. A priori power analyses were

	undertaken based on existing data on correlations among various research devices, which suggested that the correlation between consumer-level and research devices would be about 0.85. If the actual population correlation between consumer-level and research-grade devices was 0.85, then a target sample size of 21 would yield, in 95% of cases, a sample correlation between 0.65 and 0.94.
Data Collection Summary	Timing of Measurements: Data were collected continuously throughout the 48-hour wear period from all nine devices. However, data relating to physical activity were limited to the full calendar day (24-hour period midnight to midnight) following initialization, and data relating to sleep were limited to the first night of sleep (24-hour period midday to midday, excluding naps) following initialization. Dependent Variables: Physical activity parameters (step count, MVPA, sleep, and TDEE) according to Fitbit One, Fitbit Zip, Jawbone UP, Misfit Shine, Nike Fuelband, Striiv Smart Pedometer, and Withings Pulse Independent Variables: Free living conditions (e.g. participants' daily obligations, lifestyles, level of physical fitness, stress levels) Control Variables: Physical activity parameters (step count, MVPA, sleep, and TDEE) according to BodyMedia SenseWear and ActiGraph GT3X+
Description of Actual Data Sample	Initial: 21 (10 Males 11 Females) Attrition (final N): 21 Age: 20 to 59 years Ethnicity: Unclear Other relevant demographics: All participants were right hand dominant Anthropometrics: Male BMI: 27.3 ± 3.2 kg/m2, female BMI: 25.5 ± 5.2 kg/m2 Location: Metropolitan Adelaide, South Australia
Summary of Results	Key Findings: All consumer-level activity monitors measured steps, and correlations with reference devices were very strong (r = 0.94-0.99). Bland-Altman analyses suggested that three of the activity monitors

	slightly over-counted (Striiv, Zip, One) while four under-counted (Fuelband, Shine, Up, Pulse). Five of the activity monitors (Striiv, Shine, Up, Zip, One) were considered to measure a parameter similar or equivalent to MVPA time. Correlations between readings from the activity monitors and reference devices ranged from weak to strong (r = 0.52-0.91). Bland-Altman analyses showed large differences between the mean values reported. For example, the Shine under-counted (mean = 53.3 min of MVPA compared to reference device (GT3X+) mean = 58.5 min), while the Striiv over-counted (mean = 249 min of MVPA compared to reference device (GT3X+)). Of the five activity monitors (Shine, Up, Pulse, Zip, One) that measured TDEE, correlations with the reference devices were moderate to strong (r = 0.74- 0.81). Bland-Altman analyses suggest all activity monitors considerably underestimated TDEE compared to the reference device (SenseWear, mean = 3005 kcal), ranging from 475 kcal (One) to 898 kcal (UP). Of the four activity monitors (Shine, Up, Pulse, One) that measured minutes of sleep, all correlated strongly with the reference device (r = 0.82-0.92). Bland-Altman analyses showed all activity monitors overestimated minutes of sleep, most notably, the Shine (mean = 44 min) compared to reference device (SenseWear) mean = 424 min). Other Findings:
Author Conclusion	In free-living conditions, the consumer-level activity monitors showed strong validity for the measurement of steps and sleep duration, and moderate-to-strong valid for measurement of TDEE and MVPA. Median absolute differences were generally modest for sleep and steps, moderate for TDEE, and large for MVPA. Validity for each construct ranged widely between activity monitors, with the Fitbit One, Fitbit Zip, and Withings Pulse being the strongest performers.
Reviewer Comments	*Strengths: zero percent attrition, the use of numerous consumer and reference devices, testing the devices in free-living conditions as they are*

	designed for, and examining several different physical activity variables collected by the devices *Limitations: participant recruitment was not discussed, blinding was not utilized, and validity may vary if activity monitors are worn in locations other than the hip or wrist*
Funding Source	Unclear, authors declared no competing interests

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive** – Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative** – Indicates that these issues have not been adequately addressed.*
⊘	***Neutral** – indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Unclear
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	No
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Yes
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Yes
2.2. Were criteria applied equally to all study groups?	2.2	N/A
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	Yes
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Unclear

3. Were study groups comparable? 3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) 3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? 3.3. Were concurrent controls used? (Concurrent preferred over historical controls.) 3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? 3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.) 3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3 3.1 3.2 3.3 3.4 3.5 3.6	N/A N/A N/A Yes N/A N/A Yes
4. Was method of handling withdrawals described? 4.1. Were follow up methods described and the same for all groups? 4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.) 4.3. Were all enrolled subjects/patients (in the original sample) accounted for? 4.4. Were reasons for withdrawals similar across groups 4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4 4.1 4.2 4.3 4.4 4.5	N/A N/A N/A Yes N/A Yes
5. Was blinding used to prevent introduction of bias? 5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate? 5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) 5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? 5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status? 5.5. In diagnostic study, were test results blinded to patient history and other test results?	5 5.1 5.2 5.3 5.4 5.5	Unclear N/A Unclear N/A N/A Unclear
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described? 6.1. In RCT or other intervention trial, were protocols described for all regimens studied? 6.2. In observational study, were interventions, study settings, and clinicians/provider described? 6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? 6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured? 6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described? 6.6. Were extra or unplanned treatments described? 6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups?	6 6.1 6.2 6.3 6.4 6.5 6.6 6.7	Yes N/A Yes Yes Yes N/A N/A N/A

6.8. In diagnostic study, were details of test administration and replication sufficient?	6.8	Yes
7. Were <u>outcomes</u> clearly defined and the <u>measurements valid and reliable</u>?	7	Yes
7.1. Were primary and secondary endpoints described and relevant to the question?	7.1	Yes
7.2. Were nutrition measures appropriate to question and outcomes of concern?	7.2	N/A
7.3. Was the period of follow-up long enough for important outcome(s) to occur?	7.3	N/A
7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures?	7.4	Yes
7.5. Was the measurement of effect at an appropriate level of precision?	7.5	Yes
7.6. Were other factors accounted for (measured) that could affect outcomes?	7.6	No
7.7. Were the measurements conducted consistently across groups?	7.7	N/A
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?	8	Yes
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Yes
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes
8.3. Were statistics reported with levels of significance and/or confidence intervals?	8.3	Yes
8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)?	8.4	N/A
8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)?	8.5	N/A
8.6. Was clinical significance as well as statistical significance reported?	8.6	Yes
8.7. If negative findings, was a power calculation reported to address type 2 error?	8.7	N/A
9. **Are <u>conclusions supported by results</u> with biases and limitations taken into consideration?**	9	Yes
9.1. Is there a discussion of findings?	9.1	Yes
9.2. Are biases and study limitations identified and discussed?	9.2	Yes
10. Is bias due to study's <u>funding or sponsorship</u> unlikely?	10	Yes
10.1. Were sources of funding and investigators' affiliations described?	10.1	Yes
10.2. Was there no apparent conflict of interest?	10.2	Yes

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.

NEUTRAL (∅)
If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.

PLUS/POSITIVE (+)
If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Gomersall, S., Ng, N., Burton, N., Pavey, T., Gilson, N., & Brown, W. (2016). Estimating Physical Activity and Sedentary Behavior in a Free-Living Context: A Pragmatic Comparison of Consumer-Based Activity Trackers and ActiGraph Accelerometry. *Journal of Medical Internet Research*, *18*(9), e239. doi:10.2196/jmir.5531
Study design: *Use algorithm – RCT, cohort, etc*	Validity study
Study Class (A,B,C,D)	C
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅) This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**NEUTRAL (ø)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To compare Fitbit One and Jawbone UP estimates of steps, moderate-to-vigorous physical activity (MVPA), and sedentary behavior with data from the ActiGraph GT3X+ accelerometer in a free-living context
Inclusion criteria: *requirements for study eligibility*	Healthy, ambulatory, adult participants between 18 and 65 years of age, have accumulated less than 150 minutes of MVPA in the past week (assessed using the Active Australia Survey), and own or have access to a mobile phone compatible with both the Fitbit One and Jawbone UP
Exclusion criteria (conditions that make individual ineligible)	Unclear
Recruitment	Convenience sampling at three campuses of a large Australian metropolitan university via an email advertisement to staff that included study information and participant eligibility criteria
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	Data were extracted via the users' accounts and entered into an Excel spreadsheet by a research assistant
Description of study protocol *What happened in the study?*	On two occasions for seven days each, participants wore an ActiGraph GT3X+ accelerometer on their right hip and either a hip-worn Fitbit One or wrist-worn Jawbone UP activity tracker. Daily estimates of steps and very active minutes were derived from the Fitbit One, and steps, active time, and longest idle time were derived from the Jawbone UP. Daily estimates of steps, MVPA, and longest sedentary bout were derived from the corresponding days of ActiGraph data.
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	Data were collected as part of a larger, 12-week physical activity intervention study that included three groups that were randomly allocated to wear a Fitbit One, Jawbone UP, or standard pedometer. Demographic and anthropometric data were collected at baseline. Data for this substudy were collected at mid- and post-intervention when participants concurrently wore an ActiGraph GT3X+ accelerometer. On these two occasions for seven days each, participants wore an ActiGraph GT3X+ accelerometer on their right hip and either a hip-worn Fitbit One (n=14) or wrist-worn Jawbone UP (n=15) activity tracker. Participants were instructed to wear the devices during waking hours, removing them for water-based activities or contact sports, but were not required to keep wear logs in order to

	improve the free-living fidelity of the devices over the 12-week intervention. Participants were able to input activity sessions, such as swimming or contact sports, through the "log workout" function in the Jawbone UP app and the "track exercise" feature in the Fitbit app. Daily estimates of steps and very active minutes were derived from the Fitbit One, and steps, active time, and longest idle time were derived from the Jawbone UP. Daily estimates of steps, MVPA, and longest sedentary bout were derived from the corresponding days of ActiGraph data. Data were extracted via the users' accounts and entered into an Excel spreadsheet by a research assistant.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Descriptive statistics (n, mean, standard deviation, and prevalence) were calculated for demographic and physical measures. Absolute agreement was examined using intraclass correlation coefficients (ICC) and 95% confidence intervals. Correlation was assessed using Pearson correlation coefficient or Spearman rank correlation coefficient when data were non-normally distributed with 95% confidence intervals. Bland-Altman plots were used to examine the differences between all outcomes, with mean bias and 95% limits of agreement reported. Linear regression was used to examine whether mean difference and limits of agreement varied across mean values of Fitbit One or Jawbone UP and ActiGraph outcomes. Cohen's kappa statistic was used to assess the agreement between devices for classification of active versus inactive based first on achieving 10,000 steps or more per day (default step goal on both devices) and second on achieving 30 minutes per day or more of MVPA (comparable with public health guidelines). *P* values were based on two-sided tests and were considered statistically significant at $P<.05$. Post hoc power calculations determined that a sample size of N=289 daily comparisons would detect correlations as low as .17 with 80% power and 5% alpha.
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	Data was collected continuously as participants were instructed to wear either a Fitbit One or Jawbone UP activity tracker every day during the 12-week physical activity intervention. Daily estimates of steps, active time, and longest idle time were extracted from the users' accounts. Participants concurrently wore an ActiGraph GT3X+ accelerometer at mid- and post-intervention for seven days. Daily estimates of steps, MVPA, and longest sedentary bout were derived from the corresponding days of ActiGraph data.
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	Steps and very active minutes according to the Fitbit One Steps, active time, and longest idle time according to the Jawbone UP
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Free living context (e.g. participants' daily obligations, lifestyles, level of physical fitness, stress levels)
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	Steps, MVPA, and longest sedentary bout according to the ActiGraph GT3X+ accelerometer
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	32 participants (only 29 provided valid data for the current analyses)

Final n (attrition) *number of subjects that completed study*	29 participants, 3 males and 26 females
Age *usually mean or range*	Mean age: 39.6, SD: 11.0 years
Ethnicity (if given)	Unclear
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	25 participants (86%) completed tertiary education
Anthropometrics: *e.g. were groups same or different on important physical measures (BMI, fitness level)*	Mean BMI: 25.9, SD: 5.0 kg/m2
Location: *Where did the study take place? City or country*	Australia
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	Correlations for steps and MVPA were strong for both devices, although higher for Fitbit One (r=.85 for steps and ρ=.80 for MVPA) than for Jawbone UP (r=.75 for steps and ρ=.75 for MVPA). The correlation between Jawbone UP longest idle time and ActiGraph longest sedentary bout was poor (ρ=.19). Absolute agreement (ICC) was acceptable for ActiGraph and Fitbit One steps (.90) and MVPA (.72) and Jawbone UP steps (.79). However, agreement was weak between ActiGraph and Jawbone UP estimates of MVPA (.56) and longest idle time (.08). For the estimation of steps, 95% limits of agreement were unbiased for both devices, although limits were wider for Jawbone UP than for Fitbit One (5290 and 3567 steps/day). When absolute values were calculated, both devices overestimated steps (Fitbit One: mean bias 767, 95% limits of agreement –2800 to 4334; Jawbone UP: mean bias 1178, 95% limits of agreement –4112 to 6468). For the estimation of MVPA, bias was evident for both the mean difference and the limits of agreement for both the Fitbit One and the Jawbone UP. When absolute values were calculated, the Fitbit One underestimated MVPA by a mean 19.2 minutes/day (95% limits of agreement –39.2 to 5.5), whereas the Jawbone UP overestimated by a mean of 38.1 minutes/day (95% limits of agreement 5.8-65). For the estimation of longest sedentary bout, the limits of agreement were unbiased but wide (mean difference ±88 minutes), varying by up to 150% of the mean estimate according to ActiGraph. Using the criterion of at least 10,000 steps per day, agreement between the Fitbit One and ActiGraph for the classification of active versus inactive was substantial (κ=.68, P<.001). The Fitbit One correctly classified 95% of days as active and 79% of days as inactive. Agreement between the Jawbone UP and ActiGraph was moderate (κ=.52, P<.001). The Jawbone UP correctly classified 90% of days as active and 80% of days as inactive. Using the criterion of at least 30 minutes/day of MVPA, agreement between the Fitbit One and ActiGraph was fair (κ=.39, P<.001). The Fitbit One correctly classified 40% of days as active and 100% of days as inactive. Agreement between the Jawbone UP and ActiGraph was slight (κ=.14, P=.001). The Jawbone UP correctly classified 100% of days as active and 12% of days as inactive.
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	The findings reported in this study suggest that both activity trackers have utility for counting steps in free-living settings, with both devices overestimating daily steps by only 5% to 15% compared with ActiGraph (Fitbit One: 8%; Jawbone UP: 14%). Both devices were less accurate measuring MVPA than steps, with correlations of .56 to .80 for both devices against ActiGraph data. Despite reasonable correlations, the Fitbit One underestimated MVPA by 46%, while the Jawbone UP overestimated MVPA by 50%. Findings indicate that the validity of the Jawbone UP measure of sedentary behavior (longest idle time) compared with ActiGraph-determined "longest sedentary bout" was poor. Both devices accurately classified more than 80% of the sample days as active or inactive based on the

	10,000 steps criterion; however, days were frequently misclassified for meeting public health guidelines of 30 minutes/day of MVPA. Due to modest accuracy and systematic bias, both activity trackers are better suited as self-monitoring tools (e.g. for the public consumer or in behavior change interventions) rather than for evaluation of research outcomes. The outcomes that relate to sedentary behavior and MVPA should be used with caution for both consumers and researchers alike.
Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: concurrent assessment of two popular brands of activity trackers on the market and two popular wear locations (wrist and waist), large number of daily observations for comparison, free-living setting which improves ecological validity and takes previous laboratory studies into a real-world setting, sample had good wear compliance, and the thorough evaluation of systematic bias Limitations: predominantly female, healthy, middle-aged sample which limits the generalizability of the findings, and the study could not control for wear time of the activity trackers which may explain some of the large absolute differences between the devices and the ActiGraph Funding source: Start-Up Grant from The University of Queensland. Drs. Gomersall and Pavey were supported by an Australian National Health and Medical Research Council (NHMRC) program grant (NHMRC no: 569940). Authors declared no conflicts of interest.

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Gomersall, S., Ng, N., Burton, N., Pavey, T., Gilson, N., & Brown, W. (2016). Estimating Physical Activity and Sedentary Behavior in a Free-Living Context: A Pragmatic Comparison of Consumer-Based Activity Trackers and ActiGraph Accelerometry. *Journal of Medical Internet Research*, *18*(9), e239. doi:10.2196/jmir.5531		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1			X	
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3		X		
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES	NO	UNCLEAR X	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1			X	
2.2 Were criteria applied equally to all study groups?	2.2	X			
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3	X			
2.4 Were the subjects/patients in a representative sample of the relevant population?	2.4			X	

The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.					
3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES	NO	UNCLEAR X	NA
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1			X	
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2			X	
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3	X			
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6	X			
4. Was method of handling withdrawals described?		YES	NO	UNCLEAR	NA X
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2				X
4.3 Were all enrolled subjects/patients (in the original sample) accounted for?	4.3		X		

This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).						
4.4 Were reasons for withdrawals similar across groups? *If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.*	4.4				X
4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5	X			
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES X	NO	UNCLEAR	NA
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1				X
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2	X			
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3				X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4				X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5	X			
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES X	NO	UNCLEAR	NA
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1				X
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2	X			
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X			
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4	X			
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5	X			
6.6 Were extra or unplanned treatments described?	6.6				X

The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.						
6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7	X				
6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8	X				
7. Were <u>outcomes</u> clearly defined and the measurements valid and reliable?		YES X	NO	UNCLEAR	NA	
7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant.*	7.1	X				
7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X	
7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X	
7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X				
7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5	X				
7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6		X			
7.7 Were the measurements conducted consistently across groups?	7.7	X				
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?		YES X	NO	UNCLEAR	NA	
8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X				
8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X				
8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X				
8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4				X	
8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)?	8.5				X	

Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.						
8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.*	8.6	X			
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7				X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X			
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X			
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X			
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X			

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral (ø) symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Gomersall, S., Ng, N., Burton, N., Pavey, T., Gilson, N., & Brown, W. (2016). Estimating Physical Activity and Sedentary Behavior in a Free-Living Context: A Pragmatic Comparison of Consumer-Based Activity Trackers and ActiGraph Accelerometry. Journal of Medical Internet Research, 18(9), e239. doi:10.2196/jmir.5531
Study Design	Validity study
Class	C
Quality Rating	☐ + (Positive) ☐ - (Negative) ☒ ⊘ (Neutral)
Research Purpose	To compare Fitbit One and Jawbone UP estimates of steps, moderate-to-vigorous physical activity (MVPA), and sedentary behavior with data from the ActiGraph GT3X+ accelerometer in a free-living context
Inclusion Criteria	Healthy, ambulatory, adult participants between 18 and 65 years of age, have accumulated less than 150 minutes of MVPA in the past week (assessed using the Active Australia Survey), and own or have access to a mobile phone compatible with both the Fitbit One and Jawbone UP
Exclusion Criteria	Unclear
Description of Study Protocol	Recruitment: Convenience sampling at three campuses of a large Australian metropolitan university via an email advertisement to staff that included study information and participant eligibility criteria Design: On two occasions for seven days each, participants wore an ActiGraph GT3X+ accelerometer on their right hip and either a hip-worn Fitbit One or wrist-worn Jawbone UP activity tracker. Daily estimates of steps and very active minutes were derived from the Fitbit One, and steps, active time, and longest idle time were derived from the Jawbone UP. Daily estimates of steps, MVPA, and longest sedentary bout were derived from the corresponding days of ActiGraph data. Blinding used (if applicable): Data were extracted via the users' accounts and entered into an Excel spreadsheet by a research assistant Intervention (if applicable): Data were collected as part of a larger, 12-week physical activity intervention study that included three groups that were randomly allocated to wear a Fitbit One, Jawbone UP, or standard pedometer. Demographic and anthropometric data were collected at

	baseline. Data for this substudy were collected at mid- and post-intervention when participants concurrently wore an ActiGraph GT3X+ accelerometer. On these two occasions for seven days each, participants wore an ActiGraph GT3X+ accelerometer on their right hip and either a hip-worn Fitbit One (n=14) or wrist-worn Jawbone UP (n=15) activity tracker. Participants were instructed to wear the devices during waking hours, removing them for water-based activities or contact sports, but were not required to keep wear logs in order to improve the free-living fidelity of the devices over the 12-week intervention. Participants were able to input activity sessions, such as swimming or contact sports, through the "log workout" function in the Jawbone UP app and the "track exercise" feature in the Fitbit app. Daily estimates of steps and very active minutes were derived from the Fitbit One, and steps, active time, and longest idle time were derived from the Jawbone UP. Daily estimates of steps, MVPA, and longest sedentary bout were derived from the corresponding days of ActiGraph data. Data were extracted via the users' accounts and entered into an Excel spreadsheet by a research assistant. Statistical Analysis: Descriptive statistics (n, mean, standard deviation, and prevalence) were calculated for demographic and physical measures. Absolute agreement was examined using intraclass correlation coefficients (ICC) and 95% confidence intervals. Correlation was assessed using Pearson correlation coefficient or Spearman rank correlation coefficient when data were non-normally distributed with 95% confidence intervals. Bland-Altman plots were used to examine the differences between all outcomes, with mean bias and 95% limits of agreement reported. Linear regression was used to examine whether mean difference and limits of agreement varied across mean values of Fitbit One or Jawbone UP and ActiGraph outcomes. Cohen's kappa statistic was used to assess the agreement between devices for classification of active versus inactive based first on achieving 10,000 steps or more per day (default step goal on both devices) and second on

	achieving 30 minutes per day or more of MVPA (comparable with public health guidelines). P values were based on two-sided tests and were considered statistically significant at P<.05. Post hoc power calculations determined that a sample size of N=289 daily comparisons would detect correlations as low as .17 with 80% power and 5% alpha.
Data Collection Summary	Timing of Measurements: Data was collected continuously as participants were instructed to wear either a Fitbit One or Jawbone UP activity tracker every day during the 12-week physical activity intervention. Daily estimates of steps, active time, and longest idle time were extracted from the users' accounts. Participants concurrently wore an ActiGraph GT3X+ accelerometer at mid- and post-intervention for seven days. Daily estimates of steps, MVPA, and longest sedentary bout were derived from the corresponding days of ActiGraph data. Dependent Variables: Steps and very active minutes according to the Fitbit One; steps, active time, and longest idle time according to the Jawbone UP Independent Variables: Free living context (e.g. participants' daily obligations, lifestyles, level of physical fitness, stress levels) Control Variables: Steps, MVPA, and longest sedentary bout according to the ActiGraph GT3X+ accelerometer
Description of Actual Data Sample	Initial: 32 (Males Females) Attrition (final N): Only 29 participants provided valid data for the current analyses, 3 males and 26 females Age: 39.6, SD 11.0 years Ethnicity: Unclear Other relevant demographics: 25 participants (86%) completed tertiary education Anthropometrics: Mean BMI: 25.9, SD: 5.0 kg/m2 Location: Australia

Summary of Results	Key Findings: Correlations for steps and MVPA were strong for both devices, although higher for Fitbit One (r=.85 for steps and ρ=.80 for MVPA) than for Jawbone UP (r=.75 for steps and ρ=.75 for MVPA). The correlation between Jawbone UP longest idle time and ActiGraph longest sedentary bout was poor (ρ=.19). Absolute agreement (ICC) was acceptable for ActiGraph and Fitbit One steps (.90) and MVPA (.72) and Jawbone UP steps (.79). However, agreement was weak between ActiGraph and Jawbone UP estimates of MVPA (.56) and longest idle time (.08). For the estimation of steps, 95% limits of agreement were unbiased for both devices, although limits were wider for Jawbone UP than for Fitbit One (5290 and 3567 steps/day). When absolute values were calculated, both devices overestimated steps (Fitbit One: mean bias 767, 95% limits of agreement –2800 to 4334; Jawbone UP: mean bias 1178, 95% limits of agreement –4112 to 6468). For the estimation of MVPA, bias was evident for both the mean difference and the limits of agreement for both the Fitbit One and the Jawbone UP. When absolute values were calculated, the Fitbit One underestimated MVPA by a mean 19.2 minutes/day (95% limits of agreement –39.2 to 5.5), whereas the Jawbone UP overestimated by a mean of 38.1 minutes/day (95% limits of agreement 5.8-65). For the estimation of longest sedentary bout, the limits of agreement were unbiased but wide (mean difference ±88 minutes), varying by up to 150% of the mean estimate according to ActiGraph. Using the criterion of at least 10,000 steps per day, agreement between the Fitbit One and ActiGraph for the classification of active versus inactive was substantial (κ=.68, P<.001). The Fitbit One correctly classified 95% of days as active and 79% of days as inactive. Agreement between the Jawbone UP and ActiGraph was moderate (κ=.52, P<.001). The Jawbone UP correctly classified 90% of days as active and 80% of days as inactive. Using the criterion of at least 30 minutes/day of MVPA, agreement between the Fitbit One and ActiGraph was fair (κ=.39, P<.001). The Fitbit One correctly classified 40% of days as active and

	100% of days as inactive. Agreement between the Jawbone UP and ActiGraph was slight (κ=.14, P=.001). The Jawbone UP correctly classified 100% of days as active and 12% of days as inactive. Other Findings:
Author Conclusion	The findings reported in this study suggest that both activity trackers have utility for counting steps in free-living settings, with both devices overestimating daily steps by only 5% to 15% compared with ActiGraph (Fitbit One: 8%; Jawbone UP: 14%). Both devices were less accurate measuring MVPA than steps, with correlations of .56 to .80 for both devices against ActiGraph data. Despite reasonable correlations, the Fitbit One underestimated MVPA by 46%, while the Jawbone UP overestimated MVPA by 50%. Findings indicate that the validity of the Jawbone UP measure of sedentary behavior (longest idle time) compared with ActiGraph-determined "longest sedentary bout" was poor. Both devices accurately classified more than 80% of the sample days as active or inactive based on the 10,000 steps criterion; however, days were frequently misclassified for meeting public health guidelines of 30 minutes/day of MVPA. Due to modest accuracy and systematic bias, both activity trackers are better suited as self-monitoring tools (e.g. for the public consumer or in behavior change interventions) rather than for evaluation of research outcomes. The outcomes that relate to sedentary behavior and MVPA should be used with caution for both consumers and researchers alike.
Reviewer Comments	*Strengths: concurrent assessment of two popular brands of activity trackers on the market and two popular wear locations (wrist and waist), large number of daily observations for comparison, free-living setting which improves ecological validity and takes previous laboratory studies into a real-world setting, sample had good wear compliance, and the thorough evaluation of systematic bias*

	Limitations: predominantly female, healthy, middle-aged sample which limits the generalizability of the findings, and the study could not control for wear time of the activity trackers which may explain some of the large absolute differences between the devices and the ActiGraph
Funding Source	Start-Up Grant from The University of Queensland. Drs. Gomersall and Pavey were supported by an Australian National Health and Medical Research Council (NHMRC) program grant (NHMRC no: 569940). Authors declared no conflicts of interest.

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive** – Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative** – Indicates that these issues have not been adequately addressed.*
⊘	***Neutral** – indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Unclear
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	No
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Unclear
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Unclear
2.2. Were criteria applied equally to all study groups?	2.2	Yes
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	Yes

2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Unclear
3. Were <u>study groups comparable</u>?	3	Unclear
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	Unclear
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	Unclear
3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	Yes
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	N/A
3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.)	3.5	N/A
3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	Yes
4. Was method of handling <u>withdrawals</u> described?	4	N/A
4.1. Were follow up methods described and the same for all groups?	4.1	N/A
4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)	4.2	N/A
4.3. Were all enrolled subjects/patients (in the original sample) accounted for?	4.3	No
4.4. Were reasons for withdrawals similar across groups	4.4	N/A
4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4.5	Yes
5. Was <u>blinding</u> used to prevent introduction of bias?	5	Yes
5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate?	5.1	N/A
5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.)	5.2	Yes
5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	5.3	N/A
5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	5.4	N/A
5.5. In diagnostic study, were test results blinded to patient history and other test results?	5.5	Yes
6. Were <u>intervention</u>/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were <u>intervening factors</u> described?	6	Yes
6.1. In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	N/A
6.2. In observational study, were interventions, study settings, and clinicians/provider described?	6.2	Yes
6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	6.3	Yes
6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured?	6.4	Yes
6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described?	6.5	Yes
	6.6	N/A

6.6. Were extra or unplanned treatments described?	6.7	Yes
6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups?	6.8	Yes
6.8. In diagnostic study, were details of test administration and replication sufficient?		
7. Were <u>outcomes</u> clearly defined and the <u>measurements valid and reliable</u>?	7	Yes
7.1. Were primary and secondary endpoints described and relevant to the question?	7.1	Yes
7.2. Were nutrition measures appropriate to question and outcomes of concern?	7.2	N/A
7.3. Was the period of follow-up long enough for important outcome(s) to occur?	7.3	N/A
7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures?	7.4	Yes
7.5. Was the measurement of effect at an appropriate level of precision?	7.5	Yes
7.6. Were other factors accounted for (measured) that could affect outcomes?	7.6	No
7.7. Were the measurements conducted consistently across groups?	7.7	Yes
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?	8	Yes
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Yes
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes
8.3. Were statistics reported with levels of significance and/or confidence intervals?	8.3	Yes
8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)?	8.4	N/A
8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)?	8.5	N/A
8.6. Was clinical significance as well as statistical significance reported?	8.6	Yes
8.7. If negative findings, was a power calculation reported to address type 2 error?	8.7	N/A
9. **Are <u>conclusions supported by results</u> with biases and limitations taken into consideration?**	9	Yes
9.1. Is there a discussion of findings?	9.1	Yes
9.2. Are biases and study limitations identified and discussed?	9.2	Yes
10. Is bias due to study's <u>funding or sponsorship</u> unlikely?	10	Yes
10.1. Were sources of funding and investigators' affiliations described?	10.1	Yes
10.2. Was there no apparent conflict of interest?	10.2	Yes

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.

NEUTRAL (∅)
If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.

PLUS/POSITIVE (+)
If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Gualtieri, L., Rosenbluth, S., & Phillips, J. (2016). Can a Free Wearable Activity Tracker Change Behavior? The Impact of Trackers on Adults in a Physician-Led Wellness Group. *JMIR Research Protocols*, *5*(4), e237. doi:10.2196/resprot.6534
Study design: *Use algorithm – RCT, cohort, etc*	Non-randomized crossover trial
Study Class (A,B,C,D)	C
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**NEUTRAL (ø)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To investigate the use of wearable activity trackers by adults with chronic medical conditions who have never used trackers previously. Specifically, the researchers aimed to determine (1) if participants would accept and use trackers to increase their physical activity; (2) if there were barriers to use besides cost and training; (3) if trackers would educate participants on their baseline and ongoing activity levels and support behavior change; and (4) if clinical outcomes would show improvements in participants' health.
Inclusion criteria: *requirements for study eligibility*	Patients had to be part of the private practice (Family Doctors, LLC), have at least one chronic medical condition, and be over 18 years of age
Exclusion criteria (conditions that make individual ineligible)	Patients who could not comprehend and speak English, or if they had advanced dementia
Recruitment	Through the Family Doctors, LLC Facebook page, brochures in the office, word of mouth, and informal mentions from staff that patients would receive a free wearable activity tracker. Cost to patients was a US $150 program fee, plus insurance co-payments.
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	NA
Description of study protocol *What happened in the study?*	This study was conducted with patients (N=10) in a 12-week physician-led wellness group offered by Family Doctors, LLC. Patients were given Withings Pulse wearable activity trackers in the second week of the wellness group and were interviewed two to four weeks after it ended. Study investigators analyzed the interview notes to extract themes about the participants' attitudes and behavior changes and collected and analyzed participants' clinical data over the course of the study.
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	The wellness group was designed as a 12-week program with two-hour meetings every week, during which patients received guidance and teaching from health experts on physical activity, nutrition, mental health, mindfulness, and sleep. At week two of the 12-week wellness group, all participants were given a new Withings Pulse, a wearable activity tracker that measures step count, calories burned, distance walked, heart rate, and sleep. Participants were given instructions developed by the research team on the setup and use of the activity tracker. Researchers assisted seven participants in setting up their devices, while the

	remaining three felt confident in setting up their devices independently. Participants were given guidance on how to select their daily step count goal. Some used the default step goal of 10,000 steps per day, while those with significant physical limitations used a goal personalized to their needs, with instructions to slowly increase their daily and weekly step count as their health permitted. In alignment with the philosophy of the wellness group, the use of activity trackers was discussed with participants as a way to build better health habits and create lifestyle change. Researchers helped troubleshoot or answer participant questions about the activity trackers during weekly meetings, by phone, and by email. All 10 patients who completed the 12-week program participated in semi-structured phone interviews, consisting of 18 open-ended questions with potential follow-up statements to encourage further responsiveness, which occurred at weeks 14, 15, and 16. Researchers recorded age, systolic blood pressure (SBP), diastolic blood pressure (DBP), low-density lipoprotein (LDL), and body weight at the start and end of the intervention.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Thematic analyses were conducted through reviews of interview notes to identify underlying themes in participant experiences. Transcripts were manually reviewed for common language and word choice, followed by multiple discussion sessions to determine significance and prevalence of themes. Paired *t* tests and *P* values were calculated, and *P* values less than .05 were considered as significant.
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	Researchers recorded age, SBP, DBP, LDL, and body weight at the start and end of the 12-week intervention. Step count, calories burned, distance walked, heart rate, and sleep data was collected continuously throughout 11 weeks of the wellness program while participants wore their Withings Pulse activity tracker. Phone interviews occurred at weeks 14, 15, and 16, and lasted approximately 30 minutes each.
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	The amount of use, acceptance, and barriers associated with the Withings Pulse. Changes in step count, calories burned, distance walked, heart rate, and sleep data. Levels of physical activity and measurements of SBP, DBP, LDL, and body weight after the 12-week intervention.
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Two-hour meetings every week, during which patients received guidance and teaching from health experts on physical activity, nutrition, mental health, mindfulness, and sleep
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	Baseline levels of physical activity and measurements of SBP, DBP, LDL, and body weight
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	11 participants
Final n (attrition) *number of subjects that completed study*	10 participants, 2 males and 8 females
Age *usually mean or range*	39 to 77 years
Ethnicity (if given)	Unclear

Other relevant demographics: *demographics describe the population (students, athletes, etc)*	Primarily lower-income patients, five (50%) worked full-time, one (10%) worked part-time, and four (40%) were retired. All patients suffered from at least one of the following chronic medical problems: overweight or obesity, hypertension, type 2 diabetes, hyperlipidemia, and joint pain. All patients stated that they were first-time activity tracker users at the onset of the group.
Anthropometrics: *e.g. were groups same or different on important physical measures (BMI, fitness level)*	All but one of the patients was overweight or obese. Baseline levels of physical activity, as assessed through patient interviews and group counseling, ranged from almost entirely sedentary to moderately active.
Location: *Where did the study take place? City or country*	Family Doctors, LLC, a private practice in a suburban community north of Boston, Massachusetts
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	Over the 11 weeks of activity tracker use, improvements were seen in clinical outcomes, physical activity behaviors, and attitudes towards the Withings Pulse. Participants lost an average of 0.5 pounds per week (SD 0.4), with a mean total weight loss of 5.97 pounds (P=.004). Other short-term clinical outcomes included a 9.2% decrease in LDL levels (P=.038). Changes in blood pressure were non-significant. All participants reported an increase in well-being, health education, physical activity, and confidence in their ability to lead more active lives. Researchers identified the following six major themes from the qualitative analysis of the post-intervention interview notes: (1) barriers to activity tracker purchase included cost, perceived value, and choice confusion; (2) attitudes towards the activity trackers shifted for many, from half of the participants expressing excitement and hope and half expressing hesitation or trepidation, to all participants feeling positive towards their tracker at the time of the interviews; (3) activity trackers served as educational tools for baseline activity levels; (4) activity trackers provided concrete feedback on physical activity, which motivated behavior change; (5) activity tracker use reinforced wellness group activities and goals; and (6) although commitment to activity tracker use did not waver, external circumstances influenced some participants' ongoing use.
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	Findings suggest that adding activity trackers to wellness groups comprising primarily older adults with chronic medical conditions can support education and behavior change to be more physically active. Barriers need to be identified and removed. In this study, the barriers to purchase included cost, perceived value, and choice confusion, which were removed by providing participants with free activity trackers. The barriers to use were removed by providing participants with initial training and ongoing support. Overall, this study demonstrated the educational benefits to individuals of learning their baseline activity levels, the increased self-efficacy arising from concrete feedback on physical activity that motivated behavior change, the positive attitudes that developed towards activity trackers, and improvements in clinical outcomes. Findings also suggest that it may be cost-effective for physicians and other health care providers to provide free or heavily subsidized trackers, along with training and support, to their patients, especially those who may most benefit from increasing their physical activity. A US $60 activity tracker that lowers the risk of chronic conditions by facilitating changes in health behaviors would be greatly beneficial compared to the health care, medication, or intervention costs required to treat illnesses after they develop.
Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: providing all participants with free Withings Pulse activity trackers, initial training, and ongoing support to minimize barriers, and the multifaceted approach to health and wellness, which encouraged participants to incorporate gradual, evidence-based changes into their lives, promoting true lifestyle change rather than "dieting" or being on an "exercise program" Limitations: *small sample size, predominantly female sample, study design lacked a control group*, *blinding was not utilized,* and the results cannot separate the impact of the wellness group education and support from that of the activity tracker use

	Funding source: sponsorship was provided by Withings, who donated trackers to RecycleHealth, a non-profit at Tufts University, who provided them to participants in the Family Doctors, LLC Wellness Group. The authors declared no conflicts of interest.

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Gualtieri, L., Rosenbluth, S., & Phillips, J. (2016). Can a Free Wearable Activity Tracker Change Behavior? The Impact of Trackers on Adults in a Physician-Led Wellness Group. *JMIR Research Protocols*, *5*(4), e237. doi:10.2196/resprot.6534		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1	X			
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	X			
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the research question clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the selection of study subjects/patients free from bias?		YES	NO	UNCLEAR X	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1	X			
2.2 Were criteria applied equally to all study groups?	2.2				X
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3	X			
2.4 Were the subjects/patients in a representative sample of the relevant population?	2.4			X	

The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.					
3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES	NO	UNCLEAR	NA X
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1		X		
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2				X
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3	X			
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6				X
4. Was method of handling withdrawals described?		YES X	NO	UNCLEAR	NA
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2	X			
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3		X		

4.4 Were reasons for withdrawals similar across groups? *If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.*	4.4					X
4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5					X
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES	NO	UNCLEAR		NA
			X			
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1		X			
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2		X			
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3					X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4					X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5					X
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES	NO	UNCLEAR		NA
		X				
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	X				
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2					X
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X				
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4	X				
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5	X				
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6					X

	6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7				X
	6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8				X
7.	**Were <u>outcomes</u> clearly defined and the measurements valid and reliable?**		YES	NO	UNCLEAR	NA
	7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant.*	7.1	X			
	7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X
	7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X
	7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X			
	7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5			X	
	7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6			X	
	7.7 Were the measurements conducted consistently across groups?	7.7				X
8.	**Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?**		YES	NO	UNCLEAR	NA
	8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1			X	
	8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X			
	8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X			
	8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4		X		
	8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)?	8.5		X		

Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.						
8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.*	8.6	X			
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7				X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X			
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X			
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X			
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X			

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral (ø) symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Gualtieri, L., Rosenbluth, S., & Phillips, J. (2016). Can a Free Wearable Activity Tracker Change Behavior? The Impact of Trackers on Adults in a Physician-Led Wellness Group. JMIR Research Protocols, 5(4), e237. doi:10.2196/resprot.6534
Study Design	Non-randomized crossover trial
Class	C
Quality Rating	☐ + (Positive) ☐ - (Negative) ☒ ⊘ (Neutral)
Research Purpose	To investigate the use of wearable activity trackers by adults with chronic medical conditions who have never used trackers previously. Specifically, the researchers aimed to determine (1) if participants would accept and use trackers to increase their physical activity; (2) if there were barriers to use besides cost and training; (3) if trackers would educate participants on their baseline and ongoing activity levels and support behavior change; and (4) if clinical outcomes would show improvements in participants' health.
Inclusion Criteria	Patients had to be part of the private practice (Family Doctors, LLC), have at least one chronic medical condition, and be over 18 years of age
Exclusion Criteria	Patients who could not comprehend and speak English, or if they had advanced dementia
Description of Study Protocol	Recruitment: Through the Family Doctors, LLC Facebook page, brochures in the office, word of mouth, and informal mentions from staff that patients would receive a free wearable activity tracker. Cost to patients was a US $150 program fee, plus insurance co-payments. Design: This study was conducted with patients (N=10) in a 12-week physician-led wellness group offered by Family Doctors, LLC. Patients were given Withings Pulse wearable activity trackers in the second week of the wellness group and were interviewed two to four weeks after it ended. Study investigators analyzed the interview notes to extract themes about the participants' attitudes and behavior changes and collected and analyzed participants' clinical data over the course of the study. Blinding used (if applicable): NA Intervention (if applicable): The wellness group was designed as a 12-week program with two-hour meetings every week, during which patients received guidance and teaching from health experts on physical activity,

	nutrition, mental health, mindfulness, and sleep. At week two of the 12-week wellness group, all participants were given a new Withings Pulse, a wearable activity tracker that measures step count, calories burned, distance walked, heart rate, and sleep. Participants were given instructions developed by the research team on the setup and use of the activity tracker. Researchers assisted seven participants in setting up their devices, while the remaining three felt confident in setting up their devices independently. Participants were given guidance on how to select their daily step count goal. Some used the default step goal of 10,000 steps per day, while those with significant physical limitations used a goal personalized to their needs, with instructions to slowly increase their daily and weekly step count as their health permitted. In alignment with the philosophy of the wellness group, the use of activity trackers was discussed with participants as a way to build better health habits and create lifestyle change. Researchers helped troubleshoot or answer participant questions about the activity trackers during weekly meetings, by phone, and by email. All 10 patients who completed the 12-week program participated in semi-structured phone interviews, consisting of 18 open-ended questions with potential follow-up statements to encourage further responsiveness, which occurred at weeks 14, 15, and 16. Researchers recorded age, systolic blood pressure (SBP), diastolic blood pressure (DBP), low-density lipoprotein (LDL), and body weight at the start and end of the intervention. Statistical Analysis: Thematic analyses were conducted through reviews of interview notes to identify underlying themes in participant experiences. Transcripts were manually reviewed for common language and word choice, followed by multiple discussion sessions to determine significance and prevalence of themes. Paired t tests and P values were calculated, and P values less than .05 were considered as significant.
Data Collection Summary	

	Timing of Measurements: Researchers recorded age, SBP, DBP, LDL, and body weight at the start and end of the 12-week intervention. Step count, calories burned, distance walked, heart rate, and sleep data was collected continuously throughout 11 weeks of the wellness program while participants wore their Withings Pulse activity tracker. Phone interviews occurred at weeks 14, 15, and 16, and lasted approximately 30 minutes each. Dependent Variables: The amount of use, acceptance, and barriers associated with the Withings Pulse. Changes in step count, calories burned, distance walked, heart rate, and sleep data. Levels of physical activity and measurements of SBP, DBP, LDL, and body weight after the 12-week intervention. Independent Variables: Two-hour meetings every week, during which patients received guidance and teaching from health experts on physical activity, nutrition, mental health, mindfulness, and sleep Control Variables: Baseline levels of physical activity and measurements of SBP, DBP, LDL, and body weight
Description of Actual Data Sample	Initial: 11 (Males Females) Attrition (final N): 10 (2 males and 8 females) Age: 39 to 77 years Ethnicity: Unclear Other relevant demographics: Primarily lower-income patients, five (50%) worked full-time, one (10%) worked part-time, and four (40%) were retired. All patients suffered from at least one of the following chronic medical problems: overweight or obesity, hypertension, type 2 diabetes, hyperlipidemia, and joint pain. All patients stated that they were first-time activity tracker users at the onset of the group. Anthropometrics: All but one of the patients was overweight or obese. Baseline levels of physical activity, as assessed through patient interviews and group counseling, ranged from almost entirely sedentary to moderately active.

	Location: Family Doctors, LLC, a private practice in a suburban community north of Boston, Massachusetts
Summary of Results	Key Findings: Over the 11 weeks of activity tracker use, improvements were seen in clinical outcomes, physical activity behaviors, and attitudes towards the Withings Pulse. Participants lost an average of 0.5 pounds per week (SD 0.4), with a mean total weight loss of 5.97 pounds (P=.004). Other short-term clinical outcomes included a 9.2% decrease in LDL levels (P=.038). Changes in blood pressure were non-significant. All participants reported an increase in well-being, health education, physical activity, and confidence in their ability to lead more active lives. Researchers identified the following six major themes from the qualitative analysis of the post-intervention interview notes: (1) barriers to activity tracker purchase included cost, perceived value, and choice confusion; (2) attitudes towards the activity trackers shifted for many, from half of the participants expressing excitement and hope and half expressing hesitation or trepidation, to all participants feeling positive towards their tracker at the time of the interviews; (3) activity trackers served as educational tools for baseline activity levels; (4) activity trackers provided concrete feedback on physical activity, which motivated behavior change; (5) activity tracker use reinforced wellness group activities and goals; and (6) although commitment to activity tracker use did not waver, external circumstances influenced some participants' ongoing use. Other Findings:
Author Conclusion	Findings suggest that adding activity trackers to wellness groups comprising primarily older adults with chronic medical conditions can support education and behavior change to be more physically active. Barriers need to be identified and removed. In this study, the barriers to purchase included cost, perceived value, and choice confusion, which were removed by providing participants with free activity trackers. The barriers to use were removed by providing participants with initial

	training and ongoing support. Overall, this study demonstrated the educational benefits to individuals of learning their baseline activity levels, the increased self-efficacy arising from concrete feedback on physical activity that motivated behavior change, the positive attitudes that developed towards activity trackers, and improvements in clinical outcomes. Findings also suggest that it may be cost-effective for physicians and other health care providers to provide free or heavily subsidized trackers, along with training and support, to their patients, especially those who may most benefit from increasing their physical activity. A US $60 activity tracker that lowers the risk of chronic conditions by facilitating changes in health behaviors would be greatly beneficial compared to the health care, medication, or intervention costs required to treat illnesses after they develop.
Reviewer Comments	*Strengths: providing all participants with free Withings Pulse activity trackers, initial training, and ongoing support to minimize barriers, and the multifaceted approach to health and wellness, which encouraged participants to incorporate gradual, evidence-based changes into their lives, promoting true lifestyle change rather than "dieting" or being on an "exercise program"* *Limitations: Limitations: small sample size, predominantly female sample, study design lacked a control group, blinding was not utilized, and the results cannot separate the impact of the wellness group education and support from that of the activity tracker use*
Funding Source	Sponsorship was provided by Withings, who donated trackers to RecycleHealth, a non-profit at Tufts University, who provided them to participants in the Family Doctors, LLC Wellness Group. The authors declared no conflicts of interest.

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive** – Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative** – Indicates that these issues have not been adequately addressed.*

⊘	***Neutral*** *– indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Yes
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	Yes
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Unclear
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Yes
2.2. Were criteria applied equally to all study groups?	2.2	N/A
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	Yes
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Unclear
3. Were <u>study groups comparable</u>?	3	N/A
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	No
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	N/A
3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	Yes
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	N/A
3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.)	3.5	N/A
3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	N/A
4. Was method of handling <u>withdrawals</u> described?	4	Yes

4.1. Were follow up methods described and the same for all groups?	4.1	N/A
4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)	4.2	Yes
4.3. Were all enrolled subjects/patients (in the original sample) accounted for?	4.3	No
4.4. Were reasons for withdrawals similar across groups	4.4	N/A
4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4.5	N/A
5. Was <u>blinding</u> used to prevent introduction of bias?	5	No
5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate?	5.1	No
5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.)	5.2	No
5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	5.3	N/A
5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	5.4	N/A
5.5. In diagnostic study, were test results blinded to patient history and other test results?	5.5	N/A
6. Were <u>intervention</u>/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were <u>intervening factors</u> described?	6	Yes
6.1. In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	Yes
6.2. In observational study, were interventions, study settings, and clinicians/provider described?	6.2	N/A
6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	6.3	Yes
6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured?	6.4	Yes
6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described?	6.5	Yes
6.6. Were extra or unplanned treatments described?	6.6	N/A
6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups?	6.7	N/A
6.8. In diagnostic study, were details of test administration and replication sufficient?	6.8	N/A
7. Were <u>outcomes</u> clearly defined and the <u>measurements valid and reliable</u>?	7	Unclear
7.1. Were primary and secondary endpoints described and relevant to the question?	7.1	Yes
7.2. Were nutrition measures appropriate to question and outcomes of concern?	7.2	N/A
7.3. Was the period of follow-up long enough for important outcome(s) to occur?	7.3	N/A
7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures?	7.4	Yes
7.5. Was the measurement of effect at an appropriate level of precision?	7.5	Unclear
7.6. Were other factors accounted for (measured) that could affect outcomes?	7.6	Unclear
7.7. Were the measurements conducted consistently across groups?	7.7	N/A
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?	8	Unclear
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Unclear
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes

8.3. Were statistics reported with levels of significance and/or confidence intervals? 8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? 8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)? 8.6. Was clinical significance as well as statistical significance reported? 8.7. If negative findings, was a power calculation reported to address type 2 error?	8.3	Yes
	8.4	No
	8.5	No
	8.6	Yes
	8.7	N/A
9. **Are <u>conclusions supported by results</u> with biases and limitations taken into consideration?** 9.1. Is there a discussion of findings? 9.2. Are biases and study limitations identified and discussed?	9	Yes
	9.1	Yes
	9.2	Yes
10. Is bias due to study's <u>funding or sponsorship</u> unlikely? 10.1. Were sources of funding and investigators' affiliations described? 10.2. Was there no apparent conflict of interest?	10	Yes
	10.1	Yes
	10.2	Yes
MINUS/NEGATIVE (-) *If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.*		
NEUTRAL (∅) *If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.*		
PLUS/POSITIVE (+) *If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.*		

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Maher, C., Ryan, J., Ambrosi, C., & Edney, S. (2017). Users' experiences of wearable activity trackers: a cross-sectional study. *BMC Public Health*, *17*, 880. doi:10.1186/s12889-017-4888-1
Study design: *Use algorithm – RCT, cohort, etc*	Cross-sectional study
Study Class (A,B,C,D)	D
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**POSITIVE (+)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To explore users' experience of activity trackers, including the perceived usefulness of devices for tracking and modifying lifestyle behaviors (physical activity, diet, and sleep), ease of use, patterns of usages, and barriers to use
Inclusion criteria: *requirements for study eligibility*	Adults over 18 years of age, living in Australia, and either currently using or have formerly used an activity tracker
Exclusion criteria (conditions that make individual ineligible)	Use of an activity tracker smartphone app without an associated wearable activity tracker, use of a fitness watch which could not measure daily steps, and activity trackers that cannot interact with a computer or smart phone
Recruitment	Promoted using low-cost distribution methods on Facebook and Twitter, primarily; sharing the survey link with a variety of Facebook community groups in the field of health and fitness (e.g. sporting clubs, cycling interest groups). Additionally, the survey link was shared on the University of South Australia's Facebook and Twitter feeds, and shared by individual members of the research team. An incentive (a $50 voucher random prize draw) was offered to encourage people to share the social media posts. A second $50 voucher random prize draw was offered for people who completed the survey
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	NA
Description of study protocol *What happened in the study?*	A cross-sectional online survey was developed and administered to Australian adults who were current or former activity tracker users
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	A purpose-designed survey instrument was developed to address the research objectives. The online survey was delivered via Survey Monkey. All participants were asked for basic demographic characteristics, including sex, age, education level, and relationship status. Participants were also asked whether they were currently using an activity tracker or had formerly used an activity tracker, and which brand of activity tracker they currently or formerly used. In addition, participants were asked how long they had worn their activity trackers, and if they were current users of an activity tracker, how long they intended to continue wearing it into the future. Three items were included to assess how participants used and shared the data derived from their activity trackers. Four items were included to assess perceived behavior change related to use of the activity tracker.

	Participants were asked to identify whether wearing their activity tracker motivated them in eight different domains, including: 'improve fitness', 'improve health', 'improve appearance', 'lose weight', 'monitor activities', 'share activities', 'compete with family or friends', and 'keep up with technology'. Up to three questions explored practical issues related to use of activity trackers. Finally, participants who had formerly worn an activity tracker were asked to select a reason why they had ceased to use it, from 10 options including reasons such as 'it broke', 'it was difficult to understand', or 'it wasn't helping with my goals', with an open-ended 'other' option included to capture additional reasons. Twenty-one items were included to assess current users' perceptions of the ease of use, usefulness, and accuracy of seven common features of activity trackers: active minutes, step counts, stair counts, sleep, heart rate, energy burned, and energy consumed. Responses were recorded on a 5-point Likert scale.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Categorical variables were analyzed using frequency of responses and percentages, and continuous variables were analyzed using medians, means, ranges, and standard deviations. Differences between former and current users were explored using independent samples t-tests, Mann-Whitney U tests, and chi square tests. Differences in use and experience on the basis of activity tracker manufacturer were also examined.
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	Data collection took place in April to May 2016
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	Users' experience of activity trackers, including the perceived usefulness of devices for tracking and modifying lifestyle behaviors (physical activity, diet, and sleep), ease of use, patterns of use, and barriers to use
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Unclear survey questions, answers that do not apply
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	NA
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	305 participants
Final n (attrition) *number of subjects that completed study*	237 participants, 69 males and 168 females
Age *usually mean or range*	18 to 74 years
Ethnicity (if given)	Unclear
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	52 (21.9%) participants completed high school, 42 (17.7%) participants completed technical and further education/certificate/diploma/apprenticeship, 95 (40.1%) participants earned an undergraduate degree, and 48 (20.3%) participants earned a postgraduate degree. 154 (65.0%) participants were in a relationship, 64 (27.0%)

	participants were single, and 19 (8.0%) participants did not specify their relationship status.
Anthropometrics: *e.g. were groups same or different on important physical measures (BMI, fitness level)*	NA
Location: *Where did the study take place? City or country*	Australia
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	Participants included 200 current and 37 former activity tracker users (total N = 237). The most commonly used brand of activity tracker was Fitbit (67.5%, n = 160), followed by Garmin (16.5%, n = 39), Apple (3.4%, n = 8), Jawbone (2.5%, n = 6), Samsung (1.7%, n = 4), Polar (1.3%, n = 3), and other (7.1%, n = 17). Participants typically used their activity trackers for sustained periods (5–7 months) and most intended to continue usage. The majority of current users either somewhat or strongly agreed that various features on their trackers were useful, including: steps (95%), active minutes (76%), sleep (66%), heart rate (63%), stairs climbed (58%), and energy burned (57%). Fewer agreed that the food intake feature was useful (36%). Participants reported they had improved their physical activity (51–81%) more commonly than they had their diet (14–40%) or sleep (11–24%), and slightly more participants reported to value the real time feedback (89%) compared to the long-term monitoring (78%). The majority of participants reported that they did not use social features (65%) nor did they share their activity data on social media platforms (77%). A chi square test was conducted to determine whether participants who shared their activity tracker data via social media reported positive behavior change more frequently than participants who did not. The results suggested that sharing data via social media was not associated with behavior change, X2 (1) = 1.07, p = .30. Overall, 94% of current users and 65% of former users agreed that they had had a positive experience using their activity tracker. A Mann-Whitney U test revealed this differed significantly, with current users more likely to report a positive experience than former users, U = 1746.50, z = −5.79, p = < .001, r = .38. Despite this, current users reported technical issues or other complaints relating to their activity trackers, most commonly relating to the tracker not suiting their outfit (19%), low battery life (19%), difficulties with the support software (17%), or perceived inaccuracy of data collected (17%). Former users reported more issues than current users overall (U = 1648.5, z = − 2.36, p = .02, r = .18). Former users were asked to identify why they no longer use their activity tracker. The main reasons given were that they felt they had learned everything they could from their tracker (30%), their tracker was broken (22%), and/or their tracker was not helping them achieve their goals (14%). Finally, analyses were performed to determine whether users' experiences and perceptions relating to activity tracker varied by brand. Only Fitbit and Garmin were included in these comparisons, since other brands had insufficient sample sizes. The perceived usefulness and accuracy of activity data did not vary between brands. However, some aspects of ease of use did vary. Fitbit users rated the stair climbing, heart rate, and dietary intake features as being significantly easier to use than Garmin users did (p = 0.01–0.02).
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	Findings suggest that in general, activity trackers are used for a substantial period of time and are viewed positively by users. Participants predominantly use their trackers to monitor and intervene on physical activity rather than other daily activities (e.g. sleep and diet) and were slightly more likely to value the trackers' real-time feedback more than long term monitoring capabilities. The majority of users perceived they had increased their physical activity as a result of using the activity tracker. Key barriers to continued use were device breakage or loss, and technical difficulties with the device or accompanying software. Findings support activity trackers as appealing and useful tools for intervening on physical activity.

Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: *large sample size, study specifically explored activity tracker users' perspectives,* survey instrument was well-designed using a rigorous process, feedback from independent experts in the field, and underwent pilot testing Limitations: *relatively high dropout rate, predominantly female sample,* study design increased the risk of recall bias, and difficulty knowing how generalizable the results are Funding source: The authors' have no funding to declare. The authors declare that they have no competing interests.

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Maher, C., Ryan, J., Ambrosi, C., & Edney, S. (2017). Users' experiences of wearable activity trackers: a cross-sectional study. *BMC Public Health*, *17*, 880. doi:10.1186/s12889-017-4888-1		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1			X	
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3		X		
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES X	NO	UNCLEAR	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1	X			
2.2 Were criteria applied equally to all study groups?	2.2	X			
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3	X			
2.4 Were the subjects/patients in a representative sample of the relevant population? *The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.*	2.4	X			

3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES X	NO	UNCLEAR	NA
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1				X
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2	X			
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3				X
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4	X			
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6				X
4. Was method of handling withdrawals described?		YES X	NO	UNCLEAR	NA
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2	X			
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3		X		
4.4 Were reasons for withdrawals similar across groups? *If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.*	4.4	X			

4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5				X
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES	NO	UNCLEAR	NA X
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1				X
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2			X	
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3			X	
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4				X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5				X
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES X	NO	UNCLEAR	NA
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1				X
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2	X			
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X			
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4				X
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5				X
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6				X
6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7	X			
6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8				X

		YES	NO	UNCLEAR	NA
7. Were <u>outcomes</u> clearly defined and the measurements valid and reliable?		X			
7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant*.	7.1	X			
7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X
7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X
7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X			
7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5	X			
7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6			X	
7.7 Were the measurements conducted consistently across groups?	7.7	X			
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?		YES X	NO	UNCLEAR	NA
8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X			
8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X			
8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X			
8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4				X
8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)? *Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.*	8.5		X		
8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical*	8.6	X			

significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.						
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7					X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA	
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X				
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X				
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA	
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X				
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X				

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral* (ø) *symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Maher, C., Ryan, J., Ambrosi, C., & Edney, S. (2017). Users' experiences of wearable activity trackers: a cross-sectional study. BMC Public Health, 17, 880. doi:10.1186/s12889-017-4888-1
Study Design	Cross-sectional study
Class	D
Quality Rating	☒ + (Positive) ☐ - (Negative) ☐ ⊘ (Neutral)
Research Purpose	To explore users' experience of activity trackers, including the perceived usefulness of devices for tracking and modifying lifestyle behaviors (physical activity, diet, and sleep), ease of use, patterns of usages, and barriers to use
Inclusion Criteria	Adults over 18 years of age, living in Australia, and either currently using or have formerly used an activity tracker
Exclusion Criteria	Use of an activity tracker smartphone app without an associated wearable activity tracker, use of a fitness watch which could not measure daily steps, and activity trackers that cannot interact with a computer or smart phone
Description of Study Protocol	Recruitment: Promoted using low-cost distribution methods on Facebook and Twitter, primarily; sharing the survey link with a variety of Facebook community groups in the field of health and fitness (e.g. sporting clubs, cycling interest groups). Additionally, the survey link was shared on the University of South Australia's Facebook and Twitter feeds, and shared by individual members of the research team. An incentive (a $50 voucher random prize draw) was offered to encourage people to share the social media posts. A second $50 voucher random prize draw was offered for people who completed the survey Design: A cross-sectional online survey was developed and administered to Australian adults who were current or former activity tracker users Blinding used (if applicable): NA Intervention (if applicable): A purpose-designed survey instrument was developed to address the research objectives. The online survey was delivered via Survey Monkey. All participants were asked for basic demographic characteristics, including sex, age, education level, and relationship status. Participants were also asked whether they were

	currently using an activity tracker or had formerly used an activity tracker, and which brand of activity tracker they currently or formerly used. In addition, participants were asked how long they had worn their activity trackers, and if they were current users of an activity tracker, how long they intended to continue wearing it into the future. Three items were included to assess how participants used and shared the data derived from their activity trackers. Four items were included to assess perceived behavior change related to use of the activity tracker. Participants were asked to identify whether wearing their activity tracker motivated them in eight different domains, including: 'improve fitness', 'improve health', 'improve appearance', 'lose weight', 'monitor activities', 'share activities', 'compete with family or friends', and 'keep up with technology'. Up to three questions explored practical issues related to use of activity trackers. Finally, participants who had formerly worn an activity tracker were asked to select a reason why they had ceased to use it, from 10 options including reasons such as 'it broke', 'it was difficult to understand', or 'it wasn't helping with my goals', with an open-ended 'other' option included to capture additional reasons. Twenty-one items were included to assess current users' perceptions of the ease of use, usefulness, and accuracy of seven common features of activity trackers: active minutes, step counts, stair counts, sleep, heart rate, energy burned, and energy consumed. Responses were recorded on a 5-point Likert scale. Statistical Analysis: Categorical variables were analyzed using frequency of responses and percentages, and continuous variables were analyzed using medians, means, ranges, and standard deviations. Differences between former and current users were explored using independent samples t-tests, Mann-Whitney U tests, and chi square tests. Differences in use and experience on the basis of activity tracker manufacturer were also examined.

Data Collection Summary	Timing of Measurements: Data collection took place in April to May 2016 Dependent Variables: Users' experience of activity trackers, including the perceived usefulness of devices for tracking and modifying lifestyle behaviors (physical activity, diet, and sleep), ease of use, patterns of use, and barriers to use Independent Variables: Unclear survey questions, answers that do not apply Control Variables: NA
Description of Actual Data Sample	Initial: 305 (Males Females) Attrition (final N): 237 (69 males and 168 females) Age: 18 to 74 years Ethnicity: Unclear Other relevant demographics: 52 (21.9%) participants completed high school, 42 (17.7%) participants completed technical and further education/certificate/diploma/apprenticeship, 95 (40.1%) participants earned an undergraduate degree, and 48 (20.3%) participants earned a postgraduate degree. 154 (65.0%) participants were in a relationship, 64 (27.0%) participants were single, and 19 (8.0%) participants did not specify their relationship status. Anthropometrics: NA Location: Australia
Summary of Results	Key Findings: Participants included 200 current and 37 former activity tracker users (total N = 237). The most commonly used brand of activity tracker was Fitbit (67.5%, n = 160), followed by Garmin (16.5%, n = 39), Apple (3.4%, n = 8), Jawbone (2.5%, n = 6), Samsung (1.7%, n = 4), Polar (1.3%, n = 3), and other (7.1%, n = 17). Participants typically used their activity trackers for sustained periods (5–7 months) and most intended to continue usage. The majority of current users either somewhat or strongly agreed that various features on their trackers were useful, including: steps (95%), active minutes (76%), sleep (66%), heart

	rate (63%), stairs climbed (58%), and energy burned (57%). Fewer agreed that the food intake feature was useful (36%). Participants reported they had improved their physical activity (51–81%) more commonly than they had their diet (14–40%) or sleep (11–24%), and slightly more participants reported to value the real time feedback (89%) compared to the long-term monitoring (78%). The majority of participants reported that they did not use social features (65%) nor did they share their activity data on social media platforms (77%). A chi square test was conducted to determine whether participants who shared their activity tracker data via social media reported positive behavior change more frequently than participants who did not. The results suggested that sharing data via social media was not associated with behavior change, X2 (1) = 1.07, p = .30. Overall, 94% of current users and 65% of former users agreed that they had had a positive experience using their activity tracker. A Mann-Whitney U test revealed this differed significantly, with current users more likely to report a positive experience than former users, U = 1746.50, z = −5.79, p = < .001, r = .38. Despite this, current users reported technical issues or other complaints relating to their activity trackers, most commonly relating to the tracker not suiting their outfit (19%), low battery life (19%), difficulties with the support software (17%), or perceived inaccuracy of data collected (17%). Former users reported more issues than current users overall (U = 1648.5, z = −2.36, p = .02, r = .18). Former users were asked to identify why they no longer use their activity tracker. The main reasons given were that they felt they had learned everything they could from their tracker (30%), their tracker was broken (22%), and/or their tracker was not helping them achieve their goals (14%). Finally, analyses were performed to determine whether users' experiences and perceptions relating to activity tracker varied by brand. Only Fitbit and Garmin were included in these comparisons, since other brands had insufficient sample sizes. The perceived usefulness and accuracy of activity data did not vary between

	brands. However, some aspects of ease of use did vary. Fitbit users rated the stair climbing, heart rate, and dietary intake features as being significantly easier to use than Garmin users did (p = 0.01–0.02). Other Findings:
Author Conclusion	Findings suggest that in general, activity trackers are used for a substantial period of time and are viewed positively by users. Participants predominantly use their trackers to monitor and intervene on physical activity rather than other daily activities (e.g. sleep and diet) and were slightly more likely to value the trackers' real-time feedback more than long term monitoring capabilities. The majority of users perceived they had increased their physical activity as a result of using the activity tracker. Key barriers to continued use were device breakage or loss, and technical difficulties with the device or accompanying software. Findings support activity trackers as appealing and useful tools for intervening on physical activity.
Reviewer Comments	*Strengths: large sample size, study specifically explored activity tracker users' perspectives, survey instrument was well-designed using a rigorous process, feedback from independent experts in the field, and underwent pilot testing* *Limitations: relatively high dropout rate, predominantly female sample, study design increased the risk of recall bias, and difficulty knowing how generalizable the results are*
Funding Source	The authors' have no funding to declare. The authors declare that they have no competing interests.

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive** – Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative** – Indicates that these issues have not been adequately addressed.*
⊘	***Neutral** – indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Unclear
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	No
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Yes
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Yes
2.2. Were criteria applied equally to all study groups?	2.2	Yes
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	Yes
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Yes
3. Were <u>study groups comparable</u>?	3	Yes
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	N/A
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	Yes
3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	N/A
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	Yes
3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.)	3.5	N/A
3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	N/A
4. Was method of handling <u>withdrawals</u> described?	4	Yes
4.1. Were follow up methods described and the same for all groups?	4.1	N/A
	4.2	Yes

4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.) 4.3. Were all enrolled subjects/patients (in the original sample) accounted for? 4.4. Were reasons for withdrawals similar across groups 4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4.3 4.4 4.5	No Yes N/A
5. Was <u>blinding</u> used to prevent introduction of bias? 5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate? 5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) 5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? 5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status? 5.5. In diagnostic study, were test results blinded to patient history and other test results?	5 5.1 5.2 5.3 5.4 5.5	N/A N/A Unclear Unclear N/A N/A
6. Were <u>intervention</u>/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were <u>intervening factors</u> described? 6.1. In RCT or other intervention trial, were protocols described for all regimens studied? 6.2. In observational study, were interventions, study settings, and clinicians/provider described? 6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? 6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured? 6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described? 6.6. Were extra or unplanned treatments described? 6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups? 6.8. In diagnostic study, were details of test administration and replication sufficient?	6 6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8	Yes N/A Yes Yes N/A N/A N/A Yes N/A
7. Were <u>outcomes</u> clearly defined and the <u>measurements valid and reliable</u>? 7.1. Were primary and secondary endpoints described and relevant to the question? 7.2. Were nutrition measures appropriate to question and outcomes of concern? 7.3. Was the period of follow-up long enough for important outcome(s) to occur? 7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? 7.5. Was the measurement of effect at an appropriate level of precision? 7.6. Were other factors accounted for (measured) that could affect outcomes? 7.7. Were the measurements conducted consistently across groups?	7 7.1 7.2 7.3 7.4 7.5 7.6 7.7	Yes Yes N/A N/A Yes Yes Unclear Yes
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators? 8.1. Were statistical analyses adequately described the results reported appropriately? 8.2. Were correct statistical tests used and assumptions of test not violated? 8.3. Were statistics reported with levels of significance and/or confidence intervals?	8 8.1 8.2 8.3	Yes Yes Yes Yes

8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? 8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)? 8.6. Was clinical significance as well as statistical significance reported? 8.7. If negative findings, was a power calculation reported to address type 2 error?	8.4 8.5 8.6 8.7	N/A No Yes N/A
9. **Are <u>conclusions supported by results</u> with biases and limitations taken into consideration?** 9.1. Is there a discussion of findings? 9.2. Are biases and study limitations identified and discussed?	9 9.1 9.2	Yes Yes Yes
10. Is bias due to study's <u>funding or sponsorship</u> unlikely? 10.1. Were sources of funding and investigators' affiliations described? 10.2. Was there no apparent conflict of interest?	10 10.1 10.2	Yes Yes Yes
MINUS/NEGATIVE (-) *If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.*		
NEUTRAL (∅) *If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.*		
PLUS/POSITIVE (+) *If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.*		

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Naslund, J., Aschbrenner, K., Scherer, E., McHugo, G., Marsch, L., & Bartels, S. (2016). Wearable Devices and Mobile Technologies for Supporting Behavioral Weight Loss Among People with Serious Mental Illness. *Psychiatry Research*, *244*, 139–144. doi:10.1016/j.psychres.2016.06.056
Study design: *Use algorithm – RCT, cohort, etc*	Non-randomized crossover trial
Study Class (A,B,C,D)	C
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**POSITIVE (+)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To examine whether average daily step count measured using Fitbit Zip wearable devices was associated with weight loss and improved fitness among individuals with serious mental illness enrolled in a 6-month lifestyle program
Inclusion criteria: *requirements for study eligibility*	Participants were age 21 or older; had serious mental illness defined by a diagnosis of schizophrenia, schizoaffective disorder, major depressive disorder, or bipolar disorder; spoke English; were on stable pharmacological treatment defined as receiving the same psychiatric medications over the prior two months; and had obesity defined as body mass index (BMI) ≥ 30
Exclusion criteria (conditions that make individual ineligible)	Participants with any medical contraindication to weight loss; were pregnant or planning to become pregnant within the next six months; or had a current diagnosis of an active alcohol-use or substance-use disorder
Recruitment	All participants were receiving services through community mental health settings
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	NA
Description of study protocol *What happened in the study?*	Participants had a schizophrenia spectrum disorder, major depressive disorder, or bipolar disorder, and wore Fitbit Zips most of the days they were enrolled in the 6-month lifestyle program
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	Participants were enrolled in a 6-month group behavioral weight loss program targeting fitness and healthy eating through an urban community mental health center. The program was modeled after the evidence-based Diabetes Prevention Program, and included weekly group sessions led by lifestyle coaches. Prior to starting the program, participants received medical clearance from a primary care provider. Participants were given Fitbit Zip wearable devices and smartphones to use for the 6-month study duration. Participants attended two brief 30-minute training sessions with a member of the research staff for instruction in using the wearable device and synching the Fitbit Zip with the associated smartphone application. Technical support for using the Fitbit Zip or associated application was provided to participants on an as needed basis by a member of the research team over the study duration. The goal-setting component of the program was personalized to meet participants' physical abilities. In general, participants shared an activity goal of reaching 150 minutes of exercise each week. As part of the

	program, participants received 3-5 text messages from research staff each week as reminders to attend optional exercise classes, to be more active as part of their daily routines, to provide encouragement, and to support participants in reaching the program's weekly physical activity goal. Daily step count data for the 6-month study duration was exported from participants' personal Fitbit accounts into an Excel spreadsheet. Fitness was measured using the 6-Minute Walk Test (6-MWT), which measures the distance in feet that an individual can walk in six minutes. Change in fitness was calculated as the change in feet on the 6-MWT from baseline to 6-months. Weight was measured and reported as the change in body weight from baseline to 6-months.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Linear regression models were used to evaluate the association between average daily step count over the 6-month study duration and the pre-post changes in participants' weight and fitness. Penalized functional regression models were used to evaluate the time-varying association between daily step count collected and the pre-post changes in participants' weight and fitness. A p-value of 0.05 was considered statistically significant.
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	Daily step count data was collected continuously throughout the 6-month study duration. Body weight and fitness were measured at baseline and after 6-months.
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	Changes in step count data according to the Fitbit Zip, changes in body weight and fitness after the 6-month intervention
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Goal-setting component of the weight loss program, weekly group sessions led by lifestyle coaches targeting fitness and healthy eating, and several weekly text messages from research staff
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	Body weight and fitness at baseline
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	43 participants
Final n (attrition) *number of subjects that completed study*	34 participants (13 males and 21 females)
Age *usually mean or range*	Mean age: 50.2 years, SD = 11.0
Ethnicity (if given)	Non-Hispanic white
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	Clinical characteristics: eight participants (23.5%) had a schizophrenia spectrum disorder, 17 participants (50.0%) had major depressive disorder, and nine participants (26.5%) had bipolar disorder. Education: two participants (5.9%) had less than a high school education, 11 participants (32.4%) attended high school, 13 participants (38.2%) attended some college, and eight participants (23.5%) earned a college degree. Living situation: 27 participants (79.4%) lived independently, six participants (17.7%) lived with family, and one participant (2.9%) lived in supported housing/assisted living. Marital status: 11 participants (32.4%) were

	never married, three participants (8.8%) were currently married, and 20 participants (58.8%) were previously married. Employment/Insurance coverage: seven participants (20.6%) were currently employed (part or full-time), eight participants (23.5%) were enrolled in Medicaid only, eight participants (23.5%) were enrolled in Medicare only, and 18 participants (52.9%) were dual eligible (enrolled in Medicaid and Medicare). Six participants (17.7%) were current smokers.
Anthropometrics: *e.g. were groups same or different on important physical measures (BMI, fitness level)*	Mean weight: 231.9 pounds, SD = 46.7 Mean BMI: 38.5 kg/m2, SD = 9.3 Fitness: 1303.8 feet in 6-MWT, SD = 323.2
Location: *Where did the study take place? City or country*	Urban community mental health center in southern New Hampshire
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	Due to rolling enrollment at the start of the study, participants had their Fitbit Zips to use for an average of 181.7 days (SD = 34.7), with a median of 181.5 days (interquartile range = 169 to 196). Participants wore their Fitbit Zips for a mean of 86.2% (SD = 18.4%) of the days that they had the Fitbit Zip to use (median = 94.0% of the days, interquartile range = 82.0% to 97.0%). Participants achieved an average of 4453.5 (SD = 2707.4) steps each day, with average daily step counts ranging from 1037.6 (SD = 767.9) steps to 11,366.3 (SD = 3416.9) steps. In total, 21 participants (61.8%) achieved 10,000 steps or more on at least one day during the study. These 21 participants achieved 10,000 daily steps or more for a mean of 16.1% (SD = 21.0%) of the days that they had the Fitbit Zip to use (median = 25.7% of the days, interquartile range = 0.9% to 25.7%). Three participants (8.8%) achieved 10,000 daily steps or more on at least half of the days that they were enrolled in the study. There was a significant association between participants' average daily step count and weight loss. For every 1000 step increase in participants' daily average step count, they experienced a decrease in weight of 1.78 pounds ($F = 5.07$; $df = 1, 32$; $p = 0.0314$). The relationship between average daily step count and change in fitness was not significant. If participants' average daily step count increased by 1000 steps, it corresponded to an increase of 18.79 feet on the 6-Minute Walk Test ($F = 1.92$; $df = 1, 31$; $p = 0.176$). In the penalized functional regression models, the time-varying relationship between daily step count and weight loss (permutation test statistic = 0.180; $p = 0.264$) and improved fitness (permutation test statistic = 0.076; $p = 0.574$) were not significant. This suggests that there was no specific period of time for which an increase in steps was significantly associated with either weight loss or improved fitness.
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	At 6-months, higher average daily step count was associated with greater weight loss, but not improved fitness. These findings suggest that wearable devices and their associated smartphone applications may serve as valuable tools for supporting community-based weight loss efforts for people with serious mental illness. Importantly, it was observed that a higher average daily step count over the 6-month program duration was associated with greater weight loss. Therefore, it appears that providing participants with serious mental illness the recommendation to collect more steps each day and maintain a high average daily step count throughout participation in a lifestyle intervention may contribute to greater weight loss. These are preliminary findings and should be interpreted with caution, but they offer promise regarding the potential benefits of using wearable devices to support lifestyle interventions delivered through community mental health settings.
Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: *long study duration,* highly engaged participants, Fitbit Zips were used to support self-monitoring, goal-setting, and tracking progress over time Limitations: *small sample size, predominantly female sample, lacking racial or ethnic diversity, relatively high dropout rate, blinding was not utilized,* analyses were based on the participants who completed the 6-month intervention, because Fitbit Zips were integrated as part of the behavioral weight loss program, the results cannot separate the impact of group education and support from use of the wearable

	device, and all participants were receiving services through community mental health settings, thus findings are likely not representative of individuals with serious mental illness not currently receiving services Funding source: This study was supported by the National Institute of Mental Health (R01 MH089811-01) and the United States Centers for Disease Control and Prevention Health Promotion and Disease Prevention Research Center (Cooperative Agreement Number U48DP005018). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors report no conflicting interests.

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Naslund, J., Aschbrenner, K., Scherer, E., McHugo, G., Marsch, L., & Bartels, S. (2016). Wearable Devices and Mobile Technologies for Supporting Behavioral Weight Loss Among People with Serious Mental Illness. *Psychiatry Research*, *244*, 139–144. doi:10.1016/j.psychres.2016.06.056		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1	X			
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	X			
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES X	NO	UNCLEAR	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1	X			
2.2 Were criteria applied equally to all study groups?	2.2				X
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3	X			
2.4 Were the subjects/patients in a representative sample of the relevant population? *The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.*	2.4			X	

3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES X	NO	UNCLEAR	NA
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1				X
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2	X			
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3	X			
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6				X
4. Was method of handling withdrawals described?		YES X	NO	UNCLEAR	NA
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2	X			
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3		X		
4.4 Were reasons for withdrawals similar across groups? *If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.*	4.4	X			

4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5				X
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES	NO	UNCLEAR	NA
			X		
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1		X		
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2		X		
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3				X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4				X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5				X
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES	NO	UNCLEAR	NA
		X			
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	X			
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2				X
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X			
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4	X			
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5	X			
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6				X
6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7				X
6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8				X

		YES	NO	UNCLEAR	NA
7. Were <u>outcomes</u> clearly defined and the measurements valid and reliable?		X			
7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant*.	7.1	X			
7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X
7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X
7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X			
7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5	X			
7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6	X			
7.7 Were the measurements conducted consistently across groups?	7.7				X
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?		YES X	NO	UNCLEAR	NA
8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X			
8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X			
8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X			
8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4		X		
8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)? *Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.*	8.5		X		
8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical*	8.6	X			

significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.						
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7					X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA	
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X				
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X				
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA	
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X				
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X				

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral* (ø) *symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Naslund, J., Aschbrenner, K., Scherer, E., McHugo, G., Marsch, L., & Bartels, S. (2016). Wearable Devices and Mobile Technologies for Supporting Behavioral Weight Loss Among People with Serious Mental Illness. Psychiatry Research, 244, 139–144. doi:10.1016/j.psychres.2016.06.056
Study Design	Non-randomized crossover trial
Class	C
Quality Rating	☒ + (Positive) ☐ - (Negative) ☐ ⊘ (Neutral)
Research Purpose	To examine whether average daily step count measured using Fitbit Zip wearable devices was associated with weight loss and improved fitness among individuals with serious mental illness enrolled in a 6-month lifestyle program
Inclusion Criteria	Participants were age 21 or older; had serious mental illness defined by a diagnosis of schizophrenia, schizoaffective disorder, major depressive disorder, or bipolar disorder; spoke English; were on stable pharmacological treatment defined as receiving the same psychiatric medications over the prior two months; and had obesity defined as body mass index (BMI) ≥ 30
Exclusion Criteria	Participants with any medical contraindication to weight loss; were pregnant or planning to become pregnant within the next six months; or had a current diagnosis of an active alcohol-use or substance-use disorder
Description of Study Protocol	Recruitment: All participants were receiving services through community mental health settings Design: Participants had a schizophrenia spectrum disorder, major depressive disorder, or bipolar disorder, and wore Fitbit Zips most of the days they were enrolled in the 6-month lifestyle program Blinding used (if applicable): NA Intervention (if applicable): Participants were enrolled in a 6-month group behavioral weight loss program targeting fitness and healthy eating through an urban community mental health center. The program was modeled after the evidence-based Diabetes Prevention Program, and included weekly group sessions led by lifestyle coaches. Prior to starting the program, participants received medical clearance from a primary care provider. Participants were given Fitbit Zip wearable devices and

	smartphones to use for the 6-month study duration. Participants attended two brief 30-minute training sessions with a member of the research staff for instruction in using the wearable device and synching the Fitbit Zip with the associated smartphone application. Technical support for using the Fitbit Zip or associated application was provided to participants on an as needed basis by a member of the research team over the study duration. The goal-setting component of the program was personalized to meet participants' physical abilities. In general, participants shared an activity goal of reaching 150 minutes of exercise each week. As part of the program, participants received 3-5 text messages from research staff each week as reminders to attend optional exercise classes, to be more active as part of their daily routines, to provide encouragement, and to support participants in reaching the program's weekly physical activity goal. Daily step count data for the 6-month study duration was exported from participants' personal Fitbit accounts into an Excel spreadsheet. Fitness was measured using the 6-Minute Walk Test (6-MWT), which measures the distance in feet that an individual can walk in six minutes. Change in fitness was calculated as the change in feet on the 6-MWT from baseline to 6-months. Weight was measured and reported as the change in body weight from baseline to 6-months. Statistical Analysis: Linear regression models were used to evaluate the association between average daily step count over the 6-month study duration and the pre-post changes in participants' weight and fitness. Penalized functional regression models were used to evaluate the time-varying association between daily step count collected and the pre-post changes in participants' weight and fitness. A p-value of 0.05 was considered statistically significant.
Data Collection Summary	Timing of Measurements: Daily step count data was collected continuously throughout the 6-month study duration. Body weight and fitness were measured at baseline and after 6-months.

	Dependent Variables: Changes in step count data according to the Fitbit Zip, changes in body weight and fitness after the 6-month intervention Independent Variables: Goal-setting component of the weight loss program, weekly group sessions led by lifestyle coaches targeting fitness and healthy eating, and several weekly text messages from research staff Control Variables: Body weight and fitness at baseline
Description of Actual Data Sample	Initial: 43 (Males Females) Attrition (final N): 34 (13 males and 21 females) Age: Mean age: 50.2 years, SD = 11.0 Ethnicity: Non-Hispanic white Other relevant demographics: Clinical characteristics: eight participants (23.5%) had a schizophrenia spectrum disorder, 17 participants (50.0%) had major depressive disorder, and nine participants (26.5%) had bipolar disorder. Education: two participants (5.9%) had less than a high school education, 11 participants (32.4%) attended high school, 13 participants (38.2%) attended some college, and eight participants (23.5%) earned a college degree. Living situation: 27 participants (79.4%) lived independently, six participants (17.7%) lived with family, and one participant (2.9%) lived in supported housing/assisted living. Marital status: 11 participants (32.4%) were never married, three participants (8.8%) were currently married, and 20 participants (58.8%) were previously married. Employment/Insurance coverage: seven participants (20.6%) were currently employed (part or full-time), eight participants (23.5%) were enrolled in Medicaid only, eight participants (23.5%) were enrolled in Medicare only, and 18 participants (52.9%) were dual eligible (enrolled in Medicaid and Medicare). Six participants (17.7%) were current smokers. Anthropometrics: Mean weight: 231.9 pounds, SD = 46.7, mean BMI: 38.5 kg/m2, SD = 9.3, fitness: 1303.8 feet in 6-MWT, SD = 323.2 Location: Urban community mental health center in southern New Hampshire

Summary of Results	Key Findings: Due to rolling enrollment at the start of the study, participants had their Fitbit Zips to use for an average of 181.7 days (SD = 34.7), with a median of 181.5 days (interquartile range = 169 to 196). Participants wore their Fitbit Zips for a mean of 86.2% (SD = 18.4%) of the days that they had the Fitbit Zip to use (median = 94.0% of the days, interquartile range = 82.0% to 97.0%). Participants achieved an average of 4453.5 (SD = 2707.4) steps each day, with average daily step counts ranging from 1037.6 (SD = 767.9) steps to 11,366.3 (SD = 3416.9) steps. In total, 21 participants (61.8%) achieved 10,000 steps or more on at least one day during the study. These 21 participants achieved 10,000 daily steps or more for a mean of 16.1% (SD = 21.0%) of the days that they had the Fitbit Zip to use (median = 25.7% of the days, interquartile range = 0.9% to 25.7%). Three participants (8.8%) achieved 10,000 daily steps or more on at least half of the days that they were enrolled in the study. There was a significant association between participants' average daily step count and weight loss. For every 1000 step increase in participants' daily average step count, they experienced a decrease in weight of 1.78 pounds ($F = 5.07$; $df = 1, 32$; $p = 0.0314$). The relationship between average daily step count and change in fitness was not significant. If participants' average daily step count increased by 1000 steps, it corresponded to an increase of 18.79 feet on the 6-Minute Walk Test ($F = 1.92$; $df = 1, 31$; $p = 0.176$). In the penalized functional regression models, the time-varying relationship between daily step count and weight loss (permutation test statistic = 0.180; $p = 0.264$) and improved fitness (permutation test statistic = 0.076; $p = 0.574$) were not significant. This suggests that there was no specific period of time for which an increase in steps was significantly associated with either weight loss or improved fitness. Other Findings:

Author Conclusion	At 6-months, higher average daily step count was associated with greater weight loss, but not improved fitness. These findings suggest that wearable devices and their associated smartphone applications may serve as valuable tools for supporting community-based weight loss efforts for people with serious mental illness. Importantly, it was observed that a higher average daily step count over the 6-month program duration was associated with greater weight loss. Therefore, it appears that providing participants with serious mental illness the recommendation to collect more steps each day and maintain a high average daily step count throughout participation in a lifestyle intervention may contribute to greater weight loss. These are preliminary findings and should be interpreted with caution, but they offer promise regarding the potential benefits of using wearable devices to support lifestyle interventions delivered through community mental health settings.
Reviewer Comments	*Strengths: long study duration, highly engaged participants, Fitbit Zips were used to support self-monitoring, goal-setting, and tracking progress over time* *Limitations: small sample size, predominantly female sample, lacking racial or ethnic diversity, relatively high dropout rate, blinding was not utilized, analyses were based on the participants who completed the 6-month intervention, because Fitbit Zips were integrated as part of the behavioral weight loss program, the results cannot separate the impact of group education and support from use of the wearable device, and all participants were receiving services through community mental health settings, thus findings are likely not representative of individuals with serious mental illness not currently receiving services*
Funding Source	This study was supported by the National Institute of Mental Health (R01 MH089811-01) and the United States Centers for Disease Control and Prevention Health Promotion and Disease Prevention Research Center (Cooperative Agreement Number U48DP005018). The funders had no

	role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors report no conflicting interests.

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive** – Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative** – Indicates that these issues have not been adequately addressed.*
⊘	***Neutral** – indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Yes
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	Yes
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Yes
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Yes
2.2. Were criteria applied equally to all study groups?	2.2	N/A
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	Yes
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Unclear
3. Were <u>study groups comparable</u>?	3	Yes
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	N/A
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	Yes

3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	Yes
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	N/A
3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.)	3.5	N/A
3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	N/A
4. Was method of handling <u>withdrawals</u> described?	4	Yes
4.1. Were follow up methods described and the same for all groups?	4.1	N/A
4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)	4.2	Yes
4.3. Were all enrolled subjects/patients (in the original sample) accounted for?	4.3	No
4.4. Were reasons for withdrawals similar across groups	4.4	Yes
4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4.5	N/A
5. Was <u>blinding</u> used to prevent introduction of bias?	5	No
5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate?	5.1	No
5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.)	5.2	No
5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	5.3	N/A
5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	5.4	N/A
5.5. In diagnostic study, were test results blinded to patient history and other test results?	5.5	N/A
6. Were <u>intervention</u>/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were <u>intervening factors</u> described?	6	Yes
6.1. In RCT or other intervention trial, were protocols described for all regimens studied?	6.1	Yes
6.2. In observational study, were interventions, study settings, and clinicians/provider described?	6.2	N/A
6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	6.3	Yes
6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured?	6.4	Yes
6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described?	6.5	Yes
6.6. Were extra or unplanned treatments described?	6.6	N/A
6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups?	6.7	N/A
6.8. In diagnostic study, were details of test administration and replication sufficient?	6.8	N/A
7. Were <u>outcomes</u> clearly defined and the <u>measurements valid and reliable</u>?	7	Yes
7.1. Were primary and secondary endpoints described and relevant to the question?	7.1	Yes
7.2. Were nutrition measures appropriate to question and outcomes of concern?	7.2	N/A

7.3. Was the period of follow-up long enough for important outcome(s) to occur?	7.3	N/A
7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures?	7.4	Yes
7.5. Was the measurement of effect at an appropriate level of precision?	7.5	Yes
7.6. Were other factors accounted for (measured) that could affect outcomes?	7.6	Yes
7.7. Were the measurements conducted consistently across groups?	7.7	N/A
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?	8	Yes
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Yes
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes
8.3. Were statistics reported with levels of significance and/or confidence intervals?	8.3	Yes
8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)?	8.4	No
8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)?	8.5	No
8.6. Was clinical significance as well as statistical significance reported?	8.6	Yes
8.7. If negative findings, was a power calculation reported to address type 2 error?	8.7	N/A
9. **Are <u>conclusions supported by results</u> with biases and limitations taken into consideration?**	9	Yes
9.1. Is there a discussion of findings?	9.1	Yes
9.2. Are biases and study limitations identified and discussed?	9.2	Yes
10. Is bias due to study's <u>funding or sponsorship</u> unlikely?	10	Yes
10.1. Were sources of funding and investigators' affiliations described?	10.1	Yes
10.2. Was there no apparent conflict of interest?	10.2	Yes

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.

NEUTRAL (∅)
If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.

PLUS/POSITIVE (+)
If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Rosenberger, M., Buman, M., Haskell, W., McConnell, M., & Carstensen, L. (2016). 24 Hours of Sleep, Sedentary Behavior, and Physical Activity with Nine Wearable Devices. *Medicine and Science in Sports and Exercise*, *48*(3), 457–465. doi:10.1249/MSS.0000000000000778
Study design: *Use algorithm – RCT, cohort, etc*	Validity study
Study Class (A,B,C,D)	C
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**NEUTRAL (ø)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To compare the output from commercially available wearable devices using current standards for objective measurement of sleep, sedentary behaviors (SED), light-intensity physical activity (LPA), moderate-to-vigorous physical activity (MVPA), and steps in a free-living environment. The ultimate goal of this research is to determine the best ways to measure the full 24 hours of activity behavior to guide future clinical studies and recommendations.
Inclusion criteria: *requirements for study eligibility*	Unclear
Exclusion criteria (conditions that make individual ineligible)	Unclear
Recruitment	The Stanford University community and surrounding areas through word-of-mouth with an effort to include equal numbers of men and women over a wide age range
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	NA
Description of study protocol *What happened in the study?*	Participants wore nine devices for 24-hours: Actigraph GT3X+, activPAL, Fitbit One, GENEactiv, Jawbone Up, LUMOback, Nike Fuelband, Omron pedometer, and Z-Machine. Comparisons (to standards) were made for total sleep time (Z-machine), time spent in SED (activPAL), LPA duration (GT3x+), MVPA duration (GT3x+), and total steps per day (Omron).
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	Participants came to the laboratory where height, weight, age, and gender were collected and recorded. Software was used to submit participant-specific information to each device for initialization and calibration. Participants also received a study kit including device supplies, written, and oral instructions of when to put on the devices and how to wear them. Participants were asked to wear all nine devices for one full day of activity and one full night of sleep. Devices were worn from approximately the time a participant woke up until the participant woke up the next morning. Device feedback was not provided to the participant except in cases where the data was presented on the device itself. No interventions were introduced such as step goals, vibrations to interrupt sedentary behavior, or other guidelines for the participant. Device data were downloaded after the participant returned the study kit. Participants could view their data after the

	conclusion of their participation if they were willing to stay through data download. No written reports were provided to the participant. Data were either downloaded to the computer (Fitbit, GT3X+, Fuelband, and activPAL) or through the phone application (LUMOback and Jawbone) for devices that lack desktop software. Devices compared to the Z-machine for measuring sleep duration included the Fitbit, Jawbone, GENEactiv, and GT3X+. Devices compared to the activPAL for measuring SED duration included the GT3X+, GENEactiv, LUMOback, and Fitbit. Devices compared to the GT3X+ for measuring LPA duration included the Fitbit and GENEactiv. Devices compared to the GT3X+ for measuring MVPA duration included the Jawbone, Fitbit, GENEactiv, and Fuelband. Devices compared to the Omron for measuring total steps included the Jawbone, Fitbit, Fuelband, GT3X+, LUMOback, and activPAL.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Standard sample calculations were conducted to set goals for recruitment, and alpha was set at .05 with the confidence interval set to 95%. Separate sample calculations were conducted for each domain. Statistical analyses, including mean absolute percent error (MAPE), equivalence testing, and Bland-Altman plots were performed to determine statistically significant differences as well as agreement among devices.
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	Height, weight, age, and gender were collected and recorded at baseline. Sleep, SED, LPA, MVPA, and steps were collected continuously throughout the 24-hour intervention. Data were downloaded after the 24-hour intervention.
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	Total sleep time according to Fitbit, Jawbone, GENEactiv, and GT3X+; time spent in SED according to GT3X+, GENEactiv, LUMOback, and Fitbit; LPA duration according to Fitbit and GENEactiv; MVPA duration according to Jawbone, Fitbit, GENEactiv, and Fuelband; and total steps per day according to Jawbone, Fitbit, Fuelband, GT3X+, LUMOback, and activPAL
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Free living conditions (e.g. participants' daily obligations, lifestyles, level of physical fitness, stress levels)
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	Total sleep time according to Z-machine; time spent in SED according to activPAL; LPA and MVPA duration according to GT3x+; and total steps per day according to Omron
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	40 participants (19 males and 21 females)
Final n (attrition) *number of subjects that completed study*	40 participants (19 males and 21 females)
Age *usually mean or range*	21 to 76 years
Ethnicity (if given)	Unclear
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	NA
Anthropometrics: *e.g. were groups same or different on*	NA

important physical measures (BMI, fitness level)	
Location: *Where did the study take place? City or country*	Stanford University community and surrounding areas (California)
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	Mean error analyses for the devices ranged from 8.1% for GT3X+ to 16.9% for GENEactiv when measuring sleep duration; 9.5% for LUMOback to 65.8% for GENEactiv when measuring SED; 19.7% for GENEactiv to 28.0% for Fitbit when measuring LPA; 51.8% from Jawbone to 92.0% from Fuelband when measuring MVPA; and 14.1% from GT3X+ to 29.9% from Fuelband when measuring total steps per day. Equivalence analyses indicated only two comparison devices were significantly equivalent to standards: GT3X+ for sleep (90% CI), and LUMOback for SED (90% CI). Bland-Altman plots had mean differences ranging from 4 minutes for GT3X+ to 36 minutes for Fitbit and GENEactiv when measuring sleep duration; 18 minutes for LUMOback to 162 minutes for GENEactiv when measuring SED; 43 minutes for GENEactiv to 64 minutes for Fitbit when measuring LPA; 48 minutes for Jawbone to 598 minutes for Fuelband when measuring MVPA; and 698 steps for GT3X+ to 2258 steps for activPAL when measuring total steps per day.
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	Findings suggest that measurement of activity domains (sleep, sedentary behavior, and physical activity) is highly varied among wearable devices when tested outside of the laboratory. While this may sound discouraging, the ability to measure very specific behaviors has greatly increased with the introduction of a large number of wearable devices. For sleep, many of the devices can measure total sleep time with the predictable error that comes from comparing actigraphy to polysomnography. For steps, many of the devices were different from the standard, but gave similar results to each other, implying some predictable agreement among devices. Currently, 24-hour activity measurement is only possible with research-grade devices. None of the commercial wearable devices provide all the measures of the 24-hour model. The future of activity measurement should aim for accurate 24-hour measurement as a goal. Researchers should continue to select measurement devices based on their primary outcomes of interest. Evaluation of devices will be an ongoing area of research because of the rapid changes in wearable technology.
Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: *zero percent attrition, the use of numerous consumer and reference devices, testing the devices in a free-living environment as they are designed for, and examining several different activity domains collected by the devices* Limitations: *inclusion and exclusion criteria and participant demographics were not discussed, blinding was not utilized,* standards were based on common field-based measures, not gold standards used in the laboratory, therefore, both the test device and criterion device introduce substantial error into the comparisons, placement of activity monitors can affect how well these devices match up to standards, and the functions of these devices change with every software and hardware update, therefore, not every possible update can be evaluated with the research at one particular point in time Funding source: Grant R37-AG008816 from the National Institute on Aging to Laura L. Carstensen. Dr. Rosenberger was a postdoctoral fellow supported on the same grant. Stanford Cardiovascular Medicine has received in-kind mobile health research support from Apple Inc. The results of this study do not constitute endorsement by the American College of Sports Medicine. The authors have no potential conflicts-of-interest to disclose.

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Rosenberger, M., Buman, M., Haskell, W., McConnell, M., & Carstensen, L. (2016). 24 Hours of Sleep, Sedentary Behavior, and Physical Activity with Nine Wearable Devices. *Medicine and Science in Sports and Exercise*, *48*(3), 457–465. doi:10.1249/MSS.0000000000000778		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1			X	
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3		X		
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3			X	
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES	NO X	UNCLEAR	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1		X		
2.2 Were criteria applied equally to all study groups?	2.2				X
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3		X		
2.4 Were the subjects/patients in a representative sample of the relevant population? *The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.*	2.4			X	

3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES	NO	UNCLEAR	NA X
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1				X
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2				X
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3	X			
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6	X			
4. Was method of handling withdrawals described?		YES	NO	UNCLEAR	NA X
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2				X
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3	X			
4.4 Were reasons for withdrawals similar across groups? *If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.*	4.4				X

4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5	X			
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES	NO	UNCLEAR	NA
				X	
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1				X
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2			X	
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3				X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4				X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5			X	
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES	NO	UNCLEAR	NA
		X			
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1				X
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2	X			
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X			
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4			X	
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5				X
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6				X
6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7				X
6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8	X			

		YES	NO	UNCLEAR	NA
7. Were <u>outcomes</u> clearly defined and the measurements valid and reliable?		X			
7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant*.	7.1	X			
7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X
7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X
7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X			
7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5	X			
7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6	X			
7.7 Were the measurements conducted consistently across groups?	7.7				X
8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?		YES X	NO	UNCLEAR	NA
8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X			
8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X			
8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X			
8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4				X
8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)? *Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.*	8.5				X
8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical*	8.6	X			

significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.						
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7					X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA	
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X				
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X				
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA	
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X				
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X				

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral* (ø) *symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Rosenberger, M., Buman, M., Haskell, W., McConnell, M., & Carstensen, L. (2016). 24 Hours of Sleep, Sedentary Behavior, and Physical Activity with Nine Wearable Devices. Medicine and Science in Sports and Exercise, 48(3), 457–465. doi:10.1249/MSS.0000000000000778
Study Design	Validity study
Class	C
Quality Rating	☐ + (Positive) ☐ - (Negative) ☒ ⊘ (Neutral)
Research Purpose	To compare the output from commercially available wearable devices using current standards for objective measurement of sleep, sedentary behaviors (SED), light-intensity physical activity (LPA), moderate-to-vigorous physical activity (MVPA), and steps in a free-living environment. The ultimate goal of this research is to determine the best ways to measure the full 24 hours of activity behavior to guide future clinical studies and recommendations.
Inclusion Criteria	Unclear
Exclusion Criteria	Unclear
Description of Study Protocol	Recruitment: The Stanford University community and surrounding areas through word-of-mouth with an effort to include equal numbers of men and women over a wide age range Design: Participants wore nine devices for 24-hours: Actigraph GT3X+, activPAL, Fitbit One, GENEactiv, Jawbone Up, LUMOback, Nike Fuelband, Omron pedometer, and Z-Machine. Comparisons (to standards) were made for total sleep time (Z-machine), time spent in SED (activPAL), LPA duration (GT3x+), MVPA duration (GT3x+), and total steps per day (Omron). Blinding used (if applicable): NA Intervention (if applicable): Participants came to the laboratory where height, weight, age, and gender were collected and recorded. Software was used to submit participant-specific information to each device for initialization and calibration. Participants also received a study kit including device supplies, written, and oral instructions of when to put on the devices and how to wear them. Participants were asked to wear all

	nine devices for one full day of activity and one full night of sleep. Devices were worn from approximately the time a participant woke up until the participant woke up the next morning. Device feedback was not provided to the participant except in cases where the data was presented on the device itself. No interventions were introduced such as step goals, vibrations to interrupt sedentary behavior, or other guidelines for the participant. Device data were downloaded after the participant returned the study kit. Participants could view their data after the conclusion of their participation if they were willing to stay through data download. No written reports were provided to the participant. Data were either downloaded to the computer (Fitbit, GT3X+, Fuelband, and activPAL) or through the phone application (LUMOback and Jawbone) for devices that lack desktop software. Devices compared to the Z-machine for measuring sleep duration included the Fitbit, Jawbone, GENEactiv, and GT3X+. Devices compared to the activPAL for measuring SED duration included the GT3X+, GENEactiv, LUMOback, and Fitbit. Devices compared to the GT3X+ for measuring LPA duration included the Fitbit and GENEactiv. Devices compared to the GT3X+ for measuring MVPA duration included the Jawbone, Fitbit, GENEactiv, and Fuelband. Devices compared to the Omron for measuring total steps included the Jawbone, Fitbit, Fuelband, GT3X+, LUMOback, and activPAL. Statistical Analysis: Standard sample calculations were conducted to set goals for recruitment, and alpha was set at .05 with the confidence interval set to 95%. Separate sample calculations were conducted for each domain. Statistical analyses, including mean absolute percent error (MAPE), equivalence testing, and Bland-Altman plots were performed to determine statistically significant differences as well as agreement among devices.
Data Collection Summary	Timing of Measurements: Height, weight, age, and gender were collected and recorded at baseline. Sleep, SED, LPA, MVPA, and steps were

	collected continuously throughout the 24-hour intervention. Data were downloaded after the 24-hour intervention. Dependent Variables: Total sleep time according to Fitbit, Jawbone, GENEactiv, and GT3X+; time spent in SED according to GT3X+, GENEactiv, LUMOback, and Fitbit; LPA duration according to Fitbit and GENEactiv; MVPA duration according to Jawbone, Fitbit, GENEactiv, and Fuelband; and total steps per day according to Jawbone, Fitbit, Fuelband, GT3X+, LUMOback, and activPAL Independent Variables: Free living conditions (e.g. participants' daily obligations, lifestyles, level of physical fitness, stress levels) Control Variables: Total sleep time according to Z-machine; time spent in SED according to activPAL; LPA and MVPA duration according to GT3x+; and total steps per day according to Omron
Description of Actual Data Sample	Initial: 40 (19 Males 21 Females) Attrition (final N): 40 Age: 21-76 years Ethnicity: Unclear Other relevant demographics: NA Anthropometrics: NA Location: Stanford University community and surrounding areas (California)
Summary of Results	Key Findings: Mean error analyses for the devices ranged from 8.1% for GT3X+ to 16.9% for GENEactiv when measuring sleep duration; 9.5% for LUMOback to 65.8% for GENEactiv when measuring SED; 19.7% for GENEactiv to 28.0% for Fitbit when measuring LPA; 51.8% from Jawbone to 92.0% from Fuelband when measuring MVPA; and 14.1% from GT3X+ to 29.9% from Fuelband when measuring total steps per day. Equivalence analyses indicated only two comparison devices were significantly equivalent to standards: GT3X+ for sleep (90% CI), and LUMOback for SED (90% CI). Bland-Altman plots had mean differences ranging from 4 minutes for GT3X+ to 36 minutes for Fitbit and

	GENEactiv when measuring sleep duration; 18 minutes for LUMOback to 162 minutes for GENEactiv when measuring SED; 43 minutes for GENEactiv to 64 minutes for Fitbit when measuring LPA; 48 minutes for Jawbone to 598 minutes for Fuelband when measuring MVPA; and 698 steps for GT3X+ to 2258 steps for activPAL when measuring total steps per day. Other Findings:
Author Conclusion	Findings suggest that measurement of activity domains (sleep, sedentary behavior, and physical activity) is highly varied among wearable devices when tested outside of the laboratory. While this may sound discouraging, the ability to measure very specific behaviors has greatly increased with the introduction of a large number of wearable devices. For sleep, many of the devices can measure total sleep time with the predictable error that comes from comparing actigraphy to polysomnography. For steps, many of the devices were different from the standard, but gave similar results to each other, implying some predictable agreement among devices. Currently, 24-hour activity measurement is only possible with research-grade devices. None of the commercial wearable devices provide all the measures of the 24-hour model. The future of activity measurement should aim for accurate 24-hour measurement as a goal. Researchers should continue to select measurement devices based on their primary outcomes of interest. Evaluation of devices will be an ongoing area of research because of the rapid changes in wearable technology.
Reviewer Comments	*Strengths: zero percent attrition, the use of numerous consumer and reference devices, testing the devices in a free-living environment as they are designed for, and examining several different activity domains collected by the devices* *Limitations: inclusion and exclusion criteria and participant demographics were not discussed, blinding was not utilized, standards*

	were based on common field-based measures, not gold standards used in the laboratory, therefore, both the test device and criterion device introduce substantial error into the comparisons, placement of activity monitors can affect how well these devices match up to standards, and the functions of these devices change with every software and hardware update, therefore, not every possible update can be evaluated with the research at one particular point in time
Funding Source	Grant R37-AG008816 from the National Institute on Aging to Laura L. Carstensen. Dr. Rosenberger was a postdoctoral fellow supported on the same grant. Stanford Cardiovascular Medicine has received in-kind mobile health research support from Apple Inc. The results of this study do not constitute endorsement by the American College of Sports Medicine. The authors have no potential conflicts-of-interest to disclose.

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*
+	***Positive** – Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative** – Indicates that these issues have not been adequately addressed.*
⊘	***Neutral** – indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Unclear
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	No
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
	1.1	Yes

1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.2	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.3	Unclear
1.3. Were the target population and setting specified?		
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	No
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	No
2.2. Were criteria applied equally to all study groups?	2.2	N/A
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	No
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Unclear
3. Were <u>study groups comparable</u>?	3	N/A
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	N/A
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	N/A
3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	Yes
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	N/A
3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.)	3.5	N/A
3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	Yes
4. Was method of handling <u>withdrawals</u> described?	4	N/A
4.1. Were follow up methods described and the same for all groups?	4.1	N/A
4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.)	4.2	N/A
4.3. Were all enrolled subjects/patients (in the original sample) accounted for?	4.3	Yes
4.4. Were reasons for withdrawals similar across groups	4.4	N/A
4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4.5	Yes
5. Was <u>blinding</u> used to prevent introduction of bias?	5	Unclear
5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate?	5.1	N/A
5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.)	5.2	Unclear
5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded?	5.3	N/A
5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status?	5.4	N/A
5.5. In diagnostic study, were test results blinded to patient history and other test results?	5.5	Unclear
	6	Yes

6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?	6.1	N/A
6.1. In RCT or other intervention trial, were protocols described for all regimens studied?	6.2	Yes
6.2. In observational study, were interventions, study settings, and clinicians/provider described?	6.3	Yes
6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect?	6.4	Unclear
6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured?	6.5	N/A
6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described?	6.6	N/A
6.6. Were extra or unplanned treatments described?	6.7	N/A
6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups?		
6.8. In diagnostic study, were details of test administration and replication sufficient?	6.8	Yes
7. Were outcomes clearly defined and the measurements valid and reliable?	7	Yes
7.1. Were primary and secondary endpoints described and relevant to the question?	7.1	Yes
7.2. Were nutrition measures appropriate to question and outcomes of concern?	7.2	N/A
7.3. Was the period of follow-up long enough for important outcome(s) to occur?	7.3	N/A
7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures?	7.4	Yes
7.5. Was the measurement of effect at an appropriate level of precision?	7.5	Yes
7.6. Were other factors accounted for (measured) that could affect outcomes?	7.6	Yes
7.7. Were the measurements conducted consistently across groups?	7.7	N/A
8. Was the statistical analysis appropriate for the study design and type of outcome indicators?	8	Yes
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Yes
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes
8.3. Were statistics reported with levels of significance and/or confidence intervals?	8.3	Yes
8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)?	8.4	N/A
8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)?	8.5	N/A
8.6. Was clinical significance as well as statistical significance reported?	8.6	Yes
8.7. If negative findings, was a power calculation reported to address type 2 error?	8.7	N/A
9. **Are conclusions supported by results with biases and limitations taken into consideration?**	9	Yes
9.1. Is there a discussion of findings?	9.1	Yes
9.2. Are biases and study limitations identified and discussed?	9.2	Yes
10. Is bias due to study's funding or sponsorship unlikely?	10	Yes
10.1. Were sources of funding and investigators' affiliations described?	10.1	Yes
10.2. Was there no apparent conflict of interest?	10.2	Yes
MINUS/NEGATIVE (-) *If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.*		
NEUTRAL (∅)		

If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.
PLUS/POSITIVE (+) *If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.*

Evidence Worksheet for Primary RESEARCH Article

Citation: *write it in AMA format as found in JADA.*	Shcherbina, A., Mattsson, C., Waggott, D., Salisbury, H., Christle, J., Hastie, T., … Ashley, E. (2017). Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine*, *7*(2), 3. doi:10.3390/jpm7020003
Study design: *Use algorithm – RCT, cohort, etc*	Validity study
Study Class (A,B,C,D)	C
Research Quality Rating *This rating tells if the research design is good (+), bad (-) or neutral (∅)* *This is determined by the quality criteria list. Delete the ratings that do not apply (i.e. if positive, delete minus/negative and neutral).*	**POSITIVE (+)**
Purpose/Population Studied/Practice Studied	
Research purpose: *What is the research question being investigated in the study?)*	To assess the accuracy of seven commercially available wrist-worn devices in estimating heart rate (HR) and energy expenditure (EE), and to propose a wearable sensor evaluation framework
Inclusion criteria: *requirements for study eligibility*	For devices: wrist-worn watch or band, continuous measurement of HR, stated battery life greater than 24 hours, commercially available direct to consumer at the time of the study, one device per manufacturer For participants: healthy adults over 18 years
Exclusion criteria (conditions that make individual ineligible)	For devices: technical problems during pre-testing For participants: unclear
Recruitment	Through advertisements within Stanford University and local amateur sports clubs. From these interested volunteers, participants were selected to maximize demographic diversity as measured by age, height, weight, body mass index (BMI), wrist circumference, and fitness level
Blinding used: *some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results*	NA
Description of study protocol *What happened in the study?*	The Apple Watch, Basis Peak, Fitbit Surge, Microsoft Band, Mio Alpha 2, PulseOn, and Samsung Gear S2 were evaluated. Participants wore devices while being simultaneously assessed with continuous electrocardiography and indirect calorimetry while sitting, walking, running, and cycling. Error in HR and EE was computed for each subject/device/activity combination.
Intervention: *Describe interventions, regimens, risk factors, or procedures studied.*	Devices were tested in two phases; the first phase included the Apple Watch, Basis Peak, Fitbit Surge, and Microsoft Band; the second phase included the MIO Alpha 2, PulseOn, and Samsung Gear S2. Participants wore up to four devices and simultaneously underwent continuous electrocardiographic monitoring and continuous indirect calorimetry using FDA approved equipment. After being fitted with all equipment, participants performed the standardized exercise protocol in a controlled laboratory setting. The exercise protocol involved five-minute intervals of sitting, walking, fast walking, running, fast running, cycling, and intense cycling. The running and cycling stages were individualized to each participants' fitness level. Data was collected according to manufacturers' instructions or by making

	use of an Application Programming Interface. The last minute of each stage was used for data analysis.
Statistical analysis: *List tests, significance level set a priori (α=0.05; include intent to treat analysis if applicable; note if there is Power analysis.*	Statistical analyses were performed separately for HR (electrocardiography served as the gold standard) and EE (indirect calorimetry served as the gold standard). Two-way ANOVA with post-hoc Turkey honest significant difference was performed to check for a difference between groups for categorical demographic covariates. A Pearson correlation test was performed between continuous demographic variables and device error. Separate tests were performed for each device, and p-values were adjusted with the Bonferroni correction for multiple testing. Principal component analysis was performed to identify outliers and to cluster devices by error profiles. A singular value decomposition was computed over the activity error rates. Several regression approaches were applied to uncover associations in the dataset, and a Bland-Altman analysis was performed to measure device error relative to the gold standards. An error rate of 5% at a p-value of 0.05 was determined to be within acceptable limits since this approximates a widely accepted standard for statistical and clinical significance.
Timing of measurements: *when outcomes were measured; usually baseline and one or more later times*	The exercise protocol involved five-minute intervals of sitting, walking, fast walking, running, and fast running until 25 minutes had passed. Participants then sat for three minutes to rest and recover, followed by five-minute intervals of cycling and intense cycling. Participants were given one minute to sit and recover, concluding the exercise protocol at 39 minutes. HR and EE data were collected continuously throughout the 39-minute intervention. The last minute of each stage was used for data analysis.
Dependent variables: *outcomes that are measured or registered; variable whose change or different states the researcher wants to understand, explain, or predict*	HR and EE according to the Apple Watch, Basis Peak, Fitbit Surge, Microsoft Band, Mio Alpha 2, PulseOn, Samsung Gear S2
Independent variables (intervention or procedure; this variable can be manipulated; a variable whose effect upon the dependent variable one is trying to understand)	Participant demographics, such as age, height, weight, BMI, wrist circumference, skin tone, fitness level
Control Variables *Examples: 1) multivariate logistic regression controlled for age, BMI, albumin; 2) usual care; 3) isocaloric diet, etc.*	HR and EE according to electrocardiography and indirect calorimetry
Initial n (e.g. 731 (298 males, 433 females)) *Record number that entered study – not the number screened.*	60 participants (29 males and 31 females)
Final n (attrition) *number of subjects that completed study*	60 participants (29 males and 31 females)
Age *usually mean or range*	21 to 64 years
Ethnicity (if given)	Unclear, diverse sample
Other relevant demographics: *demographics describe the population (students, athletes, etc)*	Skin tone: 1 to 6 (measured by Fitzpatrick skin tone scale) Fitness level: 31.7 to 66.6 mL/kg/min (measured by VO2 max)
Anthropometrics: *e.g. were groups same or different on*	Height: 154.4 to 190 cm Weight: 47.8 to 130.6 kg

important physical measures (BMI, fitness level)	BMI: 17.2 to 39.3 kg/m2 Wrist circumference: 13.5 to 21 cm
Location: *Where did the study take place? City or country*	Stanford University, California
Summary of Results: Abstract results including *quantitative data and statistics. Include statistical significance: P-values, confidence intervals (CI), relative risk (RR), odds ratios (OR), likelihood ratio, number needed to treat, power analysis if available.*	All results presented as median and 95% confidence interval. The lowest error in measuring HR was observed for the cycling stage, 1.8% (0.9%–2.7%), while the highest error was observed for the walking stage, 5.5% (3.9%–7.1%). Six devices achieved a median error below 5% in measuring HR for the cycling stage, while the Samsung Gear S2 achieved a median error rate of 5.1% (2.3%–7.9%). For the walking stage, three devices achieved a median error rate below 5%: the Apple Watch, 2.5% (1.1%–3.9%); the PulseOn, 4.9% (1.4%–8.6%); and the Microsoft Band, 5.6% (4.9%–6.3%). The remaining four devices had median error between 6.5% and 8.8%. Error in estimation of EE was considerably higher than for HR for all devices. Median error rates across activities varied from 27.4% (24.0%–30.8%) for the Fitbit Surge to 92.6% (87.5%–97.7%) for the PulseOn. For EE, the lowest relative error rates across devices were achieved for the walking (31.8% (28.6%–35.0%)) and running (31.0% (28.0%–34.0%)) stages, and the highest relative error rates across devices were achieved on the sitting stage (52.4% (48.9%–57.0%)). The Apple Watch achieved the lowest overall error in both HR and EE, while the Samsung Gear S2 reported the highest. Device error was higher for males, greater BMI, and darker skin tone.
Author's Conclusions	
Author conclusion: *paraphrase that stated by study author in body of the report or abstract*	In a diverse group of individuals, most wrist-worn devices reported HR within acceptable error range (5%) under controlled laboratory conditions of walking, running, and cycling at low and high intensities. None of the wrist-worn devices reported EE within an acceptable error range under these conditions. Across devices and modes of activities, the Apple Watch had the most favorable error profile while the Samsung Gear S2 had the least favorable error profile. Individuals and practitioners should be aware of the strengths and limitations of consumer devices that measure heart rate and estimate energy expenditure. The authors encourage transparency from device companies and consistent release of validation data to facilitate the integration of such data into clinical care.
Reviewer comments: *Note strengths and limitations of study; identify concerns that affect study validity and generalizability—your comments should be italicized)*	Strengths: *zero percent attrition, diverse sample of participants with different ages, BMI, and skin tones, the use of numerous consumer and gold standard devices, standardized exercise protocol, and examining several different activity domains collected by the devices* Limitations: *blinding was not utilized,* only consumer devices available at the time of this study were tested, and consumer devices were assessed in a controlled laboratory setting rather than in a free-living environment Funding source: unclear, the authors declare no conflict of interest

Table 3.2.a. Quality Criteria Checklist: Primary Research

RELEVANCE QUESTIONS					
Citation: *write it in AMA format as found in JADA (copy and paste from page 1 of worksheet)* Shcherbina, A., Mattsson, C., Waggott, D., Salisbury, H., Christle, J., Hastie, T., … Ashley, E. (2017). Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine*, *7*(2), 3. doi:10.3390/jpm7020003		YES	NO	UNCLEAR	NA
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (Not Applicable for some epidemiological studies)	1			X	
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/ population group would care about?	2	X			
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3		X		
4. Is the intervention or procedure feasible (NA for some epidemiological studies)?	4	X			
If the answers to all of the above relevance questions are "yes", the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.					
VALIDITY QUESTIONS					
1. Was the <u>research question</u> clearly stated? *This is usually stated at end of the introduction and just before methods section.*		YES X	NO	UNCLEAR	NA
1.1 Was the specific intervention(s) or procedure (independent variable(s)) identified? *This is often called the treatment and explained in the methods section.*	1.1	X			
1.2 Was the outcome(s) (dependent variable(s)) clearly indicated? *These are sometimes called the endpoints; the results section reports the outcomes, but this information should be in the methods section, too*	1.2	X			
1.3 Were the target population and setting specified? *The target population is group for whom findings may be applicable; look for this in the introduction and in the methods section*	1.3	X			
2. Was the <u>selection</u> of study subjects/patients free from bias?		YES X	NO	UNCLEAR	NA
2.1 Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study? *The authors should give several points about the inclusion/exclusion criteria such as the age range of the subjects, disease condition (like hyperlipidemia) required for inclusion. Exclusion criteria should be listed, too, although some are understood. For example if the ages for inclusion are 18 to 70, the authors will probably not specifically note that children and people over age 70 were excluded. Most of the time, however, subjects may be excluded for certain characteristics such as being pregnant or having some disease (like CHD).*	2.1			X	
2.2 Were criteria applied equally to all study groups?	2.2				X
2.3 Were health, demographics, and other characteristics of subjects described? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are <u>not</u> different if the P-Value is > 0.05. If there has been a previous paper describing the study population, that paper may be referenced and you would need to go back to the original publication to see that Table 1.*	2.3	X			
2.4 Were the subjects/patients in a representative sample of the relevant population?	2.4	X			

The abstractor may have to apply a bit of clinical judgment here. Authors try to be brief and may only say that the patients came from the same clinic from people who met the inclusion criteria.					
3. Were study groups comparable? *There is usually a Table 1 summarizing demographics and characteristics at baseline. Groups are not different if the P-Value is > 0.05.*		YES	NO	UNCLEAR	NA X
3.1 Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT) *In a strong study, the authors may tell how the subjects were assigned to a group (e.g. randomized block design; or assigned by computer-generated random numbers). Look for instances that show bias; for example I once read a study where patients were randomized to receive liquid energy supplements; however, if someone disliked their supplement, they were allowed to change groups – this is not unbiased!*	3.1				X
3.2 Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline? *See Table I for this - there should be no significant differences across study groups in an intervention study.*	3.2			X	
3.3 Were concurrent controls used? (Concurrent preferred over historical controls.) *Most RCTs use a concurrent control group. Occasionally an intervention study will use a prior study as a control group; that is an example of a historical control. That is not as strong a research design as use of concurrent control group. A crossover study where the subject acts as his/her own control is use of concurrent control.*	3.3	X			
3.4 If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis? *The groups in a cohort or cross-sectional study should not be different from each other; if they are, a strong study will utilize statistical techniques such as multivariate analyses to remove the variance due to the group differences. Look for this information in the statistics and results sections.*	3.4				X
3.5 If case control study, were potential confounding factors comparable for cases and controls? If case series or trial with subjects serving as own control, this criterion is not applicable. Criterion may not be applicable in some cross-sectional studies. *Subjects are generally matched for age, gender, etc. Look for this in the statistical description and results sections.*	3.5				X
3.6 If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g. "gold standard")? *Example: comparing body fat analysis method with underwater weighing (gold standard). In studies trying to determine the best equation (like Mifflin-St. Jeor or Harris-Benedict) to predict energy needs, a gold standard measure of REE (Indirect Calorimetry) is used.*	3.6	X			
4. Was method of handling withdrawals described?		YES	NO	UNCLEAR	NA X
4.1 Were follow up methods described and the same for all groups?	4.1				X
4.2 Was the number, characteristics of withdrawals (i.e. dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80 %.) *This should be found in the results section. If there is attrition > 20%, it is important to note that on the worksheet (as a note in the results section or in the reviewer comments at the very bottom)*	4.2				X
4.3 Were all enrolled subjects/patients (in the original sample) accounted for? *This information is often presented in a figure with # recruited, # enrolled (this is initial N), # remaining at end of study period (final N). Sometimes the reasons that subjects withdrew or were dropped is given in the figure or in the text (results section).*	4.3	X			

4.4 Were reasons for withdrawals similar across groups? *If there is a large attrition from one group and not others, you would want to look for a reason why; the answer to this question would then be no.*	4.4				X
4.5 If diagnostic test, was decision to perform reference test not dependent on results of test under study? *The test under study should be compared to reference test all the time. An example of this might be using a DEXA machine to measure percent body fat only if a subject's BMI was > 35 but bioimpedance analyzer indicated body fat < 30%.*	4.5	X			
5. **Was <u>blinding</u> used to prevent introduction of bias?**		YES	NO	UNCLEAR	NA
				X	
5.1 In intervention study, were subjects, clinicians/practitioners and investigators blinded to treatment group, **<u>as appropriate</u>**? *The key term is as appropriate. For example, in the Lim et al 2008 study, the investigators studied the effect of MNT on lipid levels in hypercholesterolemic patients. It was an RCT, but obviously, the subjects and practitioners knew who was getting MNT and who was not. Therefore, you would not answer question 5.1 NO. It was appropriate for the dietitians and patients to know they were receiving MNT.*	5.1				X
5.2 Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.2			X	
5.3 In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? *Answer yes, if a lab test was used to measure an outcome. A method of blinding a diet study is to have separate people analyzing the data (not the same ones who were collecting the data).*	5.3				X
5.4 In case control study, was case definition explicit and case ascertainment not influenced by exposure status? *Establish who is a case and who is a control at the beginning of the study.*	5.4				X
5.5 In diagnostic study, were test results blinded to patient history and other test results?	5.5			X	
6. Were intervention/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were intervening factors described?		YES	NO	UNCLEAR	NA
		X			
6.1 In RCT or other intervention trial, were protocols described for all regimens studied?	6.1				X
6.2 In observational study, were interventions, study settings, and clinicians/provider described?	6.2	X			
6.3 Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? *Use clinical judgment (e.g. 12 weeks is long enough for a dietary intervention to make a difference in lab values for cholesterol; however, 12 days would not be long enough)*	6.3	X			
6.4 Was the amount of exposure and, if relevant, subject/patient compliance measured? *How long did the treatment last? Did the patient follow directions?*	6.4	X			
6.5 Were co-interventions (e.g., ancillary treatments other therapies) described? *(e.g. were patients on lipid-lowering meds at the same time as the diet therapy)*	6.5				X
6.6 Were extra or unplanned treatments described? *The text may not describe any unplanned treatments. If yes, it would likely be in the discussion section. It is likely there were no unplanned treatments, so a "no" answer is not a problem overall.*	6.6				X

	6.7 Was the information for 6.4, 6.5, 6.6 and 6.7 assessed the same way for all groups? *For a study to be valid and unbiased, it is important that this be yes.*	6.7				X
	6.8 In diagnostic study, were details of test administration and replication sufficient? *Usually answer n/a for diet study.*	6.8	X			
7.	**Were <u>outcomes</u> clearly defined and the measurements valid and reliable?**		YES X	NO	UNCLEAR	NA
	7.1 Were primary and secondary endpoints described and relevant to the question? *Primary endpoint –main result measured at the end of a study to see if the treatment worked. The primary endpoint is decided at the beginning of the study.* *Secondary endpoint - not as important as the main results; not usually analyzed if the primary endpoint is not statistically significant*.	7.1	X			
	7.2 Were nutrition measures appropriate to question and outcomes of concern? *Clinical judgment required: weight loss, changes in energy intake are relevant to MNT; Sometimes there are no nutrition measures and you should answer N/A.*	7.2				X
	7.3 Was the period of follow-up long enough for important outcome(s) to occur? *Clinical judgment required: was there enough time?*	7.3				X
	7.4 Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? *Check that surveys were validated.*	7.4	X			
	7.5 Was the measurement of effect at an appropriate level of precision? *Precision is reproducibility or repeatability.*	7.5	X			
	7.6 Were other factors accounted for (measured) that could affect outcomes? *Other factors are sometimes covered in the discussion of the strengths/limitations of the study.*	7.6	X			
	7.7 Were the measurements conducted consistently across groups?	7.7				X
8.	**Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?**		YES X	NO	UNCLEAR	NA
	8.1 Were statistical analyses adequately described and the results reported appropriately? *There should be a discussion of the statistics in the methods section.*	8.1	X			
	8.2 Were correct statistical tests used and assumptions of test not violated? *You will get better at this the more papers you abstract.EAL abstractors are expected to have some statistical and research training (minimum of master's degree).*	8.2	X			
	8.3 Were statistics reported with levels of significance and/or confidence intervals? *(P-value)* and/or confidence intervals *(mean ± CI)*	8.3	X			
	8.4 Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)? Intent to treat– analysis is based on the original treatment intent, not the treatment ultimately administered (i.e. does not matter if treatment was for 2, 6, 8 or all the weeks in the study). The analyses are done using all the subjects in the study, not just the ones who completed it. This is done in order to avoid effects of dropout that can be a threat to randomization. *Intent-to-treat analysis of outcomes applies to any intervention study.* If intent to treat analysis was done, it will be mentioned in the statistical section. If all subjects who began the trial completed it, intent-to-treat analysis was done.	8.4				X
	8.5 Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g. multivariate analyses)? *Multivariate analyses are used to adjust or control for other variables (age, sex, smoking, etc). Assumes data is valid and reduces a larger number of variables to a smaller number. Just answer yes or no that multivariate analyses were used.*	8.5	X			

8.6 Was clinical significance as well as statistical significance reported? *Example: Lim, et al 2008 reported that after 12 weeks of MNT, total cholesterol was reduced from 229.2±158 to 181.3±16.3 (P<0.001); This includes: statistical significance (P-value) and clinical significance (compare to standard of < 200 mg/do for normal cholesterol). A problem can occur when only statistical significance is reported. Reducing cholesterol from 300 to 250 might be statistically significant, but clinically it is still abnormal.*	8.6	X			
8.7 If negative findings, was a power calculation reported to address type 2 error? *Type II (β error is a false negative that happens when the investigators fail to reject the null hypothesis when the null hypothesis is false. Look for the authors to say something like "a sample size of n=xx is needed to provide 80% power."*	8.7				X
9. **Are conclusions supported by results with biases and limitations taken into consideration?**		YES X	NO	UNCLEAR	NA
9.1 Is there a discussion of findings? *Answer yes or no.*	9.1	X			
9.2 Are biases and study limitations identified and discussed? *This will be in the discussion of finding section that follows the results*	9.2	X			
10. Is bias due to study's funding or sponsorship unlikely? *Be careful here – if bias is unlikely, answer YES.*		YES X	NO	UNCLEAR	NA
10.1 Were sources of funding and investigators' affiliations described? • *Look just under the abstract, or* • *The funding may be acknowledged at the end of the paper* • *Just because the work was funded by industry does not mean the study was biased.*	10.1	X			
10.2 Was there no apparent conflict of interest? *If an investigator is testing a piece of equipment, process or drug that s/he developed, it could be a conflict of interest.*	10.2	X			

SYMBOL

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "no," the report should be designated with a minus (-) symbol on the Evidence Quality Worksheet.

NEUTRAL (ø)
*If the answers to validity criteria questions **2, 3, 6, and 7** do not indicate that the study is exceptionally strong, the report should be designated with a neutral* (ø) *symbol on the Evidence Quality Worksheet.*

PLUS/POSITIVE (+)
*If most of the answers to the above validity questions are "Yes" including criteria **2, 3, 6, and 7 and at least one additional "yes"**, (the report should be designated with a plus symbol (+) on the Evidence Quality Worksheet.*

Citation	Shcherbina, A., Mattsson, C., Waggott, D., Salisbury, H., Christle, J., Hastie, T., … Ashley, E. (2017). Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. Journal of Personalized Medicine, 7(2), 3. doi:10.3390/jpm7020003
Study Design	Validity study
Class	C
Quality Rating	☒ + (Positive) ☐ - (Negative) ☐ ⊘ (Neutral)
Research Purpose	To assess the accuracy of seven commercially available wrist-worn devices in estimating heart rate (HR) and energy expenditure (EE), and to propose a wearable sensor evaluation framework
Inclusion Criteria	For devices: wrist-worn watch or band, continuous measurement of HR, stated battery life greater than 24 hours, commercially available direct to consumer at the time of the study, one device per manufacturer For participants: healthy adults over 18 years
Exclusion Criteria	For devices: technical problems during pre-testing For participants: unclear
Description of Study Protocol	Recruitment: Through advertisements within Stanford University and local amateur sports clubs. From these interested volunteers, participants were selected to maximize demographic diversity as measured by age, height, weight, body mass index (BMI), wrist circumference, and fitness level Design: The Apple Watch, Basis Peak, Fitbit Surge, Microsoft Band, Mio Alpha 2, PulseOn, and Samsung Gear S2 were evaluated. Participants wore devices while being simultaneously assessed with continuous electrocardiography and indirect calorimetry while sitting, walking, running, and cycling. Error in HR and EE was computed for each subject/device/activity combination. Blinding used (if applicable): NA Intervention (if applicable): Devices were tested in two phases; the first phase included the Apple Watch, Basis Peak, Fitbit Surge, and Microsoft Band; the second phase included the MIO Alpha 2, PulseOn, and Samsung Gear S2. Participants wore up to four devices and simultaneously underwent continuous electrocardiographic monitoring

	and continuous indirect calorimetry using FDA approved equipment. After being fitted with all equipment, participants performed the standardized exercise protocol in a controlled laboratory setting. The exercise protocol involved five-minute intervals of sitting, walking, fast walking, running, fast running, cycling, and intense cycling. The running and cycling stages were individualized to each participants' fitness level. Data was collected according to manufacturers' instructions or by making use of an Application Programming Interface. The last minute of each stage was used for data analysis. Statistical Analysis: Statistical analyses were performed separately for HR (electrocardiography served as the gold standard) and EE (indirect calorimetry served as the gold standard). Two-way ANOVA with post-hoc Turkey honest significant difference was performed to check for a difference between groups for categorical demographic covariates. A Pearson correlation test was performed between continuous demographic variables and device error. Separate tests were performed for each device, and p-values were adjusted with the Bonferroni correction for multiple testing. Principal component analysis was performed to identify outliers and to cluster devices by error profiles. A singular value decomposition was computed over the activity error rates. Several regression approaches were applied to uncover associations in the dataset, and a Bland-Altman analysis was performed to measure device error relative to the gold standards. An error rate of 5% at a p-value of 0.05 was determined to be within acceptable limits since this approximates a widely accepted standard for statistical and clinical significance.
Data Collection Summary	Timing of Measurements: The exercise protocol involved five-minute intervals of sitting, walking, fast walking, running, and fast running until 25 minutes had passed. Participants then sat for three minutes to rest and recover, followed by five-minute intervals of cycling and intense cycling. Participants were given one minute to sit and recover, concluding the

	exercise protocol at 39 minutes. HR and EE data were collected continuously throughout the 39-minute intervention. The last minute of each stage was used for data analysis. Dependent Variables: HR and EE according to the Apple Watch, Basis Peak, Fitbit Surge, Microsoft Band, Mio Alpha 2, PulseOn, Samsung Gear S2 Independent Variables: Participant demographics, such as age, height, weight, BMI, wrist circumference, skin tone, fitness level Control Variables: HR and EE according to electrocardiography and indirect calorimetry
Description of Actual Data Sample	Initial: 60 (29 Males 31 Females) Attrition (final N): 60 Age: 21-64 years Ethnicity: Unclear, diverse sample Other relevant demographics: Skin tone: 1 to 6 (measured by Fitzpatrick skin tone scale), fitness level: 31.7 to 66.6 mL/kg/min (measured by VO2 max) Anthropometrics: Height: 154.4 to 190 cm, weight: 47.8 to 130.6 kg, BMI: 17.2 to 39.3 kg/m2, wrist circumference: 13.5 to 21 cm Location: Stanford University, California
Summary of Results	Key Findings: All results presented as median and 95% confidence interval. The lowest error in measuring HR was observed for the cycling stage, 1.8% (0.9%–2.7%), while the highest error was observed for the walking stage, 5.5% (3.9%–7.1%). Six devices achieved a median error below 5% in measuring HR for the cycling stage, while the Samsung Gear S2 achieved a median error rate of 5.1% (2.3%–7.9%). For the walking stage, three devices achieved a median error rate below 5%: the Apple Watch, 2.5% (1.1%–3.9%); the PulseOn, 4.9% (1.4%–8.6%); and the Microsoft Band, 5.6% (4.9%–6.3%). The remaining four devices had median error between 6.5% and 8.8%. Error in estimation of EE was considerably higher than for HR for all devices. Median error rates across

	activities varied from 27.4% (24.0%–30.8%) for the Fitbit Surge to 92.6% (87.5%–97.7%) for the PulseOn. For EE, the lowest relative error rates across devices were achieved for the walking (31.8% (28.6%–35.0%)) and running (31.0% (28.0%–34.0%)) stages, and the highest relative error rates across devices were achieved on the sitting stage (52.4% (48.9%–57.0%)). The Apple Watch achieved the lowest overall error in both HR and EE, while the Samsung Gear S2 reported the highest. Device error was higher for males, greater BMI, and darker skin tone. Other Findings:
Author Conclusion	In a diverse group of individuals, most wrist-worn devices reported HR within acceptable error range (5%) under controlled laboratory conditions of walking, running, and cycling at low and high intensities. None of the wrist-worn devices reported EE within an acceptable error range under these conditions. Across devices and modes of activities, the Apple Watch had the most favorable error profile while the Samsung Gear S2 had the least favorable error profile. Individuals and practitioners should be aware of the strengths and limitations of consumer devices that measure heart rate and estimate energy expenditure. The authors encourage transparency from device companies and consistent release of validation data to facilitate the integration of such data into clinical care.
Reviewer Comments	*Strengths: zero percent attrition, diverse sample of participants with different ages, BMI, and skin tones, the use of numerous consumer and gold standard devices, standardized exercise protocol, and examining several different activity domains collected by the devices* *Limitations: blinding was not utilized, only consumer devices available at the time of this study were tested, and consumer devices were assessed in a controlled laboratory setting rather than in a free-living environment*
Funding Source	Unclear, the authors declare no conflict of interest

Quality Criteria Checklist: Primary Research

Symbols Used	*Explanation*

+	***Positive*** *– Indicates that the report has clearly addressed issues of inclusion/exclusion, bias, generalizability, and data collection and analysis*
--	***Negative*** *– Indicates that these issues have not been adequately addressed.*
⊘	***Neutral*** *– indicates that the report is neither exceptionally strong nor exceptionally week*

Select a rating from the drop-down menu ↓

Relevance Questions		
1. Would implementing the studied intervention or procedure (if found successful) result in improved outcomes for the patients/clients/population group? (NA for some Epi studies)	1	Unclear
2. Did the authors study an outcome (dependent variable) or topic that the patients/clients/population group would care about?	2	Yes
3. Is the focus of the intervention or procedure (independent variable) or topic of study a common issue of concern to dietetics practice?	3	No
4. Is the intervention or procedure feasible? (NA for some epidemiological studies)	4	Yes
If the answers to all of the above relevance questions are "Yes," the report is eligible for designation with a plus (+) on the Evidence Quality Worksheet, depending on answers to the following validity questions.		
Validity Questions		
1. Was the <u>research question</u> clearly stated?	1	Yes
1.1. Was the specific intervention(s) or procedure (independent variable(s)) identified?	1.1	Yes
1.2. Was the outcome(s) (dependent variable(s)) clearly indicated?	1.2	Yes
1.3. Were the target population and setting specified?	1.3	Yes
2. Was the <u>selection</u> of study subjects/patients free from bias?	2	Yes
2.1. Were inclusion/exclusion criteria specified (e.g., risk, point in disease progression, diagnostic or prognosis criteria), and with sufficient detail and without omitting criteria critical to the study?	2.1	Unclear
2.2. Were criteria applied equally to all study groups?	2.2	N/A
2.3. Were health, demographics, and other characteristics of subjects described?	2.3	Yes
2.4. Were the subjects/patients a representative sample of the relevant population?	2.4	Yes
3. Were <u>study groups comparable</u>?	3	N/A
3.1. Was the method of assigning subjects/patients to groups described and unbiased? (Method of randomization identified if RCT)	3.1	N/A
3.2. Were distribution of disease status, prognostic factors, and other factors (e.g., demographics) similar across study groups at baseline?	3.2	Unclear
3.3. Were concurrent controls used? (Concurrent preferred over historical controls.)	3.3	Yes
3.4. If cohort study or cross-sectional study, were groups comparable on important confounding factors and/or were preexisting differences accounted for by using appropriate adjustments in statistical analysis?	3.4	N/A
3.5. If case control study, were potential confounding factors comparable for cases and controls? (If case series or trial with subjects serving as own control, this	3.5	N/A

criterion is not applicable. Criterion may not be applicable in some cross-sectional studies.) 3.6. If diagnostic test, was there an independent blind comparison with an appropriate reference standard (e.g., "gold standard")?	3.6	Yes
4. Was method of handling <u>withdrawals</u> described? 4.1. Were follow up methods described and the same for all groups? 4.2. Was the number, characteristics of withdrawals (i.e., dropouts, lost to follow up, attrition rate) and/or response rate (cross-sectional studies) described for each group? (Follow up goal for a strong study is 80%.) 4.3. Were all enrolled subjects/patients (in the original sample) accounted for? 4.4. Were reasons for withdrawals similar across groups 4.5. If diagnostic test, was decision to perform reference test not dependent on results of test under study?	4 4.1 4.2 4.3 4.4 4.5	N/A N/A N/A Yes N/A Yes
5. Was <u>blinding</u> used to prevent introduction of bias? 5.1. In intervention study, were subjects, clinicians/practitioners, and investigators blinded to treatment group, as appropriate? 5.2. Were data collectors blinded for outcomes assessment? (If outcome is measured using an objective test, such as a lab value, this criterion is assumed to be met.) 5.3. In cohort study or cross-sectional study, were measurements of outcomes and risk factors blinded? 5.4. In case control study, was case definition explicit and case ascertainment not influenced by exposure status? 5.5. In diagnostic study, were test results blinded to patient history and other test results?	5 5.1 5.2 5.3 5.4 5.5	Unclear N/A Unclear N/A N/A Unclear
6. Were <u>intervention</u>/therapeutic regimens/exposure factor or procedure and any comparison(s) described in detail? Were <u>intervening factors</u> described? 6.1. In RCT or other intervention trial, were protocols described for all regimens studied? 6.2. In observational study, were interventions, study settings, and clinicians/provider described? 6.3. Was the intensity and duration of the intervention or exposure factor sufficient to produce a meaningful effect? 6.4. Was the amount of exposure and, if relevant, subject/patient compliance measured? 6.5. Were co-interventions (e.g., ancillary treatments, other therapies) described? 6.6. Were extra or unplanned treatments described? 6.7. Was the information for 6.4, 6.5, and 6.6 assessed the same way for all groups? 6.8. In diagnostic study, were details of test administration and replication sufficient?	6 6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8	Yes N/A Yes Yes Yes N/A N/A N/A Yes
7. Were <u>outcomes</u> clearly defined and the <u>measurements valid and reliable</u>? 7.1. Were primary and secondary endpoints described and relevant to the question? 7.2. Were nutrition measures appropriate to question and outcomes of concern? 7.3. Was the period of follow-up long enough for important outcome(s) to occur? 7.4. Were the observations and measurements based on standard, valid, and reliable data collection instruments/tests/procedures? 7.5. Was the measurement of effect at an appropriate level of precision? 7.6. Were other factors accounted for (measured) that could affect outcomes? 7.7. Were the measurements conducted consistently across groups?	7 7.1 7.2 7.3 7.4 7.5 7.6 7.7	Yes Yes N/A N/A Yes Yes Yes N/A

8. Was the <u>statistical analysis</u> appropriate for the study design and type of outcome indicators?	8	Yes
8.1. Were statistical analyses adequately described the results reported appropriately?	8.1	Yes
8.2. Were correct statistical tests used and assumptions of test not violated?	8.2	Yes
8.3. Were statistics reported with levels of significance and/or confidence intervals?	8.3	Yes
8.4. Was "intent to treat" analysis of outcomes done (and as appropriate, was there an analysis of outcomes for those maximally exposed or a dose-response analysis)?	8.4	N/A
8.5. Were adequate adjustments made for effects of confounding factors that might have affected the outcomes (e.g., multivariate analyses)?	8.5	Yes
8.6. Was clinical significance as well as statistical significance reported?	8.6	Yes
8.7. If negative findings, was a power calculation reported to address type 2 error?	8.7	N/A
9. **Are <u>conclusions supported by results</u> with biases and limitations taken into consideration?**	9	Yes
9.1. Is there a discussion of findings?	9.1	Yes
9.2. Are biases and study limitations identified and discussed?	9.2	Yes
10. Is bias due to study's <u>funding or sponsorship</u> unlikely?	10	Yes
10.1. Were sources of funding and investigators' affiliations described?	10.1	Yes
10.2. Was there no apparent conflict of interest?	10.2	Yes

MINUS/NEGATIVE (-)
If most (six or more) of the answers to the above validity questions are "No," the report should be designated with a minus (-) symbol on the Evidence Worksheet.

NEUTRAL (∅)
If the answers to validity criteria questions 2, 3, 6, and 7 do not indicate that the study is exceptionally strong, the report should be designated with a neutral (∅) symbol on the Evidence Worksheet.

PLUS/POSITIVE (+)
If most of the answers to the above validity questions are "Yes" (including criteria 2, 3, 6, 7 and at least one additional "Yes"), the report should be designated with a plus symbol (+) on the Evidence Worksheet.

APPENDIX B: OVERVIEW TABLE

Author, Year, Study Design, Class, Rating	Study Purpose	Study Population	Intervention	Outcomes	Limitations
Author: Benedetto et al. Year: 2018 Study Design: Validity study Class: C Rating: +	To assess in a controlled, research environment the accuracy and precision of the Fitbit Charge 2 for measuring heart rate (HR) with respect to a gold standard electrocardiograph	15 (7 males and 8 females) healthy, Caucasian adult participants Age: 25 to 36 years Weight: 56 to 82 kg Height: 155 to 185 cm BMI: 20 to 25 kg/m2	Participants rode a stationary bike for 10 minutes with the stated goal to raise their HR as much as possible. Participants' HR was simultaneously recorded from the Fitbit Charge 2 and ProComp Infiniti T7500M	The Fitbit Charge 2 exhibited a mean bias of -5.9 bpm (95% CI). The limits of agreement (LoA) between the Fitbit Charge 2 and ProComp Infiniti T7500M were wide. The upper LoA was +16.8 bpm, whereas the lower LoA was -28.5 bpm. The intraclass correlation coefficients (ICC) between the Fitbit Charge 2 and ProComp Infiniti T7500M was 0.21 (95% CI).	Small sample size, participant recruitment was not discussed, blinding was not utilized, possible unstable positioning of the Fitbit Charge 2, lacking a defined activity pattern for the participants to simulate low, medium, and intensive exercise, and lacking a variety of participants
Author: Cadmus-Bertram et al. Year: 2017 Study Design: Validity study Class: C Rating: ø	To determine the accuracy of the heart rate measured by four commercial, light-emitting diode–dependent, wrist-worn activity trackers (Basis Peak, Fitbit Charge, Fitbit Surge, Mio Fuse)	40 (20 males and 20 females) healthy adult participants Age: 30 to 65 years Mean BMI: 25.1 kg/m2	Participants wore two activity trackers on each wrist and were connected to an electrocardiograph. Participants sat and rested for 10 minutes, then exercised on a treadmill at 65% of their maximum heart rate for 10 minutes while their heart rates were measured.	The Fitbit Surge had the best LoA (−5.1 to 4.5 beats/min) while the Basis Peak had the worst LoA (−17.1 to 22.6 beats/min) while resting. When participants exercised, the LoA were relatively poor for all the activity trackers (Mio Fuse, −22.5 to 26.0 beats/min; Basis Peak, −27.1 to 29.2 beats/min; Fitbit Surge, −34.8 to 39.0 beats/min; and Fitbit Charge, −41.0 to 36.0 beats/min)	Participant recruitment, demographics, and location of study were not discussed, and blinding was not utilized
Author: Cadmus-Bertram et al. Year: 2015 Study Design:	To evaluate the feasibility and efficacy of integrating a Fitbit tracker and website into a physical	51 (0 males and 51 females) postmenopausal, overweight or obese women	Participants were randomized to either a Fitbit or pedometer-based intervention	After the 16-week intervention, the Fitbit group increased MVPA by 62±108 min/week (p<.001), MVPA in 10-min bouts by 38±83	Small sample size, short duration, and lack of generalizability since participants were all postmenopausal,

Randomized controlled trial Class: A Rating: +	activity intervention for postmenopausal, overweight or obese women	Age: 53 to 67 years BMI: 25.7 to 32.7 kg/m2	group to determine whether the Fitbit One increased physical activity more than the pedometer	min/week (p=0.008), and steps by 789±1,979 (p=0.01), compared to non-significant increases in the pedometer group (between-group p-values: 0.11, 0.28 and 0.30).	overweight or obese women and there were several confounders such as the goal setting process, four-week follow-up call, and Fitbit website
Author: Chum et al. Year: 2017 Study Design: Randomized controlled trial Class: A Rating: ø	To understand patients' perceived benefit from the Fitbit One and explore themes associated with patient experiences. To compare perceived benefit, patient factors, Fitbit usage, and Beck's Depression Inventory (BDI) scores.	36 (18 males and 18 females) participants with major depressive disorder (MDD) Mean age: 53 years Mean BDI score: 36.27 (scale of 0-63) Mean BMI: 30.16 kg/m2	Fitbit One activity trackers were provided to all patients allocated to the behavioral activation (BA) group at the beginning of the BRAVE study. Patients were encouraged to carry the Fitbit One at all times throughout 18 weeks to track their physical activity. Interviews were conducted with 36 patients who completed the BRAVE study.	23 patients found the Fitbit One to be helpful for their physical activity. Themes of positive experiences included self-awareness, peer motivation, and goal-setting opportunities. Themes of negative experiences included inconvenience, inaccuracies, discouragement, and disinterest. There was a significant relationship between total Fitbit One usage and perceived benefit. The mean number of weeks of Fitbit One use for those who found the Fitbit helpful was 18.57 and 12.27 weeks for those who did not (p<0.001).	Small sample size compared to the number of participants who completed the BRAVE study, difficulty interpreting exploratory findings given that the study was underpowered to test effectiveness, and there were several confounders such as combined BA intervention, study setting, and participants' restricted use of Fitbit
Author: Cook et al. Year: 2017 Study Design: Validity study Class: C Rating: +	To evaluate the utility of the Fitbit Flex (FBF) to estimate sleep in adult patients with MDD relative to gold standard polysomnography (PSG) and validated actigraphy (Actiwatch-2; AW-2).	21 (4 males and 17 females) unmedicated participants with MDD Mean age: 26.5 ± 4.6 years Mean BDI-II score: 22.9 ± 6.8	Patients wore the FBF and AW-2 during in-laboratory PSG. The following sleep variables were calculated: total sleep time (TST), sleep onset latency (SOL), wake after sleep onset (WASO), and sleep efficiency (SE).	Compared to PSG, the FBF significantly overestimated TST (mean difference of 46.0 min, p<0.0001) and SE (mean difference of 8.1%, p<0.0001), while significantly underestimating WASO (mean difference of −44.0 min, p<0.0001). SOL assessed by FBF and PSG were quite similar (mean difference of −2.0 min, p=0.72). The FBF showed a high sensitivity (0.98 ±	Participants were young to middle aged and predominantly female. Study specifically examined outpatients with MDD, limiting the generalizability of findings. Results cannot be extended to other fitness trackers, or more current generations of the same model as these devices may have different

				0.02) and accuracy (0.88 ± 0.05), with low specificity (0.35 ± 0.13).	performance characteristics.
Author: Ferguson et al. Year: 2015 Study Design: Validity study Class: C Rating: +	To assess the concurrent validity of a selection of consumer-level accelerometer-based activity monitors (Fitbit One, Fitbit Zip, Jawbone UP, Misfit Shine, Nike Fuelband, Striiv Smart Pedometer, and Withings Pulse) compared to two research-grade accelerometers (BodyMedia SenseWear, and ActiGraph GT3X+) in free-living conditions	21 (10 males and 11 females) healthy adult participants Age: 20 to 59 years Male BMI: 27.3 ± 3.2 kg/m2 Female BMI: 25.5 ± 5.2 kg/m2	Participants wore seven activity monitors and two research-grade accelerometers simultaneously for 48-hours. Participants went about their daily life in free-living conditions. Four physical activity parameters were measured, including step count, moderate to vigorous physical activity (MVPA), sleep, and total daily energy expenditure (TDEE).	All activity monitors measured steps, and correlations with reference devices were very strong (r = 0.94-0.99). Five activity monitors measured MVPA, and correlations ranged from weak to strong (r = 0.52-0.91). Four activity monitors measured sleep, and all correlated strongly with the reference device (r = 0.82-0.92). Five activity monitors measured TDEE, and correlations were moderate to strong (r = 0.74- 0.81). The Fitbit One, Fitbit Zip, and Withings Pulse were the strongest performers.	Participant recruitment was not discussed, blinding was not utilized, and validity may vary if activity monitors are worn in locations other than the hip or wrist
Author: Gomersall et al. Year: 2016 Study Design: Validity study Class: C Rating: ø	To compare Fitbit One and Jawbone UP estimates of steps, MVPA, and sedentary behavior with data from the ActiGraph GT3X+ accelerometer in a free-living context	32 (3 males and 26 females) healthy adult participants Mean age: 39.6, SD: 11.0 years Mean BMI: 25.9, SD: 5.0 kg/m2	On two occasions for seven days each, participants wore an ActiGraph GT3X+ on their hip and either a hip-worn Fitbit One or wrist-worn Jawbone UP. Daily estimates of steps, MVPA, and longest sedentary time were measured.	Correlations for steps and MVPA were strong for both devices, although higher for the Fitbit One (r=.85 for steps and ρ=.80 for MVPA) than for Jawbone UP (r=.75 for steps and ρ=.75 for MVPA). The correlation between the Jawbone UP longest idle time and ActiGraph longest sedentary bout was poor (ρ=.19). Agreement between the Fitbit One and ActiGraph for the classification of active versus inactive was substantial (P<.001), while	Predominantly female, healthy, middle-aged sample which limits the generalizability of the findings, and the study could not control for wear time of the activity trackers

				agreement between the Jawbone UP and ActiGraph was moderate (P<.001).	
Author: Gualtieri et al. Year: 2016 Study Design: Non-randomized crossover trial Class: C Rating: ø	To determine (1) if participants would accept and use activity trackers to increase their physical activity; (2) if there were barriers to use besides cost and training; (3) if activity trackers would educate participants on their activity levels and support behavior change; and (4) if clinical outcomes would show improvements in participants' health	10 (2 males and 8 females) adult participants with one chronic medical condition Age: 39 to 77 years	Patients were given Withings Pulse wearable activity trackers in the physician-led wellness group and were interviewed two to four weeks after the 12-week study. Themes about participants attitudes and behavior changes were analyzed along with participants' clinical data over the course of the study.	Improvements were seen in clinical outcomes, physical activity behaviors, and attitudes towards the Withings Pulse. Participants lost an average of 0.5 pounds per week, with a mean total weight loss of 5.97 pounds (P=.004). Other clinical outcomes included a 9.2% decrease in LDL levels (P=.038). Changes in blood pressure were non-significant. All participants reported an increase in well-being, health education, physical activity, and confidence in their ability to lead more active lives.	Small sample size, predominantly female sample, study design lacked a control group, blinding was not utilized, and the results cannot separate the impact of the wellness group education and support from that of the activity tracker use
Author: Maher et al. Year: 2017 Study Design: Cross-sectional study Class: D Rating: +	To explore users' experience of activity trackers, including the perceived usefulness of devices for tracking and modifying lifestyle behaviors (physical activity, diet, and sleep), ease of use, patterns of use, and barriers to use	237 (69 males and 168 females) adult participants who were current or former activity tracker users Age: 18 to 74 years	A cross-sectional online survey was developed to address the research objectives and was administered via Survey Monkey to adults who were current or former activity tracker users	The most commonly used brand of activity tracker was Fitbit (67.5%), followed by Garmin (16.5%), Apple (3.4%), Jawbone (2.5%), Samsung (1.7%), Polar (1.3%), and other (7.1%). Participants agreed that various features on their trackers were useful, including: steps (95%), active minutes (76%), sleep (66%), heart rate (63%), stairs climbed (58%), and energy burned (57%), while fewer agreed that the food intake feature was useful (36%). Overall, 94% of current users and	Relatively high dropout rate, predominantly female sample, study design increased the risk of recall bias, and difficulty knowing how generalizable the results are

				65% of former users agreed that they had had a positive experience using their activity tracker.	
Author: Naslund et al. Year: 2016 Study Design: Non-randomized crossover trial Class: C Rating: +	To examine whether average daily step count measured using Fitbit Zip wearable devices was associated with weight loss and improved fitness among individuals with serious mental illness enrolled in a 6-month lifestyle program	34 (13 males and 21 females) non-Hispanic white, obese, adult participants receiving services for schizophrenia spectrum disorder, MDD, or bipolar disorder Mean age: 50.2 years, SD = 11.0 Mean weight: 231.9 lbs Mean BMI: 38.5 kg/m2 Fitness: 1303.8 feet in 6-Minute Walk Test (6-MWT)	Participants wore Fitbit Zips most of the days they were enrolled in the 6-month group behavioral weight loss program. Participants' weight and change in fitness was measured at baseline and 6 months. Daily step count data was extracted from participants' Fitbits Zips	Participants achieved an average of 4453.5 steps each day, with average daily step counts ranging from 1037.6 to 11,366.3 steps. There was a significant association between participants' average daily step count and weight loss. For every 1000 step increase, participants experienced a decrease in weight of 1.78 pounds ($p = 0.0314$). The relationship between average daily step count and change in fitness was not significant (increase of 18.79 feet on the 6-MWT ($p = 0.176$)).	Small sample size, predominantly female sample, lacking racial or ethnic diversity, relatively high dropout rate, blinding was not utilized, analyses were based on participants who completed the study, results cannot separate the impact of group education and support from the use of Fitbit Zips, and findings are likely not representative of individuals with serious mental illness not currently receiving services
Author: Rosenberger et al. Year: 2016 Study Design: Validity study Class: C Rating: ø	To compare the output from commercially available wearable devices using current standards for objective measurement of sleep, sedentary behaviors (SED), light-intensity physical activity (LPA), (MVPA), and steps in a free-living environment.	40 (19 males and 21 females) adult participants Age: 21 to 76 years	Participants wore nine devices for 24-hours: Actigraph GT3X+, activPAL, Fitbit One, GENEactiv, Jawbone Up, LUMOback, Nike Fuelband, Omron pedometer, and Z-Machine. Comparisons (to standards) were made for total sleep time (Z-machine), time spent in SED (activPAL), LPA duration	Mean error analyses for the devices ranged from 8.1% for GT3X+ to 16.9% for GENEactiv when measuring sleep duration; 9.5% for LUMOback to 65.8% for GENEactiv when measuring SED; 19.7% for GENEactiv to 28.0% for Fitbit when measuring LPA; 51.8% from Jawbone to 92.0% from Fuelband when measuring MVPA; and 14.1% from GT3X+ to 29.9% from Fuelband when measuring total steps per day. Equivalence analyses indicated	Inclusion and exclusion criteria and participant demographics were not discussed, blinding was not utilized, standards were based on common field-based measures rather than gold standards used in the laboratory, placement of activity monitors can affect how well these devices match up to standards, and the results cannot be extended to other fitness trackers, or more current

			(GT3x+), MVPA duration (GT3x+), and total steps per day (Omron).	only one comparison device, the LUMOback was significantly equivalent to standards for SED (90% CI).	generations of the same model as these devices may have different hardware and software updates.
Author: Shcherbina et al. Year: 2017 Study Design: Validity study Class: C Rating: +	To assess the accuracy of seven commercially available wrist-worn devices (Apple Watch, Basis Peak, Fitbit Surge, Microsoft Band, Mio Alpha 2, PulseOn, and Samsung Gear S2) in estimating heart rate (HR) and energy expenditure (EE), and to propose a wearable sensor evaluation framework	60 (29 males and 31 females) diverse, healthy adult participants Age: 21 to 64 years Height: 154.4 to 190 cm Weight: 47.8 to 130.6 kg BMI: 17.2 to 39.3 kg/m2 Skin tone: 1 to 6 (measured by Fitzpatrick skin tone scale) Fitness level: 31.7 to 66.6 mL/kg/min (measured by VO2 max)	Devices were evaluated in two phases. Participants wore up to four devices while being simultaneously assessed with continuous electrocardiography and indirect calorimetry while sitting, walking, running, and cycling. Error in HR and EE was computed for each device/activity combination.	The lowest error in measuring HR was observed for the cycling stage, 1.8% (0.9%–2.7%, 95% CI), while the highest error was observed for the walking stage, 5.5% (3.9%–7.1%, 95% CI). Error in estimation of EE was considerably higher than for HR for all devices. Median error rates across activities varied from 27.4% (24.0%–30.8%, 95% CI) for the Fitbit Surge to 92.6% (87.5%–97.7%, 95% CI) for the PulseOn. The Apple Watch achieved the lowest overall error, while the Samsung Gear S2 reported the highest.	Blinding was not utilized, only consumer devices available at the time of th study were tested, and consumer devices were assessed in a controlled laboratory setting rather than in a free-living environment

APPENDIX C: COMPARISON REPORT

Product	**Apple Watch Series 4**  https://www.bhphotovideo.com/c/product/1434890-REG/apple_mtug2ll_a_watch_series_4_gps.html	**Fitbit Charge 3**  https://www.walmart.com/ip/Fitbit-Charge-3-Advanced-Heart-Rate-Fitness-Tracker/654994366	**Fitbit Versa**  https://www.amazon.com/Fitbit-Versa-Smart-Aluminium-Included/dp/B07B48SQGT?th=1	**Garmin Vivosmart 4**  https://www.amazon.com/Garmin-v%C3%ADvosmart-Activity-Fitness-Midnight/dp/B07GM7WHBG	**Garmin Vivosport**  https://www.clevertraining.com/garmin-vivosport-gps-activity-tracker
Release Date	September 2018	October 2018	April 2018	September 2018	August 2017
Price	$399	$149.95	$199.95	$129.99	$169.99
Wear Site	Wrist	Wrist	Wrist	Wrist	Wrist
Compatibility	iOS	Android, iOS, Windows	Android, iOS, Windows	Android, iOS	Android, iOS
Display	OLED	OLED	Color LCD	OLED	Color LCD
Battery	18 hours	7 days	4+ days	7 days	7 days
Water Resistant	Yes	Yes	Yes	Yes	Yes
Functions					
Steps	Yes	Yes	Yes	Yes	Yes
Distance	Yes	Yes	Yes	Yes	Yes
Elevation/ Stairs	Yes	Yes	Yes	Yes	Yes
Heart Rate	Yes	Yes	Yes	Yes	Yes
Calories Burned	Yes	Yes	Yes	Yes	Yes
Active Time	Yes	Yes	Yes	Yes	Yes
Sleep Time	Yes	Yes	Yes	Yes	Yes
Sleep Quality	Yes	Yes	Yes	Yes	Yes
Other	Apps, music, GPS, notifications, goal setting, exercise modes, coaching, activity sharing, guided breathing	Apps, notifications, goal setting, exercise modes, activity sharing, guided breathing	Apps, music, notifications, goal setting, exercise modes, coaching, activity sharing, guided breathing	Apps, music, notifications, goal setting, VO2 max, pulse ox, exercise modes, activity sharing, guided breathing	Apps, music, GPS, notifications, goal setting, VO2 max, exercise modes, activity sharing

Product	Misfit Shine 2	Moov Now	Samsung Gear Fit2	Withings Steel HR Sport	Xiaomi Mi Band 3
	 https://www.amazon.com/Misfit-Shine-Fitness-Tracker-Monitor/dp/B01AURE4UE	 https://www.amazon.com/Fitness-Tracker-Audio-Coach-Moov/dp/B01CX26IEO	 https://www.samsung.com/global/galaxy/gear-fit2/	 https://www.smartwatchspex.com/withings-steel-hr-sport-specifications/	 https://www.amazon.co.uk/Wristband-Fitness-incoming-waterproof-forecast/dp/B07DJ9J1QZ
Release Date	November 2015	November 2015	June 2016	September 2018	May 2018
Price	$79.99	$49.99	$179.99	$199.95	$29.99
Wear Site	Wrist	Wrist	Wrist	Wrist	Wrist
Compatibility	Android, iOS, Windows	Android, iOS	Android, iOS	Android, iOS	Android
Display	12 color LED lights	None	AMOLED	Analog dial, subdial, OLED	OLED
Battery	6 months	6 months	3+ days	25 days	20 days
Water Resistant	Yes	Yes	Yes	Yes	Yes
Functions					
Steps	Yes	Yes	Yes	Yes	Yes
Distance	Yes	Yes	Yes	Yes	Yes
Elevation/ Stairs	No	Yes	Yes	No	No
Heart Rate	No	Yes	Yes	Yes	Yes
Calories Burned	Yes	Yes	Yes	Yes	Yes
Active Time	Yes	Yes	Yes	Yes	Yes
Sleep Time	Yes	Yes	Yes	Yes	Yes
Sleep Quality	Yes	Yes	Yes	Yes	Yes
Other	Music, notifications, goal setting, activity sharing	Exercise modes, coaching	Music, GPS, notifications, goal setting, exercise modes, activity sharing	Notifications, VO2 max, exercise modes	Notifications, goal setting, exercise modes, activity sharing